Learning Hidden Markov Models Using Conditional Samples

Sham Kakade SHAM@SEAS.HARVARD.EDU
Harvard University

Akshay Krishnamurthy AKSHAY@CS.UMASS.EDU
Microsoft Research NYC

Gaurav Mahajan GAURAV.MAHAJAN@YALE.EDU
Yale University

Cyril Zhang CYRILZHANG@MICROSOFT.COM
Microsoft Research NYC

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

This paper is concerned with the computational and statistical complexity of learning the Hidden Markov model (HMM). Although HMMs are some of the most widely used tools in sequential and time series modeling, they are cryptographically hard to learn in the standard setting where one has access to i.i.d. samples of observation sequences. In this paper, we depart from this setup and consider an *interactive access model*, in which the algorithm can query for samples from the *conditional distributions* of the HMMs. We show that interactive access to the HMM enables computationally efficient learning algorithms, thereby bypassing cryptographic hardness.

Specifically, we obtain efficient algorithms for learning HMMs in two settings:

1. An easier setting where we have query access to the exact conditional probabilities. Here our algorithm runs in polynomial time and makes polynomially many queries to approximate any HMM in total variation distance.

2. A harder setting where we can only obtain samples from the conditional distributions. Here the performance of the algorithm depends on a new parameter, called the fidelity of the HMM. We show that this captures cryptographically hard instances and previously known positive results.

We also show that these results extend to a broader class of distributions with latent low rank structure. Our algorithms can be viewed as generalizations and robustifications of Angluin's L^* algorithm for learning deterministic finite automata from membership queries.

1. Introduction

Hidden Markov Models (HMMs) are among the most fundamental tools for modeling temporal and sequential phenomena. These probabilistic models specify a joint distribution over a sequence of observations generated via a Markov chain of latent states. This structure enjoys the simultaneous benefits of low description complexity, sufficient expressivity to capture long-range dependencies, and efficient inference algorithms. For these reasons, HMMs have become ubiquitous building blocks for sequence modeling in varied fields, ranging from bioinformatics to natural language processing to finance. A long-standing challenge, in both theory and practice, is the computational difficulty of learning an unknown HMM from samples. In this paper, we are interested in the computational complexity of this estimation/learning task.

Although one can consider several notions of learnability, we focus on distribution learning, in total variation (TV) distance. In the standard realizable formulation, we are given observation sequences generated by an underlying HMM and are asked to efficiently compute a distribution that is close to the HMM in TV distance. Maximum likelihood estimation is known to be statistically efficient, and therefore the central challenge is computational. Indeed, HMMs can encode the parity with noise problem (Mossel and Roch, 2005), which is widely believed to be computationally hard (Blum et al., 1994b; Kearns et al., 1994; Alekhnovich, 2003). Recent works have therefore focused on obtaining computationally efficient algorithms under structural assumptions which evade these hard instances (Cryan et al., 2001; Hsu et al., 2012; Kontorovich et al., 2013; Weiss and Nadler, 2015; Huang et al., 2015; Sharan et al., 2017).

This work takes a different perspective. We ask: can we evade computational hardness by allowing the learner to access the HMM *interactively*? Specifically we consider a *conditional sampling oracle*: we allow the learner to sample a "future sequence" from the HMM conditioned on a "past sequence" or history. This approach is closely related to recent work in distribution testing (e.g., Chakraborty et al., 2013; Canonne and Rubinfeld, 2014; Canonne et al., 2015; Bhattacharyya and Chakraborty, 2018; Chen et al., 2021), which demonstrates improvements in various property testing tasks via conditional sampling. One conceptual difference is that we use conditional sampling to evade computational hardness, rather than obtaining statistical improvements.

From a practical perspective, we are motivated by potential applications of interactive learning to training language models or world models more generally. Indeed, it is quite natural to fine-tune a language model by asking annotators to complete prompts generated by the model; this precisely corresponds to conditional sampling if we view the annotators as representative of the population (Zhang et al., 2022). When training world models for decision making, it may be possible to request expert demonstrations starting from a particular state, which again is effectively conditional sampling. This latter approach is closely related to interactive imitation learning (Ross et al., 2011).

We are further motivated by two theoretical considerations. First, it is not hard to show that parity with noise can be efficiently learned in this model, as we can sample the label conditioned on each history with a single observation set to 1 and naively denoise these samples (we describe this in more detail in Appendix C). However, this approach is quite tailored to noisy parity, and so it is natural to ask if it can be generalized to arbitrary HMMs. Second, learning HMMs with conditional samples can be seen as a statistical generalization of learning deterministic finite automata (DFAs) with membership queries, for which Angluin's seminal L^* algorithm provides a strong computational separation between interactive and non-interactive PAC learning (Angluin, 1987). We believe it is natural to ask if L^* can be extended to HMMs and be made robust to sampling, thereby providing further evidence for the computational benefits of interactive learning.

Contributions. In this paper, we develop new algorithms and techniques for learning Hidden Markov models when provided with interactive access. As our first result, we show how a generalization of Angluin's L^* algorithm can efficiently learn any HMM in the stronger access model where the learner can query for exact conditional probabilities. As our main result, we consider the more natural conditional sampling access model and obtain an algorithm that is efficient for all HMMs with "high fidelity" a new property we introduce. We show that this property captures the cryptographically hard instances and the prior positive results, but we leave open the question of efficiently learning all HMMs via conditional sampling. Our results require a number of new algorithmic ideas and analysis techniques, most important among them: an efficient representation for distributions over exponentially large domains and a new perturbation argument for mitigating error amplification over long sequences. We hope these techniques find application in other settings.

1.1. Preliminaries

Notation. Let $\mathcal{O} := \{1, \ldots, O\}$ denote a finite observation space and let $\mathcal{O}^t$, $\mathcal{O}^{\leq t}$ and $\mathcal{O}^*$ denote observation sequences of length t, observation sequences of length $\leq t$ and observation sequences of arbitrary length respectively. We consider a distribution $\Pr[\cdot]$ over T random variables $\mathbf{x}_1, \ldots, \mathbf{x}_T$ with a sequential ordering, and we use $x_t \in \mathcal{O}$ to denote the value taken by the t^{th} random variable. For convenience, we often simply write $\Pr[x_1, x_2, \ldots, x_T]$ in lieu of $\Pr[\mathbf{x}_1{=}x_1, \ldots, \mathbf{x}_T{=}x_T]$, omitting explicit reference to the random variables themselves.

When considering conditionals of this distribution, we *always* condition on assignment to a prefix of the random variables and marginalize out a suffix. For example, we consider conditionals of the form $\Pr[\mathbf{x}_{t+1}{=}x_{t+1}, \ldots, \mathbf{x}_{t+k}{=}x_{t+k} | \mathbf{x}_1{=}x_1, \ldots, \mathbf{x}_t{=}x_t]$, and we will write this as $\Pr[x_{t+1}, \ldots, x_{t+k} | x_1, \ldots, x_t]$. Similarly, when considering tuples $f := (x'_1, \ldots, x'_k) \in \mathcal{O}^k$ and $h := (x_1, \ldots, x_t) \in \mathcal{O}^t$, we write $\Pr[\mathbf{x}_{t+1}{=}x'_1, \ldots, \mathbf{x}_{t+k}{=}x'_k | \mathbf{x}_1{=}x_1, \ldots, \mathbf{x}_t{=}x_t]$ as $\Pr[f|h]$, noting that the random variables assigned to f are determined by the length of h.

We lift this conditioning notation to sets of observation sequences in the following manner. If $F := \{f_1, f_2, \ldots\}$ and $H := \{h_1, h_2, \ldots\}$ where each $f_i, h_j \in \mathcal{O}^*$, we write $\Pr[F|H]$ to denote the $|F| \times |H|$ matrix whose $(i,j)^{\text{th}}$ entry is $\Pr[f_i|h_j]$. We allow the sequences in F and H to have different lengths, but always ensure that $\text{len}(f_i) + \text{len}(h_j) \leq T$ so that this matrix is well-defined. We refer to rows and columns of this matrix as $\Pr[f|H]$ and $\Pr[F|h]$ respectively.[1]

Lastly, for $h = (x_1, \ldots, x_t)$ we use $ho = (x_1, \ldots, x_t, o)$ to denote concatenation, and we lift this notation to sequences and sets. For instance, if $H = \{h_1, h_2, \ldots\}$ then $Ho = \{h_1 o, h_2 o, \ldots\}$.

1.1.1. HIDDEN MARKOV MODELS AND LOW RANK DISTRIBUTIONS

Hidden Markov Models provide a low-complexity parametrization for distributions over observation sequences. These models are defined formally as follows.

Definition 1 (Hidden Markov Models) *Let $\mathcal{S} := \{1, \ldots, S\}$. An HMM with $S \in \mathbb{N}$ hidden states is specified by (1) an initial distribution $\mu \in \Delta(\mathcal{S})$, (2) an emission matrix $\mathbb{O} \in \mathbb{R}^{O \times S}$, and (3) a state transition matrix $\mathbb{T} \in \mathbb{R}^{S \times S}$, and defines a distribution over sequences of length T via:*

$$\Pr[x_1, \ldots, x_T] := \sum_{s_1, \ldots, s_{T+1} \in \mathcal{S}^{T+1}} \mu(s_1) \prod_{t=1}^{T} \mathbb{O}[x_t, s_t] \mathbb{T}[s_{t+1}, s_t]. \tag{1}$$

Here $M[i,j]$ represents the $(i,j)^{\text{th}}$ entry of a matrix M.

As the name suggests, HMMs parameterize the distribution with a Markov chain over a hidden state sequence along with an emission function that generates observations. While this specific model is particularly natural, our analysis only leverages a certain low rank structure present in HMMs. To highlight the importance of this structure, we define the *rank* of a distribution.

[1] We always refer to rows, columns, and entries of these matrices in this manner, so no confusion arises when constructing these matrices from (unordered) sets of sequences.

Definition 2 (Rank of a distribution) *We say distribution* $\Pr[\cdot]$ *over observation sequences of length* T *has rank* r *if, for each* $t \in [T]$*, the conditional probability matrix* $\Pr[\mathcal{O}^{\leq T-t} | \mathcal{O}^t]$ *has rank at most* r*.*

It is not hard to verify that an HMM with S hidden states has rank at most S, using the fact that the hidden states form a Markov chain.[2] More generally, the rank identifies a low dimensional structure in the distribution: we have exponentially many vectors $\Pr[\mathcal{O}^{\leq T-t}|h]$, one for each history h, in an r-dimensional subspace of an exponentially larger ambient space. Thus, we are interested in algorithms that exploit the low dimensional structure and admit statistical and computational guarantees scaling polynomially with the rank.

1.1.2. LEARNING MODELS

To circumvent computational hardness, we allow the learner to access conditional distributions of the underlying distribution $\Pr[\cdot]$. We specifically consider two different access models formalized with the following two oracles:

Definition 3 (Exact conditional probability oracle) *The exact conditional probability oracle is given as input: observation sequences* h *and* f *of length* $t \leq T$ *and* $T - t$ *respectively, chosen by the algorithm, and returns the scalar* $\Pr[f|h]$*.*

Definition 4 (Conditional sampling oracle) *The conditional sampling oracle is given as input: an observation sequence* h *of length* $t \leq T$*, chosen by the algorithm, and returns an observation sequence* f *of length* $T - t$ *such that the probability that* f *is returned is* $\Pr[f|h]$*, independently of all other randomness.*

When considering the exact probability oracle, we also allow the learner to obtain independent samples from the joint distribution $\Pr[\cdot]$. Note that this oracle equivalently provides access to exact (unconditional) probabilities of length T sequences. We view this as a noiseless analog of the conditional sampling oracle, which is the main model of interest. This is analogous to noiseless oracles in distribution testing literature (e.g., Canonne and Rubinfeld, 2014).

As a learning goal, we consider distribution learning in total variation distance as studied in prior works (Kearns et al., 1994; Mossel and Roch, 2005; Hsu et al., 2012; Anandkumar et al., 2014). Given access to a target distribution $\Pr[\cdot]$ we want to efficiently compute an estimate $\widehat{\Pr}[\cdot]$ that is close in total variation distance. Formally, we want an algorithm that, when given parameters $\varepsilon, \delta > 0$, computes an estimate $\widehat{\Pr}[\cdot]$ such that with probability at least $1 - \delta$ we have

$$\mathrm{TV}(\Pr, \widehat{\Pr}) := \frac{1}{2} \sum_{x_1,\ldots,x_T \in \mathcal{O}^T} \left| \Pr[x_1, \ldots, x_T] - \widehat{\Pr}[x_1, \ldots, x_T] \right| \leq \varepsilon$$

The algorithm is efficient if its computational complexity (and hence number of oracle calls) scale polynomially in $r, T, O, 1/\varepsilon$ and $\log(1/\delta)$. As the support of $\Pr[\cdot]$ is exponentially large in T, it is not possible to write down all $\mathcal{O}^T$ values of $\widehat{\Pr}$ efficiently. Instead, the goal is to return an efficient representation from which we can evaluate $\widehat{\Pr}[x_1, \ldots, x_T]$ for any sequence $x_1, \ldots, x_T$ efficiently.

[2] In fact the rank of the HMM can be much smaller, since the decomposition alluded to above realizes the non-negative rank of the matrix, which can be exponentially larger than the rank.

1.2. Our results

Our first result studies the computational power provided by the exact probability oracle (Definition 3). We show how a generalization of Angluin's L^* algorithm can efficiently learn any HMM given access to this oracle. The result is summarized in the following theorem:[3]

Theorem 1 (Learning with exact conditional probabilities) *Assume $\mathcal{O} = \{0, 1\}$. Let $\Pr[\cdot]$ be any rank r distribution over observation sequences of length T. Pick any $0 < \varepsilon, \delta < 1$. Then Algorithm 1 with access to an exact probability oracle and samples from $\Pr[\cdot]$, runs in $\mathrm{poly}(r, T, 1/\varepsilon, \log(1/\delta))$ time and returns an efficiently represented approximation $\widehat{\Pr}[\cdot]$ satisfying $\mathrm{TV}(\Pr, \widehat{\Pr}) \leq \varepsilon$ with probability at least $1 - \delta$.*

The main technical challenge is finding a succinct and observable parametrization of the distribution, so that we can infer all conditional distributions using polynomially many queries. This observable parameterization plays a central role in our main result, and in this sense Theorem 1 can be seen as an insightful warmup.

Our main contribution is in extending these results to the more natural interactive setting where the learner only accesses conditional samples via the oracle in Definition 4. Our algorithm here can be viewed as a robust version of L^*, and we obtain the following guarantee:

Theorem 2 (Learning with conditional samples) *Let $\Pr[\cdot]$ be any rank r distribution over observation sequences of length T. Assume distribution $\Pr[\cdot]$ has fidelity Δ^*. Pick any $0 < \varepsilon, \delta < 1$. Then Algorithm 2 with access to a conditional sampling oracle runs in $\mathrm{poly}(r, T, O, 1/\Delta^*, 1/\varepsilon, \log(1/\delta))$ time and returns an efficiently represented approximation $\widehat{\Pr}[\cdot]$ satisfying $\mathrm{TV}(\Pr, \widehat{\Pr}) \leq \varepsilon$ with probability at least $1 - \delta$.*

The theorem provides a robust analog to Theorem 1 in the much weaker conditional sampling access model. The caveat is that the guarantee depends on a spectral property of a distribution, which we call the fidelity. The definition of fidelity (Definition 6) requires further development of the algebraic structure in $\Pr[\cdot]$ and is deferred to Section 2. Nevertheless, we can show that the cryptographically hard examples of HMMs and positive results from prior work on learning HMMs have favorable fidelity parameters and thus are efficiently learnable by our algorithm (see Appendix C). On the other hand, there are HMMs with exponentially small fidelity parameter, and we have no evidence that these instances are computationally intractable when provided with conditional samples. This leads to the main open question stemming from our work.

Open Problem 1 *Is there a computationally efficient algorithm for learning* any *low rank distribution given access to a conditional sampling oracle?*

We discuss this problem in more detail in Section 3 after introducing our techniques in Section 2.

2. Technical overview

To explain the central challenges with learning low rank distributions and how we overcome them, let us introduce the following notation: let $H_t := \mathcal{O}^t$ and $F_t := \mathcal{O}^{T-t}$ denote the observation sequences of length t and $T - t$ respectively. Then the matrix $\Pr[F_t | H_t]$ is a submatrix of

[3] As this result is a warmup for our main result, we focus on the setting where $\mathcal{O} = \{0, 1\}$ for simplicity.

$\Pr[\mathcal{O}^{\leq T-t}|\mathcal{O}^t]$ and hence is rank at most r by assumption. If we define these matrices for each length $t \in [T]$, then clearly we have encoded the entire distribution. Hence, estimating these matrices in an appropriate sense would suffice for distribution learning. Although the matrices all have rank at most r, they are exponentially large, so the low rank property does not immediately yield an efficient representation of the distribution. Indeed, we must leverage further structure to obtain efficient algorithms.

2.1. Background: Observable operators and hard instances

For HMMs, we can hope to leverage the explicit formula for the probability of a sequence (Equation (1)) to obtain an efficient algorithm. Indeed, this is the approach adopted by Hsu, Kakade, and Zhang (Hsu et al., 2012). Specifically, they use the *observable operator* representation (Jaeger, 2000): if we define $S \times S$ matrices $\{\mathbb{A}_o\}_{o\in\mathcal{O}}$ as $\mathbb{A}_o := \mathbb{T}\mathrm{diag}(\mathbb{O}[o,\cdot])$ then we can write the probability of any observation sequence as

$$\Pr[x_1,\dots,x_T] = \mathbf{1}^\top \mathbb{A}_{x_T}\dots\mathbb{A}_{x_1}\mu,$$

where $\mathbf{1}$ is the all-ones vector and recall that μ is the initial state distribution. Hsu, Kakade and Zhang show that these operators can be estimated, up to a linear transformation, whenever $\mathbb{T}$ and $\mathbb{O}$ have full column rank. In fact, under their assumptions, these operators can be recovered from $\Pr[\mathbf{x}_1{=}\cdot, \mathbf{x}_2{=}\cdot, \mathbf{x}_3{=}\cdot]$ alone; no higher order moments of the distribution are required.

Unfortunately, this approach fails if either $\mathbb{T}$ or $\mathbb{O}$ are (column) rank deficient, and it is conjectured that the rank deficient HMMs are precisely the hard instances (Mossel and Roch, 2005). On the other hand, many interesting HMMs *are* rank deficient. For example, any *overcomplete* HMM—one with fewer observations than states—cannot have a full column rank $\mathbb{O}$ matrix. This captures all deterministic finite automata where the alphabet size is smaller than the number of states as well as the parity with noise problem.

Learning parity with noise is a particularly interesting case. The standard formulation is that we obtain samples of the form $(\mathbf{z},\mathbf{y}) \in \{0,1\}^{T-1}\times\{0,1\}$ where $\mathbf{z}$ is uniformly distributed on the hypercube and $\mathbf{y} = \bigoplus_{i\in I}\mathbf{z}_i$ with probability $1-\alpha$ and $\mathbf{y} = 1 - \bigoplus_{i\in I}\mathbf{z}_i$ with the remaining probability. Here $\bigoplus$ denotes the parity operation, I is a secret subset of indices $I \subseteq [T-1]$, and $\alpha \in (0, 1/2)$ is a noise parameter. We want to learn the subset I, given samples from this process. This problem is widely believed to be computationally hard and can be encoded as an HMM with $\mathcal{O} = \{0,1\}$ and $4T$ states (see Appendix C). This HMM exhibits two particularly challenging features. First, many states have identical observation distributions, or are *aliased*; characterizing the learnability (as well as basic structural properties) of aliased HMMs remain long-standing open problems (Weiss and Nadler, 2015). Second, it is quite apparent that low degree moments, like those used by Hsu, Kakade, and Zhang, reveal no information about the subset I. In particular, the observable operators $\mathbb{A}_o$ are not identifiable from low degree moments. One must use higher order information, i.e., statistics about long sequences, to solve this problem.

2.2. Efficient representation

For rank deficient HMMs, it is not clear how to identify the observable operators and it is not even clear that such operators exist for the more general case of low rank distributions. So, we must return to the question of how to efficiently represent the distribution. Here, our first observation is that any submatrix of $\Pr[F_t|H_t]$ that has the same rank as the entire matrix can be used to build

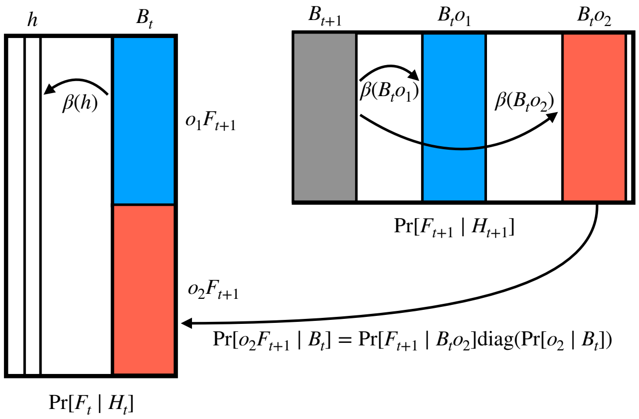

Figure 1: Schematic of the circulant structure relating the $\Pr[F_t|H_t]$ and $\Pr[F_{t+1}|H_{t+1}]$ matrices. Columns of $\Pr[F_t \mid H_t]$ can be represented linearly in basis B_t using coefficients $\beta(\cdot)$. The blocks $\Pr[oF_{t+1} \mid B_t]$ appear in the next matrix $\Pr[F_{t+1} \mid H_{t+1}]$ (up to scaling), so they can be represented in basis B_{t+1}, yielding operators $A_{o,t}$.

an efficient representation. To see why, suppose we have such a submatrix, and let us index the columns/histories of the submatrix by B_t, which we refer to as the *basis*. It follows that $\Pr[F_t|B_t]$ spans the column space of $\Pr[F_t|H_t]$, which implies that for any history $h \in H_t$ there exists coefficients $\beta(h) \in \mathbb{R}^{|B_t|}$ such that

$$\Pr[F_t|h] = \Pr[F_t|B_t]\beta(h).$$

The main observation toward obtaining an efficient representation is to exploit a certain circulant structure in the matrices $\{\Pr[F_t|H_t]\}_{t\leq T}$ to model the evolution of the coefficients (visualized in Figure 1). The circulant structure is simply that for basis B_t, observation o, and future $f \in F_{t+1}$ (i.e., of length $T - t - 1$) the vector $\Pr[B_t of]$ appears in two of the matrices (albeit with different scaling). It appears in the matrix $\Pr[F_t|H_t]$ in row of and columns B_t, and it appears in the matrix $\Pr[F_{t+1}|H_{t+1}]$ in row f and columns $B_t o$. Thus, if we learn how to represent the columns $\Pr[F_{t+1}|B_t o]$ in terms of the columns $\Pr[F_{t+1}|B_{t+1}]$—which we can do via the coefficients—the circulant property provides a connection between the matrices $\Pr[F_{t+1}|H_{t+1}]$ and $\Pr[F_t|H_t]$.

