Learning Mixtures of Markov Chains and MDPs

Chinmaya Kausik [1] **Kevin Tan** [2] **Ambuj Tewari** [2]

Abstract

We present an algorithm for learning mixtures of Markov chains and Markov decision processes (MDPs) from short unlabeled trajectories. Specifically, our method handles mixtures of Markov chains with optional control input by going through a multi-step process, involving (1) a subspace estimation step, (2) spectral clustering of trajectories using "pairwise distance estimators," along with refinement using the EM algorithm, (3) a model estimation step, and (4) a classification step for predicting labels of new trajectories. We provide end-to-end performance guarantees, where we only explicitly require the length of trajectories to be linear in the number of states and the number of trajectories to be linear in a mixing time parameter. Experimental results support these guarantees, where we attain 96.6% average accuracy on a mixture of two MDPs in gridworld, outperforming the EM algorithm with random initialization (73.2% average accuracy). We also significantly outperform the EM algorithm on real data from the LastFM song dataset.

1. Introduction

Efficiently clustering a mixture of time series data, especially with access to only short trajectories, is a problem that pervades sequential decision making and prediction (Liao (2005), Huang et al. (2021), Maharaj (2000)). This is motivated by various real-world problems, ranging through psychology (Bulteel et al. (2016)), economics (McCulloch & Tsay (1994)), automobile sensing (Hallac et al. (2017)), biology (Wong & Li (2000)), neuroscience (Albert (1991)), to name a few. One natural and important time series model is that of a mixture of K MDPs, which includes the case of a mixture of K Markov chains. We want to cluster from a set of short trajectories where (1) one does not know which MDP or Markov chain any trajectory comes from and (2) one does not know the transition structures of any of the K MDPs or Markov chains. Previous literature like Kwon et al. (2021) and Gupta et al. (2016) has stated and underlined the importance of this problem, but so far, the literature on methods to solve it with theoretical guarantees and empirical results has been sparse.

Broadly, there are three threads of literature on problems related to ours. Within reinforcement learning literature, there has been a sustained interest in frameworks very similar to mixtures of MDPs – latent MDPs (Kwon et al. (2021)), multi-task RL (Brunskill & Li (2013)), hidden model MDPs (Chades et al. (2021)), to name a few. However, most effort in this thread has been towards regret minimization in the online setting, where the agent interacts with an MDP from a set of unknown MDPs. The framework of latent MDPs in Kwon et al. (2021) is equivalent to adding reward information to ours. They have shown that one can only learn latent MDPs online with number of episodes required polynomial in states and actions to the power of trajectory length (under a reachability assumption similar to our mixing time assumption). On the other hand, our method learns latent MDPs offline with number of episodes needed only linear in the number of states (in no small part due to the subspace estimation step we make).

The other thread of literature deals with using a "subspace estimation" idea to efficiently cluster mixture models, from which we gain inspiration for our algorithm. Vempala & Wang (2004) first introduce the idea of using subspace estimation and clustering steps, with application to learning mixtures of Gaussians. Kong et al. (2020) adapt these ideas to the setting of meta-learning for mixed linear regression, adding a classification step. Chen & Poor (2022) bring these ideas to the time-series setting to learn mixtures of linear dynamical systems. They leave open the problems of (1) adapting the method to handle control inputs (mentioning mixtures of MDPs as an important example) and (2) handling other time series models (like autoregressive models and Markov chains), and state that the former is of great importance. There are many technical and algorithmic subtleties in adapting the ideas developed so far to MDPs and Markov Chains. The most obvious one comes from the following observation: in linear dynamical systems, the deviation from the predicted next-state value under the lin-

[1]Department of Mathematics, University of Michigan, Ann Arbor, USA [2]Department of Statistics, University of Michigan, Ann Arbor, USA. Correspondence to: Chinmaya Kausik <ckausik@umich.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

ear model occurs with additive i.i.d. noise. In MDPs and Markov chains, we are *sampling* from the next-state probability simplex at each timestep, and this cannot be cast as a deterministic function of the current state with additive i.i.d. noise.

Gupta et al. (2016) also provide a method for learning a mixture of Markov chains using only 3-trails, and compare its performance to the EM algorithm. While the requirement on trajectory length is as lax as can be, their method needs to estimate the distribution of 3-trails using all available data, incurring an estimation error in estimating S^3A^3 parameters, while providing no finite-sample theoretical guarantees. If the method can be shown to enjoy finite sample guarantees, the need to estimate S^3A^3 parameters indicates that the guarantees will scale poorly with S and A.

The problem that we aim to solve is the following.

Is there a method with finite-sample guarantees that can learn both mixtures of Markov chains and MDPs offline, with only data on trajectories and the number of elements in the mixture K?

1.1. Summary of Contributions

We provide such a method, with trajectory length requirements free from an S, A dependence. The method performs (1) subspace estimation, (2) spectral clustering, an optional step of using clusters to initialize the EM algorithm, (3) estimating models, and finally (4) classifying future trajectories.

Theorem (Informal)**.** *Ignoring logarithmic terms, we can recover all labels exactly with K^2S trajectories of length $K^{3/2}t_{mix}$, up to logarithmic terms and instance-dependent constants characterizing the models but not explicitly dependent on S, A, t_{mix} or K.*

Other contributions include:

- This is the first method, to our knowledge, that can cluster MDPs with finite-sample guarantees where the length of trajectories does not depend explicitly on S, A. The length only explicitly depends linearly on the mixing time t_{mix}, and the number of trajectories only explicitly depends linearly on S.
- We are able to provide theoretical guarantees while making no explicit demands on the policies and rewards used to collect the data, only relying on a difference in the transition structures at frequently occurring (s, a) pairs.
- Chen & Poor (2022) work under deterministic transitions with i.i.d. additive Gaussian noise, and we need to bring in non-trivial tools to analyse systems like ours, determined by transitions with non-i.i.d. additive noise. Our use of the blocking technique of Yu (1994) opens the door for the analysis of such systems.
- Empirical results in our experiments show that our method outperforms, outperforming the EM algorithm by a significant margin (73.2% for soft EM and 96.6% for us on gridworld).

2. Background and Problem Setup

We work in the scenario where we have K unknown models, either K Markov chains or K MDPs, and data of N_{traj} trajectories collected offline. Throughout the rest of the paper, we work with the case of MDPs, as we can think of Markov chains as an MDP where there is only one action ($A = \{*\}$) and rewards are ignored by our algorithm anyway.

We have a tuple $(\mathcal{S}, \mathcal{A}, \{\mathbb{P}_k\}_{k=1}^K, \{f_k\}_{k=1}^K, p_k)$ describing our mixture. Here, $\mathcal{S}, \mathcal{A}$ are the state and action sets respectively. $\mathbb{P}_k(s' \mid s, a)$ describes the probability of an s, a, s' transition under label k. At the start of each trajectory, we draw $k \sim \text{Categorical}(f_1, ..., f_K)$, and starting state according to p_k, and generate the rest of the trajectory under policies $\pi_k(a \mid s)$. We have stationary distributions on the state-action pairs $d_k(s, a)$ for π_k interacting with $\mathbb{P}_k$. We do not know (1) the parameters $\mathbb{P}_k, f_k, p_k, \pi_k(\cdot \mid s)$ of each model or the policies, and (2) k, i.e., which model each trajectory comes from.

This coincides with the setup in Gupta et al. (2016) in the case of Markov chains ($|\mathcal{A}| = 1$). It also overlaps with the setup of learning latent MDPs offline, in the case of MDPs. However, one difference is that we make no assumptions about the reward structure – once trajectories are clustered, we can learn the models, including the rewards. It is also possible to learn the rewards with a term in the distance measure that is alike to the model separation term. However, this would require extra assumptions on reward separation that are not necessary for clustering.

Assumption 1 (Mixing)**.** The K Markov chains on $\mathcal{S} \times \mathcal{A}$ induced by the behaviour policies π_k, each achieve mixing to a stationary distribution $d_k(s, a)$ with mixing time $t_{mix,k}$. Define the overall mixing time of the mixture of MDPs to be $t_{mix} := \max_k t_{mix,k}$.

Assumption 2 (Model Separation)**.** There exist α, Δ so that for each pair k_1, k_2 of hidden labels, there exists a state action pair (s, a) (possibly depending on k_1, k_2) so that $d_{k_1}(s, a), d_{k_2}(s, a) \geq \alpha$ and $\|\mathbb{P}_{k_1}(\cdot \mid s, a) - \mathbb{P}_{k_2}(\cdot \mid s, a)\|_2 \geq \Delta$.

Assumption 2 is merely saying that for any pair of labels, at least one visible state action pair witnesses a model difference Δ. Call this the separating state-action pair. If no visible pair witnesses a model difference between the labels, then one certainly cannot hope to distinguish them using

trajectories.

Remark 1. Why is there no assumption about policies? Notice that we make no explicit assumptions about policies. The nature of our algorithm allows us to work with the transition structure directly, and so we only demand that we observe a state action pair that witnesses a difference in transition structures. The policy is implicitly involved in this assumption through the stationary distribution $d_k(s,a)$ it induces, but our results demonstrate that this is the minimal demand we need to make in relation to the policies.

It is important to note that in Assumption 2, we assume a separation of the *MDP models themselves*, not the induced Markov chain models. To directly apply a method designed only for clustering mixtures of Markov chains to mixtures of MDPs, one would need to assume that the induced Markov chain models are separated. While our method handles Markov chains as described earlier, it does not need to use the induced Markov chain models for MDPs, and instead relies directly on the MDP models for clustering. This allows us to make a more natural assumption on separation.

Additionally, Assumption 1, which establishes the existence of a mixing time, is not a strong assumption (outside of the implicit hope that t_{mix} is small). This is because any irreducible aperiodic finite state space Markov chain mixes to a unique stationary distribution. If the Markov chain is not irreducible, it mixes to a unique distribution determined by the irreducible component of the starting distribution.

The only requirement is thus aperiodicity, which is also technically superficial, as we now clarify. If the induced Markov chains were periodic with period L, we would have a finite set of stationary distributions $d_{u,l}(s,a)$ that the chain would cycle through over a single period, indexed by $l = 1 \to L$. One can follow the proofs to verify that the guarantees continue to hold if we modify α in Assumption 2 to be a lower bound for $\min_{i,l} d_{u_i,l}(s,a)$ instead of just $\min_i d_{u_i}(s,a)$.

3. Algorithm

3.1. Setup and Notation

We have short trajectories of length T_n, divided into 4 segments of equal length. We call the second and fourth segment Ω_1 and Ω_2 respectively.[1] We further sub-divide Ω_i into G blocks, and focus only on the first state-action observation in each sub-block and its transition (discard all other observations). We often refer to these observations as "single-step sub-blocks." See Figure 1 for an illustration of this. Divide the set of trajectory indices into two sets and call them $\mathcal{N}_{sub}$ and $\mathcal{N}_{clust}$, for subspace estimation and clustering. Denote their sizes by N_{sub} and N_{clust} respectively. Let $\mathcal{N}_{traj}(s,a)$ be the set of trajectory indices in either $\mathcal{N}_{sub}$ or $\mathcal{N}_{clust}$ (to be inferred from the context) where (s,a) is observed in both Ω_1 and Ω_2. Let $N_{traj}(s,a)$ be the size of this set. Denote by $N(n,i,s,a)$ the number of times (s,a) is recorded in segment i of trajectory n, and let $\mathbf{N}(n,i,s,a,\cdot)$ be the vector of next-state counts. We denote by $\mathbb{P}_k(\cdot \mid s,a)$ the vector of next state transition probabilities. We denote by Freq_β the set of all state action pairs whose occurrence frequency in our observations is higher than β.

We will call the predicted clusters returned by the clustering algorithm $\mathcal{C}_k$. For model estimation and classification, we do not use segments, and merely split the entire trajectory into G blocks, discarding all but the last observation in each block. We call this observation the corresponding single-step sub-block. We denote the total count of s,a observations in trajectory n by $N(n,s,a)$ and that of s',s,a triples by $N(n,s,a,s')$.

In practice, we choose to not be wasteful and observations are not discarded while computing the transition probability estimates. To clarify, in that case $N(n,i,s,a)$ is just the count of (s,a) in segment i and similarly for $\mathbf{N}(n,i,s,a,\cdot), N(n,s,a)$ and $\mathbf{N}(n,s,a,\cdot)$. Estimators in both cases, that is both with and without discarding observations, are MLE estimates of the transition probabilities. One of them maximizes the likelihood of just the single-step sub-blocks and the other maximizes the likelihood of the entire segment. We need the latter for good finite-sample guarantees (using mixing). However, the former satisfies asymptotic normality, which is not enough for finite-sample guarantees, but it often makes it a good and less wasteful estimator in practice.

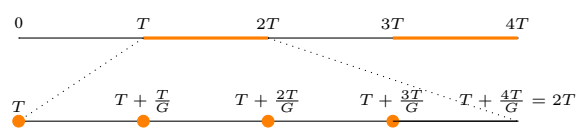

Figure 1. Breaking up a trajectory into 4 segments and G blocks per segment ($G = 4$) for the single-step estimator. Observations are only recorded at the orange points.

3.2. Overview

The algorithm amounts to (1) a PCA-like subspace estimation step, (2) spectral clustering of trajectories using "thresholded pairwise distance estimates," along with an optional step of using clusters to initialize the EM algorithm, (3) estimating models (MDP transition probailities) and finally (4) classifying any trajectories not in $\mathcal{N}_{clust}$ (for example, $\mathcal{N}_{sub}$). We provide performance guarantees for each step of the algorithm in section 4.

[1] As the proofs demonstrate, we do not technically need all segments to be equal. If the order of t_{mix} is known a priori, then we only need the first and third segment to be $\tilde{\Omega}(t_{mix})$ to allow for sufficient mixing before and after Ω_1.

3.3. Subspace Estimation

The aim of this algorithm is to estimate for each (s, a) pair a matrix $\mathbf{V}_{s,a}$ satisfying $\text{rowspan}\,\mathbf{V}_{s,a}^T \approx \text{span}(\mathbb{P}_k(\cdot|s,a))_{k=1,..,K}$. That is, we want to obtain an orthogonal projector to the subspace spanned by the next-state distributions $\mathbb{P}_k(\cdot|s,a)$ for $1 \leq k \leq K$.

Summarizing the algorithm in natural language, we perform subspace estimation via 3 steps. We first estimate the next state distribution given state and action for each trajectory. We then obtain the outer product of the next state distributions thus estimated. These outer product matrices are averaged over trajectories, and the average is used to find the orthogonal projectors $V_{s,a}^T$ to the top K eigenvectors.

Algorithm 1 Subspace Estimation

Compute $N_{traj}(s,a)$ for all s, a. Initialize the $S \times S$ matrix $\hat{\mathbf{M}}_{s,a} \leftarrow 0$ and the $SA \times SA$ matrix $\hat{\mathbf{D}} \leftarrow 0$.
$\hat{\mathbf{d}}_{n,1}, \hat{\mathbf{d}}_{n,2} \leftarrow \mathbf{0} \in \mathbb{R}^{SA}$ for all $n \in \mathcal{N}_{sub}$
for $(i, s, a) \in \{1, 2\} \times S \times A$ **do**
  Compute $\mathbf{N}(n, i, s, a, \cdot), N(n, i, s, a), \quad \forall n \in \mathcal{N}_{sub}$
  $\hat{\mathbb{P}}_{n,i}(\cdot|s,a) \leftarrow \frac{\mathbf{N}(n,i,s,a,\cdot)}{N(n,i,s,a)} \mathbb{1}_{N(n,i,s,a)\neq 0}, \quad \forall n$
  $[\hat{\mathbf{d}}_{n,i}]_{s,a} \leftarrow \frac{N(n,i,s,a)}{G}, \quad \forall n$
  $\hat{\mathbf{M}}_{s,a} \leftarrow \hat{\mathbf{M}}_{s,a} + \sum_{n \in \mathcal{N}_{traj}(s,a)} \frac{\hat{\mathbb{P}}_{n,1}(\cdot|s,a)\hat{\mathbb{P}}_{n,2}(\cdot|s,a)^T}{N_{traj}(s,a)}$
end for
$\hat{\mathbf{D}} \leftarrow \hat{\mathbf{D}} + \sum_{n \in \mathcal{N}_{sub}} \frac{1}{N_{sub}} \hat{\mathbf{d}}_{n,1} \hat{\mathbf{d}}_{n,2}^T$
Using SVD, return the orthogonal projectors $(\mathbf{V}_{s,a}^T)_{K \times S}$ to the top K eigenspaces of $\hat{\mathbf{M}}_{s,a} + \hat{\mathbf{M}}_{s,a}^T$ for each (s, a) where $N_{traj}(s, a) \neq 0$ (set the others to 0), along with the orthogonal projector $(\mathbf{U}^T)_{K \times SA}$ to the top K eigenspace of $\hat{\mathbf{D}} + \hat{\mathbf{D}}^T$.

