

Contents

1 Introduction	1
1.1 Our results	2
1.2 Intuition for Universal Sequence Preconditioning	3
1.3 Related work	4
2 Main Results	5
2.1 Universal Sequence Preconditioning Applied to Regression	5
2.2 Universal Sequence Preconditioning Applied to Spectral Filtering	5
3 Proof Overview	7
3.1 Using the Chebyshev Polynomial over the Complex Plane	8
4 Experimental Evaluation	8
4.1 Synthetic Data Generation	8
4.2 Algorithms and Preconditioning Variants	8
4.3 Results	9
4.4 ETTh1 Dataset	10
5 Discussion	10
A Related work	17
B Memory Capacity of Linear Dynamical Systems	18
C Online Version of Preconditioning	19
D Proof of Convolutional Preconditioned Regression Performance Theorem 2.1	19
E Proof of Theorem 2.2	22
E.1 Preliminaries and Notation	22
E.2 General Form of Main Result	23
E.3 Proving Approximation Lemma E.2	27
E.3.1 Proof of the Spectral Filtering Property Lemma E.4	28
A Proofs of Technical Lemmas	31
B Chebyshev Polynomials Evaluated in the Complex Plane	32
C Plots from Experiments	34
C.1 Experiments with Linear Dynamical System Data	34
C.2 Nonlinear Data	36
C.3 Data from a DNN	37

A Related work

Preconditioning for sequence prediction. Preconditioning in the context of time series analysis has roots in the classical work of Box and Jenkins [13]. In this foundational text they propose differencing as a method for making the time series stationary, and thus amenable to statistical learning techniques such as ARMA (auto-regressive moving average) [9]. The differencing operator can be applied numerous times, and for different lags, giving rise to the ARIMA family of forecasting models.

Identifying the order of an ARIMA model, and in particular the types of differencing needed to make a series stationary, is a hard problem. This is a special case of the problem we consider: differencing corresponds to certain coefficients of preconditioning the time series, whereas we consider arbitrary coefficients.

Background on control of linear dynamical systems. Linear dynamical systems are the most fundamental and basic model in control theory, and have been studied for more than a century. For a thorough introduction, see the texts [12, 38, 26].

A rigorous proof that the Cayley-Hamilton theorem implies that d_h learned closed-loop components are sufficient to learn any LDS is given in [7, 25].

The seminal work of Kalman on state-space representation and filtering [28] allows one to learn any LDS with hidden-dimension parameters under stochastic and generative assumptions. Closed-loop auto-regressive learning subsumes Kalman filtering in the presence of adversarial noise, see e.g. [29]. [21] provide a method to learn a marginally stable LDS in the presence of bounded adversarial noise and asymmetric matrices, however their regret bound depends on the hidden dimension of the system. More recently, [10] use tensor and moment-based methods to learn a LDS with stochastic noise without dependence on the system’s condition number. However, their algorithmic complexity still scales polynomially with the hidden dimension.

In this work we consider regret in the context of *learning* linear dynamical systems. This is related to, but different from, *control* of the systems. We survey regret minimization for control next.

Regret for classical control models. The first works addressing control in the machine learning community assume either no perturbation in the dynamics at all, or i.i.d. Gaussian perturbations. Much of this work has considered obtaining low regret in the online LQR setting [2, 18, 30, 17] where a fully-observed linear dynamic system is driven by i.i.d. Gaussian noise via $x_{t+1} = \mathbf{A}x_t + \mathbf{B}u_t + w_t$, and the learner incurs constant quadratic state and input cost $\ell(x, u) = \frac{1}{2}x^\top \mathbf{Q}x + \frac{1}{2}u^\top \mathbf{R}u$. The optimal policy for this setting is well-approximated by a *state feedback controller* $u_t = Kx_t$, where K is the solution to the Discrete Algebraic Riccati Equation (DARE), and thus regret amounts to competing with this controller. Recent algorithms [30, 17] attain \sqrt{T} regret for this setting, with polynomial runtime and polynomial regret dependence on relevant problem parameters. A parallel line of work by [16] establishes \sqrt{T} regret in a variant of online LQR where the system is known to the learner, noise is stochastic, but an adversary selects quadratic loss functions ℓ_t at each time t . Again, the regret is measured with respect to a best-in-hindsight state feedback controller.

Provable control in the Gaussian noise setting via the policy gradient method was studied in [20]. Other relevant work from the machine learning literature includes tracking adversarial targets [1].

The non-stochastic control problem. The most accepted and influential control model stemming from the machine learning community was established in [4], who obtain \sqrt{T} -regret in the more general and challenging setting where both the Lipschitz loss function and the perturbations are adversarially chosen. The key insight behind this result is combining an improper controller parametrization known as disturbance-based control with online convex optimization with memory due to [8]. Follow-up work by [6] achieves logarithmic pseudo-regret for strongly convex, adversarially selected losses and well-conditioned stochastic noise. Further extensions were made for linear control with partial observation [33], system identification with adversarial noise [15], and many more settings surveyed in [26].

Online learning and online convex optimization. We make extensive use of techniques from the field of online learning and regret minimization in games [14, 24]. Following previous work from

machine learning, we consider regret minimization in sequence prediction, where the underlying sequence follows a linear dynamical system.

Spectral filtering for learning linear dynamical systems. The spectral filtering technique was put forth in [27] for learning symmetric linear dynamical systems. In [25], the technique was extended for more general dynamical systems using closed-loop regression; however, this required hidden-dimension parameters and polynomial dependence on the approximation guarantee. Spectral filtering techniques were recently used in non-linear sequence prediction, notably in the context of large language models, albeit with no theoretical guarantees [7]. As convolutional models, these methods are attractive for sequence prediction due to faster generation and inference as compared to attention-based models [5].

While several methods exist that can learn in the presence of asymmetric transition matrices [28, 10, 21], their performance depends on hidden dimension. On the other hand, spectral filtering methods [27] achieve regret which is independent of hidden dimension, even for marginally stable systems. However, these spectral filtering methods were limited to systems with symmetric transition matrices. In contrast, real-world dynamical systems often have asymmetric transition matrices with large hidden dimension, necessitating a more general approach. In this paper, we provide such an approach by extending the theory of spectral filtering to handle asymmetric systems, as long as the complex component of their eigenvalues is bounded.

In this paper we dramatically improve the spectral filtering technique and broaden its applicability in two major aspects: First, for general asymmetric linear dynamical systems we remove the dependence on the hidden dimension. Second, we improve the dependence of the number of learned parameters from polynomial to logarithmic.

B Memory Capacity of Linear Dynamical Systems

The hidden dimension d_h , which is the dimension of the transition matrix \mathbf{A} , plays a significant role in the expressive power of LDS. One of the most important features of the hidden dimension is that an LDS can memorize and recall inputs from up to d_{hidden} iterations in the past. This can be seen with the system where \mathbf{B}, \mathbf{C} are identity, and \mathbf{A} is given by the permutation matrix

$$\mathbf{A}_{d_{\text{hidden}}}^{\text{perm}} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix},$$

which implements the memory operator $\mathbf{y}_t = \mathbf{u}_{t-d_h}$. Observe that any method which uses fewer than d_{hidden} parameters will fail to implement this memory operator and therefore, for general linear dynamical systems, d_{hidden} parameters are *necessary*. Seemingly, this contradicts our promised results, which allows for learning a general LDS without hidden dimension dependence. The explanation is in the spectrum of the system. Notice that the eigenvalues of the permutation matrix \mathbf{A} above are the d_{hidden} roots of unity given by

$$\lambda_1, \dots, \lambda_{d_{\text{hidden}}} \in \left\{ e^{2\pi i \frac{k}{d}}, k = 1, 2, \dots, d_{\text{hidden}} \right\}.$$

Note that these eigenvalues have complex component as large as $1 - 1/d_{\text{hidden}}$. Although in general a LDS can express signals with d_{hidden} memory, and thus intuitively might require d_{hidden} parameters, there are notable special cases that allow for efficient learning, i.e. learning the LDS with far fewer parameters. A notable case is that of spectral filtering, which allows efficient learning of a *symmetric LDS* with poly-logarithmic (in the desired accuracy ε) number of parameters. The significance of a symmetric transition matrix \mathbf{A} is that it can be diagonalized over the real numbers. The natural question that arises is **which asymmetric matrices can be learned by spectral filtering efficiently, and which characterization of their spectrum allows for efficient learning?**

The answer we offer is surprisingly broad. For a LDS with transition matrix \mathbf{A} , let $\lambda_1, \dots, \lambda_d$ be its complex eigenvalues. We show that we can learn up to ε accuracy any LDS for which the largest

eigenvalue has imaginary part bounded by $\frac{1}{\text{poly log } \frac{1}{\varepsilon}}$. We note that the spectral radius can be arbitrarily close to, or even equal to, one. The only restriction is on the complex part, which is mildly constrained as a logarithmic function of ε . As per the permutation matrix example, this dependence is necessary and nearly tight - if the imaginary component of the eigenvalues of \mathbf{A} becomes large, any learning method requires parameterization that depends on the hidden dimension of the system.

C Online Version of Preconditioning

Algorithm 4 Universal Sequence Preconditioning (Online Version)

- 1: Input: sequence prediction model f_θ with initial parameter θ^0 ; loss function $\ell(\cdot)$; preconditioning coefficients $\mathbf{c}_{1:n}$.
- 2: **for** $t = 1$ to T **do**
- 3: Receive \mathbf{u}_t
- 4: Predict

$$\hat{\mathbf{y}}_t = - \sum_{j=1}^n \mathbf{c}_j \mathbf{y}_{t-j} + f_{\theta^t}(\mathbf{u}_{1:t}, \mathbf{y}_{1:t-1})$$

- 5: Observe true output \mathbf{y}_t and suffer loss $\ell_t(\hat{\mathbf{y}}_t, \mathbf{y}_t)$.
- 6: Update via projected gradient descent

$$\theta^{t+1} \leftarrow \text{proj}_{\mathcal{K}}(\theta^t - \eta_t \nabla_{\theta} \ell_t(\hat{\mathbf{y}}_t, \mathbf{y}_t))$$

- 7: **end for**
-

D Proof of Convolutional Preconditioned Regression Performance

Theorem 2.1

We will prove Theorem 2.1 first by proving the general result of Algorithm 2 for any choice of preconditioning coefficients $\mathbf{c}_{0:n}$. Then we will apply the Chebyshev coefficients to the result to get Theorem 2.1. For convenience, we restate Theorem 2.1 here.