Formally, we can define operators $\{A_{o,t}\}$ for each observation $o \in \mathcal{O}$ and sequence length $t \in [T]$ satisfying

$$\Pr[F_{t+1}|B_{t+1}]A_{o,t} = \Pr[oF_{t+1}|B_t], \tag{2}$$

which can then be used to express sequence probabilities by iterated application. Indeed, we have

$$\begin{aligned}\Pr[x_1,\dots,x_T] = \Pr[x_1,\dots,x_T|B_0] = \Pr[x_2,\dots,x_T|B_1]A_{x_1,0} = \dots \\ \dots = \Pr[x_T|B_{T-1}]A_{x_{T-1},T-2}\dots A_{x_1,0} = A_{x_T,T-1}\dots A_{x_1,0}\,,\end{aligned} \tag{3}$$

where by an explicit choice of B_0, B_T and F_T, the matrices $A_{x_1,0}$ and $A_{x_T,T-1}$ are column and row vectors respectively, and so the right-hand side is a scalar (see Proposition 8 for details[4]). More importantly, these operators can also be viewed as evolving the coefficients via the identity:

$$\forall h \in H_t, o \in \mathcal{O} : \beta(ho) = \frac{A_{o,t}\beta(h)}{\Pr[o|h]}. \tag{4}$$

[4] We define B_0, B_T and F_T to be singleton sets. B_0 and F_T contain the empty string φ and B_T contains any length T observation sequence. These new definitions, in conjunction with Proposition 8 imply: $A_{x_T,T-1} = \Pr[x_T|B_{T-1}]$ and therefore will be a row vector. Similarly, $A_{x_1,0}$ is a solution of $\Pr[F_1|B_1]A_{x_1,0} = \Pr[x_1F_1|\varphi]$ and is therefore a column vector.

This identity is proved in Propositions 8 and 24. We highlight the scaling, which results in a nonlinear update equation and appears because the coefficients express conditional rather than joint probabilities. This viewpoint of operators evolving coefficients will play a central role in our error analysis.

Thus, it remains to find the bases $\{B_t\}_{t\leq T}$, estimate the operators $\{A_{o,t}\}_{o\in\mathcal{O},t\leq T}$, and control the error amplification from iteratively multiplying these estimates. We turn to these issues next.

Remark 5 *The approach of Hsu, Kakade, and Zhang can also be viewed as estimating operators via Equation (2) with the particular choice of basis. They show that conditional distribution of futures given any history can be written in the span of the conditional distributions of the single observation histories, so that $\mathcal{O}$ itself forms a basis. This is implied by their assumptions and it permits using only second and third degree moments to estimate the operators. However, in general we will need to use long sequences in our bases and interactive access will be crucial for estimation. Additionally, under their choice of bases and their assumptions they show that the solution of Equation (2) is related to the observable operators (Jaeger, 2000), explicitly given by $\mathbb{T}$ and $\mathbb{O}$, by an invertible and bounded transformation, which is instrumental in their error analysis. When considering general basis B, we do not have such a connection and will require a novel error propagation argument.*

2.3. Error propagation

Although finding the bases B_t and estimating corresponding operators $A_{o,t}$ is nontrivial, even if we have estimated these operators accurately, we must address the error amplification issues from repeated application of the learned operators. This challenge makes up the majority of our technical analysis. We discuss estimating operators $A_{o,t}$ in Section 2.4 and how to find the basis in Section 2.5.

To explain this challenge, suppose for now that we are given bases $\{B_t\}_{t\leq T}$ and subsequently estimate the operators $A_{o,t}$ in ℓ_2 norm, i.e., we have estimate $\widehat{A}_{o,t}$ satisfying $\|\widehat{A}_{o,t} - A_{o,t}\|_2 \leq \varepsilon$. We first define our estimated model $\widehat{\Pr}$ in terms of the estimated operators $\widehat{A}_{o,t}$. Considering Equation (3) the natural estimator is

$$\widehat{\Pr}[x_1,\ldots,x_T] = \widehat{A}_{x_T,T-1}\ldots\widehat{A}_{x_2,1}\widehat{A}_{x_1,0}, \tag{5}$$

where as before, the matrices $\widehat{A}_{x_1,0}$ and $\widehat{A}_{x_T,T-1}$ are column and row vectors respectively, so the right-hand side is a scalar. To simplify notation for this section, we omit the time indexing on the operators.

Given this estimate, the total variation distance is

$$\frac{1}{2}\sum_{x_1,\ldots,x_T\in\mathcal{O}^T}\left|\widehat{A}_{x_T}\ldots\widehat{A}_{x_1} - A_{x_T}\ldots A_{x_1}\right|.$$

Let us first discuss two strategies for bounding this expression that can work in some cases, but do not seem to work in our setting. One idea is to pass to the ℓ_2 norm and use a telescoping argument to obtain several terms of the form

$$\sum_{x_1,\ldots,x_T\in\mathcal{O}^T}\|\widehat{A}_{x_T}\ldots\widehat{A}_{x_{t+2}}\|_2\cdot\|\left(\widehat{A}_{x_{t+1}} - A_{x_{t+1}}\right)A_{x_t}\ldots A_{x_1}\|_2$$

These terms are convenient because the matrix products only disagree in the t^{th} operator. However, both the "incoming" product $A_{x_t} \dots A_{x_1}$ that pre-multiplies this difference and the "outgoing" product $\widehat{A}_{x_T} \dots \widehat{A}_{x_{t+2}}$ whose norm we must bound can be rather poorly behaved. For example, the product $A_{x_t} \dots A_{x_1}$ can have ℓ_2 norm that grows exponentially with t, since the ℓ_2 norm of the individual matrices can be much larger than 1. An even worse problem is that we have exponentially many terms in the sum, so that even bounding each term by ε (which would be possible if the incoming and outgoing products were well behaved) is grossly insufficient.

The other approach is the strategy adopted by Hsu, Kakade, and Zhang (Hsu et al., 2012), which uses the definition of the observable operators (Jaeger, 2000), $\mathbb{A}_x = \mathbb{T}\text{diag}(\mathbb{O}[x, \cdot])$, explicitly. This allows them to control the incoming and outgoing products in a decomposition based on the ℓ_1 norm version. For instance

$$\sum_{x_1,\dots,x_{t+1}} \| \left(\widehat{\mathbb{A}}_{x_{t+1}} - \mathbb{A}_{x_{t+1}} \right) \mathbb{A}_{x_t} \dots, \mathbb{A}_{x_1} \|_1 \lesssim O\varepsilon \cdot \sum_{x_1,\dots,x_t} \|\mathbb{A}_{x_t} \dots, \mathbb{A}_{x_1}\|_1 \leq O\varepsilon.$$

The idea is that each term in the final sum can be seen as a joint probability of the history $x_1, \dots, x_t$ and the hidden state s_{t+1}, so we can sum over all histories with no error amplification. Unfortunately, there is no hidden state in the more general setting (and for the rank deficient case, the observable operators can not be learned accurately as discussed in Section 2.1), so we cannot appeal to an argument of this form. Indeed, our main technical contribution is a new perturbation analysis that relies on no structural assumptions.

At a more technical level, the issue with both of these arguments is that passing to any norm, seems to be too coarse to adequately control the error amplification. Instead, our argument carefully tracks the error in the space of the coefficients. Precisely, given estimates $\widehat{A}_{o,t}$ that satisfy $\|\widehat{A}_{o,t} - A_{o,t}\|_2 \leq \varepsilon$, we can show, via an inductive argument, that for any $x_1, \dots, x_t$

$$(\widehat{A}_{x_t} \dots \widehat{A}_{x_1} - A_{x_t} \dots A_{x_1}) = \sum_{h \in H_t} \beta(h)\alpha_h + \sum_{v \in V_t^{\perp}} v\gamma_v,$$

where $V_t^{\perp}$ is an orthonormal basis for the kernel of $\Pr[F_t \mid B_t]$ and α_h, γ_v are scalars. Moreover, the TV distance between $\Pr[\cdot]$ and $\widehat{\Pr}[\cdot]$ is exactly equal to the sum of these scalars over all sequences $x_1, \dots, x_T$. Even though there could be exponentially many terms in this sum, we show that this sum is small via an inductive argument. This makes up the most technical component of our proof, and we give a more detailed overview in Appendix B.1.2 with the formal proofs in Appendix E.3.

2.4. Estimating operators

We next discuss estimating the operators $\{A_{o,t}\}_{o \in \mathcal{O}, t \leq T}$ using the conditional sampling oracle. A natural idea is to use samples to estimate both sides of the system in Equation (2) and solve the noisy version via linear regression. Unfortunately, this system may have exponentially small (in $T - t$) singular values, making it highly sensitive to perturbation. There is also a cosmetic issue when working with $\Pr[F_{t+1}|B_{t+1}]$, namely this matrix is exponentially large.

To address these challenges, we introduce a particular preconditioner that stabilizes the system. Specifically, we instead estimate and solve

$$\Pr[F_{t+1}|B_{t+1}]^{\top} D_{t+1}^{-1} \Pr[F_{t+1}|B_{t+1}] A_{o,t} = \Pr[F_{t+1}|B_{t+1}]^{\top} D_{t+1}^{-1} \Pr[oF_{t+1}|B_t] \,,$$

where D_{t+1} is a diagonal matrix with entries $d_{t+1}(f) := \frac{1}{|B_{t+1}|}\sum_{b \in B_{t+1}} \Pr[f|b]$ on the diagonal.[5] One benefit of this preconditioner is that the new matrices are of size $|B_{t+1}| \times |B_{t+1}|$ rather than exponentially large, and yet they can still be estimated efficiently using the conditional sampling oracle. To see why the latter holds, observe that the $(i,j)^{\text{th}}$ entry of the matrix on the LHS is

$$\left[\Pr[F_{t+1}|B_{t+1}]^\top D_{t+1}^{-1} \Pr[F_{t+1}|B_{t+1}]\right]_{i,j} = \sum_{f \in F_{t+1}} d_{t+1}(f) \left[\frac{\Pr[f|b_i]\Pr[f|b_j]}{d_{t+1}(f)^2}\right],$$

where $B_{t+1} = \{b_1, b_2, \ldots,\}$. Intuitively, we can estimate this entry by sampling futures f from $\Pr[\cdot|b]$ to approximate any term in the sum and sampling futures from $d_{t+1}(\cdot)$ to approximate the sum itself. While this is true, there is one technical issue to overcome: to estimate the ratio to additive accuracy, we must estimate the individual probabilities $\Pr[f \mid b_i]$, $\Pr[f \mid b_j]$ and $d_{t+1}(f)$ to relative accuracy. We can obtain $(1 \pm \zeta)$ relative error estimates using conditional samples as long as the one-step probabilities are at least $\Omega(\zeta/T)$, but this is challenging when even a single one-step probability is small. To address this issue, we show that such futures actually contribute very little to the overall sum, and we design a test to safely ignore them. See Appendix E.6 for details.

While the ability to estimate the entries is clearly important, the hope with preconditioning is that it dramatically amplifies the singular values of the matrix on the left hand side. In particular, we want that the matrix $\Pr[F_{t+1}|B_{t+1}]^\top D_{t+1}^{-1} \Pr[F_{t+1}|B_{t+1}]$ has large (non-zero) eigenvalues, as this will allow us to estimate the operators $A_{o,t}$ in the ℓ_2 norm. Our choice of preconditioner does achieve this in the important example of parity with noise: we can show that $\Pr[F_{t+1}|B_{t+1}]$ has exponentially small (in $T - t$) singular values for every choice of B_{t+1}, while there exists a basis B_{t+1} for which the eigenvalues of the preconditioned matrix are $\Omega(1)$ (see Appendix C). Unfortunately, in general, a basis which ensures the preconditioner has large eigenvalues might not exist, and we address this by introducing the notion of fidelity.

Definition 6 (Fidelity) *We say that distribution* $\Pr[\cdot]$ *has fidelity* Δ *if there exists some basis* $\{B_t\}_{t \in [T]}$*, such that* $\max_t |B_t| \le 1/\Delta$ *and*

$$\forall t \in [T] : \ \sigma_+\left(S_t^{\frac{1}{2}} \Pr[F_t|H_t]^\top D_t^{-1} \Pr[F_t|H_t] S_t^{\frac{1}{2}}\right) \ge \Delta$$

where $\sigma_+(M)$ *denotes the magnitude of the smallest non-zero singular value of* M*,* D_t *is a diagonal matrix of size* $|F_t| \times |F_t|$ *with entries* $d_t(f) := \frac{1}{|B_t|}\sum_{b \in B_t} \Pr[f|b]$*, and* S_t *is a diagonal matrix of size* $|H_t| \times |H_t|$ *with entries* $s_t(h) := \Pr[h]$*.*

Importantly, we only assume the existence of a basis with this property, not that it is given to us or otherwise known in advance. Note that, although the matrix with large singular values according to the fidelity definition is not the same as the preconditioned matrix we care about for learning operators, nevertheless when the distribution has high fidelity (i.e., Δ is large), we can find a basis for which $\Pr[F_{t+1}|B_{t+1}]^\top D_{t+1}^{-1} \Pr[F_{t+1}|B_{t+1}]$ has large singular values. This, combined with our approach for estimating entries of the preconditioned matrix, allow us to learn operators $A_{o,t}$ in the ℓ_2 norm. We provide details in Appendix E.2.

[5] This choice of D_{t+1} ensures there is no division-by-zero issue, see Remark 11.

Remark 7 *Although we still require an assumption, the parity with noise example suggests that the fidelity definition, which can lead to a favorable preconditioned system, is more appropriate than directly assuming* $\Pr[F_{t+1}|B_{t+1}]$ *has large singular values. Indeed, we can also show that fidelity captures all previously studied positive results for learning HMMs.*

2.5. Finding the basis

The only remaining challenge is to find the basis $\{B_t\}_{t\in[T]}$. Recall that, when considering the conditional sampling oracle, we want a basis for which the preconditioned matrix has large eigenvalues. It turns out that when the distribution has high fidelity a random sample of polynomially many histories will form a basis with this property with high probability. Given that the other aspects of our analysis seem to require high fidelity, this random sampling approach thus suffices to prove Theorem 2.

On the other hand, for low fidelity distributions, random sampling will fail to cover the directions with small eigenvalue, and so basis finding becomes an intriguing aspect of learning with the conditional sampling oracle. Basis finding is also the final issue to address for Theorem 1, using the exact oracle. In both cases, we provide adaptations of Angluin's L^* algorithm that finds a basis for any low rank distribution. We defer discussion of the conditional sampling version to Appendix F and hope that it serves as a starting point toward resolving Open Problem 1.

Adapting L^* for basis finding with the exact oracle. We close this section by explaining how to find a basis when provided with the exact probability oracle. As a first observation, note that we need not construct the entire system in Equation (2) to identify operators $A_{o,t}$. It suffices to find a set of futures $\Lambda_t \subset F_t$ such that $\Pr[\Lambda_t \mid H_t]$ spans the row space of $\Pr[F_t \mid H_t]$. In other words, we just need B_t and Λ_t for which $\Pr[\Lambda_t \mid B_t]$ has the same rank as $\Pr[F_t \mid H_t]$.

The difficulty is that there is no universal choice of B_t, Λ_t for general low rank distributions, and finding these sets poses a challenge search problem in an exponentially large space. We address this challenge using the exact probability oracle and an adaptation of Angluin's L^* algorithm for learning DFAs. The basic idea is as follows: given sets B_t, Λ_t whose submatrix is not of the required rank, we can still solve the (underdetermined) system

$$\Pr[\Lambda_t|B_t]A_{o,t} = \Pr[o\Lambda_t|B_t]$$

and obtain an estimate $\widehat{\Pr}[\cdot]$ via Equation (5). Then, we can sample sequences $x_1, \dots, x_t \sim \Pr[\cdot]$ and check if our estimate makes the correct predictions on these sequences. In particular, we check

$$\widehat{\Pr}[x_1, \dots, x_t, \Lambda_t] \overset{?}{=} \Pr[x_1, \dots, x_t, \Lambda_t]$$

If the prediction are accurate (i.e., these equalities hold) for each t and for polynomially many random sequences, then we can show that $\widehat{\Pr}[\cdot]$ is close $\Pr[\cdot]$ in total variation distance.

On the other hand, if these equalities do not hold for some sample $x_1, \dots, x_t$, then we can use it as a counterexample to improve our basis. Indeed, if the equalities do not hold, there must exist some index $\tau \leq t-1$ such that

$$\Pr[x_1, \dots, x_\tau, \Lambda_\tau] = \Pr[\Lambda_\tau \mid B_\tau]\widehat{A}_{x_\tau,\tau-1} \dots \widehat{A}_{x_1,0} \tag{6}$$

$$\Pr[x_1 \dots x_\tau x_{\tau+1}\Lambda_{\tau+1}] \neq \Pr[\Lambda_{\tau+1}|B_{\tau+1}]\widehat{A}_{x_{\tau+1},\tau}\widehat{A}_{x_\tau,\tau-1} \dots \widehat{A}_{x_1,0} \tag{7}$$

Let $\lambda \in \Lambda_{\tau+1}$ index a row in Equation (7) where equality does not hold. We update $\Lambda'_\tau = \Lambda_\tau \cup \{x_{\tau+1}\lambda\}$ and $B'_\tau = B_\tau \cup \{x_1, \dots, x_\tau\}$ and we claim that the rank of the submatrix $\Pr[\Lambda'_\tau \mid B'_\tau]$ is greater than that of $\Pr[\Lambda_\tau \mid B_\tau]$. To see why, note that this matrix has a new row, indexed by $x_{\tau+1}\lambda$ and a new column indexed by $x_1, \dots, x_\tau$.

We prove that this new row is linearly independent of the previous ones as follows. First note that, we have $\Pr[\Lambda_{\tau+1} \mid B_{\tau+1}]\widehat{A}_{x_{\tau+1},\tau} = \Pr[x_{\tau+1}\Lambda_{\tau+1} \mid B_\tau]$, by the way we define our estimated operators. Using this fact, the two equations above become:

$$\Pr[x_1, \dots, x_\tau, \Lambda_\tau] = \Pr[\Lambda_\tau \mid B_\tau]\widehat{A}_{x_\tau,\tau-1} \dots \widehat{A}_{x_1,0},$$
$$\Pr[x_1 \dots x_\tau x_{\tau+1}\lambda] \neq \Pr[x_{\tau+1}\lambda | B_\tau]\widehat{A}_{x_\tau,\tau-1} \dots \widehat{A}_{x_1,0}.$$

Observe that the product of estimated operators on the right hand side is identical for both equations. In the first equation, this product gives the coefficients for writing $\Pr[\Lambda_\tau \mid x_1, \dots, x_\tau]$ in the basis $\Pr[\Lambda_\tau \mid B_\tau]$ (up to scaling). But the same coefficients do not express $\Pr[x_{\tau+1}\lambda \mid x_1, \dots, x_\tau]$ in terms of the row vector $\Pr[x_{\tau+1}\lambda \mid B_\tau]$, and this implies that the row $\Pr[x_{\tau+1}\lambda \mid B'_\tau]$ cannot be in the span of the previous ones. We provide all the details in Appendix A.

3. Discussion

In this paper we show how interactive access to hidden Markov models (and more generally low rank distributions) can circumvent computational barriers to efficient learning. In particular, we show that all low rank distributions with a certain fidelity property can be efficiently learned assuming access to a conditional sampling oracle. In Appendix C, we show that fidelity captures the assumptions considered in prior work on (non-interactive) learning of HMMs, specifically:

- Parity with noise admits bases B_t each of cardinality 2 with fidelity $(1 - 2\alpha^2)/2$, where α is the noise parameter.
- Full rank HMMs, where $\mathbb{T}$ and $\mathbb{O}$ are full column rank, admit bases of size O with fidelity bounded by the minimum singular value of the second moment matrix $\Pr[\mathbf{x}_2 = \cdot, \mathbf{x}_1 = \cdot]$. This parameter also appears polynomially in the analysis of Hsu et al. (2012).
- The overcomplete setting of Sharan et al. (2017), where sequences of length $\log S$ are used for estimation, admits bases of size S with fidelity $1/\operatorname{poly}(S)$, matching their parameters.

Despite this, the reliance on the fidelity parameter is the main limitation of our results. We believe this dependence is not necessary, which leads to the main open problem, Open Problem 1. We close the paper with some final remarks regarding this open problem.

As we have mentioned previously, although fidelity greatly simplifies the basis finding aspect of our algorithm, it is not necessary for this part and refer the reader to Appendix F where we give a general algorithm for basis finding. Indeed the only place where fidelity is required is in our error propagation analysis, where our techniques require that operators $\widehat{A}_{o,t}$ are estimated in ℓ_2 norm. In the general case, we will only be able to learn operators in the directions for which the preconditioned matrix has large eigenvalues, and ideally we should be able to ignore the directions with small eigenvalues. This strategy would work if we can show that ignoring the small directions preserves the low rank property, which is the linear-algebraic analog of approximating an HMM by one with fewer states. Unfortunately, we do not know if the latter holds, and we believe this is the key challenge to resolving Open Problem 1. We look forward to further progress on this problem.

Acknowledgments

This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence. Sham Kakade acknowledges funding from the Office of Naval Research under award N00014-22-1-2377 and the National Science Foundation Grant under award #CCF-2212841. The authors thank Weihao Kong, Vatsal Sharan, Geelon So and Shachar Lovett for insightful discussions.

References

Michael Alekhnovich. More on average case vs approximation complexity. In *Symposium on Foundations of Computer Science*, 2003.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of machine learning research*, 2014.

Dana Angluin. Learning regular sets from queries and counterexamples. *Information and computation*, 1987.

Rishiraj Bhattacharyya and Sourav Chakraborty. Property testing of joint distributions using conditional samples. *ACM Transactions on Computation Theory*, 2018.

Avrim Blum, Merrick Furst, Jeffrey Jackson, Michael Kearns, Yishay Mansour, and Steven Rudich. Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In *Symposium on Theory of Computing*, 1994a.

Avrim Blum, Merrick Furst, Michael Kearns, and Richard J Lipton. Cryptographic primitives based on hard learning problems. In *Advances in Cryptology*, 1994b.

Clément Canonne and Ronitt Rubinfeld. Testing probability distributions underlying aggregated data. In *International Colloquium on Automata, Languages, and Programming*, pages 283–295. Springer, 2014.

Clément L Canonne, Dana Ron, and Rocco A Servedio. Testing probability distributions using conditional samples. *SIAM Journal on Computing*, 2015.

Sourav Chakraborty, Eldar Fischer, Yonatan Goldhirsh, and Arie Matsliah. On the power of conditional samples in distribution testing. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 561–580, 2013.

Xi Chen, Rajesh Jayaram, Amit Levi, and Erik Waingarten. Learning and testing junta distributions with sub cube conditioning. In *Conference on Learning Theory*, 2021.

Mary Cryan, Leslie Ann Goldberg, and Paul W Goldberg. Evolutionary trees can be learned in polynomial time in the two-state general Markov model. *SIAM Journal on Computing*, 2001.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, 2008.

Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, 2012.

Qingqing Huang, Rong Ge, Sham Kakade, and Munther Dahleh. Minimal realization problems for hidden markov models. *IEEE Transactions on Signal Processing*, 2015.

Herbert Jaeger. Observable operator models for discrete stochastic time series. *Neural computation*, 2000.

Michael Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert E Schapire, and Linda Sellie. On the learnability of discrete distributions. In *Symposium on Theory of Computing*, 1994.

Aryeh Kontorovich, Boaz Nadler, and Roi Weiss. On learning parametric-output hmms. In *International Conference on Machine Learning*, 2013.

Elchanan Mossel and Sébastien Roch. Learning nonsingular phylogenies and hidden markov models. In *Symposium on Theory of Computing*, 2005.

Tianyi Peng. Bound on difference of eigen projections of positive definite matrices. Mathematics Stack Exchange, 2020. URL https://math.stackexchange.com/q/3921839.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, 2011.

Vatsal Sharan, Sham M Kakade, Percy S Liang, and Gregory Valiant. Learning overcomplete HMMs. *Advances in Neural Information Processing Systems*, 2017.

Roi Weiss and Boaz Nadler. Learning parametric-output hmms with two aliased states. In *International Conference on Machine Learning*, 2015.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. A survey of active learning for natural language processing. *arXiv:2210.10109*, 2022.

Contents

Appendix A. Warmup Setting: Learning with exact conditional probabilities

Before, we show how to learn a low rank distribution in total variation distance using only conditional *samples*, we first introduce our overall approach on an easier setting: when the learner has access to *exact conditional probability oracle* (Definition 3).