Remark 2. Why do we split the trajectories? We use two approximately independent segments Ω_1 and Ω_2 time separated by a multiple of the mixing time t_{mix} to estimate the next state distributions. The reduced correlation between the two estimates obtained allows us to give theoretical guarantees for concentration, despite using dependent data within each trajectory n in the estimation of the rank 1 matrices $(\mathbb{P}_{k_n}(\cdot|s,a))(\mathbb{P}_{k_n}(\cdot|s,a))^T$. The key point is that the double estimator $\hat{\mathbb{P}}_{n,1}(\cdot \mid s,a)\hat{\mathbb{P}}_{n,2}(\cdot \mid s,a)$ is in expectation very close to this matrix.

Notice that our estimator $\hat{\mathbf{M}}_{s,a}$ is in expectation then given approximately by $\sum_{k=1}^{K} f_k(\mathbb{P}_k(\cdot|s,a))(\mathbb{P}_k(\cdot|s,a))^T$. The eigenspace of this matrix is clearly $\text{span}(\mathbb{P}_k(\cdot|s,a))_{k=1,..,K}$. The deviation from the expectation is controlled by the total number of trajectories, while the "approximation error" separating the expectation from the desired matrix is controlled by the separation between Ω_1 and Ω_2.

Remark 3. Why is this not PCA? This procedure has many linear-algebraic similarities to uncentered PCA on the dataset of (trajectories, next state frequencies), but statistically has a very different target. Crucially, (centered) PCA is concerned with the variance $\mathbb{E}[X^T X]$, while we are interested in a decent estimate of the target $\mathbb{E}[X^T]\mathbb{E}[X]$ above and thus use a double estimator. Our theoretical analysis also has nothing to do with analyses of PCA due to this difference in the statistical target.

3.4. Clustering

Using the subspace estimation algorithm's output, we can embed estimates from trajectories in a low dimensional subspace. For the clustering algorithm, we aim to compute the pairwise distances of these estimates from trajectories in this embedding. A double estimator is used yet again, to reduce the covariance between the two terms in the inner product used to compute such a distance.

This embedding is crucial because it reduces the variance of the pairwise distance estimators from a dependence on SA to a dependence on K. This is the intuition for how we can shift the onus of good clustering from being heavily dependent on the length of trajectories to being more dependent on the subspace estimate and thus on the number of trajectories.

There are many ways to use such "pairwise distance estimates" for clustering trajectories. In one successful example, we use a test: if the squared distances are below some threshold (details provided later), then we can conclude that they come from the same element of the mixture, and different ones otherwise. This allows us to construct (the adjacency matrix of) a graph with vertices as trajectories, and we can feed the results into a clustering algorithm like spectral clustering. Alternatively, one can use other graph partitioning methods or agglomerative methods on the distance estimates themselves.

We present the algorithm formally in Algorithm 2. Choosing hyperparameters β, λ and the threshold τ involve heuristic choices, much like how choosing the threshold in Chen & Poor (2022) needs heuristics. However, our methods are very different, and we describe them in more detail in Section 5.

3.4.1. Refinement using EM

Our guarantees in section 4 will show that we can recover exact clusters with high probability at the end of Algorithm 2. However, in practice, it makes sense to refine the clusters if trajectories are not long enough for exact clustering. Remember that an instance of the EM algorithm for any model is specified by choosing the observations Y, the hidden variables Z and the parameters θ.

If we consider observations to be next-state transitions from $(s, a) \in \text{Freq}_\beta$, hidden variables to be the hidden labels and the parameters θ to include both next-state transition

Algorithm 2 Clustering

Compute the set Freq_β by picking (s,a) pairs with occurrence more than β.
$\mathbf{d}_{n,1}, \mathbf{d}_{n,2} \leftarrow \mathbf{0} \in \mathbb{R}^{SA}$
for $(i,s,a) \in \{1,2\} \times S \times A$ **do**
  Compute $\mathbf{N}(n,i,s,a,\cdot), N(n,i,s,a), \quad \forall n \in \mathcal{N}_{clust}$
  $\hat{\mathbb{P}}_{n,i}(\cdot|s,a) \leftarrow \frac{\mathbf{N}(n,i,s,a,\cdot)}{N(n,i,s,a)} \mathbb{1}_{N(n,i,s,a)\neq 0}, \quad \forall n$
  $[\hat{\mathbf{d}}_{n,i}]_{s,a} \leftarrow \frac{N(n,i,s,a)}{G}, \quad \forall n$
end for
for $(n,m) \in \mathcal{N}_{clust} \times \mathcal{N}_{clust}$ **do**
  for $(i,s,a) \in \{1,2\} \times S \times A$ **do**
    $\hat{\mathbf{\Delta}}_{i,s,a} := \mathbf{V}_{s,a}^T(\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) - \hat{\mathbb{P}}_{m,i}(\cdot|s,a))$
  end for
  $\text{dist}_1(n,m) := \max_{(s,a)\in \text{Freq}_\beta} \hat{\mathbf{\Delta}}_{1,s,a}^T \hat{\mathbf{\Delta}}_{2,s,a}$
  $\text{dist}_2(n,m) := (\hat{\mathbf{d}}_{n,1} - \hat{\mathbf{d}}_{m,1})^T \mathbf{U}\mathbf{U}^T(\hat{\mathbf{d}}_{n,2} - \hat{\mathbf{d}}_{m,2})$
  $\text{dist}(n,m) := \lambda \, \text{dist}_1(n,m) + (1-\lambda)\, \text{dist}_2(n,m)$
end for
Plot a histogram of dist to determine threshold τ and cluster trajectories $\text{sim}(n,m) := \mathbb{1}_{\text{dist}(n,m)\leq\tau}$

probabilities for $(s,a) \in \text{Freq}_\beta$ and cluster weights $\hat{f}_k$, then one can now refine the clusters using the EM algorithm on this setup, which enjoys monotonicity guarantees in log-likelihood if one uses soft EM. The details of the EM algorithm are quite straightforward, described in Appendix C.

We hope that this is a step towards unifying the discussion on spectral and EM methods for learning mixture models, highlighting that we need not choose between one or the other – spectral methods can initialize the EM algorithm, in one reinterpretation of the refinement step.

Note that refinement using EM is not unique to our algorithm. The model estimation and classification steps in Kong et al. (2020) (under the special case of Gaussian noise) and Chen & Poor (2022) (who already assume Gaussian noise) are exactly the E-step and M-step of the hard EM algorithm as well.

3.5. Model Estimation and Classification

Given clusters from the clustering and refinement step, 2 tasks remain, namely those of estimating the models from them and correctly classifying any future trajectories. We can estimate the models exactly as in the M-step of hard EM.

$$\hat{\mathbb{P}}_k(s'|s,a) \leftarrow \frac{\sum_{n\in\mathcal{C}_k} N(n,s,a,s')}{\sum_{n\in\mathcal{C}_k} N(n,s,a)}$$

$$\hat{f}_k \leftarrow \frac{|\mathcal{C}_k|}{N_{clust}}$$

For classification, given a set $\mathcal{N}_{class}$ of trajectories with size N_{class} generated independently of $\mathcal{N}_{clust}$, we can run a process very similar to Algorithm 2 to identify which cluster to assign each new trajectory to. It is worth noting that we can run the classification step on the subspace estimation dataset itself and recover true labels for those trajectories, since trajectories in $\mathcal{N}_{sub}$ and $\mathcal{N}_{clust}$ are independent.

We describe the algorithm in natural language here. The algorithm is presented formally as Algorithm 3 in Appendix D. We first compute an orthogonal projector $\tilde{\mathbf{V}}_{s,a}$ to the subspace spanned by the now known approximate models $\hat{\mathbb{P}}_k(\cdot \mid s,a)$. For any new trajectory n and label k, we estimate a distance $\text{dist}(n,k)$ between the model $\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a)$ estimated from n and the model $\hat{\mathbb{P}}_k(\cdot \mid s,a)$ for k, after embedding both in the subspace mentioned above using $\tilde{\mathbf{V}}_{s,a}$. Again, we use a double estimator as hinted at by the use of the subscript i, similar to Algorithm 2. In practice $\text{dist}(n,k)$ could also include occupancy measure differences. Each trajectory n gets the label k_n that minimizes $\text{dist}(n,k)$.

Previous work like Chen & Poor (2022) and Kong et al. (2020) uses the word refinement for its model estimation and classification algorithms themselves. However, we posit that the monotonic improvement in log-likelihood offered by EM makes it well-suited for *repeated application and refinement*, while in our case, the clear theoretical guarantees for the model estimation and classification algorithms make them well-suited for *single-step classification*. Note that we can also apply repeated refinement using EM to the labels obtained by single-step classification, which should combine the best of both worlds.

4. Analysis

We have the following end-to-end guarantee for correctly classifying all data.

Theorem 1 (End-to-End Guarantee)**.** *Let both N_{sub} and N_{clust} be $\Omega\left(K^2 S \frac{\log(1/\delta)}{f_{min}^2 \alpha^3 \Delta^8}\right)$ and let $T_n = \Omega\left(K^{3/2} t_{mix} \frac{\log^4((N_{clust}+N_{sub})/\delta)\log^3(1/\Delta)\log^4(1/\alpha)}{\Delta^6\alpha^3}\right)$. If we execute algorithms 1, 2 and model estimation, and then apply algorithm 3 to $\mathcal{N}_{sub}$ with $\lambda = 1$, $\alpha/3 \leq \beta < \alpha$ and $\Delta^2/4 \leq \tau \leq \Delta^2/2$ for clustering and classification, then we can recover the true labels for the entire dataset ($\mathcal{N}_{clust} \cup \mathcal{N}_{sub}$) with probability at least $1-\delta$.*

Proof. This follows directly from Theorems 2, 3, 4 and 5 upon combining the conditions on N_{sub}, N_{clust}, and T_n in both theorems. We also use the brief discussion after the statement of Theorem 5. □

The dependence on model-specific parameters like α, Δ and f_{min} is conservative and can be easily improved upon by

following the proofs carefully. We chose the form of the guarantees in this section to present a clearer message. In one example, there are versions of these theorems that depend on both G and T_n. We choose $G = (T_n/t_{mix})^{2/3}$ to present crisper guarantees. For understanding how the guarantees would behave depending on both G and T_n, or how to improve the dependence on model-specific parameters, the reader can follow the proofs in the appendix.

4.1. Techniques and Proofs

We make a few remarks on the technical novelty of our proofs. As mentioned in Section 1, we are dealing with two kinds of non-independence. While we borrow some ideas in our analysis from Chen & Poor (2022) to deal with the temporal dependence, we crucially need new technical inputs to deal with the fact that we cannot cast the temporal evolution as a deterministic function with additive i.i.d. noise, unlike in linear dynamical systems.

We identify the blocking technique in Yu (1994) as a general method to leverage the "near-independence" in observations made in a mixing process when they are separated by a multiple of the mixing time. Our proofs involve first showing that estimates made from a single trajectory would concentrate if the observations were independent, and then we bound the "mixing error" to account for the non-independence of the observations. We first choose a distribution (often labelled as a variant of Q or Ξ) with desirable properties, and then bound the difference between probabilities of undesirable events under Q and under the true joint distribution of observations χ, using the blocking technique due to Yu (1994).

There are many other technical subtleties here. In one example, the number of (s, a) observations made in a single trajectory is itself a random variable and so our estimator takes a ratio of two random variables. To resolve this, we have to condition on the random set of (s, a) observations recorded in a trajectory and use a conditional version of Hoeffding's inequality (different from the Azuma-Hoeffding inequality), followed by a careful argument to get unconditional concentration bounds, all under Q.

4.2. Subspace Estimation

For subspace estimation, we have the following guarantee.

Theorem 2 (Subspace Estimation Guarantee)**.** *Consider* 2 *models with labels* k_1, k_2 *and a state-action pair* s, a *with* $d_{min}(s, a) \geq \alpha/3$*. Consider the output* $\boldsymbol{V}_{s,a}^T$ *of Algorithm 1. Let* $f_{min} = \min(f_{k_1}, f_{k_2})$ *be the lower of the label prevalences. Remember that each trajectory has length* T_n*.*

Then given that $N_{sub} = \Omega\left(\frac{\log(1/\delta)}{\alpha^2}\right)$*,* $T_n = \Omega(t_{mix} \log^4(1/\alpha))$*, with probability at least* $1 - \delta$*, for* $k = k_1, k_2$

$$\|\mathbb{P}_k(\cdot \mid s, a) - \boldsymbol{V}_{s,a}\boldsymbol{V}_{s,a}^T \mathbb{P}_k(\cdot \mid s, a)\|_2 \leq \epsilon_{sub}(\delta)$$

where

- *For* $T_n = \Omega\left(t_{mix} \log^3\left(\frac{f_{min} N_{sub} \alpha}{KS \log(1/\delta)}\right)\right)$

$$\epsilon_{sub}(\delta) = O\left(\sqrt{\frac{K}{f_{min}}\left(\sqrt{\frac{S}{N_{sub} \cdot \alpha^3} \log\left(\frac{1}{\delta}\right)}\right)}\right)$$

- *While for* $T_n = O\left(t_{mix} \log^3\left(\frac{f_{min} N_{sub} \alpha}{KS \log(1/\delta)}\right)\right)$

$$\epsilon_{sub}(\delta) = O\left(\left(\frac{1}{2}\right)^{\frac{1}{16}\left(\frac{T_n}{t_{mix}}\right)^{1/3}}\right)$$

Alternatively, we only need $N_{sub} = \Omega\left(\frac{K^2 S \log(1/\delta)}{f_{min}^2 \alpha^3 \epsilon^4}\right)$ *and* $T_n = \Omega\left(t_{mix} \log^3(1/\epsilon) \log^4(1/\alpha)\right)$ *trajectories for* ϵ *accuracy in subspace estimation with probability at least* $1 - \delta$*.*

Remark 4. Why are short trajectories enough? Notice that the length of trajectories only affects the bound as a multiple of t_{mix} with some logarithmic terms. This is because intuitively, the onus of estimating the correct subspace lies on aggregating information across trajectories. So, as long as there are enough trajectories, each trajectory does not have to be long.

4.3. Clustering

Remember that Δ is the model separation and α is the corresponding "stationary occupancy measure" from Assumption 2. We give guarantees for choosing $\lambda = 1$, which corresponds to using only model difference information instead of also using occupancy measure information. This is unavoidable since we have no guarantees on the separation of occupancy measures. See Section 5.2 for a discussion. Here, we provide a high-probability guarantee for exact clustering.

Theorem 3 (Exact Clustering Guarantee)**.** *Pick any pair of trajectories* n, m*. Then for* Freq_β *so that it contains* (s, a) *with* $d_{min}(s, a) \geq \Omega(\alpha)$*,* $T_n = \Omega(t_{mix} \log^4(1/\delta)/\alpha^3)$*, with probability at least* $1 - \delta$*,*

$$\left|\mathrm{dist}_1(m, n) - \|\Delta_{m,n}\|_2^2\right|$$

is

$$O\left(\sqrt{\frac{K \log(1/\delta)}{\alpha}}\left(\frac{t_{mix}}{T_n}\right)^{\frac{1}{3}}\right) + 4\epsilon_{sub}(\delta/2)$$

This means that if we choose $\lambda = 1$, then if $\epsilon_{sub}(\delta) \leq \Delta^2/32$ and $T_n = \Omega\left(K^{3/2} t_{mix} \frac{\log^4(N_{clust}/(\alpha\delta))}{\Delta^6 \alpha^3}\right)$, no distance estimate attains a value between $\Delta^2/4$ and $\Delta^2/2$. So, Algorithm 2 attains exact clustering using a threshold of say $\Delta^2/3$ with probability at least $1-\delta$.