Theorem D.1 (General Form of Theorem 2.1). *Let $\{\mathbf{u}_t\}_{t=1}^T \in \mathcal{R}^{d_{in}}$ be any sequence of inputs which satisfy $\|\mathbf{u}_t\|_2 \leq 1$ and let $\{\mathbf{y}_t\}_{t=1}^T$ be the corresponding output coming from some linear dynamical system $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ as defined per Eq. 1. Let \mathbf{P} diagonalize \mathbf{A} (note \mathbf{P} exists w.l.o.g.) and let $\kappa = \|\mathbf{P}\| \|\mathbf{P}^{-1}\|$. Assume that $\|\mathbf{B}\| \|\mathbf{C}\| \kappa \leq C_{\text{domain}}$. Assume that $\|\mathbf{B}\| \|\mathbf{C}\| \kappa \leq C_{\text{domain}}$. Then the predictions $\hat{\mathbf{y}}_1, \dots, \hat{\mathbf{y}}_T$ from Algorithm 2 satisfy*

$$\sum_{t=1}^T \|\hat{\mathbf{y}}_t - \mathbf{y}_t\|_1 \leq C_{\text{domain}} \left(\frac{3}{2} n^2 \sqrt{d_{\text{out}}} \|\mathbf{c}\|_1 \sqrt{T} + \max_{\lambda \in \lambda(\mathbf{A})} |p_n^c(\lambda)| T^2 \right).$$

Proof of Theorem D.1 For the remainder of the proof we will denote

$$\hat{\mathbf{y}}_t(\mathbf{Q}) = - \sum_{i=1}^n \mathbf{c}_i \mathbf{y}_{t-i} + \sum_{j=0}^n \mathbf{Q}_j \mathbf{u}_{t-j},$$

so that Algorithm 2 outputs $\hat{\mathbf{y}}_t = \hat{\mathbf{y}}_t(\mathbf{Q}^t)$. Recall we define the domain

$$\mathcal{K} = \{(\mathbf{Q}_0, \dots, \mathbf{Q}_{n-1}) \text{ s.t. } \|\mathbf{Q}_j\| \leq C_{\text{domain}} n \|\mathbf{c}\|_1\}.$$

For convenience, we will use the shorthand \mathbf{Q} to refer to $(\mathbf{Q}_0, \dots, \mathbf{Q}_{n-1})$. First we prove that the regret of Algorithm 2 as compared to the best $\mathbf{Q}^* \in \mathcal{K}$ in hindsight is small. Then we prove that the best $\mathbf{Q}^* \in \mathcal{K}$ in hindsight achieves small prediction error. Let

$$G = \max_{t \in [T]} \|\nabla_{\mathbf{Q}} \ell_t(\mathbf{Q}^t)\|,$$

and let

$$D = \max_{\mathbf{Q}^1, \mathbf{Q}^2 \in \mathcal{K}} \|\mathbf{Q}^1 - \mathbf{Q}^2\|.$$

By Theorem 3.1 from [24],

$$\sum_{t=1}^T \ell_t(\mathbf{Q}^t) - \min_{\mathbf{Q}^* \in \mathcal{K}} \sum_{t=1}^T \ell_t(\mathbf{Q}^*) \leq \frac{3}{2} G D \sqrt{T}.$$

Therefore it remains to bound D and G . First we bound D . By definition of \mathcal{K} ,

$$\begin{aligned} D &= \max_{\mathbf{Q}^1, \mathbf{Q}^2 \in \mathcal{K}} \|\mathbf{Q}^1 - \mathbf{Q}^2\| \\ &= \max_{(\mathbf{Q}_0^1, \dots, \mathbf{Q}_{n-1}^1), (\mathbf{Q}_0^2, \dots, \mathbf{Q}_{n-1}^2) \in \mathcal{K}} \|(\mathbf{Q}_0^1, \dots, \mathbf{Q}_{n-1}^1) - (\mathbf{Q}_0^2, \dots, \mathbf{Q}_{n-1}^2)\| \\ &\leq \sum_{j=0}^{n-1} \|\mathbf{Q}_j^1 - \mathbf{Q}_j^2\| \\ &\leq 2C_{\text{domain}} n^2 \|\mathbf{c}\|_1. \end{aligned}$$

Next we bound the gradient. Recall that

$$\begin{aligned} \ell_t(\mathbf{Q}) &= \|\hat{\mathbf{y}}_t(\mathbf{Q}) - \mathbf{y}_t\|_1 \\ &= \left\| -\sum_{i=1}^n \mathbf{c}_i \mathbf{y}_{t-i} + \sum_{j=0}^n \mathbf{Q}_j \mathbf{u}_{t-j} - \mathbf{y}_t \right\|_1. \end{aligned}$$

Note that, in general, $\nabla_{\mathbf{A}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1 = \text{sign}(\mathbf{A}\mathbf{x} - \mathbf{b})\mathbf{x}^\top$. Since $\|\mathbf{x}\mathbf{y}^\top\|_F \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$ we have

$$\|\nabla_{\mathbf{A}} f(\mathbf{A})\|_F \leq \sqrt{d} \|\mathbf{x}\|_2,$$

where d is the dimension of \mathbf{b} . Using this and the assumption that for any $t \in [T]$, $\|\mathbf{u}_t\|_2 \leq 1$, we have

$$\|\nabla_{\mathbf{Q}_i} \ell_t(\mathbf{Q})\|_F \leq \sqrt{d_{\text{out}}} \|\mathbf{u}_{t-i}\|_2 \leq \sqrt{d_{\text{out}}}.$$

Therefore,

$$G = \max_{t \in [T]} \|\nabla_{\mathbf{Q}} \ell_t(\mathbf{Q}^t)\|_F \leq n \sqrt{d_{\text{out}}}.$$

Thus we have a final regret bound

$$\sum_{t=1}^T \ell_t(\mathbf{Q}^t) - \min_{\mathbf{Q}^* \in \mathcal{K}} \ell_t(\mathbf{Q}^*) \leq \frac{3}{2} C_{\text{domain}} n^2 \sqrt{d_{\text{out}}} \|\mathbf{c}\|_1 \sqrt{T}.$$

Next we show that if $(\mathbf{u}_{1:T}, \mathbf{y}_{1:T})$ is a linear dynamical system parameterized by $(\mathbf{A}, \mathbf{B}, \mathbf{C})$, then for any $t \in [T]$,

$$\min_{\mathbf{Q}^* \in \mathcal{K}} \ell_t(\mathbf{Q}^*) \leq \|\mathbf{C}\| \|\mathbf{B}\| \cdot \max_{\lambda \in \lambda(\mathbf{A})} |p_n^{\mathbf{c}}(\lambda)| \cdot T,$$

where $p_n^{\mathbf{c}}$ denotes the polynomial

$$p_n^{\mathbf{c}}(x) \stackrel{\text{def}}{=} \sum_{i=0}^n \mathbf{c}_i x^{n-i},$$

and $\lambda(\mathbf{A})$ denotes the set of eigenvalues of \mathbf{A} . Indeed, if $(\mathbf{u}_{1:T}, \mathbf{y}_{1:T})$ is a linear dynamical system parameterized by $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ then

$$\mathbf{y}_t = \sum_{s=1}^t \mathbf{C} \mathbf{A}^{t-s} \mathbf{B} \mathbf{u}_s.$$

With some linear algebra we get that convolving $\mathbf{y}_{t:t-n}$ with coefficients $\mathbf{c}_{0:n}$ results in

$$\sum_{i=0}^n \mathbf{c}_i \mathbf{y}_{t-i} = \sum_{s=0}^{n-1} \sum_{i=0}^s \mathbf{c}_i \mathbf{C} \mathbf{A}^{s-i} \mathbf{B} \mathbf{u}_{t-s} + \sum_{s=0}^{t-n-1} \mathbf{C} p_n^{\mathbf{c}}(\mathbf{A}) \mathbf{A}^s \mathbf{B} \mathbf{u}_{t-n-s},$$

or equivalently (since $\mathbf{c}_0 = 1$ due to the assertion in Algorithm 2),

$$\mathbf{y}_t = - \sum_{i=1}^n \mathbf{c}_i \mathbf{y}_{t-i} + \sum_{s=0}^{n-1} \sum_{i=0}^s \mathbf{c}_i \mathbf{C} \mathbf{A}^{s-i} \mathbf{B} \mathbf{u}_{t-s} + \sum_{s=0}^{t-n-1} \mathbf{C} p_n^{\mathbf{c}}(\mathbf{A}) \mathbf{A}^s \mathbf{B} \mathbf{u}_{t-n-s}.$$

Set $\hat{\mathbf{Q}}_s = \sum_{i=0}^s \mathbf{c}_i \mathbf{C} \mathbf{A}^{s-i} \mathbf{B}$ and set $\hat{\mathbf{Q}} = (\hat{\mathbf{Q}}_0, \dots, \hat{\mathbf{Q}}_{n-1})$. Since we assumed $C_{\text{domain}} \geq \|\mathbf{C}\| \|\mathbf{B}\|$,

$$\|\hat{\mathbf{Q}}_i\| \leq \sum_{j=0}^i |\mathbf{c}_j| \|\mathbf{C} \mathbf{A}^{i-j} \mathbf{B}\| \leq \|\mathbf{C}\| \|\mathbf{B}\| \sum_{j=0}^i |\mathbf{c}_j| \leq C_{\text{domain}} n \|\mathbf{c}\|_1.$$

Therefore $\hat{\mathbf{Q}} \in \mathcal{K}$. Moreover,

$$\begin{aligned} \hat{\mathbf{y}}_t(\hat{\mathbf{Q}}) - \mathbf{y}_t &= \left(- \sum_{i=1}^n \mathbf{c}_i \mathbf{y}_{t-i} + \sum_{j=0}^{n-1} \hat{\mathbf{Q}}_j \mathbf{u}_{t-j} \right) \\ &\quad - \left(- \sum_{i=1}^n \mathbf{c}_i \mathbf{y}_{t-i} + \sum_{s=0}^{n-1} \sum_{i=0}^s \mathbf{c}_i \mathbf{C} \mathbf{A}^{s-i} \mathbf{B} \mathbf{u}_{t-s} + \sum_{s=0}^{t-n-1} \mathbf{C} p_n^{\mathbf{c}}(\mathbf{A}) \mathbf{A}^s \mathbf{B} \mathbf{u}_{t-n-s} \right) \\ &= \sum_{s=0}^{t-n-1} \mathbf{C} p_n^{\mathbf{c}}(\mathbf{A}) \mathbf{A}^s \mathbf{B} \mathbf{u}_{t-n-s}. \end{aligned}$$