We first recall some definitions. Here our notation will differ from the technical ideas section (Section 2) since we will be building basis for each sequence length $t \in [T]$ separately. We will refer to set $H_t := \mathcal{O}^t$ as the set of histories of length t. Similarly, we will refer to set $F_t := \mathcal{O}^{\leq T-t}$ as the set of futures (notice the discrepancy of t and $T-t$). Note that one could append elements from futures F_t in front of elements from histories H_t and get a valid observation sequence of total length $\leq T$. Throughout, we will denote the target low rank distribution by $\mathbf{p}$. Furthermore, our low rank assumption on $\mathbf{p}$ implies that the conditional probability matrix $\Pr[F_t|H_t]$ has rank r_t at most r. We can interpret this geometrically, as the vectors $\Pr[F_t|h]$ for history $h \in H_t$ span an r_t-dimensional subspace.

A crucial implication of this assumption is that there exists r_t special histories of length t, henceforth denoted by B_t, such that the submatrix $\Pr[F_t|B_t]$ is also rank r_t. This means, for any history $x \in \mathcal{O}^t$, there exists coefficients $\beta(x) \in \mathbb{R}^{|B_t|}$ such that

$$\Pr[F_t|x] = \Pr[F_t|B]\beta(x)\,.$$

As discussed in Section 2.2, even though there are exponentially many histories for which we may have to learn the coefficients, because of the circulant structure of the conditional probability matrix, we can generate all of them using a small matrix with $O(r^2)$ entries.

Notation. To clean up the notation, we define $B_0 = \{\varphi\}$ where φ is the empty string. We define probabilities associated to empty string as: $\beta(\varphi) = 1$, $\Pr[x_1 \ldots x_T|\varphi] = \Pr[x_1 \ldots x_T]$ and $\Pr[\varphi|x_1 \ldots x_T] = 1$ for any T-length sequence $x_1, \ldots, x_T$. Let $F_T = \Lambda_T = \{\varphi\}$. Then, because $\Pr[F_T|H_T]$ is all ones matrix, we can set $B_T = \{h\}$ for any observation sequence $h \in H_T$. These new definitions, in conjunction with the proposition below imply: $A_{x_T,T-1} = \Pr[x_T|B_{T-1}]$ and therefore will be a row vector. Similarly, $A_{x_1,0}$ is a solution of $\Pr[F_1|B_1]A_{x_1,0} = \Pr[x_1F_1|\varphi]$ and is therefore a column vector.

Proposition 8 (Existence of efficient representation) *Let B_0, B_T and F_T be as defined above. For $t \in [T-1]$, let $B_t \subset H_t$ be any set of histories of length t such that column vectors $\Pr[F_t|B_t]$ span the column space of $\Pr[F_t|H_t]$. Then, the probability distribution $\mathbf{p}$ can be written as*[6]*:*

$$\Pr_{\mathbf{p}}[x_1 \ldots x_T] = A_{x_T,T-1}A_{x_{T-1},T-2} \ldots A_{x_1,0}$$

where matrices $A_{o,t}$ for every $o \in \mathcal{O}$ and $t+1 \in [T]$ which satisfy[7]

$$\Pr[F_{t+1}|B_{t+1}]A_{o,t} = \Pr[oF_{t+1}|B_t] \tag{8}$$

[6] Here by choice of basis B_0 and B_T, $A_{x_1,0}$ and $A_{x_T,T-1}$ are column and row vectors respectively

[7] and we will see that this equation always has a solution.

Proof We first show there exists a solution $A_{o,t}$ for Equation (8). We claim $A_{o,t}$ defined using basis $B_t = \{b_1, \ldots, b_n\}$ and B_{t+1} as follows is a solution:

$$A_{o,t} = \begin{bmatrix} \beta(b_1 o) & \beta(b_2 o) & \cdots & \beta(b_n o) \end{bmatrix} \begin{bmatrix} \Pr[o|b_1] & 0 & \cdots & 0 \\ 0 & \Pr[o|b_2] & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \Pr[o|b_n] \end{bmatrix}$$

Here $\beta(x)$ and $\beta(xo)$ are the coefficients associated to history x of length t under B_t and history xo of length $t+1$ under B_{t+1} respectively. By definition of $A_{o,t}$,

$$\begin{aligned} &\Pr[F_{t+1}|B_{t+1}]A_{o,t} \\ =&\Pr[F_{t+1}|B_{t+1}] \begin{bmatrix} \beta(b_1 o) & \beta(b_2 o) & \cdots & \beta(b_n o) \end{bmatrix} \begin{bmatrix} \Pr[o|b_1] & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \Pr[o|b_n] \end{bmatrix} \\ =&\Pr[F_{t+1}|B_t o] \begin{bmatrix} \Pr[o|b_1] & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \Pr[o|b_n] \end{bmatrix} && \text{(by definition of } \beta(b_i o)) \\ =&\Pr[oF_{t+1}|B_t] && \text{(by Bayes rule)} \end{aligned}$$

Since oF_{t+1} is a subset of F_t, by repeatedly applying this equation, we get

$$\Pr[F_T|B_T]A_{x_T,T-1}A_{x_{T-1},T-2}\ldots A_{x_1,0} = \Pr[x_1 x_2 \ldots x_T|\varphi]$$

Noting $\Pr[F_T|B_T] = 1$ and $\Pr[x_1 x_2 \ldots x_T|\varphi] = \Pr[x_1 x_2 \ldots x_T]$ completes the proof.

We do point here that $A_{x_T,T-1} = \Pr[x_T|B_{T-1}]$ by definition and is therefore a row vector. Similarly, $A_{x_1,0}$ is a solution of $\Pr[F_1|B_1]A_{x_1,0} = \Pr[x_1 F_1|\varphi]$ and is therefore a column vector. ■

With exact conditional probabilities, we can learn $A_{o,t}$ exactly if we know any set of histories B_t which form r independent columns of $\Pr[F_t|H_t]$. So the goal is to find such a set of histories. This presents an issue: since there are exponentially many columns in $\Pr[F_t|H_t]$, finding the independent columns is tricky. In fact, even checking if two histories form independent columns would require checking exponentially many entries in the matrix.

To solve these issues, we iteratively build set of basis histories B_t and tests Λ_t. Basis histories B_t as the name suggests is a set of independent columns in $\Pr[F_t|H_t]$ we have identified till now. Tests are the subset of futures/rows which witness the independence. We will always maintain that the submatrix $\Pr[\Lambda_t|B_t]$ is an invertible square matrix and therefore Equation (8) will always have a solution.

How do we improve the set of basis histories and tests? Suppose, we start with sets B_t, Λ_t such that submatrix $\Pr[\Lambda_t|B_t]$ is not rank r_t. We can then compute matrices $\widehat{A}_{o,t}$ using our current guess of basis and tests:

$$\Pr[\Lambda_{t+1}|B_{t+1}]\widehat{A}_{o,t} = \Pr[o\Lambda_{t+1}|B_t]\,,$$

and let $\widehat{\mathbf{p}}$ be the induced distribution:

$$\Pr_{\widehat{\mathbf{p}}}[x_1 \ldots x_T] = \widehat{A}_{x_T,T-1}\widehat{A}_{x_{T-1},T-2}\ldots\widehat{A}_{x_1,0}$$

We will often use $\widehat{\Pr}$ and $\Pr_{\widehat{\mathbf{p}}}$ interchangeably. The main observation here is that if we observe a counterexample to $\widehat{\mathbf{p}}$ being close to $\mathbf{p}$ in TV distance, then we can use this counterexample to improve our basis and tests.

Proposition 9 (Finding basis histories and futures) *If $x_1 \dots x_t$ is a counterexample, that is, it satisfies the following:*

$$\Pr_{\widehat{\mathbf{p}}}[x_1, \dots, x_t, \Lambda_t] \neq \Pr_{\mathbf{p}}[x_1, \dots, x_t, \Lambda_t] \tag{9}$$

then for some $\tau \in [t]$, we find a new test λ' and representative history b' and set $\Lambda'_\tau = \Lambda_\tau \cup \{\lambda\}$ and $B'_\tau = B_\tau \cup \{\lambda'\}$ such that $\operatorname{rank}(\Pr[\Lambda'_\tau | B'_\tau]) = \operatorname{rank}(\Pr[\Lambda_\tau | B_\tau]) + 1$.

Proof For clarity, in the poof, we will abuse notation and not explicitly mention the sequence length when writing the operator $A_{o,t}$ i.e. we will use A_{x_t} instead of $A_{x_t,t-1}$. First, we find a time $\tau \in [t]$ where the following equations hold:

$$\Pr[x_1 \dots x_\tau \Lambda_\tau] = \Pr[\Lambda_\tau | B_\tau] \widehat{A}_{x_\tau} \dots \widehat{A}_{x_1}$$
$$\Pr[x_1 \dots x_\tau x_{\tau+1} \Lambda_{\tau+1}] \neq \Pr[\Lambda_{\tau+1} | B_{\tau+1}] \widehat{A}_{x_{\tau+1}} \widehat{A}_{x_\tau} \dots \widehat{A}_{x_1}$$

This is true because the first equation is true for $\tau = 0$ by definition, and the second equation is true for $\tau = t-1$ because of the counterexample (Equation (9)). Now, we can simplify the equations above by substituting vector $\mathbf{v} = (\Pr[x_1 \dots x_\tau])^{-1} \widehat{A}_{x_\tau} \dots \widehat{A}_{x_1}$ which gives

$$\Pr[\Lambda_\tau | x_1 \dots x_\tau] = \Pr[\Lambda_\tau | B_\tau] \mathbf{v} \tag{10}$$
$$\Pr[x_{\tau+1} \Lambda_{\tau+1} | x_1 \dots x_\tau] \neq \Pr[\Lambda_{\tau+1} | B_{\tau+1}] \widehat{A}_{x_{\tau+1}} \mathbf{v} = \Pr[x_{\tau+1} \Lambda_{\tau+1} | B_\tau] \mathbf{v} \tag{11}$$

where the last step holds by definition of $\widehat{A}_{x_{\tau+1}}$ (Equation (13)). Let $x_{\tau+1}\lambda_{\tau+1}$ be the row where the inequality holds. Define $\lambda' = x_{\tau+1}\lambda_{\tau+1}$ and $b' = x_1 \dots x_\tau$. We will now show that the equations above imply vector $\Pr[x_{\tau+1}\lambda_{\tau+1} | B'_\tau]$ is independent of rows of $\Pr[\Lambda_\tau | B'_\tau]$. This will be enough to prove our claim i.e. $\operatorname{rank}(\Pr[\Lambda'_\tau | B'_\tau]) = \operatorname{rank}(\Pr[\Lambda_\tau | B_\tau]) + 1$.

We prove this independence by contradiction. Assume, they are linearly dependent, that is there exists a vector $\mathbf{w}$ such that :

$$\Pr[x_{\tau+1}\lambda_{\tau+1} | B'_\tau] = \mathbf{w}^\top \Pr[\Lambda_\tau | B'_\tau]\,. \tag{12}$$

Then, we reach a contradiction as

$$\begin{aligned} \Pr[x_{\tau+1}\lambda_{\tau+1} | x_1 \dots x_\tau] &= \mathbf{w}^\top \Pr[\Lambda_\tau | x_1 \dots x_\tau] \\ &= \mathbf{w}^\top \Pr[\Lambda_\tau | B_\tau] \mathbf{v} \\ &= \Pr[x_{\tau+1}\lambda_{\tau+1} | B_\tau] \mathbf{v} \\ &\neq \Pr[x_{\tau+1}\lambda_{\tau+1} | x_1 \dots x_\tau] \end{aligned}$$

where the first and third equality follows from linear dependence (Equation (12)), second equality follows from Equation (10) and last inequality follows from Equation (11). ■

Since, each B_t can be at most size r, and this process expands our basis every iteration, it must terminate after rT iterations, hence proving Theorem 1. We present the remaining proofs in Appendix D.

Algorithm 1: Learning low rank distributions using exact conditional probabilities.

```
1 Set $B_t = \{\underbrace{0 \ldots 0}_{t \text{ times}}\}$ and $\Lambda_t = \{0\}$ for all $t \in [T]$.
2 for round $1, 2, \ldots$ do
3     Choose $\widehat{A}_{o,t}$ for each $o \in \mathcal{O}$ and $t \in [T-1]$ which satisfies
          $$\Pr[\Lambda_{t+1}|B_{t+1}]\widehat{A}_{o,t} = \Pr[o\Lambda_{t+1}|B_t] \quad (13)$$
4     Let $\widehat{\Pr}$ be a function defined on observation sequence $(x_1 \ldots x_t)$ for $t \in [T]$ as,
          $$\widehat{\Pr}[x_1, \ldots, x_t, \Lambda_t] = \Pr[\Lambda_t|B_t]\widehat{A}_{x_t,t-1} \ldots \widehat{A}_{x_1,0}$$
5     Sample $n$ sequences $(x_1, \ldots x_t)$ each of length $t \in [T]$ and check if any one of them is a
        "counterexample" i.e. satisfies
          $$\widehat{\Pr}[x_1, \ldots, x_t, \Lambda_t] \neq \Pr[x_1, \ldots, x_t, \Lambda_t]$$
6     if we find such a counterexample $(x_1, \ldots, x_t)$ then
7         Use Proposition 9 to find a new test $\lambda'$ and representative history $b'$ and update
            $\Lambda_\tau := \Lambda_\tau \cup \{\lambda\}$ and $B_\tau := B_\tau \cup \{\lambda'\}$ for some $\tau \in [t]$.
8     else
9         return $\{\widehat{A}_{o,t}\}_{o \in \mathcal{O}, t \in [T-1]}$
```

A.1. Algorithm

We now present our algorithm. The user furnishes ε, the accuracy with which the distribution is to be learned; and δ, a confidence parameter. The parameter n depends on the input and is detailed in the proof of Theorem 1 below. Note that Line 3 of the algorithm is valid as we will always maintain that $\Pr[\Lambda_{t+1}|B_{t+1}]$ is invertible.

Appendix B. Main Setting: Learning with conditional samples

In this section, we consider our main setting: learn a low rank distribution in total variation distance with only access to conditional sampling oracle. We use the same notation as the previous section, except we abuse notation for the set of futures F_t and redefine it as $F_t := \mathcal{O}^{T-t}$ (so instead of all futures of length up to $T - t$, now its only futures of length exactly $T - t$). Note that this only decreases the rank of matrix $\Pr[F_t|H_t]$. We now formally define a basis for distribution $\mathbf{p}$.

Definition 10 (Basis of a distribution) *Fix a distribution* $\mathbf{p}$ *over observation sequences of length* T. *Consider a set* $\{B_t\}_{t\in[T]}$ *where each* B_t *is a subset of histories of length* t. *Then, we say* $\{B_t\}_{t\in[T]}$ *forms a basis for distribution* $\mathbf{p}$, *if for every observation sequence* $x \in \mathcal{O}^t$, *there exists coefficients* $\beta(x)$ *such that:*

$$\Pr[F_t|x] = \Pr[F_t|B_t]\beta(x)$$

with $||\beta(x)||_2 \leq c$ *for some universal constant* $c \leq 1$[8].

In the previous setting, when we had access to exact conditional probability oracle, the main challenge was finding a set of basis histories. Now, in comparison to our warmup setting, we will also have errors coming from the samples, and therefore even if we know a set of basis histories, we will only be able to learn $A_{o,t}$ approximately by solving a noisy version of this equation:

$$\Pr[F_{t+1}|B_{t+1}]\widehat{A}_{o,t} = \Pr[oF_{t+1}|B_t]\,,$$

This presents us with three difficulties: (i) how do we learn a good estimate for $A_{o,t}$, (ii) how does error in estimates of $A_{o,t}$ propagate to error in estimating $\mathbf{p}$ in TV distance, and (iii) how do we find a good basis to do the above tasks robustly?

B.1. Error propagation

In this section, we show how to estimate the operators $A_{o,t}$ so that the induced probability distribution is not too far from $\mathbf{p}$ in total variation distance. For this, we would need the basis to be robust to small errors. We explain this in detail now.

Suppose, we already have a set of histories B_t of length t such that columns of $\Pr[F_t|B_t]$ span the column space of $\Pr[F_t|H_t]$ or in other words $\{B_t\}_{t\in[T]}$ forms a basis for distribution $\mathbf{p}$. Then, as noted in Proposition 8, $A_{o,t}$ can be written as the solution of the following equation:

$$\Pr[F_{t+1}|B_{t+1}]A_{o,t} = \Pr[oF_{t+1}|B_t]$$

To estimate this operator accurately, we need to learn the subspace spanned by the right singular vectors of $\Pr[F_{t+1}|B_{t+1}]$. The issue however is that the entries in this matrix can be very small: and therefore almost all of its singular values can be pretty small. For example, as noted in the techniques section (Section 2.4), for HMM simulating noisy parity, all its singular values would be exponentially small in T. And therefore, we will not be able to estimate its right singular vectors accurately. Instead, we consider the following conditioned matrix

$$\Pr[F_{t+1}|B_{t+1}]^\top D_{t+1}^{-1}\Pr[F_{t+1}|B_{t+1}]\,,$$

where D_{t+1} is a diagonal matrix of size $|F_{t+1}| \times |F_{t+1}|$ with entries $d_{t+1}(f) := \mathbb{E}_{b\in B_{t+1}}\Pr[f|b]$ on the diagonal. Here $\mathbb{E}_{B_{t+1}}[\cdot]$ for set B_{t+1} refers to expectation under uniform distribution on set B_{t+1}.

Remark 11 *We note here that since we invert the diagonal matrix* D_{t+1}*, we need to be careful about futures* f *where* $d_{t+1}(f) = 0$*. This is not an issue however as by definition of a basis,* $d_{t+1}(f) = 0$ *implies* $\Pr[f|h] = 0$ *for all histories* $h \in H_{t+1}$*. As we will see in Appendix E.6, these futures do not affect the spectrum of above matrix.*

This matrix shares the same right singular vectors as the original matrix and is more likely to have large singular values because of the preconditioning. As an example, for the noisy parity, there exists a basis for which this matrix has only large non-zero singular values (see Appendix C for details). We further define such basis as robust basis of distribution $\mathbf{p}$.

[8] Note that by repeating elements in the basis, we can always make the norm smaller than 1. So, this norm bound is without loss of generality.

Definition 12 (Robust basis of a distribution) *Fix a distribution* $\mathbf{p}$ *over observation sequences of length* T*. Consider a set* $\{B_t\}_{t\in[T]}$ *which forms a basis for distribution* $\mathbf{p}$*. Then, we say* $\{B_t\}_{t\in[T]}$ *is* Δ*-robust basis for distribution* $\mathbf{p}$ *if for every sequence length* $t \in [T]$*:*

$$\sigma_+\left(\Pr[F_t|B_t]^\top D_t^{-1}\Pr[F_t|B_t]\right) \geq \Delta$$

where $\sigma_+(M)$ *denotes the minimum non-zero eigenvalue of* M *and* D_t *is a diagonal matrix of size* $|F_t| \times |F_t|$ *with entries* $d_t(f) := \mathbb{E}_{b\in B_t}\Pr[f|b]$ *on the diagonal.*

A priori, it is unclear if such a basis exists for arbitrary low rank distributions. Moreover, even if such a robust basis exists, how do we find it? For now, we will ignore these issues and assume we have access to a Δ-robust basis B_t.

B.1.1. ESTIMATING THE OPERATORS

With access to a Δ-robust basis B_t, we first show that we can estimate $A_{o,t}$ in ℓ_2 norm.

Lemma 13 (Estimating operators $A_{o,t}$) *Fix a distribution* $\mathbf{p}$ *over observation sequences of length* T*. Assume the distribution* $\mathbf{p}$ *has rank* r*, and we know a* Δ*-robust basis* $\{B_t\}_{t\in[T]}$ *for distribution* $\mathbf{p}$*. Then, using* $\mathrm{poly}(r, |\mathcal{O}|, T, 1/\varepsilon, 1/\Delta, \log(1/\delta))$ *queries to the conditional sampling oracle, we can learn an approximation* $\widehat{A}_{o,t}$ *such that with probability* $1-\delta$*, for all observations* $o \in \mathcal{O}$ *and* $t \in [T]$

$$||\widehat{A}_{o,t} - A_{o,t}||_2 \leq \varepsilon\,.$$

We provide proofs in Appendices E.2 and E.6. This lemma follows from standard arguments once we can estimate the matrix $\Pr[F_t|B_t]^\top D^{-1}\Pr[F_t|B_t]$ accurately. To estimate this matrix, first note that each entry of this matrix can be written as:

$$\left[\Pr[F_t|B_t]^\top D^{-1}\Pr[F_t|B_t]\right]_{i,j} = \mathbb{E}_{f\sim d}\left[\frac{\Pr[f|b_i]\Pr[\lambda|b_j]}{d(f)^2}\right]\,,$$

Because using conditional sampling oracle, we can estimate each $\Pr[f|b_i]$ to multiplicative accuracy, we can estimate each entry of abovementioned matrix to additive accuracy. There are some new technical issues with ignoring futures and histories with low probabilities, which can be easily tested for and ignored.

B.1.2. ERROR PROPAGATION

The main technical challenge is in analyzing how the error in estimating $A_{o,t}$ matrices propagate to errors in induced distributions. In the previous section, we learned an estimate $\widehat{A}_{o,t}$ of operator $A_{o,t}$ such that they were close in ℓ_2 sense. Now, we show that this implies that the induced distributions are also close in total variation error.

We first define some more notation. Let V_t be the subspace formed by the right singular vectors (with non-zero singular value) of $\Pr[F_t|B_t]^\top D^{-1}\Pr[F_t|B_t]$ and $V_t^\perp$ be its orthogonal complement. In Lemma 13, we learned an estimate $\widehat{A}_{o,t}$ of operator $A_{o,t}$ such that they were close in ℓ_2 sense. We can actually do much better!

Let $B_{t+1} = \{b_1, b_2, \ldots, b_n\}$ and $V_{t+1}^{\perp} = \{v_1, v_2, \ldots, v_m\}$. Then, we can show that for any unit vector $u \in \mathbb{R}^{|B_t|}$

$$(\widehat{A}_{o,t} - A_{o,t})u = \sum_{i=1}^{n} \alpha_i \beta(b_i) + \sum_{j=1}^{m} \alpha_j^{\perp} v_j \tag{14}$$

where the coefficients α_i and $\alpha_i^{\perp}$ are bounded in ℓ_1 norm: i.e $\|\alpha\|_1, \|\alpha^{\perp}\|_1 \leq \varepsilon$. Using this structured error, we can show how to bound the TV distance between the induced distributions.

Lemma 14 (Pertubation argument) *Assume for each sequence length $t \in [T]$ and observation $o \in \mathcal{O}$, we have an operator $\widehat{A}_{o,t}$ which is close to $A_{o,t}$ as defined above in Equation (14). Let $\widehat{\mathbf{p}}$ be the distribution induced by $\widehat{A}_{o,t}$ given by*

$$\Pr_{\widehat{\mathbf{p}}}[x_1 \ldots x_T] = \widehat{A}_{x_T} \ldots \widehat{A}_{x_1} \beta(\varphi)$$

Then, for small enough ε, the induced distributions $\widehat{\mathbf{p}}$ and $\mathbf{p}$ are close in TV distance:

$$TV(\mathbf{p}, \widehat{\mathbf{p}}) \leq 2|\mathcal{O}|T\varepsilon$$

We provide the proof in Appendix E.3. We now give a rough idea of the main ideas. As mentioned in the technical section, the main approach will be to analyze how the error written in terms of coefficients evolves. To do this more formally, let's define some more notation. For any set of histories S of sequence length t, let $\beta(S)$ denote a matrix whose columns are given by the coefficient $\beta(s)$ under basis B_t for sequence $s \in S$. Moreover, recall we assume $B_0 = \{\varphi\}$ and therefore $\beta(\varphi) = 1$. Also, abusing notation, $V_{t+1}^{\perp}$ is also matrix representing the orthogonal complement of subspace formed by the right singular vectors (with non-zero singular value) of $\Pr[F_t|B_t]^{\top} D^{-1} \Pr[F_t|B_t]$.

