A Memory Efficient Unified Algorithm for Online Learning of Linear Dynamical Systems

Yuval Ran-Milo* Tel Aviv University

Angelos Assos* Princeton University

Elad Hazan Princeton University

Abstract

Motivated by the challenge of stabilizing a *general* unknown linear dynamical system (LDS) from observations, we study the natural prerequisite of online prediction. Our goal is to achieve sublinear regret with a memory footprint that adapts to the intrinsic complexity of the dynamics rather than the full hidden-state dimension. We focus on the practically central regime of systems with low *instability complexity*—eigenvalues outside the real stable interval that do not decay rapidly, together with non-semisimple modes—potentially embedded in an otherwise stable real spectrum of much higher dimension; we write k for this count. This regime is the primary setting in which stabilization is plausible: we show that many systems with high instability complexity cannot be stabilized without exponentially large controls. Thus, prediction is meaningful for stabilization precisely when the instability complexity is small. Within this regime, we introduce a unified online algorithm that handles every LDS (including non-diagonalizable systems with complex or exploding modes) with a learnable parameter count of $\widetilde{O}(k)$. Finally, we prove a lower bound showing that k is a valid complexity measure: any filter-based predictor needs at least k filters. Experiments corroborate our theory: on a high-dimensional system, our predictor sharply outperforms prior methods at an equal parameter budget.

1 Introduction

A long-standing goal in control theory is to *stabilize* an unknown dynamical system from observations: design an input policy that keeps the system in a bounded region, even when the open-loop dynamics are unstable [Doyle et al., 1989, Tsiamis et al., 2022, Foster and Simchowitz, 2020, Agarwal et al., 2019, Abbasi-Yadkori and Szepesvári, 2011, Faradonbeh et al., 2018, Simchowitz et al., 2018, Sarkar and Rakhlin, 2019]. Prediction is a natural prerequisite for stabilization: an algorithm that cannot track the system's output trajectory has little basis for choosing inputs that would bring it under control. We study this prediction problem for linear dynamical systems (LDS) in the online learning framework: an algorithm observes inputs and outputs sequentially and at each step must predict the next output before seeing it, achieving sublinear regret against the best LDS predictor in hindsight, with no knowledge of the system parameters [Hazan et al., 2017, Hardt et al., 2019, Kuznetsov and Mohri, 2018, Tsiamis and Pappas, 2020].

A central challenge—and the main contribution of this work—is achieving sublinear regret with a memory footprint that adapts to the intrinsic complexity of the dynamics rather than the full hidden-state dimension. We focus on a regime that is both practically relevant and theoretically interesting: systems with few modes that are unstable, non-semisimple, or not rapidly decaying. Such systems arise naturally—a physical plant with mild resonance, a single unstable direction embedded in a large stable state space, or a controlled process with a few runaway modes—yet prior work offers no prediction guarantees that adapt to this structure. Moreover, as we show in Section 5, a broad

*Equal contribution. Correspondence to: <yuvalmilo@mail.tau.ac.il> and <aa2279@princeton.edu>.

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

class of systems with many exploding modes relative to the number of input channels is essentially impossible to keep bounded in practice (Proposition 9). This makes the small-k regime more than a natural special case: it delineates the setting in which one can practically hope to maintain the system under control.

Existing predictors fall into three useful families, each suited to a different regime. Spectral filtering [Hazan et al., 2017] achieves the strongest known guarantees for stable systems: a parameter footprint polylogarithmic in T and completely independent of the state dimension d. Spectral filtering, however, is limited to systems whose state matrix is diagonalizable with all eigenvalues in $(-1, 1)$, a provably significant restriction [Ran-Milo et al., 2024]. Finite-memory input filters (FIR) are also independent of the state dimension, but apply only to systems whose modes all decay rapidly. Autoregressive (AR) predictors [Ho and Kalman, 1966, Kailath, 1980], and system identification methods [Overschee, 1996, Ljung, 1999] can represent any observable LDS exactly, but at the cost of memory proportional to d—even if only one out of d modes is non-stable. None of these tools alone handles the mixed case: a system with $d - k$ easily filtered modes and $k \ll d$ hard ones.

We give a single, *unified* algorithm (Algorithm 1) that closes this gap, simultaneously achieving what neither prior family could alone. Our algorithm handles every LDS—including non-diagonalizable systems and systems with complex or exploding modes—and achieves sublinear regret with a learnable parameter count of $\widetilde{O}(k)$, where k is the instability complexity (Definition 1). It thus combines the two benefits that no prior method enjoyed simultaneously: the *generality* of autoregressive predictors and the *dimension-free memory footprint* of spectral and finite-memory filtering. Constructing a stabilizing controller from the learned predictor is left to future work.

Throughout, we study system-input pairs that keep the dynamics under control: although the open-loop system may be unstable, the chosen input sequence keeps the resulting outputs bounded. This is the regime in which prediction is useful for learning control; once the realized trajectory has already exploded, the rollout has left the setting where predictions can be used to keep the system stabilized.

Finally, in Section 5 we show that the *instability complexity* k is a natural and well-motivated parameter from both a theoretical and practical perspective. On the theoretical side, we prove a lower bound showing that the dependence on k is unavoidable: any filter-based predictor with bounded coefficient mass needs at least k filter directions to achieve sublinear regret uniformly over the LDS class (Proposition 11). On the practical side, we show that, in a broad class of systems, many exploding modes make stabilization impractical: with high probability over a uniformly random first nonzero input direction, every stabilizing continuation must span an exponentially large range of magnitudes—its input dynamic range (Definition 8) grows exponentially in k (Proposition 9)—making such controllers acutely sensitive to noise, actuator limits, and finite precision. These results motivate the regime where k is small compared with d, making our algorithm highly efficient.

We complement these results with experiments (Section 6) on a high-dimensional system with small instability complexity. Given the same number of learned parameters, the unified predictor attains orders-of-magnitude lower error than spectral filtering, finite-memory, and autoregressive predictors, confirming that it captures the few hard modes without paying for the full state dimension, unlike existing methods.

2 Related Work

Learning stable dynamical systems. The cleanest online learning guarantees for dynamical systems are known in stable or non-explosive regimes. General sequential forecasting work studies prediction under temporal dependence and non-stationarity without relying on i.i.d. assumptions [Kuznetsov and Mohri, 2018]. For linear dynamical systems, spectral filtering [Hazan et al., 2017] gives a particularly relevant result: stable LDS outputs can be predicted online using a small universal family of filters, with parameter count independent of the hidden-state dimension. Hazan and Marsden [2026] extend spectral filtering to systems with complex-valued eigenvalues, although their regret scales linearly with the sequence length. Ghai et al. [2020] extend no-regret prediction guarantees to marginally stable systems. Related work on learning Kalman filters gives regret and sample-complexity guarantees for partially observed linear systems under stochastic and non-explosive assumptions [Tsiamis and Pappas, 2020, Zhang et al., 2023]. More recently, Dogariu et al. [2025] extended spectral-filtering ideas toward marginally stable nonlinear dynamics. These works show

that stability or marginal stability enables compact prediction without full system identification, but they do not address systems with genuinely exploding[2], complex, or non-semisimple modes.

Unstable dynamical systems. Several works study learning problems in which the underlying dynamics may be unstable, but mostly in settings different from the partially observed prediction problem considered here. Classical work of White [1958] studies the one-dimensional explosive autoregressive process and shows that instability changes the asymptotic behavior of standard estimators; in this fully observed scalar setting, the exploding signal can make the parameter easier to estimate. Faradonbeh et al. [2018] study finite-time system identification for unstable linear systems under full-state observation. More broadly, Simchowitz et al. [2018], Sarkar and Rakhlin [2019], Jedra and Proutiere [2020] give finite-sample identification guarantees for arbitrary or poorly mixing LDSs, targeting recovery of the system parameters rather than compact prediction from outputs. Wagenmaker and Jamieson [2020] study active input design for identifying LDSs efficiently. Thus, while unstable dynamics have been studied in identification and input-design settings, these works do not provide the partially observed, parameter-adaptive online prediction guarantees considered here.

Controlling linear dynamical systems from data. Hazan and Singh [2026] surveys regret-based formulations connecting online convex optimization to control. The nonstochastic/adversarial line [Agarwal et al., 2019, Chen and Hazan, 2021, Gradu et al., 2022] studies sublinear regret under adversarial disturbances and time variation; Chen and Hazan [2021] in particular shows that black-box online control can incur unavoidable exponential-in-dimension overhead, complementing minimax hardness for underactuated plants in Tsiamis et al. [2022]. For quadratic costs, Foster and Simchowitz [2020] achieves logarithmic regret when the dynamics are known but disturbances are adversarial, whereas Simchowitz and Foster [2023] establishes that adaptive LQR with unknown models has $\tilde{\Theta}(\sqrt{T})$ minimax regret—contrasting with the polylogarithmic prediction rates available in some filtering settings. On the synthesis side, Dean et al. [2018] couples identification with robust control, and Mania et al. [2019] shows certainty-equivalent LQ/LQG controllers inherit particularly favorable (quadratic-in-error) suboptimality scalings when estimates are accurate.

3 Problem Setup

Linear dynamical system. We consider a bounded input sequence $u_1, u_2, \ldots \in \mathbb{R}^m$ with $\|u_t\|_2 \le K$ for all t, and the corresponding output sequence $y_1, y_2, \ldots \in \mathbb{R}^p$ generated by an unknown linear dynamical system described below. The hidden state $x_t \in \mathbb{R}^d$ summarizes the system's memory of past inputs, with dynamics

$$x_t \;=\; A\,x_{t-1} + B\,u_t, \qquad y_t \;=\; C\,x_t, \qquad x_0 \;=\; 0, \tag{1}$$

where $A \in \mathbb{R}^{d\times d}$, $B \in \mathbb{R}^{d\times m}$, $C \in \mathbb{R}^{p\times d}$. We assume the pair (A, C) is observable. We use the shorthand $y_{1:t} := (y_1, \ldots, y_t)$ and likewise for other sequences.

Stable modes and instability complexity. The eigenvalues $\lambda_1, \ldots, \lambda_d$ of A in its Jordan decomposition over $\mathbb{C}$ are counted with algebraic multiplicity. We use the following complexity measure to isolate the modes which are hardest to learn.

Definition 1 (Instability complexity)**.** The *instability complexity* of an LDS, denoted k, is the smallest integer $k \ge 0$ such that at most k eigenvalues of A (counted with algebraic multiplicity) either lie outside $(-1, 1)$ with $|\lambda_i| > 1 - 1/\log(k+2)$, or are non-semisimple. We set $k = 1$ when no eigenvalue satisfies these conditions.

We also write

$$\rho_{\mathrm{st}} := \max\{|\lambda_i| : \lambda_i \in (-1,1) \text{ and semisimple}\},$$

with the convention $\rho_{\mathrm{st}} = 0$ if there are no such eigenvalues.

Intuitively, the instability complexity counts the modes whose impulse responses are not guaranteed to lie close to a common low-dimensional subspace spanned by a few fixed filters. Rapidly decaying semisimple modes can be approximated using a short input window, and stable real semisimple modes using a few spectral filters, even when they decay slowly. Other modes count toward k.

[2]Exploding dynamics do not by themselves preclude bounded mean-square estimation error: classical Kalman filtering can achieve this under standard assumptions [Anderson and Moore, 1979, Sec. 4.4]. The challenge addressed here is compact online prediction, with sublinear regret and only $\widetilde{O}(k)$ learned parameters, independent of the full hidden-state dimension d.

Prediction goal. Our goal is to design an online algorithm that, given $(u_t,\, y_{1:t-1})$ at each step, produces a prediction $\widehat{y}_t$ whose average squared loss $\frac{1}{T}\sum_{t=1}^{T}\ell_2^2(\widehat{y}_t, y_t)$ vanishes as $T \to \infty$, where $\ell_2(\widehat{y}_t, y_t) = \|\widehat{y}_t - y_t\|_2$ denotes the standard L_2 loss.

Controlled inputs. Our motivating goal is to learn to stabilize unknown systems. Prediction is useful for this purpose only when the trajectory remains bounded: once the system has already exploded, the rollout has left the setting where predictions can be used to keep it stabilized. We therefore focus on system-input pairs that keep the dynamics *under control*: although the open-loop dynamics may be unstable, the chosen inputs keep the resulting outputs within a fixed bounded region. The input-generating process need not be known to the learner; prediction uses only the observed inputs and outputs. Moreover, systems whose open-loop dynamics are already strictly stable (including those with complex eigenvalues) are inherently under control for bounded inputs; our algorithm naturally handles this setting as well[3]. This is formally defined as follows:

Assumption 2 (Controlled outputs)**.** There exists a constant $B_y < \infty$, independent of the horizon T, such that for the LDS (A, B, C) and the input sequence $(u_t)_{t\geq 1}$ under consideration, the resulting outputs satisfy

$$\max_{1\leq t\leq T}\|y_t\|_2 \leq B_y \qquad \text{for every } T \geq 1.$$

We call such an input sequence B_y*-controlling* for the system.[4]

One motivating setting is learning a system while a stabilizing operator keeps it bounded, as when a pilot flies an open-loop unstable helicopter. Stabilization and control optimization can be pursued jointly, with the goal of improving control without losing stability. Prediction remains useful because, even after the unstable component is controlled, the remaining dynamics may determine the accumulated cost. Using the learned predictor to construct a stabilizing controller is left to future work.

4 Algorithm and Main Result

4.1 Predictor Classes

We begin by describing the predictor class from which our unified algorithm (Algorithm 1) will learn online. The central goal of this section is to construct, in Section 4.1.4, a predictor class that can approximate *any* LDS to arbitrary accuracy using only $O(k \cdot \mathrm{polylog}(T, k))$ parameters, a quantity that depends linearly in k and only logarithmically in k and T.

Our construction draws on three well-known predictor classes. The first, *spectral filtering* (Section 4.1.1), approximates stable real-diagonalizable LDS outputs using a number of parameters independent of the state dimension, but fails on systems with non-stable or complex modes. The second, *finite-memory input filters* (FIR) (Section 4.1.2), handle modes whose impulse responses decay quickly. The third, *autoregressive* (AR) *predictors* (Section 4.1.3), represent any observable LDS exactly, but at the cost of a parameter count proportional to the full state dimension. None of these components alone suffices for our setting: we want a predictor that captures every LDS, including non-diagonalizable systems and systems with exploding or complex eigenvalues, while paying a parameter cost only in the instability complexity. Combining the three families gives precisely such a class.

4.1.1 Spectral filters for stable real-diagonalizable LDS

For stable real-diagonalizable LDS (equivalently, all modes are real, semisimple, and lie in $(-1, 1)$), spectral filtering [Hazan et al., 2017] approximates the output arbitrarily well using a small, fixed set of universal basis filters—independent of the system parameters—applied to the input history. Write $\tilde{u}_t := (u_t, u_{t-1}, \ldots, u_1, 0, \ldots, 0) \in \mathbb{R}^{m\times T}$ for the zero-padded input history, and let $\phi_1, \ldots, \phi_T \in \mathbb{R}^T$ denote the universal spectral filters as defined in Definition 13. The following proposition follows from the spectral filtering framework [Hazan et al., 2017, Hazan and Singh, 2026].

[3] Conversely, systems that are not controlled rapidly explode. Such trajectories are statistically benign: the output sequence becomes dominated by the system's largest mode, typically making prediction easy.

[4] Bounded inputs can produce growing outputs in marginally stable systems, for example through resonance. Our assumption excludes such trajectories. We leave extending the result to this case to future work.

Proposition 3 (Informal; see Proposition 18 for the formal version)**.** *Consider a stable real-diagonalizable LDS as in Eq.* (1)*. For every $\varepsilon > 0$, there exist $h = O\big(\log T \cdot \log \frac{T}{\varepsilon}\big)$ and matrices $M_1, \ldots, M_h \in \mathbb{R}^{p \times m}$ depending only on (A, B, C), such that*

$$\sum_{t=1}^{T} \Big\| y_t - \sum_{s=1}^{h} M_s(\tilde{u}_t \phi_s) \Big\|^2 \;\le\; \varepsilon.$$

In particular, h has no polynomial dependence on the state dimension d.

4.1.2 Finite-memory filters for rapidly decaying modes

Finite-memory input filters approximate the output using only a short window of past inputs. They apply whenever the relevant modes are semisimple and strictly stable, as formalized in the following theorem.

Proposition 4 (Informal; see Lemma 19 for the formal version)**.** *Consider an observable LDS as in Eq.* (1) *whose state matrix A is diagonalizable and satisfies $|\lambda_i| \le 1 - \frac{1}{L}$ for every eigenvalue λ_i. For every $\varepsilon > 0$, there exist $L_\varepsilon = O\big(L \log \frac{T}{\varepsilon}\big)$ and matrices $Q_0, \ldots, Q_{L_\varepsilon - 1} \in \mathbb{R}^{p \times m}$, depending only on (A, B, C), such that*

$$\sum_{t=1}^{T} \Big\| y_t - \sum_{\ell=0}^{L_\varepsilon - 1} Q_\ell \, u_{t-\ell} \Big\|^2 \;\le\; \varepsilon.$$

In particular, L_ε is independent of the state dimension d.

4.1.3 Autoregressive predictors for general LDS

For any observable LDS with state dimension d, the output satisfies an exact autoregressive relation: the last d past outputs and the d most recent inputs determine y_t exactly. Unlike spectral filtering and FIR, this representation requires no approximation, but its parameter count is proportional to d. This exact representation is a standard consequence of realization theory and the Cayley–Hamilton theorem [Hazan and Singh, 2026, Kailath, 1980, Cayley, 1858]

Proposition 5 (Informal; see Proposition 17 for the formal version)**.** *Consider any observable LDS as in Eq.* (1) *with state dimension d. Then there exist coefficients $\alpha_1, \ldots, \alpha_d \in \mathbb{R}$ and matrices $N_0, \ldots, N_{d-1} \in \mathbb{R}^{p \times m}$ depending only on (A, B, C), such that for all $t \ge d + 1$,*

$$y_t \;=\; \sum_{i=1}^{d} \alpha_i \, y_{t-i} \;+\; \sum_{i=0}^{d-1} N_i \, u_{t-i}.$$

4.1.4 The unified predictor class

In this section we introduce a hybrid predictor class that combines the strengths of the spectral filtering, finite-memory filtering, and autoregressive predictor classes introduced in the previous sections. While it might seem that these classes can be trivially combined, the difficulty is that their approximation errors interact. The AR block acts on the observed outputs with coefficients that can be exponentially large in k, amplifying the errors of the other two components. We therefore need exponentially more accurate approximations and a learner whose regret depends only logarithmically on the coefficient norms.

Below we prove that any dynamical system with instability complexity k can be approximated using only $O(k \cdot \mathrm{polylog}(T, k, 1/\varepsilon))$ parameters, depending only on k and not on the dimension of the state space.

Theorem 6 (Informal; see Theorem 20 for the formal version)**.** *Consider any LDS as in Eq.* (1) *with instability complexity k. For every $\varepsilon > 0$, there exist $h = O\big(k \cdot \log T \cdot \log \frac{T \cdot k}{\varepsilon}\big)$ and $L = O(k \cdot \mathrm{polylog}(T, k, 1/\varepsilon))$, coefficients $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$, and matrices $W_s, Q_\ell \in \mathbb{R}^{p \times m}$, such that*

$$\widehat{y}_t \;=\; \sum_{i=1}^{k} \alpha_i \, y_{t-i} \;+\; \sum_{s=1}^{h} W_s(\tilde{u}_t \phi_s) \;+\; \sum_{\ell=0}^{L+k} Q_\ell \, u_{t-\ell}$$

satisfies $\sum_{t=1}^{T} \|y_t - \widehat{y}_t\|^2 \le \varepsilon$, using $O(k \cdot \mathrm{polylog}(T, k, 1/\varepsilon))$ parameters.