Therefore, $\|\hat{\mathbf{y}}_t(\hat{\mathbf{Q}}) - \mathbf{y}_t\|_1 = \|\sum_{s=0}^{t-n-1} \mathbf{C} p_n^{\mathbf{c}}(\mathbf{A}) \mathbf{A}^s \mathbf{B} \mathbf{u}_{t-n-s}\|_1$. Let \mathbf{A} be diagonalized by some \mathbf{P} so that

$$\mathbf{A} = \mathbf{P} \mathbf{D} \mathbf{P}^{-1},$$

where \mathbf{D} is the diagonalization of \mathbf{A} and let $\kappa = \|\mathbf{P}\| \|\mathbf{P}^{-1}\|$, note that we can assume this w.l.o.g. since the set of diagonalizable matrices over the complex numbers is dense and therefore if \mathbf{A} is not diagonalizable we may apply an arbitrarily small perturbation to it. Then since $\max_{j \in [d_h]} |\lambda_j(\mathbf{A})| \leq 1$,

$$\begin{aligned} \|\mathbf{C} p_n^{\mathbf{c}}(\mathbf{A}) \mathbf{A}^j \mathbf{B}\| &= \|\mathbf{C} \mathbf{P} p_n^{\mathbf{c}}(\mathbf{D}) \mathbf{D}^j \mathbf{P}^{-1} \mathbf{B}\| \\ &\leq \max_{k \in [d_h]} |p_n^{\mathbf{c}}(\mathbf{D}_{kk})| \cdot \|\mathbf{C}\| \|\mathbf{P}\| \|\mathbf{P}^{-1}\| \|\mathbf{B}\| \\ &\leq \max_{\lambda \in \lambda(\mathbf{A})} |p_n^{\mathbf{c}}(\lambda)| \cdot \|\mathbf{C}\| \|\mathbf{B}\| \kappa. \end{aligned}$$

Thus,

$$\|\hat{\mathbf{y}}_t(\hat{\mathbf{Q}}) - \mathbf{y}_t\|_1 \leq \|\mathbf{C}\| \|\mathbf{B}\| \kappa \cdot \max_{\lambda \in \lambda(\mathbf{A})} |p_n^{\mathbf{c}}(\lambda)| \cdot T,$$

and so (recalling that we showed $\hat{\mathbf{Q}} \in \mathcal{K}$),

$$\begin{aligned} \min_{\mathbf{Q}^* \in \mathcal{K}} \sum_{t=1}^T \ell_t(\mathbf{Q}^*) &\leq \sum_{t=1}^T \ell_t(\hat{\mathbf{Q}}) \\ &= \sum_{t=1}^T \|\hat{\mathbf{y}}_t(\hat{\mathbf{Q}}) - \mathbf{y}_t\|_1 \\ &\leq \|\mathbf{C}\| \|\mathbf{B}\| \kappa \cdot \max_{\lambda \in \lambda(\mathbf{A})} |p_n^{\mathbf{c}}(\lambda)| \cdot T^2. \end{aligned}$$

Since $C_{\text{domain}} \geq \|\mathbf{C}\| \|\mathbf{B}\| \kappa$ we conclude,

$$\sum_{t=1}^T \ell_t(\mathbf{Q}^t) \leq C_{\text{domain}} \left(\frac{3}{2} n^2 \sqrt{d_{\text{out}}} \|\mathbf{c}\|_1 \sqrt{T} + \max_{\lambda \in \lambda(\mathbf{A})} |p_n^{\mathbf{c}}(\lambda)| T^2 \right).$$

■

Next we choose \mathbf{c} to be the coefficients of the n -th monic Chebyshev polynomial to get the original theorem, Theorem 2.1 restated here for convenience.

Theorem (Restatement of Theorem 2.1). Let $\{\mathbf{u}_t\}_{t=1}^T \in \mathcal{R}^{d_{\text{in}}}$ be any sequence of inputs which satisfy $\|\mathbf{u}_t\|_2 \leq 1$ and let $\{\mathbf{y}_t\}_{t=1}^T$ be the corresponding output coming from some linear dynamical system $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ as defined per Eq. 1. Let \mathbf{P} diagonalize \mathbf{A} (note \mathbf{P} exists w.l.o.g.) and let $\kappa = \|\mathbf{P}\| \|\mathbf{P}^{-1}\|$. Assume that $\|\mathbf{B}\| \|\mathbf{C}\| \kappa \leq C_{\text{domain}}$. Let $\lambda_1, \dots, \lambda_{d_h}$ denote the spectrum of \mathbf{A} . If

$$\max_{j \in [d_h]} |\text{Arg}(\lambda_j)| \leq \left(64 \log_2 \left(\frac{8T^{3/2}}{3\sqrt{d_{\text{out}}}} \right) \right)^{-2},$$

then the predictions $\hat{\mathbf{y}}_1, \dots, \hat{\mathbf{y}}_T$ from Algorithm 2 where the preconditioning coefficients $\mathbf{c}_{0:n}$ are chosen to be the coefficients of the n -th monic Chebyshev polynomial satisfy

$$\frac{1}{T} \sum_{t=1}^T \|\hat{\mathbf{y}}_t - \mathbf{y}_t\|_1 \leq \frac{9C_{\text{domain}} \sqrt{d_{\text{out}}} \log_2^2(3T)}{T^{2/13}}.$$

Proof. From Theorem D.1 we have

$$\sum_{t=1}^T \ell_t(\mathbf{Q}^t) \leq C_{\text{domain}} \left(\frac{3}{2} n^2 \sqrt{d_{\text{out}}} \|\mathbf{c}\|_1 \sqrt{T} + \max_{\lambda \in \lambda(\mathbf{A})} |p_n^c(\lambda)| T^2 \right).$$

By Lemma 3.1 if for any eigenvalue λ of \mathbf{A} , $|\arg(\lambda)| \leq 1/64n^2$ then

$$\max_{\lambda \in \lambda(\mathbf{A})} |p_n^c(\lambda)| \leq \frac{1}{2^{n-2}}.$$

Moreover, by Lemma 3.2, $\|\mathbf{c}\|_1 \leq 2^{0.3n}$. Thus the Chebyshev-conditioned predictor class satisfies

$$\sum_{t=1}^T \ell_t(\mathbf{Q}^t) \leq C_{\text{domain}} \left(\frac{3}{2} n^2 \sqrt{d_{\text{out}}} 2^{0.3n} \sqrt{T} + 2^{-(n-2)} T^2 \right).$$

Picking

$$n = \frac{10}{13} \log_2 \left(\frac{8}{3\sqrt{d_{\text{out}}}} T^{3/2} \right),$$

we get

$$\begin{aligned} \sum_{t=1}^T \ell_t(\mathbf{Q}^t) &\leq 2 \left(\frac{3}{2} \left(\frac{10}{13} \log_2 \left(\frac{8}{3\sqrt{d_{\text{out}}}} T^{3/2} \right) \right)^2 \sqrt{d_{\text{out}}} \right)^{10/13} 4^{3/13} T^{11/13} \\ &\leq 9C_{\text{domain}} \sqrt{d_{\text{out}}} \log_2^2(3T)^2 T^{11/13}. \end{aligned}$$

Dividing both sides by T we get the stated result. \blacksquare

E Proof of Theorem 2.2

E.1 Preliminaries and Notation

We analyze the output sequence $\{\mathbf{y}_t\}_{t=1}^T$ generated by a linear dynamical system $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ with inputs $\{\mathbf{u}_t\}_{t=1}^T \in \mathcal{R}^{d_{\text{in}}}$ satisfying $\|\mathbf{u}_t\|_2 \leq 1$. We assume that the dynamics matrix \mathbf{A} is diagonalizable. Let \mathbf{P} be the matrix that diagonalizes \mathbf{A} (which exists w.l.o.g. as diagonalizable matrices are dense in $\mathbb{C}^{d \times d}$), and let $\kappa = \|\mathbf{P}\| \|\mathbf{P}^{-1}\|$ denote the condition number of the eigenbasis.

Recall the spectral domain $\mathbb{C}_\beta = \{z \in \mathbb{C} \mid |z| \leq 1, |\arg(z)| \leq \beta\}$ defined in Section 2.2. For a given polynomial $p_n^c(x)$ with coefficients \mathbf{c} , let $B_n = \max_{\alpha \in \mathbb{C}_\beta} |p_n^c(\alpha)|$.

Recall the definitions of $\tilde{p}_n^c(\alpha)$, $\tilde{\mu}(\alpha)$, and the spectral covariance matrix \mathbf{Z} from Section 2.2 (Eqs. 4–5).

Let ϕ_1, \dots, ϕ_T be the eigenvectors of \mathbf{Z} . Let $\mathbf{u}_{t:1}$ be the concatenated inputs up to time t which are padded to create a length T vector,

$$\mathbf{u}_{t:1} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{u}_t & \mathbf{u}_{t-1} & \dots & \mathbf{u}_1 & 0 & \dots & 0 \end{bmatrix}^\top. \quad (6)$$

E.2 General Form of Main Result

First, we prove a more general form that holds for any choice of $\mathbf{c}_{0:n}$ and resulting polynomial p_n^c .

Theorem E.1. *Let the system and notation be defined as in Section E.1. If the radius parameters of Algorithm 3 are set to:*

$$\begin{aligned} R_Q &= \|\mathbf{C}\| \|\mathbf{B}\| \|\mathbf{c}\|_1 \\ R_M &= 2\|\mathbf{C}\| \|\mathbf{B}\| \kappa \log(T) \beta^{4/3} T^{7/6} B_n \end{aligned}$$

then the predictions $\hat{\mathbf{y}}_1, \dots, \hat{\mathbf{y}}_T$ from Algorithm 3 satisfy

$$\sum_{t=1}^T \ell_t(\mathbf{Q}^t, \mathbf{M}^t) \leq 18\kappa \|\mathbf{C}\| \|\mathbf{B}\| \log(T)(n+k) \left(T^3 \beta^{4/3} B_n k + n \|\mathbf{c}\|_1 T^{1/2} \right).$$

Critically, we observe that the guaranteed regret bound does not depend on the hidden dimension of the dynamics matrix \mathbf{A} . While the general version of Algorithm 3 is interesting in its own right, we show that by choosing the coefficients of the algorithm to be the Chebyshev coefficients we obtain sublinear absolute error.

Theorem (Detailed Version of Theorem 2.2). Let the assumptions of Theorem E.1 hold. Furthermore, suppose that for any eigenvalue, λ_j , of \mathbf{A}

$$\max_{j \in [d_h]} |\arg(\lambda_j)| \leq T^{-1/4} \cdot T^{-13p/4},$$

then the predictions $\hat{\mathbf{y}}_1, \dots, \hat{\mathbf{y}}_T$ from Algorithm 3 where the preconditioning coefficients $\mathbf{c}_{0:n}$ are chosen to be the coefficients of the n -th monic Chebyshev polynomial satisfy

$$\frac{1}{T} \sum_{t=1}^T \|\hat{\mathbf{y}}_t - \mathbf{y}_t\|_1 \leq \tilde{O} \left(\frac{\kappa \|\mathbf{C}\| \|\mathbf{B}\|}{T^p} \right),$$

where $\tilde{O}(\cdot)$ hides polylogarithmic factors in T .