To prove our claim inductively, assume

$$(\widehat{A}_{x_{1:t-1}} - A_{x_{1:t-1}})\beta(\varphi) = \beta(B_{x_{1:t-1}})\gamma_{x_{1:t-1}} + V_{t-1}^{\perp}\gamma_{x_{1:t-1}}^{\perp} \tag{15}$$

where $B_{x_{1:t-1}}$ is some subset of observation sequences of length $t-1$. Our goal is to understand how the ℓ_1 norm of coefficients $\gamma_{x_{1:t-1}}$ grows. Ideally, we would want to show

$$\sum_{x_{1:t} \in \mathcal{O}^t} \|\gamma_{x_{1:t}}\|_1 \approx O(\varepsilon) + (1 + O(\varepsilon)) \sum_{x_{1:t-1} \in \mathcal{O}^{t-1}} \|\gamma_{x_{1:t-1}}\|_1 \tag{16}$$

Couple of remarks. First, the sum over all sequences $x_{1:t}$ is crucial: since the number of sequences is exponentially growing, bounding this error for each sequence separately and summing will be suboptimal. Second, we will also get some terms in the expression above from the error in the orthogonal subspace $V_{t-1}^{\perp}\gamma_{x_{1:t-1}}^{\perp}$. Fortunately, we will show it also grows similar to expression (Equation (16)) above.

We now show why Equation (16) holds true. Specifically, consider the following standard decomposition of this error

$$\begin{aligned}&(\widehat{A}_{x_{1:t}} - A_{x_{1:t}})\beta(\varphi)\\ &= (\widehat{A}_{x_t} - A_{x_t})A_{x_{1:t-1}}\beta(\varphi) + A_{x_t}(\widehat{A}_{x_{1:t-1}} - A_{x_{1:t-1}})\beta(\varphi) + (\widehat{A}_{x_t} - A_{x_t})(\widehat{A}_{x_{1:t-1}} - A_{x_{1:t-1}})\beta(\varphi)\end{aligned}$$

We look at the three terms separately.

The first term is much simpler to bound: in words, the first term is looking at how the one-step error $(\widehat{A}_{x_t} - A_{x_t})$ acts on the *true* coefficients $A_{x_{1:t-1}}\beta(\varphi)$. Here the main argument will be to note that $A_{x_{1:t-1}}\beta(\varphi) = \Pr[x_{1:t-1}]\beta(x_{1:t-1})$. This scaling by the marginal distribution will ensure that summing over all observation sequences will not blow up.

Bounding the other two terms are a bit more involved. The second term is looking at how the true operator A_{x_t} acts on the propagation error $(\widehat{A}_{x_{1:t-1}} - A_{x_{1:t-1}})\beta(\varphi)$ till now. Since, the propagation error can be written in terms of coefficients $\beta(B_{x_{1:t-1}})$ (Equation (15)), we can actually show that A_{x_t} will keep the error bounded.

The third term is looking at how $(\widehat{A}_{x_t} - A_{x_t})$ acts on the propagation error. The most troublesome component of this error is: $(\widehat{A}_{x_t} - A_{x_t})(\beta(B_{x_{1:t-1}})\gamma_{x_{1:t-1}})$. Here, we can think of each column of $\beta(B_{x_{1:t-1}})$ as a vector and use how the one-step error propagates for each column. We fill the details in Appendix E.3.

B.2. Finding robust basis

The last remaining step is to find a Δ-robust basis. Such a basis might not exist, so we first define a class of distributions where such basis exists.

Definition 6 (Fidelity) *We say that distribution $\Pr[\cdot]$ has fidelity Δ if there exists some basis $\{B_t\}_{t\in[T]}$, such that $\max_t |B_t| \le 1/\Delta$ and*

$$\forall t \in [T]:\ \sigma_+\left(S_t^{\frac{1}{2}} \Pr[F_t|H_t]^\top D_t^{-1} \Pr[F_t|H_t] S_t^{\frac{1}{2}}\right) \ge \Delta$$

where $\sigma_+(M)$ denotes the magnitude of the smallest non-zero singular value of M, D_t is a diagonal matrix of size $|F_t| \times |F_t|$ with entries $d_t(f) := \frac{1}{|B_t|}\sum_{b\in B_t} \Pr[f|b]$, and S_t is a diagonal matrix of size $|H_t| \times |H_t|$ with entries $s_t(h) := \Pr[h]$.

In Appendix C, we will go over few common examples of HMMs where fidelity is large. Next, we show how existence of a basis under which a distribution has high fidelity is enough for us to find a robust basis.

Lemma 15 (Finding robust basis) *Fix a distribution $\mathbf{p}$ over observation sequences of length T. Assume distribution $\mathbf{p}$ has rank r and fidelity Δ^*. Pick $0 < \delta < 1$. Let $n = O(\log r\Delta^{*-8})$ and $\Delta = \Omega(\log r(\Delta^*)^{-11/2})$. Then, we can find sets $\{S_t\}_{t\in[T]}$, each of size n, using $n\log(T/\delta)$ conditional samples such that with probability $1-\delta$, $\{S_t\}_{t\in[T]}$ is a Δ-robust basis for distribution $\mathbf{p}$.*

We provide a proof in Appendix E.4. According to this lemma, a random sample from a high fidelity distribution forms a robust basis. This is contrast to our exact setting (Appendix A), where it seemed necessary to search for a basis. Since, our motivation is to find algorithms which do not require high fidelity, in Appendix F, we show how to build a robust basis for arbitrary low rank distributions (without the high fidelity assumption).

B.3. Algorithm

We are now ready to present our algorithm. For our algorithm, the user furnishes ε, the accuracy with which the distribution is to be learned; δ, a confidence parameter; Δ^*, the fidelity of the distribution and r, rank of the distribution. The parameters Δ, λ, n and m depend on the input and are detailed in the proof of Theorem 2 in Appendix E.5.

Algorithm 2: Learning low rank distributions using conditional samples.

```
for sequence length $t = 0, 1, 2, \ldots, T$ do
    Sample set $B_t = \{b_1, \ldots, b_n\}$ of $n$ observation sequences of length $t$ using Lemma 15.
    Build empirical estimates $\widehat{q}(bo)$ and $\widehat{\Sigma}_{B_t}$ for all $b \in B_t$, observations $o \in \mathcal{O}$ using
      Corollary 35 with $m$ conditional samples.
    Compute SVD of $\widehat{\Sigma}_{B_t}$, and let $\widehat{V}_t$ be the matrix of eigenvectors corresponding to
      eigenvalues $> \Delta/2$.
    Compute the coefficients $\widehat{\beta}(b'_i o)$ for each observation $o \in \mathcal{O}$ and sequence $b'_i \in B_{t-1}$ by
      solving the program:
          $\widehat{\beta}(b'_i o) = \operatorname*{argmin}_z ||\widehat{\Sigma}_{B_t} z - \widehat{q}(b'_i o)||_2^2 + \lambda ||z||_2^2.$
    Compute model parameters $\widehat{A}_{o,t-1}$ for each observation $o \in \mathcal{O}$:
          $\widehat{A}_{o,t-1} = \widehat{V}_t \widehat{V}_t^\top \begin{bmatrix} \widehat{\beta}(b'_1 o) & \widehat{\beta}(b'_2 o) & \cdots & \widehat{\beta}(b'_n o) \end{bmatrix} \begin{bmatrix} \widehat{\Pr}[o|b'_1] & 0 & \cdots & 0 \\ 0 & \widehat{\Pr}[o|b'_2] & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \widehat{\Pr}[o|b'_n] \end{bmatrix} \widehat{V}_{t-1} \widehat{V}_{t-1}^\top.$
return model parameters $\{\widehat{A}_{o,t}\}$.
```

Appendix C. Examples

In this section, we show that our parity with noise and all previously known positive results: full rank HMMs from (Mossel and Roch, 2005; Hsu et al., 2012) and overcomplete HMMs from (Sharan et al., 2017) can be learned by our algorithm using conditional sampling oracle. Note that we will use the alternate form of fidelity which is more amenable to analysis given by

$$\sigma_+\left(S_t^{\frac{1}{2}} \Pr[F_t|H_t]^\top D_t^{-1} \Pr[F_t|H_t] S_t^{\frac{1}{2}}\right) = \sigma_+\left(D_t^{-1/2} \mathbb{E}_{x_{1:t} \sim \mathbf{p}} \left[\Pr[F_t|x_{1:t}] \Pr[F_t|x_{1:t}]^\top\right] D_t^{-1/2}\right)$$

C.1. Parity with noise

We first formally define the distribution induced by parity with noise which has been extensively studied in the computational learning theory (Blum et al., 1994a).

Definition 16 (Parity with noise) *Let $(x_1, \ldots, x_{T-1})$ be a vector in $\{0,1\}^{T-1}$, S a subset of $[T-1]$ and $0 < \alpha < 1/2$. The parity of $(x_1, \ldots, x_{T-1})$ on S is the boolean function $\phi(x_1, \ldots, x_{T-1})$*

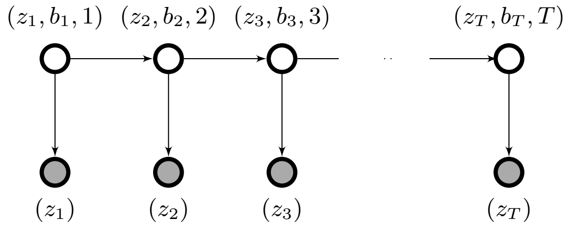

Figure 2: Hidden Markov model for noisy parity. Each hidden state is of the form (z_t, b_t, t) where z_t represents the current bit to be output, b_t is the parity of a secret subset of previous bits and t is the bit position. b_1 is always set to 0 and z_T is set to b_T with probability α and $1 - b_T$ otherwise for some $\alpha \in (0, 1/2)$. For other positions $t \in [T-1]$, transition from hidden state (z_t, b_t, t) goes uniformly randomly to hidden states $(1, b_{t+1}, t+1)$ and $(0, b_{t+1}, t+1)$, where $b_{t+1} = b_t \oplus z_t$ if $t \in I$ and $b_{t+1} = b_t$ otherwise.

which outputs 0 *if the number of ones in the sub-vector* $(x_i)_{i \in S}$ *is even and* 0 *otherwise. Then the distribution induced by HMM for parity with noise is such that the first* $T-1$ *bits are uniform over* $\{0,1\}^{T-1}$ *and the last bit is* $\phi(x_1, \ldots, x_{T-1})$ *with probability* $1-\alpha$ *and* $1 - \phi(x_1, \ldots, x_{T-1})$ *otherwise.*

We now show that Parity with noise HMM satisfies the conditions of Theorem 2. Note that for $\alpha = 1/2$, each bit becomes a random bit. And then as can be seen in the proof below, the fidelity is 1 (since the second eigenvalue goes to 0).

Proposition 17 *The distribution induced by Parity with noise HMM has rank* $\leq 2T$ *and fidelity* $(1-2\alpha)^2/2$ *under a basis of size* ≤ 2 *for every sequence length* $t \in T$.

Proof The claim about rank follows from noting that $\Pr[\cdot|x] = \Pr[\cdot|y]$ is same if both x and y have the same length t; and the subvectors $(x_i)_{i \in S \cap [t]}$ and $(y_i)_{i \in S \cap [t]}$ have the same number of ones modulo 2.

We only need to show that the distribution has large fidelity. The proof is same for all $t \in [T]$, so we prove this for a particular t. For a fixed t, $\Pr[F|x]$ only depends on the parity of the secret subset $(x_i)_{S \cap [t]}$, so there are only two options for $\Pr[F|x]$. Let those be v_1 and v_2 and V be the matrix with these vectors as columns. We choose the basis B_t to be any histories with probability vector v_1 and v_2. Note that if the last bit in the future f matches the parity, then the corresponding probability entry is $(1-\alpha)/2^{T-t-1}$, otherwise it is $\alpha/2^{T-t-1}$. Also, $v_1(f) = (1-\alpha)/2^{T-t-1} \iff v_2(f) = \alpha/2^{T-t-1}$. Therefore, each entry in $d(f)$ is $1/2^{T-t}$.

Our goal is to show that following matrix has large non-zero eigenvalues

$$\mathrm{diag}(d)^{-1/2}\mathbb{E}\left[\Pr[F|x]\Pr[F|x]^\top\right]\mathrm{diag}(d)^{-1/2} = \frac{1}{2}\mathrm{diag}(d)^{-1/2}VV^\top\mathrm{diag}(d)^{-1/2}.$$

Since $\mathrm{diag}(d)^{-1/2}VV^\top\mathrm{diag}(d)^{-1/2}$ has same eigenvalues as $V^\top\mathrm{diag}(d)^{-1}V$, we compute its eigenvalues. The diagonal entries of $V^\top\mathrm{diag}(d)^{-1}V$ are $\alpha^2 + (1-\alpha)^2$ and the off-diagonal entries are $2\alpha(1-\alpha)$. Therefore, the eigenvalues of $V^\top\mathrm{diag}(d)^{-1}V$ are 1 and $(1-2\alpha)^2$. This gives us a lower bound on $\Delta \geq (1-2\alpha)^2/2$. ■

C.2. Full rank HMMs

We first define full rank HMMs studied in (Mossel and Roch, 2005; Hsu et al., 2012). Recall the definition of HMMs (Definition 1).

Definition 18 (Full rank HMMs) *We say an HMM is full rank, if its emission matrix $\mathbb{O}$ and state transition matrix $\mathbb{T}$ have full column rank.*

We next show the distribution induced by full rank HMM can be learned by our algorithm in Theorem 2. Let $P_{2,1}$ be an $O \times O$ matrix with (i,j)th entry $\Pr[o_2 = i, o_1 = j]$. Note that the previous result (Hsu et al., 2012) in this setting depend on smallest eigenvalues of $P_{2,1}$.

Proposition 19 *The distribution induced by full rank HMM has rank S and fidelity $\sigma_{\min}(P_{2,1})^2$ under a basis of size $\leq O$ for every sequence length $t \in T$.*

Proof We note that in the full rank case, we can simplify our algorithm considerably. By our assumptions, $\mathrm{rank}(\Pr[\mathcal{O}|\mathcal{O}]) = S$. And therefore instead of picking all distributions, we can replace $\Pr[F_t|x]$ by one-step probabilities $\Pr[\mathcal{O}|x]$. We can do this because $\Pr[\mathcal{O}|x] = \Pr[\mathcal{O}|\mathcal{O}]\beta(x)$ implies $\Pr[F_t|x] = \Pr[F_t|\mathcal{O}]\beta(x)$ as $\mathrm{rank}(\Pr[F_t|H_t]) = \mathrm{rank}(\Pr[F_t|\mathcal{O}]) = \mathrm{rank}(\Pr[\mathcal{O}|\mathcal{O}]) = S$ and $\mathcal{O} \subset F_t$ where H_t and F_t is the set of all observation sequences of length $t \geq 1$ and $\leq T - t$ respectively. This also means we do not need a different basis for each $t \in [T]$. And only need to show that the following matrix has large eigenvalues:

$$D^{-1/2}\mathbb{E}_{o\sim\mathbf{p}}\left[\Pr[\mathcal{O}|o]\Pr[\mathcal{O}|o]^\top\right]D^{-1/2}$$

where D is a diagonal matrix with $d(o') = \mathbb{E}_{o\in\mathcal{O}}[\Pr[o'|o]]$. Since, eigenvalues of $D^{-1/2}$ are ≥ 1, we are interested in eigenvalues of

$$\mathbb{E}_{o\sim\mathbf{p}}\left[\Pr[\mathcal{O}|o]\Pr[\mathcal{O}|o]^\top\right] = \Pr[\mathcal{O}|\mathcal{O}]U\Pr[\mathcal{O}|\mathcal{O}]^\top$$

where U is a diagonal matrix of size $|\mathcal{O}| \times |\mathcal{O}|$ with its diagonal entries given by $\Pr[o]$. Then, by definition, $P_{2,1} = \Pr[\mathcal{O}|\mathcal{O}]U$. Using this we can lower bound Δ by $\sigma_{\min}(P_{2,1})^2$.

■

C.3. Overcomplete HMMs

We first define the class of overcomplete HMMs which can be learned using techniques in (Sharan et al., 2017). In (Sharan et al., 2017), the authors were concerned with the stationary distribution induced by HMMs. To define their assumptions, let S be the number of hidden states and $\tau = O(\lceil \log_{|\mathcal{O}|} S \rceil)$. Moreover, let H_τ be the set of histories of length τ of the form $x_{-\tau}, \ldots, x_{-1}$, F_τ be the set of histories of length τ of the form $x_0, \ldots, x_\tau$, and $\mathcal{S} = \{1, \ldots, S\}$ be the set of hidden states. Note that by our setting of τ, H_τ, F_τ and $\mathcal{S}$ are all size $O(S)$. Define $\Pr[F_\tau|\mathcal{S}]$ to be a matrix of size $|F_\tau| \times S$ whose $((x_0, \ldots, x_\tau), s)$ entry is given by probability $\Pr[x_0, \ldots, x_\tau | s_0 = s]$. Similarly define $\Pr[H_\tau|\mathcal{S}]$ as the equivalent matrix for time-reversed Markov chain whose $((x_{-\tau}, \ldots, x_{-1}), s)$ entry is given by probability $\Pr[x_{-\tau}, \ldots, x_{-1} | s_0 = s]$.

Sharan et al. (2017) showed efficient algorithms for HMMs under assumptions which imply (a) $\Pr[F_\tau|\mathcal{S}]$ and $\Pr[H_\tau|\mathcal{S}]$ matrices are rank S, (b) the condition number of $\Pr[F_\tau|\mathcal{S}]$ and $\Pr[H_\tau|\mathcal{S}]$ is $\operatorname{poly}(S)$ and (c) every hidden state has stationary probability at least $1/\operatorname{poly}(S)$.

We next show the distribution induced by these HMMs can be learned by our algorithm in Theorem 2. We will not require the uniqueness of columns of $\Pr[\mathcal{O}, \mathcal{S}]$.

Proposition 20 *The distribution induced by HMMs defined above has rank S and fidelity $(\operatorname{poly}(S))^{-1}$ under a basis of size $O(S)$ for every sequence length $t \in T$.*

Proof Just like the full rank case (Proposition 19), we can simplify our algorithm considerably. We choose B to be the set H_τ. Let F be the set of all observation sequences of length $T - \tau$ for some $T > 2\tau$. Now, we can replace $\Pr[F|x]$ by probabilities $\Pr[F_\tau|x]$ in our algorithm. We can do this because $\Pr[F_\tau|x] = \Pr[F_\tau|B]\beta(x)$ implies $\Pr[F|x] = \Pr[F|B]\beta(x)$ as $\operatorname{rank}(\Pr[F|B]) = \operatorname{rank}(\Pr[F_\tau|B]) = S$ and $F_\tau \subset F$. Second, we do not need a different basis for each $t \in [T]$ as $\Pr[F_\tau|x]$ lives in the span of $\Pr[F_\tau|B]$ for every history x as $\operatorname{rank}(\Pr[F_\tau|B]) = \operatorname{rank}(\Pr[F_\tau|H]) = S$ and $B \subset H$ where H is the set of all histories of length $\leq T$. This means we only need to show that the following matrix has large eigenvalues:

$$D^{-1/2}\mathbb{E}_{x_{1:\tau}\sim p}\left[\Pr[F_\tau|x_{1:\tau}]\Pr[F_\tau|x_{1:\tau}]^\top\right]D^{-1/2}$$

where D is a diagonal matrix with entries $d(f) := \mathbb{E}_{b\in B}\Pr[f|b]$ on the diagonal. Since, eigenvalues of $D^{-1/2}$ are ≥ 1, we are interested in eigenvalues of

$$\mathbb{E}_{x_{1:\tau}\sim p}\left[\Pr[F_\tau|x_{1:\tau}]\Pr[F_\tau|x_{1:\tau}]^\top\right] = \Pr[F_\tau|B]K\Pr[F_\tau|B]^\top$$

where K is a diagonal matrix of size $|B| \times |B|$ with diagonal entries given by $k(b) = \Pr[b]$. Define $\Pr[F_\tau H_\tau]$ to be a matrix of size $|F_\tau| \times |H_\tau|$ whose $((x_0, \ldots, x_\tau), (x_{-\tau}, \ldots, x_{-1}))$ entry is given by probability $\Pr[x_{-\tau}, \ldots, x_{-1}, x_0, \ldots, x_\tau]$. Then, by definition, $\Pr[F_\tau|B]K = \Pr[F_\tau H_\tau]$. Using this, each entry of $K < 1$ and every hidden state has stationary probability at least $1/\operatorname{poly}(S)$, we can lower bound Δ by $\sigma_{\min}(\Pr[F_\tau H_\tau])^2 = 1/\operatorname{poly}(S)$.

■

Appendix D. Proofs for Appendix A

Theorem 1 (Learning with exact conditional probabilities) *Assume $\mathcal{O} = \{0, 1\}$. Let $\Pr[\cdot]$ be any rank r distribution over observation sequences of length T. Pick any $0 < \varepsilon, \delta < 1$. Then Algorithm 1 with access to an exact probability oracle and samples from $\Pr[\cdot]$, runs in $\operatorname{poly}(r, T, 1/\varepsilon, \log(1/\delta))$ time and returns an efficiently represented approximation $\widehat{\Pr}[\cdot]$ satisfying $\operatorname{TV}(\Pr, \widehat{\Pr}) \leq \varepsilon$ with probability at least $1 - \delta$.*

Proof First, the total number of rounds are at most rT. This is because by Proposition 9, we increase $\operatorname{rank}(\Pr[\Lambda_\tau|B_\tau])$ by 1 in every round for some $\tau \in [T]$, and $\operatorname{rank}(\Pr[\Lambda_\tau|B_\tau])$ can be at most r by our low rank assumption. So we only need to show that when the algorithm ends, we have found a good estimate. Note that this happens when we can not find a counterexample in Line 5. By Hoeffding's inequality, for $n = O(\log(Tr/\delta)/\varepsilon^2)$, we get with probability $1 - \delta/Tr$ for all $t \in [T]$:

$$\Pr_{x_{1:t}}\left[\Pr[x_{1:t}] \neq \widehat{\Pr}[x_{1:t}]\right] \leq \varepsilon \tag{17}$$

Moreover, define a probability distribution $\overline{\Pr}$ over sequences of length up to T using $\widehat{\Pr}$ as follows: for any $t \in [T]$,

$$\overline{\Pr}[0|x_1 \ldots x_t] = \Pi_{[0,1]} \left[\frac{\boldsymbol{b}_\infty \widehat{\Pr}[x_1 \ldots x_t \Lambda_t]}{\overline{\Pr}[x_1 \ldots x_t]} \right]$$
$$\overline{\Pr}[1|x_1 \ldots x_t] = 1 - \overline{\Pr}[0|x_1 \ldots x_t]$$

where $\Pi_{[0,1]}$ projects onto interval $[0,1]$ and where $\boldsymbol{b}_\infty$ is an indicator vector such that (since each Λ_t contains 0 string)

$$\boldsymbol{b}_\infty \widehat{\Pr}[x_1 \ldots x_t \Lambda_t] = \widehat{\Pr}[x_1 \ldots x_t 0].$$

Note that for a sequence $(x_1 \ldots x_T)$, if for all $t \in [T]$, $\Pr[x_1 \ldots x_t \Lambda_t] = \widehat{\Pr}[x_1 \ldots x_t \Lambda_t]$, then $\Pr[x_1 \ldots x_T] = \overline{\Pr}[x_1 \ldots x_T]$. Therefore, together with Equation (17), we get for each $t \in [T]$ and $o \in \mathcal{O}$,

$$\Pr_{x_{1:t}} \left[\Pr[o|x_{1:t}] \neq \overline{\Pr}[o|x_{1:t}] \right] \leq 2T\varepsilon$$

which implies

$$\mathbb{E}_{x_{1:t}} \left[\left| \Pr[o|x_{1:t}] - \overline{\Pr}[o|x_{1:t}] \right| \right] \leq 2T\varepsilon$$

Using Lemma 21, we get for distribution $\overline{\mathbf{p}}$ corresponding to probability function $\overline{\Pr}$:

$$TV(\mathbf{p}, \overline{\mathbf{p}}) \leq 2T(T+1)\varepsilon$$

Re-substituting the value of ε, we get $TV(\mathbf{p}, \overline{\mathbf{p}}) \leq \varepsilon$ using at most $O(rT^5 \log(Tr/\delta)/\varepsilon^2)$ samples from $\Pr[\cdot]$ and queries to the exact conditional probability oracle. ■

Next, we need a technical lemma which allows us to test for TV distance using just conditional samples. This lemma will imply that if our algorithm does not find a violation, then with high probability our estimate should be close to true distribution in TV distance.

