Provable Length Generalization in Sequence Prediction via Spectral Filtering

Annie Marsden [*1] **Evan Dogariu** [*1] **Naman Agarwal** [1] **Xinyi Chen** [1] **Daniel Suo** [1] **Elad Hazan** [1 2]

Abstract

We consider the problem of length generalization in sequence prediction. We define a new metric of performance in this setting – the Asymmetric-Regret– which measures regret against a benchmark predictor with longer context length than available to the learner. We continue by studying this concept through the lens of the spectral filtering algorithm. We present a gradient-based learning algorithm that provably achieves length generalization for linear dynamical systems. We conclude with proof-of-concept experiments which are consistent with our theory.

1. Introduction

Sequence prediction is a fundamental problem in machine learning, with applications that span natural language processing (NLP), forecasting, and control systems. A common challenge in these tasks is to determine how much past information, known as context length, should be used to make accurate predictions. Although longer contexts often improve performance, they come with increased computational and memory costs, making it impractical to store and process entire sequences, especially during training.

This challenge raises a key question: Can we design learning algorithms that effectively operate with limited context during training while still generalizing to longer sequences at test time? This property, which we call length generalization, is particularly relevant in large-scale models, such as large language models (LLMs), which struggle to extrapolate beyond the context lengths seen during training. Despite extensive empirical research on this issue, formal theoretical guarantees on length generalization remain elusive.

To address this, we introduce a novel performance metric, Asymmetric-Regret, which quantifies the regret of a predictor with a limited context length compared to an ideal predictor with a longer context. Unlike standard regret, which assumes both the learner and benchmark operate under the same conditions, Asymmetric-Regret explicitly accounts for the context length discrepancy, offering a more realistic evaluation of length generalization in resource-constrained settings.

We study length generalization through the lens of spectral filtering algorithms, a class of methods known for their effectiveness in learning linear dynamical systems (LDS) with long memory (Hazan et al., 2017a). Spectral filtering methods have recently been used to develop state space models that achieve state-of-the-art (SOTA) performance in LLMs, improving both efficiency and scalability (Gu et al., 2021b; Poli et al., 2023; Gu & Dao, 2023). In this work, we prove that spectral filtering algorithms can generalize across context lengths while maintaining strong theoretical guarantees on regret. This is the first result, to the authors' knowledge, providing such provable guarantees in the setting of learning linear dynamical systems.

1.1. Our Contributions

Consider **online sequence prediction** in which the predictor iteratively receives input $u_t \in \mathcal{R}^{d_{\text{in}}}$ and then makes a prediction $\hat{y}_t \in \mathcal{R}^{d_{\text{out}}}$ of the output, after which the true output y_t is revealed. The goal of the predictor is to minimize error according to a given convex and Lipschitz loss function $\ell_t(y_t, \hat{y}_t)$. In this work we consider the class of *spectral filtering* predictors, introduced by Hazan et al. (2017b). A spectral filtering predictor is characterized by parameters $(T, M_i{}_{i=1}^{k}, k)$ and outputs predictions $\hat{y}_t$ of the form

$$\hat{y}_t = y_{t-1} + \sum_{i=1}^{k} M_i u_{(t-1):0} \phi_i,$$

where $u_{(t-1):0} \in \mathbb{R}^{d_{\text{in}} \times T}$ is a matrix whose columns are the previous inputs $u_{t-1}, u_{t-2}, \ldots, u_0$ (possibly zero-padded as necessary), $\{\phi_j\}_{j=1}^{k}$ are the T-dimensional spectral filters, $\{M_i\}_{i=1}^{k} \subset \mathcal{R}^{d_{\text{out}} \times d_{\text{in}}}$ are matrices which are learned online, and k is the number of filters used. Hazan et al. (2017b) provide an algorithm to learn $\{M_i\}_{i=1}^{k}$ and show this achieves nearly optimal regret bounds when measured against the best Linear Dynamical System (LDS) predictor. We explore whether the full history $u_{(t-1):0}$ is needed to learn $\{M_i\}_{i=1}^{k}$. More broadly, we explore whether predic-

[*]Equal contribution [1]Google DeepMind [2]Princeton University. Correspondence to: Annie Marsden <anniemarsden@google.com>, Evan Dogariu <dogariu@google.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

tor classes and corresponding online learning algorithms exist that can achieve context length generalization—that is, they use only a short recent history during learning but perform nearly as well as if they had used the full, much longer history length. Of course, predictors which perform poorly on systems that require long memory can trivially achieve context length generalization if their performance is poor regardless of the context length used. Notably, spectral filtering predictors excel in systems with long memory (Hazan et al., 2017b).

To properly understand context length generalization, we introduce the notion of *Asymmetric-Regret*. The idea is to consider the regret of learning a predictor from a class which is only allowed to use context length L' against the best predictor which is allowed to use (potentially much longer and therefore asymmetric) context length L. Let Π_L denote the class of predictors in Π which use context length L. Given an algorithm $\mathcal{A}(L')$ which learns over predictors from some class $\Pi_{L'}$, the Asymmetric-Regret over horizon T is

$$\text{Regret}_{\text{Asym},T}\left(\mathcal{A}(L'), \Pi_L\right) \stackrel{\text{def}}{=} \sum_{t=1}^{T} \ell_t(y_t, \hat{y}_t^{\mathcal{A}(L')}) - \min_{\pi \in \Pi_L} \ell_t(y_t, \hat{y}_t^{\pi}).$$

Our first result shows that spectral filtering generalizes from a history of T^q, where $q \in [0,1]$, to T for certain linear dynamical systems. It is formally given in the following theorem.

Theorem 1. *Let $T \in \mathbb{Z}_{\geq 0}$ and $q \in [0,1]$. Consider a sequence $(y_1, \ldots, y_T)$ generated by an unknown and noiseless linear dynamical system defined by matrices (A, B, C, D) as per Eq. 1. Assume the input sequence $u_{0:(t-1)}$ is sufficiently well-conditioned, satisfying $\sum_{t=0}^{T-1}(T-t)u_t u_t^\top \succeq \left(\frac{2|C||B|}{\sqrt{T}}\right) I$. Suppose the eigenvalues of A lie within the range $\left[0, 1 - \frac{\log(T)}{8T^q}\right] \cup \left[1 - \frac{1}{2T^{5/4}}, 1\right]$.*

Let $\mathcal{A}(L)$ denote Algorithm 1 operating with context length L, and let Π_L^{SF} denote the class of spectral filtering predictors using context length L. For the squared loss $\ell_t(y, y') = |y - y'|^2$ and sufficiently large T, it holds that:

$$\textit{Regret}_{\text{Asym},T}\left(\mathcal{A}(T^q), \Pi_T^{\text{SF}}\right) \leq \tilde{O}(\sqrt{T}).$$

This theorem indicates that for any $q \in [0,1]$, the Asymmetric-Regret is bounded by $\tilde{O}(\sqrt{T})$. However, as q decreases, the class of linear dynamical systems for which this bound holds becomes more restricted due to the eigenvalue conditions on A. The spectrum of A determines the memory of the system; when the eigenvalues of A are 1, the system is only marginally-stable and standard predictors which aim to use low memory typically fail. Critically, Theorem 1 holds even for these marginally-stable systems. When interpreting this result, it's important to note that the class of spectral filtering predictors Π_T^{SF} which use the full context length are provably able to predict well on marginally-stable Linear Dynamical Systems (Hazan et al., 2017b)[1]. Therefore, this result implies that spectral filtering predictors are able to context length generalize in a nontrivial way.

Inspired by the way in which Theorem 1 is sensitive to the spectrum of A, we develop a novel variation on the Spectral Filtering algorithm, presented in Algorithm 2, which achieves robust length generalization without added assumptions on the spectrum of A (whenever the context-length is at least $T^{1/4}$). Algorithm 2 achieves this by using two autoregressive components y_{t-1} and y_{t-2} to construct its prediction $\hat{y}_t$ of y_t. We provide our main theorem of this work.

Theorem 2. *Let $T \in \mathbb{Z}_{\geq 0}$ and $q \in \left[\frac{1}{4} + \frac{\log(\log(T)/8)}{\log(T)}, 1\right]$. Consider a sequence $(y_1, \ldots, y_T)$ generated by an unknown and noiseless linear dynamical system defined by matrices (A, B, C, D) as per Eq. 1. Assume the input sequence $u_{0:(t-1)}$ is sufficiently well-conditioned, satisfying $\sum_{t=0}^{T-1}(T-t)u_t u_t^\top \succeq \left(\frac{2|C||B|}{\sqrt{T}}\right) I$. Let $\mathcal{A}(L)$ denote Algorithm 2 operating with context length L, and let Π_L^{SF} denote the class of spectral filtering predictors using context length L. For the squared loss $\ell_t(y, y') = |y - y'|^2$ and sufficiently large T, it holds that:*

$$\textit{Regret}_{\text{Asym},T}\left(\mathcal{A}(T^q), \Pi_T^{\text{SF}}\right) \leq \tilde{O}(\sqrt{T}).$$

Finally, we experimentally confirm the results of Theorem 1 and Theorem 2 on synthetic data generated by an LDS. Interestingly, we find that Theorem 1 accurately predicts when length generalization is possible; indeed, when the data is generated by an LDS which has eigenvalues in the "bad" range $[1 - \log(T)/(8T^q), 1 - 1/(2T^{5/4})]$ we find that the limited context length spectral filtering predictors are unable to length generalize. However, when the data is generated by an LDS which has eigenvalues "hugging" this bad range (i.e. either just smaller than $1 - \log(T)/(8T^q)$ or just larger than $1 - 1/(2T^{5/4})$), the limited context length spectral filtering predictors successfully length generalize, demonstrating the sharpness of our analysis. Next, we see that adding the second autoregressive term allows for robust length generalization on marginally-stable systems with no

[1]The only LDS's for which there can be any useful results are those with A's eigenvalues in $[-1, 1]$, i.e. marginally-stable systems. We recall that the spectral filtering principle can be readily applied to handle negative eigenvalues in $[-1, 0]$ (see Appendix D of (Agarwal et al., 2023), for example). For ease of presentation, we focus on capturing the length generalization effects of eigenvalues in $[0, 1]$ in the sequel, and so we suppose without loss of generality that $A \succeq 0$.

spectral assumption. Lastly, we conduct experiments using the STU neural architecture to test the hypothesis that this architecture should simply length generalize without any task-specific engineering. We consider the induction heads synthetic task and find that the out-of-the-box STU neural architecture does indeed enjoy some level of length generalization. This suggests that incorporating spectral filtering into neural architectures, like the STU, may provide improved length generalization in deep learning applications. We leave further empirical study on this for future work.

1.2. Related Work

The literature for sequence prediction is too broad to survey in detail, so we give a few highlights of the recent rapid advancements. The most notable progress includes the Transformer model (Vaswani et al., 2017) that incorporates an attention mechanism for accurate sequence prediction in many domains (Brown et al., 2020; Dosovitskiy et al., 2020; Jumper et al., 2021). Transformer models and their attention layers have memory/computation requirements that scale quadratically with context length. Many approximations have been proposed (see (Tay et al., 2022) for a recent survey).

Motivated by the high memory and compute requirements of transformers, state space models were revisited starting from (Gu et al., 2020; 2021b) who propose and develop the HiPPO theory. Gu et al. (2021a) develop the S4 parameterization to address the bottlenecks of training efficiency, performance and numerical stability. Further works in the area show SOTA performance and include (Gupta et al., 2022; Smith et al., 2023; Orvieto et al., 2023; Gu & Dao, 2023). State space models are very efficient for training and inference, but can suffer in long-context applications. This motivated the use of spectral filtering technique for learning marginally-stable linear dynamical systems (Hazan et al., 2017b; 2018). This technique was incorporated to a neural architecture in (Agarwal et al., 2023), that was recently shown to perform well across several modalities (Liu et al., 2024).

From an applied perspective, generalization in sequence prediction has been studied in (Hou et al., 2024) through the theoretical lens of Turing programs. They propose a methodology that empirically improves length generalization across a diverse set of tasks. There are also many architecture-specific approaches to improving length generalization, usually in the context of language models, such as ALiBi positional embeddings for transformers (Press et al., 2022), recursive application of deep networks (Schwarzschild et al., 2021; Bansal et al., 2022), a self-attentive recurrent sequence model called the "Universal Transformer" (Dehghani et al., 2018), and the Neural GPU (Kaiser & Sutskever, 2015). However, such methods lack provable guarantees and can have varying empirical performance (Kazemnejad et al., 2024). There have also been extensive investigations into the ability of neural networks to length generalize and the role that various features play in this, for example see (Anil et al., 2022; Murray & Chiang, 2018; Yehudai et al., 2021; Zhang et al., 2022; Newman et al., 2020).

In contrast, our investigation starts from the theory of regret minimization in games and online learning. Regret minimization has the advantage that it implies generalization in the statistical learning setting, see e.g. (Cesa-Bianchi et al., 2004) and is usually accompanied by efficient algorithms such as online gradient descent, see e.g. (Hazan et al., 2016). Our new notion of Asymmetric-Regret incorporates asymmetric information access between the online learner and the benchmark class.