Proof sketch (full proof in Appendix F). Write $y_t = y_t^{(\mathrm{st})} + y_t^{(\mathrm{fast})} + y_t^{(\mathrm{ar})}$ for the contributions of, respectively, the real semisimple modes with $|\lambda| > 1 - 1/\log(k+2)$ (the *spectral group*), the remaining semisimple modes (all strictly stable, with modulus at most $1 - 1/\log(k+2)$), and the k modes counted by the instability complexity. By Cayley–Hamilton applied to the AR block, the third piece $y_t^{(\mathrm{ar})}$ satisfies an exact AR relation of order at most k. Since $y_t^{(\mathrm{ar})}$ is not directly observed, we substitute $y_{t-i}^{(\mathrm{ar})} = y_{t-i} - y_{t-i}^{(\mathrm{st})} - y_{t-i}^{(\mathrm{fast})}$, rewriting y_t as observable AR lags of y plus the AR residuals of the stable and fast components. The AR residual of the stable component can be approximated, at time t, by spectral filters applied to the current input window together with a few of the most recent inputs $u_t, \ldots, u_{t-k}$. The fast residual is approximated by a single FIR convolution of length $L + k$. Young's inequality then bounds how the FIR error propagates through the AR coefficients, yielding the theorem. □

4.2 The Unified Algorithm

We now turn the approximation results of Section 4.1 into an online learning algorithm. The goal is to compete with an unknown LDS without knowing its matrices or spectrum, while using only $O(k \cdot \mathrm{polylog}(T, k))$ parameters. This is the sense in which the algorithm is a unified one: the same procedure applies to stable systems, systems with complex eigenvalues, and systems with exploding modes.

The algorithm learns the coefficients of the unified predictor class from Section 4.1.4. At time t, the feature vector contains the k autoregressive output lags, a finite-memory input window of length $L + k + 1$, and the spectral-filter features of the current input window. We run one copy of the Vovk–Azoury–Warmuth forecaster per output coordinate; see Appendix H for the definition of the algorithm and its regret guarantee.

In each iteration, the VAW forecaster for the i-th output coordinate receives a feature vector $a_{t,i}$ and is asked to make a prediction $\widehat{y}_t^{(i)} = a_{t,i}^\top \widehat{x}_t$. The algorithm then observes the true output $y_t^{(i)}$ and suffers squared loss $(\widehat{y}_t^{(i)} - y_t^{(i)})^2$. The feature vectors at time t for the i-th output coordinate are defined as follows:

$$a_{t,i} \;=\; \begin{bmatrix} Y_{t,i}^\top & F_t^\top & \tilde{U}_t^\top \end{bmatrix}^\top \in \mathbb{R}^{d_{\mathrm{feat}}},$$

where $d_{\mathrm{feat}} = k + m(L + k + 1) + hm$ and, for $i \in [p]$,

$$Y_{t,i} = \begin{bmatrix} (y_{t-1})_i & \cdots & (y_{t-k})_i \end{bmatrix}^\top \in \mathbb{R}^k, \tag{2}$$

$$F_t = \begin{bmatrix} u_t^\top & \cdots & u_{t-(L+k)}^\top \end{bmatrix}^\top \in \mathbb{R}^{m(L+k+1)}, \tag{3}$$

$$\tilde{U}_t = \begin{bmatrix} (\tilde{u}_t \phi_1)^\top & \cdots & (\tilde{u}_t \phi_h)^\top \end{bmatrix}^\top \in \mathbb{R}^{hm}. \tag{4}$$

The window F_t subsumes the k recent inputs originally used by the AR component and additionally carries the lags needed by the finite-memory predictor for strictly stable modes. These features are exactly the parameters of a predictor in the unified class of Theorem 6.

Algorithm 1 Unified LDS prediction.

Input: AR order k, filter count h, finite-memory length L.
Initialize VAW instances $\mathcal{A}_1, \ldots, \mathcal{A}_p$ with dimension d_{feat}.
for $t = 1, 2, \ldots, T$ **do**
  Observe $u_t \in \mathbb{R}^m$.
  Set vectors $Y_{t,i}, F_t, \tilde{U}_t, \forall i \in [p]$ as defined in Eq. (2).
  Set $a_{t,i} \leftarrow \begin{bmatrix} Y_{t,i}^\top & F_t^\top & \tilde{U}_t^\top \end{bmatrix}^\top$.
  For each $i \in [p]$, obtain $m_t^i \leftarrow \mathcal{A}_i.\mathrm{predict}(a_{t,i})$.
  Predict $\widehat{y}_t \leftarrow (m_t^1, \ldots, m_t^p)^\top$.
  Observe y_t and suffer loss $\ell_2^2(\widehat{y}_t, y_t) = \|\widehat{y}_t - y_t\|_2^2$.
  For each $i \in [p]$, update $\mathcal{A}_i.\mathrm{update}\big((y_t)_i\big)$.
end for

Theorem 7 (Informal; see Theorem 22 for the formal version)**.** *Consider any observable LDS with state dimension d, of which the instability complexity is k, and any input sequence whose generated*

outputs satisfy Assumption 2. With filter count $h = O(k \cdot \log T \cdot \log(\mathrm{poly}(T,k)))$, finite-memory length $L = O(k \cdot \mathrm{polylog}(T,k))$, Algorithm 1 achieves loss

$$\frac{1}{T}\sum_{t=1}^{T} \ell_2^2(\widehat{y}_t, y_t) = O\left(\frac{k^2 \log^3(k) \cdot \log^3(T)}{T}\right)$$

assuming the parameters of the LDS are constants except T, k[5], and uses $O(k \cdot \mathrm{polylog}(T,k))$ learnable parameters.

Proof sketch (full proof in Appendix G). We decompose the cumulative loss into two parts. Let $h^\star$ be the value of h from the statement and let $f^\star$ be the predictor from Theorem 6 with that $h^\star$.

$$\sum_{t=1}^{T} \ell_2^2(\widehat{y}_t, y_t) \;=\; \underbrace{\Big(\sum_t \ell_2^2(\widehat{y}_t, y_t) - \sum_t \ell_2^2(f_t^\star, y_t)\Big)}_{(I)\text{: VAW regret}} + \underbrace{\sum_t \ell_2^2(f_t^\star, y_t)}_{(II)\text{: approximation error}} \quad .$$

The first term, (I), is the estimation error of the chosen predictor class: it is the regret of the per-coordinate VAW forecaster on the features $a_{t,i}$ defined in Algorithm 1. Since the features $a_{t,i}$, the comparator class from Theorem 20, and the labels $y_t^{(i)}$ are bounded by constants, the standard VAW analysis (see Lemma 24 in Appendix H) gives logarithmic regret in T.

The second term, (II), is the approximation error: it measures how well the class can approximate the real output sequence. This term is controlled by Theorem 20, which gives that $f^\star$ achieves total squared error $\sum_t \ell_2^2(f_t^\star, y_t) \le \varepsilon$ against $y_{1:T}$. Choosing the optimal ε and combining the two terms yields the claimed $o(T)$ regret. □

4.2.1 Extensions

The theorem above assumes a correctly specified spectral cutoff and known T and k; the following extensions address cutoff misspecification and remove the need to know these two parameters.

Robustness to the spectral cutoff. Allowing the finite-memory block's spectral radius to exceed its nominal cutoff by a sufficiently small δ adds only $O(\delta)$ to the cumulative loss guarantee, for fixed approximation accuracy and problem parameters, including k. The predictor size and VAW regret bound are unchanged (Proposition 21).

Adapting to unknown horizon and complexity. The horizon T and instability complexity k need not be known in advance, given the known input, output, and system bounds specified in Appendix I. This costs only an additional $\log(T)$ factor in the loss bound while retaining $\widetilde{O}(k)$ learnable parameters (see Theorem 25 for details).

5 Lower Bounds and Motivation

Our unified algorithm, Algorithm 1, has learnable parameter count $\widetilde{O}(k)$, depending on the instability complexity rather than the full state dimension. In the following subsections we explain why k is the correct parameter to depend on. First, we show that in a broad diagonalizable class, many exploding modes relative to the number of input channels create a practical obstruction to stabilization, motivating the small-k regime under limited actuation (Section 5.1). Second, we show that at least a linear dependence on k is unavoidable for filter-based predictors (Section 5.2).

5.1 Systems with many exploding modes and few inputs are practically uncontrollable

Our motivating goal is to stabilize systems whose open-loop dynamics may be unstable. The results above show that prediction—a natural prerequisite for control—can be done with a memory footprint depending only on k, the instability complexity, rather than the full state dimension d. We now

[5] As in prior work [Hazan et al., 2017, Ghai et al., 2020], we hold the input and output dimensions and the bounds on inputs, outputs, and system-matrix norms fixed as T, k, d vary. The dependence on these quantities is explicit in Theorem 20 and Lemma 24.

argue that a dependence on k is a minor concession when actuation is limited: already in a broad diagonalizable class, many exploding modes relative to the number of input channels can make stabilization impractical.

To formalize this, we introduce the *input dynamic range* of a control sequence u: the ratio of the largest to the smallest nonzero input magnitude, denoted $\mathrm{IDR}(u)$. A large input dynamic range means the actuator must simultaneously operate at vastly different scales—an extreme sensitivity to rounding, actuator limits, and noise.

Definition 8. For a nonzero input sequence $u = (u_t)_{t\geq 1}$ with $u_t \in \mathbb{R}^m$, its *dynamic range*, denoted $\mathrm{IDR}(u)$, is defined as follows:

$$\mathrm{IDR}(u) \;:=\; \frac{\sup_{t\geq 1} \|u_t\|_2}{\inf\{\|u_t\|_2 : u_t \neq 0\}},$$

with $\mathrm{IDR}(u) = \infty$ if the numerator is infinite or the denominator is zero.

We show this obstruction in a simple diagonalizable setting with k exploding modes and m input channels. With high probability over a uniformly random first nonzero input direction, every stabilizing continuation has dynamic range growing exponentially in k/m (Proposition 9). For a single input channel, the conclusion holds for every nonzero stabilizing sequence. We state the result in diagonal coordinates; any diagonalizable system with distinct real eigenvalues can be put in this form by a change of basis.

Proposition 9. *Let $1 \leq m \leq k$ and $A = \mathrm{diag}(\lambda_1, \ldots, \lambda_k) \in \mathbb{R}^{k\times k}$ have distinct real entries satisfying $|\lambda_i| \geq 1 + \eta$ for some $\eta > 0$. Let $B \in \mathbb{R}^{k\times m}$ have no zero rows,[6] and consider*

$$x_t = Ax_{t-1} + Bu_t, \qquad x_0 = 0.$$

For any $\delta \in (0, 1)$, draw v uniformly from the unit sphere $\mathbb{S}^{m-1}$. With probability at least $1 - \delta$, every nonzero input sequence whose first nonzero input is a positive multiple of v and which stabilizes the system, meaning $x_t \to 0$, satisfies

$$\mathrm{IDR}(u) \geq \frac{\delta}{\sqrt{m}} \frac{\eta}{2(1+\eta)} \left(1 + \frac{\eta}{2}\right)^{k/m}. \tag{5}$$

Proof sketch (full proof in Appendix K.1). We construct a positive definite Gramian G from the system matrices A and B, and show that $B^\top G^{-1} B$ has an eigenvalue that grows exponentially in k/m. Stabilization requires subsequent controls to cancel the effect of the first nonzero control u_1. Using this eigenvalue bound, we show that, with high probability over a uniformly random direction of u_1, every stabilizing continuation must contain a control whose norm is exponentially larger than that of u_1. This gives the claimed lower bound on the input dynamic range. □

5.2 Filter-count lower bound

Our algorithm uses a filter class whose size grows linearly with k up to logarithmic factors. We now show that this linear dependence is unavoidable for predictors of the same broad type, even when the filters themselves are chosen arbitrarily. The lower bound applies to bounded-coefficient filter predictors: predictors that form bounded linear combinations of fixed causal filters applied to past inputs and outputs. For such predictors, fewer than k filter directions cannot guarantee sublinear regret against the LDS class. Since Section 5.1 motivates the regime of small k, the resulting dependence is not on the hidden dimension or an expressivity obstruction.

We formalize this by defining a general filter class (Definition 10) that applies arbitrary fixed ℓ_1-bounded filters to past outputs and inputs, with a bound on the total learned coefficient mass. Our unified predictor (Theorem 6) is an instance of this template: its output filters are delay operators, and its input filters consist of spectral filters and finite-memory delay operators. We then show (Proposition 11) that if the total filter count is below k, there is an LDS whose output no predictor in the class can track with vanishing average loss.

[6] This excludes input-decoupled modes, which can be ignored.

Definition 10 (General filter class)**.** Let $\Psi^y = \{\psi_1^y, \dots, \psi_{q_y}^y\} \subset \ell_1(\mathbb{N})$ be strictly causal output filters with $\|\psi_r^y\|_1 \le 1$, and let $\Psi^u = \{\psi_1^u, \dots, \psi_{q_u}^u\} \subset \ell_1(\mathbb{N}_0)$ be input filters with $\|\psi_a^u\|_1 \le 1$. For scalar input and output sequences, write

$$(\psi_r^y * y)_t \;:=\; \sum_{\tau=1}^{t-1} \psi_{r,\tau}^y \, y_{t-\tau}, \qquad (\psi_a^u * u)_t \;:=\; \sum_{\tau=0}^{t-1} \psi_{a,\tau}^u \, u_{t-\tau}.$$

For $R < \infty$, the associated bounded-coefficient filter predictor class is

$$\mathrm{GUF}^R_{\Psi^y,\Psi^u} \;:=\; \left\{ f \;:\; \begin{array}{c} f_t = \sum_{r=1}^{q_y} \alpha_r (\psi_r^y * y)_t + \sum_{a=1}^{q_u} \beta_a (\psi_a^u * u)_t, \\ \sum_{r=1}^{q_y} |\alpha_r| + \sum_{a=1}^{q_u} |\beta_a| \le R \end{array} \right\}.$$

We call $q_y + q_u$ the *total filter count*.

Proposition 11. *Fix $k \ge 1$ and $R < \infty$, and let Ψ^y, Ψ^u be any filters as in Definition 10 with total filter count $q_y + q_u < k$. Then there exist a k-dimensional real LDS (A, B, C) with $\|A\|_\infty, \|B\|_\infty, \|C\|_\infty \le 1$, and a bounded input sequence $(u_t)_{t\ge 1}$ that is 1-controlling such that the resulting outputs satisfy*

$$\liminf_{T\to\infty} \; \inf_{f \in \mathrm{GUF}^R_{\Psi^y,\Psi^u}} \frac{1}{T} \sum_{t=1}^{T} \ell_2^2(f_t, y_t) \;>\; 0.$$

Proof sketch (full proof in Appendix J). We prove this directly by explicitly constructing such an LDS[7]. The construction uses a standard delay-line system: an input impulse enters the first state coordinate, shifts one coordinate at a time, and is observed as output only after a specifically chosen delay. Since the predictor has strictly fewer than k input-filter directions, linear algebra dictates that at least one of the k possible delays cannot be represented by the first k coefficients of these filters. We then feed the system a sequence of isolated impulses separated by long gaps. The ℓ_1 coefficient bound ensures that contributions from old impulses decay and become negligible. Thus, in every block, the predictor makes a constant error on the unrepresentable delay, giving a strictly positive average squared loss. □

6 Experimental Validation

We empirically test the prediction suggested by Theorem 7: for a high-dimensional LDS with small instability complexity, the unified predictor should learn with a parameter budget tied to k rather than to the full state dimension. We construct a 503-dimensional controlled system with $k = 3$, train several online linear predictors with the same number of learned scalar parameters using the same Vovk–Azoury–Warmuth forecaster, and compare their normalized mean-squared error[8] on a single training trajectory of length $T = 200{,}000$. To measure the asymptotic behavior, we report the average loss over the last $10{,}000$ time steps. More extensive details for the experimental setting are in Appendix L. Code for reproducing all demonstrations can be found in `https://github.com/YuvMilo/UnifiedLearningLDS`.

Table 1: Controlled $k = 3$ system with equal parameter budget. We report the best and worst normalized mean-squared error over five random seeds, measured on the last $10{,}000$ time steps.

Predictor class	Best seed	Worst seed
Finite-memory input filters	$3.76 \cdot 10^{-1}$	$9.60 \cdot 10^{-1}$
Autoregressive predictor	$1.73 \cdot 10^{-5}$	$9.08 \cdot 10^{-5}$
Spectral filtering	$5.20 \cdot 10^{-3}$	$6.13 \cdot 10^{-2}$
Unified predictor	$7.88 \cdot 10^{-9}$	$1.65 \cdot 10^{-8}$

[7] Our construction is heavily influenced by the lower bound argument in Hazan and Marsden [2026].

[8] We normalize the system by a scalar so that the zero predictor has normalized mean-squared error 1.

System. The system is scalar-input, scalar-output, and has state dimension $d = 503$. It contains one exploding scalar mode with eigenvalue 1.3 and one slow complex conjugate pair with eigenvalues $0.98e^{\pm i1.57}$; these three eigenvalues are the only modes counted by the instability complexity. The other modes are 300 real stable modes in $(-1, 1)$ and 100 fast-decaying complex conjugate pairs of the form $re^{\pm i\theta}$ with $r \in [0, 0.25]$. The real stable part is chosen so that its input-output response has large components in many spectral-filter directions, making it difficult to learn from a short input window alone. The input controls the exploding mode explicitly. We draw independent signs $z_t \in \{-1, +1\}$ along the trajectory and set $u_t = z_t - 1.3z_{t-1}$, with $z_0 = 0$. Thus the open-loop system is unstable, but the observed trajectory remains bounded.

Predictor classes. The comparison is an ablation study of the three ingredients in the unified predictor. To make the comparison fair, each predictor has exactly 16 learned scalar parameters and is trained by the same Vovk–Azoury–Warmuth learner from Appendix H. The finite-memory input-filter predictor spends all 16 parameters on recent inputs, as in Section 4.1.2. The autoregressive predictor splits the budget evenly between 8 output lags and 8 input lags, as in Section 4.1.3. The spectral-filtering predictor spends all 16 parameters on spectral filters, as in Section 4.1.1. The unified predictor uses all three components: 3 output lags, 7 input lags, and 6 spectral filters, matching the construction in Section 4.1.4.

Results. Table 1 reports the final normalized mean-squared error range over five random seeds. The unified predictor gives the lowest error by a wide margin, showing that the advantage comes from combining the autoregressive, finite-memory, and spectral-filtering features.

7 Limitations and Future Work

Several limitations of this work suggest natural directions for future research. First, while our upper and lower bounds match up to polylogarithmic factors in their dependence on k for parameter count, they do not yet settle the optimal logarithmic dependence or the sharp regret constants. Second, our results concern prediction rather than control. A natural next step is to use the predictor as a component of a closed-loop stabilization policy, and to understand how its regret guarantees translate into guarantees for control objectives.

8 Conclusion

Motivated by the goal of stabilizing unknown linear dynamical systems from observations, we studied the online prediction problem that precedes it. The existing theory leaves a fundamental gap: available methods either rely on restrictive assumptions on the dynamics, or else pay memory proportional to the full hidden-state dimension. We identified a new regime that captures an important middle ground: systems with only k modes outside the real stable interval that do not decay rapidly, together with non-semisimple modes, embedded in an arbitrarily large stable real semisimple component. We show that this regime is central not only because it is broader than the classical stable setting, but also because it is strongly motivated by an obstruction to control: already in a broad diagonalizable class, many exploding modes relative to the number of input channels can force stabilizing inputs to span exponentially many scales (Proposition 9), making such dynamics practically uncontrollable due to acute sensitivity to finite precision, actuator limits, and noise.

Within this important regime, we showed that prediction does not need to pay for the full hidden dimension. Our algorithm combines dimension-free spectral filtering for stable real semisimple modes, finite-memory input filtering for fast decaying semisimple modes, and a small autoregressive correction for the k exceptional modes. This yields a unified online predictor for general LDSs that achieves sublinear regret with a learnable parameter count of $\widetilde{O}(k)$, completely independent of the full hidden-state dimension. The lower bound of Proposition 11 shows that at least linear dependence on k is unavoidable for a fixed filter family covering the LDS class, suggesting the *instability complexity* k as a natural complexity measure for this problem.

We hope that this perspective serves as a foundation for future work that turns such prediction guarantees into closed-loop stabilization methods.

Acknowledgments

This work was supported by the European Research Council (ERC) grant NN4C 101164614, a Google Research Scholar Award, a Google Research Gift, Meta, the Yandex Initiative in Machine Learning, the Israel Science Foundation (ISF) grant 1780/21, the Tel Aviv University Center for AI and Data Science, the Adelis Research Fund for Artificial Intelligence, Len Blavatnik and the Blavatnik Family Foundation, and Amnon and Anat Shashua. Elad Hazan gratefully acknowledges support from the Office of Naval Research, Open Philanthropy, and ARIA.

References

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In Sham M. Kakade and Ulrike von Luxburg, editors, *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 1–26, Budapest, Hungary, 09–11 Jun 2011. PMLR. URL https://proceedings.mlr.press/v19/abbasi-yadkori11a.html.