Lemma 21 (Substitute for TV oracle) *Let $\mathbf{p}$ and $\overline{\mathbf{p}}$ be two probability distributions over observation sequences of length T with probability functions $\Pr$ and $\overline{\Pr}$ respectively. Suppose we have for all $t \in [T]$ and for all $o \in \mathcal{O}$*

$$\mathbb{E}_{x_{1:t} \sim \mathbf{p}} \left[\left| \overline{\Pr}[o|x_{1:t}] - \Pr[o|x_{1:t}] \right| \right] \leq \varepsilon\,.$$

Then

$$TV(\mathbf{p}, \overline{\mathbf{p}}) = \frac{1}{2} \sum_{x_{1:T}} |(\Pr[x_{1:T}] - \overline{\Pr}[x_{1:T}])| \leq \frac{(T+1)|O|\varepsilon}{2}$$

Proof We prove this by induction. Assume

$$\sum_{x_{1:t-1}} \left| (\Pr[x_{1:t-1}] - \overline{\Pr}[x_{1:t-1}]) \right| \leq t|O|\varepsilon$$

Then,

$$\begin{aligned}
&\sum_{x_{1:t}} \left|(\Pr[x_{1:t}] - \overline{\Pr}[x_{1:t}])\right| \\
&= \sum_{x_{1:t}} \left|\Pr[x_{1:t-1}] \Pr[x_t|x_{1:t-1}] - \overline{\Pr}[x_{1:t-1}]\overline{\Pr}[x_t|x_{1:t-1}]\right| \\
&\leq \sum_{x_{1:t}} \Pr[x_{1:t-1}] \cdot \left|\Pr[x_t|x_{1:t-1}] - \overline{\Pr}[x_t|x_{1:t-1}]\right| + \sum_{x_{1:t}} \left|(\Pr[x_{1:t-1}] - \overline{\Pr}[x_{1:t-1}])\right| \cdot \overline{\Pr}[x_t|x_{1:t-1}]
\end{aligned}$$

We handle the two terms separately. The first term

$$\begin{aligned}
&\sum_{x_{1:t}} \Pr[x_{1:t-1}] \cdot \left|\Pr[x_t|x_{1:t-1}] - \overline{\Pr}[x_t|x_{1:t-1}]\right| \\
&= \sum_{x \in \mathcal{O}} \sum_{x_{1:t-1}} \Pr[x_{1:t-1}] \cdot \left|\Pr[x|x_{1:t-1}] - \overline{\Pr}[x|x_{1:t-1}]\right| \\
&= \sum_{x \in \mathcal{O}} \mathbb{E}_{x_{1:t-1}} \left[\left|\Pr[x|x_{1:t-1}] - \overline{\Pr}[x|x_{1:t-1}]\right|\right] \\
&\leq |O|\varepsilon
\end{aligned}$$

where the last step follows from our assumption. The second term

$$\begin{aligned}
&\sum_{x_{1:t}} \left|(\Pr[x_{1:t-1}] - \overline{\Pr}[x_{1:t-1}])\right| \cdot \overline{\Pr}[x_t|x_{1:t-1}] \\
&= \sum_{x_{1:t-1}} \left|\Pr[x_{1:t-1}] - \overline{\Pr}[x_{1:t-1}]\right| \cdot \sum_{x \in \mathcal{O}} \overline{\Pr}[x|x_{1:t-1}] \\
&\leq \sum_{x_{1:t-1}} \left|\Pr[x_{1:t-1}] - \overline{\Pr}[x_{1:t-1}]\right| \\
&\leq t|O|\varepsilon
\end{aligned}$$

■

Appendix E. Proofs for Appendix B

In this section, we introduce conditions for efficient learnability of low rank distributions using only conditional sampling oracle. We will fill out missing details from Appendix B.

E.1. Properties of operator A_o

Let $\{B_t\}_{t \in [T]}$ be some basis of distribution $\mathbf{p}$ (as defined in Definition 10). In the previous section, we defined the operators $A_{o,t}$ under basis $\{B_t\}_{t \in [T]}$ in Proposition 8. We now prove some properties of this operator and the associated coefficients. We use the same notation as the previous section, except we abuse notation for the set of futures F_t and redefine it as $F_t := \mathcal{O}^{T-t}$ (so instead of all futures of length up to $T - t$, now its only futures of length exactly $T - t$). Consider the eigenvalue decomposition of the covariance matrix associated to B_t where D_t is a diagonal matrix of size $|F_t| \times |F_t|$ with entries $d_t(f) := \mathbb{E}_{b \in B_t} \Pr[f|b]$ on the diagonal:

$$\Pr[F_t|B_t]^\top D_t^{-1} \Pr[F_t|B_t] = \begin{bmatrix} V_t & V_t^\perp \end{bmatrix} \begin{bmatrix} M_t & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_t^\top \\ V_t^{\perp\top} \end{bmatrix}$$

Here, M_t is a set of all non-zero eigenvalues, V_t is the eigenspace corresponding to non-zero eigenvalues and $V_t^\perp$ is the eigenspace corresponding to zero eigenvalues. A basic property of $\operatorname{span}(V_t^\perp)$ follows from the definition of basis $\{B_t\}_{t\in[T]}$ as $d_t(f) = 0$ implies $\Pr(f|x) = 0$ for all f of length $T - t$ and x of length t. For such future f, we define $\Pr(f|x)/d_t(f) = 1$ for the proposition below to make sense. And we will see in Appendix E.6 how these futures do not matter in the estimation.

Proposition 22 $\operatorname{span}(V_t^\perp) = \ker(\Pr[F_t|B_t]^\top D^{-1} \Pr[F_t|B_t]) = \ker(\Pr[F_t|B_t])$

Recall we denote the coefficients associated to history x of length t under basis B_t by $\beta(x)$ given by:

$$\Pr[F_t|B_t]\beta(x) = \Pr[F_t|x]\,. \tag{18}$$

By Proposition 22, we can assume that $\beta(x) \in \operatorname{span}(V_t)$ without loss of generality. We now show that the coefficients $\beta(x)$ satisfy some nice properties.

Proposition 23 *Let $\beta(x) \in \operatorname{span}(V_t)$ be coefficients associated to history x. Then, the following statements are true:*

1. *The coefficients $\beta(x)$ are uniquely defined in $\operatorname{span}(V_t)$. Formally, Let $\beta'(x)$ be any other coefficients which satisfy Equation (18). Then,*

$$P_{V_t}\beta'(x) = \beta(x)\,,$$

where P_{V_t} is the projection matrix onto subspace V_t.

2. *The coefficients sum to one, even though some of the entries could be negative*

$$\mathbf{1}^\top \beta(x) = 1$$

where $\mathbf{1}$ is all ones column vector.

Proof First, recall by definition, the coefficients satisfy Equation (18). The first claim follows from $\operatorname{span}(V_t^\perp) = \ker(\Pr[F_t|B_t])$ (Proposition 22). Finally, the last claim follows by multiplying both sides in Equation (18) by all ones row vector $\mathbf{1}^\top$

$$1 = \mathbf{1}^\top \Pr[F_t|x] = \mathbf{1}^\top \Pr[F_t|B_t]\beta(x) = \mathbf{1}^\top \beta(x)$$

where the last equation follows by noting that $\mathbf{1}^\top \Pr[F_t|x]$ for any probability vector $\Pr[F_t|x]$ is 1 (recall F_t is set of all futures of length exactly $T - t$). ■

Even though, these are exponentially many coefficients, we next show existence of operators which can be used to construct these coefficients.

Proposition 24 *Let $A_{o,t}$ be defined using basis $B_t = \{b_1, \ldots, b_n\}$ and B_{t+1} as:*

$$A_{o,t} = \begin{bmatrix} \beta(b_1 o) & \beta(b_2 o) & \cdots & \beta(b_n o) \end{bmatrix} \begin{bmatrix} \Pr[o|b_1] & 0 & \cdots & 0 \\ 0 & \Pr[o|b_2] & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \Pr[o|b_n] \end{bmatrix}.$$

Then, it satisfies the following:

1. *$span(A_{o,t}) \subset span(V_{t+1})$*
2. *$\ker(A_{o,t}) \supset span(V_t^{\perp})$*
3. *$A_{o,t}\beta(x) = \Pr[o|x]\beta(xo)$*
4. *$\mathbf{1}^\top A_{o,t}\beta(x) = \Pr[o|x]$*

Proof The first two properties follow from the definition of $A_{o,t}$. Let F_t be all observation sequences of length $T - t$. Similar to proof of Proposition 8, $A_{o,t}$ satisfies

$$\Pr[F_{t+1}|B_{t+1}]A_{o,t} = \Pr[oF_{t+1}|B_t]P_{V_t}$$

Since oF_{t+1} is a subset of F_t, we get by multiplying $\beta(x)$ on both sides, we get

$$\begin{aligned} \Pr[F_{t+1}|B_{t+1}]A_{o,t}\beta(x) &= \Pr[oF_{t+1}|B_t]\beta(x) && \text{(as } \beta(x) \in \text{span}(V_t)\text{)} \\ &= \Pr[oF_{t+1}|x] && \text{(as } oF_{t+1} \text{ is a subset of } F_t\text{)} \\ &= \Pr[F_{t+1}|xo]\Pr[o|x] && \text{(by Bayes rule)} \\ &= \Pr[F_{t+1}|B_{t+1}]\beta(xo)\Pr[o|x] \end{aligned}$$

By uniqueness of $\beta(xo)$ (Item 1) and that $\text{span}(A_{o,t}) \subset \text{span}(V_{t+1})$ (Item 1), we get that

$$\frac{A_{o,t}\beta(x)}{\Pr[o|x]} = P_{V_{t+1}} \frac{A_{o,t}\beta(x)}{\Pr[o|x]} = \beta(xo)$$

The last one follows from multiplying both sides above by all ones row vector and then using $\mathbf{1}^\top\beta(x) = 1$ (Item 2).

$$\mathbf{1}^\top \frac{A_{o,t}\beta(x)}{\Pr[o|x]} = \mathbf{1}^\top\beta(xo) = 1$$

■

E.2. Learning operators

In this section, we will assume knowledge of a Δ-robust basis $\{B_t\}_{t\in[T]}$ (Definition 12) and show how to learn an approximation of operators A_o using it. For this, we need to learn approximations of projections P_{V_t} and coefficients $\beta(b_i o)$ for the one-step extensions of basis $B_t = \{b_1, \ldots, b_n\}$. Towards this end, we would have to solve for $\beta(x)$ in the following linear equation: $\Pr[F|x] = \Pr[F_t|B_t]\beta(x)$. This requires estimating the following:

$$q(x) = \Pr[F_t|B_t]^\top \text{diag}(d_t)^{-1}\Pr[F|x] \quad \text{and} \quad \Sigma_{B_t} = \Pr[F_t|B_t]^\top \text{diag}(d_t)^{-1}\Pr[F_t|B_t]$$

which we show how to do in Appendix E.6. Now, we use these estimates to learn approximations of projection P_{V_t}. Recall Σ_{B_t} is the covariance matrix with $r_t < r$ non-zero eigenvalues and let $\widehat{\Sigma}_{B_t}$ be an approximation for covariance matrix Σ_{B_t} from Corollary 35. Compute SVD of $\widehat{\Sigma}_{B_t}$, and let $\widehat{V_t}$ be the matrix of top r_t eigenvectors (or equivalently corresponding to eigenvalues $> \Delta/2$ according to the proof of proposition below). Then, $\widehat{V_t}$ is close to V_t using Davis-Kahan $\sin(\theta)$ theorem (Corollary 47):

Proposition 25 *Let* $\mathbf{p}$ *be any rank* r *distribution over observation sequences of length* T*. Assume knowledge of a* Δ*-robust basis* $\{B_t\}_{t\in[T]}$ *for distribution* $\mathbf{p}$*. Let* $\Sigma_{B_t}, \widehat{\Sigma}_{B_t}, V_t$ *and* $\widehat{V_t}$ *be as defined above. Then, we can build an approximate projection* $P_{\widehat{V_t}}$ *such that*

$$||P_{V_t} - P_{\widehat{V_t}}||_F \leq O\left(\frac{\sqrt{r}\cdot||\widehat{\Sigma}_{B_t} - \Sigma_{B_t}||_2}{\Delta}\right)$$

for $||\widehat{\Sigma}_{B_t} - \Sigma_{B_t}||_2 \leq \Delta/2$.

Proof Let estimate Σ_{B_t} be such that

$$||\widehat{\Sigma}_{B_t} - \Sigma_{B_t}||_2 \leq ||\widehat{\Sigma}_{B_t} - \Sigma_{B_t}||_F \leq \alpha \tag{19}$$

The claim now follows from Davis-Kahan $\sin(\theta)$ theorem. From our assumptions, eigenvalues of V_t are $> \Delta$ and rest are all 0. Moreover, from Equation (19) and Weyl's inequality, all the eigenvalues associated to $\widehat{V}_t^{\perp}$ are $< \alpha$. Then, for $\alpha < \Delta/2$, we get using Corollary 47,

$$||P_{V_t} - P_{\widehat{V_t}}||_F \leq \frac{2\sqrt{2r}\alpha}{\Delta}$$

■

To compute an approximation for operators A_o, we also need to get the coefficients for one-step extensions of elements in the basis.

Proposition 26 *Let* $\mathbf{p}$ *be any rank* r *distribution over observation sequences of length* T*. Assume knowledge of a* Δ*-robust basis* $\{B_t\}_{t\in[T]}$ *for distribution* $\mathbf{p}$*. Also, assume the coefficients under basis* $\{B_t\}_{t\in[T]}$ *are bounded i.e.* $||\beta(h)||_2 \leq c$ *for all histories* $h \in \mathcal{O}^{\leq T}$*. Suppose we have estimates* $\widehat{q}(b_i o)$ *and* $\widehat{\Sigma}_{B_t}$ *such that*

$$\max\left\{||\widehat{q}(b_i o) - q(b_i o)||_2\,, ||\widehat{\Sigma}_{B_t} - \Sigma_{B_t}||_2\right\} \leq \alpha\,.$$

Define $\widehat{\beta}(b_i o)$*, approximation of* $\beta(b_i o)$*, for* $\lambda = 4\alpha^2/c^2$*:*

$$\widehat{\beta}(b_i o) = \operatorname*{argmin}_z ||\widehat{\Sigma}_{B_t} - \widehat{q}(b_i o)||_2^2 + \lambda||z||_2^2\,.$$

Then, the following are true:

1. *the norm of the approximation* $\widehat{\beta}(b_i o)$ *is bounded:*

$$||\widehat{\beta}(b_i o)||_2 \leq \sqrt{2}c$$

2. *the approximation* $\widehat{\beta}(b_io)$ *is accurate in the* $\mathrm{span}(V_{t+1})$*:*

$$||P_{V_{t+1}}\beta(b_io) - P_{V_{t+1}}\widehat{\beta}(b_io)||_2 \leq O\left(\frac{\alpha}{\Delta}\right) .$$

Proof Let

$$||\widehat{q}(b_io) - q(b_io)||_2 \, , ||\widehat{\Sigma}_{B_t} - \Sigma_{B_t}||_2 \leq \alpha$$
$$\lambda = 4\alpha^2/c^2$$

We know that $\beta(b_io)$ satisfies

$$\Sigma_{B_t}\beta(b_io) = q(b_io)$$

and approximate it by the following program (for $\lambda = 4\alpha^2$):

$$\widehat{\beta}(b_io) = \underset{z}{\mathrm{argmin}}\, ||\widehat{\Sigma}_{B_t} z - \widehat{q}(b_io)||_2^2 + \lambda||z||^2$$

First, we can see that the above error is small as

$$||\widehat{\Sigma}_{B_t}\widehat{\beta}(b_io) - \widehat{q}(b_io)||_2^2 + \lambda||\widehat{\beta}(b_io)||_2^2 \leq ||\widehat{\Sigma}_{B_t}\beta(b_io) - \widehat{q}(b_io)||_2^2 + \lambda||\beta(b_io)||_2^2 \leq 8\alpha^2$$

as $||\beta(b_io)||_2 \leq c$ and

$$||\widehat{\Sigma}_{B_t}\beta(b_io) - \widehat{q}(b_io)||_2 \leq ||\widehat{\Sigma}_{B_t}\beta(b_io) - \Sigma_{B_t}\beta(b_io)||_2 + ||q(b_io) - \widehat{q}(b_io)||_2 \leq 2\alpha$$

Note that this also means

$$||\widehat{\beta}(b_io)|| \leq \sqrt{\frac{8\alpha^2c^2}{4\alpha^2}} = \sqrt{2}c$$

Using these, we get

$$\begin{aligned}
&||\Sigma_{B_t}\beta(b_io) - \Sigma_{B_t}\widehat{\beta}(b_io)|| \\
&\leq ||\Sigma_{B_t}\beta(b_io) - \widehat{\Sigma}_{B_t}\widehat{\beta}(b_io) + \widehat{\Sigma}_{B_t}\widehat{\beta}(b_io) - \Sigma_{B_t}\widehat{\beta}(b_io)|| \\
&\leq ||\Sigma_{B_t}\beta(b_io) - \widehat{\Sigma}_{B_t}\widehat{\beta}(b_io)|| + ||\widehat{\Sigma}_{B_t}\widehat{\beta}(b_io) - \Sigma_{B_t}\widehat{\beta}(b_io)|| \\
&\leq ||\Sigma_{B_t}\beta(b_io) - q(b_io) + q(b_io) - \widehat{q}(b_io) + \widehat{q}(b_io) - \widehat{\Sigma}_{B_t}\widehat{\beta}(b_io)|| + ||\widehat{\Sigma}_{B_t} - \Sigma_{B_t}||||\widehat{\beta}(b_io)|| \\
&\leq ||\Sigma_{B_t}\beta(b_io) - q(b_io)|| + ||q(b_io) - \widehat{q}(b_io)|| + ||\widehat{q}(b_io) - \widehat{\Sigma}_{B_t}\widehat{\beta}(b_io)|| + ||\widehat{\Sigma}_{B_t} - \Sigma_{B_t}||||\widehat{\beta}(b_io)|| \\
&\leq \alpha + 2\sqrt{2}\alpha + \sqrt{2}c\alpha \leq 6\alpha
\end{aligned}$$

for $c < 1$. Now, using that our assumption on Σ_{B_t},

$$||P_{V_{t+1}}\beta(b_io) - P_{V_{t+1}}\widehat{\beta}(b_io)||_2 \leq \frac{6\alpha}{\Delta}$$

■

Proposition 27 *Let* $\mathbf{p}$ *be any rank* r *distribution over observation sequences of length* T *and* $\{B_t\}_{t\in[T]}$ *be some basis of distribution* $\mathbf{p}$*. Then, we can build an estimate* $\mathrm{diag}(\widehat{\Pr}[o|B_t])$ *for every* $o \in \mathcal{O}$ *and* $t \in [T]$*, such that with probability* $1-\delta$

$$\|\mathrm{diag}(\Pr[o|B_t]) - \mathrm{diag}(\widehat{\Pr}[o|B_t])\|_2 \leq \varepsilon$$

using $\widetilde{O}(|\mathcal{O}|Tn^3/\varepsilon^2 \log(1/\delta))$ *conditional samples.*

Proof This follows from Hoeffding's inequality Proposition 45. ■

Using the approximations above, we are in position to present the approximation error for our estimate of operator $A_{o,t}$. Let $\beta(B_t)$ denote a matrix with the coefficients $\beta(b_i)$ as its columns. We also show that the error can be written with small coefficients on the columns on matrix $\beta(B_{t+1})$ and $V_{t+1}^{\perp}$.