2. Background and Setting

In the **online sequence prediction** setting the predictor iteratively receives input u_t and makes prediction $\hat{y}_t$ of the output, after which the true output y_t is revealed. The goal is to minimize error according to a given (convex Lipschitz) loss function $\ell_t(y_t, \hat{y}_t)$.

In online learning, we usually do not make statistical assumptions about the generation of the input sequence. As such, performance is measured relative to a certain benchmark class of predictors. A prediction algorithm $\mathcal{A}$ is measured by regret, or difference in total loss, vs. a class of reference predictors Π^{ref} (such as linear predictors), i.e.

$$\text{Regret}_T(\mathcal{A}, \Pi) = \sum_{t=1}^{T} \ell_t(y_t, \hat{y}_t^{\mathcal{A}}) - \min_{\pi \in \Pi} \sum_{t=1}^{T} \ell_t(y_t, \hat{y}_t^{\pi}).$$

This formulation is valid for online sequence prediction of any signal. We are particularly interested in signals that are generated by dynamical systems. A time-invariant linear dynamical system is given by the dynamics equations

$$x_{t+1} = Ax_t + Bu_t + w_t \ , \ \ y_{t+1} = Cx_t + Du_t + \zeta_t, \quad (1)$$

where x_t is the (hidden) state, u_t is the input or control to the system, and y_t is the observation. The terms w_t, ζ_t are noise terms, and the matrices A, B, C, D are called the system matrices.

Many methods exist for linear dynamical systems and their performance guarantees rely heavily on the spectrum of A. The system is *unstable* whenever $|\lambda_{\max}(A)| > 1$ because the norm of the observations tends towards infinity, *stable* when $|\lambda_{\max}(A)| < 1$ and *marginally-stable* if $|\lambda_{\max}(A)| = 1$. In the stable setting, if $|\lambda_{\max}(A)| = 1 - \delta < 1$, typical methods (i.e., Kalman filtering) must use a history of at least $\gg \frac{1}{\delta}$ previous states to accurately capture the dynamics. As δ gets

smaller (i.e. long memory), it therefore becomes difficult for methods to directly learn these relationships. Methods that learn system matrices require knowledge of the dimension of the hidden state (which may be very large) and can also be unstable for systems with long memory. Through a particular parameterization and convex relaxation, however, the spectral filtering algorithm is able to efficiently predict observations from marginally-stable systems with sublinear regret. We provide more background on spectral filtering in Section 2.2, and more details on the rich theory of linear dynamical systems may be found in (Hazan et al., 2020).

2.1. Context Length Generalization and the Asymmetric-Regret metric

We say that an online predictor has context length L if it bases its prediction $\hat{y}_t$ only on information from the previous L timesteps, i.e. $u_{t:t-L}$ and $y_{t:t-L}$. The key question in our work is whether there are algorithms which learn and predict using a short context length, but perform as well as had they been allowed to use long context length. To formalize this notion, we introduce Asymmetric-Regret whose definition we restate here:

Definition 3 (Asymmetric-Regret)**.** Let $\Pi^{\text{learn}}_{L'}$ be a class of predictors which use context length L' and let Π^{ref}_{L} be a reference class of predictors which use context length L. The *Asymmetric-Regret* with respect to (convex Lipschitz) loss ℓ_t over horizon T of an algorithm $\mathcal{A}(L')$ which tries to learn a predictor from $\Pi^{\text{learn}}_{L'}$ is

$$\text{Regret}_{\text{Asym},T}\left(\mathcal{A}(L'), \Pi^{\text{ref}}_{L}\right) \stackrel{\text{def}}{=} \sum_{t=1}^{T} \ell_t(y_t, \hat{y}_t^{\mathcal{A}(L')}) - \min_{\pi \in \Pi_L} \sum_{t=1}^{T} \ell_t(y_t, \hat{y}_t^{\pi}).$$

To gain a better understanding of Asymmetric-Regret, note that the typical notion of regret in sequence prediction sets $L' = T$ for the given class of predictors and sets $L = T$ for the given reference class of predictors Π^{ref} by default. In this case Asymmetric-Regret recovers typical regret,

$$\text{Regret}\left(\mathcal{A}, \Pi^{\text{ref}}\right) = \text{Regret}_{\text{Asym},T}\left(\mathcal{A}(T), \Pi^{\text{ref}}_{T}\right).$$

However, if $L' < T$, any upper bound on $\text{Regret}_{\text{Asym},T}\left(\mathcal{A}(L'), \Pi^{\text{ref}}_{T}\right)$ immediately implies an upper bound on $\text{Regret}\left(\mathcal{A}, \Pi^{\text{ref}}\right)$ since the algorithm $\mathcal{A}(T)$ can choose to only use context length L' and ignore the rest. Therefore, Asymmetric-Regret is a stronger notion than typically used. Another possible notion of regret that aligns more with the literature of length generalization for language models is to restrict the algorithm $\mathcal{A}(L)$ to update itself using context length L' but to allow it to use full context length when making predictions. Note that any bound on the Asymmetric-Regret would immediately apply to this alternative notion of regret since the algorithm is only given more power. That said, this alternative notion of regret could provide a different landscape of results. For instance, there may be signals (like a linear dynamical system with asymmetric transition matrix) that can still be learned with short context length but require the full history to be predicted accurately.

2.2. Spectral Filtering

Spectral filtering is a notable deviation from the standard theory of linear dynamical systems that allows efficient learning in the presence of arbitrarily long memory (Hazan et al., 2017b). The idea is to project the sequence of inputs to a small subspace that is constructed using the special structure of discrete linear dynamical systems. The output of the spectral filtering predictor is represented as

$$\hat{y}_t = y_{t-1} + \sum_{i=1}^{k} M_i u_{(t-1):0} \phi_i, \tag{2}$$

where $u_{(t-1):0} \in \mathbb{R}^{d_{\text{in}} \times T}$ is a matrix whose columns are the previous inputs $u_{t-1}, \ldots, u_0$ (possibly zero-padded as necessary), $\{\phi_j\}_{j=1}^{k}$ are the T-dimensional spectral filters that can be computed offline given the target sequence length T, and $\{M_i\}_{i=1}^{k} \subset \mathcal{R}^{d_{\text{out}} \times d_{\text{in}}}$ are the matrices parameterizing the model. These spectral filters are the eigenvectors of the matrix constructed as the average of outer products of the discrete impulse-response functions as we now detail.

Let $\mu_{\alpha,T} = (1-\alpha)[1, \alpha, \alpha^2, ..., \alpha^T]$ be the (weighted) impulse-response vector corresponding to a one dimensional linear dynamical system with parameter α unfolded to T time steps, and consider the symmetric matrix

$$H_T \stackrel{\text{def}}{=} \int_0^1 \mu_{\alpha,T} \mu_{\alpha,T}^{\top} d\alpha. \tag{3}$$

Since H_T is a real PSD matrix, it admits a real spectral decomposition, and the (non-negative) eigenvalues can be ordered naturally by their value. Let $\{(\sigma_j \in \mathbb{R}, \phi_j \in \mathbb{R}^L)\}_{j=1}^{L}$ be the eigenvalue-eigenvector pairs of H_T ordered to satisfy $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d$. The spectral filters $\phi_1, ..., \phi_k$ are exactly those first k eigenvectors corresponding to the largest eigenvalues. The spectral filtering class is further parameterized by matrices $M_1, ..., M_k \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. The output at time t is then given by equation (2).

The following theorem establishes that the spectral filtering class of predictors approximately contains bounded linear dynamical systems with positive semi-definite A. The exact constants are left out for simplicity of presentation, but appear in the original work.

Theorem 4 (Simplified from (Hazan et al., 2017a))**.** *Given any linear dynamical system parametrized by A, B, C, D*

such that A is a PSD matrix with $\|A\| \leq 1$, there exists matrices $M_1, ..., M_k$, such that for all T and all sequences $u_{1:T}, \|u_t\| \leq 1$, the following holds. Let $y^{\text{LDS}}_{1:T}$ be the sequence generated by execution of the LDS via (1) and $y^{\text{SF}}_{1:T}$ be the sequence generated by Spectral Filtering via (2). Then for all $t \in [T]$,

$$\|y^{\text{LDS}}_t - y^{\text{SF}}_t\| \sim e^{-\frac{k}{\log(L)}}.$$

Theorem 4 establishes that Spectral Filtering can predict long memory sequences since the statements holds even over marginally stable linear dynamical systems.

3. Learning with a Short Context—Provable Length Generalization for Linear Dynamical Systems

In Algorithm 1, we modify the classical online learning algorithm for spectral filtering to use a shorter context window. To properly define our notion of length generalization, we need to distinguish between context lengths. Thus we introduce the notation for the loss observed with a context length L: letting $\hat{y}(M, L)$ denotes the prediction of y_t using $M = [M_1, \ldots, M_k]$ and context window size L as in Eq. 4 of Algorithm 1 we have

$$\ell_t(M, L) \stackrel{\text{def}}{=} \|\hat{y}(M, L) - y_t\|^2.$$

Note that this is overloaded notation compared with $\ell_t(y, y')$ which measures the loss of the true y with the predicted y' as used in our definition of regret. To provide a precise statement on length generalization, we present the following performance guarantee. Note that we prove the following for a (A, B, C, I)-LDS rather than (A, B, C, D) which is without loss of generality since we can consider the input as $Du_1, \ldots, Du_T$.

Algorithm 1 Spectral Filtering with Limited Context

1: **Input:** $k > 0, T > 0, L > 0, r > 0$. Initialize $M^1_i \in \mathcal{R}^{d_{\text{out}} \times d_{\text{in}}}$ for $i \in [k]$ and set $M^1 = [M^1_1, \ldots, M^1_k]$. Let $\phi_{1:k}$ be the largest eigenvectors of H_T defined in Eq. 3 with corresponding eigenvalues $\sigma_{1:k}$, and let $\pi_{\mathcal{K}}(\cdot)$ denote the projection to convex set $\mathcal{K}$.

2: **for** $t = 1, 2, ..., T$ **do**

3: Compute and predict

$$\hat{y}_t = y_{t-1} + \sum_{i=1}^{k} M^t_i u_{(t-1):(t-L)} (\sigma_i^{1/4} \phi_i). \quad (4)$$

4: Observe y_t, denote $\ell_t(M^t, L) = \|\hat{y}_t - y_t\|^2$ and update and project onto the low Frobenius norm ball

$$\hat{M}^{t+1} \leftarrow M^t - \eta_t \nabla_M \ell_t(M^t)$$

$$M^{t+1} = \pi_{\mathcal{K}} \left(\hat{M}^{t+1} \right),$$

where $\mathcal{K}_r = \left\{ M \in \mathbb{R}^{k \times d_{\text{out}} \times d_{\text{in}}} : \|M_i\| \leq r, \forall i \in [k] \right\}$.

5: **end for**

Theorem 5. *Let $T \in \mathbb{Z}_{\geq 0}$ and $q \in [0, 1]$. Consider a sequence $(y_1, \ldots, y_T)$ generated by an unknown and noiseless linear dynamical system defined by matrices (A, B, C, I) as per Eq. 1. Assume the input sequence $u_{0:(t-1)}$ is sufficiently well-conditioned, satisfying $\sum_{t=0}^{T-1} (T - t) u_t u_t^\top \succeq \left(\frac{2|C||B|}{\sqrt{T}} \right) I$. Suppose the eigenvalues of A lie within the range $\left[0, 1 - \frac{\log(T)}{8T^q}\right] \cup \left[1 - \frac{1}{2T^{5/4}}, 1\right]$. Let $k = \Omega\left(\log(T) \cdot \log\left(T d_A\right)\right)$, $r \geq \|B\|\|C\|$, and assume $T \geq (4k \log(T)/\|C\|\|B\|)^4$. Algorithm 1 satisfies:*

$$\begin{aligned} &\textit{Regret}_{\text{Asym},T} \left(\mathcal{A}(T^q), \Pi^{\text{SF}}_T \right) \\ &\quad \leq O\left(\|B\|^2 \|C\|^2 k^{3/2} \log(T) \sqrt{T} \right). \end{aligned}$$

The proof of Theorem 5 is in Appendix B with a high-level overview at the end of this section. This theorem shows that the sequence $M^1, \ldots, M^T$ constructed by Algorithm 1, even when using a reduced context length of size T^q, is able to achieve regret $O(\sqrt{T})$ when compared to the best spectral filter that uses full context length T. To gain better understanding of the needed assumption on the spectrum of A, first suppose that all the eigenvalues of A are bounded by $1 - \delta$. Then the extent to which the input u_{t-t_0} affects the value of y_t is roughly $(1 - \delta)^{t_0}$, since the hidden state is multiplied by A t_0 many times. This becomes negligible when t_0 is much larger than $1/\delta$ and implies that u_{t-t_0} may be forgotten. This intuition explains why length generalization is possible for the first region of eigenvalues $[0, 1 - \log(T)/(8T^q)]$. Indeed, letting $\delta = \log(T)/8T^q$ and $t_0 = T^q$ (which is much bigger than $8T^q/\log(T)$ for large enough T) we see that when the spectrum of A is smaller than $1 - \delta$, after t_0 many steps we can forget about the previous inputs u_{t-t_0}. The second part of the range – i.e. that the spectrum of A can lie between $[1 - 1/(2T^{5/4}), 1]$– is a special feature of spectral filtering's ability to efficiently capture long memory effects and is rather technical. The "bad region" is exactly the range where the eigenvalues aren't small enough that u_{t-t_0} can be forgotten for $t_0 \geq T^q$, but also aren't large enough that spectral filtering is naturally able to capture them. Numerically, the range is very small for large T and reasonable q.