Naman Agarwal, Brian Bullins, Elad Hazan, Sham M. Kakade, and Karan Singh. Online control with adversarial disturbances, 2019. URL https://arxiv.org/abs/1902.08721.

Brian D. O. Anderson and John B. Moore. *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, NJ, 1979.

Arthur Cayley. Ii. a memoir on the theory of matrices. *Philosophical Transactions of the Royal Society of London*, (148):17–37, 12 1858. ISSN 0261-0523. doi: 10.1098/rstl.1858.0002. URL https://doi.org/10.1098/rstl.1858.0002.

Xinyi Chen and Elad Hazan. Black-box control for linear dynamical systems, 2021. URL https://arxiv.org/abs/2007.06650.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator, 2018. URL https://arxiv.org/abs/1710.01688.

Evan Dogariu, Anand Brahmbhatt, and Elad Hazan. Universal learning of nonlinear dynamics, 2025. URL https://arxiv.org/abs/2508.11990.

John Doyle, Keith Glover, P. Khargonekar, and Bruce Francis. State space solution to standard H_2 and H_∞ control problem. *Automatic Control, IEEE Transactions on*, 34:831 – 847, 09 1989. doi: 10.1109/9.29425.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time identification in unstable linear systems, 2018. URL https://arxiv.org/abs/1710.01852.

Dylan J. Foster and Max Simchowitz. Logarithmic regret for adversarial online control, 2020. URL https://arxiv.org/abs/2003.00189.

Udaya Ghai, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. No-regret prediction in marginally stable systems. 2020. URL https://arxiv.org/abs/2002.02064.

Paula Gradu, Elad Hazan, and Edgar Minasyan. Adaptive regret for control of time-varying dynamics, 2022. URL https://arxiv.org/abs/2007.04393.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems, 2019. URL https://arxiv.org/abs/1609.05191.

Elad Hazan and Annie Marsden. Spectral filtering for complex linear dynamical systems, 2026. URL https://arxiv.org/abs/2601.22400.

Elad Hazan and Karan Singh. Introduction to online control, 2026. URL https://arxiv.org/abs/2211.09619.

Elad Hazan, Karan Singh, and Cyril Zhang. Learning linear dynamical systems via spectral filtering, 2017. URL https://arxiv.org/abs/1711.00946.

B. L. Ho and R. E. Kalman. Editorial: Effective construction of linear state-variable models from input/output functions. *at - Automatisierungstechnik*, 14(1-12):545–548, 1966. doi: doi: 10.1524/auto.1966.14.112.545. URL https://doi.org/10.1524/auto.1966.14.112.545.

Yassir Jedra and Alexandre Proutiere. Finite-time identification of stable linear systems: Optimality of the least-squares estimator, 2020. URL https://arxiv.org/abs/2003.07937.

Thomas Kailath. *Linear Systems*. 01 1980.

Vitaly Kuznetsov and Mehryar Mohri. Theory and algorithms for forecasting time series, 2018. URL https://arxiv.org/abs/1803.05814.

Lennart Ljung. *System Identification: Theory for the User*. Prentice Hall PTR, Upper Saddle River, NJ, 2nd edition, 1999.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control, 2019. URL https://arxiv.org/abs/1902.07826.

VP Overschee. Subspace identification for linear systems, 1996.

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, 2024. URL https://arxiv.org/abs/2410.14067.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems, 2019. URL https://arxiv.org/abs/1812.01251.

Max Simchowitz and Dylan J. Foster. Naive exploration is optimal for online lqr, 2023. URL https://arxiv.org/abs/2001.09576.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I. Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification, 2018. URL https://arxiv.org/abs/1802.08334.

Anastasios Tsiamis and George Pappas. Online learning of the kalman filter with logarithmic regret, 2020. URL https://arxiv.org/abs/2002.05141.

Anastasios Tsiamis, Ingvar Ziemann, Manfred Morari, Nikolai Matni, and George J. Pappas. Learning to control linear systems can be hard, 2022. URL https://arxiv.org/abs/2205.14035.

Volodya Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001. doi: https://doi.org/10.1111/j.1751-5823.2001.tb00457.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.2001.tb00457.x.

Andrew Wagenmaker and Kevin Jamieson. Active learning for identification of linear dynamical systems, 2020. URL https://arxiv.org/abs/2002.00495.

John S. White. The Limiting Distribution of the Serial Correlation Coefficient in the Explosive Case. *The Annals of Mathematical Statistics*, 29(4):1188 – 1197, 1958. doi: 10.1214/aoms/1177706450. URL https://doi.org/10.1214/aoms/1177706450.

Xiangyuan Zhang, Bin Hu, and Tamer Başar. Learning the kalman filter with fine-grained sample complexity, 2023. URL https://arxiv.org/abs/2301.12624.

A Spectral filtering approximation: stable real-diagonalizable case

In this section we prove the approximation guarantee of the spectral filtering class for stable real-diagonalizable LDS, restated as Proposition 3 in the main text.

Theorem 12. *If we set the number of filters to be*

$$h(\varepsilon, T) = 2 \cdot \log T \cdot \log\left(\frac{C_0 \cdot T^6}{\varepsilon}\right)$$

where C_0 is a constant depending on $\|B\|_F, \|C\|_F, K$ and the norm of the matrix V, where $A = VDV^{-1}$ and $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$, then the best policy from $\mathrm{SF}_{h(\varepsilon,T)}$ class achieves ℓ_2 loss at most ε. That is,

$$\sum_{t=1}^{T} \|\widehat{y}_t - y_t\|_2^2 \leq \varepsilon.$$

Before we jump into the proof, we will need to introduce some machinery.

Definition 13 (Spectral Filters)**.** Suppose $\mu_\alpha = \begin{bmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{T-1} \end{bmatrix}^\top$ for $\alpha \in (-1, 1)$. Define matrix Z_α as:

$$Z_\alpha = \int_{-1}^{1} \mu_\alpha \mu_\alpha^\top d\alpha \in \mathbb{R}^{T \times T}$$

Then the spectral filters are the orthonormal eigenvectors $\phi_1, \phi_2, \cdots, \phi_T$ of Z_α and the corresponding eigenvalues are $\lambda_1, \lambda_2, \cdots, \lambda_T$ of Z_α where $1 > |\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_T| > 0$.

We make use of the following two theorems from the book Hazan and Singh [2026]:

Theorem 14. *We have that for all $k \in [T]$*

$$\lambda_k \leq 4\pi^2 e^{-\frac{k}{\log T}}$$

Theorem 15. *We have that for all $i \in [T]$ and for all $\alpha \in (-1, 1)$:*

$$|\phi_i^\top \mu_\alpha| \leq \left(4T^2 \lambda_i\right)^{\frac{1}{4}} \overset{\text{Thm 14}}{\Longrightarrow} |\phi_i^\top \mu_\alpha| \leq 2\sqrt{\pi} \cdot T^{\frac{1}{2}} \cdot e^{-\frac{i}{4 \log T}}$$

The proof of Theorem 12 will be to essentially show that there exists a predictor in the class SF_h, where $h = O\left(\log T \cdot \log(T/\varepsilon)\right)$, which ε-approximates any stable real-diagonalizable LDS. Below we provide the full proof:

Proof of Theorem 12. Notice that since we are in the stable real-diagonalizable regime, where we can write:

$$A = VDV^{-1} = V \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{bmatrix} V^{-1}$$

where $1 > |\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_d| > 0$. We have that:

$$\begin{aligned} y_t &= \sum_{\tau=0}^{t-1} CA^\tau B u_{t-\tau} = \sum_{\tau=0}^{t-1} CVD^\tau V^{-1} B u_{t-\tau} = \sum_{\tau=0}^{t-1} CV \left(\sum_{i=1}^{d} \lambda_i^\tau e_i e_i^\top\right) V^{-1} B u_{t-\tau} \\ &= \sum_{\tau=0}^{t-1} \sum_{i=1}^{d} \lambda_i^\tau (CVe_i)(e_i^\top V^{-1} B) u_{t-\tau} = \sum_{i=1}^{d} c_i b_i^\top \sum_{\tau=0}^{t-1} \lambda_i^\tau u_{t-\tau} \end{aligned}$$

where we have defined $c_i = CVe_i$ (the i-th column of CV) and $b_i = (e_i^\top V^{-1} B)^\top$ (the i-th row of $V^{-1}B$). We can write

$$S_i = c_i b_i^\top \in \mathbb{R}^{p \times m} \text{ for } i \in [d], \text{ and } \tilde{u}_t = \begin{bmatrix} u_t & u_{t-1} & u_{t-2} & \cdots & u_1 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

where $\tilde{u}_t \in \mathbb{R}^{m \times T}$. Notice that we can bound the norm of M_i as follows:

$$\|S_i\|_2 = \|c_i b_i^\top\|_2 = \|CVe_i e_i^\top V^{-1} B\|_2 \leq \|CVe_i\|_2 \cdot \|e_i^\top V^{-1} B\|_2 \tag{6}$$

We can also bound the norm of $\tilde{u}_t$ as follows:

$$\|\tilde{u}_t\|_2^2 = \sum_{\tau=0}^{t-1} \|u_{t-\tau}\|_2^2 \leq \sum_{\tau=0}^{t-1} K^2 \leq K^2 \cdot T \tag{7}$$

That means that the following quantity can be written as:

$$\sum_{\tau=0}^{t-1} \lambda_i^\tau u_{t-\tau} = \begin{bmatrix} u_t & u_{t-1} & u_{t-2} & \cdots & u_1 & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} 1 \\ \lambda_i \\ \lambda_i^2 \\ \vdots \\ \lambda_i^{T-1} \end{bmatrix} = \tilde{u}_t \cdot \mu_{\lambda_i} \in \mathbb{R}^m$$

We will now perform a change of basis:

$$\begin{aligned} y_t &= \sum_{i=1}^{d} S_i \left(\tilde{u}_t \cdot \mu_{\lambda_i}\right) = \sum_{i=1}^{d} S_i \tilde{u}_t \left(\sum_{\tau=1}^{T} \phi_\tau \phi_\tau^\top\right) \mu_{\lambda_i} \\ &= \sum_{i=1}^{d} \sum_{\tau=1}^{T} \left(S_i \cdot \left(\phi_\tau^\top \mu_{\lambda_i}\right)\right) \tilde{u}_t \phi_\tau = \sum_{\tau=1}^{T} \sum_{i=1}^{d} M_{i,\tau} \tilde{u}_t \phi_\tau \\ &= \sum_{\tau=1}^{T} M_\tau \tilde{u}_t \phi_\tau \end{aligned}$$

where $M_\tau = \sum_{i=1}^{d} M_{i,\tau}$ and $M_{i,\tau} = S_i \cdot \left(\phi_\tau^\top \mu_{\lambda_i}\right)$ – recall that $\phi_\tau^\top \mu_{\lambda_i}$ is just a scalar. We can bound the norm of M_τ which will be useful later:

$$\|M_\tau\|_2 = \left\| \sum_{i=1}^{d} M_{i,\tau} \right\|_2 = \left\| \sum_{i=1}^{d} S_i \cdot \left(\phi_\tau^\top \mu_{\lambda_i}\right) \right\|_2 \leq \sum_{i=1}^{d} \|S_i\|_2 \cdot |\phi_\tau^\top \mu_{\lambda_i}| \tag{8}$$

$$\overset{\text{Ineq 6}}{\leq} \sum_{i=1}^{d} \|CVe_i\|_2 \cdot \|e_i^\top V^{-1} B\|_2 \cdot |\phi_\tau^\top \mu_{\lambda_i}| \tag{9}$$

$$\leq \|B\|_2 \cdot \|C\|_2 \sum_{i=1}^{d} \|Ve_i\|_2 \cdot \|e_i^\top V^{-1}\|_2 \cdot |\phi_\tau^\top \mu_{\lambda_i}| \tag{10}$$

$$\overset{\text{Thm 15}}{\leq} 2\sqrt{\pi} \cdot \|B\|_2 \cdot \|C\|_2 \cdot \sum_{i=1}^{d} \|Ve_i\|_2 \cdot \|e_i^\top V^{-1}\|_2 \cdot T^{\frac{1}{2}} e^{-\frac{\tau}{4 \log T}} \tag{11}$$

$$\leq 2\sqrt{\pi} \cdot \|B\|_2 \cdot \|C\|_2 \cdot \|V\|_F \cdot \|V^{-1}\|_F \cdot T^{\frac{1}{2}} e^{-\frac{\tau}{4 \log T}} \tag{12}$$

Notice that y_t is the true output of the LDS. Consider a prediction $\hat{y}_t$ from the class $\mathrm{SF}_{h,R}$ which predicts:

$$\hat{y}_t = \sum_{\tau=1}^{h} M_\tau \left(\tilde{u}_t \cdot \phi_\tau\right)$$

We have that:

$$\begin{aligned} \|\hat{y}_t - y_t\|_2 &= \left\| \sum_{\tau=1}^{T} M_\tau \left(\tilde{u}_t \cdot \phi_\tau\right) - \sum_{\tau=1}^{h} M_\tau \left(\tilde{u}_t \cdot \phi_\tau\right) \right\|_2 \\ &= \left\| \sum_{\tau=h+1}^{T} M_\tau \left(\tilde{u}_t \cdot \phi_\tau\right) \right\|_2 \\ &\leq \sum_{\tau=h+1}^{T} \|M_\tau\|_2 \cdot \|\tilde{u}_t \cdot \phi_\tau\|_2 \leq \sum_{\tau=h+1}^{T} \|M_\tau\|_2 \cdot \|\tilde{u}_t\|_2 \cdot \|\phi_\tau\|_2 \end{aligned}$$

$$\overset{\text{Ineq. 7}}{\leq} K \cdot \sqrt{T} \cdot \sqrt{T} \cdot \sum_{\tau=h+1}^{T} \|M_\tau\|_2$$

$$\overset{\text{Ineq. 8}}{\leq} T \cdot K \sum_{\tau=h+1}^{T} 2\sqrt{\pi} \cdot \|B\|_2 \cdot \|C\|_2 \cdot \|V\|_F \cdot \|V^{-1}\|_F \cdot T^{\frac{1}{2}} e^{-\frac{\tau}{4 \log T}}$$

$$\leq 2\sqrt{\pi} \cdot T^{\frac{5}{2}} \cdot K \cdot \|B\|_F \cdot \|C\|_F \cdot \|V\|_F \cdot \|V^{-1}\|_F \cdot e^{-\frac{h}{4 \log T}}$$

Thus:

$$L = \sum_{t=1}^{T} \|\hat{y}_t - y_t\|^2 \leq \sum_{t=1}^{T} \left(2\sqrt{\pi} \cdot T^{\frac{5}{2}} \cdot K \cdot \|B\|_F \cdot \|V\|_F \cdot \|V^{-1}\|_F \cdot \|C\|_F \cdot e^{-\frac{h}{4 \log T}}\right)^2$$
$$= 4\pi \cdot K^2 \cdot \|C\|_F^2 \cdot \|B\|_F^2 \cdot \|V\|_F^2 \cdot \|V^{-1}\|_F^2 \cdot T^6 \cdot e^{-\frac{h}{2 \log T}}$$

We want $L \leq \varepsilon$ thus we need:

$$4\pi \cdot K^2 \cdot \|C\|_F^2 \cdot \|B\|_F^2 \cdot \|V\|_F^2 \cdot \|V^{-1}\|_F^2 \cdot T^6 \cdot e^{-\frac{h}{2 \log T}} \leq \varepsilon$$
$$\iff e^{\frac{h}{2 \log T}} \geq \frac{4\pi \cdot K^2 \cdot \|C\|_F^2 \cdot \|B\|_F^2 \cdot \|V\|_F^2 \cdot \|V^{-1}\|_F^2 \cdot T^6}{\varepsilon}$$
$$\iff h \geq 2 \log T \cdot \log\left(\frac{4\pi \cdot K^2 \cdot \|C\|_F^2 \cdot \|B\|_F^2 \cdot \|V\|_F^2 \cdot \|V^{-1}\|_F^2 \cdot T^6}{\varepsilon}\right)$$

Using the constant $C_0 = 4\pi \cdot K^2 \cdot \|C\|_F^2 \cdot \|B\|_F^2 \cdot \|V\|_F^2 \cdot \|V^{-1}\|_F^2$ we get that we just need:

$$h \geq 2 \cdot \log T \cdot \log\left(\frac{C_0 \cdot T^6}{\varepsilon}\right)$$

□

B Diagonalization with separated real eigenvalues

Lemma 16. *Let $A \in \mathbb{R}^{d \times d}$. Let $S \subseteq \mathbb{R}$ be a set of real eigenvalues, and assume that every eigenvalue in S is semisimple. Let n be the total algebraic multiplicity of the eigenvalues in S. Write $\lambda_1, \ldots, \lambda_n$ for the eigenvalues in S, counted with algebraic multiplicity, and write $\mu_1, \ldots, \mu_{d-n}$ for the remaining eigenvalues of A, counted with algebraic multiplicity. We denote by $\sigma(A) = \{\lambda_1, \ldots, \lambda_n, \mu_1, \ldots, \mu_{d-n}\}$ the spectrum of A.*

Then there exist $P \in \mathrm{GL}_d(\mathbb{R})$ and $A_{\mathrm{ustb}} \in \mathbb{R}^{(d-n)\times(d-n)}$ such that

$$A = P \begin{pmatrix} \mathrm{diag}(\lambda_1, \ldots, \lambda_n) & 0 \\ 0 & A_{\mathrm{ustb}} \end{pmatrix} P^{-1}.$$

Moreover, the eigenvalues of A_{ustb} are precisely $\mu_1, \ldots, \mu_{d-n}$, counted with algebraic multiplicity.