Since we already have high probability guarantees for exact clustering before refinement of the clusters, guarantees for the EM step analogous to the single-step guarantees for refinement in Chen & Poor (2022) are not useful here. However, we do still present single-step guarantees for the EM algorithm in our case using a combination of Theorem 4 for the M-step and Theorem 6 in Appendix G.

4.4. Model Estimation and Classification

We also have guarantees for correctly estimating the relevant parts of the models and classifying sets of trajectories different from $\mathcal{N}_{clust}$.

Theorem 4 (Model Estimation Guarantee)**.** *For any state action pair (s,a) with $d_{min}(s,a) \geq \alpha/3$, and for $GN_{clust} \geq \Omega\left(\frac{\log(1/\delta)}{f_{min}^2 \alpha^2}\right)$ and $T_n \geq \Omega(Gt_{mix}\log(G/\delta))$, with probability greater than $1-\delta$,*

$$\|\hat{\mathbb{P}}_k(\cdot \mid s,a) - \mathbb{P}_k(\cdot \mid s,a)\|_1$$

is bounded above by

$$O\left(\left(\frac{t_{mix}}{T_n}\right)^{1/3} \sqrt{\frac{1}{N_{clust} f_{min} \alpha}(S + \log(\frac{1}{\delta}))}\right)$$

Note that since the 1-norm is greater than the 2-norm, the same bound holds in the 2-norm as well. Also notice that since our assumptions do not say anything about observing all (s,a) pairs often enough, we can only given guarantees in terms of the occurrence frequency of (s,a) pairs.

Theorem 5 (Classification Guarantee)**.** *Let $\epsilon_{mod}(\delta)$ be a high probability bound on the model estimation error $\|\hat{\mathbb{P}}_k(\cdot \mid s,a) - \mathbb{P}_k(\cdot \mid s,a)\|_2$. Then there is a universal constant C_3 so that Algorithm 3 can identify the true labels for trajectories in $\mathcal{N}_{class}$ with probability at least $1-\delta$ for $T_n = \Omega\left(K^{3/2} t_{mix} \frac{\log^4(N_{class}/(\alpha\delta))}{\Delta^6 \alpha^3}\right)$, whenever $\epsilon_{mod}(\delta/2) \leq \frac{C_3 \Delta^4 f_{min} \alpha}{K}$ and $N_{clust} \geq \Omega\left(\frac{\log(1/\delta)}{f_{min}^2 \alpha^2}\right)$.*

Note that by Theorem 4, a sufficient condition for $\epsilon_{mod}(\delta/2) \leq \frac{C_3 \Delta^4 f_{min} \alpha}{K}$ is $N_{clust} T_n^{2/3} \geq \Omega\left(K^2 t_{mix}^{2/3} S \frac{\log(1/\delta)}{\Delta^8 f_{min}^3 \alpha^3}\right)$. Under the conditions on T_n in Theorem 5, a suboptimal but sufficient condition on N_{clust} is $N_{clust} = \Omega\left(K^2 S \frac{\log(1/\delta)}{f_{min}^2 \alpha^3 \Delta^8}\right)$, which matches that for N_{sub}.

5. Practical Considerations

5.1. Subspace Estimation

Heuristics for choosing K: One often does not know K beforehand and often wants temporal features to guide the process of determining K, for example in identifying the number of groups of similar people represented in a medical study. We suggest a heuristic for this. One can examine how many large eigenvalues there are in the decomposition, via (1) ordering the eigenvalues of $\hat{\mathbf{M}}_{sa}$ $\forall s,a$ by magnitude, (2) taking the square of each to obtain the eigenvalue energy, (3) taking the mean or average over states and actions, and (4) plotting a histogram. See Figure 6 in the appendix.

One can also consider running the whole process with different values of K and choose the value of K that maximises the likelihood or the AIC of the data (if one wishes the mixture to be sparse). However, Fitzpatrick & Stewart (2022) points out that such likelihood-based methods can lead to incorrect predictions for K even with infinite data.

5.2. Clustering

Picking β: Choosing β involves heuristically picking state-action pairs that have high frequency and "witness" enough model separation. We propose one method for this. For each (s,a) pair, one first executes subspace estimation and then averages the value of $\mathrm{dist}_1(m,n)$ across all pairs of trajectories. Call this estimate $\Delta_{s,a}$, since it is a measure of how much model separation (s,a) can "witness". We then compute the occupancy measure value $d(s,a)$ of (s,a) in the entire set of observations. Making a scatter-plot of $\Delta_{s,a}$ against $d(s,a)$, we want a value of β so that there are enough pairs from Freq_β in the top right.

Picking thresholds τ: The histogram of dist plotted will have many modes. The one at 0 reflects distance estimates between trajectories belonging to the same hidden label, while all the other modes reflect distance between trajectories coming from various pairs of hidden labels. The threshold should thus be chosen between the first two modes. See Figure 8 in the appendix.

Picking λ: In general, occupancy measures are different for generic policies interacting with MDPs and should be included in the implementation by choosing $\lambda < 1$. The histogram for dist_2 should indicate whether or not occupancy measures allow for better clustering (if they have the right number of well-separated modes).

Versions of the EM algorithm: In our description of the EM algorithm, we only use next-state transitions as observations instead of the whole trajectory. So, we do not learn other parameters like the policy and the starting state's distribution for the EM algorithm. This makes sense in principle, because our minimal assumptions only talk about separation in next-state transition probabilities, and there is no guarantee that other information will help with classification. In

practice, one should make a domain-specific decision on whether or not to include them.

Initializing soft EM with cluster labels: We also recommend that when one initializes the soft EM algorithm with results from the clustering step, one introduces some degree of uncertainty instead of directly feeding in the 1-0 clustering labels. That is, for trajectory m, instead of assigning $\mathbb{1}(i = k_m)$ to be the responsibilities, make them say $0.8 \cdot \mathbb{1}(i \in \mathcal{C}_k) + 0.2/K$ instead. We find that this can aid convergence to the global maximum, and do so in our experiments.

6. Experiments

We perform two sets of experiments, one considering a mixture of MDPs, and another considering a mixture of Markov chains.[2] In all experiments, the error is determined by matching clusters to labels in a way that minimizes the proportion of misclassified trajectories, and then reporting this proportion.

6.1. Gridworld MDPs, $K = 2$

We perform our experiments for MDPs on an 8x8 gridworld with $K = 2$ elements in the mixture (from (Bruns-Smith, 2021)). Unlike Bruns-Smith (2021), the behavior policy here is the same across both elements of the mixture to eliminate any favorable effects that a different behavior policy might have on clustering, so that we evaluate the algorithm on fair grounds. The mixing time of this system is roughly $t_{mix} \approx 25$. We only use dist_1 for the clustering, omitting the occupancy measures to parallel the theoretical guarantees. Including them would likely improve performance. We chose to perform the experiments with 1000 trajectories, given the difficulty of obtaining large numbers of trajectories in important real-life scenarios that often arise in areas like healthcare.

Figure 2 plots the error at the end of Algorithm 2 (before refinement) while either using the projectors $\mathbf{V}_{s,a}^T$ determined in Algorithm 1 ("With Subspaces"), replacing them with a random projector ("Random Subspaces") or with the identity matrix ("Without Subspaces"). The difference in performance demonstrates the importance of our structured subspace estimation step. Also note that past a certain point, between $T_n = 60$ and $T_n = 70 \sim 3t_{mix}$, the performance of our method drastically improves, showing that the dependence of our theoretical guarantees on the mixing time is reflected in practice as well. We briefly discuss the poor performance of choosing a random subspace in Appendix B.

In Figure 3, we benchmark our method's end-to-end performance against the most natural benchmark, the randomly initialized EM algorithm. We use the version of the soft EM algorithm that considers the entire trajectory to be our observation, and thus also includes policies and starting state distributions. So, we are comparing our method against the full power of the EM algorithm. We have three different plots, corresponding to (1) soft EM with random initialization, (2) refining models obtained from the model estimation step applied to $\mathcal{N}_{clust}$ using soft EM on $\mathcal{N}_{clust} \cup \mathcal{N}_{sub}$, and (3) refining labels for $\mathcal{N}_{clust}$ and $\mathcal{N}_{sub}$ using soft EM (the latter obtained from applying Algorithm 3 to $\mathcal{N}_{sub}$). We report the final label accuracies over the entire dataset, $\mathcal{N}_{clust} \cup \mathcal{N}_{sub}$. Remember that we can view refinement using soft EM as initializing soft EM with the outputs of our algorithms. Note that the plot for (3), which reflects the true end-to-end version of our algorithm, almost always outperforms randomly initialized soft EM. Also, for $T_n > 60$, both variants of our method outperform randomly initialized soft EM. We present a variant of Figure 3 with hard EM included as Figure 10 in the appendix.

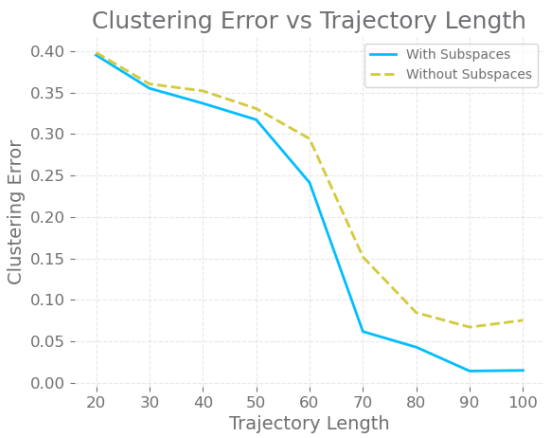

Figure 2. Clustering error v.s. trajectory length on 1000 trajectories in gridworld, with a comparison between using $\mathbf{V}_{s,a}^T$ in Algorithm 2 and using $I_{S\times S}$. The same threshold was used for each trajectory length. Results averaged over 30 trials. The mixing time of this system is roughly $t_{mix} \approx 25$.

6.2. Last.fm Markov chains, $K = 10$

For our experiments with real-life data, we work with the Last.fm 1K dataset (Celma, 2010b; Lamere, 2008; Celma, 2010a). Like Gupta et al. (2016), we consider the listening history of individual users, modeled as a Markov chain with states given by the top 100 genres ($S = 100$). For each of the top 10 users, we chop up their listening history into 75 trajectories ($N_{sub} + N_{clust} = 75$) each, of varying horizons. The user generating a trajectory is then the hidden label to be inferred. We try to infer both the user corresponding to each trajectory and a Markovian model of each user's listening dynamics, using the Hungarian algorithm to compute the

[2]Code is available at https://github.com/hetankevin/mdpmix.

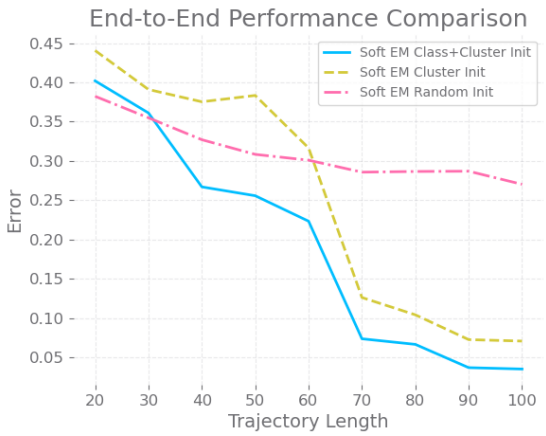

Figure 3. End-to-end error v.s. trajectory length on 1000 trajectories in gridworld, comparing initializations of the soft EM algorithm using (1) random initializations, (2) models from $\mathcal{N}_{clust}$, and (3) classification and clustering labels from $\mathcal{N}_{clust}$ and $\mathcal{N}_{sub}$. Results averaged over 30 trials, with 30 random initializations for randomly-initialized EM within each trial.

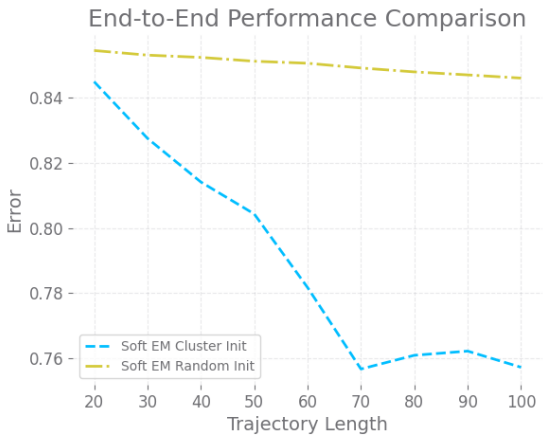

Figure 5. End-to-end error v.s. trajectory length on 750 trajectories obtained from the Last.fm 1K dataset, comparing initializations of the soft EM algorithm using (1) random initializations and (2) models from $\mathcal{N}_{clust}$. Results averaged over 30 trials, with 30 random initializations for randomly-initialized EM within each trial.

clustering error. The results are qualitatively similar to the results in the gridworld experiment above. Figure 4 demonstrates the importance of the subspace estimation step, and Figure 5 demonstrates that our method's end-to-end performance improves upon that of randomly initialized EM algorithm.

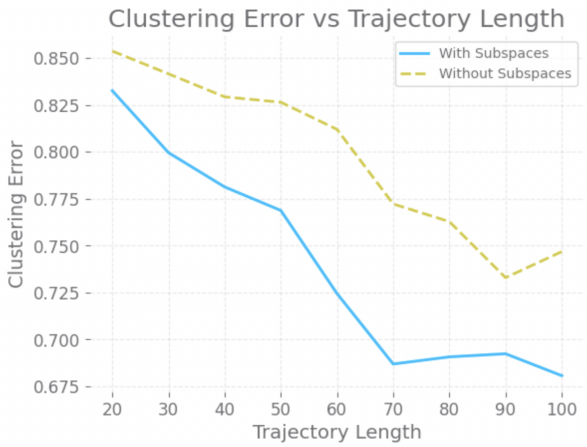

Figure 4. Clustering error v.s. trajectory length on 750 trajectories obtained from the Last.fm 1K dataset, comparing an implementation of Algorithm 2 using $\mathbf{V}_{s,a}^T$ with one using $I_{S\times S}$. Results averaged over 30 trials.

7. Discussion

We have shown that we can recover the true trajectory labels with (1) the number of trajectories having only a linear dependence in the size of the state space, and (2) the length of the trajectories depending only linearly in the mixing time – even before initializing the EM algorithm with these clusters (which would further improve the log-likelihood, and potentially cluster accuracy). End-to-end performance guarantees are provided in Theorem 1, and experimental results are both promising and in line with the theory.

7.1. Future Work

Matrix sketching: The computation of $\mathrm{dist}_1(m,n)$ is computationally intensive, amounting to computing about $S\times A$ distance matrices. We could alternatively approximate the thresholded version of the matrix $\mathrm{dist}(m,n)$ (which in the ideal case is a rank-K binary matrix) with ideas from Musco & Musco (2016).

Function approximation: The question of the right extension of our ideas to Markov chains and MDPs with large, infinite, or uncountable state spaces is very much open (at least, those whose transition kernel is not described by a linear dynamical systems). This is important, as many applications often rely on continuous state spaces.

Other controlled processes: Chen & Poor (2022) learn a mixture of linear dynamical systems without control input. An extension to the case with control input will be very valuable. We believe that the techniques used in our work may prove useful in this, as well as for extensions to other controlled processes that may neither be linear nor Gaussian.

8. Acknowledgements

Ambuj Tewari would like to acknowledge the support of NSF via grant IIS-2007055.

References

Albert, P. S. A two-state markov mixture model for a time series of epileptic seizure counts. *Biometrics*, 47(4):1371–1381, 1991. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2532392.

Bruns-Smith, D. A. Model-free and model-based policy evaluation when causality is uncertain. In *International Conference on Machine Learning*, pp. 1116–1126. PMLR, 2021.