Motivated by the limitations of Theorem 5, in order to provide a length generalization that is robust to the spectrum of A, we introduce a variation on the classical Spectral Filtering algorithm, presented as Algorithm 2. This algorithm uses the two most previous outputs y_{t-1} and y_{t-2} when making a prediction $\hat{y}_t$ of y_t.

This algorithm has a slightly different construction of spectral filters. Indeed, they are the eigenvectors of the following matrix

$$N_T \stackrel{\text{def}}{=} \int_0^1 \tilde{\mu}_{\alpha,T}\tilde{\mu}_{\alpha,T}^\top d\alpha, \tag{5}$$

where $\tilde{\mu}_{\alpha,T} \stackrel{\text{def}}{=} (1-\alpha)^2[1,\alpha,\alpha^2,\ldots,\alpha^T]$. Interestingly, just by using one extra autoregressive term, our adapted algorithm is able to enjoy *robust* length generalization in the sense that whenever the context window is at least $T^{1/4+\epsilon}$ then no extra assumptions on the spectrum of A are necessary to achieve our notion of length generalization. We state this formally in the following theorem.

Algorithm 2 Spectral Filtering with Limited Context and Two Autogressive Components

1: **Input:** $k > 0, T > 0, L > 0, r > 0$. Initialize $M_i^1 \in \mathcal{R}^{d_{\text{out}} \times d_{\text{in}}}$ for $i \in [k]$ and set $M^1 = [M_1^1, \ldots, M_k^1]$. Let $\tilde{\phi}_{1:k}$ be the largest eigenvectors of N_{T-2} defined in Eq. 5 with corresponding eigenvalues $\tilde{\sigma}_{1:k}$, and let $\pi_{\mathcal{K}}(\cdot)$ denote the projection to convex set $\mathcal{K}$.

2: **for** $t = 1, 2, ..., T$ **do**

3: Compute and predict

$$\hat{y}_t = 2y_{t-1} - y_{t-2} + M_1^t u_{t-1} + M_2^t u_{t-2} + \sum_{i=3}^{k} M_i^t u_{(t-3):(t-L)}(\tilde{\sigma}_i^{1/4}\tilde{\phi}_i).$$

4: Observe y_t, denote $\ell_t(M^t, L) = \|\hat{y}_t - y_t\|^2$ and update and project onto the low Frobenius norm ball

$$\hat{M}^{t+1} \leftarrow M^t - \eta_t \nabla_M \ell_t(M^t)$$

$$M^{t+1} = \pi_{\mathcal{K}}\left(\hat{M}^{t+1}\right),$$

where $\mathcal{K}_r = \{[M_1, \ldots, M_k] \text{ s.t. } \|M_i\| \le r, \forall i \in [k]\}$.

5: **end for**

Theorem 6. *Let $T \in \mathbb{Z}_{\ge 0}$ and $q \in \left[\frac{1}{4} + \frac{\log(\log(T)/8)}{\log(T)}, 1\right]$. Consider a sequence $(y_1, \ldots, y_T)$ generated by an unknown and noiseless linear dynamical system defined by matrices (A, B, C, I) as per Eq. 1. Assume the input sequence $u_{0:(t-1)}$ is sufficiently well-conditioned, satisfying $\sum_{t=0}^{T-1}(T-t)u_t u_t^\top \succeq \left(\frac{2|C||B|}{\sqrt{T}}\right) I$. Let $k = \Omega\left(\log(T) \cdot \log\left(Td_A\right)\right)$, $r \ge \|B\|\|C\|$ and assume $T \ge (4k\log^2(T)/\|C\|\|B\|)^4$. Algorithm 2 satisfies:*

$$\textit{Regret}_{\text{Asym},T}\left(\mathcal{A}(T^q), \Pi_T^{\text{SF}}\right) \le O\left(\|B\|^2\|C\|^2 k^{3/2}\log(T)^2\sqrt{T}\right).$$

The proof of Theorem 6 is in Appendix C and we now give a high-level overview.

High-Level Proof Overview. The general proof technique for both Theorem 5 and Theorem 6 is the same. First, using standard online gradient descent results from (Hazan et al., 2017b) we prove that the iterates M^t achieve $O(\sqrt{T})$ regret as measured by the context-length restricted loss $\sum_{t=1}^T \ell_t(M, L)$. That is,

$$\sum_{t=1}^T \ell_t(M^t, L) \le \min_{M \in \mathcal{K}_r} \sum_{t=1}^T \ell_t(M, L) + O(\sqrt{T}). \tag{6}$$

Next we prove that there is a unique M_T^* which minimizes the loss on the full T-length context and this M_T^* achieves length generalization in the sense that it achieves small loss even when only allowed to use context length L. That is

$$\sum_{t=1}^T \ell_t(M_T^*, L) \le \sum_{t=1}^T \ell_t(M_T^*, T) + O(\sqrt{T}). \tag{7}$$

We combine Eq. 6 and Eq. 7 to get the final notion of length generalization that

$$\begin{aligned}\sum_{t=1}^T \ell_t(M^t, L) &\le \min_{M \in \mathcal{K}_r} \sum_{t=1}^T \ell_t(M, L) + O(\sqrt{T}) \\ &\le \sum_{t=1}^T \ell_t(M_T^*, L) + O(\sqrt{T}) \\ &\le \sum_{t=1}^T \ell_t(M_T^*, T) + O(\sqrt{T}).\end{aligned}$$

The difficult result to prove is Eq. 7. The high level idea is that when $y_{1:t}$ evolves as a noiseless LDS and when the input $u_{0:(t-1)}$ is sufficiently well-conditioned, then $\sum_{t=1}^T \ell_t(M, T)$ is strongly convex and the minimizer approximately recovers a collection of "true" matrices which are generated by the underlying linear dynamical system. The second key idea is that if an algorithm had access to these "true" matrices then it would be able to achieve small loss even when restricted to a small context-length $L \ll T$. The extent to which these recovered matrices can achieve small loss when restricted to the small context-length depends on the way the algorithm chooses to predict y_t. In the case of Algorithm 1 where y_t is predicted based only using only one autoregressive term, even having access to the true matrices is not enough to accurately predict y_t. However, in the case of Algorithm 2, having access to the true matrices as well as a second autoregressive term allows accurate prediction of y_t even when restricted to small context-length window.

4. Experiments

4.1. Linear Dynamical System

We can empirically verify Theorem 5 and Theorem 6 in an online sequence prediction task where the data is gen-

erated by a noiseless LDS. We refer to a "bad" region of eigenvalues $\left(1-\log(T)/(8T^{7/8}),\ 1-1/(2T^{5/4})\right)$ as Region B, and we define Region A to hug Region B on both sides as shown in Figure 1.

Figure 1. Region B is the interval of eigenvalues for which Theorem 5 does not provide length generalization. Region A hugs both sides of Region B (Region A is $\left[0.9\cdot\left(1-\log(T)/(8T^{7/8})\right),1\right]\setminus$ `Region B`). This ensures that Region A will contain bad eigenvalues as q decreases from $7/8$ and eigenvalues in Region B are bad for $q<7/8$.

Theorem 5 predicts that if all the eigenvalues lie outside Region B, then spectral filtering will length generalize from $T^{7/8}$ to T. To confirm this, we generate a random LDS (hidden dimension: 512) with half of its eigenvalues sampled from each part of **Region A**. The online prediction losses are plotted in Figure 2 for different choices of context length T^q, where $T=2^{14}$ and $k=24$. As expected from the theory, context lengths approaching $T^{7/8}$ closely match the performance of the optimal spectral filtering predictor with full context.

Interestingly, we see that context length $T^{1/2}$ consistently fails in a qualitatively worse fashion – indeed, some of the values in Region A are actually "bad" for $q=1/2$. This seems to suggest that such eigenvalues can actually cause instabilities with length generalization and are not limitations of our proof – if true, such a fact could be seen as a partial converse to Theorem 5. To check this conjecture empirically, we run another experiment where we generate a random LDS of hidden dimension 512 with all eigenvalues in **Region B** and plot the prediction losses $\ell_t(M^t,T^q)$ for M^t from Algorithm 1 in Figure 3 (averaged over random seeds and smoothed). These results confirm that (some subset of) this bad region is indeed what impedes the length generalization capability of spectral filtering.

Next we apply our novel Algorithm 2, which uses two autoregressive components. Theorem 6 predicts that this algorithm should be robust to this bad region of eigenvalues and instead achieve length generalazation for any (symmetric, marginally-stable) LDS. We verify this experimentally in Figure 4 – to be as adversarial as we can, this experiment is run with all eigenvalues sampled from **Region B**. As predicted by Theorem 6, the second autoregressive component allows for robust length generalization even with context lengths as small as $\sqrt{T}$.

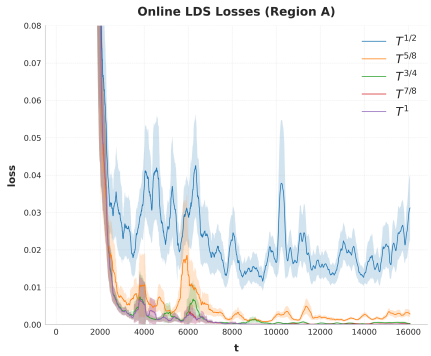

Figure 2. Loss for Algorithm 1 with eigenvalues in **Region A**.

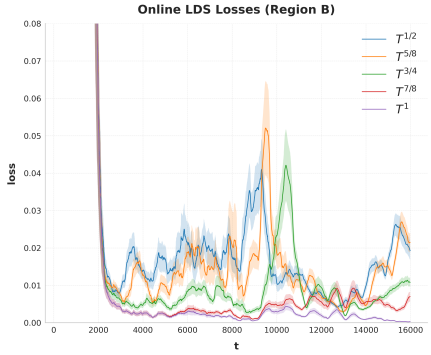

Figure 3. Loss for Algorithm 1 with eigenvalues in **Region B**.

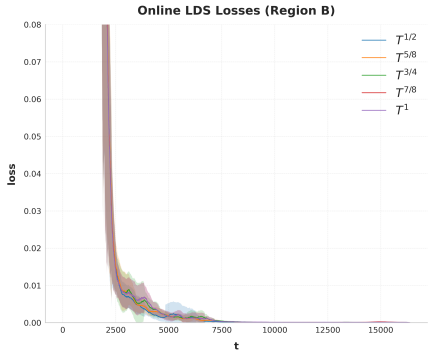

Figure 4. Loss for Algorithm 2 with eigenvalues in **Region B**.

4.2. Induction Heads

So far, we have demonstrated length generalization of spectral filtering on linear systems: when trained with a shorter context length of T^q it is able to compete with methods that have access to the full context T (even on marginally-stable systems that can have arbitrarily large effective memory lengths). This length generalization property is most crucial in deep learning applications, in which multi-layer models are stacked (with added nonlinearities) to solve non-LDS sequence prediction task

As an empirical proof-of-concept to demonstrate that STU's length generalization capability extends to this regime, we evaluate it on the induction heads synthetic sequence modeling task, which is commonplace in the language modeling literature (see (Gu & Dao, 2023)) and was experimentally shown in (Liu et al., 2024) to be efficiently solved by a two-layer STU. In the induction heads task, the model is required to recall one token (sampled uniformly from a vocabulary) immediately after a special `flag` token; the rest of the sequence consists of the same special `blank` token, which the model should learn to ignore.

The STU architecture we use is composed of an embedding layer, two "tensordot" STU layers with MLPs and ReLU nonlinearities, and an output projection layer (the same as in (Liu et al., 2024)) with filters of length $T = 256$.

Following prior STU architecture implementations we use **no autoregressive components**, and so any length generalization observed here comes directly from the filtering mechanism itself. We train these models until convergence with a tuned Adam optimizer and various context lengths T^q. The vocabulary size is set to 4.

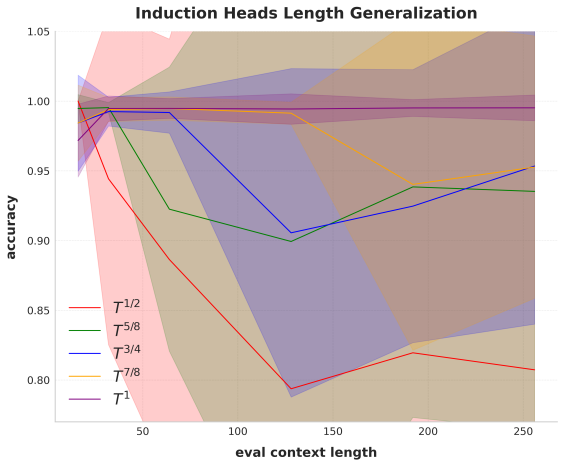

Figure 5. Accuracies for STU models trained on an induction heads task of length T^q and evaluated on sequence lengths increasing up to T, averaged over random seeds.