Proof. For each eigenvalue $\nu \in \sigma(A)$, let

$$G_\nu := \ker\big((A - \nu I)^d\big)$$

denote the generalized eigenspace corresponding to ν. By the primary decomposition theorem,

$$\mathbb{C}^d = \bigoplus_{\nu \in \sigma(A)} G_\nu.$$

Set

$$U := \bigoplus_{\rho \in S} G_\rho, \qquad W := \bigoplus_{\nu \in \sigma(A) \setminus S} G_\nu.$$

Thus W is the generalized eigenspace corresponding to the eigenvalues not in S. Then

$$\mathbb{C}^d = U \oplus W, \qquad \dim_{\mathbb{C}} U = n.$$

Since every eigenvalue in S is semisimple, we have

$$G_\rho = \ker(A - \rho I) \qquad \text{for every } \rho \in S.$$

Also, each $\rho \in S$ is real, and $\ker(A - \rho I)$ is the kernel of a real matrix. Hence it has a basis of real vectors. Choosing such bases and concatenating them, we get real eigenvectors

$$q_1, \dots, q_n \in \mathbb{R}^d$$

forming a basis of U, ordered so that

$$Aq_i = \lambda_i q_i, \qquad i = 1, \dots, n.$$

Now A is real, so for every eigenvalue ν,

$$\overline{G_\nu} = G_{\overline{\nu}}.$$

Since A is real, non-real eigenvalues appear together with their complex conjugates. Since $S \subseteq \mathbb{R}$ contains only real eigenvalues, if $\nu \notin S$ then also $\overline{\nu} \notin S$. Hence $\sigma(A) \setminus S$ is closed under complex conjugation, and by the identity above the subspace W is stable under complex conjugation. Set

$$U_{\mathbb{R}} := \operatorname{span}_{\mathbb{R}}(q_1, \dots, q_n), \qquad W_{\mathbb{R}} := W \cap \mathbb{R}^d.$$

If $x \in \mathbb{R}^d$, write uniquely

$$x = u + w, \qquad u \in U, \quad w \in W.$$

Since

$$x = \overline{x} = \overline{u} + \overline{w},$$

with $\overline{u} \in U$ and $\overline{w} \in W$, uniqueness of the decomposition $\mathbb{C}^d = U \oplus W$ gives

$$u = \overline{u}, \qquad w = \overline{w}.$$

Thus $u \in U_{\mathbb{R}}$ and $w \in W_{\mathbb{R}}$, so

$$\mathbb{R}^d = U_{\mathbb{R}} \oplus W_{\mathbb{R}}.$$

Choose a real basis

$$w_1, \dots, w_{d-n}$$

of $W_{\mathbb{R}}$, and define

$$P := [q_1 \quad \cdots \quad q_n \quad w_1 \quad \cdots \quad w_{d-n}] \in \mathrm{GL}_d(\mathbb{R}).$$

Both $U_{\mathbb{R}}$ and $W_{\mathbb{R}}$ are A-invariant. Hence, in this basis,

$$P^{-1}AP = \begin{pmatrix} \operatorname{diag}(\lambda_1, \dots, \lambda_n) & 0 \\ 0 & A_{\mathrm{ustb}} \end{pmatrix}$$

for some real matrix

$$A_{\mathrm{ustb}} \in \mathbb{R}^{(d-n)\times(d-n)}.$$

Equivalently,

$$A = P \begin{pmatrix} \operatorname{diag}(\lambda_1, \dots, \lambda_n) & 0 \\ 0 & A_{\mathrm{ustb}} \end{pmatrix} P^{-1}.$$

Finally, A_{ustb} represents the restriction of A to the invariant subspace W. Therefore its eigenvalues are exactly the remaining eigenvalues

$$\mu_1, \dots, \mu_{d-n},$$

counted with algebraic multiplicity. This proves the result. □

C Autoregressive representation of any LDS

Proposition 17. *Let $\{u_t\}_{t=1}^T \subset \mathbb{R}^m$ be any input sequence, and consider the noiseless LDS*

$$x_t = Ax_{t-1} + Bu_t, \qquad y_t = Cx_t,$$

where

$$A \in \mathbb{R}^{d\times d}, \qquad B \in \mathbb{R}^{d\times m}, \qquad C \in \mathbb{R}^{p\times d}.$$

Then there exist coefficients

$$\{\alpha_i\}_{i=1}^d \subset \mathbb{R}, \qquad \{M_i\}_{i=0}^{d-1} \subset \mathbb{R}^{p\times m},$$

depending only on (A, B, C), such that for every $t = d+1, \ldots, T$,

$$\left\| y_t - \sum_{i=1}^{d} \alpha_i \, y_{t-i} - \sum_{i=0}^{d-1} M_i \, u_{t-i} \right\|_2 = 0.$$

Moreover, if $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of A (counted with algebraic multiplicity), then the coefficients may be chosen so that, for all relevant i,

$$|\alpha_i| \le L, \qquad \|M_i\|_\infty \le L,$$

where

$$L := \left(\prod_{j=1}^{d} (1 + |\lambda_j|) \right) \left(1 + d\|C\|_\infty \|B\|_\infty \max\{1, \|A\|_\infty\}^{\,d-1}\right).$$

Proof. Let

$$\chi_A(z) = z^d + c_1 z^{d-1} + \cdots + c_d = \prod_{j=1}^{d} (z - \lambda_j)$$

be the characteristic polynomial of A, and set $c_0 := 1$. By Cayley–Hamilton,

$$A^d + c_1 A^{d-1} + \cdots + c_d I = 0.$$

For any $s \ge 1$, repeated substitution in the state recursion gives

$$x_t = A^s x_{t-s} + \sum_{j=0}^{s-1} A^j B u_{t-j}.$$

Applying this with $s = d$ yields

$$x_t = A^d x_{t-d} + \sum_{j=0}^{d-1} A^j B u_{t-j},$$

and, for each $i = 1, \ldots, d$,

$$x_{t-i} = A^{d-i} x_{t-d} + \sum_{j=0}^{d-i-1} A^j B u_{t-i-j}.$$

Multiplying the latter identity by c_i, summing over $i = 1, \ldots, d$, and adding the first identity, we obtain

$$x_t + \sum_{i=1}^{d} c_i x_{t-i} = \Big(A^d + \sum_{i=1}^{d} c_i A^{d-i}\Big) x_{t-d} + \sum_{j=0}^{d-1} \Big(\sum_{\ell=0}^{j} c_\ell A^{\,j-\ell}\Big) B u_{t-j}.$$

The term multiplying x_{t-d} vanishes by Cayley–Hamilton, hence

$$x_t + \sum_{i=1}^{d} c_i x_{t-i} = \sum_{j=0}^{d-1} \Big(\sum_{\ell=0}^{j} c_\ell A^{\,j-\ell}\Big) B u_{t-j}.$$

Multiplying by C and using $y_t = Cx_t$, we get

$$y_t + \sum_{i=1}^{d} c_i y_{t-i} = \sum_{j=0}^{d-1} C\Big(\sum_{\ell=0}^{j} c_\ell A^{\,j-\ell}\Big) B u_{t-j}.$$

Therefore the claimed identity holds with

$$\alpha_i := -c_i, \qquad i = 1, \dots, d,$$

and

$$M_j := C\Big(\sum_{\ell=0}^{j} c_\ell A^{\,j-\ell}\Big) B, \qquad j = 0, \dots, d-1.$$

It remains to prove the bounds. Set

$$K := \prod_{j=1}^{d} (1 + |\lambda_j|), \qquad \beta := \max\{1, \|A\|_\infty\}.$$

Since $c_i = (-1)^i e_i(\lambda_1, \dots, \lambda_d)$, where e_i is the ith elementary symmetric polynomial,

$$|c_i| \le e_i(|\lambda_1|, \dots, |\lambda_d|) \le \prod_{j=1}^{d} (1 + |\lambda_j|) = K.$$

Hence

$$|\alpha_i| = |c_i| \le K.$$

Also, for each $j = 0, \dots, d-1$,

$$\|M_j\|_\infty \le \|C\|_\infty \|B\|_\infty \sum_{\ell=0}^{j} |c_\ell|\, \|A\|_\infty^{j-\ell} \le \|C\|_\infty \|B\|_\infty \sum_{\ell=0}^{j} K\, \beta^{\,j-\ell} \le dK \|C\|_\infty \|B\|_\infty \beta^{\,d-1}.$$

Thus both families of coefficients are bounded by

$$L := K\Big(1 + d\|C\|_\infty \|B\|_\infty \beta^{\,d-1}\Big) = \left(\prod_{j=1}^{d} (1 + |\lambda_j|)\right) \left(1 + d\|C\|_\infty \|B\|_\infty \max\{1, \|A\|_\infty\}^{\,d-1}\right).$$

This proves the result. □

D Spectral filtering approximation: stable real-diagonalizable case

Proposition 18. *Let $\{u_t\}_{t=1}^{T} \subset \mathbb{R}^m$ be as in Section 3, and consider the noiseless LDS*

$$x_t = Ax_{t-1} + Bu_t, \qquad y_t = Cx_t, \qquad x_0 = 0,$$

where $A = V D V^{-1}$ for some $V \in \mathrm{GL}_d(\mathbb{R})$ and $D = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$ with $\lambda_1, \dots, \lambda_d \in (-1, 1)$, $B \in \mathbb{R}^{d \times m}$, and $C \in \mathbb{R}^{p \times d}$. Let $\phi_1, \dots, \phi_T \in \mathbb{R}^T$ denote the spectral filters as defined in Definition 13, and for each $t = 1, \dots, T$, let $\tilde{u}_t \in \mathbb{R}^{m \times T}$ denote the zero-padded input history.

Fix $\varepsilon > 0$ and set the number of filters to

$$h \;=\; h(\varepsilon, T) \;:=\; 2\,\log T \,\cdot\, \log\!\left(\frac{C_0\, T^6}{\varepsilon}\right),$$

where $C_0 = 4\pi \cdot K^2 \cdot \|C\|_F^2 \cdot \|B\|_F^2 \cdot \|V\|_F^2 \cdot \|V^{-1}\|_F^2$. Then there exist coefficients

$$\{M_s \mid s \in [h]\} \;\subset\; \mathbb{R}^{p \times m},$$

depending only on (A, B, C), such that the predictor

$$\widehat{y}_t \;:=\; \sum_{s=1}^{h} M_s \left(\tilde{u}_t\, \phi_s\right)$$

satisfies

$$\sum_{t=1}^{T} \| y_t - \widehat{y}_t \|_2^2 \;\le\; \varepsilon.$$

Moreover, the coefficients may be chosen so that

$$\|M_s\|_2 \;\le\; \|V\|_F \cdot \|V^{-1}\|_F \cdot \|C\|_F \cdot \|B\|_F \cdot \min\left(\frac{1}{1-\max_{1\le i\le d}|\lambda_i|}, 2\sqrt{\pi}\cdot\sqrt{T}\cdot e^{-\frac{s}{4\log T}}\right)$$

where V and D are the real diagonalization above.

Proof. This lemma is identical to Theorem 12 but we still have to prove the bounds on the norm of M_s. Following the notation we used in the proof of Theorem 12, we have that:

$$\|M_s\|_2 \le \sum_{r=1}^{d} \|M_{r,s}\|_2 = \sum_{r=1}^{d} \left\|S_r\left(\phi_s^\top \mu_{\lambda_r}\right)\right\|_2 = \sum_{r=1}^{d} \|S_r\|_2 \cdot |\phi_s^\top \mu_{\lambda_r}|$$

Notice that:

$$\begin{aligned}\sum_{r=1}^{d} \|S_r\|_2 &= \sum_{r=1}^{d} \|CVe_i(e_i^\top V^{-1}B)\|_2 \\ &\le \sum_{r=1}^{d} \|CVe_i\|_2 \cdot \|e_i^\top V^{-1}B\|_2 \\ &\le \|V\|_F \cdot \|V^{-1}\|_F \cdot \|C\|_F \cdot \|B\|_F\end{aligned}$$

Also, we have that:

$$|\phi_s^\top \mu_{\lambda_r}| \le \|\phi_s\|_2 \cdot \|\mu_{\lambda_r}\|_2 \le 1 \cdot \frac{1}{1-|\lambda_r|}$$

since ϕ_s has norm 1 and the entries of μ_{λ_r} sum up to be at most $\frac{1}{1-|\lambda_r|}$. Combining this with Theorem 15 we get:

$$|\phi_s^\top \mu_{\lambda_r}| \le 2\sqrt{\pi}\cdot T^{\frac{1}{2}} \cdot e^{-\frac{s}{4\log T}}$$

Combining the above inequalities we get the desired bound. □

E Finite-memory filters for strictly stable LDS

The following lemma formalises Section 4.1.2: a strictly stable semisimple LDS admits a finite-memory input filter of length independent of the state dimension.

Lemma 19. *Consider an LDS $x_t = Ax_{t-1} + Bu_t$, $y_t = Cx_t$ with $x_0 = 0$. Suppose A is semisimple over $\mathbb{C}$ and every eigenvalue satisfies $|\lambda| \le 1 - 1/L$ for some $L \ge 1$. Write $A = VDV^{-1}$ and set $C_{\mathrm{fir},0} := \|CV\|\,\|V^{-1}B\|$. Then, for every $\varepsilon > 0$, setting*

$$L_\varepsilon \;=\; \left\lceil L\,\log\left(\frac{K\,C_{\mathrm{fir},0}\,L\sqrt{T}}{\sqrt{\varepsilon}}\right)\right\rceil,$$

implies that there exists matrices $Q_0, \ldots, Q_{L_\varepsilon - 1}$ such that

$$\sum_{\ell=0}^{L_\varepsilon-1} \|Q_\ell\| \le C_{\mathrm{fir},0} L,$$

and

$$\sum_{t=1}^{T} \left\| y_t - \sum_{\ell=0}^{L_\varepsilon-1} Q_\ell\, u_{t-\ell} \right\|_2^2 \;\le\; \varepsilon.$$

Proof. Since D is diagonal and $\rho(D) \le 1 - 1/L$,

$$\|CA^\tau B\| \le \|CV\| \, \|V^{-1}B\| \, (1 - 1/L)^\tau,$$

Set $Q_\ell := CA^\ell B$ for $\ell = 0, \dots, L_\varepsilon - 1$ and recall that $y_t = \sum_{\ell \ge 0} CA^\ell B \, u_{t-\ell}$ is the exact convolution. The truncation error at time t is at most

$$\sum_{\tau \ge L_\varepsilon} \|CA^\tau B\| \, K \;\le\; K \, C_{\mathrm{fir},0} \, L \, (1 - 1/L)^{L_\varepsilon} \;\le\; K \, C_{\mathrm{fir},0} \, L \, e^{-L_\varepsilon / L}.$$

Choosing L_ε as in the statement gives

$$e^{-L_\varepsilon / L} \le \frac{\sqrt{\varepsilon}}{K \, C_{\mathrm{fir},0} \, L\sqrt{T}}.$$

Thus, for every $t \le T$,

$$\Big\| y_t - \sum_{\ell=0}^{L_\varepsilon - 1} Q_\ell u_{t-\ell} \Big\|_2 \le \sqrt{\varepsilon / T},$$

and summing over $t = 1, \dots, T$ gives the claimed total error bound. Finally,

$$\sum_{\ell=0}^{L_\varepsilon - 1} \|Q_\ell\| \le C_{\mathrm{fir},0} \sum_{\ell=0}^{L_\varepsilon - 1} (1 - 1/L)^\ell \le C_{\mathrm{fir},0} L,$$

which proves the lemma. □

F Unified approximation: AR plus spectral and finite-memory filters

Theorem 20. *Consider the following LDS*

$$x_t = Ax_{t-1} + Bu_t, \qquad y_t = Cx_t, \qquad x_0 = 0,$$

where $A \in \mathbb{R}^{d \times d}$, $B \in \mathbb{R}^{d \times m}$, and $C \in \mathbb{R}^{p \times d}$. Extend the input and output sequences by $u_t := 0$ and $y_t := 0$ for all $t \le 0$, and define $\tilde{u}_t$ using this zero-padded input sequence.

Define the split parameter

$$\tau_k := \log(k + 2).$$

Let $\mu_1, \dots, \mu_q$ be the eigenvalues of A counted by the instability complexity Definition 1—the non-semisimple eigenvalues together with the semisimple eigenvalues outside $(-1, 1)$ with $|\lambda_i| > 1 - 1/\tau_k$, so that $q \le k$; if $q < k$ we pad the AR coefficients below by zeros. Let $\lambda_1, \dots, \lambda_r$ be the real semisimple eigenvalues of A with

$$1 - \frac{1}{\tau_k} \;<\; |\lambda_a| \;<\; 1, \qquad a = 1, \dots, r,$$

counted with algebraic multiplicity; we call these the spectral group*. The remaining $d - r - q$ eigenvalues are semisimple with modulus at most $1 - 1/\tau_k$ and form the* strictly stable *block.*

Using the real invariant subspaces associated with these three spectral groups (Lemma 16 applied separately to the spectral and to the strictly stable parts), there exist $P \in \mathrm{GL}_d(\mathbb{R})$ and matrices $A_{\mathrm{ustb}} \in \mathbb{R}^{q \times q}$ and $A_{\mathrm{fast}} \in \mathbb{R}^{(d-r-q) \times (d-r-q)}$ (with A_{fast} semisimple of spectral radius at most $1 - 1/\tau_k$) such that

$$A = P \begin{pmatrix} \Lambda & 0 & 0 \\ 0 & A_{\mathrm{ustb}} & 0 \\ 0 & 0 & A_{\mathrm{fast}} \end{pmatrix} P^{-1}, \qquad \Lambda := \mathrm{diag}(\lambda_1, \dots, \lambda_r).$$

The eigenvalues of A_{ustb} are precisely $\mu_1, \dots, \mu_q$ (counted with algebraic multiplicity).

Define the real block matrices

$$P^{-1}B = \begin{bmatrix} B_{\mathrm{st}} \\ B_{\mathrm{ustb}} \\ B_{\mathrm{fast}} \end{bmatrix}, \qquad CP = \begin{bmatrix} C_{\mathrm{st}} & C_{\mathrm{ustb}} & C_{\mathrm{fast}} \end{bmatrix},$$

where $B_{\mathrm{st}} \in \mathbb{R}^{r\times m}$, $B_{\mathrm{ustb}} \in \mathbb{R}^{q\times m}$, $C_{\mathrm{st}} \in \mathbb{R}^{p\times r}$, and $C_{\mathrm{ustb}} \in \mathbb{R}^{p\times q}$.

Define

$$L_{\mathrm{ustb}} := \left(\prod_{j=1}^{q}(1+|\mu_j|)\right)\left(1 + k\,\|C_{\mathrm{ustb}}\|_\infty\,\|B_{\mathrm{ustb}}\|_\infty\,\max\{1,\|A_{\mathrm{ustb}}\|_\infty\}^{\,\max\{q-1,0\}}\right)$$

and

$$\widetilde{L}_{\mathrm{ustb}} := (1+kL_{\mathrm{ustb}})\left(1-\frac{1}{\tau_k}\right)^{-k},$$

so that $\log \widetilde{L}_{\mathrm{ustb}} = O(\log(1+kL_{\mathrm{ustb}}) + k/\log(k+2))$ after absorbing the finitely many small values of k into the constant. Also set

$$\rho_{\mathrm{st}} := \max_{1\le a\le r}|\lambda_a|,$$

with the convention $\rho_{\mathrm{st}} = 0$ if $r = 0$. Define $C_{\mathrm{fast},0} := 0$ if the fast block is empty. Otherwise, write $A_{\mathrm{fast}} = V_{\mathrm{fast}} D_{\mathrm{fast}} V_{\mathrm{fast}}^{-1}$ over $\mathbb{C}$ and set

$$C_{\mathrm{fast},0} := \|C_{\mathrm{fast}}V_{\mathrm{fast}}\|\,\|V_{\mathrm{fast}}^{-1}B_{\mathrm{fast}}\|.$$

Define

$$h := \left\lceil 2\log T\cdot\log\left(\frac{4\pi K^2\,\widetilde{L}_{\mathrm{ustb}}^2\,\|C_{\mathrm{st}}\|_F^2\|B_{\mathrm{st}}\|_F^2\,T^6\,(1+kL_{\mathrm{ustb}})^2}{\varepsilon}\right)\right\rceil$$

and

$$L := \left\lceil \tau_k\,\log\left(\frac{(1+C_{\mathrm{fast},0})\,K\,(1+\tau_k)\,T\,(1+kL_{\mathrm{ustb}})^2}{\varepsilon}\right)\right\rceil.$$

For every $s = 1,\dots,h$, define

$$R_s := \|C_{\mathrm{st}}\|_F\,\|B_{\mathrm{st}}\|_F\cdot\min\left(\frac{1}{1-\rho_{\mathrm{st}}}, 2\sqrt{\pi}\sqrt{T}\,e^{-s/(4\log T)}\right).$$

Then there exist coefficients

$$\{\alpha_i\}_{i=1}^{k}\subset\mathbb{R},\qquad \{W_s\}_{1\le s\le h}\subset\mathbb{R}^{p\times m},\qquad \{Q_\ell\}_{0\le\ell\le L+k}\subset\mathbb{R}^{p\times m},$$

such that, for every $t = 1,\dots,T$, the predictor

$$\widehat{y}_t = \sum_{i=1}^{k}\alpha_i\,y_{t-i} + \sum_{s=1}^{h}W_s\,(\tilde{u}_t\phi_s) + \sum_{\ell=0}^{L+k}Q_\ell\,u_{t-\ell}$$

satisfies

$$\sum_{t=1}^{T}\|y_t-\widehat{y}_t\|_2^2 \le \varepsilon.$$

Moreover, for all relevant indices,

$$|\alpha_i|\le L_{\mathrm{ustb}},\qquad \|W_s\|_F \le \widetilde{L}_{\mathrm{ustb}}\,R_s,$$

and the Q_ℓ satisfy

$$\|Q_\ell\|_2 \le L_Q := mL_{\mathrm{ustb}} + \widetilde{L}_{\mathrm{ustb}}\,\|C_{\mathrm{st}}\|_F\,\|B_{\mathrm{st}}\|_F + (1+kL_{\mathrm{ustb}})\,C_{\mathrm{fast},0},$$

a constant depending only on the fixed system parameters.

Proof. Write $z_t := P^{-1}x_t = (s_t, e_t, f_t)$ according to the spectral, AR, and strictly stable blocks. The transformed dynamics are

$$s_t = \Lambda s_{t-1} + B_{\mathrm{st}}u_t,\qquad e_t = A_{\mathrm{ustb}}e_{t-1} + B_{\mathrm{ustb}}u_t,\qquad f_t = A_{\mathrm{fast}}f_{t-1} + B_{\mathrm{fast}}u_t,$$

and

$$y_t = C_{\mathrm{st}}s_t + C_{\mathrm{ustb}}e_t + C_{\mathrm{fast}}f_t =: y_t^{\mathrm{st}} + y_t^{\mathrm{ustb}} + y_t^{\mathrm{fast}}.$$

We may assume $k\ge 1$; otherwise the AR component is empty and the proof collapses to a direct application of Proposition 18 and Lemma 19.