Brunskill, E. and Li, L. Sample complexity of multi-task reinforcement learning. *Uncertainty in Artificial Intelligence - Proceedings of the 29th Conference, UAI 2013*, 09 2013.

Bulteel, K., Tuerlinckx, F., Brose, A., and Ceulemans, E. Clustering vector autoregressive models: Capturing qualitative differences in within-person dynamics. *Frontiers in Psychology*, 7, 2016. ISSN 1664-1078. doi: 10.3389/fpsyg.2016.01540. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2016.01540.

Celma, O. *Music Recommendation and Discovery in the Long Tail*. Springer, 2010a.

Celma, O. Last.fm Dataset – 1K users. http://www.dtic.upf.edu/~ocelma/MusicRecommendationDataset/lastfm-1K.html, 2010b.

Chades, I., Carwardine, J., Martin, T., Nicol, S., Sabbadin, R., and Buffet, O. Momdps: A solution for modelling adaptive management problems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 26(1):267–273, Sep. 2021. doi: 10.1609/aaai.v26i1.8171. URL https://ojs.aaai.org/index.php/AAAI/article/view/8171.

Chen, Y. and Poor, H. V. Learning mixtures of linear dynamical systems. *CoRR*, abs/2201.11211, 2022. URL https://arxiv.org/abs/2201.11211.

Fitzpatrick, M. and Stewart, M. Asymptotics for markov chain mixture detection. *Econometrics and Statistics*, 22:56–66, 2022. ISSN 2452-3062. doi: https://doi.org/10.1016/j.ecosta.2021.11.004. URL https://www.sciencedirect.com/science/article/pii/S2452306221001337. The 2nd Special issue on Mixture Models.

Gupta, R., Kumar, R., and Vassilvitskii, S. On mixtures of markov chains. In *NIPS*, pp. 3441–3449, 2016. URL http://papers.nips.cc/paper/6078-on-mixtures-of-markov-chains.

Hallac, D., Vare, S., Boyd, S., and Leskovec, J. Toeplitz inverse covariance-based clustering of multivariate time series data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pp. 215–223, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3098060. URL https://doi.org/10.1145/3097983.3098060.

Huang, L., Sudhir, K., and Vishnoi, N. Coresets for time series clustering. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 22849–22862. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/c115ba9e04ab27fbbb664f932112246d-Paper.pdf.

Kong, W., Somani, R., Song, Z., Kakade, S. M., and Oh, S. Meta-learning for mixed linear regression. *CoRR*, abs/2002.08936, 2020. URL https://arxiv.org/abs/2002.08936.

Kwon, J., Efroni, Y., Caramanis, C., and Mannor, S. RL for latent mdps: Regret guarantees and a lower bound. *CoRR*, abs/2102.04939, 2021. URL https://arxiv.org/abs/2102.04939.

Lamere, P. LastFM-ArtistTags2007 dataset. http://musicmachinery.com/2010/11/10/lastfm-artisttags2007/, 2008.

Larsen, K. G. and Nelson, J. Optimality of the johnson-lindenstrauss lemma. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 633–638, 2017. doi: 10.1109/FOCS.2017.64.

Liao, T. W. Clustering of time series data—a survey. *Pattern Recognition*, 38(11):1857–1874, nov 2005. doi: 10.1016/j.patcog.2005.01.025. URL https://doi.org/10.1016%2Fj.patcog.2005.01.025.

Maharaj, E. A. Cluster of time series. *Journal of Classification*, 17(2):297–314, Jul 2000. ISSN 1432-1343. doi: 10.1007/s003570000023. URL https://doi.org/10.1007/s003570000023.

McCulloch, R. and Tsay, R. Statistical analysis of economic time series via markov switching models. *Journal of Time Series Analysis*, 15(5):523–539, 1994. ISSN 0143-9782. doi: 10.1111/j.1467-9892.1994.tb00208.x.

Musco, C. and Musco, C. Recursive sampling for the nyström method. 2016. doi: 10.48550/ARXIV.1605.07583. URL https://arxiv.org/abs/1605.07583.

Vempala, S. and Wang, G. A spectral algorithm for learning mixture models. *J. Comput. Syst. Sci*, 68:2004, 2004.

Vidyasagar, M. *Learning and Generalization: With Applications to Neural Networks*. Springer Publishing Company, Incorporated, 2nd edition, 2010. ISBN 1849968675.

Wong, C. S. and Li, W. K. On a mixture autoregressive model. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(1):95–115, 2000. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/2680680.

Wong, W. H. and Shen, X. Probability Inequalities for Likelihood Ratios and Convergence Rates of Sieve MLES. *The Annals of Statistics*, 23(2):339 – 362, 1995. doi: 10.1214/aos/1176324524. URL https://doi.org/10.1214/aos/1176324524.

Yu, B. Rates of Convergence for Empirical Processes of Stationary Mixing Sequences. *The Annals of Probability*, 22(1):94 – 116, 1994. doi: 10.1214/aop/1176988849. URL https://doi.org/10.1214/aop/1176988849.

A. Additional Figures

All figures here pertain to the gridworld experiment in Section 6.1.

A.1. Determining K

See Figure 6 below, following the discussion in section 5.1.

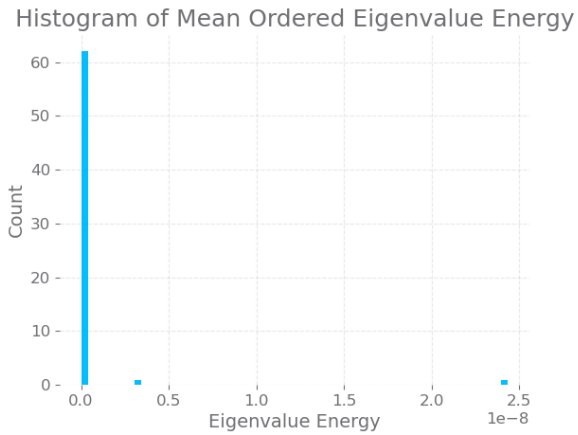

Figure 6. Histogram of the average ordered eigenvalue energy (the square of the eigenvalue) where the mean is taken over states and actions. There are two large eigenvalues, corresponding to $K = 2$.

A.2. Block Matrix of Raw Distance Estimates

See Figure 7 below, which presents the raw distance matrix before thresholding, to provide a sense of the quality of the pairwise distance estimates themselves. These could also be used for agglomerative clustering, for example.

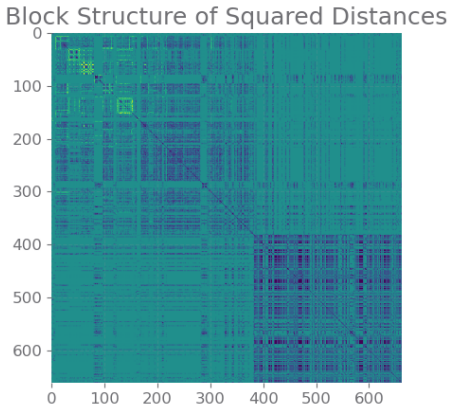

Figure 7. Block structure of the matrix of squared pairwise distance estimates (after sorting).

A.3. Determining The Threshold τ

See Figure 8 below, following the discussion in section 5.2.

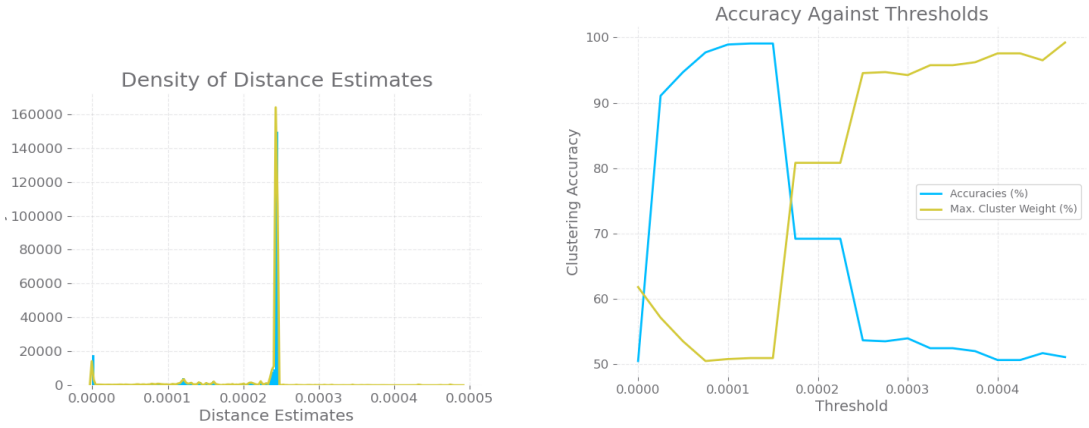

Figure 8. Histogram (and KDE) of pairwise squared distance estimates in projected subspace above, and accuracy against thresholds below. Note how there is a spurious mode around the 0.00015 mark, and picking any threshold past it yields a significant drop in accuracy.

A.4. Local Extrema in EM

See Figure 9 below, illustrating how EM often gets stuck in suboptimal local extrema, given by the low final log-likelihood values recorded in the scatterplot.

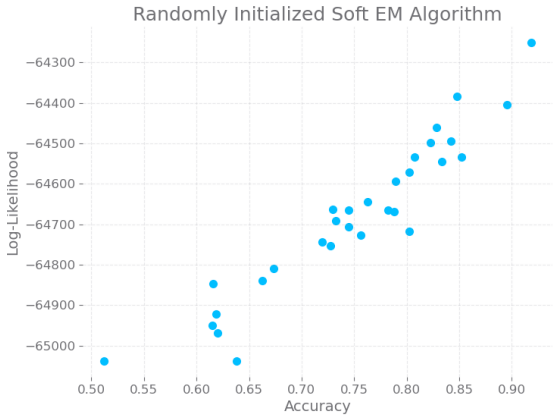

Figure 9. Scatter-plot of likelihoods v.s. clustering accuracy achieved by the randomly-initialized soft EM algorithm over 30 trials on gridworld. Randomly-initialized soft EM does not achieve the global maximum all of the time.

A.5. Comparing End-To-End Performance Using Soft and Hard EM

We compare various initializations of EM – (1) random initializations, (2) models from $\mathcal{N}_{clust}$, and (3) classification and clustering labels from $\mathcal{N}_{clust}$ and $\mathcal{N}_{sub}$ – this time using both soft and hard EM.

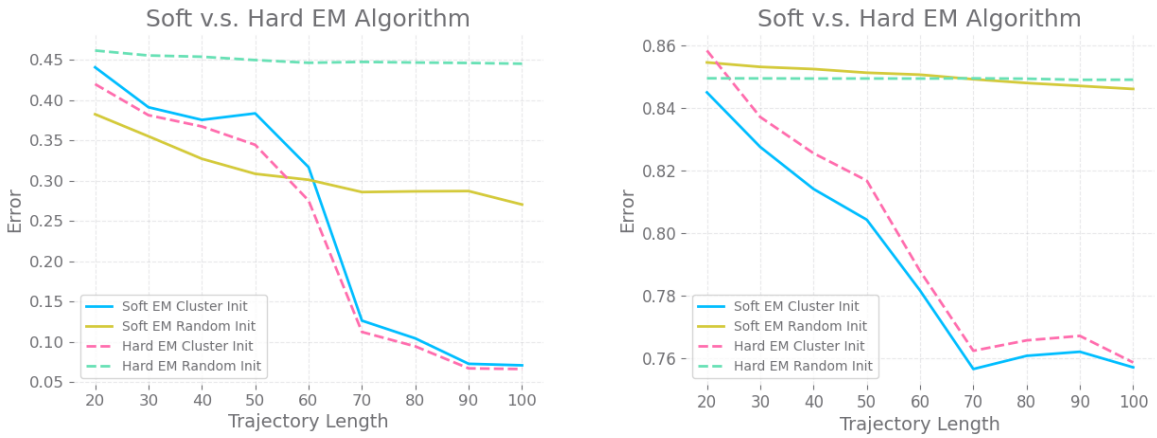

Figure 10. End-to-end error v.s. trajectory length on (left) 1000 MDP trajectories from the gridworld dataset and (right) 750 Markov chain trajectories from the Last.fm dataset, comparing various initializations of the soft and the hard EM algorithm. Results averaged over 30 trials, with 30 random initializations for randomly-initialized EM within each trial.

B. Discussion on Using Random Projections

We note that those familiar with the intuition behind the Johnson-Lindenstrauss lemma would guess that a projection to a random n-dimensional subspace for low n would preserve distances with good accuracy. However, note that the bound on the dimension n needed to preserve distances between our N_{clust} estimators up to a multiplicative distortion of $1 \pm \epsilon$ is $\frac{\log(N_{clust})}{\epsilon^2}$. This bound is known to be tight, see for example Larsen & Nelson (2017). Upon thought, this shows that to get good distortion bounds (which will contribute to the deviation between distance estimates and the thresholds), we need a large dimension, interpreted as being affected by the $1/\epsilon^2$. In fact, as soon as $\log(N_{clust})$ exceeds 1, we will need a dimension of order $1/\Delta^2$, while K can be arbitrarily small compared to this.

In the gridworld case, $K = 2$, and we see that we don't get good performance using a random subspace until we hit dimension 50, where the maximum dimension is $S = 64$. Clearly, the $1/\epsilon^2$ term in the Johnson-Lindenstrauss lemma drastically affects the performance of using random subspaces. Using a random subspace of dimension 50 for $S = 64$ is much closer to not projecting at all than to using a subspace of dimension 2.

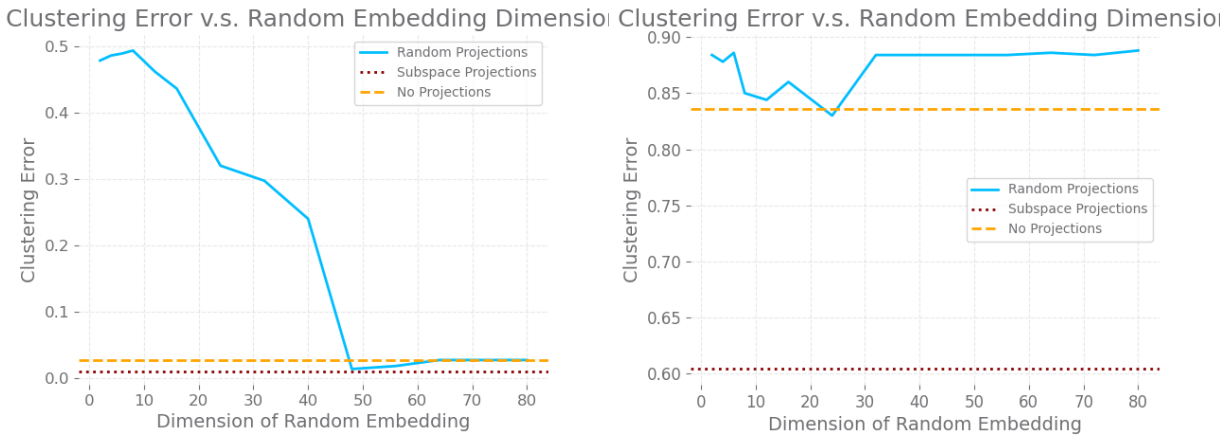

Figure 11. Clustering error using random projections of varying dimension for a trajectory length of 100, benchmarked against the performance of the "with subspace" and "without subspace" versions. The gridworld MDP dataset is on the left, while the Last.fm Markov chain dataset is on the right.

C. Details of the EM Algorithm

We describe the E and M steps for hard EM below first, for simplicity.