Accuracies are plotted[2] in Figure 5 for evaluation task lengths increasing up to T. As we see, vanilla STU models are able to nontrivially length generalize and occasionally retain good accuracy beyond their training context lengths, though inconsistently. Importantly, unlike algorithms that achieve length generalization through architectural modification, we simply train with filters longer than the train context. As such, this method allows for the convolutional mode during training and inherits all the benefits of STU that are demonstrated in (Liu et al., 2024). For example, the nonlinear selection mechanism of (Gu & Dao, 2023) allows for extreme length generalization on induction heads without prior knowledge of the evaluation length, though at a cost to training efficiency, implementation simplicity, and optimization complexity. We reiterate that our goal is not to navigate such a tradeoff by modifying the STU model so that it length generalizes on induction heads, but rather to exhibit a provable length generalization capability of the STU that comes for free from its natural structure.

[2]Even though the accuracy cannot go above 1, the error bars (1.96 times standard deviation) are still well defined above this value. For example, $T^{7/8}$ at eval length 256 has average accuracy of 95% with error interval roughly $[85\%, 105\%]$, indicating that most trials achieved perfect accuracy and length generalization.

5. Discussion

In review, we first introduced the notion of Asymmetric-Regret as a way to describe length generalization through the lens of online learning and regret minimization in games. We then proved that the class of spectral filtering predictors naturally enjoys sublinear Asymmetric-Regret thereby exhibiting length generalization without any change to the algorithm, albeit with some restrictions on the underlying data (i.e. the spectrum of A). We introduced a new variant of spectral filtering which uses two autoregressive components and achieves length generalization which is more robust to the assumptions of the underlying data. Next, we used experiments on synthetic data generated by an LDS to demonstrate the validity and sharpness of our theory and provided proof-of-concept length generalization experiments on a synthetic nonlinear sequence prediction task.

Our theoretical results and initial empirical findings reveal that some type of length generalization comes naturally with the spectral filtering algorithm. This result implies that spectral filtering is powerful in its ability to learn the dynamics of a complicated underlying system with long memory – it naturally handles the issue of what aspects in a sequence should be memorized for the future and what aspects can be forgotten, whereas many existing methods are hand engineered depending on the specific task. This adds to the already-exciting list of its useful (and provable) properties, including: robustness to systems with long memory and large hidden dimension, efficient training via convolutions, optimization convexity, and the existence of good parameter-efficient approximations. Given recent successful applications of spectral filtering as the building block for STU models in deep learning (Agarwal et al., 2023; Liu et al., 2024), it would be valuable to research how to best take advantage of their length generalization capacity at scale – we leave this for future work.

Impact Statement

This paper is primarily theoretical, proving that spectral filtering learns LDS's efficiently under a stronger notion of regret which describes a type of length generalization. We believe there are no societal consequences of our work that require specific highlighting here.

Acknowledgements

Elad Hazan acknowledges support from the Office of Naval Research and Open Philanthropy.

References

Agarwal, N., Suo, D., Chen, X., and Hazan, E. Spectral state space models. *arXiv preprint arXiv:2312.06837*, 2023.

Anil, C., Wu, Y., Andreassen, A., Lewkowycz, A., Misra, V., Ramasesh, V., Slone, A., Gur-Ari, G., Dyer, E., and Neyshabur, B. Exploring length generalization in large language models. *Advances in Neural Information Processing Systems*, 35:38546–38556, 2022.

Bansal, A., Schwarzschild, A., Borgnia, E., Emam, Z., Huang, F., Goldblum, M., and Goldstein, T. End-to-end algorithm synthesis with recurrent networks: Extrapolation without overthinking. *Advances in Neural Information Processing Systems*, 35:20232–20242, 2022.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Cesa-Bianchi, N., Conconi, A., and Gentile, C. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., and Kaiser, Ł. Universal transformers. *arXiv preprint arXiv:1807.03819*, 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Gu, A., Dao, T., Ermon, S., Rudra, A., and Ré, C. Hippo: Recurrent memory with optimal polynomial projections. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1474–1487. Curran Associates, Inc., 2020.

Gu, A., Goel, K., and Ré, C. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021a.

Gu, A., Johnson, I., Goel, K., Saab, K., Dao, T., Rudra, A., and Ré, C. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34: 572–585, 2021b.

Gupta, A., Gu, A., and Berant, J. Diagonal state spaces are as effective as structured state spaces. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=RjS0j6tsSrf.

Hazan, E. and Singh, K. Introduction to online nonstochastic control. *arXiv preprint arXiv:2211.09619*, 2022.

Hazan, E., Singh, K., and Zhang, C. Efficient regret minimization in non-convex games. In *International Conference on Machine Learning*, pp. 1433–1441. PMLR, 2017a.

Hazan, E., Singh, K., and Zhang, C. Learning linear dynamical systems via spectral filtering. *Advances in Neural Information Processing Systems*, 30, 2017b.

Hazan, E., Lee, H., Singh, K., Zhang, C., and Zhang, Y. Spectral filtering for general linear dynamical systems. *Advances in Neural Information Processing Systems*, 31, 2018.

Hazan, E., Kakade, S., and Singh, K. The nonstochastic control problem. In Kontorovich, A. and Neu, G. (eds.), *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pp. 408–421, San Diego, California, USA, 08 Feb–11 Feb 2020. PMLR. URL http://proceedings.mlr.press/v117/hazan20a.html.

Hazan, E. et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Hou, K., Brandfonbrener, D., Kakade, S., Jelassi, S., and Malach, E. Universal length generalization with turing programs. *arXiv preprint arXiv:2407.03310*, 2024.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kaiser, Ł. and Sutskever, I. Neural gpus learn algorithms. *arXiv preprint arXiv:1511.08228*, 2015.

Kazemnejad, A., Padhi, I., Natesan Ramamurthy, K., Das, P., and Reddy, S. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems*, 36, 2024.

Liu, Y. I., Nguyen, W., Devre, Y., Dogariu, E., Majumdar, A., and Hazan, E. Flash stu: Fast spectral transform units. *arXiv preprint arXiv:2409.10489*, 2024.

Murray, K. and Chiang, D. Correcting length bias in neural machine translation. *arXiv preprint arXiv:1808.10006*, 2018.

Newman, B., Hewitt, J., Liang, P., and Manning, C. D. The eos decision and length extrapolation. *arXiv preprint arXiv:2010.07174*, 2020.

Orvieto, A., Smith, S. L., Gu, A., Fernando, A., Gulcehre, C., Pascanu, R., and De, S. Resurrecting recurrent neural networks for long sequences. *arXiv preprint arXiv:2303.06349*, 2023.

Poli, M., Massaroli, S., Nguyen, E., Fu, D. Y., Dao, T., Baccus, S., Bengio, Y., Ermon, S., and Ré, C. Hyena hierarchy: Towards larger convolutional language models. *arXiv preprint arXiv:2302.10866*, 2023.

Press, O., Smith, N. A., and Lewis, M. Train short, test long: Attention with linear biases enables input length extrapolation, 2022. URL https://arxiv.org/abs/2108.12409.

Schwarzschild, A., Borgnia, E., Gupta, A., Huang, F., Vishkin, U., Goldblum, M., and Goldstein, T. Can you learn an algorithm? generalizing from easy to hard problems with recurrent networks. *Advances in Neural Information Processing Systems*, 34:6695–6706, 2021.

Smith, J. T., Warrington, A., and Linderman, S. Simplified state space layers for sequence modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Tay, Y., Dehghani, M., Bahri, D., and Metzler, D. Efficient transformers: A survey. *ACM Comput. Surv.*, 55(6), dec 2022. ISSN 0360-0300. doi: 10.1145/3530811. URL https://doi.org/10.1145/3530811.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yehudai, G., Fetaya, E., Meirom, E., Chechik, G., and Maron, H. From local structures to size generalization in graph neural networks. In *International Conference on Machine Learning*, pp. 11975–11986. PMLR, 2021.

Zhang, Y., Backurs, A., Bubeck, S., Eldan, R., Gunasekar, S., and Wagner, T. Unveiling transformers with lego: a synthetic reasoning task. *arXiv preprint arXiv:2206.04301*, 2022.

A. General Length Generalization

In this section we introduce a general algorithm which we will use to prove length generalization for both Algorithm 1 and Algorithm 2.

Algorithm 3 General Spectral Filtering

1: **Input:** $k > 0, L > 0, r > 0$, functions $p_t(\cdot)$, vectors $v_{1:k}$. Initialize $M_i = 0$ for $i \in [k]$.
2: **for** $t = 1, 2, ..., T$ **do**
3: Compute and predict

$$\hat{y}_t = p_t(y_{t-1:1}) + \sum_{i=1}^{k} M_i u_{t-1:t-L} v_i.$$

4: Observe y_t, denote $\ell_t(M^t, L) = \|\hat{y}_t - y_t\|^2$ and update and project update and project onto the low Frobenius norm ball

$$\hat{M}^{t+1} \leftarrow M^t - \eta_t \nabla_M \ell_t(M^t)$$

$$M_{t+1} = \text{Proj}_{\mathcal{K}} \left(\hat{M}_{t+1} \right),$$

where $\mathcal{K}_r = \{M \text{ s.t. } \|M_i\| \leq r\}$.
5: **end for**

Our workhorse theorem is presented below. We will use this theorem to prove length generalization for our special cases in the following sections.

Theorem 7. *Suppose $y_{1:t}$ evolves as a noiseless (A, B, C, I)-LDS and the input $u_{(t-1):0}$ is such that $\sum_{t=0}^{T-1}(T-t)u_t u_t^\top \succeq (2\|C\|\|B\|/\sqrt{T})I$. Let k, L, r, $\{v_i\}_{i=1}^k$, $p_t(\cdot)$, and $\ell_t(\cdot)$ all be as defined in Algorithm 3. Suppose $\{v_i\}_{i=1}^k$ is orthonormal with $\|v_i\|_1 \leq \log^p(T)$. Suppose that $p_t(\cdot)$ is such that there exists some function $h(\cdot)$, constant $\ell > 0$, and some $M^{true} \in \mathcal{K}_r$ such that*

$$y_t - p_t(y_{t-1:1}) = \sum_{i=1}^{T} M_i^{true} u_{t-1:0} v_i = \sum_{i=1}^{\ell_1} M_i^{true} u_{t-i} + \sum_{i=1}^{t-\ell_1} CA^i h(A) B u_{t-\ell_1-i},$$

where

$$\| \sum_{i=k+1}^{T} M_i^{true} u_{t-1:t-L} v_i \| \leq \|C\|\|B\|/T,$$

and

$$\max_{\alpha(A)} \left\{ h(\alpha)\alpha^{L-\ell_1-1}(1-\alpha^{T-L+1})(1-\alpha)^{-1} \right\} \leq \frac{1}{T^{1/4}}.$$

Then if M^t are the iterates of Algorithm 3 and $T \geq (4k\log^p(T)/\|C\|\|B\|)^4$,

$$\sum_{t=1}^{T} \ell_t(M^t, L) - \min_{M^* \in \mathcal{K}_r} \sum_{t=1}^{T} \ell_t(M^*, T) \leq \left(12k^{3/2}r^2 \log^p(T) + 8\|C\|^2\|B\|^2 \right) \sqrt{T}.$$

The proof of this theorem requires several technical lemmas which we present and prove in the subsequent subsections. In Lemma 8 we essentially prove the standard result showing that Online Gradient Descent implemented in Algorithm 3 achieves $O(\sqrt{T})$ regret. In Lemma 9 we prove the more nuanced result which shows that the optimal M which minimizes the loss on the full T-length context achieves length generalization in the sense that it achieves small loss even when only allowed to use context length L. Combining these two lemmas gives the proof of Theorem 7.