Proof. We introduce parameter β to denote the maximum argument of the eigenvalues of \mathbf{A} so that the spectrum of \mathbf{A} lies in \mathbb{C}_β . From Theorem D.1 we have

$$\sum_{t=1}^T \ell_t(\mathbf{Q}^t, \mathbf{M}^t) \leq 18\kappa \|\mathbf{C}\| \|\mathbf{B}\| \log(T)(n+k) \left(T^3 \beta^{4/3} B_n k + n \|\mathbf{c}\|_1 T^{1/2} \right).$$

By Lemma 3.1 if for any eigenvalue λ of \mathbf{A} , $|\arg(\lambda)| \leq 1/64n^2$ then

$$\max_{\lambda \in \lambda(\mathbf{A})} |p_n^c(\lambda)| \leq \frac{1}{2^{n-2}} = 4 \cdot 2^{-n}. \quad (7)$$

We will choose

$$n = \frac{10}{13} \log_2 \left(T^{-1/2} T^3 \beta^{4/3} \right), \quad (8)$$

and so if $\beta < 1/T^{1/4}$ then $\beta < 1/64n^2$, meaning that Eq. 7 holds. Moreover, by Lemma 3.2 $\|\mathbf{c}\|_1 \leq 2^{0.3n}$. Thus the Chebyshev-conditioned predictor class satisfies

$$\sum_{t=1}^T \ell_t(\mathbf{Q}^t) \leq 60\kappa \|\mathbf{C}\| \|\mathbf{B}\| \log(T)(n+k) \left(T^3 \beta^{4/3} 2^{-n} k + n 2^{0.3n} T^{1/2} \right)$$

Then for n as chosen above in Eq. 8

$$T^3 \beta^{4/3} 2^{-n} + 2^{0.3n} T^{1/2} = 2 \left(T^{14} \beta^4 \right)^{1/13}.$$

Assuming that $\beta < 1$ we have that $n < 3 \log_2(T)$ and therefore

$$\begin{aligned} \sum_{t=1}^T \ell_t(\mathbf{Q}^t) &\leq 120\kappa \|\mathbf{C}\| \|\mathbf{B}\| \log(T) n^2 k^2 \left(T^3 \beta^{4/3} 2^{-n} + 2^{0.3n} T^{1/2} \right) \\ &\quad \text{(Factoring out } n \text{ and } k \text{ and using } n+k \leq 2nk) \\ &\leq 240\kappa \|\mathbf{C}\| \|\mathbf{B}\| \log(T) (3 \log_2(T))^2 k^2 \left(T^{14} \beta^4 \right)^{1/13} \\ &\quad \text{(Plugging in chosen value for } n.) \\ &\leq 720\kappa \|\mathbf{C}\| \|\mathbf{B}\| \log(T)^3 k^2 \left(T^{14} \beta^4 \right)^{1/13}. \end{aligned}$$

Therefore if $\beta = T^{-1/4} \cdot T^{-13p/4}$ then the final accumulated error is

$$\sum_{t=1}^T \ell_t(\mathbf{Q}^t) \leq 720\kappa \|\mathbf{C}\| \|\mathbf{B}\| \log(T)^3 k^2 T^{1-p},$$

which is sublinear as long as $p > 0$. ■

Remark on the loss function. The reader may notice we use the ℓ_1 , or absolute loss, rather than Euclidean or other loss functions. All norms are equivalent up to the (output) dimension, and thus learning to predict as well as the best linear dynamical system in hindsight is meaningful in any norm. However, we make this technical choice since it greatly simplifies the regret bounds, and in particular the bound on the gradient norms, which is technically involved. We conjecture that sublinear regret bounds are attainable in other norms as well, and leave it for future work. We prove Theorem 2.2 on an (equivalent) algorithm, where we rescale the parameter \mathbf{M}_j by \sqrt{T} and the input $\langle \phi_j, \mathbf{u}_{(t-n-i):1} \rangle$ by $1/\sqrt{T}$. We account for this rescaling by increasing the size of the domain for \mathbf{M} by \sqrt{T} . The proof of Theorem 2.2 proceeds in two parts. The first is to show that any linear dynamical signal is well approximated by a predictor of the form in Algorithm 3 $\hat{\mathbf{y}}(\mathbf{Q}, \mathbf{M})$ where $(\mathbf{Q}, \mathbf{M}) \in \mathcal{K}$.

Lemma E.2 (Approximation Lemma). Suppose $\mathbf{y}_{1:T}$ evolves as a linear dynamical system characterized by $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ as in Eq. 1 satisfying the assumptions in Section E.1. Consider domain

$$\mathcal{K} = \{(\mathbf{Q}_1, \dots, \mathbf{Q}_n, \mathbf{M}_1, \dots, \mathbf{M}_k) \text{ s.t. } \|\mathbf{Q}_i\| \leq R_Q, \text{ and } \|\mathbf{M}_i\| \leq R_M\}.$$

If

$$\begin{aligned} R_Q &\geq \|\mathbf{C}\| \|\mathbf{B}\| \|\mathbf{c}\|_1, \\ R_M &\geq 2\|\mathbf{C}\| \|\mathbf{B}\| \kappa \log(T) \beta^{4/3} T^{7/6} B_n, \end{aligned}$$

then there exists $(\hat{\mathbf{Q}}_1, \dots, \hat{\mathbf{Q}}_n, \hat{\mathbf{M}}_1, \dots, \hat{\mathbf{M}}_k) \in \mathcal{K}$ such that for prediction (as in Algorithm 3)

$$\hat{\mathbf{y}}_t(\hat{\mathbf{Q}}, \hat{\mathbf{M}}) = - \sum_{i=1}^n c_i \mathbf{y}_{t-i} + \sum_{j=0}^n \hat{\mathbf{Q}}_j^t \mathbf{u}_{t-j} + \frac{1}{\sqrt{T}} \sum_{j=1}^k \hat{\mathbf{M}}_j^t \phi_j^\top \tilde{\mathbf{u}}_{t-n-1:1},$$

it holds that

$$\|\hat{\mathbf{y}}_t - \mathbf{y}_t\|_1 \leq 6\kappa \|\mathbf{C}\| \|\mathbf{B}\| \log(T) T^2 \beta^{4/3} B_n.$$

Proof of Lemma E.2 Suppose $\mathbf{y}_{1:T}$ evolves as a linear dynamical system characterized by $(\mathbf{A}, \mathbf{B}, \mathbf{C})$. Then given inputs $\mathbf{u}_{1:t}$,

$$\mathbf{y}_t = \sum_{s=1}^t \mathbf{C} \mathbf{A}^{t-s} \mathbf{B} \mathbf{u}_s.$$

With some linear algebra we get that convolving $\mathbf{y}_{t:t-n}$ with coefficients $\mathbf{c}_{0:n}$ results in

$$\mathbf{y}_t = - \sum_{i=1}^n c_i \mathbf{y}_{t-i} + \sum_{s=0}^{n-1} \sum_{i=0}^s c_i \mathbf{C} \mathbf{A}^{s-i} \mathbf{B} \mathbf{u}_{t-s} + \sum_{s=0}^{t-n-1} C p_n^c(\mathbf{A}) \mathbf{A}^s \mathbf{B} \mathbf{u}_{t-n-s}. \quad (9)$$

Set $\hat{\mathbf{Q}}_s = \sum_{i=0}^s c_i \mathbf{C} \mathbf{A}^{s-i} \mathbf{B}$ and set $\hat{\mathbf{Q}} = (\hat{\mathbf{Q}}_0, \dots, \hat{\mathbf{Q}}_{n-1})$. Then

$$\|\hat{\mathbf{Q}}_i\| \leq \sum_{j=0}^i |c_j| \|\mathbf{C} \mathbf{A}^{i-j} \mathbf{B}\| \leq \|\mathbf{C}\| \|\mathbf{B}\| \sum_{j=0}^i |c_j| \leq \|\mathbf{C}\| \|\mathbf{B}\| \|\mathbf{c}\|_1.$$

Next we turn our attention to the spectral filtering parameters. Using the notation $\tilde{p}_n^c(\alpha)$ and $\tilde{\mu}(\alpha)$ from Section E.1, we define the combined vector $\mu_{\tilde{p}_n}(\alpha) \stackrel{\text{def}}{=} p_n^c(\alpha) \tilde{\mu}(\alpha)$. Let \mathbf{D} be the diagonalization

of \mathbf{A} . Then,

$$\begin{aligned}
\sum_{s=0}^{t-n-1} \mathbf{C} \tilde{p}_n^{\mathbf{c}}(\mathbf{A}) \mathbf{A}^s \mathbf{B} \mathbf{u}_{t-n-s} &= \sum_{s=0}^{t-n-1} \mathbf{C} \mathbf{P} \tilde{p}_n^{\mathbf{c}}(\mathbf{D}) \mathbf{D}^s \mathbf{P}^{-1} \mathbf{B} \mathbf{u}_{t-n-s} \\
&= \sum_{s=0}^{t-n-1} \mathbf{C} \mathbf{P} \left(\sum_{m=1}^{d_h} \tilde{p}_n^{\mathbf{c}}(\alpha_m) \alpha_m^s \mathbf{e}_m \mathbf{e}_m^{\top} \right) \mathbf{P}^{-1} \mathbf{B} \mathbf{u}_{t-n-s} \\
&= \sum_{m=1}^{d_h} \sum_{s=0}^{t-n-1} \mathbf{C} \mathbf{P} \mathbf{e}_m \mathbf{e}_m^{\top} \mathbf{P}^{-1} \mathbf{B} \tilde{p}_n^{\mathbf{c}}(\alpha_m) \alpha_m^s \mathbf{u}_{t-n-s} \\
&= \sum_{m=1}^{d_h} \mathbf{C} \mathbf{P} \mathbf{e}_m \mathbf{e}_m^{\top} \mathbf{P}^{-1} \mathbf{B} \mu_{\tilde{p}_n}(\alpha_m)^{\top} \tilde{\mathbf{u}}_{t-n-1:1} \\
&= \sum_{m=1}^{d_h} \mathbf{C} \mathbf{P} \mathbf{e}_m \mathbf{e}_m^{\top} \mathbf{P}^{-1} \mathbf{B} \mu_{\tilde{p}_n}(\alpha_m)^{\top} \left(\sum_{j=1}^{T-n} \phi_j \phi_j^{\top} \right) \tilde{\mathbf{u}}_{t-n-1:1} \\
&\quad \text{(Orthonormality of the filters)} \\
&= \sum_{j=1}^{T-n} \left(\sum_{m=1}^{d_h} \mathbf{C} \mathbf{P} \mathbf{e}_m \mathbf{e}_m^{\top} \mathbf{P}^{-1} \mathbf{B} \mu_{\tilde{p}_n}(\alpha_m)^{\top} \phi_j \right) \phi_j^{\top} \tilde{\mathbf{u}}_{t-n-1:1}.
\end{aligned}$$