Lemma 28 (Restatement of Lemma 13) *Let* $\mathbf{p}$ *be any rank* r *distribution over observation sequences of length* T*. Assume knowledge of a* Δ*-robust basis* $\{B_t\}_{t\in[T]}$ *for distribution* $\mathbf{p}$*. Also, assume the coefficients under basis* $\{B_t\}_{t\in[T]}$ *are bounded i.e.* $||\beta(h)||_2 \le c$ *for all histories* $h \in \mathcal{O}^{\le T}$*. Then, we can build an estimate* $\widehat{A}_{o,t}$ *using*

$$\widetilde{O}\left(\frac{c^8 r^3 n^{21}|\mathcal{O}|^3 T^5}{\Delta^6 \varepsilon^6}\log^2\left(\frac{1}{\delta}\right)\right)$$

many conditional samples such that with probability $1-\delta$*, for any unit vector* v

$$(\widehat{A}_{o,t} - A_{o,t})v = \beta(B_{t+1})\alpha(o,v) + V_{t+1}^{\perp}\alpha^{\perp}(o,v)$$

with bounded errors (here $\alpha(o,v)$ *and* $\alpha^{\perp}(o,v)$ *are column vectors of size* $|B_{t+1}|$ *and* $|V_{t+1}^{\perp}|$ *respectively):*

$$\|\alpha(o,v)\|_1, \|\alpha^{\perp}(o,v)\|_1 \le \varepsilon\,.$$

Proof Using $P_{\widehat{V}_t}$, $\widehat{\beta}(B_t o)$ and $\mathrm{diag}(\widehat{\Pr}[o|B_t])$ from Propositions 25 to 27, define operator $\widehat{A}_{o,t}$, an approximation of $A_{o,t}$, as

$$\widehat{A}_{o,t} = P_{\widehat{V}_{t+1}}\widehat{\beta}(oB_t)\mathrm{diag}(\widehat{\Pr}[o|B_t])P_{\widehat{V}_t}$$

Let

$$||P_{\widehat{V}_{t+1}} - P_{V_{t+1}}||_2\,, ||P_{\widehat{V}_t} - P_{V_t}||_2 \le \alpha_1 = O\left(\frac{\sqrt{r}n\varepsilon}{\Delta}\right)$$

$$||P_{V_{t+1}}\widehat{\beta}(B_t o) - P_{V_{t+1}}\beta(B_t o)||_2 \le \alpha_2 \le \sqrt{n}\max_{b\in B_t}||P_{V_{t+1}}\widehat{\beta}(bo) - P_{V_{t+1}}\beta(bo)||_2 \le O\left(\frac{n^{3/2}\varepsilon}{\Delta}\right)$$

$$||\mathrm{diag}(\widehat{\Pr}[o|B_t]) - \mathrm{diag}(\Pr[o|B_t])||_2 \le \alpha_3 = \varepsilon$$

Also, note that $||\widehat{\beta}(b_i o)||_2 \le \sqrt{2}c$, $||\mathrm{diag}(\widehat{\Pr}[o|B_t])||_2 \le 2$ and $||P_{\widehat{V}_{t+1}}||_2\,, ||P_{\widehat{V}_t}||_2 \le 1$. To prove our main claim, we will first show that for any unit vector v

$$(\widehat{A}_{o,t} - A_{o,t})v = V_{t+1}e(o,v) + V_{t+1}^{\perp}\alpha^{\perp}(o,v)$$

with $\|e(o,v)\|_2, \|\alpha^{\perp}\|_2 \le 4\sqrt{2}c\alpha_1 + \sqrt{2}c\alpha_3 + \alpha_2$. To prove this, we will show that

$$\begin{aligned}
&\left\|\widehat{A}_{o,t} - A_{o,t}\right\| \\
&= \left\|P_{\widehat{V}_{t+1}}\widehat{\beta}(oB_t)\mathrm{diag}(\widehat{\Pr}[o|B_t])P_{\widehat{V}_t} - P_{V_{t+1}}\beta(oB_t)\mathrm{diag}(\Pr[o|B_t])P_{V_t}\right\| \\
&= \left\|(P_{\widehat{V}_{t+1}} - P_{V_{t+1}})\widehat{\beta}(oB_t)\mathrm{diag}(\widehat{\Pr}[o|B_t])P_{\widehat{V}_t}\right\| + \left\|P_{V_{t+1}}\widehat{\beta}(oB_t)\mathrm{diag}(\widehat{\Pr}[o|B_t])(P_{\widehat{V}_t} - P_{V_t})\right\| \\
&+ \left\|P_{V_{t+1}}\widehat{\beta}(oB_t)(\mathrm{diag}(\widehat{\Pr}[o|B_t]) - \mathrm{diag}(\Pr[o|B_t]))P_{V_t}\right\| \\
&+ \left\|P_{V_{t+1}}(\widehat{\beta}(oB_t) - \beta(oB_t))\mathrm{diag}(\Pr[o|B_t])P_{V_t}\right\|
\end{aligned}$$

The first term is bounded by $2\sqrt{2}c\alpha_1$, second term by $2\sqrt{2}c\alpha_1$, similarly third term by $\sqrt{2}c\alpha_3$ and the last term by α_2. Therefore, we get that

$$\left\|\widehat{A}_{o,t} - A_{o,t}\right\|_2 \leq 4\sqrt{2}c\alpha_1 + \sqrt{2}c\alpha_3 + \alpha_2 \tag{20}$$

which implies for any unit vector v

$$(\widehat{A}_{o,t} - A_{o,t})v = V_{t+1}e(o,v) + V_{t+1}^{\perp}\alpha^{\perp}(o,v) \tag{21}$$

with $\|e(o,v)\|_2, \|\alpha^{\perp}(o,v)\|_2 \leq \|\widehat{A}_{o,t} - A_{o,t}\|_2 \leq 4\sqrt{2}c\alpha_1 + \sqrt{2}c\alpha_3 + \alpha_2$. To complete our proof, we will show that

$$V_{t+1} = \beta(B_{t+1})V_{t+1} \tag{22}$$

This is enough, as this gives,

$$(\widehat{A}_{o,t} - A_{o,t})v = \beta(B_{t+1})\alpha(o,v) + V_{t+1}^{\perp}\alpha^{\perp}(o,v)$$

where $\alpha(o,v) = V_{t+1}e(o,v)$ as $||\alpha(o,v)||_2 \leq ||V_{t+1}e(o,v)|| \leq ||e(o,v)||$ as V_{t+1} is a unit norm matrix.

We now prove our claim. First note that $\Pr[F|B_{t+1}]I = \Pr[F|B_{t+1}]$ and therefore by uniqueness of $\beta(B_{t+1})$,

$$\begin{aligned} & \beta(B_{t+1}) = V_{t+1}V_{t+1}^{\top}I \\ \implies & \beta(B_{t+1})V_{t+1} = V_{t+1} \qquad \text{(by right multiplying by } V_{t+1} \text{ and } V_{t+1}^{\top}V_{t+1} = I) \end{aligned}$$

This whole process requires $O(n^2|\mathcal{O}|Tm_{\varepsilon})$ many conditional samples with m_{ε} defined in Proposition 32. Substituting in $\varepsilon = \Delta\varepsilon'/(c\sqrt{r}n^{3/2})$ gives the result. ■

E.3. Perturbation analysis: Error in coefficients

Our approach for learning the distribution p is to learn approximations of operators $\widehat{A}_{o,t}$ and use them to compute probabilities

$$\widehat{\Pr}[x_{1:T}] = \widehat{A}_{x_T,T-1}\widehat{A}_{x_{T-1},T-2}\ldots\widehat{A}_{x_1,0}$$

For clarity, let $A_{x_{1:t}}$ and $\widehat{A}_{x_{1:t}}$ represent the product of matrices $A_{x_t,t-1}\ldots A_{x_1,0}$ and $\widehat{A}_{x_t,t-1}\ldots\widehat{A}_{x_1,0}$ respectively. Similarly, we use $x_{1:t}$ to represent the sequence $(x_1,\ldots,x_t)$. In this section, we now present a technical lemma showing that the errors in our estimates of $\widehat{\Pr}[x_{1:T}]$ scales linearly with errors in our approximate operators $\widehat{A}_{o,t}$.

Proposition 29 *Assume for every $t \in [T]$ and observation $o \in \mathcal{O}$, we have operator $A_{o,t}$ which satisfies for any vector v*

$$(\widehat{A}_{o,t} - A_{o,t})v = \beta(B_{t+1})\alpha(o,v) + V_{t+1}^{\perp}\alpha^{\perp}(o,v) \tag{23}$$

with $\|\alpha(o,v)\|_1, \|\alpha^{\perp}(o,v)\|_1 \leq \varepsilon||v||_2$. Then,

$$(\widehat{A}_{x_{1:t}} - A_{x_{1:t}})\beta(\varphi) = \beta(B_{x_{1:t}})\gamma_{x_{1:t}} + V^{\perp}\gamma_{x_{1:t}}^{\perp}$$

where $B_{x_{1:t}}$ is a set of observation sequences of length $t+1$ of size exponential in t but the ℓ_1 norm of these coefficients grows nicely:

$$\sum_{x_t \in \mathcal{O}} \|\gamma_{x_{1:t}}\|_1 \le |\mathcal{O}|\varepsilon \Pr[x_{1:t-1}] + (1+|\mathcal{O}|\varepsilon)\|\gamma_{x_{1:t-1}}\|_1 + |\mathcal{O}|\varepsilon\|\gamma^{\perp}_{x_{1:t-1}}\|_1 \tag{24}$$

$$\sum_{x_t \in \mathcal{O}} \|\gamma^{\perp}_{x_{1:t}}\|_1 \le |\mathcal{O}|\varepsilon \Pr[x_{1:t-1}] + |\mathcal{O}|\varepsilon\|\gamma_{x_{1:t-1}}\|_1 + |\mathcal{O}|\varepsilon\|\gamma^{\perp}_{x_{1:t-1}}\|_1 \tag{25}$$

Proof First, by our assumption, for any observation sequence $x = (x_1, \ldots, x_{t-1})$

$$(\widehat{A}_{o,t-1} - A_{o,t-1})\beta(x) = \beta(B_t)\alpha(o,x) + V_t^{\perp}\alpha^{\perp}(o,x) \tag{26}$$

with $\|\alpha(o,x)\|_1, \|\alpha^{\perp}(o,x)\|_1 \le \varepsilon$ as $||\beta(x)||_2 \le 1$. Next, the following recursive relation holds

$$\begin{aligned}
&(\widehat{A}_{x_{1:t}} - A_{x_{1:t}})\beta(\varphi)\\
&= (\widehat{A}_{x_{1:t}} - \widehat{A}_{x_t}A_{x_{1:t-1}} + \widehat{A}_{x_t}A_{x_{1:t-1}} - A_{x_{1:t}})\beta(\varphi)\\
&= (\widehat{A}_{x_t} - A_{x_t})A_{x_{1:t-1}}\beta(\varphi) + \widehat{A}_{x_t}(\widehat{A}_{x_{1:t-1}} - A_{x_{1:t-1}})\beta(\varphi)\\
&= (\widehat{A}_{x_t} - A_{x_t})A_{x_{1:t-1}}\beta(\varphi) + A_{x_t}(\widehat{A}_{x_{1:t-1}} - A_{x_{1:t-1}})\beta(\varphi) + (\widehat{A}_{x_t} - A_{x_t})(\widehat{A}_{x_{1:t-1}} - A_{x_{1:t-1}})\beta(\varphi)
\end{aligned}$$

We will bound the three terms separately for each x_t. We start by bounding the first term

$$\begin{aligned}
&(\widehat{A}_{x_t} - A_{x_t})A_{x_{1:t-1}}\beta(\varphi)\\
&= (\widehat{A}_{x_t} - A_{x_t})\Pr[x_{1:t-1}]\beta(x_{1:t-1}) && \text{(by Proposition 24)}\\
&= \Pr[x_{1:t-1}]\beta(B_t)\alpha(x_t, x_{1:t-1}) + \Pr[x_{1:t-1}]V_t^{\perp}\alpha^{\perp}(x_t, x_{1:t-1}) && \text{(by Equation (26))}\\
&= \beta(B_t)\left(\Pr[x_{1:t-1}]\alpha(x_t, x_{1:t-1})\right) + V_t^{\perp}\left(\Pr[x_{1:t-1}]\alpha^{\perp}(x_t, x_{1:t-1})\right)
\end{aligned}$$

and here we can see that

$$\begin{aligned}
\|\Pr[x_{1:t-1}]\alpha(x_t, x_{1:t-1})\|_1 \le \Pr[x_{1:t-1}]\|\alpha(x_t, x_{1:t-1})\|_1 \le \varepsilon \Pr[x_{1:t-1}]\\
\|\Pr[x_{1:t-1}]\alpha^{\perp}(x_t, x_{1:t-1})\|_1 \le \Pr[x_{1:t-1}]\|\alpha^{\perp}(x_t, x_{1:t-1})\|_1 \le \varepsilon \Pr[x_{1:t-1}]
\end{aligned}$$

where we used $||\alpha^{\perp}(x_t, x_{1:t-1})||_1, ||\alpha(x_t, x_{1:t-1})||_1 \le \varepsilon||\beta(x_{1:t-1})||_2 \le \varepsilon$. This gives the first term in Equations (24) and (25) (with the $|\mathcal{O}|$ factor for sum over x_t). We bound the remaining terms by induction. Assume

$$(\widehat{A}_{x_{1:t-1}} - A_{x_{1:t-1}})\beta(\varphi) = \beta(B_{x_{1:t-1}})\gamma_{x_{1:t-1}} + V_{t-1}^{\perp}\gamma^{\perp}_{x_{1:t-1}} \tag{27}$$

where $B_{x_{1:t-1}}$ is a set of observation sequences of length $t-1$. Let's first bound the second term.

$$\begin{aligned}
&A_{x_t}(\widehat{A}_{x_{1:t-1}} - A_{x_{1:t-1}})\beta(\varphi)\\
&= A_{x_t}(\beta(B_{x_{1:t-1}})\gamma_{x_{1:t-1}} + V_{t-1}^{\perp}\gamma^{\perp}_{x_{1:t-1}}) && \text{(by Equation (27))}\\
&= A_{x_t}(\beta(B_{x_{1:t-1}})\gamma_{x_{1:t-1}}) && \text{(by Proposition 24)}\\
&= \beta(x_t B_{x_{1:t-1}})\mathrm{diag}(\Pr[x_t|B_{x_{1:t-1}}])\gamma_{x_{1:t-1}} && \text{(by Proposition 24)}
\end{aligned}$$

We can bound the ℓ_1 norm of the coefficients as

$$\sum_{x_t\in\mathcal{O}}\sum_{b_i\in B_{x_{1:t-1}}} |\Pr[x_t|b_i]\gamma_i| = \sum_{x_t\in\mathcal{O}}\sum_{b_i\in B_{x_{1:t-1}}} \Pr[x_t|b_i]|\gamma_i| = \sum_{b_i\in B_{x_{1:t-1}}} |\gamma_i| = \|\gamma_{x_{1:t-1}}\|_1$$

where the second last step follows from $\sum_{x_t\in\mathcal{O}}\Pr[x_t|b_i]=1$. For clarity, let $\alpha(x_t, B_{x_{1:t-1}})$ represent a matrix with its column given by $\alpha(x_t, b)$ for $b\in B_{x_{1:t-1}}$. Similarly, define $\alpha^{\perp}(x_t, B_{x_{1:t-1}})$, $\alpha(x_t, V_{t-1}^{\perp})$ and $\alpha^{\perp}(x_t, V_{t-1}^{\perp})$. We can then similarly bound the remaining term.

$$\begin{aligned}
&(\widehat{A}_{x_t}-A_{x_t})(\widehat{A}_{x_{1:t-1}}-A_{x_{1:t-1}})\beta(\varphi)\\
&=(\widehat{A}_{x_t}-A_{x_t})(\beta(B_{x_{1:t-1}})\gamma_{x_{1:t-1}}+V_{t-1}^{\perp}\gamma_{x_{1:t-1}}^{\perp}) && \text{(by Equation (27))}\\
&=\Big[\beta(B_t)\alpha(x_t,B_{x_{1:t-1}})\gamma_{x_{1:t-1}}+V_t^{\perp}\alpha^{\perp}(x_t,B_{x_{1:t-1}})\gamma_{x_{1:t-1}}\Big] && \text{(by Equation (26))}\\
&+\Big[\beta(B_t)\alpha(x_t,V_t^{\perp})\gamma_{x_{1:t-1}}^{\perp}+V_t^{\perp}\alpha^{\perp}(x_t,V_t^{\perp})\gamma_{x_{1:t-1}}^{\perp}\Big] && \text{(by Lemma 28)}
\end{aligned}$$

Each of these terms can be bounded similarly. We show how to bound the first term:

$$\|\sum_{b_i\in B_{x_{1:t-1}}}\gamma_i\alpha(x_t,b_i)\|_1 \le \sum_{b_i\in B_{x_{1:t-1}}}|\gamma_i|\cdot\|\alpha(x_t,b_i)\|_1 \le \varepsilon\|\gamma_{t-1}\| \tag{28}$$

where we used $||\alpha(x_t,b_i)||_1 \le \varepsilon||\beta(b_i)||_2 \le \varepsilon$. This gives the remaining terms in Equations (24) and (25) (with the $|\mathcal{O}|$ factor for sum over x_t). ■

We next give a solution for the recursion from Proposition 29. This is standard, but we give a proof for completeness.

Proposition 30 *Consider the following recursions:*

$$\begin{aligned}
f(0)&=0; g(0)=0\\
f(t)&=d\varepsilon+(1+d\varepsilon)f(t-1)+d\varepsilon g(t-1)\\
g(t)&=d\varepsilon+d\varepsilon f(t-1)+d\varepsilon g(t-1)
\end{aligned}$$

Let $T\in\mathbb{Z}^+$. Then, the following holds for all $t\le T$ and $\varepsilon\le 1/12dT$:

$$f(t)\le 3dT\varepsilon$$

Proof We first claim: $g(t)\le 6d\varepsilon$ for all $t\le T$. We prove this by strong induction. This is true for $t=1$. Let's assume $g(i)\le 6\varepsilon$ for all $i\le t-1$. Then, first we unroll the recursion for $f(t)$.

$$\begin{aligned}
f(t)&=d\varepsilon+(1+d\varepsilon)f(t-1)+d\varepsilon g(t-1)\\
&=d\varepsilon+(1+d\varepsilon)\left[d\varepsilon+(1+d\varepsilon)f(t-2)+d\varepsilon g(t-2)\right]+d\varepsilon g(t-1)\\
&=d\varepsilon+(1+d\varepsilon)d\varepsilon+(1+d\varepsilon)^2f(t-2)+d\varepsilon(1+d\varepsilon)g(t-2)+d\varepsilon g(t-1)\\
&=(1+d\varepsilon)^t-1+d\varepsilon\left(\sum_{i=1}^{t-1}(1+d\varepsilon)^{i-1}g(t-i)\right)
\end{aligned}$$

where the last equation follows from observing the first few terms form a geometric series. Using our induction hypothesis, we get

$$\begin{aligned} f(t-1) \leq f(t) &\leq (1+2dT\varepsilon) - 1 + (d\varepsilon)(6d\varepsilon)\left(\sum_{i=1}^{t-1}(1+d\varepsilon)^{i-1}\right) \\ &\leq 2dT\varepsilon + 6d\varepsilon\left((1+d\varepsilon)^{t-1} - 1\right) \\ &\leq 2dT\varepsilon + 6d\varepsilon(2dT\varepsilon) \\ &\leq 3dT\varepsilon \end{aligned}$$

where we used that $(1+a)^t \leq 1+2at$ for $a < 1/2t$ and $\varepsilon < 1/12dT$. And therefore, we can bound $g(t)$ as

$$\begin{aligned} g(t) &= d\varepsilon + d\varepsilon f(t-1) + d\varepsilon g(t-1) \\ &\leq d\varepsilon + 3d^2T\varepsilon^2 + 6d^2\varepsilon^2 \\ &\leq 3d\varepsilon \end{aligned}$$

where we used $g(t-1) \leq 6d\varepsilon$ and $\varepsilon \leq 1/12dT$. This proves that $g(t) \leq 6d\varepsilon$ for all $t < T$. Moreover, by arguments above, in that case $f(t) \leq 3dT\varepsilon$ for all $t \leq T$. ■

Lemma 31 (Restatement of Lemma 14) *Let $\widehat{A}_{o,t}$ be the approximation of $A_{o,t}$ from Lemma 28. Furthermore, for sequence $x_{1:T}$ of length T, define*

$$\widehat{\Pr}[x_{1:T}] = \widehat{A}_{x_{1:T}}\beta(\varphi)\,.$$

Then, for the values of $\varepsilon < (12|\mathcal{O}|T)^{-1}$, we get that the distribution $\mathbf{p}$ and $\widehat{\mathbf{p}}$ given by probability functions $\Pr$ and $\widehat{\Pr}$ are close in TV distance:

$$TV(\mathbf{p}, \widehat{\mathbf{p}}) \leq 2|\mathcal{O}|T\varepsilon\,.$$

Proof Recall $F_T = \{\varphi\}$. Then,

$$\begin{aligned} 2\cdot TV(p,\widehat{p}) &= \sum_{x_{1:T}} \left|\widehat{\Pr}[x_{1:T}] - \Pr[x_{1:T}]\right| \\ &= \sum_{x_{1:T}} \left|\widehat{A}_{x_{1:T}} - A_{x_{1:T}}\right| \\ &= \sum_{x_{1:T}} \left|\beta(B_{x_{1:T}})\gamma_{x_{1:T}} + V_T^{\perp}\gamma_{x_{1:T}}^{\perp}\right| && \text{(by Proposition 29)} \\ &= \sum_{x_{1:T}} \left|\beta(B_{x_{1:T}})\gamma_{x_{1:T}}\right| && \text{(as } V_T^{\perp} = \ker(\Pr[F_T|B_T]) = [0]\text{(Proposition 22))} \\ &= \sum_{x_{1:T}} \left|\mathbf{1}^\top\gamma_{x_{1:T}}\right| && \text{(as } \beta(x) = 1 \text{ for all } x \in H_T\text{)} \\ &\leq \sum_{x_{1:T}} \|\gamma_{x_{1:T}}\|_1 && (29) \end{aligned}$$

The claim follows from Proposition 29 and Proposition 30. ■

E.4. Finding robust basis

We now show how to find a robust basis. We first recall some definitions: the covariance matrix for a basis is defined as

$$\Sigma_{B_t} = \Pr[F_t|B_t]^\top D_t^{-1} \Pr[F_t|B_t]$$

where D is a diagonal matrix of size $|F_t| \times |F_t|$ with entries $d_t(f) := \mathbb{E}_{b \in B_t} \Pr[f|b]$ on the diagonal. We further define the inner covariance as

$$\bar{\Sigma}_{B_t} = D_t^{-1/2} \Pr[F_t|B_t] \Pr[F_t|B_t]^\top D_t^{-1/2}$$

Note that the matrices Σ_B and $\bar{\Sigma}_B$ share their non-zero eigenvalues (with extra eigenvalues all 0). As we will see, random sampling from a high fidelity distribution gives a robust basis. We show in Appendix F more efficient ways of building a basis.

Lemma 15 (Finding robust basis) *Fix a distribution* $\mathbf{p}$ *over observation sequences of length* T*. Assume distribution* $\mathbf{p}$ *has rank* r *and fidelity* Δ^**. Pick* $0 < \delta < 1$*. Let* $n = O(\log r \Delta^{*-8})$ *and* $\Delta = \Omega(\log r (\Delta^*)^{-11/2})$*. Then, we can find sets* $\{S_t\}_{t \in [T]}$*, each of size* n*, using* $n \log(T/\delta)$ *conditional samples such that with probability* $1 - \delta$*,* $\{S_t\}_{t\in[T]}$ *is a* Δ*-robust basis for distribution* $\mathbf{p}$*.*

Proof Let S_t be a random sample of size n of observation sequences of length t from distribution $\mathbf{p}$. We ignore the t dependence in notation in this proof for clarity.

Let B^* be the unknown basis of length t sequences under which distribution $\mathbf{p}$ has high fidelity and $|B^*| = n^* \leq 1/\Delta^*$. Define a distribution d^* over futures given by $d^*(f) = \mathbb{E}_{b\in B^*}[\Pr[f|b]]$. Define $\operatorname{diag}(d^*)$ to be a diagonal matrix with diagonal entries given by $d^*(f)$.

Before, we prove that S is a robust basis at length t, we first show some properties of d^* that will come in handy. First, the norm of $\Pr[F|x]$ under d is upper bounded:

$$||\operatorname{diag}(d^*)^{-1/2} \Pr[F|x]||_2^2 = \mathbb{E}\left[\frac{\Pr[f|x]}{d^*(f)}\right]^2 = \mathbb{E}\left[\frac{\sum_i \alpha_i(x) \Pr[f|b_i^*]}{d^*(f)}\right]^2 \leq \mathbb{E}\left[||\alpha(x)||_1 n^*\right]^2 \leq n^{*3} \tag{30}$$

since $\Pr[f|b_i^*]/d^*(f) \leq n^*$ by definition and $||\alpha(x)||_1 \leq \sqrt{n^*}||\alpha(x)||_2 \leq \sqrt{n^*}$ as B^* is an basis.

Our next step is to show that the eigenvalues under S and B^* are not so different. For clarity, let

$$H = \operatorname{diag}(d^*)^{-1/2}\left(\mathbb{E}\left[\Pr[F|x]\Pr[F|x]^\top\right] - \mathbb{E}_{s\in S}\left[\Pr[F|s]\Pr[F|s]^\top\right]\right)\operatorname{diag}(d^*)^{-1/2}$$

Then, by matrix Bernstein inequality and Equation (30), for $|S| = n = O(n^{*6}\log 2r(\Delta^*)^{-2})$ and $\Delta^* < 1$, we get the following bound on H:

$$\mathbb{E}||H||_2 \leq \sqrt{\frac{2n^{*6}\log r}{n}} + \frac{2n^{*3}\log 2r}{3n} \leq \frac{\Delta^*}{4}$$

which by Markov's inequality shows $||H||_2 \leq \Delta^*/2$ with probability $1/2$. And by Weyl's inequality, shows that

$$\frac{1}{n}\sigma_{r_t}\left(\operatorname{diag}(d^*)^{-1/2}\left(\mathbb{E}_{s\in S}\left[\Pr[F|s]\Pr[F|s]^\top\right]\right)\operatorname{diag}(d^*)^{-1/2}\right) > \frac{\Delta^*}{2} \tag{31}$$

and all other eigenvalues are 0. We can repeat this sampling $\log(T/\delta)$ times to find the random set with these properties with probability $1-\delta$ for all $t \in [T]$.