Set

$$\eta := \frac{\varepsilon}{4(1+kL_{\mathrm{ustb}})^2}.$$

AR component via Cayley–Hamilton. Apply Lemma 17 to the q-dimensional subsystem $e_t = A_{\mathrm{ustb}} e_{t-1} + B_{\mathrm{ustb}} u_t$, $y_t^{\mathrm{ustb}} = C_{\mathrm{ustb}} e_t$, then pad the resulting AR coefficients by zeros so that they have length k. This yields

$$\{\alpha_i\}_{i=1}^{k} \subset \mathbb{R}, \qquad \{N_i\}_{i=0}^{k-1} \subset \mathbb{R}^{p\times m},$$

with $|\alpha_i| \le L_{\mathrm{ustb}}$ and $\|N_i\|_\infty \le L_{\mathrm{ustb}}$, satisfying (after extending u_t, y_t^{ustb} by zero for $t \le 0$)

$$y_t^{\mathrm{ustb}} = \sum_{i=1}^{k} \alpha_i y_{t-i}^{\mathrm{ustb}} + \sum_{i=0}^{k-1} N_i u_{t-i}, \qquad t = 1, \dots, T.$$

In particular $\sum_{i=1}^{k} |\alpha_i| \le k L_{\mathrm{ustb}}$. We perform this step first because the coefficients α_i are used in the treatment of the spectral block below. Write

$$p_\alpha(z) := z^k - \sum_{j=1}^{k} \alpha_j z^{k-j}$$

for the associated monic polynomial, so that $|p_\alpha(z)| \le 1 + k L_{\mathrm{ustb}}$ whenever $|z| \le 1$.

A shift identity for a single decay mode. The goal of this paragraph and the next is to express the AR residual $y_t^{\mathrm{st}} - \sum_{i=1}^{k} \alpha_i y_{t-i}^{\mathrm{st}}$ of the spectral component using only features available at time t (the current input history), with no reference to the lagged histories at times $t-1, \dots, t-k$. We first prove the required identity for a single scalar decay mode; the next paragraph applies it to each eigenvalue of Λ and sums.

For $\lambda \in (-1, 1) \setminus \{0\}$, define the geometric convolution of the input sequence with the decaying weights $1, \lambda, \lambda^2, \dots$:

$$G_\lambda(t) := \sum_{\tau=0}^{t-1} \lambda^\tau \, u_{t-\tau} \in \mathbb{R}^m, \qquad t \ge 1,$$

and $G_\lambda(t) := 0$ for $t \le 0$. This is the convention-consistent extension: for $t \le 0$ every input appearing in the sum is zero by the zero-padding $u_s = 0$, $s \le 0$. Splitting off the $\tau = 0$ term and re-indexing the remaining sum shows that G_λ obeys the one-step recurrence

$$G_\lambda(t) = u_t + \sum_{\tau=1}^{t-1} \lambda^\tau u_{t-\tau} = u_t + \lambda \sum_{\sigma=0}^{t-2} \lambda^\sigma u_{(t-1)-\sigma} = u_t + \lambda \, G_\lambda(t-1), \qquad t \ge 1.$$

Solving this recurrence for the lagged value gives $G_\lambda(t-1) = \lambda^{-1}\big(G_\lambda(t) - u_t\big)$: the value of the convolution one step in the past is recovered from its current value by dividing out one factor of λ and removing the contribution of the most recent input. Iterating this observation expresses *any* lag of G_λ through its current value:

Claim 20.1. For every $j \ge 0$ and every $t \in \mathbb{Z}$,

$$G_\lambda(t-j) = \lambda^{-j} G_\lambda(t) - \sum_{i=0}^{j-1} \lambda^{i-j} \, u_{t-i}. \tag{13}$$

Proof of claim. Induction on j. For $j = 0$ both sides equal $G_\lambda(t)$ (the empty sum is zero). Assume (13) holds for some $j \ge 0$. Applying the relation $G_\lambda(s-1) = \lambda^{-1}\big(G_\lambda(s) - u_s\big)$ at $s = t - j$ and then substituting the inductive hypothesis for $G_\lambda(t-j)$,

$$G_\lambda(t-j-1) = \lambda^{-1} G_\lambda(t-j) - \lambda^{-1} u_{t-j} = \lambda^{-(j+1)} G_\lambda(t) - \sum_{i=0}^{j-1} \lambda^{i-(j+1)} u_{t-i} - \lambda^{-1} u_{t-j}.$$

The last term is the $i = j$ summand $\lambda^{i-(j+1)} u_{t-i}\big|_{i=j}$, so the two negative terms merge into $\sum_{i=0}^{j} \lambda^{i-(j+1)} u_{t-i}$, which is (13) for $j+1$. The zero-padding conventions make the recurrence, and hence every step above, valid for all $t \in \mathbb{Z}$, including $t - j \le 0$; no boundary cases arise. $\diamond$

In words, (13) says: a lagged convolution $G_\lambda(t-j)$ equals the *current* convolution $G_\lambda(t)$ rescaled by λ^{-j}, minus a correction involving only the j most recent inputs $u_t, \dots, u_{t-j+1}$ (which the current convolution contains but the lagged one does not).

We now apply (13) to the AR residual of G_λ. Substituting (13) for each lag $j = 1, \dots, k$,

$$G_\lambda(t) - \sum_{j=1}^{k} \alpha_j \, G_\lambda(t-j) = G_\lambda(t) - \sum_{j=1}^{k} \alpha_j \lambda^{-j} \, G_\lambda(t) + \sum_{j=1}^{k} \alpha_j \sum_{i=0}^{j-1} \lambda^{i-j} \, u_{t-i}.$$

Two collections of terms appear. First, the multiplier of $G_\lambda(t)$ is

$$1 - \sum_{j=1}^{k} \alpha_j \lambda^{-j} = \frac{\lambda^k - \sum_{j=1}^{k} \alpha_j \lambda^{k-j}}{\lambda^k} = \frac{p_\alpha(\lambda)}{\lambda^k},$$

by the definition of p_α. Second, in the double sum, exchanging the order of summation (the pairs (j, i) with $1 \le j \le k$, $0 \le i \le j-1$ are exactly the pairs with $0 \le i \le k-1$, $i+1 \le j \le k$) collects the coefficient of each fixed input u_{t-i}:

$$\sum_{j=1}^{k} \alpha_j \sum_{i=0}^{j-1} \lambda^{i-j} \, u_{t-i} = \sum_{i=0}^{k-1} \underbrace{\Big(\sum_{j=i+1}^{k} \alpha_j \, \lambda^{\,i-j} \Big)}_{=:\, r_i(\lambda)} \, u_{t-i}.$$

Altogether we have shown, for every t,

$$G_\lambda(t) - \sum_{j=1}^{k} \alpha_j G_\lambda(t-j) = \frac{p_\alpha(\lambda)}{\lambda^k} \, G_\lambda(t) + \sum_{i=0}^{k-1} r_i(\lambda) \, u_{t-i}. \tag{14}$$

The left-hand side involves the convolution at $k+1$ different times $t, t-1, \dots, t-k$; the right-hand side involves it only at time t, plus the k most recent inputs. This is the mechanism that will remove the shifted spectral-filter features.

The rescaling factor and the input coefficients are controlled as follows. Since $|\lambda| < 1$, each power satisfies $|\lambda|^{k-j} \le 1$, so $|p_\alpha(\lambda)| \le 1 + \sum_{j=1}^{k} |\alpha_j| \le 1 + k L_{\mathrm{ustb}}$, using the bound $|\alpha_j| \le L_{\mathrm{ustb}}$ from the AR step. If in addition $|\lambda| > 1 - 1/\tau_k$, then $|\lambda|^{-k} \le \left(1 - \frac{1}{\tau_k}\right)^{-k}$, hence

$$\begin{aligned} \left| \frac{p_\alpha(\lambda)}{\lambda^k} \right| &\le (1 + k L_{\mathrm{ustb}}) \left(1 - \frac{1}{\tau_k}\right)^{-k} = \widetilde{L}_{\mathrm{ustb}}, \\ |r_i(\lambda)| &\le \sum_{j=i+1}^{k} |\alpha_j| \, |\lambda|^{\,i-j} \le \Big(\sum_{j=1}^{k} |\alpha_j| \Big) |\lambda|^{-k} \le \widetilde{L}_{\mathrm{ustb}}, \end{aligned} \tag{15}$$

where in the bound on $r_i(\lambda)$ we used $|\lambda|^{\,i-j} \le |\lambda|^{-k}$, valid because $|\lambda| < 1$ and $0 < j - i \le k$. It is exactly here that the spectral group's lower threshold $1 - 1/\tau_k$ is used: without it, the factor $|\lambda|^{-k}$ would be unbounded as $\lambda \to 0$.

Spectral component: AR residual first, then spectral filters. We first express y^{st} in the form required by the previous paragraph, namely as a weighted sum of single-mode convolutions. Unrolling the recursion $s_t = \Lambda s_{t-1} + B_{\mathrm{st}} u_t$ from $s_0 = 0$ gives $s_t = \sum_{\tau=0}^{t-1} \Lambda^\tau B_{\mathrm{st}} \, u_{t-\tau}$, and hence

$$y^{\mathrm{st}}_t = C_{\mathrm{st}} \, s_t = \sum_{\tau=0}^{t-1} C_{\mathrm{st}} \Lambda^\tau B_{\mathrm{st}} \, u_{t-\tau}.$$

Because $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_r)$ is diagonal, its powers decompose as $\Lambda^\tau = \sum_{a=1}^{r} \lambda_a^\tau \, e_a e_a^\top$, where $e_1, \dots, e_r$ is the standard basis of $\mathbb{R}^r$. Therefore

$$C_{\mathrm{st}} \Lambda^\tau B_{\mathrm{st}} = \sum_{a=1}^{r} \lambda_a^\tau \underbrace{(C_{\mathrm{st}} e_a)}_{a\text{-th column of } C_{\mathrm{st}}} \underbrace{(e_a^\top B_{\mathrm{st}})}_{a\text{-th row of } B_{\mathrm{st}}} = \sum_{a=1}^{r} \lambda_a^\tau \, c_a, \qquad c_a := (C_{\mathrm{st}})_{:,a} \, (B_{\mathrm{st}})_{a,:} \in \mathbb{R}^{p \times m},$$

where each c_a is a fixed rank-one matrix recording how the input enters mode a and how mode a appears in the output. Substituting this into the expression for y_t^{st} and exchanging the two finite sums,

$$y_t^{\mathrm{st}} = \sum_{\tau=0}^{t-1} \sum_{a=1}^{r} \lambda_a^\tau \, c_a \, u_{t-\tau} = \sum_{a=1}^{r} c_a \sum_{\tau=0}^{t-1} \lambda_a^\tau \, u_{t-\tau} = \sum_{a=1}^{r} c_a \, G_{\lambda_a}(t).$$

That is, the spectral component of the output is a weighted sum of r single-mode convolutions, one per eigenvalue of Λ — precisely the objects to which the shift identity (14) applies. Every λ_a in the spectral group satisfies $|\lambda_a| > 1 - 1/\tau_k$ (in particular $\lambda_a \neq 0$), so (14) applies to each mode. Summing (14) against c_a over $a = 1, \dots, r$,

$$y_t^{\mathrm{st}} - \sum_{i=1}^{k} \alpha_i y_{t-i}^{\mathrm{st}} = \sum_{a=1}^{r} \widetilde{c}_a \, G_{\lambda_a}(t) + \sum_{i=0}^{k-1} D_i \, u_{t-i}, \qquad \widetilde{c}_a := \frac{p_\alpha(\lambda_a)}{\lambda_a^k} \, c_a, \quad D_i := \sum_{a=1}^{r} c_a \, r_i(\lambda_a). \tag{16}$$

The first sum on the right-hand side of (16) is, at the same time t, the output of the LDS $(s_t, B_{\mathrm{st}}, \widetilde{C}_{\mathrm{st}}, \Lambda)$, where

$$\widetilde{C}_{\mathrm{st}} := C_{\mathrm{st}} \, \mathrm{diag}\Big(\frac{p_\alpha(\lambda_a)}{\lambda_a^k} \Big)_{a=1}^{r} :$$

it has the same eigenvalues $\lambda_1, \dots, \lambda_r$ as the original spectral block, and by (15) its readout satisfies $\|\widetilde{C}_{\mathrm{st}}\|_F \leq \widetilde{L}_{\mathrm{ustb}} \|C_{\mathrm{st}}\|_F$. By the choice of h and Proposition 18 applied to $(s_t, B_{\mathrm{st}}, \widetilde{C}_{\mathrm{st}}, \Lambda)$ with target error η, there are matrices $W_s \in \mathbb{R}^{p \times m}$ such that

$$\sum_{t=1}^{T} \Big\| \sum_{a=1}^{r} \widetilde{c}_a \, G_{\lambda_a}(t) - \sum_{s=1}^{h} W_s (\tilde{u}_t \phi_s) \Big\|_2^2 \leq \eta, \tag{17}$$

and, for every s, $\|W_s\|_F \leq \widetilde{L}_{\mathrm{ustb}} R_s$ (the bound of Proposition 18 with $\|C\|_F$ replaced by $\|\widetilde{C}_{\mathrm{st}}\|_F \leq \widetilde{L}_{\mathrm{ustb}} \|C_{\mathrm{st}}\|_F$). The constant in h is exactly Proposition 18's requirement for this rescaled system: the factor $\widetilde{L}_{\mathrm{ustb}}^2$ absorbs the rescaled readout, the factor $(1 + k L_{\mathrm{ustb}})^2$ absorbs $1/\eta$, and K^2 is the correct power (Proposition 18's C_0 carries K^2). For the boundary matrices, (15) and Cauchy–Schwarz over a give

$$\|D_i\|_2 \leq \widetilde{L}_{\mathrm{ustb}} \sum_{a=1}^{r} \|c_a\|_F \leq \widetilde{L}_{\mathrm{ustb}} \, \|C_{\mathrm{st}}\|_F \, \|B_{\mathrm{st}}\|_F, \qquad i = 0, \dots, k-1. \tag{18}$$

Note that, in contrast with the shifted construction, no approximation of the lagged outputs y_{t-i}^{st} is performed: the AR residual of y^{st} is approximated directly at time t, so no shifted features $\tilde{u}_{t-i}\phi_s$ appear and the spectral-filter error (17) will not be amplified by the AR coefficients.

Strictly stable component via finite-memory filter. The eigenvalues of A_{fast} have modulus at most $1 - 1/\tau_k$, so in Lemma 19 the stability parameter is τ_k. Applying that lemma with target error η gives a required memory length

$$L_\eta = \left\lceil \tau_k \log\left(\frac{K \, C_{\mathrm{fast},0} \, \tau_k \sqrt{T}}{\sqrt{\eta}} \right) \right\rceil = \left\lceil \tau_k \log\left(\frac{2K \, C_{\mathrm{fast},0} \, \tau_k \sqrt{T}(1 + k L_{\mathrm{ustb}})}{\sqrt{\varepsilon}} \right) \right\rceil.$$

The theorem uses the slightly larger, cleaner choice of L in the statement, which dominates L_η up to harmless constants and leaves room for the later multiplication by $(1 + k L_{\mathrm{ustb}})$ in the AR error propagation for this block. By Lemma 19 applied to the subsystem $f_t = A_{\mathrm{fast}} f_{t-1} + B_{\mathrm{fast}} u_t$, $y_t^{\mathrm{fast}} = C_{\mathrm{fast}} f_t$, with the choice of L in the theorem statement there are matrices $\widetilde{Q}_0, \dots, \widetilde{Q}_{L-1} \in \mathbb{R}^{p \times m}$ such that, with

$$\hat{y}_t^{\mathrm{fast}} := \sum_{\ell=0}^{L-1} \widetilde{Q}_\ell \, u_{t-\ell}, \qquad \sum_{t=1}^{T} \|y_t^{\mathrm{fast}} - \hat{y}_t^{\mathrm{fast}}\|_2^2 \leq \eta,$$

and each $\|\widetilde{Q}_\ell\| \leq C_{\mathrm{fast},0}$. Set $\hat{y}_t^{\mathrm{fast}} := 0$ for $t \leq 0$.

It will be convenient to record now the AR residual of this FIR approximation, which is again a finite-memory filter, of length $L + k$:

$$\hat{y}_t^{\mathrm{fast}} - \sum_{i=1}^{k} \alpha_i \hat{y}_{t-i}^{\mathrm{fast}} \; = \; \sum_{\ell=0}^{L+k} F_\ell \, u_{t-\ell}, \qquad F_\ell \; := \; \widetilde{Q}_\ell \; - \sum_{\substack{i=1,\dots,k \\ 0 \leq \ell - i \leq L-1}} \alpha_i \, \widetilde{Q}_{\ell - i},$$

for $\ell = 0, \dots, L+k$, with the conventions $\widetilde{Q}_\ell := 0$ for $\ell \geq L$ and that the sum is zero if no i satisfies the conditions. Using $\|\widetilde{Q}_\ell\|_2 \leq C_{\mathrm{fast},0}$ and $|\alpha_i| \leq L_{\mathrm{ustb}}$, we obtain

$$\|F_\ell\|_2 \leq (1 + kL_{\mathrm{ustb}})\, C_{\mathrm{fast},0}.$$

Combining the three pieces. Substituting $y^{\mathrm{ustb}}_{t-i} = y_{t-i} - y^{\mathrm{st}}_{t-i} - y^{\mathrm{fast}}_{t-i}$ into the AR identity gives

$$\begin{aligned} y_t = &\sum_{i=1}^{k} \alpha_i y_{t-i} + \sum_{i=0}^{k-1} N_i u_{t-i} \\ &+ \Big(y^{\mathrm{st}}_t - \sum_{i=1}^{k} \alpha_i y^{\mathrm{st}}_{t-i}\Big) + \Big(y^{\mathrm{fast}}_t - \sum_{i=1}^{k} \alpha_i y^{\mathrm{fast}}_{t-i}\Big), \qquad t = 1, \dots, T. \end{aligned}$$

Define

$$\widehat{y}_t := \sum_{i=1}^{k} \alpha_i y_{t-i} + \sum_{i=0}^{k-1} N_i u_{t-i} + \Big(\sum_{s=1}^{h} W_s(\tilde{u}_t \phi_s) + \sum_{i=0}^{k-1} D_i\, u_{t-i}\Big) + \Big(\hat{y}^{\mathrm{fast}}_t - \sum_{i=1}^{k} \alpha_i \hat{y}^{\mathrm{fast}}_{t-i}\Big),$$

i.e. the AR residual of y^{st} is replaced by its time-t approximation from (16)–(17), while the AR residual of y^{fast} is replaced by the AR residual of its FIR approximation. Let

$$\delta_t := \sum_{a=1}^{r} \widetilde{c}_a\, G_{\lambda_a}(t) - \sum_{s=1}^{h} W_s(\tilde{u}_t \phi_s), \qquad \gamma_t := y^{\mathrm{fast}}_t - \hat{y}^{\mathrm{fast}}_t$$

(with the convention that both are zero for $t \leq 0$). By (16),

$$y_t - \widehat{y}_t = \delta_t + \Big(\gamma_t - \sum_{i=1}^{k} \alpha_i \gamma_{t-i}\Big), \qquad t = 1, \dots, T.$$

The first term is bounded directly by (17); only the second term passes through the AR filter. By the triangle inequality (in $\ell_2(\{1, \dots, T\})$), Young's inequality for convolution, and the bound $|\alpha_i| \leq L_{\mathrm{ustb}}$,

$$\begin{aligned} \Big(\sum_{t=1}^{T} \|y_t - \widehat{y}_t\|_2^2\Big)^{1/2} &\leq \Big(\sum_t \|\delta_t\|_2^2\Big)^{1/2} + (1 + kL_{\mathrm{ustb}})\Big(\sum_t \|\gamma_t\|_2^2\Big)^{1/2} \\ &\leq \sqrt{\eta} + (1 + kL_{\mathrm{ustb}})\sqrt{\eta} \leq \sqrt{\varepsilon}, \end{aligned}$$

so $\sum_{t=1}^{T} \|y_t - \widehat{y}_t\|_2^2 \leq \varepsilon$.