Proof of Theorem 7. Let

$$M_T^* \stackrel{\text{def}}{=} \min_{M^* \in \mathcal{K}_r} \sum_{t=1}^{T} \ell_t(M^*, T)$$

and observe that

$$\min_{M^* \in \mathcal{K}_r} \sum_{t=1}^{T} \ell_t(M^*, L) \;\leq\; \sum_{t=1}^{T} \ell_t(M_T^*, L). \tag{8}$$

Combining this with Lemma 8 and Lemma 9, we conclude

$$\begin{aligned}
\sum_{t=1}^{T} \ell_t(M_t, L) \;&\leq\; \min_{M^* \in \mathcal{K}_r} \sum_{t=1}^{T} \ell_t(M^*, L) + 12k^{3/2}r^2 \overset{p}{\log}(T)\sqrt{T} && \text{OGD Regret Lemma 8} \\
&\leq \sum_{t=1}^{T} \ell_t(M_T^*, L) + 12k^{3/2}r^2 \overset{p}{\log}(T)\sqrt{T} && \text{Eq. 8} \\
&\leq \sum_{t=1}^{T} \ell_t(M_T^*, T) + (12k^{3/2}r^2 \overset{p}{\log}(T) + 8\|C\|^2\|B\|^2)\sqrt{T} && \text{Length Generalization Lemma 9} \\
&= \min_{M^* \in \mathcal{K}_r} \sum_{t=1}^{T} \ell_t(M, T) + (12k^{3/2}r^2 \overset{p}{\log}(T) + 8\|C\|^2\|B\|^2)\sqrt{T}. && \text{Definition of } M_T^*
\end{aligned}$$

□

A.1. OGD Regret for Generalized Spectral Filtering

Lemma 8. *Suppose the input $u_{1:t}$ satisfies $\|u_t\|_2 \leq 1$. Suppose the true output y_t evolves such that for some polynomial $p_t(y_{t-1:1})$ there exists some $M^{true} \in \mathcal{K}_r$*

$$y_t = p_t(y_{t-1:1}) + \sum_{i=1}^{T} M_i^{true} u_{t-1:0} v_i,$$

and for

$$E_{m,T} \overset{def}{=} \sum_{i=k+1}^{T} M_i^{true} u_{t-1:0} v_i,$$

we have $\|E_{m,T}\| \leq 1$. Further suppose $v_1, \ldots, v_k$ satisfy $\|v_i\|_1 \leq c_i \log^p(T)$. Let

$$\ell_t(M, L) \overset{def}{=} \|y_t - p_t(y_{t-1:1}) - \sum_{i=1}^{k} M_i u_{t-1:t-L} v_i\|^2,$$

Then if M^t are the iterates of Algorithm 3

$$\sum_{t=1}^{T} \ell_t(M^t, L) - \min_{M^* \in \mathcal{K}_r} \sum_{t=1}^{T} \ell_t(M^*, L) \;\leq\; 12k^{3/2}r^2 \overset{p}{\log}(T)\sqrt{T}.$$

Proof of Lemma 8. This proof is a near copy of the proof in (Hazan et al., 2017b), the difference is that we derive several equations that we will use later and we handle the varying context length.

Let $G = \max_{t \in [T]} \|\nabla_M \ell_t(M_t, L)\|$ and let $D = \max_{M_1, M_2 \in \mathcal{K}_r} \|M_1 - M_2\|$. By Theorem A.1 from (Hazan & Singh, 2022),

$$\sum_{t=1}^{T} \ell_t(M^t, L) - \min_{M^* \in \mathcal{K}_r} \sum_{t=1}^{T} \ell_t(M^*, L) \;\leq\; \frac{3}{2} GD\sqrt{T}.$$

Therefore it remains to bound G and D.

First we bound D. By definition of $\mathcal{K}_r$, we have that for any $M \in \mathcal{K}_r$,

$$\|M_i\| \;\leq\; r.$$

Therefore, we also have that

$$\|M\| \;\le\; \sqrt{k}r.$$

Therefore

$$D \stackrel{\text{def}}{=} \max_{M,M' \in \mathcal{K}_r} \|M - M'\| \;\le\; 2\sqrt{k}r.$$

Next we bound the gradient norm G. Using the definition of $\mathcal{K}_r$,

$$\max_{M \in \mathcal{K}_r} \max_{i \in [k]} \|M_i\| \;\le\; r.$$

We bound the gradient norm as follows,

$$\begin{aligned}
\|\nabla_{M_j} \ell_t(M, L)\| &= \|2 \left(\sum_{i=1}^{k} M_i^{\text{true}} u_{t-1:0} v_i + E_{m,T} - \sum_{i=1}^{k} M_i u_{t-1:t-L} v_i \right) \left(u_{t-1:t-L} v_j\right)^\top \| \\
&\le\; 2 \left(\sum_{i=1}^{k} \|M_i^{\text{true}}\| \|u_{t-1:0}\|_\infty \|v_i\|_1 + \|E_{m,T}\| + \sum_{i=1}^{k} \|M_i\| \|u_{t-1:t-L}\|_\infty \|v_i\|_1 \right) \|u_{t:t-L}\|_\infty \|v_j\|_1 \\
&\le\; 2\left(1 + \|E_{m,T}\|\right) \sum_{i=1}^{k} \max_{M \in \mathcal{K}_r} \|M_i\| \cdot \|u_{t-1:0}\|_\infty^2 \cdot \|v_i\|_1^2 \\
&\le\; 4kr \overset{p}{\log}(T).
\end{aligned}$$

Putting everything together we have

$$\begin{aligned}
\sum_{t=1}^{T} \ell_t(M_t, L) - \min_{M^* \in \mathcal{K}_r} \sum_{t=1}^{T} \ell_t(M^*, L) \;&\le\; \frac{3}{2} \left(4kr \overset{p}{\log}(T) \right) \left(2\sqrt{k}r \right) \sqrt{T} \\
&= 12 k^{3/2} r^2 \overset{p}{\log}(T) \sqrt{T}.
\end{aligned}$$

□

A.2. Length Generalization on the Best Optimizer in Hindsight

Lemma 9. *Let input $u_{(t-1):0}$, $\{v_i\}_{i=1}^k$, $p_t(\cdot)$, and $\ell_t(M, L)$ all be as defined in Algorithm 3. Suppose the input $u_{(t-1):0}$ is such that $\sum_{t=0}^{T-1} (T-t) u_t u_t^\top \succeq (2\|C\|\|B\|/\sqrt{T}) I$, $\{v_i\}_{i=1}^k$ is orthonormal with $\|v_i\|_1 \;\le\; \log^p(T)$, and that there exists some M^{true} such that*

$$y_t - p_t(y_{t-1:1}) = \sum_{i=1}^{T} M_i^{true} u_{t-1:0} v_i = \sum_{i=1}^{\ell_1} M_i^{true} u_{t-i} + \sum_{i=1}^{t-\ell_1-1} C A^i h(A) B u_{t-\ell_1-i},$$

where

$$\| \sum_{i=k+1}^{T} M_i^{true} u_{t-1:t-L} v_i \| \;\le\; \|C\|\|B\|/T,$$

and

$$\max_{\alpha(A)} \left\{ h(\alpha) \alpha^{L-\ell_1-1} (1 - \alpha^{T-L+1}) (1-\alpha)^{-1} \right\} \;\le\; \frac{1}{T^{1/4}}.$$

Let

$$M_T^* \stackrel{def}{=} \arg\min_{M \in \mathcal{K}_r} \sum_{t=1}^{T} \ell_t(M, T).$$

Then for $T \;\ge\; (4k \log^p(T)/\|C\|\|B\|)^4$, the loss with context L well approximates the loss with context T on M_T^*,

$$|\sum_{t=1}^{T} \ell_t(M_T^*, L) - \ell_t(M_T^*, T)| \;\le\; 8\|C\|^2 \|B\|^2 \sqrt{T}.$$

The proof of Lemma 9 requires two key helper lemmas which we develop in the following subsections. The first is Lemma 10 which establishes that when $y_{1:t}$ evolves as a noiseless LDS and if the input $u_{1:t}$ is sufficiently well-conditioned, then the minimizer for $\sum_{t=1}^{T} \ell_t(M, T)$ approximately recovers a collection of matrices (we denote as M^{true}) which is generated by the true linear dynamical system. The second key helper Lemma is Lemma 11 which establishes that an algorithm which uses the collection of matrices that are generated by the true linear dynamical system, i.e. M^{true}, is able to achieve small loss even when restricted to a small context-length $L << T$. The proof of Lemma 9 combines these two insights to establish that this implies that the minimizer for $\sum_{t=1}^{T} \ell_t(M, T)$ also achieves small loss even when restricted to small context-length L.

Proof of Lemma 9. First we show that M^{true} is a $(\|C\|^2\|B\|^2/T)$-approximate minimizer to $\sum_{t=1}^{T} \ell_t(M, T)$. Indeed,

$$\begin{aligned}\sum_{t=1}^{T} \ell_t(M^{\text{true}}, T) &= \sum_{t=1}^{T} \|y_t - p_t(y_{t-1:1}) - \sum_{i=1}^{k} M_i^{\text{true}} u_{t-1:0} v_i\|^2 \\ &= \sum_{t=1}^{T} \| \sum_{i=k+1}^{T} M_i^{\text{true}} u_{t-1:0} v_i\|^2 \\ &\leq \|C\|^2\|B\|^2/T.\end{aligned}$$

By assumption $\sum_{t=0}^{T-1}(T-t)u_t u_t^\top \succeq (2\|C\|\|B\|/\sqrt{T})I$. Therefore, by Lemma 10 with $\epsilon = \|C\|\|B\|/\sqrt{T}$ we have

$$M_T^* \in \mathcal{B}_{\|C\|\|B\|/\sqrt{T}}\left(M^{\text{true}}\right).$$

Since we assumed $T \geq (4k\log^p(T)/\|C\|\|B\|)^4$ we have

$$\|C\|\|B\|/\sqrt{T} \leq \|C\|^2\|B\|^2/(4kT^{1/4}\log^p(T)).$$

Therefore by Lemma 11 we have

$$\sum_{t=1}^{T} \ell_t(M_T^*, L) \leq 4\|C\|^2\|B\|^2\sqrt{T}.$$

Moreover note that

$$0 \leq \ell_t(M_T^*, T) \leq \ell_t(M^{\text{true}}, T) \leq \|C\|^2\|B\|^2/T^2.$$

Combining these we conclude,

$$|\sum_{t=1}^{T} \ell_t(M_T^*, L) - \sum_{t=1}^{T} \ell_t(M_T^*, T)| \leq 4\|C\|^2\|B\|^2\sqrt{T} + \|C\|^2\|B\|^2/T \leq 8\|C\|^2\|B\|^2\sqrt{T}.$$

□

A.2.1. Minimization is Recovery

Lemma 10. *Suppose $\sum_{t=0}^{T-1}(T-t)u_t u_t^\top \succeq 2\epsilon I$ and $\{v_i\}_{i=1}^{k}$ is orthonormal. Then there is a unique point M^* which minimizes the function $\sum_{t=1}^{T} \ell_t(M, T)$ from Algorithm 3. Moreover, suppose some k satisfies*

$$\sum_{t=1}^{T} \ell_t(M, T) \leq \epsilon^2.$$

Then there is a matrix E_M such that $\|E_M\| \leq \epsilon$ and

$$M^* = M + E_M.$$

Proof. For convenience, let X_t be the kd_{in}-dimensional vector which stacks the filters,

$$X_t = \begin{bmatrix} u_{t-1:t-T} v_1 \\ u_{t-1:t-T} v_2 \\ \vdots \\ u_{t-1:t-T} v_k \end{bmatrix} = \begin{bmatrix} u_{t-1:0} v_1 \\ u_{t-1:0} v_2 \\ \vdots \\ u_{t-1:0} v_k \end{bmatrix},$$

where the second inequality holds since we only consider $t \leq T$. Assume k is written as $M = \begin{bmatrix} M_1 & M_2 & \dots & M_k \end{bmatrix} \in \mathbb{R}^{d_{\text{out}} \times k d_{\text{in}}}$ and let $Y_t = y_t - p_t(y_{t-1:1})$. Let $Y = \begin{bmatrix} Y_1 & Y_2 & \dots & Y_T \end{bmatrix}$ and $X = \begin{bmatrix} X_1 & X_2 & \dots & X_T \end{bmatrix}$. Then we can express the loss as

$$f(M) \stackrel{\text{def}}{=} \sum_{t=1}^{T} \ell_t(M, T) = \|Y - MX\|^2.$$

Note that this function is twice differentiable and

$$\nabla_M^2 f(M) = XX^\top.$$

Therefore, if $\lambda_{\min}\left(XX^\top\right) \geq \mu$ we have that $f(M)$ is μ-strongly convex. Then if M^* is the optimum of $f(M)$ we have

$$f(M) \geq f(M^*) + \frac{\mu}{2}\|M - M^*\|^2, \text{ or equivalently, } \|M - M^*\| \leq \frac{2}{\mu}\left(f(M) - f(M^*)\right).$$

Now suppose k is such that $f(M) \leq \epsilon^2$. Then since $f(\cdot) \geq 0$ we have

$$\|M - M^*\| \leq 2\epsilon^2/\mu.$$

Therefore we can write

$$M^* = M + E_{M^*} \text{ where } \|E_{M^*}\| \leq 2\epsilon^2/\mu. \tag{9}$$

Next we must understand the eigenvalues of $XX^\top$ and how they relate to the input $u_{T:1}$. For notational convenience, let $U = u_{T:1}$ and let D_t denote the block-diagonal $T \times T$ matrix

$$D_t \stackrel{\text{def}}{=} \begin{bmatrix} 0_{T-t \times T-t} & \\ & I_t \end{bmatrix}.$$

Finally, let

$$V = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{bmatrix} \in \mathcal{R}^{Tm \times 1}$$

Then we have $X_t = \left(I_k \otimes U D_t\right) V$ and we observe

$$\begin{aligned} XX^\top &= \sum_{t=1}^{T} X_t X_t^\top = \sum_{t=1}^{T} \left((I_k \otimes U D_t)V\right)\left((I_k \otimes U D_t)V\right)^\top \\ &= \sum_{t=1}^{T} (I_k \otimes U D_t U^\top) \\ &= I_k \otimes U \left(\sum_{t=1}^{T} D_t\right) U^\top. \end{aligned}$$

Observe that

$$\sum_{t=1}^{T} D_t = \operatorname{diag}\left(\begin{bmatrix} 1 & 2 & \dots & T \end{bmatrix}\right).$$

Using this we can further refine

$$U \left(\sum_{t=1}^{T} D_t\right) U^\top = \sum_{t=0}^{T-1} (T-t) u_t u_t^\top.$$

By assumption, this matrix has minimum eigenvalue bounded below by 2ϵ. Therefore $\lambda_{\min}(XX^\top) \geq 2\epsilon$. Plugging this value in for μ in Eq. 9 concludes the proof.