We now show that S forms an basis. For clarity, let $\Pr[F|S]$ be a matrix with $\Pr[F|s_i]$ as columns. Then, $\Pr[F|S]^\top \mathrm{diag}(d^*)^{-1} \Pr[F|S]$ has r_t eigenvalues $> n\Delta^*/2$. Let V_t be the eigenvectors corresponding to non-zero eigenvalues of the latter matrix. Next we note that $\mathrm{span}(\{\Pr[F|s_1], \ldots, \Pr[F|s_n]\})$ has dimension r and therefore there exists coefficients $\beta(x) \in \mathrm{span}(V_t)$ such that

$$\Pr[F|x] = \sum_i \beta_i(x) \Pr[F|s_i] = \Pr[F|S]\beta(x)$$

Multiplying both sides by $\mathrm{diag}(d^*)^{-1/2}$, we get

$$n^{*3/2} \geq ||\mathrm{diag}(d^*)^{-1/2} \Pr[F|x]||_2 = ||\mathrm{diag}(d^*)^{-1/2} \Pr[F|S]\beta(x)||_2 \geq \sqrt{\frac{n\Delta^*}{2}}||\beta(x)||_2$$

where the first inequality follows from Equation (30) and the last inequality follows from Equation (31). Simplifying this shows S forms an basis with the following upper bound on the coefficients

$$||\beta(x)||_2 \leq \sqrt{\frac{2n^{*3}}{n\Delta^*}} < 1 \tag{32}$$

The only remaining part is to show that the two distributions d and d^* are only a small factor apart. For this, we first note from Equation (30)

$$\frac{d(f)}{d^*(f)} = \mathbb{E}_{s\in S}\left[\frac{\Pr[F|s_i]}{d^*(f)}\right] \leq n^{*3/2}$$

Since,

$$\begin{aligned} &\mathrm{diag}(d^*/d)^{1/2}\left(\mathrm{diag}(d^*)^{-1/2}\Pr[F|S]\Pr[F|S]^\top\mathrm{diag}(d^*)^{-1/2}\right)\mathrm{diag}(d^*/d)^{1/2} \\ &= \mathrm{diag}(d)^{-1/2}\Pr[F|S]\Pr[F|S]^\top\mathrm{diag}(d)^{-1/2} = \Sigma_S\,, \end{aligned}$$

it follows from discussion above and Equation (31) that Σ_S has r_t eigenvalues greater than

$$\frac{n\Delta^*}{2n^{*3/2}}$$

Substituting $n^* \leq 1/\Delta^*$ proves the claim. ■

E.5. Main result

Theorem 2 (Learning with conditional samples) *Let* $\Pr[\cdot]$ *be any rank* r *distribution over observation sequences of length* T*. Assume distribution* $\Pr[\cdot]$ *has fidelity* Δ^**. Pick any* $0 < \varepsilon, \delta < 1$*. Then Algorithm 2 with access to a conditional sampling oracle runs in* $\mathrm{poly}(r, T, O, 1/\Delta^*, 1/\varepsilon, \log(1/\delta))$ *time and returns an efficiently represented approximation* $\widehat{\Pr}[\cdot]$ *satisfying* $\mathrm{TV}(\Pr, \widehat{\Pr}) \leq \varepsilon$ *with probability at least* $1-\delta$*.*

Proof From Lemma 15, wp $1/2$, we built a Δ-robust basis of size n where

$$\Delta = \Omega\left(\frac{\log r}{\Delta^{*11/2}}\right)$$
$$n = O\left(\frac{\log r}{\Delta^{*8}}\right)$$
$$c = O\left(\sqrt{\frac{\Delta^{*4}}{\log r}}\right)$$

using n samples from p. From Lemma 31, to get TV error ε w.p. $1/2$, when given access to a Δ-robust basis of size n, we need to learn the estimate operators $\widehat{A}_{o,t}$ using Lemma 28 to accuracy

$$\varepsilon' = \frac{\varepsilon}{2|\mathcal{O}|T}$$

Substituting this in Lemma 28 gives the required sample complexity as

$$\widetilde{O}\left(\frac{c^8 r^3 n^{21} |\mathcal{O}|^3 T^5}{\Delta^6 \varepsilon'^6} \log^2\left(\frac{1}{\delta}\right)\right) = \widetilde{O}\left(\frac{c^8 r^3 n^{21} |\mathcal{O}|^9 T^{11}}{\Delta^6 \varepsilon^6} \log^2\left(\frac{1}{\delta}\right)\right)$$

Substituting the values of Δ, n and c above, we need

$$m = \widetilde{O}\left(\frac{r^3 |\mathcal{O}|^9 T^{11}}{\varepsilon^6 \Delta^{*119}} \log^2\left(\frac{1}{\delta}\right)\right)$$

many queries to conditional sampling oracle. ■

E.6. Estimating covariance matrix in Frobenius norm

In this section, we would show how to estimate the following objects:

$$q(x) = \Pr[F_t|B_t]^\top D_t^{-1} \Pr[F|x] \quad \text{and} \quad \Sigma_{B_t} = \Pr[F_t|B_t]^\top D_t^{-1} \Pr[F_t|B_t]\,,$$

which we need for estimating the operator $A_{o,t}$. We ignore t subscript when clear from context.

Proposition 32 *Let $B_t = \{b_1, \ldots, b_n\}$ be a basis of size n and c be some upper bound on the coefficients under basis B_t. Define $s(b^*, x)$ as the following sum where $b^* \in B_t$ and x is a history of length t:*

$$s(b^*, x) = \sum_{f \in F_t} \frac{\Pr[f|b^*] \Pr[f|x]}{d(f)}\,.$$

Then we can learn estimate $\widehat{s}(b^*, x)$ such that with probability $1 - \delta$,

$$|s(b^*, x) - \widehat{s}(b^*, x)| \leq \varepsilon$$

using at most

$$m_\varepsilon = O\left(\frac{c^2 n^{10} |\mathcal{O}|^2 T^4}{\varepsilon^6} \log^2\left(\frac{1}{\delta}\right)\right)$$

conditional samples.

Proof We start by writing $s(b^*, x)$ in terms of expectation under $\Pr[\cdot|x]$:

$$s(b^*, x) = \sum_{f \in F_t} \frac{\Pr[f|b^*]\Pr[f|x]}{d(f)} = \mathbb{E}_{f\sim\Pr[\cdot|x]}\left[\frac{\Pr[f|b^*]}{d(f)}\right]$$

With this, we define (un-normalized) probability functions $\widehat{\Pr}[\cdot|b]$ for $b \in B_t$ which set probability to 0 if f is part of a set F_b (to be defined) which will depend only on history b i.e.

$$\widehat{\Pr}[f|b] = \begin{cases} 0 & f \in F_b \\ \Pr[f|b] & \text{otherwise} \end{cases}$$

We define $\widehat{d}$ as mixture distribution of $\widehat{\Pr}[\cdot|b]$ for $b \in B_t$ i.e. $\widehat{d}(f) = \mathbb{E}_{b\in B_t}[\widehat{\Pr}[f|b]]$. An important aspect of our definitions is that for any future f,

$$0 \le \frac{\widehat{\Pr}[f|b^*]}{\widehat{d}(f)} \le n \quad \text{and} \quad 0 \le \frac{\Pr[f|x]}{d(f)} \le cn^{3/2} \tag{33}$$

are upper bounded. Now, suppose $\max_b \Pr[F_b|b] \le p$. Then,

$$\begin{aligned}
&\left|\mathbb{E}_{f\sim\Pr[\cdot|x]}\left[\frac{\Pr[f|b^*]}{d(f)}\right] - \mathbb{E}_{f\sim\Pr[\cdot|x]}\left[\frac{\widehat{\Pr}[f|b^*]}{\widehat{d}(f)}\right]\right| \\
&\le \sum_f \left|\frac{\Pr[f|b^*]\Pr[f|x]}{d(f)} - \frac{\widehat{\Pr}[f|b^*]\Pr[f|x]}{\widehat{d}(f)}\right| \\
&= \sum_f \left|\frac{\Pr[f|b^*]\Pr[f|x]}{d(f)} \pm \frac{\widehat{\Pr}[f|b^*]\Pr[f|x]}{d(f)} - \frac{\widehat{\Pr}[f|b^*]\Pr[f|x]}{\widehat{d}(f)}\right| \\
&\le \sum_f \left|\left(\Pr[f|b^*] - \widehat{\Pr}[f|b^*]\right)\frac{\Pr[f|x]}{d(f)}\right| + \sum_f \left|\frac{\widehat{\Pr}[f|b^*]\Pr[f|x]}{\widehat{d}(f)d(f)}\left(\widehat{d}(f) - d(f)\right)\right| \\
&\le cn^{3/2}\Pr[F_{b^*}|b^*] + cn^{5/2}\frac{\sum_{k=1}^n \Pr[F_{b_k}|b_k]}{n} \\
&\le O(cn^{5/2}p)
\end{aligned}$$

where the last step follows from $\max_b \Pr[F_b|b] \le p$. Now, we can sample $m = (1/2c^2n^3p^2)\log(2/\delta)$ random futures from $\Pr[\cdot|x]$ and call this set S. Then, again by Equation (33) and Hoeffding's inequality (Proposition 45), we get that

$$\Pr\left[\left|\mathbb{E}_{f\sim\Pr[\cdot|x]}\left[\frac{\widehat{\Pr}[f|b^*]}{\widehat{d}(f)}\right] - \mathbb{E}_{f\sim S}\left[\frac{\widehat{\Pr}[f|b^*]}{\widehat{d}(f)}\right]\right| \ge cn^{5/2}p\right] \le \delta$$

Together, with $(1/2c^2n^3p^2)\log(2/\delta)$ conditional samples, we get the following guarantee with probability $1-\delta$,

$$\left|\mathbb{E}_{f\sim\Pr[\cdot|x]}\left[\frac{\Pr[f|b^*]}{d(f)}\right] - \mathbb{E}_{f\sim S}\left[\frac{\widehat{\Pr}[f|b^*]}{\widehat{d}(f)}\right]\right| \le O(cn^{5/2}p)$$

To estimate $\widehat{\Pr}[f|b^*]$ and $\widehat{d}(f)$, we need to identify the case when f is "irregular".

Definition 33 (α-regular future) *We define a future f to be α-regular for history b if for all $\tau \in [t]$*

$$\Pr[f_\tau | b f_{1:\tau-1}] > \alpha\,.$$

Otherwise, f is α-irregular for history b.

To do this, define empirical estimates $\widetilde{\Pr}[f_\tau | b f_{1:\tau-1}]$ for every future f and basis $b \in B_t$ using $\widetilde{O}(n|S|T/\alpha^2 \log(1/\delta))$ many samples. Then, we perform the following test $A(f, b)$ for each future f and basis history b using these estimates:

Definition 34 (Test $A(f, b)$) *Test $A(f, b)$ passes if the empirical estimate $\widetilde{\Pr}[f_\tau | b f_{1:\tau-1}] > 2\alpha$ for all $\tau \in [t]$ and fails otherwise.*

Note that with probability $1 - \delta$, (i) if test $A(f, b)$ passes for future f and history b, then f is α-regular for b, and (ii) if test $A(f, b)$ fails for future f and history b, then f is 3α-irregular for b. In the rest of the proof, we condition on the event that this relationship between test $A(f, b)$ and irregular futures holds for all futures $f \in S$ and $b \in B_t$. We set F_b to be the set of futures f which are 3α-regular for basis b and removing the ones where test $A(f, b)$ passes. By Proposition 37,

$$p = \Pr[F_b | b] \le O(|\mathcal{O}| T \alpha) \tag{34}$$

Now, we define estimates $\widetilde{\Pr}[f|b]$ for each future $f \in S$ and basis history $b \in B_t$ by first running test $A(f, b)$ on future f and history b. If test $A(f, b)$ fails we set $\widetilde{\Pr}[f|b] = 0$. Note that otherwise $\widetilde{\Pr}[f|b]$ to be the estimate from Proposition 36 i.e. with probability $1 - \delta$,

$$\left| \widetilde{\Pr}[f|b] - \Pr[f|b] \right| \le \gamma \Pr[f|b]$$

This requires $\widetilde{O}(n|S|T^2/(\gamma\alpha)^2 \log(1/\delta))$ many samples. Moreover, because $\widehat{\Pr}[f|b^*] = \Pr[f|b^*]$ for futures where tests passes, we can estimate the probability ratios with additive error:

$$\frac{\widetilde{\Pr}[f|b^*]}{\widetilde{d}(f)} \le \frac{(1+\gamma)\widehat{\Pr}[f|b^*]}{(1-\gamma)\widehat{d}(f)} \le \frac{(1+4\gamma)\widehat{\Pr}[f|b^*]}{\widehat{d}(f)} \le \frac{\widehat{\Pr}[f|b^*]}{\widehat{d}(f)} + 4\gamma n$$

where the second inequality holds for $\gamma < 1/2$ and

$$\frac{\widetilde{\Pr}[f|b^*]}{\widetilde{d}(f)} \ge \frac{(1-\gamma)\Pr[f|b^*]}{(1+\gamma)d(f)} \ge \frac{(1-2\gamma)\Pr[f|b^*]}{d(f)} \ge \frac{\Pr[f|b^*]}{d(f)} - 2\gamma n$$

where the second inequality holds for $\gamma < 1/2$. This means

$$\left| \mathbb{E}_{f\sim S}\left[\frac{\widehat{\Pr}[f|b^*]}{\widehat{d}(f)}\right] - \mathbb{E}_{f\sim S}\left[\frac{\widetilde{\Pr}[f|b^*]}{\widetilde{d}(f)}\right] \right| \le 4\gamma n \tag{35}$$

Combining Equations (33) to (35), we can build estimate $[\widehat{q}(x)]_i$ for $[q(x)]_i$ such that with probability $1 - O(\delta)$

$$|[\widehat{q}(x)]_i - [q(x)]_i| \le \varepsilon$$

using

$$\widetilde{O}(c^2 n^{10}|\mathcal{O}|^2 T^4 \frac{1}{\varepsilon^6} \log^2(\frac{1}{\delta}))$$

many conditional samples.

■

Corollary 35 *We can learn approximations $\widehat{q}(bo)$ and $\widehat{\Sigma}_{B_t}$ for all $b \in B_t$, observations o and time $t \in [T]$ such that with probability $1-\delta$,*

$$||\widehat{q}(bo) - q(bo)||_F \le \varepsilon\sqrt{n}\,, ||\widehat{\Sigma}_{B_t} - \Sigma_{B_t}||_F \le \varepsilon n$$

using at most $\widetilde{O}(n^2|\mathcal{O}|Tm_\varepsilon)$ conditional samples.

Proof Each entry of Σ_{B_t} and $q(bo)$ is given by sums of the form $s(b*, x)$ where b^* is a basis history and x is arbitrary history of length t. Therefore, we can estimate each of them using estimates given by Proposition 32:

$$||\widehat{\Sigma}_{B_t} - \Sigma_{B_t}||_F^2 \le \sum_{i,j\in[n]} |s(b_i, b_j) - \widehat{s}(b_i, b_j)|^2 \le n^2\varepsilon^2$$

The result similarly holds for $q(bo)$. There are $O(n^2)$ entries in each matrix with T many Σ_{B_t} matrices and $O(n)$ entries in each vector with $n|\mathcal{O}|T$ many $q(bo)$ vectors. ■

Proposition 36 *Consider a future $f_{1:t}$ of length t and history x. Fix $\gamma > 0$. Suppose $\Pr[f_\tau|xf_{1:\tau-1}] > \alpha$ for each $\tau \in [t]$. Then, we can build estimate $\widehat{\Pr}[f_{1:t}|x]$ such that with probability $1-\delta$*

$$\left|\widehat{\Pr}[f_{1:t}|x] - \Pr[f_{1:t}|x]\right| \le \gamma \Pr[f_{1:t}|x]$$

using at most $O(t^2/(\gamma\alpha)^2 \log(t/\delta))$ conditional samples.

Proof By Hoeffding's inequality, using $m = 16t^2\log(t/\delta)/(\gamma\alpha)^2$ samples, we have, with probability greater than $1-\delta$, that for all $\tau \in [t]$,

$$\left|\Pr[f_\tau|xf_{1:\tau-1}] - \widehat{\Pr}[f_\tau|xf_{1:\tau-1})\right| \le \frac{\gamma\alpha}{2t} \le \frac{\gamma}{2t}\Pr[f_\tau|xf_{1:\tau-1}),$$

where the last step uses our assumption above. For an upper bound, we have,

$$\begin{aligned}\widehat{\Pr}[f_{1:t}|x] = \Pi_{\tau=1}^t \widehat{\Pr}[f_\tau|xf_{1:\tau-1}] &\le (1+\frac{\gamma}{2t})^t \Pi_{\tau=1}^t \Pr[f_\tau|xf_{1:\tau-1}] \\ &= (\frac{\gamma}{2t})^t \Pr[f_{1:t}|x] \le \Pr[f_{1:t}|x) + \gamma \Pr[f_{1:t}|x]\end{aligned}$$

where the last step follows with $(1+a)^t \le 1 + 2at$ for $a < 1/2t$. Similarly, for a lower bound we have:

$$\widehat{\Pr}[f_{1:t}|x] \ge (1-\frac{\gamma}{2t})^t \Pr[f_{1:t}|x) \ge \widehat{\Pr}[f_{1:t}|x] - \gamma\Pr[f_{1:t}|x],$$

where the last step follows with $(1+a)^t \ge 1 + at$ for $a \ge -1$. ■

Proposition 37 *Define a future f to be α-irregular for history $b \in B_t$ if there exists some $\tau \in [t]$ (τ can depend on b) such that*

$$\Pr[f_\tau | b f_{1:\tau-1}] < \alpha\,.$$

Let F_b be the set of futures f where f is α-irregular for history b. Then,

$$\Pr[F_b | b] \leq |\mathcal{O}| T \alpha$$

Proof Let future f be of length T. We first partition the set F_b into T sets: $F_{b,1}, \ldots, F_{b,T}$ based on the first time irregular is observed i.e. $f \in F_{b,t} \iff t = \min_\tau \Pr[f_\tau | b f_{1:\tau-1}] < \alpha$. Now,

$$\begin{aligned}
&\sum_{f_{1:T} \in F_{b,t}} \Pr[f|b] \\
&= \sum_{f_{1:t-1} f_t f_{t+1:T} \in F_{b,t}} \Pr[f_{t+1:T} | b f_{1:t-1} f_t] \Pr[f_t | b f_{1:t-1}] \Pr[f_{1:t-1} | b] \\
&\leq \sum_{f_{1:t-1} f_t \in F_{b,t}} \Pr[f_t | b f_{1:t-1}] \Pr[f_{1:t-1} | b] \sum_{\text{futures } g \text{ of length } T-t-1} \Pr[g | b f_{1:t-1} f_t] \\
&\leq |\mathcal{O}| \alpha \sum_{f_{1:t-1} \in F_{b,t}} \Pr[f_{1:t-1} | b] \qquad \text{(as } \Pr[f_t | b f_{1:t-1}] \leq \alpha) \\
&\leq |\mathcal{O}| \alpha
\end{aligned}$$

Summing over all T of these sets gives the claim. ■

Appendix F. General algorithm for finding approximate basis

In this section, we learn an approximate version of basis for probability vectors. Throughout this section, we ignore t subscript in F_t and B_t when clear from context. We define an approximate basis, which allows us to ignore histories which have very low probability under the distribution $\mathbf{p}$:

Definition 38 (Approximate Basis) *Fix $0 < \varepsilon < 1$. For a distribution $\mathbf{p}$ over observation sequences of length T, we say a subset of observations sequences B forms an ε-basis for $\mathbf{p}$ at length $t \in [T]$, if for every observation sequence $x = (x_1, \ldots, x_t)$, there exists coefficients $\beta(x)$ with ℓ_2 norm $||\beta(x)||_2 \leq 1$ such that:*

$$\mathbb{E}_{x \sim \mathbf{p}} \left[|| \Pr[F|x] - \Pr[F|B]\beta(x) ||_1 \right] \leq \varepsilon\,.$$

We first define regular distributions.

Definition 39 (Regular distribution) *We say a distribution $\mathbf{p}$ is α-regular if $\min \Pr_{\mathbf{p}}[o|x] \geq \alpha$ where the minimum is over all histories x and observations o where $\Pr[o|x] \neq 0$.*

We now present the main result in this section: how to build an approximate basis for a regular low rank distribution.

Theorem 3 *Let $\mathbf{p}$ be an α-regular distribution over observation sequences of length T with rank r. Fix $0 < \varepsilon < \alpha/Tr$ and $0 < \delta < 1$. Then, in $\mathrm{poly}(r, T, 1/\varepsilon, 1/\alpha, \log(1/\delta))$ time, with probability $1 - \delta$, we can find an ε-basis of size at most $O(r^2 T^3 \log(1/\alpha\varepsilon))$ using conditional sampling oracle.*

We believe the regularity assumption on the distribution can be removed using the ideas from Appendix E.6 but leave it as future work.

F.1. Learning coefficients

Towards this goal, we first show how to check given an observation sequence x and set of observation sequences B if there exists $\beta(x)$ such that

$$|| \Pr[F|x] - \Pr[F|B]\beta(x)||_1 \leq \varepsilon$$

Instead of directly working with the ℓ_1 loss, we first define an ℓ_2 loss which we can use as its proxy.

Definition 40 *For set of observation sequences $B = \{b_1, \dots, b_h\}$, observation sequence x, column vector $\beta \in \mathbb{R}^{|B|}$, we define the ℓ_2 approximation error as:*

$$L_{B,x}(\beta) := \mathbb{E}_{f \sim d}\left[\left(\frac{\Pr[f|x]}{d(f)} - \sum_{j=1}^{h} \beta_j \frac{\Pr[f|b_j]}{d(f)}\right)^2\right],,$$

where d is the mixture distribution for B:

$$d(f) = \frac{1}{2}\Pr[f|x] + \frac{1}{2h}\sum_{i=1}^{h} \Pr[f|b_i],$$

When clear from context, we drop the B, x superscript.

We will use our ability to simulate relative probabilities for regular distributions (Proposition 36) to build our guess for approximate basis.

Proposition 41 *Let $\mathbf{p}$ be an α-regular distribution, x be an observation sequence of length $t \in [T]$ and B be any set of observation sequences of length t. Suppose $f_1, \dots f_m$ are i.i.d. samples from d. Then, using $\mathrm{poly}(T, 1/\varepsilon, 1/\alpha, \log(1/\delta))$ many conditional samples, we can have estimates $\widehat{\Pr}[f_i|b]$ for all $i \in [m]$ and $b \in B \cup \{x\}$ such that with probability $1 - \delta$,*

$$\sup_{\|\beta\|_2 \leq C, \widehat{L}_{B,x}(\beta) \leq \widehat{L}_{B,x}(0)} \left|L_{B,x}(\beta) - \widehat{L}_{B,x}(\beta)\right| \leq \varepsilon$$

where the estimated ℓ_2 error function $\widehat{L}_{B,x}$ is defined as

$$\widehat{L}_{B,x}(\beta) := \frac{1}{m}\sum_{i \in [m]} \left(\frac{\widehat{\Pr}[f_i|x]}{\widehat{d}(f_i)} - \sum_{j=1}^{h} \beta_j \frac{\widehat{\Pr}[f_i|b_j]}{\widehat{d}(f_i)}\right)^2.$$

and $\widehat{d}(f_i)$ is the mixture distribution defined with the estimated probabilities.