Putting everything together. Collecting all coefficient matrices that multiply lagged inputs, define

$$Q_\ell := \mathbf{1}\{\ell \leq k-1\}\,\big(N_\ell + D_\ell\big) + F_\ell, \qquad \ell = 0, \dots, L+k.$$

Then

$$\sum_{i=0}^{k-1} N_i u_{t-i} + \sum_{i=0}^{k-1} D_i u_{t-i} + \sum_{\ell=0}^{L+k} F_\ell\, u_{t-\ell} = \sum_{\ell=0}^{L+k} Q_\ell\, u_{t-\ell},$$

so $\widehat{y}_t$ has exactly the claimed form

$$\widehat{y}_t = \sum_{i=1}^{k} \alpha_i y_{t-i} + \sum_{s=1}^{h} W_s(\tilde{u}_t \phi_s) + \sum_{\ell=0}^{L+k} Q_\ell\, u_{t-\ell}.$$

The bound $|\alpha_i| \leq L_{\mathrm{ustb}}$ was established in the AR step, and $\|W_s\|_F \leq \widetilde{L}_{\mathrm{ustb}} R_s$ in the spectral step. Finally, by the triangle inequality, $\|N_\ell\|_2 \leq mL_{\mathrm{ustb}}$ (from $\|N_\ell\|_\infty \leq L_{\mathrm{ustb}}$), (18), and the bound on $\|F_\ell\|_2$,

$$\|Q_\ell\|_2 \leq mL_{\mathrm{ustb}} + \widetilde{L}_{\mathrm{ustb}}\, \|C_{\mathrm{st}}\|_F\, \|B_{\mathrm{st}}\|_F + (1 + kL_{\mathrm{ustb}})\, C_{\mathrm{fast},0} = L_Q.$$

This completes the proof.

Memory complexity. The memory complexity of the algorithm is $O(k+h+L) = O(k+h+L)$. Since $\log L_{\mathrm{ustb}} = \tilde{O}(k)$, we get that the memory complexity is $O(k+h+L) = \tilde{O}(k)$. □

F.1 Robustness to the spectral cutoff

Here k is a fixed AR budget, which need not equal the instability complexity of the system under Definition 1. We relax the decay assumption on the finite-memory block while retaining the other block and norm conditions. As in Theorem 20, ε denotes the target cumulative squared approximation error; δ denotes the cutoff shift.

Proposition 21 (Robustness to the cutoff)**.** *Fix $T \geq 2$, $k \geq 1$, and $0 < \varepsilon \leq T$, and write*

$$\tau := \log(k+2), \qquad r_0 := 1 - \frac{1}{\tau}.$$

Consider a noiseless LDS with $x_0 = 0$ and $\|u_t\|_2 \leq K$, where $K > 0$. Suppose it admits the real block decomposition of Theorem 20, with an AR block of dimension $q \leq k$, a diagonal real spectral block whose eigenvalues satisfy $r_0 < |\lambda_a| < 1$, and a semisimple finite-memory block satisfying

$$\rho(A_{\mathrm{fast}}) \leq r_0 + \delta, \qquad 0 \leq \delta \leq \frac{1}{2\tau}.$$

Define L_{ustb}, $\widetilde{L}_{\mathrm{ustb}}$, R_s, L_Q, and $G := C_{\mathrm{fast},0}$ from these blocks as in Theorem 20. Choose h, L by that theorem's formulas with approximation level ε and the nominal split parameter τ.

Then the same unified predictor class, with AR order k, filter count h, and input lags $0, \ldots, L+k$, contains a predictor $f^\star$ satisfying

$$\sum_{t=1}^{T} \|y_t - f_t^\star\|_2^2 \leq \varepsilon\big[1 + 16KG\delta \log^2(k+2)\big]. \tag{19}$$

Here ε is the target cumulative squared approximation error, as in Theorem 20: at the nominal cutoff $\delta = 0$, the guarantee is

$$\sum_{t=1}^{T} \|y_t - f_t^\star\|_2^2 \leq \varepsilon.$$

For fixed ε, G, K, k, the perturbed bound is $\varepsilon + O(\delta)$: the cutoff shift adds only $O(\delta)$ to the cumulative approximation error guarantee. All coefficient norm bounds in Theorem 20 remain valid.

Proof. Use the construction and error decomposition from the proof of Theorem 20. Only the finite-memory error estimate changes. Write

$$M := 1 + kL_{\mathrm{ustb}} \geq 2, \qquad \theta := \delta\tau \in [0, 1/2].$$

With zero-padded inputs, the error from retaining only the first L impulse-response terms of the fast block is

$$\begin{aligned}\gamma_t &:= y_t^{\mathrm{fast}} - \sum_{\ell=0}^{L-1} C_{\mathrm{fast}} A_{\mathrm{fast}}^{\ell} B_{\mathrm{fast}}\, u_{t-\ell} \\ &= \sum_{\ell=L}^{t-1} C_{\mathrm{fast}} A_{\mathrm{fast}}^{\ell} B_{\mathrm{fast}}\, u_{t-\ell}.\end{aligned}$$

Thus γ_t consists of the omitted contributions from older inputs. An empty sum is zero, and we set $\gamma_t = 0$ for $t \leq 0$.

Put $R := r_0 + \delta = 1 - (1-\theta)/\tau < 1$. In the diagonalization $A_{\mathrm{fast}} = V_{\mathrm{fast}} D_{\mathrm{fast}} V_{\mathrm{fast}}^{-1}$, we have $\|D_{\mathrm{fast}}^{\ell}\|_2 \leq R^{\ell}$. The definition of G therefore gives

$$\begin{aligned}\|C_{\mathrm{fast}} A_{\mathrm{fast}}^{\ell} B_{\mathrm{fast}}\|_2 &\leq \|C_{\mathrm{fast}} V_{\mathrm{fast}}\|_2 \|D_{\mathrm{fast}}^{\ell}\|_2 \|V_{\mathrm{fast}}^{-1} B_{\mathrm{fast}}\|_2 \\ &\leq GR^{\ell}.\end{aligned}$$

Applying the triangle inequality and $\|u_s\|_2 \le K$ to the omitted tail,

$$\begin{aligned}\|\gamma_t\|_2 &\le \sum_{\ell=L}^{t-1} \|C_{\mathrm{fast}} A_{\mathrm{fast}}^{\ell} B_{\mathrm{fast}}\|_2 \|u_{t-\ell}\|_2 \\ &\le KG \sum_{\ell=L}^{t-1} R^{\ell} \le KG \sum_{\ell=L}^{\infty} R^{\ell} = \frac{KGR^L}{1-R}.\end{aligned}$$

Since $1 - R = (1-\theta)/\tau$ and $R^L \le e^{-(1-\theta)L/\tau}$, this gives the uniform bound

$$\|\gamma_t\|_2 \le b_\theta, \qquad b_\theta := \frac{KG\tau}{1-\theta} e^{-(1-\theta)L/\tau}.$$

The AR coefficients satisfy $1 + \sum_{i=1}^{k} |\alpha_i| \le 1 + kL_{\mathrm{ustb}} = M$. Hence, at each time t, another application of the triangle inequality gives

$$\begin{aligned}\left\| \gamma_t - \sum_{i=1}^{k} \alpha_i \gamma_{t-i} \right\|_2 &\le \|\gamma_t\|_2 + \sum_{i=1}^{k} |\alpha_i| \, \|\gamma_{t-i}\|_2 \\ &\le \left(1 + \sum_{i=1}^{k} |\alpha_i|\right) b_\theta \le M b_\theta.\end{aligned}$$

Squaring this bound and summing over the T rounds yields

$$\begin{aligned}\sum_{t=1}^{T} \left\| \gamma_t - \sum_{i=1}^{k} \alpha_i \gamma_{t-i} \right\|_2^2 &\le T M^2 b_\theta^2 \\ &= \frac{T(KGM\tau)^2}{(1-\theta)^2} e^{-2(1-\theta)L/\tau} =: \mathcal{B}(\theta).\end{aligned} \tag{20}$$

Keeping the nominal memory length,

$$L \ge \tau \log\!\left(\frac{(1+G)K(1+\tau)TM^2}{\varepsilon} \right), \tag{21}$$

gives $\mathcal{B}(0) \le \varepsilon^2/(M^2T)$ and $\mathcal{B}(1/2) \le 4KG\varepsilon\tau$. For fixed L, $\mathcal{B}$ is convex on $[0, 1/2]$, so

$$\mathcal{B}(\theta) \le (1-2\theta)\mathcal{B}(0) + 2\theta\mathcal{B}(1/2) \le \frac{\varepsilon^2}{M^2 T} + 8KG\varepsilon\delta\tau^2. \tag{22}$$

The same spectral tail estimate and choice of h give $\sum_t \|d_t\|_2^2 \le \varepsilon/M^2$ for the spectral residual error d_t. To combine the errors, write

$$e_t := \gamma_t - \sum_{i=1}^{k} \alpha_i \gamma_{t-i}.$$

The original error decomposition is $y_t - f_t^\star = d_t + e_t$, and (20) gives $\sum_{t=1}^{T} \|e_t\|_2^2 \le \mathcal{B}(\theta)$. Applying $\|a+b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$ with $a = d_t$ and $b = e_t$ at each time, then summing, gives

$$\begin{aligned}\sum_{t=1}^{T} \|y_t - f_t^\star\|_2^2 &= \sum_{t=1}^{T} \|d_t + e_t\|_2^2 \\ &\le 2\sum_{t=1}^{T} \|d_t\|_2^2 + 2\sum_{t=1}^{T} \|e_t\|_2^2 \\ &\le \frac{2\varepsilon}{M^2} + 2\mathcal{B}(\theta) \\ &\le \frac{2\varepsilon}{M^2} + \frac{2\varepsilon^2}{M^2 T} + 16KG\varepsilon\delta\tau^2 \\ &\le \varepsilon + 16KG\varepsilon\delta\tau^2,\end{aligned}$$

where the penultimate step substitutes (22). For the last step, $\varepsilon \le T$ and $M \ge 2$ imply

$$\frac{2\varepsilon}{M^2} + \frac{2\varepsilon^2}{M^2 T} \le \frac{4\varepsilon}{M^2} \le \varepsilon.$$

Finally, $R < 1$ preserves the bound G on each FIR impulse-response coefficient, so the coefficient norm argument in Theorem 20 applies unchanged. □

G Formal version of Theorem 7 and its proof

We first restate Theorem 7 with the precise filter count and the constants made explicit, then prove it.

Theorem 22 (Formal version of Theorem 7)**.** *Fix $d \ge 1$. Consider any observable LDS as in Eq. (1) with state dimension d and instability complexity k. Let $(u_t)_{t\ge1}$ be any input sequence whose generated outputs $(y_t)_{t\ge1}$ satisfy Assumption 2. Write $k := k$. Choose the approximation level $\varepsilon_\star$ as in the proof, and set the filter count h and the finite-memory length L to the corresponding values from Theorem 20. Initialize the VAW instances with the coefficient-norm bound R_x defined in Lemma 24. Logarithms in the asymptotic rates are truncated below at 1 to include small T and k. Treating the per-mode LDS parameters as constants, the filter count satisfies*

$$h \;=\; O(k \cdot \log T \cdot \log(\mathrm{poly}(T, k, 1/\varepsilon))) \;=\; \widetilde{O}(k).$$

Because the autoregressive coefficients can be exponentially large ($\log L_{\mathrm{ustb}} = O(k \log k)$), the finite-memory length satisfies

$$L \;=\; O\!\left(\log(k+2) \log \frac{\mathrm{poly}(T,k)\, L_{\mathrm{ustb}}^2}{\varepsilon}\right) \;=\; O(k \cdot \mathrm{polylog}(T, k, 1/\varepsilon)) \;=\; \widetilde{O}(k).$$

Then Algorithm 1 achieves

$$\frac{1}{T}\sum_{t=1}^{T} \ell_2^2(\widehat{y}_t, y_t) \le O\!\left(\frac{k^2 \log^3 k\, \log^3 T}{T}\right)$$

using $O\big(p\big(k + m(L+k+1) + hm\big)\big)$ or $O(k \cdot \log T \cdot \mathrm{polylog}(T,k)) = \widetilde{O}(k)$ learnable parameters in total if we are treating the parameters of the LDS as constants except T, k.

Proof of Theorem 22. Let $\varepsilon > 0$ be an approximation level to be chosen below, and let $f^\star$ be the comparator from Theorem 20, instantiated with the corresponding values of h and L. By construction, $f^\star$ belongs to the linear comparator class used by Algorithm 1. Decompose the cumulative squared loss as

$$\sum_{t=1}^{T} \ell_2^2(\widehat{y}_t, y_t) = \underbrace{\left(\sum_{t=1}^{T} \|\widehat{y}_t - y_t\|_2^2 - \sum_{t=1}^{T} \|f_t^\star - y_t\|_2^2\right)}_{(I)} + \underbrace{\sum_{t=1}^{T} \|f_t^\star - y_t\|_2^2}_{(II)}.$$

Term (I) is exactly the regret term controlled by Lemma 24. Treating the LDS parameters as constants we have,

$$(I) \le O\left((k + L + h)\left[\log T + \log(k + L + Th) + \log(k + L + h) + k \log k\right]\right).$$

Term (II) is the approximation error of the unified comparator. By Theorem 20,

$$(II) = \sum_{t=1}^{T} \|f_t^\star - y_t\|_2^2 \le \varepsilon.$$

Combining the two bounds gives

$$\sum_{t=1}^{T} \ell_2^2(\widehat{y}_t, y_t) \le O\left((k + L + h)\left[\log T + \log(k + L + Th) + \log(k + L + h) + k \log k\right]\right) + \varepsilon.$$

We now substitute the asymptotic dependence of h and L on T, k, ε. From Theorem 20 and $\log L_{\mathrm{ustb}} = O(k \log k)$,

$$h = O\left(\log T \left[\log \frac{T}{\varepsilon} + k \log(k+2) \right] \right),$$

and

$$L = O\left(\log(k+2) \left[\log \frac{T}{\varepsilon} + k \log(k+2) \right] \right).$$

Set

$$A_T := \log T + k \log(k+2), \qquad B_T := \log T + \log(k+2), \qquad S_\varepsilon := A_T + \log \frac{1}{\varepsilon}.$$

Then $h + L = O(B_T S_\varepsilon)$, and the logarithmic factor in the regret bound is $O(S_\varepsilon)$. Therefore

$$\sum_{t=1}^{T} \ell_2^2(\widehat{y}_t, y_t) \le O((k + B_T S_\varepsilon)\, S_\varepsilon) + \varepsilon.$$

The right-hand side is minimized, up to constants hidden in the big-O notation, when the marginal decrease from reducing the logarithmic regret term balances the marginal increase in approximation error:

$$\varepsilon_\star = \Theta(k + B_T A_T).$$

With this choice, $S_{\varepsilon_\star} = O(A_T)$ up to lower-order logarithmic terms, and hence

$$\frac{1}{T} \sum_{t=1}^{T} \ell_2^2(\widehat{y}_t, y_t) \le \frac{O((k + B_T A_T)\, A_T)}{T}.$$

Equivalently, writing $k = k$ and using the crude bounds $\log T + \log k = O(\log T \log k)$ and $\log T + k \log k = O(k \log T \log k)$, this is

$$\frac{1}{T} \sum_{t=1}^{T} \ell_2^2(\widehat{y}_t, y_t) \le O\left(\frac{k^2 \log^3 k \, \log^3 T}{T} \right).$$

Finally, the number of learnable parameters is the dimension of the unified feature vector times the number of output coordinates:

$$p \cdot d_{\mathrm{feat}} = p\big(k + m(L + k + 1) + hm\big) = O\big(p\big(k + mL + mh\big)\big).$$

Substituting $h = O(k \cdot \log T \cdot \log(\mathrm{poly}(T, k))) = \widetilde{O}(k)$ and, using $\log L_{\mathrm{ustb}} = O(k \log k)$, $L = O(k \cdot \mathrm{polylog}(T, k)) = \widetilde{O}(k)$, the feature dimension is $\widetilde{O}(k)$ and gives the stated $\widetilde{O}(k)$ parameter count. □

H The VAW forecaster and its per-coordinate regret

This appendix collects the standard online-linear-regression machinery used by Algorithm 1. The Vovk–Azoury–Warmuth (VAW) forecaster [Vovk, 2001] is an online algorithm for least-squares prediction with logarithmic regret against any fixed linear comparator. At each time step it maintains a covariance-like matrix A_t and a target vector v_t, and predicts via a forward-step ridge regression solve.

Algorithm 2 Vovk–Azoury–Warmuth forecaster

Input: feature dimension d, coefficient-norm bound $S \ge 1$.
Initialize $A_0 \leftarrow S^{-2} I_d$, $v_0 \leftarrow 0 \in \mathbb{R}^d$, $t \leftarrow 0$.
function PREDICT($a \in \mathbb{R}^d$)
    $t \leftarrow t + 1$; store $a_t \leftarrow a$; $A_t \leftarrow A_{t-1} + a_t a_t^\top$.
    return $\widehat{b}_t \leftarrow a_t^\top A_t^{-1} v_{t-1}$.
end function
function UPDATE($b_t \in \mathbb{R}$)
    $v_t \leftarrow v_{t-1} + b_t a_t$.
end function

The input S is an available upper bound on the Euclidean norm of the comparator coefficients. Its use in the initialization is part of the VAW subroutine; the prediction algorithm needs no separate regularization parameter.

Theorem 23 (Standard VAW regret bound [Hazan and Singh, 2026, Theorem A.2])**.** *Run Algorithm 2 with a coefficient-norm bound $S \geq 1$. Suppose $\|a_t\|_2 \leq R$ and $|b_t| \leq Y$ for every t. Then, for every $x^\star \in \mathbb{R}^d$ with $\|x^\star\|_2 \leq S$,*

$$\sum_{t=1}^{T}(\widehat{b}_t - b_t)^2 - \sum_{t=1}^{T}(a_t^\top x^\star - b_t)^2 \leq 1 + Y^2 \log\det\left(I_d + S^2 \sum_{t=1}^{T} a_t a_t^\top\right) \leq 1 + dY^2 \log\left(1 + \frac{TR^2S^2}{d}\right).$$

Proof. The standard VAW inequality with initial matrix $S^{-2}I_d$ has comparator penalty $S^{-2}\|x^\star\|_2^2 \leq 1$, which gives the first bound. For the second, the matrix $\sum_t a_t a_t^\top$ has trace at most TR^2. Concavity of $\log(1+x)$ applied to its d eigenvalues bounds the log-determinant by $d\log(1 + TR^2S^2/d)$. □

We apply this guarantee to the comparator supplied by Theorem 20. The bound on its norm is used only to initialize VAW; the comparator coefficients themselves are unknown.

Lemma 24 (VAW per-coordinate regret)**.** *Let $f^\star$ be the comparator from Theorem 20, and set*

$$D := k + m(L + k + 1) + mh.$$

Define the feature and coefficient-norm bounds

$$\begin{aligned} R_a^2 &:= kB_y^2 + K^2(L + k + 1 + T), \\ R_x^2 &:= 1 + kL_{\mathrm{ustb}}^2 + (L + k + 1)L_Q^2 \\ &\quad + 4\pi Th\,\widetilde{L}_{\mathrm{ustb}}^2 \|C_{\mathrm{st}}\|_F^2 \|B_{\mathrm{st}}\|_F^2. \end{aligned} \tag{23}$$

Initialize each VAW instance in Algorithm 1 with dimension D and coefficient-norm bound $S = R_x$. Then the total regret against $f^\star$ satisfies

$$R_T := \sum_{t=1}^{T} \|\widehat{y}_t - y_t\|_2^2 - \sum_{t=1}^{T} \|f_t^\star - y_t\|_2^2 \leq p + pB_y^2 D \log\left(1 + \frac{TR_a^2R_x^2}{D}\right).$$

With the remaining norm bounds and input/output dimensions fixed, and $\log L_{\mathrm{ustb}}, \log \widetilde{L}_{\mathrm{ustb}}, \log L_Q = O(k\log(k+2))$, this is

$$R_T = O((k + L + h)\left[\log(T+2) + \log(k + L + h + 2) + k\log(k+2)\right]).$$

Any available upper bounds on the system quantities in (23) may be used to form R_x, just as for the choices of h and L in Theorem 20.

Proof. For each output coordinate i, the label satisfies $|y_t^{(i)}| \leq B_y$. The squared norms of the AR and finite-input feature blocks are at most kB_y^2 and $(L + k + 1)K^2$, respectively. For the spectral block, orthonormality gives

$$\sum_{s=1}^{h} \|\tilde{u}_t \phi_s\|_2^2 \leq \|\tilde{u}_t\|_F^2 \leq TK^2.$$

Thus $\|a_{t,i}\|_2 \leq R_a$.