M-step: Given the cluster labels, we can estimate each model with the MLE as:

$$\hat{\mathbb{P}}_k(s'|s,a) \leftarrow \frac{\sum_{n\in\mathcal{N}_{clust}} \mathbb{1}_{n\in\mathcal{C}_k} N(n,s,a,s')}{\sum_{n\in\mathcal{N}_{clust}} \mathbb{1}_{n\in\mathcal{C}_k} N(n,s,a)}$$
$$\hat{f}_k \leftarrow \frac{\sum_{n\in\mathcal{N}_{clust}} \mathbb{1}_{n\in\mathcal{C}_k}}{N_{clust}} = \frac{|\mathcal{C}_k|}{N_{clust}}$$

Readers can convince themselves that this is truly the MLE estimate by making the following observation. We can write the log-likelihood of the predicted clusters $\mathcal{C}_k$ and estimated models as $\sum_{k=1}^{K}\sum_{n\in\mathcal{N}_{clust}} \mathbb{1}_{n\in\mathcal{C}_k}\ell(\hat{\mathbb{P}}_k, \hat{f}_k, n)$, where $\ell(\hat{\mathbb{P}}_k, \hat{f}_k, n) = \log\left(f_k \prod_{s,s',a}(\hat{\mathbb{P}}_k(s' \mid s,a))^{N(n,s,a,s')}\right)$. The rest of the derivation mimics the well-known and straightforward computation for Markov chains, using Lagrange multipliers to constrain the estimates to probability distributions.

E-step: On new or unseen data, assign cluster membership according to the following rule:

$$k_m \leftarrow \text{argmax}_k\, \ell(\hat{\mathbb{P}}_k, \hat{f}_k, m) + \log(\hat{f}_i) \tag{1}$$

where $\ell(\hat{\mathbb{P}}_k, m)$ is as above.

Note that for soft EM, we can replace every occurrence of $\mathbb{1}_{n\in\mathcal{C}_k}$ in the M-step with $p_n(k)$, where $p_n(\cdot)$ is the posterior for trajectory n having label k, which is constantly updated during soft EM. For the E-step, we replace the argmax computation by a computation of $p_n(k) = \mathbb{P}(k_n = k \mid \hat{\mathbb{P}}_k, \hat{f}_k, 1 \le k \le K)$. Intuitively described, in hard EM, we recompute the values of $\mathbb{1}_{n\in\mathcal{C}_k}$ using the argmax during the E-step, while in soft EM, we recompute the values of $p_n(k)$.

D. The Classification Algorithm

Note that we define a new quantity, $\hat{f}_{k,s,a}$, which is the proportion of trajectories with label k among all trajectories in $\mathcal{N}_{clust}$ where s, a is observed. Quantities $N(n, s, a), \mathbf{N}(n, i, s, a, \cdot)$ and $N(n, i, s, a)$ carry their usual meanings with respect to either $\mathcal{N}_{clust}$ until step 5 and with respect to $\mathcal{N}_{class}$ after that.

Algorithm 3 Classification

Input: Clusters $\mathcal{C}_k \subset \mathcal{N}_{clust}$, models $\hat{\mathbb{P}}_k(\cdot \mid s, a)$ estimated from $\mathcal{C}_k$, and a set $\mathcal{N}_{class}$ of trajectories to classify.
Compute $\hat{f}_{k,s,a}$ for all k, s, a.
Compute $\tilde{\mathbf{M}}_{s,a} = \sum_{k=1}^{K} \hat{f}_{k,s,a} \hat{\mathbb{P}}_k(\cdot|s,a) \hat{\mathbb{P}}_k(\cdot|s,a)^T$ and store the orthogonal projector $\tilde{\mathbf{V}}_{s,a}^T$ to its top-K eigenspace, for each (s, a).
Compute $\hat{\mathbf{d}}_k = \frac{1}{|\mathcal{C}_k|} \sum_{n \in \mathcal{C}_k} \frac{N(n,s,a)}{G}$ for all k.
Compute $\tilde{D} = \sum_{k=1}^{K} \hat{\mathbf{d}}_k \hat{\mathbf{d}}_k^T$ and store the orthogonal projector $\tilde{\mathbf{U}}^T$ to its top-K eigenspace.
Compute the set SA_β by picking (s, a) pairs with occurrence more than β
$\mathbf{d}_{n,1}, \mathbf{d}_{n,2} \leftarrow \mathbf{0} \in \mathbb{R}^{SA}$
for $(i, s, a) \in \{1, 2\} \times S \times A$ **do**
  Compute $\mathbf{N}(n, i, s, a, \cdot)$, $N(n, i, s, a), \quad \forall n$
  $\hat{\mathbb{P}}_{n,i}(\cdot|s,a) \leftarrow \frac{\mathbf{N}(n,i,s,a,\cdot)}{N(n,i,s,a)} \mathbb{1}_{N(n,i,s,a) \neq 0}, \quad \forall n$
  $[\hat{\mathbf{d}}_{n,i}]_{s,a} \leftarrow \frac{N(n,i,s,a)}{G}, \quad \forall n$
end for
for $(n, k) \in \mathcal{N}_{clust} \times \{1, 2, \ldots K\}$ **do**
  for $(i, s, a) \in \{1, 2\} \times S \times A$ **do**
    $\hat{\mathbf{\Delta}}_{i,s,a} := (\hat{\mathbb{P}}_{n,i}(\cdot \mid s, a) - \hat{\mathbb{P}}_k(\cdot|s,a)) \tilde{\mathbf{V}}_{s,a}^T$
  end for
  $\mathrm{dist}_1(n, k) := \max_{s,a} \hat{\mathbf{\Delta}}_{1,s,a}^T \hat{\mathbf{\Delta}}_{2,s,a}$
  $\mathrm{dist}_2(n, k) := (\hat{\mathbf{d}}_{n,1} - \hat{\mathbf{d}}_k)^T \mathbf{U}\mathbf{U}^T (\hat{\mathbf{d}}_{n,2} - \hat{\mathbf{d}}_k)$
  $\mathrm{dist}(n, k) := \lambda \, \mathrm{dist}_1(n, k) + (1 - \lambda) \, \mathrm{dist}_2(n, k)$
end for
Assign $k_n \leftarrow \mathrm{argmin}_k \, \mathrm{dist}(n, k)$ for each n.

E. Proof of Theorem 2

E.1. Proof of the theorem

We recall the theorem here.

Theorem 2 (Subspace Estimation Guarantee)**.** *Consider* 2 *models with labels* k_1, k_2 *and a state-action pair* s, a *with* $d_{min}(s,a) \geq \alpha/3$*. Consider the output* $\boldsymbol{V}_{s,a}^T$ *of Algorithm 1. Let* $f_{min} = \min(f_{k_1}, f_{k_2})$ *be the lower of the label prevalences. Remember that each trajectory has length* T_n*.*

Then given that $N_{sub} = \Omega\left(\frac{\log(1/\delta)}{\alpha^2}\right)$*,* $T_n = \Omega(t_{mix}\log^4(1/\alpha))$*, with probability at least* $1-\delta$*, for* $k = k_1, k_2$

$$\|\mathbb{P}_k(\cdot \mid s,a) - \boldsymbol{V}_{s,a}\boldsymbol{V}_{s,a}^T\mathbb{P}_k(\cdot \mid s,a)\|_2 \leq \epsilon_{sub}(\delta)$$

where

- *For* $T_n = \Omega\left(t_{mix}\log^3\left(\frac{f_{min}N_{sub}\alpha}{KS\log(1/\delta)}\right)\right)$

$$\epsilon_{sub}(\delta) = O\left(\sqrt{\frac{K}{f_{min}}\left(\sqrt{\frac{S}{N_{sub}\cdot\alpha^3}\log\left(\frac{1}{\delta}\right)}\right)}\right)$$

- *While for* $T_n = O\left(t_{mix}\log^3\left(\frac{f_{min}N_{sub}\alpha}{KS\log(1/\delta)}\right)\right)$

$$\epsilon_{sub}(\delta) = O\left(\left(\frac{1}{2}\right)^{\frac{1}{16}\left(\frac{T_n}{t_{mix}}\right)^{1/3}}\right)$$

Alternatively, we only need $N_{sub} = \Omega\left(\frac{K^2S\log(1/\delta)}{f_{min}^2\alpha^3\epsilon^4}\right)$ *and* $T_n = \Omega\left(t_{mix}\log^3(1/\epsilon)\log^4(1/\alpha)\right)$ *trajectories for* ϵ *accuracy in subspace estimation with probability at least* $1-\delta$*.*

Remark 5. We can convert the α^3 in the denominator to an α at the cost of making T_n more heavily dependent on α (more than just $\log(1/\alpha)$). Intuitively, α accounts for the probability of not observing s, a, so this is just saying that we can shift the onus for that from the number of trajectories to their length. We chose not to do that since we are trying to minimize the length of trajectories needed, and assume that we have access to many trajectories.

Proof. The main input is the proposition below, proved in the next section.

Proposition 1. *Consider* $L < K$ *models with labels* j_l*,* $1 \leq l \leq L$*, with* $d_{min}(s,a) := \min_l d_{j_l}(s,a)$*. Consider the output* $\boldsymbol{V}_{s,a}^T$ *of Algorithm 1. Let* $f_{min} = \min_l f_{j_l}$ *be the minimum frequency across these models in the mixture. Remember that each trajectory has length* T_n*. Then we have the guarantee that with probability at least* $1-\delta$

$$\|\mathbb{P}_j(\cdot \mid s,a) - \boldsymbol{V}_{s,a}\boldsymbol{V}_{s,a}^T\mathbb{P}_j(\cdot \mid s,a)\|_2$$

is bounded above by

$$\sqrt{\frac{4K}{f_{min}d_{min}(s,a)}\left(\sqrt{\frac{128}{N_{sub}\cdot d_{min}(s,a)}(2S\log(12)+\log(4/\delta))} + \left(\frac{1}{2}\right)^{\frac{T_n}{8Gt_{mix}}}\right)}$$

for all $j \in \{j_l \mid 1 \leq l \leq L\}$*, when* $N_{sub} \geq \frac{32}{d_{min}(s,a)^2}\log\left(\frac{1}{\delta}\right)$ *and* $\frac{T_n}{8t_{mix}} > \frac{G\log(48G/d_{min}(s,a))}{\log 2}$*.*

For a state-action pair with $d_{min}(s,a) \geq \alpha/3$, the conditions simplify to $N_{sub} \geq \Omega\left(\frac{\log(1/\delta)}{\alpha^2}\right)$ and $T_n \geq \Omega(Gt_{mix}\log(G/\alpha))$. We set $G = \left(\frac{T_n}{t_{mix}}\right)^{\frac{2}{3}}$ to get bounds that only depend on T_n. Note that this means a sufficient

condition on T_n is $T_n \geq \Omega(t_{mix} \log^4(1/\alpha))$ (one can show this with an elementary computation). Also note that

$$\sqrt{\frac{S + \log(1/\delta)}{N_{sub} \cdot \alpha}} \leq \sqrt{\frac{S \log(1/\delta)}{N_{sub} \cdot \alpha}}$$

Then with probability at least $1 - \delta$, the following bound holds for any label $j = j_l$ for some l.

$$\|\mathbb{P}_j(\cdot \mid s, a) - \mathbf{V}_{s,a}\mathbf{V}_{s,a}^T \mathbb{P}_j(\cdot \mid s, a)\|_2 \leq O\left(\sqrt{\frac{K}{f_{min}\alpha}\left(\sqrt{\frac{S\log(1/\delta)}{N_{sub} \cdot \alpha}} + \left(\frac{1}{2}\right)^{\frac{1}{8}\left(\frac{T_n}{t_{mix}}\right)^{1/3}}\right)}\right)$$

So, there is a universal constant C_2 so that for $T_n > C_2 t_{mix} \log^3\left(\frac{f_{min} N_{sub}\alpha}{KS\log(1/\delta)}\right)$,

$$\left(\frac{1}{2}\right)^{\frac{1}{8}\left(\frac{T_n}{t_{mix}}\right)^{1/3}} \leq C' \frac{K}{f_{min}}\left(\sqrt{\frac{S}{N_{sub} \cdot \alpha^3}\log\left(\frac{1}{\delta}\right)}\right)$$

While for $T_n = O\left(t_{mix} \log^3\left(\frac{f_{min} N_{sub}\alpha}{KS\log(1/\delta)}\right)\right)$,

$$\frac{K}{f_{min}}\left(\sqrt{\frac{S}{N_{sub} \cdot \alpha^3}\log\left(\frac{1}{\delta}\right)}\right) \leq O\left(\left(\frac{1}{2}\right)^{\frac{1}{8}\left(\frac{T_n}{t_{mix}}\right)^{1/3}}\right)$$

So, combining all these, for $N_{sub} = \Omega\left(\frac{\log(1/\delta)}{\alpha^2}\right)$, $T_n = \Omega(t_{mix} \log^4(1/\alpha))$

- For $T_n = \Omega\left(t_{mix} \log^3\left(\frac{f_{min} N_{sub}\alpha}{KS\log(1/\delta)}\right)\right)$

$$\epsilon_{sub}(\delta) = O\left(\sqrt{\frac{K}{f_{min}}\left(\sqrt{\frac{S}{N_{sub} \cdot \alpha^3}\log\left(\frac{1}{\delta}\right)}\right)}\right)$$

- While for $T_n = O\left(t_{mix} \log^3\left(\frac{f_{min} N_{sub}\alpha}{KS\log(1/\delta)}\right)\right)$

$$\epsilon_{sub}(\delta) = O\left(\left(\frac{1}{2}\right)^{\frac{1}{16}\left(\frac{T_n}{t_{mix}}\right)^{1/3}}\right)$$

□

E.2. Proof of the Proposition 1

We recall the proposition here.

Proposition 1. *Consider $L < K$ models with labels j_l, $1 \leq l \leq L$, with $d_{min}(s, a) := \min_l d_{j_l}(s, a)$. Consider the output $\boldsymbol{V}_{s,a}^T$ of Algorithm 1. Let $f_{min} = \min_l f_{j_l}$ be the minimum frequency across these models in the mixture. Remember that each trajectory has length T_n. Then we have the guarantee that with probability at least $1 - \delta$*

$$\|\mathbb{P}_j(\cdot \mid s, a) - \boldsymbol{V}_{s,a}\boldsymbol{V}_{s,a}^T \mathbb{P}_j(\cdot \mid s, a)\|_2$$

is bounded above by

$$\sqrt{\frac{4K}{f_{min}d_{min}(s,a)}\left(\sqrt{\frac{128}{N_{sub}\cdot d_{min}(s,a)}(2S\log(12)+\log(4/\delta))}+\left(\frac{1}{2}\right)^{\frac{T_n}{8Gt_{mix}}}\right)}$$

for all $j \in \{j_l \mid 1 \le l \le L\}$*, when* $N_{sub} \ge \frac{32}{d_{min}(s,a)^2}\log\left(\frac{1}{\delta}\right)$ *and* $\frac{T_n}{8t_{mix}} > \frac{G\log(48G/d_{min}(s,a))}{\log 2}$.

Remark 6. We should point out that we will only need $L = 2$ for subsequent theorems. Also, remember that only s, a with $d_{min}(s,a) > \alpha$ will be relevant in subsequent theorems, with α as in our assumption.

Proof. For brevity of notation, we will denote $c_{n,i} := N(n,i,s,a)$, $\mathbf{w}_{n,i} := N(n,i,s,a,\cdot)$ and suppress the (s,a) dependence. We will first need the following lemma which guarantees that we can get past mixing and concentration hurdles with our estimator, modulo actually observing s, a in both segments.

Lemma 1. *Let* $\mathcal{B}_n$ *be the event given by* $n \in \mathcal{N}_{traj}(s,a)$*, which is the same as* $c_{n,1}c_{n,2} \neq 0$ *and let*

$$\boldsymbol{M}_{s,a} = \sum_{j=1}^{K}\mathbb{P}(k_n = j \mid \mathcal{B}_n)\mathbb{P}_j(\cdot \mid s,a)\mathbb{P}_j(\cdot \mid s,a)^T$$

.

Call our estimator $\hat{\boldsymbol{M}}_{s,a}$*. Then we know that*

$$\hat{\boldsymbol{M}}_{s,a} = \frac{1}{N_{traj}(s,a)}\sum_n \hat{\mathbb{P}}_{n,1}(\cdot \mid s,a)\hat{\mathbb{P}}_{n,2}(\cdot \mid s,a)^T$$

and we have

$$\|\hat{\boldsymbol{M}}_{s,a} - \boldsymbol{M}_{s,a}\| < \sqrt{\frac{32}{N_{traj}(s,a)}(2S\log(12)+\log(\frac{2}{\delta}))} + \frac{48G}{d_{min}(s,a)}\left(\frac{1}{4}\right)^{\frac{T_n}{8Gt_{mix}}}$$

Remark 7. Note that since all trajectories are generated independently of each other and the process that generates them is identical, $\mathbb{P}(k_n = j \cap \mathcal{B}_n)$ is the same for all n. A similar observation holds for many conditional/unconditional probabilities and conditional/unconditional expectations in this proof, and will not be stated again.