Proof For notational convenience, we will drop the B, x superscript and simply write $L(\cdot)$ in the proof. Using $\mathrm{poly}(1/\gamma, 1/\alpha, T, \log(1/\delta))$ conditional samples (Proposition 36), with probability $1 - \delta$, we can have estimates $\widehat{\Pr}[f_i|b]$ such that, for all $i \in [m]$ and $b \in B \cup \{x\}$,

$$|\Pr[f_i|b] - \widehat{\Pr}[f_i|b]| \leq \gamma \Pr[f_i|b].$$

Define:

$$\overline{L}(\beta) := \frac{1}{m}\sum_{i\in[m]}\left(\frac{\Pr[f_i|x]}{d(f_i)} - \sum_{j=1}^{h}\beta_j\frac{\Pr[f_i|b_j]}{d(f_i)}\right)^2.$$

We have that:

$$\big|L(\beta) - \widehat{L}(\beta)\big| \le \big|L(\beta) - \overline{L}(\beta)\big| + \big|\overline{L}(\beta) - \widehat{L}(\beta)\big|.$$

We will handle the two terms above separately.

To bound $|L(\beta) - \overline{L}(\beta)|$, let us first show that for any observation sequence f:

$$\sum_{j=1}^{h}\left(\frac{\Pr[f|b_j]}{d(f)}\right)^2 \le 4h^2$$

To see this, observe that:

$$\max_{i\in[m],j\in[h]}\left|\frac{\Pr[f_i|b_j]}{d(f_i)}\right| \le 2h,$$

which implies:

$$\sum_{j=1}^{h}\left(\frac{\Pr[f|b_j]}{d(f)}\right)^2 \le 2h\sum_{j=1}^{h}\frac{\Pr[f|b_j]}{d(f)} = 4h^2.$$

Now using that the square loss is a 2-smooth function and a standard uniform convergence argument, we have that

$$\sup_{\|\beta\|_2\le C}\big|L(\beta) - \widehat{L}(\beta)\big| \le 16(Ch+1)\sqrt{\frac{\log(1/\delta)}{m}},$$

holds with probability greater than $1-\delta$.

For the second term, define

$$\Delta(f_i) = \frac{\Pr[f_i|x]}{d(f_i)} - \sum_{j=1}^{h}\beta_j\frac{\Pr[f_i|b_j]}{d(f_i)} - \left(\frac{\widehat{\Pr}[f_i|x]}{\widehat{d}(f_i)} - \sum_{j=1}^{h}\beta_j\frac{\widehat{\Pr}[f_i|b_j]}{\widehat{d}(f_i)}\right).$$

Let us first show that, for all $i \in [m]$,

$$|\Delta(f_i)| \le 4\gamma h(1 + C\sqrt{h}).$$

We have that:

$$\begin{aligned}|\Delta(f_i)| &\le \left|\frac{\Pr[f_i|x]}{d(f_i)} - \frac{\widehat{\Pr}[f_i|x]}{\widehat{d}(f_i)}\right| + \sum_{j\in[h]}\beta_j\left|\frac{\Pr[f_i|b_j]}{d(f_i)} - \frac{\widehat{\Pr}[f_i|b_j]}{\widehat{d}(f_i)}\right| \\ &\le (1+\|\beta\|_1)\max_{b\in B\cup\{x\}}\left|\frac{\Pr[f_i|b]}{d(f_i)} - \frac{\widehat{\Pr}[f_i|b]}{\widehat{d}(f_i)}\right| \\ &\le (1+C\sqrt{h})\max_{b\in B\cup\{x\}}\left|\frac{\Pr[f_i|b]}{d(f_i)} - \frac{\widehat{\Pr}[f_i|b]}{\widehat{d}(f_i)}\right|.\end{aligned}$$

The claim would follow provided we have that, for all $i \in [m]$ and all $b \in B \cup \{x\}$,

$$\left| \frac{\Pr[f_i|b]}{d(f_i)} - \frac{\widehat{\Pr}[f_i|b]}{\widehat{d}(f_i)} \right| \leq 4\gamma h, \tag{36}$$

To see this, using that $\gamma \leq 1/2$, we have the following upper bound that:

$$\frac{\widehat{\Pr}[f_i|b]}{\widehat{d}(f_i)} \leq \frac{(1+\gamma)\Pr[f_i|b]}{(1-\gamma)d(f_i)} \leq \frac{(1+4\gamma)\Pr[f_i|b]}{d(f_i)} \leq \frac{\Pr[f_i|b]}{d(f_i)} + 4\gamma h$$

and the lower bound:

$$\frac{\widehat{\Pr}[f_i|b]}{\widehat{d}(f_i)} \geq \frac{(1-\gamma)\Pr[f_i|b]}{(1+\gamma)d(f_i)} \geq \frac{(1-2\gamma)\Pr[f_i|b]}{d(f_i)} \geq \frac{\Pr[f_i|b]}{d(f_i)} - 2\gamma h.$$

This completes the proof of Equation (36).

Now consider any β such that $\widehat{L}(\beta) \leq \widehat{L}(0)$. Also, it is straightforward that $\widehat{L}(0) \leq 4$. This implies that:

$$\begin{aligned}
\overline{L}(\beta) - \widehat{L}(\beta) &= \frac{2}{m}\sum_{i\in[m]} \left(\frac{\widehat{\Pr}[f_i|x]}{\widehat{d}(f_i)} - \sum_{j=1}^{h} \beta_j \frac{\widehat{\Pr}[f_i|b_j]}{\widehat{d}(f_i)} \right) \Delta(f_i) + \frac{1}{m}\sum_{i\in[m]} \Delta(f_i)^2 \\
&\leq 2\sqrt{\widehat{L}(\beta)\cdot \frac{1}{m}\sum_{i\in[m]} \Delta(f_i)^2} + \frac{1}{m}\sum_{i\in[m]} \Delta(f_i)^2 \\
&\leq 4\sqrt{\frac{1}{m}\sum_{i\in[m]} \Delta(f_i)^2} + \frac{1}{m}\sum_{i\in[m]} \Delta(f_i)^2 \\
&\leq 4\cdot 4\gamma h(1+C\sqrt{h}) + 16\gamma^2 h^2 (1+C\sqrt{h})^2 \\
&\leq 16\gamma h(1+\gamma h)(1+C\sqrt{h})^2.
\end{aligned}$$

where the first step follows from $a^2 - b^2 = 2b(a-b) + (a-b)^2$ and the second step from Cauchy–Schwarz inequality. Combining the bounds for the first and second terms completes the proof. ■

F.2. Algorithm

We now ready to present our algorithm. The user furnishes ε, the accuracy with which approximate basis is to be learned; and δ, a confidence parameter. The parameter n and H depend on the input.

By Hoeffding's inequality, it is clear that if this algorithm ends then we have found an approximate basis. We know show that with high probability, this algorithm ends in small number of rounds.

Proposition 42 *Let* $\mathbf{p}$ *be an* α*-regular distribution over observation sequences of length* T *with rank* r*. Fix* $0 \leq \varepsilon \leq \alpha^2/T^2r^2$*. Let* $C = \sqrt{2Tr\log(1/\alpha\varepsilon)}$ *and let* H *be any natural number provided* $H \geq 8rT^2\log(1/\varepsilon\alpha)$*. Consider any sequence of observation sequences* $b_1, b_2, \ldots, b_H$*. Let* $B_h = \{b_1, \ldots, b_h\}$*. Then, there exists* $h \leq H$ *such that:*

$$\min_{\beta\in\mathbb{R}^h, ||\beta||_2 \leq C} L_{B_h, b_{h+1}}(\beta) \leq \varepsilon$$

Algorithm 3: Learning approximate basis using conditional samples.

```
for round $i = 1, 2, \dots, H$ do
    Sample $n$ samples $x = (x_1, \dots, x_t)$ of length $t$ from distribution $\mathbf{p}$.
    Check using Proposition 41 if any of the samples $x$ above is a "counterexample" i.e.
      satisfies
        $$\min_{\beta \in \mathbb{R}^h, ||\beta||_2 \le C} \widehat{L}_{B,x}(\beta) > \frac{\varepsilon^2}{8}$$
    if we find such a counterexample $x = (x_1, \dots, x_t)$ then
        Add $x$ to $B$
    else
        return $B$
```

Proof For $\beta \in \mathbb{R}^h$, define:

$$L_\lambda^{(h)}(\beta) := L_{B_h, b_{h+1}}(\beta) + \lambda \sum_{j=1}^{h} \beta_j^2$$

and $d^{(h)}$ to be the mixture distribution corresponding to B_h. Define:

$$\overline{\Pr}[f|b_j] := \frac{\Pr[f|b_j]}{\sqrt{d^{(h)}(f)}}.$$

It will be helpful to overload notation and view $P_j = \Pr[F|b_j]$ as a vector of length $|F|$ and $P_{1:h} = \Pr[F|B_h]$ as a matrix of size $|F| \times h$, whose columns are $P_1, \dots P_h$. We overload notation analogously for the vector $\overline{P}_j$ and the matrix $\overline{P}_{1:h}$. Also, let $D^{(h)}$ be a diagonal matrix of size $|F| \times |F|$, whose diagonal entries are $d^{(h)}$, where we drop the h superscript when clear from context. With our notation, we have that $\overline{P}_{h+1} = D^{-1/2} P_{h+1}$ and $\overline{P}_{1:h} = D^{-1/2} P_{1:h}$. We will write P and $\overline{P}$ in lieu of $P_{1:h}$ and $\overline{P}_{1:h}$, when clear from context.

We have:

$$\begin{aligned}
L^{(h)}(\beta) &= \mathbb{E}_{x \sim d^{(h)}} \left[\left(\frac{\Pr[f|b_{h+1}]}{d^{(h)}(f)} - \sum_{j=1}^{h} \beta_j \frac{\Pr[f|b_j]}{d^{(h)}(f)} \right)^2 \right] \\
&= (D^{-1} P_{h+1} - D^{-1} P_{1:h} \beta)^\top D (D^{-1} P_{h+1} - D^{-1} P_{1:h} \beta) \\
&= \|\overline{P}_{h+1} - \overline{P} \beta\|^2,
\end{aligned}$$

where we have used our matrix notation. Let us consider the following ridge regression estimator:

$$\begin{aligned}
\beta_\lambda^{(h)} &= \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \left(L^{(h)}(\beta) + \lambda \|\beta\|^2 \right) \\
&= \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \left(\|\overline{P}_{h+1} - \overline{P} \beta\|^2 + \lambda \|\beta\|^2 \right) \\
&= (\overline{P}^\top \overline{P} + \lambda I)^{-1} \overline{P}^\top \overline{P}_{h+1}.
\end{aligned}$$

For $p_{\text{min}} = \alpha^T$, define Σ_h to be the $|F| \times |F|$ sized matrix, as follows

$$\Sigma_h := p_{\text{min}}\lambda I + P_{1:h}P_{1:h}^\top = p_{\text{min}}\lambda I + \sum_{j=1}^{h} P_j P_j^\top.$$

We will now show that:

$$\min_{\beta\in\mathbb{R}^h} L_\lambda^{(h)}(\beta) \le \lambda P_{h+1}^\top \Sigma_h^{-1} P_{h+1}. \tag{37}$$

Define:

$$\overline{\Sigma}_h := \lambda I + \overline{P}_{1:h}\overline{P}_{1:h}^\top.$$

One can verify that:

$$\overline{P}\beta_\lambda^{(h)} = \overline{P}(\overline{P}^\top\overline{P} + \lambda I)^{-1}\overline{P}^\top\overline{P}_{h+1} = \overline{PP}^\top(\overline{PP}^\top + \lambda I)^{-1}\overline{P}_{h+1} = \overline{PP}^\top\overline{\Sigma}_h^{-1}\overline{P}_{h+1},$$

and that:

$$\begin{aligned}\|\beta_\lambda^{(h)}\|^2 = \overline{P}_{h+1}^\top\overline{\Sigma}_h^{-1}\overline{PP}^\top\overline{\Sigma}_h^{-1}\overline{P}_{h+1} &= \overline{P}_{h+1}^\top\overline{\Sigma}_h^{-1}(\overline{\Sigma}_h - \lambda I)\overline{\Sigma}_h^{-1}\overline{P}_{h+1}\\ &= \overline{P}_{h+1}^\top\overline{\Sigma}_h^{-1}\overline{P}_{h+1} - \lambda\overline{P}_{h+1}^\top\overline{\Sigma}_h^{-2}\overline{P}_{h+1}.\end{aligned}$$

Using this, we have:

$$\begin{aligned}\min_{\beta\in\mathbb{R}^h} L_\lambda^{(h)}(\beta) &= \|\overline{P}_{h+1} - \overline{P}\beta_\lambda^{(h)}\|^2 + \lambda\|\beta_\lambda^{(h)}\|^2\\ &= \left\|\left(I - \overline{PP}^\top\overline{\Sigma}_h^{-1}\right)\overline{P}_{h+1}\right\|^2 + \lambda\|\beta_\lambda^{(h)}\|^2\\ &= \left\|\left(\overline{PP}^\top + \lambda I - \overline{PP}^\top\right)\overline{\Sigma}_h^{-1}\overline{P}_{h+1}\right\|^2 + \lambda\|\beta_\lambda^{(h)}\|^2\\ &= \lambda^2\overline{P}_{h+1}^\top\overline{\Sigma}_h^{-2}\overline{P}_{h+1} + \lambda\|\beta_\lambda^{(h)}\|^2\\ &= \lambda\overline{P}_{h+1}^\top\overline{\Sigma}_h^{-1}\overline{P}_{h+1}.\end{aligned}$$

By our assumption on p_{min}, we have that $d^{(h)}(f) \ge p_{\text{min}}$, which implies $D \succeq \frac{1}{p_{\text{min}}}I$. Using this, we have

$$\begin{aligned}\lambda\overline{P}_{h+1}^\top\overline{\Sigma}_h^{-1}\overline{P}_{h+1} &= \lambda(D^{-1/2}P_{h+1})^\top(\lambda I + D^{-1/2}PP^\top D^{-1/2})^{-1}D^{-1/2}P_{h+1}\\ &= \lambda P_{h+1}(\lambda D + PP^\top)^{-1}P_{h+1}\\ &\le \lambda P_{h+1}(p_{\text{min}}\lambda I + PP^\top)^{-1}P_{h+1}\\ &= \lambda P_{h+1}^\top\Sigma_h^{-1}P_{h+1},\end{aligned}$$

where we have used the definition of Σ_h in the last step. This proves the claim in Equation (37).

This implies:

$$\min_{h\le H}\min_{\beta\in\mathbb{R}^h} L_\lambda^{(h)}(\beta) \le \frac{1}{H}\sum_{h=1}^{H}\min_{\beta\in\mathbb{R}^h} L_\lambda^{(h)}(\beta) \le \frac{\lambda}{H}\sum_{h=1}^{H} P_{h+1}^\top\Sigma_h^{-1}P_{h+1}.$$

Using that $\|P_{h+1}\|^2 \leq 1$ (since P_{h+1} is a probability distribution), Lemma 43 implies:

$$\frac{1}{H}\sum_{h=1}^{H} \log(1 + P_{h+1}^\top \Sigma_h^{-1} P_{h+1}) \leq \frac{r}{H} \log\left(1 + H/(p_{\min}\lambda)\right),$$

which implies there exists an $h \leq H$ such that:

$$\log(1 + P_{h+1}^\top \Sigma_h^{-1} P_{h+1}) \leq \frac{r}{H} \log\left(1 + H/(p_{\min}\lambda)\right).$$

For $H \geq 4r\log(1 + H/(p_{\min}\lambda))$, exponentiating leads to:

$$P_{h+1}^\top \Sigma_h^{-1} P_{h+1} \leq \exp\left(\frac{r}{H}\log\left(1 + H/(p_{\min}\lambda)\right)\right) - 1 \leq 2\frac{r}{H}\log\left(1 + H/(p_{\min}\lambda)\right)$$

where the last step follows due to our choice of H. This shows that there exists an $h \leq H$ such that:

$$\min_{\beta \in \mathbb{R}^h} L_\lambda^{(h)}(\beta) \leq \frac{\lambda r}{H} \log\left(1 + H/(p_{\min}\lambda)\right).$$

Choosing $\lambda = \varepsilon^2$, setting $H = 8rT^2\log(1/\varepsilon\alpha)$ suffices to satisfy our assumptions. This implies we get the minimum loss is achieved at β whose norm is bounded by

$$||\beta||_2 \leq C := \sqrt{2Tr\log\left(\frac{1}{\alpha\varepsilon}\right)}$$

and therefore for $\varepsilon < \alpha^2/T^2r^2$, we get

$$\min_{\beta \in \mathbb{R}^h, ||\beta|| \leq C} L^{(h)}(\beta) \leq \varepsilon$$

■

The following is a variant of the Elliptical Potential Lemma, from the analysis of linear bandits (Dani et al., 2008).

Lemma 43 (Elliptical potential) *Consider a sequence of vectors $\{x_1, \ldots, x_T\}$ where, for all $i \in [T]$, each $x_i \in \mathcal{V}$, where $\mathcal{V}$ is a d-dimensional subspace of a Hilbert space, and $\|x_i\| \leq B$. Let $\lambda \in \mathbb{R}^+$. Denote $\Sigma_t = \Sigma_0 + \sum_{i=1}^t x_i x_i^\top$. We have that:*

$$\min_{i \in [T]} \ln\left(1 + x_i^\top \Sigma_i^{-1} x_i\right) \leq \frac{1}{T}\sum_{i=1}^{T} \ln\left(1 + x_i^\top \Sigma_i^{-1} x_i\right) = \frac{1}{T}\ln\frac{\det(\Sigma_T)}{\det(\lambda I)} \leq \frac{d}{T}\log\left(1 + \frac{TB^2}{d\lambda}\right)$$

We now finish the proof of Theorem 3 by adding the missing details. We first show that ℓ_1 loss is upper bounded by our ℓ_2 proxy loss.

Proposition 44 *Let x be an observation sequence of length $t \in [T]$ and $B = \{b_1, \ldots, b_h\}$ be any set of observation sequences of length t. We have that:*

$$|| \Pr[F|x] - \sum_{j=1}^{h} \beta_j \Pr[F|b_j]||_1 \leq \sqrt{L_{B,x}(\beta)}.$$

Proof We have that:

$$|| \Pr[F|x] - \sum_{j=1}^{h} \beta_j \Pr[F|b_j]||_1 = \mathbb{E}_{f\sim d}\left[\left|\frac{\Pr[f|x]}{d(f)} - \sum_{j=1}^{h}\beta_j \frac{\Pr[f|b_j]}{d(f)}\right|\right]$$
$$\leq \sqrt{\mathbb{E}_{f\sim d}\Big[\Big(\frac{\Pr[f|x]}{d(f)} - \sum_{j=1}^{h}\beta_j \frac{\Pr[f|b_j]}{d(f)}\Big)^2\Big]},$$

where the last step uses Jensen's inequality. ■

Proof [Proof of Theorem 3] We choose $C = \sqrt{2Tr\log(16/\alpha\varepsilon^2)}$, $H = 8rT^2\log(16/\varepsilon^2\alpha)$ and $n = O(1/\varepsilon^2\log(H/\delta))$. Let $\beta_h(x)$ be the coefficients such that

$$\beta_h(x) = \operatorname*{argmin}_{\beta\in\mathbb{R}^h, ||\beta||_2\leq C} \widehat{L}_{B_h,x}(\beta)$$

From $Proposition$ 41, using $\mathrm{poly}(r, T, 1/\varepsilon, 1/\alpha, \log(1/\delta))$ many conditionally samples, we can get with probability $1-\delta/2$, for all $h\in[H]$ and observation sequence x in our random sample

$$\left|L_{B_h,x}(\beta(x)) - \widehat{L}_{B_h,x}(\beta(x))\right| \leq \frac{\varepsilon^2}{32} \tag{38}$$

In our algorithm, when we find a counterexample, it means for some observation sequence x in the random sample:

$$\widehat{L}_{B_h,x}(\beta(x)) > \frac{\varepsilon^2}{8}$$

This means, by Equation (38), for that observation sequence x,

$$L_{B_h,x}(\beta(x)) > \frac{\varepsilon^2}{16}.$$

However, by our choice of C and H, by Proposition 42 this can not happen H times and therefore, our algorithm should end in at most H rounds.

We will now show that the overall error of our basis is small. When our algorithm ends, then we should have for all observation sequence x in our random sample:

$$\min_{\beta\in\mathbb{R}^h, ||\beta||_2\leq C} \widehat{L}_{B_h,x}(\beta) \leq \frac{\varepsilon^2}{8}$$

This means by Equation (38), we should have for all observation sequence x in our random sample:

$$\min_{\beta\in\mathbb{R}^h, ||\beta||_2\leq C} L_{B_h,x}(\beta) \leq \frac{\varepsilon^2}{4}.$$

By Proposition 44, this implies for all observation sequence x in our random sample:

$$\min_{\beta\in\mathbb{R}^h, ||\beta||_2\leq C} || \Pr[F|x] - \Pr[F|B]\beta||_1 \leq \frac{\varepsilon}{2}.$$

Therefore, for our choice of n, by Hoeffding inequality, we get with probability $1-\delta/2$,

$$\Pr_{x\sim p}\left[\min_{\beta\in\mathbb{R}^h,||\beta||_2\le C}||\Pr[F|x]-\Pr[F|B]\beta||_1>\varepsilon/2\right]\le\varepsilon/2$$

Since, for all histories x

$$\min_{\beta\in\mathbb{R}^h,||\beta||_2\le C}||\Pr[F|x]-\Pr[F|B]\beta||_1\le 1,$$

we get that

$$\mathbb{E}_{x\sim p}\left[\min_{\beta\in\mathbb{R}^h,||\beta||_2\le C}||\Pr[F|x]-\Pr[F|B]\beta(x)||_1\right]\le(1-\varepsilon/2)\varepsilon/2+\varepsilon/2\le\varepsilon$$

Let B' be the set where we repeat C times every $b\in B$, then we get that

$$\mathbb{E}_{x\sim p}\left[\min_{\beta\in\mathbb{R}^h,||\beta||_2\le 1}||\Pr[F|x]-\Pr[F|B]\beta(x)||_1\right]\le(1-\varepsilon/2)\varepsilon/2+\varepsilon/2\le\varepsilon$$

This completes the proof. ■

Appendix G. Helper propositions

Proposition 45 (Hoeffding's inequality) *Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $a\le X_i\le b$ almost surely. Consider the sum of these random variables,*

$$S_n = X_1+\cdots+X_n.$$

Then for all $t>0$,

$$\Pr\left[|S_n-\mathbb{E}\left[S_n\right]|\ge nt\right]\le 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

Here $E[S_n]$ is the expected value of S_n.

In our work, we care about how far are the projection operators onto top eigenspace for two symmetric matrices which are close to each other. This follows as a corollary of Davis-Kahan theorem.

Proposition 46 (Davis-Kahan theorem) *Let Σ and $\widehat{\Sigma}$ be real symmetric matrices with the eigenvalue decomposition: $V\Lambda_0V^\top + V^\perp\Lambda_1V^{\perp\top}$ and $\widehat{V}\widehat{\Lambda}_0\widehat{V}^\top+\widehat{V}^\perp\widehat{\Lambda}_1\widehat{V}^{\perp\top}$. If the eigenvalues of Λ_0 are contained in an interval $[a,b]$, and the eigenvalues of $\widehat{\Lambda}_1$ are excluded from the interval $[a-\gamma,b+\gamma]$ for some $\gamma>0$, then*

$$||\widehat{V}^{\perp\top}V||_F\le\frac{||\widehat{V}^{\perp\top}(\widehat{\Sigma}-\Sigma)V||_F}{\gamma}$$

Corollary 47 ((Peng, 2020)) *Let Σ, $\widehat{\Sigma}$, V, $\widehat{V}$ and γ be as defined above. Assume $V,\widehat{V}\in\mathbb{R}^{n\times r}$. Then,*

$$||VV^\top-\widehat{V}\widehat{V}^\top||_F\le\frac{\sqrt{2r}||\Sigma-\widehat{\Sigma}||_2}{\gamma}$$