Let $x_i^\star$ concatenate the AR coefficients and the i-th rows of the input and spectral coefficient matrices representing $f_t^{\star(i)}$. The bounds in Theorem 20 imply

$$|\alpha_j| \leq L_{\mathrm{ustb}}, \qquad \|(Q_\ell)^{(i)}\|_2 \leq L_Q, \qquad \|(W_s)^{(i)}\|_2 \leq \|W_s\|_F \leq 2\sqrt{\pi T}\,\widetilde{L}_{\mathrm{ustb}} \|C_{\mathrm{st}}\|_F \|B_{\mathrm{st}}\|_F.$$

The last inequality uses the finite-horizon alternative in the definition of R_s in that theorem. Summing the squared norms of these blocks gives $\|x_i^\star\|_2^2 \leq R_x^2 - 1 \leq R_x^2$.

Apply Theorem 23 with $S = R_x$, $R = R_a$, $Y = B_y$, and dimension D to each coordinate. Each copy has regret at most $1 + B_y^2 D \log(1 + T R_a^2 R_x^2 / D)$; summing over the p coordinates proves the explicit bound.

Finally, $D = O(k + L + h)$ and the definitions give

$$\log(1 + R_a) + \log R_x = O(\log(T + 2) + \log(k + L + h + 2) + k \log(k + 2)) .$$

Substituting into the explicit regret bound proves the stated rate. □

I Unknown horizon and instability complexity

The extension is a restart wrapper around Algorithm 1. Run it with guesses q for the instability complexity and H for the horizon. Double q if the current run exceeds its loss budget, and double H when time reaches H. The argument is that a sufficient guess cannot fail, and the costs of smaller guesses form a geometric sum.

I.1 The algorithm and its loss budget

Consider the noiseless LDS of Theorem 20 with unknown $k \geq 1$. Assume known input and output bounds $K, B_y \geq 1$. Let $M_{\mathrm{sys}} \geq 1$ be a known bound on $C_{\mathrm{fast},0}$, $\|C_{\mathrm{st}}\|_F \|B_{\mathrm{st}}\|_F$, and the absolute values of all entries of $A_{\mathrm{ustb}}, B_{\mathrm{ustb}}, C_{\mathrm{ustb}}$ in the true-k decomposition of Theorem 20. These bounds and the input/output dimensions are held fixed.

For guesses $q, H \geq 2$, let $\mathcal{A}(q, H)$ be Algorithm 1 configured using Theorem 20 and Lemma 24 with $T = H$, AR order q, and approximation target $1/4$. Evaluate their formulas using the known upper bounds on the system quantities, conservatively for every $k \leq q$.[9] Cap the prescribed filter count at H. Write D, R_a, R_x for the resulting dimension and norm bounds from Lemma 24, and define the computable loss budget

$$\mathcal{B}(q, H) := 1 + p + p B_y^2 D \log\left(1 + \frac{H R_a^2 R_x^2}{D}\right) . \tag{24}$$

This is the approximation allowance 1 plus the existing VAW regret bound. Substituting the parameter choices into Theorem 20 and Lemma 24, as in the proof of Theorem 22, gives

$$\begin{aligned} D &= O\big(q \log^2(q + 2) \log^2(H + 2)\big) , \\ \mathcal{B}(q, H) &= O\big(q^2 \log^3(q + 2) \log^3(H + 2)\big) . \end{aligned} \tag{25}$$

A restart resets the learner and its loss counter. It retains the original input/output history, so all features continue to use global time. Project each reported prediction onto the radius-B_y output ball; this cannot increase squared loss and bounds each round's loss by $4B_y^2$. The VAW instances continue to receive the true labels.

Algorithm 3 Restarting the unified predictor

```
Set q ← 2, H ← 2, and global time t ← 0.
while predictions are required do
    Start a fresh copy of A(q, H); set L ← 0.
    repeat
        t ← t + 1; run one round of the current copy, clipping its reported prediction as above.
        L ← L + ‖ŷ_t − y_t‖_2^2.
    until L > B(q, H) or t = H
    if L > B(q, H) then
        q ← 2q.
    end if
    if t = H then
        H ← 2H.
    end if
end while
```

[9] For an explicit choice, substitute $C_q = [16(m + 1)(1 + qM_{\mathrm{sys}})]^{3q+4}$ for each of $L_{\mathrm{ustb}}, \widetilde{L}_{\mathrm{ustb}}, L_Q$, and M_{sys} for $C_{\mathrm{fast},0}$ and $\|C_{\mathrm{st}}\|_F \|B_{\mathrm{st}}\|_F$. These substitutions dominate the true quantities when $q \geq k$: Theorem 20 gives $L_{\mathrm{ustb}} \leq (m + 1)(1 + qM_{\mathrm{sys}})^{2q+1}$, $\widetilde{L}_{\mathrm{ustb}} \leq (1 + qL_{\mathrm{ustb}})12^q$, and $L_Q \leq (m + 2M_{\mathrm{sys}})\widetilde{L}_{\mathrm{ustb}}$. In particular, $\log C_q = O(q \log(q + 2))$.

Thus a horizon restart preserves the current complexity guess. If both conditions occur together, both guesses are doubled before the next run.

I.2 Why the restarts are inexpensive

Theorem 25 (Prediction without knowing the horizon or complexity)**.** *Under the assumptions above, Algorithm 3 needs neither T nor k. It makes at most $O(\log(k+2))$ loss-triggered restarts and $O(\log(T+2))$ horizon-triggered restarts by time T, and satisfies*

$$\sum_{t=1}^{T} \|\widehat{y}_t - y_t\|_2^2 = O\big(k^2 \log^3(k+2) \log^4(T+2)\big) .$$

Its active predictor uses $O(k \log^2(k+2) \log^2(T+2)) = \widetilde{O}(k)$ learnable coefficients.

Proof. **A sufficient guess never fails.** If $q \geq k$, the class used by $\mathcal{A}(q, H)$ contains the true-k predictor from Theorem 20: pad its extra AR, input, and spectral coefficients with zeros. The conservative parameter choices dominate its required feature counts and coefficient bounds. When all H filters are used, the spectral expansion is exact. Thus there is a comparator with total approximation error at most 1 on $[1, H]$ and coefficient norms at most R_x. The same comparator has error at most 1 on any run prefix $I \subseteq [1, H]$. Since the features retain their original history, Lemma 24 applies to a fresh learner on I and gives

$$\sum_{t \in I} \|\widehat{y}_t - y_t\|_2^2 \leq \mathcal{B}(q, H).$$

Consequently, q can double only while $q < k$. Starting at $q = 2$, this gives $q \leq 2k$ and at most $O(\log(k+2))$ loss-triggered restarts. An insufficient guess may continue indefinitely if its loss stays within budget; the algorithm need not identify k.

Each failed run has a bounded cost. Just before a loss test fails, the cumulative loss is at most $\mathcal{B}(q, H)$. The triggering round adds at most $4B_y^2$, so that run costs at most $\mathcal{B}(q, H) + 4B_y^2$. Every run that ends without crossing the budget, including the final unfinished run, costs at most $\mathcal{B}(q, H)$, even when $q < k$.

The restart costs sum geometrically. For a fixed H, the complexity guesses are distinct powers of two, bounded by $2k$. Therefore

$$\sum_{\substack{q=2,4,8,\ldots \\ q \leq 2k}} q^2 = O(k^2),$$

and (25) bounds the sum of the run budgets at that horizon by $O(k^2 \log^3(k+2) \log^3(H+2))$. The horizon guesses are also powers of two, with $H \leq 2T$ on all observed rounds. There are $O(\log(T+2))$ such horizons, so summing adds one factor of $\log(T+2)$. The triggering rounds contribute only $O(B_y^2 \log(k+2))$ in total, proving the claimed loss bound. Finally, substitute $q \leq 2k$ and $H \leq 2T$ into the dimension bound in (25) to obtain the coefficient count. □

The coefficient count refers to the active predictor; the stored observation history, spectral filters, and VAW covariance matrices are additional storage, as in the implementation described above.

J Proof of Proposition 11

Proof of Proposition 11. Fix $k \geq 1$, $R < \infty$, and filters

$$\Psi^y = \{\psi_1^y, \ldots, \psi_{q_y}^y\} \subset \ell_1(\mathbb{N}), \qquad \Psi^u = \{\psi_1^u, \ldots, \psi_{q_u}^u\} \subset \ell_1(\mathbb{N}_0),$$

with $q_y + q_u < k$. In particular, $q_u < k$.

For $s = 0, \ldots, k-1$ and $a = 1, \ldots, q_u$, define

$$U_{s,a} := \psi_{a,s}^u.$$

Thus $U \in \mathbb{R}^{k \times q_u}$ is the matrix whose columns are the first k coefficients of the input filters.

We first show that one of the k coordinate shifts cannot be generated by the input filters. For $j \in \{0, \dots, k-1\}$, let $P_j : \mathbb{R}^k \to \mathbb{R}^{j+1}$ denote the projection onto the first $j+1$ coordinates. We claim that there exists $j \in \{0, \dots, k-1\}$ such that

$$P_j e_j \notin \operatorname{span}\{P_j U_{\cdot,1}, \dots, P_j U_{\cdot,q_u}\}.$$

Suppose not. Then for every $j = 0, \dots, k-1$, there exists $\beta^{(j)} \in \mathbb{R}^{q_u}$ such that

$$P_j U \beta^{(j)} = P_j e_j.$$

Equivalently, the vector $w^{(j)} := U\beta^{(j)} \in \mathbb{R}^k$ satisfies

$$w_i^{(j)} = 0 \quad \text{for all } i < j, \qquad w_j^{(j)} = 1.$$

Therefore the $k \times k$ matrix with columns $w^{(0)}, w^{(1)}, \dots, w^{(k-1)}$ is lower triangular with all diagonal entries equal to 1. Hence these k columns are linearly independent. But all of them lie in the column span of U, whose dimension is at most $q_u < k$, a contradiction.

Fix such a j, and define

$$\delta := \inf_{\beta \in \mathbb{R}^{q_u}} \|P_j e_j - P_j U \beta\|_2.$$

By the choice of j, $\delta > 0$.

We now construct the LDS. Let $e_0, \dots, e_{k-1}$ be the standard basis of $\mathbb{R}^k$. Define $A \in \mathbb{R}^{k \times k}$ by

$$A e_i = e_{i+1} \quad \text{for } i = 0, \dots, k-2, \qquad A e_{k-1} = 0.$$

Set $B = e_0$ and $C = e_j^\top$. Then

$$\|A\|_\infty, \|B\|_\infty, \|C\|_\infty \le 1.$$

Choose a block length $L > 2k$, to be fixed below. Let

$$t_b := bL + 1, \qquad b = 0, 1, 2, \dots.$$

Set $u_{t_b} = 1$ for all $b \ge 0$, and set $u_t = 0$ for all other t. Thus $|u_t| \le 1$ for all t. Since the system forgets an input after k steps and $L > 2k$, different blocks do not interact through the LDS state. Therefore, in every block,

$$y_{t_b+s} = \begin{cases} 1, & s = j, \\ 0, & s \ne j, \end{cases} \qquad s = 0, \dots, L-1.$$

In particular, $|y_t| \le 1$ for all t, so the constructed input sequence is 1-controlling in the sense of Assumption 2.

We now choose L large enough so that old blocks are invisible to the filters. Since all filters are in ℓ_1, and there are only finitely many of them, we may choose $L > 2k$ so large that, for every input filter a and every output filter r,

$$\sum_{\tau \ge L-k} |\psi^u_{a,\tau}| \le \frac{\delta}{4 \max\{R, 1\}\sqrt{k}}, \qquad \sum_{\tau \ge L-k} |\psi^y_{r,\tau}| \le \frac{\delta}{4 \max\{R, 1\}\sqrt{k}}.$$

Fix any predictor $f \in \mathrm{GUF}^R_{\Psi^y, \Psi^u}$, and write

$$f_t = \sum_{r=1}^{q_y} \alpha_r (\psi^y_r * y)_t + \sum_{a=1}^{q_u} \beta_a (\psi^u_a * u)_t, \qquad \sum_{r=1}^{q_y} |\alpha_r| + \sum_{a=1}^{q_u} |\beta_a| \le R.$$

Consider the first $j+1$ times of block b, $t_b, t_b+1, \dots, t_b+j$. On these times, the true output vector is

$$(y_{t_b}, y_{t_b+1}, \dots, y_{t_b+j})^\top = P_j e_j.$$

Consider first an input filter ψ^u_a at a time $t_b + s$, where $0 \le s \le j$. The current block contains a single input pulse at time t_b, so its contribution is exactly the s-th filter coefficient, $\psi^u_{a,s}$. Every other nonzero input comes from an earlier block. Since blocks are separated by L time steps and the current prefix has length at most k, all such earlier inputs appear at lag at least $L - k$. Hence

$$(\psi^u_a * u)_{t_b+s} = \psi^u_{a,s} + \varepsilon^u_{a,b,s}, \qquad |\varepsilon^u_{a,b,s}| \le \sum_{\tau \ge L-k} |\psi^u_{a,\tau}|.$$

Now consider an output filter ψ_r^y on the same prefix $t_b, \dots, t_b + j$. The current block has no nonzero output before time $t_b + j$, and at time $t_b + j$ the filter is strictly causal, so it cannot yet use the value y_{t_b+j}. Thus the output-filter feature on this prefix is entirely due to outputs from previous blocks. Those outputs also have lag at least $L - k$, and therefore

$$|(\psi_r^y * y)_{t_b+s}| \le \sum_{\tau \ge L-k} |\psi_{r,\tau}^y|, \qquad s = 0, \dots, j.$$

Collect the predictions on this prefix of the block into the vector

$$F_b := (f_{t_b}, f_{t_b+1}, \dots, f_{t_b+j})^\top \in \mathbb{R}^{j+1}.$$

The discussion above shows that, on this prefix, the predictor is equal to the linear combination of the first $j + 1$ input-filter coefficients, up to the negligible contribution of old blocks:

$$F_b = P_j U\beta + \xi_b, \qquad \|\xi_b\|_2 \le \frac{\delta}{4}.$$

The target vector on the same prefix is $P_j e_j$. By the definition of δ, every linear combination of the input-filter prefixes is at least δ away from this target:

$$\|P_j e_j - P_j U\beta\|_2 \ge \delta.$$

The perturbation ξ_b can reduce this distance by at most $\delta/4$, so

$$\|P_j e_j - F_b\|_2 \ge \|P_j e_j - P_j U\beta\|_2 - \|\xi_b\|_2 \ge \delta - \frac{\delta}{4} = \frac{3\delta}{4}.$$

Thus every predictor in $\mathrm{GUF}^R_{\Psi^y,\Psi^u}$ incurs a constant amount of squared error on the prefix of every block:

$$\sum_{s=0}^{j} |y_{t_b+s} - f_{t_b+s}|^2 = \|P_j e_j - F_b\|_2^2 \ge \frac{9\delta^2}{16}.$$

If the horizon contains N complete blocks, summing the preceding inequality over these blocks gives

$$\sum_{t=1}^{T} |y_t - f_t|^2 \ge N \cdot \frac{9\delta^2}{16}.$$

Since $N/T \to 1/L$ as $T \to \infty$, we obtain

$$\liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} |y_t - f_t|^2 \ge \frac{9\delta^2}{16L} > 0.$$

The constant on the right depends only on the fixed filters, R, and k, not on the predictor f. Taking the infimum over the bounded-coefficient filter class therefore preserves the positive lower bound:

$$\liminf_{T\to\infty} \inf_{f \in \mathrm{GUF}^R_{\Psi^y,\Psi^u}} \frac{1}{T} \sum_{t=1}^{T} \ell_2^2(f_t, y_t) > 0.$$

This proves the theorem. □

K Proof of Proposition 9

K.1 Proof for general input dimension

Proof. We first identify an input direction along which canceling the first input requires large future input energy. We then convert this energy requirement into a lower bound on the largest input norm relative to the first, and hence on the input dynamic range. This gives a deterministic bound depending on the first input direction; a probability estimate then gives the stated result for a uniformly random direction. Throughout, A, B, and η are fixed as in Proposition 9. Write

$$\kappa := 1 + \eta, \qquad r := \frac{1 + \kappa^{-1}}{2}, \qquad a := r\kappa = 1 + \frac{\eta}{2}, \qquad D := rA.$$

Thus $0 < r < 1$ and every eigenvalue of D has modulus at least $a > 1$. In particular, $\|D^{-j}\|_2 \le a^{-j}$, so powers of D^{-1} decay geometrically. The choice of r will also give a summable discount on the future inputs.

A Gramian and its determinant. Our goal is to construct an $m \times m$ positive semidefinite matrix M with a large eigenvalue. Begin with the $k \times k$ Gramian

$$G := \sum_{j=1}^{\infty} D^{-j} B B^\top (D^{-j})^\top.$$

Each matrix $D^{-j}B$ maps an input vector into state coordinates; the Gramian collects these maps through their products with their transposes. The series converges in operator norm because

$$\|D^{-j} B B^\top (D^{-j})^\top\|_2 \le a^{-2j} \|B\|_2^2,$$

and $a > 1$. For every $z \in \mathbb{R}^k$,

$$z^\top G z = \sum_{j=1}^{\infty} \|B^\top (D^{-j})^\top z\|_2^2 \ge 0.$$

To show that G is invertible, we prove that equality forces $z = 0$. If $z^\top G z = 0$, every nonnegative summand is zero, so $z^\top D^{-j} B = 0$ for every $j \ge 1$. Fix an input coordinate $\ell \in [m]$. The first k of these equations read

$$\sum_{i=1}^{k} z_i B_{i\ell} (r\lambda_i)^{-j} = 0, \qquad j = 1, \dots, k.$$

These are k linear equations in the k unknowns $z_i B_{i\ell}$. Writing $\xi_i = (r\lambda_i)^{-1}$, their coefficient matrix is $[\xi_i^j]_{j,i=1}^k$, a Vandermonde matrix multiplied by $\operatorname{diag}(\xi_1, \dots, \xi_k)$. The numbers ξ_i are distinct and nonzero, so this matrix is invertible. Thus $z_i B_{i\ell} = 0$ for every i, ℓ. Every row of B contains a nonzero entry, which forces every $z_i = 0$. This proves $G \succ 0$.

We can therefore define

$$M := B^\top G^{-1} B \in \mathbb{R}^{m \times m}.$$

This matrix is positive semidefinite: for every $v \in \mathbb{R}^m$, $v^\top M v = (Bv)^\top G^{-1} (Bv) \ge 0$.

The definition of G gives a useful identity. Multiplying by D and $D^\top$ shifts each power in the series by one:

$$\begin{aligned} DGD^\top &= \sum_{j=1}^{\infty} D^{-(j-1)} B B^\top (D^{-(j-1)})^\top \\ &= BB^\top + \sum_{j=1}^{\infty} D^{-j} B B^\top (D^{-j})^\top = BB^\top + G. \end{aligned}$$

Taking determinants of the left-hand side gives

$$\det(DGD^\top) = |\det D|^2 \det G.$$

For the right-hand side, the matrix determinant lemma gives

$$\det(G + BB^\top) = \det G \, \det(I_m + B^\top G^{-1} B) = \det G \, \det(I_m + M).$$

Equating these expressions and canceling $\det G > 0$, we obtain

$$\begin{aligned} \det(I_m + M) &= |\det D|^2 \\ &= r^{2k} \prod_{i=1}^{k} |\lambda_i|^2 \ge a^{2k}, \end{aligned} \tag{26}$$

where the inequality uses $|r\lambda_i| \ge a$ for each of the k state coordinates. On the other hand, M has only m eigenvalues. Writing them as $\mu_1, \dots, \mu_m \ge 0$ and setting $\mu = \max_\ell \mu_\ell$, we have

$$a^{2k} \le \det(I_m + M) = \prod_{\ell=1}^{m} (1 + \mu_\ell) \le (1 + \mu)^m.$$

Taking the m-th root yields

$$\mu \ge a^{2k/m} - 1 = \left(1 + \frac{\eta}{2}\right)^{2k/m} - 1.$$

This is where the ratio k/m enters the proof. Fix a unit eigenvector $w \in \mathbb{R}^m$ corresponding to μ. This direction depends only on A, B, and η, not on the input sequence.