Therefore defining

$$\hat{\mathbf{M}}_j \stackrel{\text{def}}{=} T^{1/2} \sum_{m=1}^{d_h} \mathbf{C} \mathbf{P} \mathbf{e}_m \mathbf{e}_m^{\top} \mathbf{P}^{-1} \mathbf{B} \mu_{\tilde{p}_n}(\alpha_m)^{\top} \phi_j,$$

we have

$$\sum_{s=0}^{t-n-1} \mathbf{C} \tilde{p}_n^{\mathbf{c}}(\mathbf{A}) \mathbf{A}^s \mathbf{B} \mathbf{u}_{t-n-s} = \sum_{j=1}^{T-n} \hat{\mathbf{M}}_j \frac{\phi_j^{\top} \tilde{\mathbf{u}}_{t-n-1:1}}{\sqrt{T}}.$$

Next we bound the norm of $\hat{\mathbf{M}}_j$. Let \mathbf{S} be the diagonal matrix with entries $\mathbf{S}_{mm} = \mu_{\tilde{p}_n}(\alpha_m)^{\top} \phi_j$. Note that $\hat{\mathbf{M}}_j = \mathbf{C} \mathbf{P} \mathbf{S} \mathbf{P}^{-1} \mathbf{B}$. For short, let $C_{\kappa} = \|\mathbf{C}\| \|\mathbf{B}\| \kappa$. Recalling that $\mu_{\tilde{p}_n}(\alpha) = p_n^{\mathbf{c}}(\alpha) \mu(\alpha)$,

$$\|\hat{\mathbf{M}}_j\| = \|T^{1/2} \mathbf{C} \mathbf{P} \mathbf{S} \mathbf{P}^{-1} \mathbf{B}\| \leq C_{\kappa} T^{1/2} \max_{m \in [d_h]} |p_n^{\mathbf{c}}(\alpha_m)| \cdot \max_{m \in [d_h]} |\tilde{\mu}(\alpha_m)^{\top} \phi_j|. \quad (10)$$

By Lemma [E.4](#),

$$\max_{\alpha \in \mathbb{C}_{\beta}} |\tilde{\mu}(\alpha)^{\top} \phi_j| \leq 2 \log(T) \beta^{4/3} T^{2/3}.$$

Therefore,

$$\|\hat{\mathbf{M}}_j\| \leq 2C \log(T) \beta^{4/3} T^{7/6} \max_{\alpha \in \mathbb{C}_{\beta}} |p_n^{\mathbf{c}}(\alpha)|.$$

Therefore $(\hat{\mathbf{Q}}, \hat{\mathbf{M}}) \in \mathcal{K}$ for the chosen radius of \mathcal{K} (i.e. R_Q and R_M). Finally, we bound the truncation error:

$$\hat{\mathbf{y}}_t(\hat{\mathbf{Q}}, \hat{\mathbf{M}}) - \mathbf{y}_t = \sum_{j=k+1}^{T-n} \hat{\mathbf{M}}_j \frac{\phi_j^{\top} \tilde{\mathbf{u}}_{t-n-1:1}}{\sqrt{T}}.$$

$$\begin{aligned}
\|\hat{\mathbf{y}}_t(\hat{\mathbf{Q}}, \hat{\mathbf{M}}) - \mathbf{y}_t\|_2 &\leq \sum_{j=k+1}^{T-n} \|\hat{\mathbf{M}}_j\| \|\phi_j^\top\|_1 \|\mathbf{u}_{(t-n-1):1}\|_\infty T^{-1/2} \\
&\leq \sum_{j=k+1}^{T-n} \|\hat{\mathbf{M}}_j\| \quad (\|\phi_j\|_2 = 1 \implies \|\phi_j\|_1 \leq \sqrt{T}) \\
&\leq C_\kappa T^{1/2} B_n \cdot \sum_{j=k+1}^{T-n} \max_{\alpha \in \mathbb{C}_\beta} |\tilde{\mu}(\alpha)^\top \phi_j| \quad (\text{Bound on } \|\hat{\mathbf{M}}_j\| \text{ from Eq. 10}) \\
&\leq C_\kappa T^{1/2} B_n \cdot (6 \log(T) \beta^{4/3} T^{3/2}) \quad (\text{Lemma E.4}) \\
&= 6 \|\mathbf{C}\| \|\mathbf{B}\| \kappa \log(T) T^2 \beta^{4/3} B_n. \quad (\text{Plugging in value for } C_\kappa.)
\end{aligned}$$

The next result provides the regret of Online Gradient Descent when compared to the best $(\mathbf{Q}^*, \mathbf{M}^*) \in \mathcal{K}$.

Lemma E.3 (Online Gradient Descent). Recall the domain \mathcal{K} in Algorithm 3 be

$$\mathcal{K} = \{(\mathbf{Q}_1, \dots, \mathbf{Q}_n, \mathbf{M}_1, \dots, \mathbf{M}_k) \text{ s.t. } \|\mathbf{Q}_i\| \leq R_Q \text{ and } \|\mathbf{M}_i\| \leq R_M\}.$$

The iterates output by Algorithm 3 satisfy

$$\sum_{t=1}^T \ell_t(\mathbf{Q}^t, \mathbf{M}^t) - \min_{(\mathbf{Q}^*, \mathbf{M}^*) \in \mathcal{K}} \sum_{t=1}^T \ell_t(\mathbf{Q}^*, \mathbf{M}^*) \leq \frac{3}{2} (nR_Q + kR_M)(n+k) \sqrt{d_{\text{out}}} \sqrt{T}.$$

Proof of Lemma E.3 Let $G = \max_{t \in [T]} \|\nabla_{\mathbf{Q}, \mathbf{M}} \ell_t(\mathbf{Q}^t, \mathbf{M}^t)\|$ and let

$$D = \max_{(\mathbf{Q}^1, \mathbf{M}^1), (\mathbf{Q}^2, \mathbf{M}^2) \in \mathcal{K}} \|(\mathbf{Q}^1, \mathbf{M}^1) - (\mathbf{Q}^2, \mathbf{M}^2)\|.$$

By Theorem 3.1 from [24],

$$\sum_{t=1}^T \ell_t(\mathbf{Q}^t, \mathbf{M}^t) - \min_{(\mathbf{Q}^*, \mathbf{M}^*) \in \mathcal{K}} \sum_{t=1}^T \ell_t(\mathbf{Q}^*, \mathbf{M}^*) \leq \frac{3}{2} G D \sqrt{T}.$$

Therefore it remains to bound G and D . By definition of \mathcal{K} we have

$$D \leq nR_Q + kR_M.$$

For G we compute the subgradient at any \mathbf{Q}_i and \mathbf{M}_i . Note that, in general, $\nabla_{\mathbf{A}} \|\mathbf{Ax} - \mathbf{b}\|_1 = \text{sign}(\mathbf{Ax} - \mathbf{b}) \mathbf{x}^\top$. Since $\|\mathbf{xy}^\top\|_F \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$ we have

$$\|\nabla_{\mathbf{A}} f(\mathbf{A})\|_F \leq \sqrt{d} \|\mathbf{x}\|_2,$$

where d is the dimension of \mathbf{b} . Using this and the assumption that for any $t \in [T]$, $\|\mathbf{u}_t\|_2 \leq 1$, we have

$$\begin{aligned}
\|\nabla_{\mathbf{M}_j} \ell_t(\mathbf{Q}, \mathbf{M})\| &\leq \sqrt{d_{\text{out}}} \|\phi_j^\top \tilde{\mathbf{u}}_{t-n-1:1} T^{-1/2}\| \\
&\leq \sqrt{d_{\text{out}}} \|\phi_j\|_1 \|\tilde{\mathbf{u}}_{t-n-1:1}\|_\infty T^{-1/2} \\
&\leq \sqrt{d_{\text{out}}}. \quad (\|\tilde{\mathbf{u}}_{t-n-1:1}\|_\infty \leq 1 \text{ and } \|\phi_j\|_1 \leq \sqrt{T} \text{ since } \|\phi_j\|_2 \leq 1)
\end{aligned}$$

Next,

$$\|\nabla_{\mathbf{Q}_i} \ell_t(\mathbf{Q})\|_F \leq \sqrt{d_{\text{out}}} \|\mathbf{u}_{t-i}\|_2 \leq \sqrt{d_{\text{out}}}.$$

Therefore,

$$G = \max_{t \in [T]} \|\nabla_{(\mathbf{Q}, \mathbf{M})} \ell_t(\mathbf{Q}^t, \mathbf{M}^t)\|_F \leq (n+k) \sqrt{d_{\text{out}}}.$$

Therefore, we have

$$\sum_{t=1}^T \ell_t(\mathbf{Q}^t, \mathbf{M}^t) - \min_{(\mathbf{Q}^*, \mathbf{M}^*) \in \mathcal{K}} \sum_{t=1}^T \ell_t(\mathbf{Q}^*, \mathbf{M}^*) \leq \frac{3}{2} (nR_Q + kR_M)(n+k) \sqrt{d_{\text{out}}} \sqrt{T}.$$

Combining Lemma E.2 and Lemma E.3 proves Theorem 2.2.

Proof of Theorem 2.2 By Lemma E.3 the iterates from Algorithm 3 $(\mathbf{Q}^1, \mathbf{M}^1), \dots, (\mathbf{Q}^T, \mathbf{M}^T)$ satisfy

$$\sum_{t=1}^T \ell_t(\mathbf{Q}^t, \mathbf{M}^t) - \min_{(\mathbf{Q}^*, \mathbf{M}^*) \in \mathcal{K}} \sum_{t=1}^T \ell_t(\mathbf{Q}^*, \mathbf{M}^*) \leq \frac{3}{2}(nR_Q + kR_M)(n+k)\sqrt{d_{\text{out}}}\sqrt{T}.$$

By Lemma E.2 if $\mathbf{y}_{1:T}$ comes from a linear dynamical system then for large enough radius parameters of \mathcal{K} (R_Q and R_M), there exists $(\hat{\mathbf{Q}}, \hat{\mathbf{M}}) \in \mathcal{K}$ such that

$$\sum_{t=1}^T \ell_t(\hat{\mathbf{Q}}, \hat{\mathbf{M}}) \leq T \left(6\kappa \|\mathbf{C}\| \|\mathbf{B}\| \log(T) T^2 \beta^{4/3} B_n \right).$$

Let $C_T = 6\kappa \|\mathbf{C}\| \|\mathbf{B}\| \log(T)$ for short. Since $(\hat{\mathbf{Q}}, \hat{\mathbf{M}}) \in \mathcal{K}$ we must also have that

$$\min_{(\mathbf{Q}^*, \mathbf{M}^*) \in \mathcal{K}} \sum_{t=1}^T \ell_t(\mathbf{Q}^*, \mathbf{M}^*) \leq C_T T^3 \beta^{4/3} B_n.$$