□

A.2.2. Uniform Length Generalization Around LDS Generated Solutions

The following lemma shows that any k in an (appropriately defined) ϵ-ball around M^{true} obtains length generalization in the sense that it achieves $O(\sqrt{T})$ L-context-length-limited loss $\sum_{t=1}^{T} \ell_t(\cdot, L)$.

Lemma 11. *Suppose y_t evolves as a noiseless (A, B, C, I)-LDS with input u_t. Suppose $p_t(\cdot)$ and M^{true} is such that*

$$y_t - p_t(y_{t-1:1}) = \sum_{i=1}^{T} M_i^{true} u_{t-1:0} v_i = \sum_{i=1}^{\ell_1} M_i^{true} u_{t-i} + \sum_{i=1}^{t-\ell_1-1} CA^i h(A) B u_{t-\ell_1-i}.$$

Suppose for a given $k > 0$,

$$\| \sum_{i=k+1}^{T} M_i^{true} u_{t-1:t-L} v_i \| \ \leq \ \frac{\|C\|\|B\|}{T}.$$

Suppose

$$\max_{\alpha(A)} \left\{ h(\alpha) \alpha^{L-\ell_1-1} (1 - \alpha^{T-L+1})(1-\alpha)^{-1} \right\} \ \leq \ \frac{1}{T^{1/4}}.$$

If

$$\delta \ \leq \ \frac{1}{4m} \frac{\|C\|^2 \|B\|^2}{T^{1/4} \log^p(T)},$$

then we have for any $M \in \mathcal{B}_\delta(M^{true})$

$$\sum_{t=1}^{T} \ell_t(M, L) \ \leq \ 4\|C\|^2 \|B\|^2 \sqrt{T}.$$

Proof of Lemma 11. Let $M = M^{\text{true}} + E_M$, where $\|E_M\| \ \leq \ \delta$. By definition,

$$\begin{aligned}
\ell_t(M^{\text{true}} + E_M, L) &= \|y_t - p_t(y_{t-1:1}) - \sum_{i=1}^{k} \left(M^{\text{true}} + E_M\right)_i u_{t-1:t-L} v_i\|^2 \\
&= \|y_t - p_t(y_{t-1:1}) - \sum_{i=1}^{k} M_i^{\text{true}} u_{t-1:t-L} v_i - \sum_{i=1}^{k} E_{M_i} u_{t-1:t-L} v_i\|^2 \\
&\leq \ \|y_t - p_t(y_{t-1:1}) - \sum_{i=1}^{k} M_i^{\text{true}} u_{t-1:t-L} v_i\|^2 \\
&\quad + 2\|y_t - p_t(y_{t-1:1}) - \sum_{i=1}^{k} M_i^{\text{true}} u_{t-1:t-L} v_i\| \| \sum_{i=1}^{k} E_{M_i} u_{t-1:t-L} v_i\| \\
&\quad + \| \sum_{i=1}^{k} E_{M_i} u_{t-1:t-L} v_i\|^2.
\end{aligned}$$

Observe that

$$\| \sum_{i=1}^{k} E_{M_i} u_{t-1:t-L} v_i\| \ \leq \ \sum_{i=1}^{k} \|E_{M_i}\| \|u_{t-1:t-L}\|_\infty \|v_i\|_1 \ \leq \ k\delta \overset{p}{\log}(T).$$

For the remainder of the proof we work towards bounding $\|y_t - p_t(y_{t-1:1}) - \sum_{i=1}^{k} M_i^{\text{true}} u_{t-1:t-L} v_i\|$. We replace $y_t - p_t(y_{t-1:1})$ with $\sum_{i=1}^{T} M_i^{\text{true}} u_{t-1:0} v_i$ and we replace $\sum_{i=1}^{k} M_i^{\text{true}} u_{t-1:t-L} v_i$ with $\sum_{i=1}^{T} M_i^{\text{true}} u_{t-1:t-L} v_i -$

$\sum_{i=k+1}^{T} M_i^{\text{true}} u_{t-1:t-L} v_i$ to get

$$\begin{aligned}\|y_t - p_t(y_{t-1:1}) - \sum_{i=1}^{k} M_i^{\text{true}} u_{t-1:t-L} v_i\|^2 &= \|\left(\sum_{i=1}^{T} M_i^{\text{true}} u_{t-1:0} v_i\right) - \left(\sum_{i=1}^{T} M_i^{\text{true}} u_{t-1:t-L} v_i - \sum_{i=k+1}^{T} M_i^{\text{true}} u_{t-1:t-L} v_i\right)\|^2 \\ &\le \|\sum_{i=1}^{T} M_i^{\text{true}} (u_{t-1:0} - u_{t-1:t-L}) v_i\|^2 \\ &\quad + 2\|\sum_{i=1}^{T} M_i^{\text{true}} (u_{t-1:0} - u_{t-1:t-L}) v_i\| \|\sum_{i=k+1}^{T} M_i^{\text{true}} u_{t-1:t-L} v_i\| \\ &\quad + \|\sum_{i=k+1}^{T} M_i^{\text{true}} u_{t-1:t-L} v_i\|^2.\end{aligned}$$

Next we note that $\|\sum_{i=k+1}^{T} M_i^{\text{true}} u_{t-1:t-L} v_i\|$ is assumed to be at most $\|C\|\|B\|/T$ and so we now focus on bounding the norm:

$$\|\sum_{i=1}^{T} M_i^{\text{true}} (u_{t-1:0} - u_{t-1:t-L}) v_i\|. \tag{10}$$

Towards bounding Eq. 10, assume $L > \ell_1$ so that

$$\begin{aligned}\sum_{i=1}^{T} M_i^{\text{true}} (u_{t-1:0} - u_{t-1:t-L}) v_i &= \sum_{i=L-\ell_1+1}^{t-\ell_1-1} C A^i h(A) B u_{t-\ell_1-i} \\ &= \sum_{i=L-\ell_1+1}^{t-\ell_1-1} \sum_{j=1}^{d_A} \alpha_j^i h(\alpha_j) C_j B_j^\top u_{t-\ell_1-i}.\end{aligned}$$

Then

$$\begin{aligned}\|\sum_{i=L-\ell_1+1}^{t-\ell_1-1} C A^i h(A) B u_{t-\ell_1-i}\| &\le \max_{j\in[d_A]} \alpha_j^i h(\alpha_j) \sum_{i=L-\ell_1+1}^{t-\ell_1-1} \|C_j B_j^\top u_{t-\ell_1-i}\| \\ &\le \max_{\alpha(A)} \sum_{i=L-\ell_1+1}^{t-\ell_1-1} \alpha^i h(\alpha) \|C\|\|B\|.\end{aligned}$$

Next we have

$$\begin{aligned}\left(\max_{\alpha(A)} \sum_{i=L-\ell_1+1}^{t-\ell_1-1} \alpha^i h(\alpha)\right) &\le h(\alpha)\alpha^{L-\ell_1-1} \sum_{i=0}^{T-L} \alpha^i \\ &= h(\alpha)\alpha^{L-\ell_1-1} \frac{1-\alpha^{T-L+1}}{1-\alpha} \\ &\le T^{-1/4},\end{aligned}$$

where the last inequality holds by assumption. Therefore Eq. 10 is at most

$$\|\sum_{i=1}^{T} M_i^{\text{true}} (u_{t-1:0} - u_{t-1:t-L}) v_i\| \le \|C\|\|B\| T^{-1/4}.$$

Then we have

$$\|y_t - p_t(y_{t-1:1}) - \sum_{i=1}^{k} M_i^{\text{true}} u_{t-1:t-L} v_i\|^2 \le \frac{\|C\|^2\|B\|^2}{T^{1/2}} + 2\frac{\|C\|^2\|B\|^2}{T^{3/4}} + \frac{\|C\|^2\|B\|^2}{T^2} \le 2\frac{\|C\|^2\|B\|^2}{T^{1/2}}.$$

Finally we conclude

$$\begin{aligned}\ell_t(M^{\text{true}} + E_M, L) \;&\leq\; 2\frac{\|C\|^2\|B\|^2}{T^{1/2}} + 2\left(2\frac{\|C\|^2\|B\|^2}{T^{1/2}}\right)^{1/2}\left(k\delta\overset{p}{\log}(T)\right) + \left(k\delta\overset{p}{\log}(T)\right)^2 \\ &\leq\; 4\frac{\|C\|^2\|B\|^2}{T^{1/2}},\end{aligned}$$

where the last inequality holds since we assumed

$$\delta \;\leq\; \frac{1}{4m}\frac{\|C\|^2\|B\|^2}{T^{1/4}\log^p(T)}.$$

□

B. Length Generalization for Vanilla Spectral Filtering

The proof of Theorem 5 ultimately comes from Theorem 7 and its proof in Appendix A. Theorem 7 abstracts the necessary assumptions needed to obtain a length generalization guarantee. In Lemma 12 we prove that Algorithm 1 satisfies these assumptions.

Proof of Theorem 5. By Lemma 12 and the assumptions made in the statement of Theorem 5, we may apply Theorem 7 to Algorithm 1 to get that

$$\sum_{t=1}^{T}\ell_t(M^t, L) - \min_{M^*\in\mathcal{K}_{\|C\|\|B\|}}\sum_{t=1}^{T}\ell_t(M^*, T) \;\leq\; \left(12k^{3/2}\|C\|^2\|B\|^2\log(T) + 8\|C\|^2\|B\|^2\right)\sqrt{T}.$$

□

Lemma 12 (Length Generalization for Vanilla Spectral Filtering)**.** *Recall that in Algorithm 1 we define*

$$\mu_\alpha \overset{def}{=} (\alpha-1)\begin{bmatrix}1 & \alpha & \dots & \alpha^{T-1}\end{bmatrix}^\top \in \mathbb{R}^{T-1}$$

and $H_{T-1} = \int_{\alpha\in[0,1]}\mu_\alpha\mu_\alpha^\top d\alpha$ *and we let* $\phi_1, \dots, \phi_{T-1}$ *be the orthonormal eigenvectors of* H_{T-1} *with eigenvalues* $\sigma_1, \dots, \sigma_{T-1}$*. Algorithm 1 is equivalent to Algorithm 3 with the following:*

(a) $p_t(y_{t-1:1}) = y_{t-1}$

(b) $v_1 = e_1$

(c) $v_i = (0, \sigma_{i-1}^{1/4}\phi_{i-1})$ *for* $i = 2, \dots, T$

Define M^{true} *as follows:*

$$M_1^{true} \overset{def}{=} CB,$$

and for $i \geq 2$

$$M_i^{true} \overset{def}{=} \sum_{n=1}^{d_A}\sigma_{i-1}^{-1/4}\phi_{i-i}^\top\mu_{\alpha_n}(C_nB_n^\top).$$

Then the following properties hold

1. *For* $h(A) = A - I$ *and* $\ell_1 = 1$

$$y_t - p_t(y_{t-1:1}) = \sum_{i=1}^{\ell_1}M_i^{true}u_{t-i} + \sum_{i=1}^{t-\ell_1}CA^ih(A)Bu_{t-\ell_1-i}.$$

2. $y_t - p_t(y_{t-1:1}) = \sum_{i=1}^{T} M_i^{\textit{true}} u_{t-1:1} v_i$.
3. *For* $k = \Omega(\log(T d_A \|C\| \|B\| / \epsilon))$,
$$\| \sum_{i=k+1}^{T} M_i^{\textit{true}} u_{t-1:1} v_i \| \ \leq \ \epsilon / T.$$
4. *For any* $i \in [T]$
$$\|M_i^{\textit{true}}\| \ \leq \ \|C\| \|B\|.$$
5. *For any* $i \in [T]$, $\|v_i\|_1 \ \leq \ \log(T)$ *and* $\{v_i\}_{i \in [T]}$ *are orthonormal.*
6. *Finally if the spectrum of* A *lies in the interval*
$$\left[0, 1 - \frac{\log(T)}{2(L-2)}\right] \cup \left[1 - \frac{1}{2T^{5/4}}, 1\right],$$
then
$$\max_{\alpha(A)} \left\{ |h(\alpha) \alpha^{L-\ell_1 - 1} (1 - \alpha^{T-L+1})(1-\alpha)^{-1}| \right\} \ \leq \ \frac{1}{T^{1/4}}.$$