Assume the lemma for now. The proof is delayed to after the proof of the theorem. We will combine this lemma with Lemma 3 from Chen & Poor (2022). In the context of their lemma, $p^{(j)} = \mathbb{P}(k_n = j \mid \mathcal{B}_n)$, $\mathbf{y}^{(j)} = \mathbb{P}_j(\cdot \mid s,a)$. Now, we can use the first term on the right-hand side of the bound in Lemma 3 of Chen & Poor (2022) to get that for any $1 \le l \le L$

$$\|\mathbb{P}_{j_l}(\cdot \mid s,a) - \mathbf{V}_{s,a}\mathbf{V}_{s,a}^T\mathbb{P}_{j_l}(\cdot \mid s,a)\|_2 \le \sqrt{\frac{2K}{\min_l(\mathbb{P}(k_n = j_l \mid \mathcal{B}_n))}\|\hat{\mathbf{M}}_{s,a} - \mathbf{M}_{s,a}\|} \tag{2}$$

E.2.1. **Lower Bounding** $\mathbb{P}(k_n = j_l \mid \mathcal{B}_n)$

Note that

$$\mathbb{P}(k_n = j_l \mid \mathcal{B}_n) = \frac{\mathbb{P}(k_n = j_l)\mathbb{P}(\mathcal{B}_n \mid k_n = j_l)}{\mathbb{P}(\mathcal{B}_n)} \ge f_{j_l}\mathbb{P}(\mathcal{B}_n \mid k_n = j_l)$$

So, we need only lower bound $\mathbb{P}(\mathcal{B}_n \mid k_n = j_l)$, for which we will need a lemma. We will use the following crucial lemma several times in our proofs. This is where we use (Yu, 1994)'s blocking technique.

Lemma 2. *Consider a function h on segments of a Markov chain with mixing time $t_{mix} = t_{mix}(\frac{1}{4})$ with $C = \sup h$. Consider the joint distribution χ over the product of the σ-algebras of n such segments, with marginals χ_i. Let the product distribution of the marginals χ_i be called Ξ. Then for $\lambda = (\frac{1}{4})^{\frac{1}{t_{mix}}}$ and for the minimum distance between consecutive segments being a_n, we have*

$$|\mathbb{E}_\chi h - \mathbb{E}_\Xi h| \leq 4C(n-1)\lambda^{a_n}$$

Proof. Remember that each of our Markov processes is mixing, so there exists $t_{mix,j} = t_{mix,j}(\frac{1}{4})$ and a stationary distribution d_j so that $TV(P_j^n(x,\cdot), d_j) < \frac{1}{4}$ for $n \geq t_{mix,j}$. Let $t_{mix} = \max_j t_{mix,j}$. Since the decay in total variation distance is multiplicative, $TV(P_j^n(x,\cdot), d_j) < (\frac{1}{4})^c$ for all j and $n \geq ct_{mix}$. This implies that

$$\max_j TV(P_j^n(x,\cdot), d_j) < \left(\frac{1}{4}\right)^{\frac{T_n}{4t_{mix}}-1} = 4\lambda^n$$

where $\lambda = (\frac{1}{4})^{\frac{1}{t_{mix}}}$

This means that we satisfy the definition of V-geometric ergodicity from Vidyasagar (2010), with V being the constant function with value 1, $\mu = 4$ and λ as above. That means that any of our processes is beta-mixing by (the proof of) Theorem 3.10 from the text and

$$\beta_n \leq \mu\lambda^n = 4\lambda^n$$

we employ an argument analogous to the setup and argument used to prove Lemma 4.1 of Yu (1994), merely with H_i's replaced by the segments of arbitrary length instead of a_n-sized blocks while T_i's stay at a_n sized blocks. Then, Q from Corollary 2.7 is the probability distribution of the segments here, Ω_i from Corollary 2.7 is the real vector space of the same dimension as the length of the ith segment, Σ_i is the product Borel field on this vector space and m in the theorem is the number of segments n here (note that n is called μ_n in Lemma 4.1). $\tilde{Q}$ is the product distribution over the marginals of Q, as in the theorem. Note that $\beta(Q)$ from Corollary 2.7 used in the proof remains less than β_{a_n}. Now we can directly quote Corollary 2.7 to conclude that

$$|\mathbb{E}_\chi h - \mathbb{E}_\Xi h| \leq C(n-1)\beta_{a_n} \leq 4C(n-1)\lambda^{a_n}$$

□

Define

$$h = \mathbb{1}_{(c_{n,1}c_{n,2}=0)}$$

We are now ready to bound $P(\mathcal{B}_n \mid k_n = j) = P(c_{n,1}c_{n,2} = 0 \mid k_n = j)$. Consider the joint distribution over the segments Ω_1 and Ω_2 of a trajectory sampled from hidden label j. Call this χ and let its marginals on Ω_i be χ_i. Let the product distribution of its marginals be $\Xi := \chi_1 \times \chi_2$. Notice that then

$$\mathbb{E}_\Xi h = P(c_{n,1} = 0 \mid k_n = j)P(c_{n,2} = 0 \mid k_n = j)$$

by definition of Ξ. Also, clearly we have

$$\mathbb{E}_\chi h = P(c_{n,1}c_{n,2} = 0 \mid k_n = j)$$

Now, using Lemma 2, we get that for $C = \sup h = 1$ and $n = 2$, we have the following inequality.

$$|P(c_{n,1}c_{n,2} = 0 \mid k_n = j) - P(c_{n,1} = 0 \mid k_n = j)P(c_{n,2} = 0 \mid k_n = j)| = |\mathbb{E}_\chi h - \mathbb{E}_\Xi h| \leq 4\lambda^{\frac{T_n}{4}} \tag{3}$$

Additionally, for $i = 1, 2$, if $d_{t,j}(s,a)$ is the distribution at time t, the following is obtained by the definition of mixing times.

$$\begin{aligned}\mathbb{P}(c_{n,i} = 0 \mid k_n = j) &\leq (1 - d_{(2i-1)T,j}(s,a)) \\ &\leq (1 - d_j(s,a) + TV(d_{(2i-1)T,j}, \pi))\end{aligned}$$

$$\leq (1 - d_{min}(s,a) + 4\lambda^{\frac{T_n}{4}})$$
$$\leq \left(1 - \frac{d_{min}(s,a)}{2}\right) \tag{4}$$

where the last inequality holds for $T_n > 4t_{mix}\frac{\log(8/d_{min}(s,a))}{\log 4}$. This allows us to use inequality 3 and

$$\begin{aligned}
P(c_{n,1}c_{n,2} = 0 \mid k_n = j) &\leq 4\lambda^{\frac{T_n}{4}} + P(c_{n,1} = 0 \mid k_n = j)P(c_{n,2} = 0 \mid k_n = j) \\
&\leq 4\lambda^{\frac{T_n}{4}} + \left(1 - \frac{d_{min}(s,a)}{2}\right)^2 \\
&\leq 1 - d_{min}(s,a) + \frac{d_{min}(s,a)^2}{4} + 4\lambda^{\frac{T_n}{4}} \\
&\leq 1 - d_{min}(s,a) + \frac{d_{min}(s,a)}{4} + 4\lambda^{\frac{T_n}{4}} \\
&\leq 1 - \frac{d_{min}(s,a)}{2}
\end{aligned} \tag{5}$$

where the last inequality holds for $T_n > 4t_{mix}\frac{\log(16/d_{min}(s,a))}{\log 4}$. We conclude that for $T_n > 4t_{mix}\frac{\log(16/d_{min}(s,a))}{\log 4}$, and $j = j_l$ for some l,

$$P(\mathcal{B}_n \mid k_n = j) \geq \frac{d_{min}(s,a)}{2}$$

And so,

$$\begin{aligned}
\min_l f_{j_l}(\mathbb{P}(k_n = j_l \mid \mathcal{B}_n)) &\geq \min_l f_{j_l}(\mathbb{P}(k_n = j_l \cap \mathcal{B}_n)) \\
&\geq \min_l f_{j_l}(\mathbb{P}(\mathcal{B}_n \mid k_n = j)\mathbb{P}(k_n = j)) \\
&\geq \frac{f_{min}d_{min}(s,a)}{2}
\end{aligned}$$

We can thus conclude that for $T_n > 4t_{mix}\frac{\log(16/d_{min}(s,a))}{\log 4}$,

$$\|\mathbb{P}_{j_l}(\cdot \mid s,a) - \mathbf{V}_{s,a}\mathbf{V}_{s,a}^T\mathbb{P}_{j_l}(\cdot \mid s,a)\|_2 \leq \sqrt{\frac{4K}{f_{min}d_{min}(s,a)}\|\hat{\mathbf{M}}_{s,a} - \mathbf{M}_{s,a}\|} \tag{6}$$

E.2.2. ABSORBING THE EXTRA TERMS INTO THE EXPONENT OF $1/4$

Now remember from Lemma 1 that

$$\|\hat{\mathbf{M}}_{s,a} - \mathbf{M}_{s,a}\| < \sqrt{\frac{32}{N_{traj}(s,a)}(2S\log(12) + \log(\frac{2}{\delta}))} + \frac{48G}{d_{min}(s,a)}\left(\frac{1}{4}\right)^{\frac{T_n}{8Gt_{mix}}}$$

Notice that for $\frac{T_n}{8t_{mix}} > \frac{G\log(48G/d_{min}(s,a))}{\log 2} > \frac{\log(16/d_{min}(s,a))}{2\log 4}$, we have that

$$\begin{aligned}
\frac{48G}{d_{min}(s,a)}\left(\frac{1}{4}\right)^{\frac{T_n}{8Gt_{mix}}} &= \frac{48G}{d_{min}(s,a)}\left(\frac{1}{4}\right)^{\frac{T_n}{16Gt_{mix}}}\left(\frac{1}{4}\right)^{\frac{T_n}{16Gt_{mix}}} \\
&= \frac{48G}{d_{min}(s,a)}\left(\frac{1}{2}\right)^{\frac{T_n}{8Gt_{mix}}}\left(\frac{1}{2}\right)^{\frac{T_n}{8Gt_{mix}}} \\
&\leq \left(\frac{1}{2}\right)^{\frac{T_n}{8Gt_{mix}}}
\end{aligned}$$

E.2.3. BOUNDING THE CONCENTRATION TERM

We finally need to bound $N_{traj}(s,a)$ from below to bound the first term in this sum. Note that $\mathbb{E}[N_{traj}(s,a)] \geq N_{sub}(1 - P(c_{n,1}c_{n,2} = 0)) \geq N_{sub}\frac{d_{min}(s,a)}{2}$ from equation 5 above. Now, by Hoeffding's inequality, we have

$$\begin{aligned}
\mathbb{P}\left(N_{traj}(s,a) < N_{sub}\frac{d_{min}(s,a)}{4}\right) &= \mathbb{P}\left(N_{traj}(s,a) < N_{sub}\frac{d_{min}(s,a)}{2} - N_{sub}\frac{d_{min}(s,a)}{4}\right) \\
&\leq \mathbb{P}\left(N_{traj}(s,a) < \mathbb{E}[N_{traj}(s,a)] - N_{sub}\frac{d_{min}(s,a)}{4}\right) \\
&= \mathbb{P}\left(\sum_{n\in\mathcal{N}_{sub}} \mathbb{1}_{c_{n,1}c_{n,2}\neq 0} < N_{sub}\mathbb{E}[\mathbb{1}_{c_{n,1}c_{n,2}\neq 0}] - N_{sub}\frac{d_{min}(s,a)}{4}\right) \\
&\leq \exp\left(-\frac{d_{min}(s,a)^2 N_{sub}}{8}\right)
\end{aligned}$$

This is less than δ for $N_{sub} \geq \frac{8}{d_{min}(s,a)^2}\log\left(\frac{1}{\delta}\right)$.

Combining this with equation 6 and splitting the two δ, we have our result that

$$\|\mathbb{P}_j(\cdot \mid s,a) - \mathbf{V}_{s,a}\mathbf{V}_{s,a}^T\mathbb{P}_j(\cdot \mid s,a)\|_2$$

is bounded above by

$$\sqrt{\frac{4K}{f_{min}d_{min}(s,a)}\left(\sqrt{\frac{128}{N_{sub}\cdot d_{min}(s,a)}(2S\log(12) + \log(4/\delta))} + \left(\frac{1}{2}\right)^{\frac{T_n}{8Gt_{mix}}}\right)}$$

for $N_{sub} \geq \frac{8}{d_{min}(s,a)^2}\log\left(\frac{1}{\delta}\right)$ and $\frac{T_n}{8t_{mix}} > \frac{G\log(48G/d_{min}(s,a))}{\log 2}$. □

E.3. Proof of Lemma 1

We recall Lemma 1.

Lemma 1. *Let $\mathcal{B}_n$ be the event given by $n \in \mathcal{N}_{traj}(s,a)$, which is the same as $c_{n,1}c_{n,2} \neq 0$ and let*

$$\boldsymbol{M}_{s,a} = \sum_{j=1}^{K}\mathbb{P}(k_n = j \mid \mathcal{B}_n)\mathbb{P}_j(\cdot \mid s,a)\mathbb{P}_j(\cdot \mid s,a)^T$$

.

Call our estimator $\hat{\boldsymbol{M}}_{s,a}$. Then we know that

$$\hat{\boldsymbol{M}}_{s,a} = \frac{1}{N_{traj}(s,a)}\sum_n \hat{\mathbb{P}}_{n,1}(\cdot \mid s,a)\hat{\mathbb{P}}_{n,2}(\cdot \mid s,a)^T$$

and we have

$$\|\hat{\boldsymbol{M}}_{s,a} - \boldsymbol{M}_{s,a}\| < \sqrt{\frac{32}{N_{traj}(s,a)}(2S\log(12) + \log(\frac{2}{\delta}))} + \frac{48G}{d_{min}(s,a)}\left(\frac{1}{4}\right)^{\frac{T_n}{8Gt_{mix}}}$$

Proof. We divide the proof into subsections. We first remind ourselves that the estimator $\hat{\mathbf{M}}_{s,a}$ is given by the matrix

$$\hat{\mathbf{M}}_{s,a} = \frac{1}{N_{traj}(s,a)}\sum_{n\in\mathcal{N}_{traj}(s,a)}\left(\hat{\mathbb{P}}_{n,1}(\cdot|s,a)\hat{\mathbb{P}}_{n,2}(\cdot|s,a)^T\right)$$

E.3.1. ESTIMATING $\mathbb{E}[\hat{\mathbf{M}}_{s,a}]$

We will split the expectation into the desired term and the error coming from correlation between the two segments Ω_1 and Ω_2. Remember that for brevity of notation, let $c_{n,i} := N(n,i,s,a)$, $\mathbf{w}_{n,i} := N(n,i,s,a,\cdot)$. Call the estimate from each trajectory a random variable $\hat{\mathbf{M}}_{n,s,a}$, that is

$$\hat{\mathbf{M}}_{n,s,a} = \frac{\mathbf{w}_{n,1}\mathbf{w}_{n,2}^T}{c_{n,1}c_{n,2}}$$

Now

$$\hat{\mathbf{M}}_{s,a} = \sum_{n\in\mathcal{N}_{traj}(s,a)} \frac{1}{N_{traj}(s,a)}\hat{\mathbf{M}}_{n,s,a}$$

Remember that

$$\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) := \frac{\mathbf{w}_{n,i}}{c_{n,i}}\mathbb{1}_{c_{n,i}\neq 0}$$

Let k_n be the hidden label for trajectory n, as usual. Define the event $\mathcal{B}_n$ to be $n \in \mathcal{N}_{traj}(s,a)$, which is the same as $c_{n,1}c_{n,2} \neq 0$. We establish the following equality, essentially just defining the quantity $\mathrm{Mix}(j)$.