Therefore,

$$\begin{aligned} \sum_{t=1}^T \ell_t(\mathbf{Q}^t, \mathbf{M}^t) &\leq \min_{(\mathbf{Q}^*, \mathbf{M}^*) \in \mathcal{K}} \sum_{t=1}^T \ell_t(\mathbf{Q}^*, \mathbf{M}^*) + \frac{3}{2}(nR_Q + kR_M)(n+k)\sqrt{d_{\text{out}}}\sqrt{T} \\ &\leq C_T T^3 \beta^{4/3} B_n + \frac{3}{2}(nR_Q + kR_M)(n+k)\sqrt{d_{\text{out}}}\sqrt{T}. \end{aligned}$$

Recall that we set $R_Q = \|\mathbf{C}\| \|\mathbf{B}\| \|\mathbf{c}\|_1$ and $R_M = 2\|\mathbf{C}\| \|\mathbf{B}\| \kappa \log(T) \beta^{4/3} T^{7/6} B_n$. Using C_T we bound $R_M \leq C_T \beta^{4/3} T^{7/6} B_n$. We conclude,

$$\begin{aligned} \sum_{t=1}^T \ell_t(\mathbf{Q}^t, \mathbf{M}^t) &\leq C_T T^3 \beta^{4/3} B_n + \frac{3}{2} C_T \left(n \|\mathbf{c}\|_1 + k \beta^{4/3} T^{7/6} B_n \right) (n+k) \sqrt{d_{\text{out}}}\sqrt{T} \\ &\leq C_T \left(T^3 \beta^{4/3} B_n + \frac{3}{2} n(n+k) \|\mathbf{c}\|_1 T^{1/2} + \frac{3}{2} k(n+k) \beta^{4/3} T^{5/3} B_n \right) \\ &\leq C_T \left(3T^3 \beta^{4/3} B_n k(n+k) + \frac{3}{2} n(n+k) \|\mathbf{c}\|_1 T^{1/2} \right) \\ &\quad \text{(Bounding the right-most term by the left-most term times } k(n+k)) \\ &\leq 6\kappa \|\mathbf{C}\| \|\mathbf{B}\| \log(T) \left(3T^3 \beta^{4/3} B_n k(n+k) + \frac{3}{2} n(n+k) \|\mathbf{c}\|_1 T^{1/2} \right) \\ &\quad \text{(Plugging in } C_T = 5\kappa \|\mathbf{C}\| \|\mathbf{B}\| \log(T)) \\ &\leq 18\kappa \|\mathbf{C}\| \|\mathbf{B}\| \log(T) (n+k) \left(T^3 \beta^{4/3} B_n k + n \|\mathbf{c}\|_1 T^{1/2} \right). \end{aligned}$$

■

E.3 Proving Approximation Lemma E.2

Lemma E.4. Let $\tilde{\mu}(\alpha)$ and \mathbf{Z} be as defined in Section E.1 and let ϕ_1, \dots, ϕ_T be the eigenvectors of \mathbf{Z}_T .

$$\max_{\alpha \in \mathbb{C}_\beta} |\tilde{\mu}(\alpha)^\top \phi_j| \leq 2 \log(T) \beta^{4/3} T^{2/3}.$$

Moreover,

$$\sum_{j=1}^T |\tilde{\mu}(\alpha)^\top \phi_j| \leq 6 \log(T) \beta^{4/3} T^{3/2}.$$

E.3.1 Proof of the Spectral Filtering Property Lemma E.4

In order to prove Lemma E.4 we require two further helper lemmas. The first is Lemma E.5 which roughly argues that the Lipschitz constant of a function $f : \mathbb{C}_\beta \rightarrow \mathbb{R}$ can be bounded by a polynomial of the expectation of the function on \mathbb{C}_β . The second is Lemma E.6, which argues that the matrix \mathbf{Z} from Eq. 5 that we use to derive the new spectral filters has small eigenvalues.

Lemma E.5. Let $L > 0$, $g_{\max} > 0$ and $0 \leq \beta \leq 1$. Define

$$\mathbb{C}_\beta = \{z \in \mathbb{C} : |z| \leq 1, |\arg(z)| \leq \beta\},$$

and let \mathcal{F} be the set of non-negative, L -Lipschitz functions $f : \mathbb{C}_\beta \rightarrow \mathbb{R}$ such that $\max_{z \in \mathbb{C}_\beta} f(z) = g_{\max}$. Then

$$\min_{f \in \mathcal{F}} \int_{\mathbb{C}_\beta} f(z) dz \geq \frac{\beta g_{\max}^3}{24 L^2}.$$

Proof of Lemma E.5 **1. The extremal function.** Fix $f \in \mathcal{F}$ and choose z_* with $f(z_*) = g_{\max}$. Among admissible functions we consider only those that decrease as fast as the Lipschitz constraint allows in every radial direction, replacing f by the *extremal cone*. Set

$$h(z) := [g_{\max} - L|z - z_*|]_+, \quad r := \frac{g_{\max}}{L}.$$

Because $0 \leq h \leq f$ on \mathbb{C}_β , it suffices to lower bound $\int_{\mathbb{C}_\beta} h$.

2. Geometry around z_* . It can be seen by Euclidean geometry, the point $z_* = 0$ minimize the volume of the intersection with \mathbb{C}_β , up to a factor of 4, to make it the extremal function. It is depicted in Figure 2. Notice that the area of the intersection is colored in blue, and it is particularly easy to integrate over.

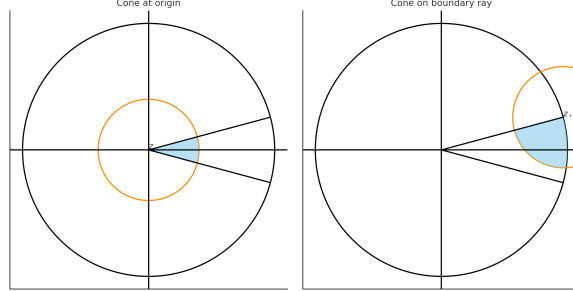


Figure 2: The origin minimizes the intersection area up to factor $\frac{1}{4}$.

3. Integrating h .

$$\int_{\mathbb{C}_\beta} h \geq \int_0^\beta \int_0^r (g_{\max} - L\rho) \rho d\rho d\varphi = \frac{\beta g_{\max}^3}{6 L^2}.$$

Since $h \in \mathcal{F}$, and is the extremal function, we obtain the Lemma statement. ■

Lemma E.6. Recall from Section E.1

$$\mathbf{Z} \stackrel{\text{def}}{=} \int_{\alpha \in \mathbb{C}_\beta} \tilde{\mu}(\alpha) \tilde{\mu}(\bar{\alpha})^\top d\alpha.$$

There are at most $2 \log(T)$ eigenvalues with absolute value greater than β .

Proof of Lemma E.6 The (j, k) -th entry of \mathbf{Z} is

$$\mathbf{Z}_{jk} = \int_{r \in [0,1]} \int_{\theta \in \mathbb{C}_\beta} (1 - r^2 e^{2i\theta}) (r e^{i\theta})^j (1 - r^2 e^{-2i\theta}) (r e^{-i\theta})^k r dr d\theta.$$

Evaluating this integral for $j = k$ we get

$$\mathbf{Z}_{jj} = \beta \left(\frac{1}{j+1} + \frac{1}{j+3} \right) - \frac{\sin(2\beta)}{j+2} \leq \frac{3\beta}{j+1}.$$

Therefore,

$$\text{tr}(\mathbf{Z}) \leq 6\beta \log(T).$$

Using the fact that \mathbf{Z} has nonnegative eigenvalues, we have that the number of eigenvalues larger than β , n_β , satisfies

$$\sum_{i=1}^T \lambda_i \geq \sum_{\lambda_i \geq \beta} \lambda_i \geq \beta n_\beta.$$

Since the sum of \mathbf{Z} 's eigenvalues is bounded by its trace we see

$$\text{tr}(\mathbf{Z}) \geq \beta n_\beta.$$

Using our bound on the trace of \mathbf{Z} we have

$$n_\beta \leq \frac{2\beta \log(T)}{\beta} = 6 \log(T).$$

■

With Lemma E.5 and Lemma E.6 in hand, we are ready to prove Lemma E.4.

Proof of Lemma E.4 Let

$$f_j(\alpha) \stackrel{\text{def}}{=} |\phi_j^\top \tilde{\mu}(\alpha)|^2,$$

If $f_j(\alpha)$ is L -Lipschitz, letting $g_{\max} = \max_{\alpha \in \mathbb{C}_\beta} f_j(\alpha)$ by Lemma E.5

$$\int_{\alpha \in \mathbb{C}_\beta} f_j(\alpha) d\alpha \geq \frac{\beta g_{\max}^3}{24L^2},$$

or equivalently,

$$\max_{\alpha \in \mathbb{C}_\beta} f_j(\alpha) \leq \left(\frac{24L^2}{\beta} \int_{\alpha \in \mathbb{C}_\beta} f_j(\alpha) d\alpha \right)^{1/3}.$$

Observe that

$$\begin{aligned} \int_{\alpha \in \mathbb{C}_\beta} f_j(\alpha) d\alpha &= \int_{\alpha \in \mathbb{C}_\beta} |\phi_j^\top \tilde{\mu}(\alpha)|^2 d\alpha \\ &= \phi_j^\top \left(\int_{\alpha \in \mathbb{C}_\beta} \tilde{\mu}(\alpha) \tilde{\mu}(\bar{\alpha})^\top d\alpha \right) \bar{\phi}_j \\ &= \sigma_j. \end{aligned}$$

Therefore conclude we have shown,

$$\max_{\alpha \in \mathbb{C}_\beta} |\tilde{\mu}(\alpha)^\top \phi_j| = \max_{\alpha \in \mathbb{C}_\beta} \sqrt{f_j(\alpha)} \leq \left(\frac{24L^2}{\beta} \sigma_j \right)^{1/6}. \quad (11)$$

The remainder of the proof consists of bounding the Lipschitz constant L and bounding the eigenvalue σ_j . To bound the Lipschitz constant of f_j ,

$$\begin{aligned} L &\leq \max_{\alpha \in S} |f_j'(\alpha)| \\ &= \max_{\alpha \in S} 2 |\text{Re}(\phi_j^\top \tilde{\mu}'(\alpha) \cdot \phi_j^\top \tilde{\mu}(\alpha))| \\ &\leq \max_{\alpha \in S} \|\phi_j\|_2^2 \cdot \|\tilde{\mu}(\alpha)\|_2 \cdot \|\tilde{\mu}'(\alpha)\|_2 \\ &= \max_{\alpha \in S} \|\tilde{\mu}(\alpha)\|_2 \cdot \|\tilde{\mu}'(\alpha)\|_2. \end{aligned}$$