Proof. Points $(a) - (c)$ are evident by definition of Algorithm 1. Now suppose y_t evolves as an LDS. By definition, there exist matrices (A, B, C, D) such that
$$y_t = \sum_{i=1}^{t} C A^{i-1} B u_{t-i},$$
where we assume $D = I$ and A is diagonal without loss of generality. Let $\alpha_1, \ldots, \alpha_{d_A}$ denote the eigenvalues of A. and let $u_{t:0}$ be the $d_{\text{in}} \times T$ (padded) matrix $u_{t:0} = \begin{bmatrix} u_t & u_{t-1} & \ldots & u_0 & 0 \end{bmatrix}$. Then we have
$$\begin{aligned} y_t - y_{t-1} &= \sum_{i=1}^{t} C A^{i-1} B u_{t-i} - \sum_{i=1}^{t-1} C A^{i-1} B u_{t-1-i} \\ &= C B u_{t-1} + \sum_{i=1}^{t-1} C \left(A^i - A^{i-1}\right) B u_{t-1-i}. \end{aligned}$$
We pause here to note this proves (1). We continue rearranging the equation to finish the derivation of (2).
$$\begin{aligned} y_t - y_{t-1} &= C B u_{t-1} + \sum_{i=1}^{t-1} C \left(A^i - A^{i-1}\right) B u_{t-1-i} \\ &= C B u_{t-1} + \sum_{n=1}^{d_A} C e_n e_n^\top B \sum_{i=1}^{t-1} \left(\alpha_n^i - \alpha_n^{i-1}\right) u_{t-1-i} \\ &= C B u_{t-1} + \sum_{n=1}^{d_A} (C_n B_n^\top) u_{(t-2):0} \mu_{\alpha_j}. \end{aligned}$$
Observe that
$$\sum_{i=1}^{T-1} \phi_i \phi_i^\top = I.$$
Using this we have,
$$\begin{aligned} y_t - y_{t-1} &= C B u_{t-1} + \sum_{n=1}^{d_A} (C_n B_n^\top) u_{(t-2):0} \mu_{\alpha_n} \\ &= C B u_{t-1} + \sum_{n=1}^{d_A} (C_n B_n^\top) u_{(t-2):0} \left(\sum_{i=1}^{T} \phi_i \phi_i^\top \right) \mu_{\alpha_n} \\ &= C B u_{t-1} + \sum_{i=1}^{T} \sum_{n=1}^{d_A} \phi_i^\top \mu_{\alpha_n} (C_n B_n^\top) u_{(t-2):0} \phi_i. \end{aligned}$$

Recalling the definition of M^{true} and $v_i = \sigma_{i-1}^{1/4}\phi_{i-1}$ we therefore have established (2):

$$y_t - y_{t-1} = M_1^{\text{true}} u_{(t-1):0} e_1 + \sum_{i=2}^{T-1} M_i^{\text{true}} u_{(t-1):0} v_i.$$

Next we aim to prove (3). We consider

$$\| \sum_{i=k+1}^{T} M_i^{\text{true}} u_{(t-2):0} v_i \|.$$

By Lemma 13.4 in (Hazan & Singh, 2022) there is some universal constant c' such that,

$$\max_{\alpha \in [0,1]} |\phi_i^\top \mu_\alpha| \ \le \ c' T^2 \exp(-i/\log(T)).$$

So,

$$\begin{aligned}
\|M_i^{\text{true}} u_{(t-2):0} v_i\| &= \| \sum_{n=1}^{d_A} \sigma_{i-1}^{-1/4} \phi_{i-i}^\top \mu_{\alpha_n} (C_n B_n^\top) u_{(t-2):0} \left(\sigma_{i-1}^{1/4} \phi_{i-1}\right) \| \\
&= \| \sum_{n=1}^{d_A} \phi_{i-i}^\top \mu_{\alpha_n} (C_n B_n^\top) u_{(t-2):0} \phi_{i-1} \| \\
&\le \ d_A (c' T^2 \exp(-(i-1)/\log(T))) \|C_n B_n^\top\| \|\phi_{i-1}\|_1 \\
&\le \ c' d_A T^{3/2} \exp(-(i-1)/\log(T))) \|C\| \|B\|.
\end{aligned}$$

Therefore,

$$\| \sum_{i=k+1}^{T} M_i^{\text{true}} u_{(t-1):0} \phi_i \| \ \le \ c' d_A T^{5/2} \exp(-k/\log(T))) \|C\| \|B\|$$

Therefore as long as

$$k \ \ge \ \log(T) \log\left(\frac{T^{5/2} c' d_A \|C\| \|B\|}{\epsilon} \right),$$

then

$$\| \sum_{i=k+1}^{T} M_i^{\text{true}} u_{(t-1):0} \phi_i \| \ \le \ \frac{\epsilon}{T}.$$

Next we note that the proof of (4) that $\|M_i^{\text{true}}\| \ \le \ \|C\| \|B\|$ is proven in Lemma D.1 of (Hazan et al., 2017b). Similarly, the proof of (5) that $\|v_i\|_1 \ \le \ \log(T)$ is proven by Lemma 13 from (Hazan et al., 2017b). Finally we prove (6). Since $h(\alpha) = \alpha - 1$ and $\ell_1 = 1$, we have

$$\max_{\alpha(A)} \left\{ |h(\alpha) \alpha^{L-\ell_1-1} (1 - \alpha^{T-L+1})(1-\alpha)^{-1}| \right\} = \max_{\alpha(A)} \alpha^{L-2} (1 - \alpha^{T-L+1}). \tag{11}$$

To bound Eq. 11, consider the case where α is bounded away from 1. Suppose $\alpha = 1 - \delta$, then

$$(1-\delta)^{L-2} \ \le \ \frac{1}{T^p} \iff \log\left(\frac{1}{1-\delta}\right) \ \ge \ \frac{p \log(T)}{L-2}.$$

Observe that for $\delta \in [0,1]$, $\log(1/(1-\delta)) \ \ge \ \delta/2$. Therefore, if

$$\delta \ \ge \ \frac{2p \log(T)}{L-2},$$

we are guaranteed that $\alpha^{L-2} \ \le \ 1/T^p$. Next consider when α is very close to 1; suppose $\alpha \ \ge \ 1 - \frac{1}{T^p T}$ for $p < 1/2$. Then using that $(1-x)^q \ \ge \ 1 - 2qx$ for $x \in [0,1]$ we have

$$\alpha^{T-L+1} \ \ge \ \left(1 - \frac{1}{T^p T}\right)^{T-L+1} \ \ge \ 1 - 2\frac{T-L+1}{T^p T} \implies 1 - \alpha^{T-L+1} \ \le \ 2\frac{T-L+1}{T^p T} \ \le \ \frac{2}{T^p}.$$

Plugging in $p = 1/4$ we conclude that

$$\alpha^{L-2}(1-\alpha^{T-L+1}) \;\le\; T^{-1/4} \qquad \text{for any } \alpha \in \left[0, 1-\frac{\log(T)}{2(L-2)}\right] \cup \left[1-\frac{1}{2T^{5/4}}, 1\right].$$

□

The following lemma comes from (Hazan et al., 2017b).

Lemma 13 (Hazan, Singh, Zhang)**.** *Let (σ_j, ϕ_j) be the j-th largest eigenvalue-eigenvector pair of the $T \times T$ Hankel matrix. Then,*

$$\|\phi_j\|_1 \;\le\; O\left(\frac{\log(T)}{\sigma_j^{1/4}}\right).$$

C. Length Generalization for Spectral Filtering Using Two Autoregressive Components

The proof of Theorem 6 ultimately comes from Theorem 7 and its proof in Appendix A. Theorem 7 abstracts the necessary assumptions needed to obtain a length generalization guarantee. In Lemma 14 we prove that Algorithm 2 satisfies these assumptions.

Proof of Theorem 6. By Lemma 14 and the assumptions made in the statement of Theorem 6, we may apply Theorem 7 to Algorithm 2 to get that

$$\sum_{t=1}^{T} \ell_t(M^t, L) - \min_{M^* \in \mathcal{K}_{\|C\|\|B\|}} \sum_{t=1}^{T} \ell_t(M^*, T) \;\le\; \left(12k^{3/2}\|C\|^2\|B\|^2 \overset{2}{\log}(T) + 8\|C\|^2\|B\|^2\right)\sqrt{T}.$$

□

Lemma 14 (Length Generalization Using Two Autoregressive Components)**.** *Recall that in Algorithm 2 we define*

$$\tilde{\mu}_{\alpha,T} \overset{def}{=} (\alpha-1)^2 \begin{bmatrix} 1 & \alpha & \dots & \alpha^T \end{bmatrix}^\top \in \mathbb{R}^T$$

and and $N_T = \int_{\alpha\in[0,1]} \tilde{\mu}_{\alpha,T}\tilde{\mu}_{\alpha,T}^\top d\alpha$ and we let $\tilde{\phi}_1, \dots, \tilde{\phi}_{T-2}$ be the orthonormal eigenvectors of N_{T-2} with eigenvalues $\tilde{\sigma}_1, \dots, \tilde{\sigma}_{T-2}$. Algorithm 2 is equivalent to Algorithm 3 with the following:

(a) $p_t(y_{t-1:1}) = 2y_{t-1} - y_{t-2}$

(b) $v_1 = e_1$, $v_2 = e_2$ *and for* $i \;\ge\; 3$, $v_i = (0, 0, \sigma_{i-2}^{1/4}\tilde{\phi}_{i-2})$

Define M^{true} as follows:

$$M_1^{true} \overset{def}{=} CB,$$

$$M_2^{true} \overset{def}{=} C(A-2I)B,$$

and for $i \;\ge\; 3$,

$$M_i^{true} \overset{def}{=} \sum_{n=1}^{d_A} \left(\sigma_i^{-1/4}\tilde{\phi}_i^\top \tilde{\mu}_{\alpha_n}\right)(C_n B_n^\top).$$

Then the following properties hold

1. *For $h(A) = (A-I)^2$ and $\ell_1 = 2$*

$$y_t - p_t(y_{t-1:1}) = \sum_{i=1}^{\ell_1} M_i^{true} u_{t-i} + \sum_{i=1}^{t-\ell_1} CA^i h(A) B u_{t-\ell_1-i}.$$

2. $y_t - p_t(y_{t-1:1}) = \sum_{i=1}^{T} M_i^{true} u_{t-1:1} v_i$.
3. *For* $k = \Omega(\log(T d_A \|C\| \|B\| / \epsilon))$,
$$\| \sum_{i=k+1}^{T} M_i^{true} u_{t-1:1} v_i \| \ \leq \ \epsilon / T.$$
4. *For any* $i \in [T]$
$$\|M_i^{true}\| \ \leq \ \|C\| \|B\|.$$
5. *For any* $i \in [T]$, $\|v_i\|_1 \ \leq \ \log(T)$ *and* $\{v_i\}_{i \in [T]}$ *are orthonormal.*
6. *Finally if the spectrum of* A *lies in the interval*
$$\left[0, 1 - \frac{\log(T)}{2(L-2)}\right] \cup \left[1 - \frac{1}{2T^{1/4}}, 1\right],$$
then
$$\max_{\alpha(A)} \left\{ |h(\alpha) \alpha^{L-\ell_1-1} (1 - \alpha^{T-L+1})(1-\alpha)^{-1}| \right\} \ \leq \ \frac{1}{T^{1/4}}.$$

Proof. Suppose y_t evolves as an LDS. By definition, there exist matrices (A, B, C, D) such that

$$y_t = \sum_{i=1}^{t} C A^{i-1} B u_{t-i},$$

where we assume $D = I$ and A is diagonal without loss of generality. Let $\alpha_1, \ldots, \alpha_{d_A}$ denote the eigenvalues of A. and let $u_{t:0}$ be the $d_{\text{in}} \times T$ (padded) matrix $u_{t:0} = \begin{bmatrix} u_t & u_{t-1} & \ldots & u_0 & 0 \end{bmatrix}$. Then we have (1):

$$y_t - 2y_{t-1} + y_{t-2} = CBu_{t-1} + C(A - 2I)Bu_{t-2} + \sum_{i=0}^{t-3} CA^i(A^2 - 2A + I)Bu_{t-3-i}.$$

Let $\alpha_1, \ldots, \alpha_{d_A}$ denote the eigenvalues of A. We observe the following equality:

$$\begin{aligned}
\sum_{i=0}^{t-3} CA^i(A^2 - 2A + I)Bu_{t-3-i} &= \sum_{i=0}^{t-3} C \sum_{n=1}^{d_A} \alpha_n^i (\alpha_n - 1)^2 e_n e_n^\top B u_{t-3-i} \\
&= \sum_{n=1}^{d_A} \left(C e_n e_n^\top B\right) \sum_{i=0}^{t-3} \alpha_n^i (\alpha_n - 1)^2 u_{t-3-i} \\
&= \sum_{n=1}^{d_A} \left(C_n B_n^\top\right) u_{(t-3):0} \tilde{\mu}_{\alpha_n}.
\end{aligned}$$