$$\begin{aligned}
\mathbb{E}[\hat{\mathbf{M}}_{n,s,a} \mid \mathcal{B}_n] &= \mathbb{E}\left[\frac{\mathbf{w}_{n,1}\mathbf{w}_{n,2}^T}{c_{n,1}c_{n,2}}\middle|\mathcal{B}_n\right] \\
&= \sum_{j=1}^{K}\mathbb{P}(k_n = j \mid \mathcal{B}_n)\mathbb{E}\left[\frac{\mathbf{w}_{n,1}\mathbf{w}_{n,2}^T}{c_{n,1}c_{n,2}}\middle| k_n = j, \mathcal{B}_n\right] \\
&= \sum_{j=1}^{K}\mathbb{P}(k_n = j \mid \mathcal{B}_n)\mathbb{P}_j(\cdot \mid s,a)\mathbb{P}_j(\cdot \mid s,a)^T + \sum_{j=1}^{K}\mathbb{P}(k_n = j \mid \mathcal{B}_n)\,\mathrm{Mix}(j) \qquad (7)
\end{aligned}$$

where $\mathrm{Mix}(j) = \mathbb{E}\left[\frac{\mathbf{w}_{n,1}\mathbf{w}_{n,2}^T}{c_{n,1}c_{n,2}}\middle| k_n = j, \mathcal{B}_n\right] - \mathbb{P}_j(\cdot \mid s,a)\mathbb{P}_j(\cdot \mid s,a)^T$. Notice that this has connotations of covariance. Now note the following chain of equations.

$$\begin{aligned}
\mathbb{E}[\hat{\mathbf{M}}_{s,a} \mid \mathcal{N}_{traj}(s,a)] &= \mathbb{E}\left[\sum_{n\in\mathcal{N}_{traj}(s,a)}\frac{1}{N_{traj}(s,a)}\hat{\mathbf{M}}_{n,s,a}\middle|\mathcal{N}_{traj}(s,a)\right] \\
&= \sum_{n\in\mathcal{N}_{traj}(s,a)}\frac{1}{N_{traj}(s,a)}\mathbb{E}\left[\hat{\mathbf{M}}_{n,s,a}\middle|\mathcal{N}_{traj}(s,a)\right] \\
&= \sum_{n\in\mathcal{N}_{traj}(s,a)}\frac{1}{N_{traj}(s,a)}\mathbb{E}\left[\hat{\mathbf{M}}_{n,s,a}\middle|\mathbb{1}_{\mathcal{B}_1}, \mathbb{1}_{\mathcal{B}_2}, \ldots \mathbb{1}_{\mathcal{B}_{N_{sub}}}\right] \\
&= \sum_{n\in\mathcal{N}_{traj}(s,a)}\frac{1}{N_{traj}(s,a)}\mathbb{E}\left[\hat{\mathbf{M}}_{n,s,a}\middle|\mathcal{B}_n\right] \\
&= \mathbb{E}\left[\hat{\mathbf{M}}_{n,s,a}\middle|\mathcal{B}_n\right] \\
&= \sum_{j=1}^{K}\mathbb{P}(k_n = j \mid \mathcal{B}_n)\mathbb{P}_j(\cdot \mid s,a)\mathbb{P}_j(\cdot \mid s,a)^T + \sum_{j=1}^{K}\mathbb{P}(k_n = j \mid \mathcal{B}_n)\,\mathrm{Mix}(j) \\
&= \mathbf{M}_{s,a} + \sum_{j=1}^{K}\mathbb{P}(k_n = j \mid \mathcal{B}_n)\,\mathrm{Mix}(j) \qquad (8)
\end{aligned}$$

Here, the third equality is because the set $\mathcal{N}_{traj}(s,a)$ is exactly described by the indicators listed, and they generate the same σ-algebra, The fourth equality holds since all trajectories are independent and so conditioning on events in other trajectories doesn't affect the expectation of $\hat{\mathbf{M}}_{n,s,a}$. The fifth equality is because $\mathbb{E}[\hat{\mathbf{M}}_{n,s,a} \mid \mathcal{B}_n]$ is the same for all n as determined above (in fact, we have shown that it is a constant random variable).

E.3.2. SETUP FOR THE MAIN BOUND

We have that

$$\mathbf{M}_{s,a} = \sum_{j=1}^{K} \mathbb{P}(k_n = j \mid \mathcal{B}_n)\mathbb{P}_j(\cdot \mid s,a)\mathbb{P}_j(\cdot \mid s,a)^T$$

By equation 8,

$$\begin{aligned}
\|\hat{\mathbf{M}}_{s,a} - \mathbf{M}_{s,a}\| &\leq \|\hat{\mathbf{M}}_{s,a} - \mathbb{E}[\hat{\mathbf{M}}_{s,a} \mid \mathcal{N}_{traj}(s,a)]\| + \sum_{j=1}^{K} \mathbb{P}(k_n = j \mid \mathcal{B}_n)\, \|\mathrm{Mix}(j)\| \\
&\leq \|\hat{\mathbf{M}}_{s,a} - \mathbb{E}[\hat{\mathbf{M}}_{s,a} \mid \mathcal{N}_{traj}(s,a)]\| + \left(\sum_{j=1}^{K} \mathbb{P}(k_n = j \mid \mathcal{B}_n)\right) \max_j \|\mathrm{Mix}(j)\| \\
&= \|\hat{\mathbf{M}}_{s,a} - \mathbb{E}[\hat{\mathbf{M}}_{s,a} \mid \mathcal{N}_{traj}(s,a)]\| + \max_j \|\mathrm{Mix}(j)\|
\end{aligned}$$

The first term represents the error in concentration across trajectories and the second term represents the correlation between the two segments Ω_1 and Ω_2 in the same trajectory. We bound the first using a covering argument and use Bin Yu's work to bound the other.

E.3.3. COVERING ARGUMENT TO BOUND $\hat{\mathbf{M}}_{s,a} - \mathbb{E}[\hat{\mathbf{M}}_{s,a}]$

We will need this conditional version of Hoeffding's inequality for this section. Note that this is not quite the Azuma-Hoeffding inequality with a constant filtration due to the conditional probability involved, as well as due to the conditional independence needed.

Lemma 3. *Consider a σ-algebra $\mathcal{F}$ and let $A_i \leq B_i$ be random variables measurable over it. If random variables X_i are almost surely bounded in $[A_i, B_i]$ and are conditionally independent over some σ-algebra $\mathcal{F}$, then the following inequalities hold for $S_n = \sum_{i=1}^{n} X_i$*

$$\begin{aligned}
\mathbb{P}\left(S_n - \mathbb{E}[S_n \mid \mathcal{F}] > \epsilon | \mathcal{F}\right) &\leq \exp\left(-\frac{2\epsilon}{\sum_i (B_i - A_i)^2}\right) \\
\mathbb{P}\left(S_n - \mathbb{E}[S_n \mid \mathcal{F}] < -\epsilon | \mathcal{F}\right) &\leq \exp\left(-\frac{2\epsilon^2}{\sum_i (B_i - A_i)^2}\right)
\end{aligned}$$

Proof. The proof is essentially a repeat of one of the standard proofs of Hoeffding's inequality. Note that we have the conditional Markov inequality $\mathbb{P}(X \geq a \mid \mathcal{F}) \leq \frac{1}{a}\mathbb{E}[X \geq a \mid \mathcal{F}]$, shown exactly the way Markov's inequality is shown. Now, we have the following chain of inequalities.

$$\begin{aligned}
\mathbb{P}((S_n - \mathbb{E}[S_n \mid \mathcal{F}] > \epsilon | \mathcal{F}) &= e^{-s\epsilon}\mathbb{E}[e^{S_n - \mathbb{E}[S_n|\mathcal{F}]} \mid \mathcal{F}] \\
&= e^{-s\epsilon}\prod_{i=1}^{n} \mathbb{E}[e^{X_i - \mathbb{E}[X_i|\mathcal{F}]} \mid \mathcal{F}]
\end{aligned}$$

We now show a conditional expectation version of Hoeffding's lemma by repeating the steps for a standard proof to show that $\mathbb{E}[e^{X - \mathbb{E}[X|\mathcal{F}]} \mid \mathcal{F}] \leq \frac{\lambda^2 (B-A)^2}{8}$ for random variables $A \leq B$ measurable over $\mathcal{F}$ and $X \in [A, B]$ almost surely. Note that by convexity of $e^{\lambda x}$, we have the following for $x \in [A, B]$ at any value of A and B.

$$e^{\lambda x} \leq \frac{B - x}{B - A} e^{\lambda A} + \frac{x - A}{B - A} e^{\lambda B}$$

WLOG, we can replace X by $X - \mathbb{E}[X \mid \mathcal{F}]$ and assume $\mathbb{E}[X \mid \mathcal{F}] = 0$. In that case, we note the following inequality, where we define for any fixed value of A and B the function $L(y) := \frac{yA}{B-A} + \log\left(1 + \frac{A - e^y B}{B - A}\right)$.

$$\mathbb{E}[e^{\lambda X} \mid \mathcal{F}] \leq \frac{B - \mathbb{E}[X \mid \mathcal{F}]}{B - A} e^{\lambda A} + \frac{\mathbb{E}[X \mid \mathcal{F}] - A}{B - A} e^{\lambda B}$$

$$= \frac{B}{B-A}e^{\lambda A} + \frac{-A}{B-A}e^{\lambda B}$$
$$= e^{L(\lambda(B-A))}$$

Basic computations involving Taylor's theorem from a standard proof of Hoeffding's inequality show that $L(y) \leq \frac{y^2}{8}$ for any value of A, B. This gives us the condition version of Hoeffding's lemma, $\mathbb{E}[e^{X-\mathbb{E}[X|\mathcal{F}]} \mid \mathcal{F}] \leq \frac{\lambda^2(B-A)^2}{8}$. This allows us to establish the following chain of inequalities.

$$\begin{aligned}
\mathbb{P}((S_n - \mathbb{E}[S_n \mid \mathcal{F}] > \epsilon|\mathcal{F}) &= e^{-s\epsilon}\prod_{i=1}^{n}\mathbb{E}[e^{X_i-\mathbb{E}[X_i|\mathcal{F}]} \mid \mathcal{F}] \\
&\leq \exp(-s\epsilon)\prod_{i=1}^{n}\exp\left(\frac{s^2(B_i-A_i)^2}{8}\right) \\
&= \exp\left(-s\epsilon + \sum_{i=1}^{n}\frac{s^2(B_i-A_i)^2}{8}\right)
\end{aligned}$$

Since s is arbitrary, we can pick $s = \frac{4\epsilon}{\sum_i (B_i-A_i)^2}$ above to get an upper bound of $\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (B_i-A_i)^2}\right)$. The other inequality is proved analogously. □

We now show that the first term from the previous section concentrates. Pick $\mathbf{u}, \mathbf{v} \in \mathbb{S}^{S-1}$, that is they lie in the unit Euclidean norm sphere in $\mathbb{R}^S$. We need only bound this term when $N_{traj}(s,a) \neq 0$, as otherwise the lemma holds vacuously. Note that

$$\hat{\mathbf{M}}_{s,a} - \mathbb{E}[\hat{\mathbf{M}}_{s,a} \mid \mathcal{N}_{traj}(s,a)] = \sum_{n\in\mathcal{N}_{traj}(s,a)} \frac{\hat{\mathbf{M}}_{n,s,a} - \mathbb{E}[\hat{\mathbf{M}}_{n,s,a} \mid \mathcal{N}_{traj}(s,a)]}{N_{traj}(s,a)}$$

Now we set up our covering argument. Consider a covering of $\mathbb{S}^{S-1}$ by balls of radius $\frac{1}{4}$. We will need at most 12^S such balls and if C is the set of their centers, then for any matrix X, the following holds in regard to its norm.

$$\|X\| = \sup_{\mathbf{u},\mathbf{v}\in\mathbb{S}^{S-1}} |\mathbf{u}^T X\mathbf{v}| \leq 2\max_{\mathbf{u},\mathbf{v}\in C} |\mathbf{u}^T X\mathbf{v}| \leq 2\|X\| \tag{9}$$

For any pair **u,v** $\in C$, note that

$$\begin{aligned}
|\mathbf{u}^T\hat{\mathbf{M}}_{n,s,a}\mathbf{v}| &= \left|\mathbf{u}^T\frac{\mathbf{w}_{n,1}}{c_{n,1}}\right|\left|\frac{\mathbf{w}_{n,2}^T}{c_{n,2}}\mathbf{v}\right|\mathbb{1}_{c_{n,1}c_{n,2}\neq 0} \\
&\leq \|\mathbf{u}\|_2\|\mathbf{v}\|_2\left\|\frac{\mathbf{w}_{n,1}}{c_{n,1}}\right\|_2\left\|\frac{\mathbf{w}_{n,2}}{c_{n,2}}\right\|_2 \\
&\leq \left\|\frac{\mathbf{w}_{n,1}}{c_{n,1}}\right\|_1\left\|\frac{\mathbf{w}_{n,2}}{c_{n,2}}\right\|_1 \\
&\leq 1
\end{aligned}$$

and so $|\mathbf{u}^T\mathbb{E}[\hat{\mathbf{M}}_{n,s,a}]\mathbf{v}| \leq \mathbb{E}[|\mathbf{u}^T\hat{\mathbf{M}}_{n,s,a}\mathbf{v}|] \leq 1$. A little thought shows that the estimates $\hat{\mathbf{M}}_{n,s,a}$ are independent for $n \in \mathcal{N}_{traj}(s,a)$ when conditioned on the $\mathcal{N}_{traj}(s,a)$.

$$\mathbb{P}\left(\left|\sum_{n\in\mathcal{N}_{traj}(s,a)}\frac{1}{N_{traj}(s,a)}\mathbf{u}^T(\hat{\mathbf{M}}_{n,s,a} - \mathbb{E}[\hat{\mathbf{M}}_{n,s,a} \mid \mathcal{N}_{traj}(s,a)])\mathbf{v}\right| > \frac{\epsilon}{4}\middle|\mathcal{N}_{traj}(s,a)\right) < 2e^{-\frac{\epsilon^2 N_{traj}(s,a)}{32}}$$

Doing this for all 12^{2S} pairs **u,v**, we use inequality 9 to get that the conditionalprobability given by

$$\mathbb{P}\left(\left\|\sum_{n\in\mathcal{N}_{traj}(s,a)}\frac{1}{N_{traj}(s,a)}\hat{\mathbf{M}}_{n,s,a}-\mathbb{E}[\hat{\mathbf{M}}_{n,s,a}\mid\mathcal{N}_{traj}(s,a)]\right\|>\frac{\epsilon}{2}\middle|\mathcal{N}_{traj}(s,a)\right)$$

is bounded above by the following expression.

$$\begin{aligned}
&\mathbb{P}\left(\exists\mathbf{u},\mathbf{v}\in C;\left|\sum_{n\in\mathcal{N}_{traj}(s,a)}\frac{1}{N_{traj}(s,a)}\mathbf{u}^T(\hat{\mathbf{M}}_{n,s,a}-\mathbb{E}[\hat{\mathbf{M}}_{n,s,a}\mid\mathcal{N}_{traj}(s,a)])\mathbf{v}\right|>\frac{\epsilon}{4}\middle|\mathcal{N}_{traj}(s,a)\right)\\
&\leq\sum_{\mathbf{u},\mathbf{v}\in C}\mathbb{P}\left(\left|\sum_{n\in\mathcal{N}_{traj}(s,a)}\frac{1}{N_{traj}(s,a)}\mathbf{u}^T(\hat{\mathbf{M}}_{n,s,a}-\mathbb{E}[\hat{\mathbf{M}}_{n,s,a}\mid\mathcal{N}_{traj}(s,a)])\mathbf{v}\right|>\frac{\epsilon}{4}\middle|\mathcal{N}_{traj}(s,a)\right)\\
&<2*12^{2S}*e^{-\frac{\epsilon^2 N_{traj}(s,a)}{32}}
\end{aligned}$$