Using Lemma A.1, we have $L \leq 12\beta^4 T^2$. By Lemma E.6, for any $j \in [T]$,

$$\sigma_j \leq 2\beta \log T.$$

We conclude,

$$\max_{\alpha \in \mathbb{C}_\beta} |\tilde{\mu}(\alpha)^\top \phi_j| \leq (24\beta^8 T^4 \log(T))^{1/6} < 2\log(T)\beta^{4/3} T^{2/3}.$$

We also bound

$$\begin{aligned} \sum_{j=1}^T |\tilde{\mu}(\alpha)^\top \phi_j| &\leq \sum_{j=1}^T \left(\frac{24L^2}{\beta} \sigma_j \right)^{1/6} \\ &= \left(\frac{24L^2}{\beta} \right)^{1/6} \sum_{j=1}^T \sigma_j^{1/6} \\ &= T \left(\frac{24L^2}{\beta} \right)^{1/6} \sum_{j=1}^T \frac{1}{T} \sigma_j^{1/6} \\ &\leq T \left(\frac{24L^2}{\beta} \right)^{1/6} \left(\sum_{j=1}^T \frac{1}{T} \sigma_j \right)^{1/6} \\ &\quad (\mathbb{E}[f(X)] \leq f(\mathbb{E}[X]) \text{ for concave } f(\cdot) \text{ and } f(x) = x^{1/6} \text{ is concave}) \\ &\leq T^{5/6} \left(\frac{24L^2}{\beta} \right)^{1/6} (\text{tr}(\mathbf{Z}))^{1/6} \\ &\leq T^{5/6} \left(\frac{24L^2}{\beta} \right)^{1/6} (6\beta \log(T))^{1/6} \\ &\quad (\text{tr}(\mathbf{Z}) \leq 6\beta \log(T) \text{ by proof of Lemma E.6}) \\ &\leq (24T^5 \cdot 12^2 \beta^8 T^4 \cdot 6 \log T)^{1/6} \quad (\text{Plugging in } L \leq 12\beta^4 T^2 \text{ from above.}) \\ &\leq 6\beta^{4/3} T^{3/2} \log(T). \end{aligned}$$

■

A Proofs of Technical Lemmas

Lemma A.1. For $\beta \ll 1$ and T large enough, the ℓ_2 norms are bounded by:

$$\|\tilde{\mu}(\alpha)\|_2 \leq \max\left(1, 4\beta^2\sqrt{T}\right), \quad \|\tilde{\mu}'(\alpha)\|_2 \leq \max\left(1, 3\beta^2T^{3/2}\right).$$

Proof of Lemma A.1 **1. Bound on $\|\tilde{\mu}(\alpha)\|_2$.** The squared magnitude of the entries depends on $|1 - \alpha^2|$. Using the triangle inequality and the small angle approximation $|\sin(x)| \leq |x|$:

$$|1 - \alpha^2| = |1 - r^2 e^{2i\theta}|$$

The magnitude is maximized when $r = 1$, pushing the value to the boundary of the unit circle.

$$|1 - e^{2i\theta}| = 2|\sin(\theta)| \leq 2|\theta| \leq 2\beta$$

Thus, we have the critical inequality for the kernel term:

$$|1 - \alpha^2| \leq 2\beta \quad (\text{for } r \approx 1)$$

The squared ℓ_2 norm is:

$$\|\tilde{\mu}(\alpha)\|_2^2 = \sum_{j=0}^{T-1} |\tilde{\mu}_j(\alpha)|^2 = |1 - \alpha^2|^4 \sum_{j=0}^{T-1} |\alpha|^{2j}$$

We analyze the maximum over the domain \mathcal{D} . The maximum of a modulus of a holomorphic function (multiplied by polynomial terms) generally occurs at the boundaries or critical points (here, $z = 0$).

Case A: The Origin ($r = 0$). At $\alpha = 0$, the zeroth term is $(1 - 0)^2 \cdot 1 = 1$. All other terms are 0, so $\|\tilde{\mu}(0)\|_2 = 1$.

Case B: The Boundary ($r = 1$). At $r = 1$, the geometric sum is simply $\sum_{j=0}^{T-1} 1 = T$. Using our preliminary bound $|1 - \alpha^2| \leq 2\beta$:

$$\|\tilde{\mu}(\alpha)\|_2^2 \leq (2\beta)^4 \cdot T = 16\beta^4 T.$$

Taking the square root, $\|\tilde{\mu}(\alpha)\|_2 \leq 4\beta^2\sqrt{T}$. Combining the cases we have,

$$\|\tilde{\mu}(\alpha)\|_2 \leq \max(1, 4\beta^2\sqrt{T})$$

2. Bound on $\|\tilde{\mu}'(\alpha)\|_2$. First, we compute the derivative using the product rule:

$$\begin{aligned} \tilde{\mu}'_j(\alpha) &= \frac{d}{d\alpha} [(1 - \alpha^2)^2 \alpha^j] \\ &= 2(1 - \alpha^2)(-2\alpha)\alpha^j + (1 - \alpha^2)^2(j\alpha^{j-1}) \\ &= \alpha^{j-1}(1 - \alpha^2) [-4\alpha^2 + j(1 - \alpha^2)] \end{aligned}$$

We seek to maximize the squared norm $\|\tilde{\mu}'(\alpha)\|_2^2 = \sum_{j=0}^{T-1} |\tilde{\mu}'_j(\alpha)|^2$.

Case A: The Origin ($r = 0$). For $j = 1$, the term is $\alpha^0(1 - 0)[-0 + 1(1)] = 1$. For $j \neq 1$, the term vanishes due to the α factor. Thus $\|\tilde{\mu}'(0)\|_2 = 1$.

Case B: The Boundary ($r = 1$). For large T , the sum is dominated by large j . For large j , the term $j(1 - \alpha^2)$ dominates the constant $-4\alpha^2$.

$$|\tilde{\mu}'_j(\alpha)| \approx |\alpha|^{j-1} |1 - \alpha^2| \cdot j |1 - \alpha^2| = j |1 - \alpha^2|^2$$

Summing the squares:

$$\|\tilde{\mu}'(\alpha)\|_2^2 \approx \sum_{j=0}^{T-1} (j |1 - \alpha^2|^2)^2 = |1 - \alpha^2|^4 \sum_{j=0}^{T-1} j^2$$

Using the summation formula $\sum_{j=0}^n j^2 \approx \frac{n^3}{3}$ and the bound $|1 - \alpha^2| \leq 2\beta$:

$$\|\tilde{\mu}'(\alpha)\|_2^2 \leq (2\beta)^4 \cdot \frac{T^3}{3} = 16\beta^4 \frac{T^3}{3}$$

Taking the square root:

$$\|\tilde{\mu}'(\alpha)\|_2 \leq \frac{4}{\sqrt{3}} \beta^2 T^{3/2}$$

■

B Chebyshev Polynomials Evaluated in the Complex Plane

In this section we let T_n denote the n -th Chebyshev polynomial and let M_n denote the monic form.

Proof of Lemma 3.1. We use that $M_n(z) = T_n(z)/2^{n-1}$ and

$$T_n(z) = \cos(n \arccos(z)). \quad (12)$$

If $|\operatorname{Im}(z)| \leq 1/64n^2$ then by Lemma B.2, $\arccos(z) \leq 1/n$. Therefore $n \arccos(z) \leq 1$ and so by Fact B.1

$$T_n(z) = \cos(n \arccos(z)) \leq 2 \quad (13)$$

Now we turn to the derivative $M'_n(z)$. It's a fact that

$$M'_n(z) = \frac{n}{2^{n-1}} U_{n-1}(z), \quad (14)$$

where U_{n-1} is the Chebyshev polynomial of the second kind. We next use the fact that

$$U_{n-1}(z) = \begin{cases} 2 \sum_{\substack{j \geq 0 \\ j \text{ even}}}^n T_j(z), & n \text{ even}, \\ 2 \sum_{\substack{j \geq 0 \\ j \text{ odd}}}^n T_j(z), & n \text{ odd}. \end{cases}$$

By Eq (13), $|T_j(z)| \leq 2$ for any j and therefore

$$|U_{n-1}(z)| \leq n. \quad (15)$$

Therefore,

$$|M'_n(z)| \leq \frac{n^2}{2^{n-1}}.$$

■

Fact B.1. Let $z \in \mathbb{C}$. Then $|\cos(z)| \leq 2$ whenever $|\operatorname{Im} z| \leq 1$.

Proof of Fact B.1

$$\begin{aligned} |\cos(x + iy)| &= (\cos^2 x \cosh^2 y + \sin^2 x \sinh^2 y)^{1/2} \quad (\text{Uses standard complex cosine identity.}) \\ &= (\cos^2 x + \cos^2 x (\cosh^2 y - 1) + \sin^2 x \sinh^2 y)^{1/2} \\ &= (\cos^2 x + \cos^2 x \sinh^2 y + \sin^2 x \sinh^2 y)^{1/2} \quad (\cosh^2 y - \sinh^2 y = 1) \\ &= (\cos^2 x + \sinh^2 y)^{1/2} \quad (\cos^2 x + \sin^2 x = 1) \\ &\leq (1 + \sinh^2 y)^{1/2} \quad (\sinh^2 y \leq 2 \text{ when } |y| \leq 1.) \\ &\leq 2. \end{aligned}$$

■

Lemma B.2. Let $z \in \mathbb{C}$ with $|z| \leq 1$. Then $|\operatorname{Im}(\arccos(z))| \leq 1/n$ whenever $|\arg(z)| \leq 1/64n^2$.