Observe that

$$\sum_{i=1}^{T-2} \tilde{\phi}_i \tilde{\phi}_i^\top = I.$$

Using this we have,

$$\begin{aligned}
\sum_{i=0}^{t-3} CA^i(A^2 - 2A + I)Bu_{t-3-i} &= \sum_{n=1}^{d_A} \left(C_n B_n^\top\right) u_{(t-3):0} \tilde{\mu}_{\alpha_n} \\
&= \sum_{n=1}^{d_A} \left(C_n B_n^\top\right) u_{(t-3):0} \left(\sum_{i=1}^{T-2} \tilde{\phi}_i \tilde{\phi}_i^\top \right) \tilde{\mu}_{\alpha_n} \\
&= \sum_{i=1}^{T-2} \left(\sum_{n=1}^{d_A} \tilde{\phi}_i^\top \tilde{\mu}_{\alpha_n} \left(C_n B_n^\top\right) \right) u_{(t-3):0} \tilde{\phi}_i \\
&= \sum_{\ell=3}^{T} M_\ell^{\text{true}} u_{(t-1):0} v_\ell.
\end{aligned}$$

Therefore we have established (2). Next we aim to prove (3). We consider

$$\| \sum_{i=k+1}^{T} M_i^{\text{true}} u_{(t-1):0} v_i \|.$$

Combining Lemma 15 and Lemma 16 gives us that there is some constant c' such that,

$$\max_{\alpha \in [0,1]} |\tilde{\phi}_i^\top \tilde{\mu}_\alpha| \ \leq \ c' \exp(-i/4 \log(T)).$$

So,

$$\begin{aligned}
\|M_i^{\text{true}} u_{(t-1):0} v_i\| &= \| \sum_{n=1}^{d_A} \sigma_{i-1}^{-1/4} \tilde{\phi}_{i-i}^\top \tilde{\mu}_{\alpha_n} (C_n B_n^\top) u_{(t-1):0} \left(\sigma_{i-1}^{1/4} \tilde{\phi}_{i-1} \right) \| \\
&= \| \sum_{n=1}^{d_A} \tilde{\phi}_{i-i}^\top \tilde{\mu}_{\alpha_n} (C_n B_n^\top) u_{(t-2):0} \tilde{\phi}_{i-1} \| \\
&\leq \ d_A \exp(-(i-1)/4 \log(T)) \|C_n B_n^\top\| \|\phi_{i-1}\|_1 \\
&\leq \ c' d_A \sqrt{T} \exp(-(i-1)/4 \log(T)) \|C\| \|B\|.
\end{aligned}$$

Therefore,

$$\| \sum_{i=k+1}^{T} M_i^{\text{true}} u_{(t-1):0} v_i \| \ \leq \ c' d_A T^{3/2} \exp(-i/4 \log(T)) \|C\| \|B\|.$$

Therefore as long as

$$k \ \geq \ 4 \log(T) \log \left(\frac{T^{3/2} c' d_A \|C\| \|B\|}{\epsilon} \right),$$

then

$$\| \sum_{i=k+1}^{T} M_i^{\text{true}} u_{(t-1):0} v_i \| \ \leq \ \frac{\epsilon}{T}.$$

To prove (4) we note that the statement is obvious for $i \ \leq \ 2$. For $i \ \geq \ 3$ the proof from Lemma D.1 of (Hazan et al., 2017b) directly applies due to Lemma 15. Next, Lemma 17 proves (5). Finally we prove (6). Next, Lemma 17 proves (5). Finally we prove (6). Since we have $h(\alpha) = (\alpha - 1)^2$ and $\ell = 2$,

$$\max_{\alpha(A)} \left\{ |h(\alpha) \alpha^{L-3} (1 - \alpha^{T-L+1}) (1 - \alpha)^{-1}| \right\} = \max_{\alpha(A)} \left\{ (1 - \alpha) \alpha^{L-3} (1 - \alpha^{T-L+1}) \right\}. \tag{12}$$

To bound Eq. 12, consider the case where α is bounded away from 1. Suppose $\alpha = 1 - \delta$, then

$$(1 - \delta)^{L-3} \ \leq \ \frac{1}{T^p} \iff \log \left(\frac{1}{1 - \delta} \right) \ \geq \ \frac{p \log(T)}{L - 3}.$$

Observe that for $\delta \in [0, 1]$, $\log(1/(1 - \delta)) \ \geq \ \delta/2$. Therefore, if

$$\delta \ \geq \ \frac{2p \log(T)}{L - 3},$$

we are guaranteed that $\alpha^{L-3} \ \leq \ 1/T^p$. Next consider when α is very close to 1. To ensure that Eq. 12 is bounded by $1/T^p$ we only require

$$\alpha \ \geq \ 1 - \frac{1}{T^p}.$$

Plugging in $p = 1/4$, we conclude that Eq. 12 is bounded by $T^{-1/4}$ if

$$\alpha_n \in \left[0, 1 - \frac{\log(T)}{2(L-3)} \right] \cup \left[1 - \frac{1}{T^{1/4}}, 1 \right] \text{ for all } n \in [d_A].$$

□

C.1. Properties of the Hankel Matrix for Two Autoregressive Terms

In Algorithm 2 we define

$$\tilde{\mu}_\alpha \stackrel{\text{def}}{=} (\alpha-1)^2 \begin{bmatrix} 1 & \alpha & \dots & \alpha^T \end{bmatrix}^\top \in \mathbb{R}^T$$

and

$$N_T = \int_{\alpha\in[0,1]} \tilde{\mu}_\alpha \tilde{\mu}_\alpha^\top d\alpha.$$

In what follows we present and prove several lemmas needed for the proof of Theorem 6.

Lemma 15 (Properties of N_T)**.** *For any $\alpha \in [0,1]$ and $1 \le i \le T$,*

$$\max_{\alpha\in[0,1]} |\phi_i^\top \tilde{\mu}_\alpha| \le 6^{1/4}\sigma_i^{1/4}.$$

Proof. We have

$$\begin{aligned}\int_{\alpha\in[0,1]} \left(\phi_i^\top \tilde{\mu}_\alpha\right)^2 d\alpha &= \phi_i^\top \left(\int_{\alpha\in[0,1]} \tilde{\mu}_\alpha \tilde{\mu}_\alpha^\top d\alpha\right) \phi_i \\ &= \phi_i^\top N_T \phi_i = \sigma_i.\end{aligned}$$

Next we observe that for $f_w(\alpha) \stackrel{\text{def}}{=} \left(w^\top \tilde{\mu}_\alpha\right)^2$, where w is any unit-norm vector, we have that f_w is 6-Lipschitz on $[0,1]$. Indeed,

$$\begin{aligned}f_w'(\alpha) &= \frac{d}{d\alpha}(\alpha-1)^4 \left(\sum_{i=1}^T w_i \alpha^{i-1}\right)^2 \\ &= 2(\alpha-1)^4 \left(\sum_{i=1}^T w_i \alpha^{i-1}\right)\left(\sum_{i=2}^T (i-1) w_i \alpha^{i-2}\right) + 4\left(\sum_{i=1}^T w_i \alpha^{i-1}\right)^2 (\alpha-1)^3 \\ &\le 2(\alpha-1)^4 \left(\frac{1-\alpha^T}{1-\alpha}\right)\left(\sum_{i=1}^{T-1} i\alpha^{i-1}\right) + 4\left(\frac{1-\alpha^T}{1-\alpha}\right)^2 (\alpha-1)^3 \\ &= 2(\alpha-1)^4 \left(\frac{1-\alpha^T}{1-\alpha}\right)\left(\frac{1-T\alpha^{T-1}+(T-1)\alpha^T}{(1-\alpha)^2}\right) + 4\left(\frac{1-\alpha^T}{1-\alpha}\right)^2 (\alpha-1)^3 \\ &= 2\left(1-\alpha^T\right)\left(1-T\alpha^{T-1}+(T-1)\alpha^T\right) + 4\left(1-\alpha^T\right)^2(\alpha-1) \\ &\le 2+4=6.\end{aligned}$$

Consider any non-negative L-Lipschitz function f that reaches some maximum value $g_{\max}$ over $[0,1]$. The function f which satisfies L-Lipschitzness, attains $g_{\max}(f)$ and also has minimum possible area $A(f) \stackrel{\text{def}}{=} \int_{\alpha\in[0,1]} f(\alpha)d\alpha$ is

$$\begin{aligned}f^*(\alpha) &= \begin{cases} L\alpha, & \text{for } \alpha \in [0,\alpha^*] \\ \max\left\{g_{\max} - L(\alpha-\alpha^*), 0\right\}, & \text{for } \alpha \in [\alpha^*,1] \end{cases} \\ &= \begin{cases} L\alpha, & \text{for } \alpha \in [0,\alpha^*] \\ g_{\max} - L(\alpha-\alpha^*), & \text{for } \alpha \in [\alpha^*, \alpha^* + \frac{g_{\max}}{L}] \\ 0, & \text{for } \alpha \in [\alpha^* + \frac{g_{\max}}{L}, 1] \end{cases}.\end{aligned}$$

Indeed, any oscillation away from this piecewise linear function would either increase the total area or violate the Lipschitz constraint. For this to be a valid construction we must have $L\alpha^* = g_{\max}$ and therefore the minimum corresponding area is

$$A(f^*) = \int_{\alpha\in[0,1]} f^*(\alpha)d\alpha = \frac{1}{2}(\alpha^*)(L\alpha^*) + \frac{1}{2}(g_{\max}/L)g_{\max} = \frac{g_{\max}^2}{L}.$$

And therefore for any function f we have $g_{\max}(f) \leq \sqrt{LA(f)}$. Using this for $f_{\phi_i}(\alpha)$ we have

$$\max_{\alpha\in[0,1]} f_{\phi_i}(\alpha) = \max_{\alpha\in[0,1]} (\phi_i^\top \tilde{\mu}_\alpha)^2 \leq \sqrt{6\int_{\alpha\in[0,1]} \left(\phi_i^\top \tilde{\mu}_\alpha\right)^2 d\alpha} = \sqrt{6\sigma_i}.$$

We conclude by noting

$$\max_{\alpha\in[0,1]} |\phi_i^\top \tilde{\mu}_\alpha| = \sqrt{\max_{\alpha\in[0,1]} (\phi_i^\top \tilde{\mu}_\alpha)^2} \leq 6^{1/4}\sigma_i^{1/4}.$$

□

Lemma 16 (Adapted from Lemma $E.2$ from (Hazan et al., 2017b))**.** *Let σ_j be the j-th top singular value of N_T. Then for all $T \geq 10$ we have*

$$\sigma_j \leq \min\left(\frac{3}{2}, K\cdot c^{-j/\log(T)}\right),$$

where $c = e^{\pi^2/4} \approx 11.79$ and $K \leq 10^6$ is an absolute constant.

Proof. The proof provided in (Hazan et al., 2017b) applies directly to N_T with only one necessary modification to bound the trace. Observe that we have

$$\begin{aligned}(N_T)_{ij} &= \int_{\alpha\in[0,1]} (\alpha-1)^4\alpha^{i+j-2}d\alpha \\ &= \int_{\alpha\in[0,1]} \alpha^{i+j} - 2\alpha^{i+j-1} + \alpha^{i+j-2}d\alpha \\ &= \frac{24}{(i+j-1)(i+j)(i+j+1)(i+j+2)(i+j+3)}.\end{aligned}$$

Therefore,

$$\sigma_j \leq \operatorname{tr}(N_T) = \sum_{i=1}^{T} \frac{24}{(2i-1)(2i)(2i+1)(2i+2)(2i+3)} \leq \sum_{i=1}^{T} \frac{24}{(2i)^5} = \frac{3}{4}\sum_{i=1}^{T}\frac{1}{i^5} < \frac{3}{2}.$$

The remainder of the proof is an exact copy of the proof of Lemma $E.2$ with $3/4$ replaced by $3/2$. □

Lemma 17 (Controlling the ℓ_1 norm of the filters)**.** *Let (σ_j, ϕ_j) be the j-th largest eigenvalue-eigenvector pair of N_T. Then for $T \geq 4$,*

$$\|\phi_j\|_1 \leq O\left(\frac{\log T}{\sigma_j^{1/4}}\right).$$

Proof. This proof is a copy from the proof of Lemma E.5 in (Hazan et al., 2017b) with only one noted modification. We note that E as defined in their proof is entrywise bounded (for $T \geq 4$) by $24/T^5 \leq 2/T^3$ (which is the stated bound they use for their matrix of interest). We also must show the base case is true for $T_0 = 4$ instead of $T_0 = 2$. We have

$$\|N_4^{1/4}\|_{2\to1} = \sup_{x:\|x\|_2\leq 1} \|N_4^{1/4}x\|_1 \leq \sum_{i,j=1}^{4} |\left(N_4^{1/4}\right)_{ij}| < 2.$$

We note that a tighter result is actually true for N_T in that $\|\phi_j\|_1 \leq O\left(\frac{\log T}{\sigma_j^{1/8}}\right)$. However, we omit this statement and proof because we don't leverage it for a tighter result overall.

□