Stabilization requires cancellation. Consider any nonzero stabilizing input sequence. If its dynamic range is infinite, the desired inequality is immediate. Otherwise set $U := \sup_t \|u_t\|_2 < \infty$. Since $x_0 = 0$, the state stays zero until the first nonzero input. We may therefore shift time so that this input is u_1. Unrolling the state recursion displays the contribution of each input:

$$x_t = \sum_{s=1}^{t} A^{t-s} B u_s = A^{t-1} B u_1 + A^{t-2} B u_2 + \cdots + B u_t.$$

Multiplying by $A^{-(t-1)}$ removes the growth affecting the first input:

$$A^{-(t-1)} x_t = B u_1 + \sum_{j=1}^{t-1} A^{-j} B u_{j+1}.$$

The left side tends to zero, since stabilization gives $x_t \to 0$ and

$$\|A^{-(t-1)} x_t\|_2 \le \kappa^{-(t-1)} \|x_t\|_2 \longrightarrow 0.$$

The series on the right converges absolutely because

$$\|A^{-j} B u_{j+1}\|_2 \le \kappa^{-j} \|B\|_2 U, \qquad \kappa > 1.$$

Taking the limit therefore gives

$$0 = B u_1 + \sum_{j=1}^{\infty} A^{-j} B u_{j+1}.$$

Thus the future inputs must cancel the first input after accounting for their application times through the factors A^{-j}. Finally, $D = rA$ implies $A^{-j} = r^j D^{-j}$, so this necessary cancellation identity can be written as

$$B u_1 = -\sum_{j=1}^{\infty} A^{-j} B u_{j+1} = -\sum_{j=1}^{\infty} D^{-j} B \, (r^j u_{j+1}). \tag{27}$$

The matrices $D^{-j}B$ are precisely those appearing in G, while $r^j u_{j+1}$ are the discounted future inputs.

From discounted energy to input dynamic range. We first lower-bound the future discounted input energy by a quadratic form in the first input. Set

$$Q := u_1^\top M u_1 \ge 0, \qquad z := G^{-1} B u_1.$$

This choice of z ensures that

$$z^\top B u_1 = (B u_1)^\top G^{-1} (B u_1) = Q.$$

Taking the inner product of (27) with z, then applying Cauchy–Schwarz to the sum of inner products, gives

$$\begin{aligned} Q &= \left| \sum_{j \ge 1} \left\langle B^\top (D^{-j})^\top z, \, r^j u_{j+1} \right\rangle \right| \\ &\le \left(\sum_{j \ge 1} \|B^\top (D^{-j})^\top z\|_2^2 \right)^{1/2} \left(\sum_{j \ge 1} r^{2j} \|u_{j+1}\|_2^2 \right)^{1/2}. \end{aligned}$$

By the definition of G, the first factor squared is

$$\sum_{j \ge 1} \|B^\top (D^{-j})^\top z\|_2^2 = z^\top G z = (B u_1)^\top G^{-1} (B u_1) = Q.$$

Hence

$$Q \le \sqrt{Q} \left(\sum_{j \ge 1} r^{2j} \|u_{j+1}\|_2^2 \right)^{1/2}.$$

If $Q > 0$, dividing by $\sqrt{Q}$ and squaring shows that the discounted energy is at least Q; this also holds when $Q = 0$. To bound that energy from above, use $\|u_{j+1}\|_2 \le U$ and sum the geometric weights:

$$\begin{aligned} u_1^\top M u_1 &\le \sum_{j\ge 1} r^{2j} \|u_{j+1}\|_2^2 \\ &\le U^2 \sum_{j\ge 1} r^{2j} = \frac{r^2}{1-r^2} U^2. \end{aligned} \tag{28}$$

Thus a large value of $u_1^\top M u_1$ forces a large peak input norm:

$$U \ge \frac{\sqrt{1-r^2}}{r} \sqrt{u_1^\top M u_1}.$$

Write $s = \|u_1\|_2 > 0$ and $v = u_1/s$, so $\|v\|_2 = 1$ and $u_1^\top M u_1 = s^2 v^\top M v$. To use the large eigenvalue μ, decompose $v = (w^\top v)w + v_\perp$ with $w^\top v_\perp = 0$. Since $Mw = \mu w$ and $M \succeq 0$,

$$v^\top M v = \mu |w^\top v|^2 + v_\perp^\top M v_\perp \ge \mu |w^\top v|^2.$$

Dividing the peak bound by s and substituting the lower bound on μ therefore gives

$$\begin{aligned} \frac{U}{s} &\ge \frac{\sqrt{1-r^2}}{r} \sqrt{v^\top M v} \\ &\ge |w^\top v| \frac{\sqrt{1-r^2}}{r} \sqrt{a^{2k/m} - 1}. \end{aligned} \tag{29}$$

The factor $|w^\top v|$ measures how strongly the first input points along the direction with the large energy requirement.

To isolate the exponential dependence, set $q = k/m \ge 1$ and write

$$\frac{\sqrt{1-r^2}}{r} \sqrt{a^{2q} - 1} = a^q \sqrt{\frac{1-r^2}{r^2}(1 - a^{-2q})}.$$

Since $a > 1$ and $q \ge 1$, we have $1 - a^{-2q} \ge 1 - a^{-1} = (1-r)/r$. Consequently

$$\frac{1-r^2}{r^2}(1 - a^{-2q}) \ge \frac{1-r^2}{r^2}(1 - a^{-1}) = (1-r)^2 \frac{1+r}{r^3} \ge (1-r)^2.$$

Taking square roots and multiplying by a^q gives

$$\frac{U}{s} \ge |w^\top v| (1-r) a^{k/m}.$$

Finally, the first input is one of the nonzero inputs over which the infimum in $\mathrm{IDR}(u)$ is taken. Therefore

$$\inf\{\|u_t\|_2 : u_t \ne 0\} \le \|u_1\|_2 = s, \qquad \mathrm{IDR}(u) = \frac{U}{\inf\{\|u_t\|_2 : u_t \ne 0\}} \ge \frac{U}{s}.$$

This step does not assume that the first input has the smallest norm: a smaller later input only increases the dynamic range. Substituting $1 - r = \eta/(2(1+\eta))$ and $a = 1 + \eta/2$ proves the deterministic directional inequality

$$\mathrm{IDR}(u) \ge |w^\top v| \, \frac{\eta}{2(1+\eta)} \left(1 + \frac{\eta}{2}\right)^{k/m}. \tag{30}$$

This holds for every stabilizing sequence, using the same w. In particular, when $m = 1$, both w and v are unit scalars, so $|w^\top v| = 1$. This proves the deterministic scalar-input conclusion.

A uniformly random first input direction. It remains to show that for fixed unit w and uniform $v \in \mathbb{S}^{m-1}$,

$$\Pr\left\{ |w^\top v| < \frac{\delta}{\sqrt{m}} \right\} \le \delta, \qquad 0 < \delta < 1. \tag{31}$$

The case $m = 1$ is immediate. For $m \geq 2$, rotational invariance lets us assume $w = e_1$ and write $v = g/\|g\|_2$, where $g \sim \mathcal{N}(0, I_m)$. Set $b = \delta/\sqrt{m} < 1$. Then

$$|v_1| < b \quad \Longleftrightarrow \quad |g_1| < \frac{b}{\sqrt{1-b^2}} \|g_{2:m}\|_2.$$

The variable g_1 is independent of $g_{2:m}$, and its density is at most $1/\sqrt{2\pi}$. Thus $\Pr(|g_1| < c) \leq \sqrt{2/\pi}\, c$ for $c \geq 0$. Using also $\mathbb{E}\|g_{2:m}\|_2 \leq \sqrt{m-1}$ gives

$$\Pr(|v_1| < b) \leq \sqrt{\frac{2}{\pi}} \frac{b\sqrt{m-1}}{\sqrt{1-b^2}} = \sqrt{\frac{2}{\pi}}\, \delta \sqrt{\frac{m-1}{m-\delta^2}} \leq \delta.$$

Combining (31) with (30) proves (5). The probability event depends only on the first input direction and the fixed system; the first input magnitude and all subsequent controls are unrestricted. Hence the conclusion holds simultaneously over every stabilizing continuation beginning in that direction, as claimed. □

Why the first-direction qualification is necessary. Even with fewer inputs than exploding modes, a specially chosen first direction can avoid a large dynamic range. For example, take $m = 2$ and $k > 2$, choose distinct positive exploding eigenvalues, and let $B = [b, Ab]$ with every entry of b nonzero. This B has rank two and no zero rows. The inputs $u_1 = (1, 0)^\top$, $u_2 = (0, -1)^\top$, and $u_t = 0$ for $t \geq 3$ give $x_1 = b$, $x_2 = Ab - Ab = 0$, and $\mathrm{IDR}(u) = 1$. Thus the proposition does not assert a lower bound uniform over all first input directions. Moreover, when $m \geq k$ and B has full row rank, any state can be canceled in one step. Thus the obstruction depends on the number of input channels as well as on the first input direction.

K.2 Alternative proof for a single input

Proof. We prove the deterministic conclusion of Proposition 9 for $m = 1$, writing $B = (b_1, \ldots, b_k)^\top$. Write $x_t^{(i)}$ for the i-th coordinate of x_t. Since

$$x_t^{(i)} = \lambda_i x_{t-1}^{(i)} + b_i u_t,$$

we have

$$b_i u_t = x_t^{(i)} - \lambda_i x_{t-1}^{(i)}.$$

If u stabilizes the system, then $x_t^{(i)} \to 0$ and $x_{t-1}^{(i)} \to 0$. Hence

$$b_i u_t \to 0.$$

Since $b_i \neq 0$, it follows that $u_t \to 0$. Therefore, if u has infinitely many nonzero values, then

$$\inf\{|u_t| : u_t \neq 0\} = 0,$$

and so $\mathrm{IDR}(u) = \infty$. The desired lower bound is then immediate.

It remains to consider the case where $\mathrm{IDR}(u) < \infty$. Since $u_t \to 0$ and the nonzero values of u_t are bounded away from zero, u can have only finitely many nonzero entries. Let

$$t_0 := \min\{t : u_t \neq 0\}, \qquad T := \max\{t : u_t \neq 0\}.$$

Define the polynomial

$$q(z) := \sum_{t=t_0}^{T} u_t z^{t-t_0}.$$

Its constant coefficient is nonzero:

$$q(0) = u_{t_0} \neq 0.$$

We claim that

$$q(1/\lambda_i) = 0 \qquad \text{for every } i \in [k].$$

Indeed, for every coordinate i,

$$x_T^{(i)} = \sum_{t=t_0}^{T} \lambda_i^{T-t} b_i u_t.$$

After time T, the input is zero, so

$$x_s^{(i)} = \lambda_i^{s-T} x_T^{(i)} \qquad \text{for all } s \geq T.$$

Since $|\lambda_i| > 1$ and $x_s^{(i)} \to 0$, we must have

$$x_T^{(i)} = 0.$$

Therefore

$$\sum_{t=t_0}^{T} \lambda_i^{T-t} b_i u_t = 0.$$

Dividing by $b_i \neq 0$, we obtain

$$\sum_{t=t_0}^{T} \lambda_i^{T-t} u_t = 0.$$

Dividing by $\lambda_i^{T-t_0} \neq 0$, we obtain

$$\sum_{t=t_0}^{T} u_t \lambda_i^{-(t-t_0)} = 0,$$

which is exactly

$$q(1/\lambda_i) = 0.$$

Now set

$$z_i := \frac{1}{\lambda_i}.$$

Then $z_1, \ldots, z_k$ are distinct roots of q, and

$$|z_i| \leq \frac{1}{1+\eta}.$$

Let

$$U := \sup_{t \geq 1} |u_t|, \qquad m := \inf\{|u_t| : u_t \neq 0\}.$$

Thus

$$\mathrm{IDR}(u) = \frac{U}{m}.$$

Choose

$$R := \frac{1 + \frac{1}{1+\eta}}{2} = \frac{2+\eta}{2(1+\eta)}.$$

Then

$$\frac{1}{1+\eta} < R < 1.$$

Hence all roots $z_1, \ldots, z_k$ lie inside the disk $\{|z| < R\}$.

By Jensen's formula applied to q,

$$\prod_{i=1}^{k} \frac{R}{|z_i|} \leq \frac{\max_{|z|=R} |q(z)|}{|q(0)|}.$$

We now bound the numerator and denominator. First,

$$|q(0)| = |u_{t_0}| \geq m.$$

Second, for every $|z| = R$,

$$|q(z)| \leq \sum_{t=t_0}^{T} |u_t| R^{t-t_0} \leq U \sum_{j=0}^{\infty} R^j = \frac{U}{1-R}.$$

Therefore

$$\prod_{i=1}^{k} \frac{R}{|z_i|} \le \frac{U}{m(1-R)} = \frac{\mathrm{IDR}(u)}{1-R}.$$

Rearranging gives

$$\mathrm{IDR}(u) \ge (1-R) \prod_{i=1}^{k} \frac{R}{|z_i|}.$$

Since $|z_i| = 1/|\lambda_i|$, this becomes

$$\mathrm{IDR}(u) \ge (1-R) R^k \prod_{i=1}^{k} |\lambda_i|.$$

Using $|\lambda_i| \ge 1+\eta$, we get

$$\mathrm{IDR}(u) \ge (1-R)\big(R(1+\eta)\big)^k.$$

By our choice of R,

$$1 - R = \frac{\eta}{2(1+\eta)}$$

and

$$R(1+\eta) = \frac{2+\eta}{2} = 1 + \frac{\eta}{2}.$$

Therefore

$$\mathrm{IDR}(u) \ge \frac{\eta}{2(1+\eta)} \left(1 + \frac{\eta}{2}\right)^k.$$

This proves the claim. □

L Experimental details

This appendix specifies the realization (A, B, C) used in Section 6 and how the predictors are trained on it.

L.1 System construction

The system is the direct sum of four scalar-input, scalar-output subsystems, so A is block diagonal while B and C are the concatenations of the per-subsystem input vectors and readout rows,

$$\begin{aligned} A &= \mathrm{diag}(A_{\mathrm{exp}}, A_{\mathrm{slow}}, A_{\mathrm{fast}}, A_{\mathrm{st}}), \\ B &= \big(B_{\mathrm{exp}}^\top, B_{\mathrm{slow}}^\top, B_{\mathrm{fast}}^\top, B_{\mathrm{st}}^\top\big)^\top, \\ C &= (C_{\mathrm{exp}}, C_{\mathrm{slow}}, C_{\mathrm{fast}}, C_{\mathrm{st}}), \end{aligned}$$

where $(A_{\mathrm{exp}}, B_{\mathrm{exp}}, C_{\mathrm{exp}})$ is the exploding scalar mode, $(A_{\mathrm{slow}}, B_{\mathrm{slow}}, C_{\mathrm{slow}})$ the slow complex pair, $(A_{\mathrm{fast}}, B_{\mathrm{fast}}, C_{\mathrm{fast}})$ the fast-decaying bank, and $(A_{\mathrm{st}}, B_{\mathrm{st}}, C_{\mathrm{st}})$ the stable block, described in turn below. Their state matrices have sizes $A_{\mathrm{exp}} \in \mathbb{R}^{1\times 1}$, $A_{\mathrm{slow}} \in \mathbb{R}^{2\times 2}$, $A_{\mathrm{fast}} \in \mathbb{R}^{200\times 200}$ (a stack of 100 blocks of size 2×2), and $A_{\mathrm{st}} \in \mathbb{R}^{300\times 300}$, for a total state dimension $d = 503$.

Exploding scalar. The unstable mode is $A_{\mathrm{exp}} = [1.3]$, with input $B_{\mathrm{exp}} = 1$ and readout $C_{\mathrm{exp}} = 1$. The input is $u_t = z_t - 1.3 z_{t-1}$, where the z_t are independent uniform signs and $z_0 = 0$. Starting from zero, the scalar state therefore satisfies $x_t^{\mathrm{exp}} = z_t$ and remains bounded.

Slow complex pair. The slowly decaying pair $0.98\, e^{\pm i 1.57}$ is realized as the real 2×2 block

$$A_{\mathrm{slow}} = 0.98 \begin{pmatrix} \cos(1.57) & -\sin(1.57) \\ \sin(1.57) & \cos(1.57) \end{pmatrix},$$

with input $B_{\mathrm{slow}} = (1, 0)^\top$ and readout $C_{\mathrm{slow}} = (c, 0)$, where $c > 0$ is set so that the peak magnitude of its impulse response equals one.

Fast-decaying bank. The bank A_{fast} consists of 100 real 2×2 rotation–scaling blocks

$$A_{\mathrm{fast},j} = r_j \begin{pmatrix} \cos(\theta_j) & -\sin(\theta_j) \\ \sin(\theta_j) & \cos(\theta_j) \end{pmatrix},$$

with radius r_j uniform in $[0, 0.25]$ and angle θ_j uniform in $[0, 2\pi]$. Each block has input $B_{\mathrm{fast},j} = (1, 0)^\top$ and a readout $C_{\mathrm{fast},j}$ that is a Gaussian row with i.i.d. $\mathcal{N}(0, 1)$ entries scaled by 0.1.

Stable block. The stable block is $A_{\mathrm{st}} = \mathrm{diag}(\lambda_1, \ldots, \lambda_{300})$ with input $B_{\mathrm{st}} = \mathbf{1}$, the all-ones vector. We place the eigenvalues on a logarithmic grid clustered near the stability boundary: the eigenvalues are $\pm(1 - \delta_i)$, where the gaps $\delta_1, \ldots, \delta_{150}$ are geometrically spaced between 10^{-5} and 1, giving 150 positive and 150 negative nodes.[10] We choose the stable readout C_{st} so that, after the autoregressive component cancels the hard-to-learn modes (the $k = 3$ eigenvalues counted by the instability complexity), the residual of the remaining stable response aligns with the spectral filters, and is thus captured by the unified predictor but not by autoregression alone. Let $\alpha_1, \alpha_2, \alpha_3$ be the autoregressive coefficients whose roots are the hard modes 1.3 and $0.98\, e^{\pm i 1.57}$, and define the residual of the stable impulse response $h_t = \sum_i C_{\mathrm{st},i} \lambda_i^t$ under this order-three recurrence by

$$r_t = h_t - \alpha_1 h_{t-1} - \alpha_2 h_{t-2} - \alpha_3 h_{t-3}, \qquad t \geq 3.$$

For each seed we form a target $g = \kappa \sum_{j=1}^{6} s_j w_j \phi_j$ from the first six spectral filters of Definition 13, with random signs $s_j \in \{-1, +1\}$, weights w_j equally spaced between 1 and 50, and κ chosen so that $\|g\|_\infty = 25$. We then fit C_{st} by least squares so that r_t matches g_t for $t = 3, \ldots, T$.

L.2 Training

Each predictor is trained by the Vovk–Azoury–Warmuth forecaster of Appendix H with regularization $\lambda = 0.1$.

[10] Spacing the distance-to-the-circle $1 - |\lambda|$ geometrically places many real stable nodes close to ± 1, exactly where long-memory stable responses are hardest for a short finite-memory predictor and must instead be captured by spectral filters.

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 abstract and introduction explicitly state our unified algorithm's regret and parameter-count properties, matching the theoretical guarantees proved in Section 4.

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss limitations in a dedicated Limitations section.

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: All formal theorems are stated in the main text with their full statements and proofs deferred to the appendix.

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: Section 6 and Appendix L describe the synthetic system, predictors, training procedure, and evaluation. The implementation, including the random seeds and reproduction instructions, is available at `https://github.com/YuvMilo/UnifiedLearningLDS`.

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: [Yes]

   Justification: The repository cited in Section 6 provides the code to generate the synthetic trajectories and reproduce all four predictors in Table 1, together with setup instructions and the command `python main.py`.

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: Section 6 specifies the trajectory length, evaluation window, parameter budgets, and predictor classes. Appendix L gives the system construction and Vovk–Azoury–Warmuth training procedure with regularization $\lambda = 0.1$.

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: [No]

   Justification: Table 1 reports the best and worst normalized mean-squared error over five random seeds. These ranges describe the observed variability; we do not report confidence intervals or statistical significance tests.

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: The paper does not specify the hardware, memory usage, or execution time for the experiments.

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: The research complies with the NeurIPS Code of Ethics. The experiments use synthetic linear dynamical systems and do not involve human subjects or personal data.

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 is foundational theoretical research on linear dynamical systems, with no direct societal impact.

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: The released code generates synthetic LDS trajectories and does not include data or models with a high risk for misuse.

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: [No]

    Justification: The experiments use NumPy and SciPy, which are listed in the requirements file of the released code. The paper does not explicitly state their licenses.

13. **New assets**

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

    Answer: [Yes]

    Justification: The released experimental code includes a README with environment setup and reproduction instructions, and a requirements file. The synthetic system and training procedure are documented in Appendix L.

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: Our research does not involve crowdsourcing or human subjects.

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: Our research does not involve human subjects.

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: The core methodology does not involve LLMs.