Proof of Lemma B.2 Let $re^{i\theta} = z$ and assume $|\theta| \leq 1/64n^2$. We use the Taylor series for $\arccos(\cdot)$,

$$\begin{aligned} \arccos(re^{i\theta}) &= \frac{\pi}{2} - \sum_{k=0}^{\infty} a_k (re^{i\theta})^{2k+1} \quad (\text{For } a_k = \frac{(2k)!}{4^k (k!)^2 (2k+1)}) \\ &= \frac{\pi}{2} - \sum_{k=0}^{\infty} a_k r^{2k+1} e^{i(2k+1)\theta} \quad (\text{De Moivre's Theorem}) \\ &= \frac{\pi}{2} - \sum_{k=0}^{\infty} a_k r^{2k+1} \cos((2k+1)\theta) - i \sum_{k=0}^{\infty} a_k r^{2k+1} \sin((2k+1)\theta). \\ &\quad (e^{i\theta} = \cos \theta + i \sin \theta) \end{aligned}$$

Therefore,

$$\operatorname{Im}(\arccos(re^{i\theta})) = \sum_{k=0}^{\infty} a_k r^{2k+1} \sin((2k+1)\theta).$$

Then

$$\begin{aligned} |\operatorname{Im}(\arccos(re^{i\theta}))| &\leq \sum_{k=0}^{\infty} a_k |r|^{2k+1} |\sin((2k+1)\theta)| \\ &\leq \sum_{k=0}^{\infty} a_k |r|^{2k+1} \min(1, (2k+1)|\theta|) \quad (|\sin(x)| \leq \min(|x|, 1)) \\ &\leq \sum_{k=0}^{\infty} a_k \min(1, (2k+1)|\arg(z)|) \quad (|r| \leq 1) \\ &\leq \sum_{k=0}^K a_k (2k+1)|\arg(z)| + \sum_{k=K+1}^{\infty} a_k. \quad (\text{For any arbitrary } K \geq 0) \end{aligned}$$

Now we bound a_k .

$$\begin{aligned} a_k(2k+1) &= \frac{(2k)!}{4^k (k!)^2} \\ &\leq \frac{\sqrt{2\pi(2k)}(2k/e)^{2k} (1 + \frac{1}{2k})}{4^k \left(\sqrt{2\pi k} (k/e)^k \right)^2} \quad (\text{Stirling's Formula}) \\ &= \left(1 + \frac{1}{2k}\right) / \sqrt{\pi k} \\ &\leq 1/\sqrt{k}. \end{aligned}$$

Therefore we also have that $a_k \leq 1/k^{3/2}$. Using this (and noting that $a_0 = 1$) we see,

$$\begin{aligned} |\operatorname{Im}(\arccos(z))| &\leq |\arg(z)| \left(1 + \sum_{k=1}^K \frac{1}{\sqrt{k}}\right) + \sum_{k=K}^{\infty} \frac{1}{k^{3/2}} \\ &\leq 4 \left(|\arg(z)| \sqrt{K} + \frac{1}{\sqrt{K}} \right) \\ &\leq \frac{8}{\sqrt{K}} \quad (\text{For } |\arg(z)| \leq 1/K) \\ &\leq \frac{1}{n} \quad (\text{For } K \geq 64n^2.) \end{aligned}$$

Therefore, for $|\arg(z)| \leq 1/64n^2$, we have that

$$|\operatorname{Im}(\arccos(z))| \leq 1/n.$$

■

Proof of Lemma 3.2 We bound the coefficients of the Chebyshev polynomial. From Chapter 22 of [3],

$$T_n(x) = \frac{n}{2} \sum_{m=0}^{\lfloor n/2 \rfloor} (-1)^m \frac{(n-m-1)!}{m!(n-2m)!} (2x)^{n-2m}. \quad (16)$$

Therefore

$$M_n(x) = \frac{1}{2^{n-1}} T_n(x) = n \sum_{m=0}^{\lfloor n/2 \rfloor} (-1)^m \frac{(n-m-1)!}{m!(n-2m)!} 2^{-2m} x^{n-2m}.$$

Let $c_m = \frac{(n-m-1)!}{m!(n-2m)!} 2^{-2m}$. Then

$$\begin{aligned}
\max_{m=0,\dots,n} c_m &\leq \max_{m=0,\dots,n} \binom{n-m}{m} 4^{-m} \\
&\leq \max_{m=0,\dots,n} \left(\frac{(n-m)e}{4m} \right)^m && \left(\binom{n}{k} \leq (ne/k)^k \right) \\
&\leq \max_{c \in [0,1]} \left(\frac{(1-c)e}{4c} \right)^{cn} && \text{(Letting } m = cn) \\
&\leq 2^{0.3n}. && (\max_{c \in [0,1]} ((1-c)e/4c)^c \leq 2^{0.3})
\end{aligned}$$

■

C Plots from Experiments

The details of the experiments are in Section [4](#), please refer to that for specifics.

C.1 Experiments with Linear Dynamical System Data

As our theory shows, for Chebyshev polynomials (the same can be shown for Legendre polynomials), the coefficients of the polynomial grow exponentially and therefore the performance gains vanish after the degree is too high. However, learning the optimal coefficients is able to sidestep this issue.

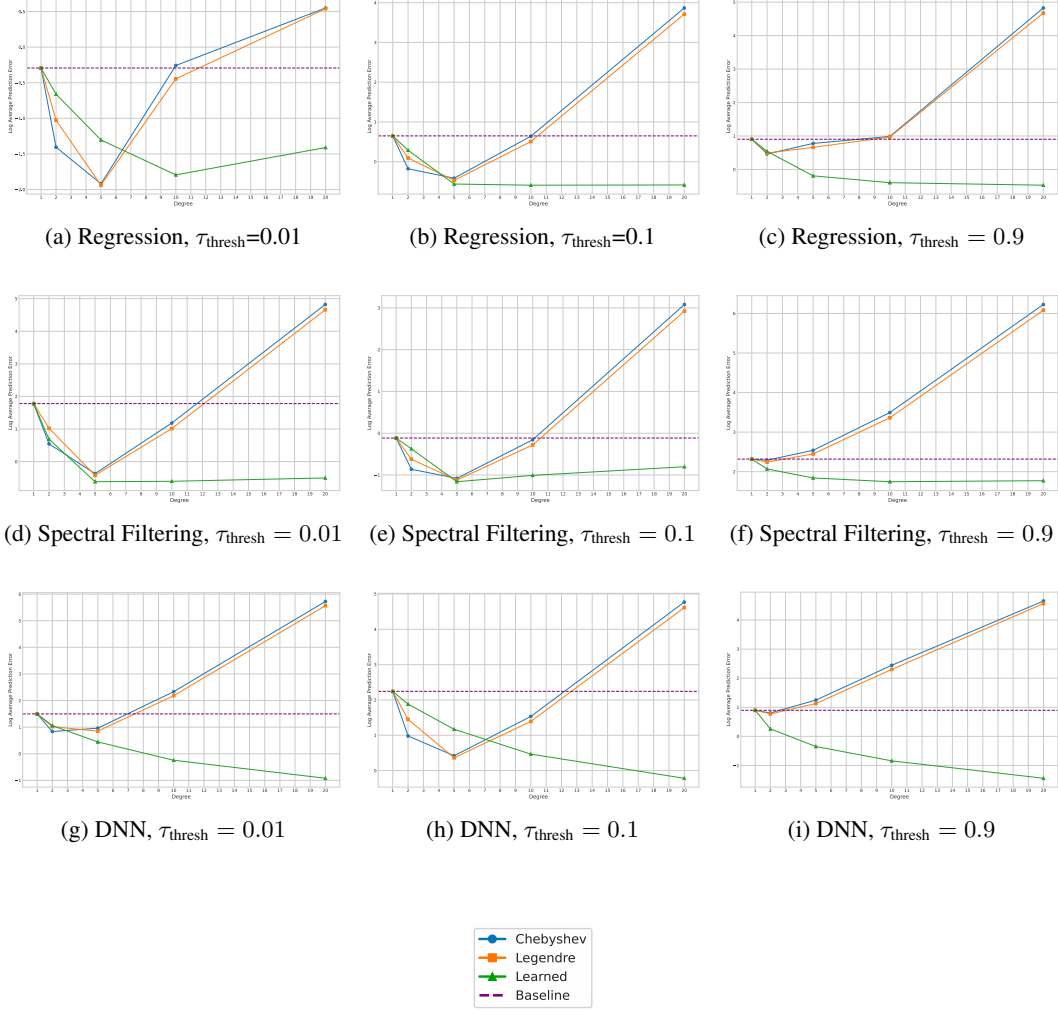


Figure 3: Absolute prediction error averaged over 200 independent runs with data generated from a linear dynamical system with varying complex threshold.

C.2 Nonlinear Data

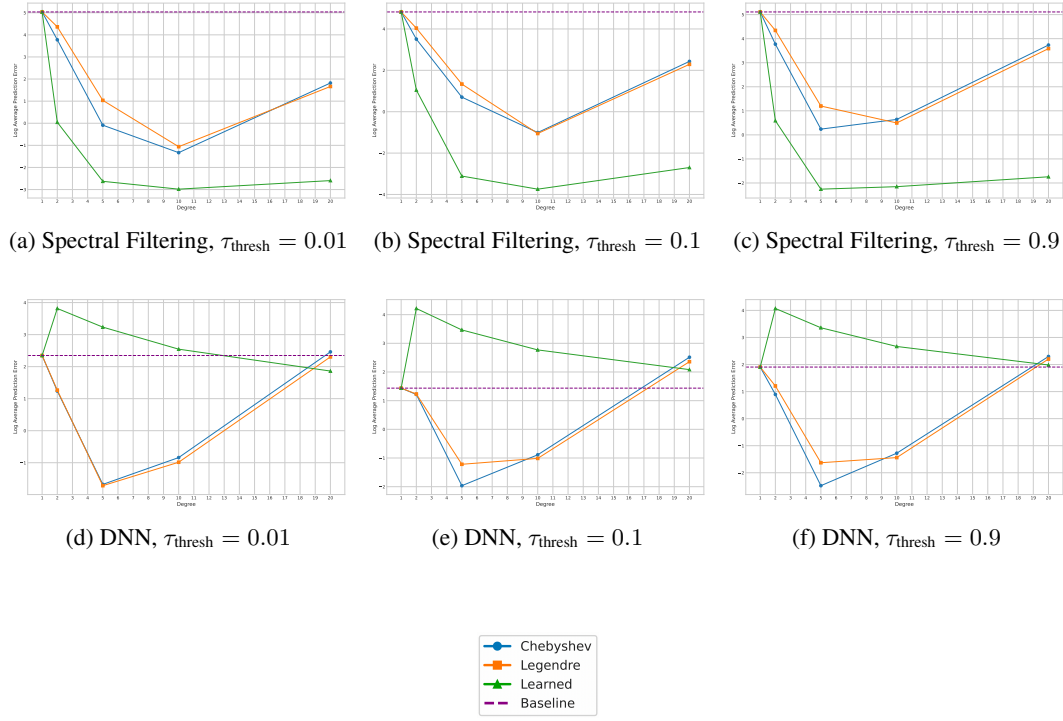


Figure 4: Absolute prediction error averaged over 200 independent runs with data generated from a nonlinear dynamical system with varying complex threshold.

C.3 Data from a DNN

Finally we generate data from a 10-layer sparse neural network which stacks 100-dimensional LSTMs with ReLU nonlinear activations.

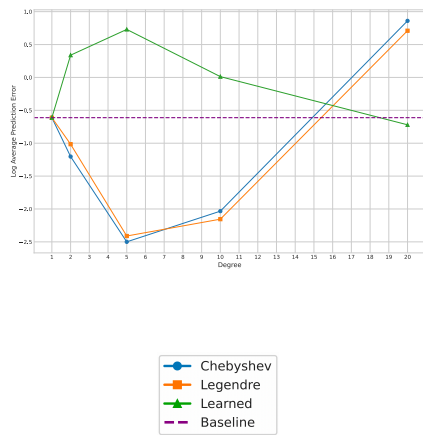


Figure 5: Absolute prediction error of a 10-layer DNN model averaged over 200 independent runs.