This is less than δ for $N_{traj}(s,a)>\frac{32}{\epsilon^2}(2S\log(12)+\log(\frac{2}{\delta}))$. Since this holds for such values of $N_{traj}(s,a)$ irrespective of $\mathcal{N}_{traj}(s,a)$, we can conclude that for $N_{traj}(s,a)>\frac{32}{\epsilon^2}(2S\log(12)+\log(\frac{2}{\delta}))$, with probability universally greater than $1-\delta$,

$$\|\hat{\mathbf{M}}_{s,a}-\mathbf{M}_{s,a}\|<\frac{\epsilon}{2}+\max_j\|\mathrm{Mix}(j)\|$$

Alternatively, this establishes that with probability greater than $1-\delta$, we have the following inequality involving the random variables $\hat{\mathbf{M}}_{s,a}$ and $N_{traj}(s,a)$.

$$\|\hat{\mathbf{M}}_{s,a}-\mathbf{M}_{s,a}\|<\sqrt{\frac{32}{N_{traj}(s,a)}(2S\log(12)+\log(\frac{2}{\delta}))}+\max_j\|\mathrm{Mix}(j)\|$$

E.3.4. BOUNDING THE MIXING TERM

We now resolve the last remaining thread, which is that of bounding the mixing term. Let's fix a j for this section, since proving our upper bounds for arbitrary j is sufficient. Let the joint distribution of the observations under label j be χ. Let its marginal on the segment Ω_i be χ_i. Let the marginals on each of the G single-step sub-blocks be $\chi_{i,g}$. Denote the product distribution $\prod_g\chi_{i,g}$ by Q_i.

$$\begin{aligned}
\|\mathrm{Mix}(j)\|&=\left\|\mathbb{E}\left[\frac{\mathbf{w}_{n,1}\mathbf{w}_{n,2}^T}{c_{n,1}c_{n,2}}\middle|k_n=j,\mathcal{B}_n\right]-\mathbb{P}_j(\cdot\mid s,a)\mathbb{P}_j(\cdot\mid s,a)^T\right\|\\
&=\left\|\frac{1}{\mathbb{P}(\mathcal{B}_n)}\mathbb{E}_\chi\left[\frac{\mathbf{w}_{n,1}\mathbf{w}_{n,2}^T}{c_{n,1}c_{n,2}}\mathbb{1}_{\mathcal{B}_n}\right]-\mathbb{P}_j(\cdot\mid s,a)\mathbb{P}_j(\cdot\mid s,a)^T\right\|\\
&\leq\frac{1}{\mathbb{P}(\mathcal{B}_n)}\left\|\mathbb{E}_\chi\left[\frac{\mathbf{w}_{n,1}\mathbf{w}_{n,2}^T}{c_{n,1}c_{n,2}}\mathbb{1}_{\mathcal{B}_n}\right]-\mathbb{E}_{\chi_1}\left[\frac{\mathbf{w}_{n,1}}{c_{n,1}}\mathbb{1}_{c_{n,1}\neq0}\right]\mathbb{E}_{\chi_2}\left[\frac{\mathbf{w}_{n,2}^T}{c_{n,2}}\mathbb{1}_{c_{n,2}\neq0}\right]\right\|\\
&\quad+\frac{1}{\mathbb{P}(\mathcal{B}_n)}\left\|\mathbb{E}_{\chi_1}\left[\frac{\mathbf{w}_{n,1}}{c_{n,1}}\mathbb{1}_{c_{n,1}\neq0}\right]\mathbb{E}_{\chi_2}\left[\frac{\mathbf{w}_{n,2}^T}{c_{n,2}}\mathbb{1}_{c_{n,2}\neq0}\right]-\mathbb{P}_{Q_1}(c_{n,1}\neq0)\mathbb{P}_{Q_2}(c_{n,2}\neq0)\mathbb{P}_j(\cdot\mid s,a)\mathbb{P}_j(\cdot\mid s,a)^T\right\|\\
&\quad+\frac{1}{\mathbb{P}(\mathcal{B}_n)}\left\|(\mathbb{P}_{Q_1}(c_{n,1}\neq0)\mathbb{P}_{Q_2}(c_{n,2}\neq0)-\mathbb{P}(c_{n,1}\neq0)\mathbb{P}(c_{n,2}\neq0))\mathbb{P}_j(\cdot\mid s,a)\mathbb{P}_j(\cdot\mid s,a)^T(\right\|\\
&\quad+\frac{1}{\mathbb{P}(\mathcal{B}_n)}\left\|(\mathbb{P}(c_{n,1}\neq0)\mathbb{P}(c_{n,2}\neq0)-\mathbb{P}(\mathcal{B}_n))\,\mathbb{P}_j(\cdot\mid s,a)\mathbb{P}_j(\cdot\mid s,a)^T\right\|\\
&\leq\frac{1}{\mathbb{P}(\mathcal{B}_n)}\left\|\mathbb{E}_\chi\left[\frac{\mathbf{w}_{n,1}\mathbf{w}_{n,2}^T}{c_{n,1}c_{n,2}}\mathbb{1}_{\mathcal{B}_n}\right]-\mathbb{E}_{\chi_1}\left[\frac{\mathbf{w}_{n,1}}{c_{n,1}}\mathbb{1}_{c_{n,1}\neq0}\right]\mathbb{E}_{\chi_2}\left[\frac{\mathbf{w}_{n,2}^T}{c_{n,2}}\mathbb{1}_{c_{n,2}\neq0}\right]\right\|
\end{aligned}$$

$$+ \frac{1}{\mathbb{P}(\mathcal{B}_n)} \left\| \mathbb{E}_{\chi_1}\left[\frac{\mathbf{w}_{n,1}}{c_{n,1}}\mathbb{1}_{c_{n,1}\neq 0}\right] \mathbb{E}_{\chi_2}\left[\frac{\mathbf{w}_{n,2}^T}{c_{n,2}}\mathbb{1}_{c_{n,2}\neq 0}\right] - \mathbb{P}_{Q_1}(c_{n,1}\neq 0)\mathbb{P}_{Q_2}(c_{n,2}\neq 0)\mathbb{P}_j(\cdot \mid s,a)\mathbb{P}_j(\cdot \mid s,a)^T \right\|$$
$$+ \frac{1}{\mathbb{P}(\mathcal{B}_n)} \left|\mathbb{P}_{Q_1}(c_{n,1}\neq 0)\mathbb{P}_{Q_2}(c_{n,2}\neq 0) - \mathbb{P}_{\chi_1}(c_{n,1}\neq 0)\mathbb{P}_{\chi_2}(c_{n,2}\neq 0)\right|$$
$$+ \frac{1}{\mathbb{P}(\mathcal{B}_n)} \left|\mathbb{P}_{\chi_1}(c_{n,1}\neq 0)\mathbb{P}_{\chi_2}(c_{n,2}\neq 0) - \mathbb{P}(\mathcal{B}_n)\right| \tag{10}$$

Here, in the last inequality, we used the fact that $\|\mathbb{P}_j(\cdot \mid s,a)\|_2 \leq \|\mathbb{P}_j(\cdot \mid s,a)\|_1 = 1$ and $\|\mathbf{a}\mathbf{a}^T\| \leq \|\mathbf{a}\|_2\|\mathbf{a}\|_2$. Also note that $\mathbb{P}_{\chi_i}(c_{n,i}\neq 0) = \mathbb{P}_{\chi}(c_{n,i}\neq 0) = \mathbb{P}(c_{n,i}\neq 0)$.

Intuitively, the first term represents mixing of the expectation across the two segments, the second term represents mixing of the expectations across the single-step sub-blocks inside segments, the third term represents mixing of the observation probabilities across the single-step sub-blocks inside segments, and the fourth term represents mixing of the observation probabilities across the two segments. In short, the first and fourth represent segment-level mixing while the second and third represent sub-block-level mixing.

Bounding the first term (segment-level mixing)

We will use (Yu, 1994)'s blocking technique again, invoking Lemma 2. Pick an arbitrary $\mathbf{u},\mathbf{v} \in \mathbb{S}^{S-1}$. Recall that

$$\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) := \frac{\mathbf{N}(n,i,s,a,\cdot)}{N(n,i,s,a)}\mathbb{1}_{N(n,i,s,a)\neq 0} = \frac{\mathbf{w}_{n,i}}{c_{n,i}}\mathbb{1}_{c_{n,i}\neq 0}$$

Consider the real-valued random variable

$$h_{\mathbf{u},\mathbf{v}} := \mathbf{u}^T\left(\frac{\mathbf{w}_{n,1}\mathbf{w}_{n,2}^T}{c_{n,1}c_{n,2}}\mathbb{1}_{\mathcal{B}_n}\right)\mathbf{v}$$

We have the following basic computations for expectations. Remember that $\mathbb{1}_{\mathcal{B}_n} = \mathbb{1}_{c_{n,1}c_{n,2}\neq 0} = \mathbb{1}_{c_{n,1}\neq 0}\mathbb{1}_{c_{n,2}\neq 0}$.

$$\mathbb{E}_{\chi}h_{\mathbf{u},\mathbf{v}} = \mathbf{u}^T\left(\mathbb{E}_{\chi}\left[\frac{\mathbf{w}_{n,1}\mathbf{w}_{n,2}^T}{c_{n,1}c_{n,2}}\mathbb{1}_{\mathcal{B}_n}\right]\right)\mathbf{v}$$

and

$$\mathbb{E}_{\chi_1\times\chi_2}h_{\mathbf{u},\mathbf{v}} = \mathbf{u}^T\left(\mathbb{E}_{\chi_1}\left[\frac{\mathbf{w}_{n,1}}{c_{n,1}}\mathbb{1}_{c_{n,1}\neq 0}\right]\mathbb{E}_{\chi_2}\left[\frac{\mathbf{w}_{n,2}^T}{c_{n,2}}\mathbb{1}_{c_{n,2}\neq 0}\right]\right)\mathbf{v}$$

This allows us to establish the following relation.

$$\sup_{\mathbf{u},\mathbf{v}\in\mathbb{S}^{S-1}} |\mathbb{E}_{\chi}h_{\mathbf{u},\mathbf{v}} - \mathbb{E}_{\chi_1\times\chi_2}h_{\mathbf{u},\mathbf{v}}| = \left\|\mathbb{E}_{\chi}\left[\frac{\mathbf{w}_{n,1}\mathbf{w}_{n,2}^T}{c_{n,1}c_{n,2}}\mathbb{1}_{\mathcal{B}_n}\right] - \mathbb{E}_{\chi_1}\left[\frac{\mathbf{w}_{n,1}}{c_{n,1}}\mathbb{1}_{c_{n,1}\neq 0}\right]\mathbb{E}_{\chi_2}\left[\frac{\mathbf{w}_{n,2}^T}{c_{n,2}}\mathbb{1}_{c_{n,2}\neq 0}\right]\right\|$$

Now, we want to use Lemma 2. Note the following upper bound.

$$\begin{aligned}
|h_{\mathbf{u},\mathbf{v}}| &\leq \|\mathbf{u}\|_2 \left\|\frac{\mathbf{w}_{n,1}}{c_{n,1}}\right\|_2 \left\|\frac{\mathbf{w}_{n,2}}{c_{n,2}}\right\|_2 \|\mathbf{v}\|_2 \\
&\leq \left\|\frac{\mathbf{w}_{n,1}}{c_{n,1}}\right\|_1 \left\|\frac{\mathbf{w}_{n,2}}{c_{n,2}}\right\|_1 \\
&= 1
\end{aligned}$$

So, we can use Lemma 2 with $C = C_{\mathbf{u},\mathbf{v}} := \sup h_{\mathbf{u},\mathbf{v}}$ and $n = 2$ for any $\mathbf{u},\mathbf{v} \in \mathbb{S}^{S-1}$, giving us the following inequality.

$$|\mathbb{E}_{\chi}h_{\mathbf{u},\mathbf{v}} - \mathbb{E}_{\chi_1\times\chi_2}h_{\mathbf{u},\mathbf{v}}| \leq 4\lambda^{\frac{T_n}{4}} \tag{11}$$

Since inequality 11 holds for any $\mathbf{u}, \mathbf{v} \in \mathbb{S}^{S-1}$, we can take the supremum over such $\mathbf{u}, \mathbf{v}$ to get the desired inequality below. We also recall that $\mathbb{P}(\mathcal{B}_n) \geq \frac{d_{min}s,a}{2}$ from equation 5.

$$\frac{1}{\mathbb{P}(\mathcal{B}_n)} \left\| \mathbb{E}_{\chi} \left[\frac{\mathbf{w}_{n,1}\mathbf{w}_{n,2}^T}{c_{n,1}c_{n,2}} \mathbb{1}_{\mathcal{B}_n} \right] - \mathbb{E}_{\chi_1} \left[\frac{\mathbf{w}_{n,1}}{c_{n,1}} \mathbb{1}_{c_{n,1}\neq 0} \right] \mathbb{E}_{\chi_2} \left[\frac{\mathbf{w}_{n,2}^T}{c_{n,2}} \mathbb{1}_{c_{n,2}\neq 0} \right] \right\| \leq \frac{1}{\mathbb{P}(\mathcal{B}_n)} 4\lambda^{\frac{T_n}{4}}$$
$$\leq \frac{8\lambda^{\frac{T_n}{4}}}{d_{min}(s,a)}$$

Bounding the second term (sub-block-level mixing)

Remember that the product distribution $\prod_g \chi_{i,g}$ is Q_i. First note that, since under Q_i, each observation is independent, we have the following expectation.

$$\begin{aligned}
\mathbb{E}_{Q_i} \left[\frac{\mathbf{w}_{n,i}}{c_{n,i}} \mathbb{1}_{c_{n,i}\neq 0} \right] &= \mathbb{E}_{Q_i} \left[\frac{\mathbb{E}_{Q_i}[\mathbf{w}_{n,i} \mid c_{n,i}]}{c_{n,i}} \mathbb{1}_{c_{n,i}\neq 0} \right] \\
&= \mathbb{E}_{Q_i} \left[\frac{\mathbb{P}_j(\cdot \mid s,a) c_{n,i}}{c_{n,i}} \mathbb{1}_{c_{n,i}\neq 0} \right] \\
&= \mathbb{P}_j(\cdot \mid s,a) \mathbb{P}_{Q_i}(c_{n,i} \neq 0)
\end{aligned} \tag{12}$$

Remark 8. Note that this holds crucially because we are working with the product distribution Q_i over the single-step sub-blocks.

Also, let $h_{\mathbf{u}} = \mathbf{u}^T \frac{\mathbf{w}_{n,1}}{c_{n,1}} \mathbb{1}_{c_{n,1}}$ and let $g_{\mathbf{v}} = \frac{\mathbf{w}_{n,2}^T}{c_{n,2}} \mathbb{1}_{c_{n,2}} \mathbf{v}$. Then the second term is exactly given by the following expression.

$$\frac{1}{\mathbb{P}(\mathcal{B}_n)} \sup_{\mathbf{u},\mathbf{v}\in\mathbb{S}^{S-1}} |\mathbb{E}_{\chi_1}[h_{\mathbf{u}}]\mathbb{E}_{\chi_2}[g_{\mathbf{v}}] - \mathbb{E}_{Q_1}[h_{\mathbf{u}}]\mathbb{E}_{Q_2}[g_{\mathbf{v}}]|$$

Also note that both $|h_{\mathbf{u}}|$ and $|g_{\mathbf{v}}|$ are bounded by 1. We then have the following chain of inequalities.

$$\begin{aligned}
|\mathbb{E}_{\chi_1}[h_{\mathbf{u}}]\mathbb{E}_{\chi_2}[g_{\mathbf{v}}] - \mathbb{E}_{Q_1}[h_{\mathbf{u}}]\mathbb{E}_{Q_2}[g_{\mathbf{v}}]| &\leq |\mathbb{E}_{\chi_1}[h_{\mathbf{u}}] - \mathbb{E}_{Q_1}[h_{\mathbf{u}}]|\,|\mathbb{E}_{\chi_2}[g_{\mathbf{v}}]| + |\mathbb{E}_{\chi_2}[g_{\mathbf{v}}] - \mathbb{E}_{Q_2}[g_{\mathbf{v}}]||\mathbb{E}_{Q_1}[h_{\mathbf{u}}]| \\
&\leq |\mathbb{E}_{\chi_1}[h_{\mathbf{u}}] - \mathbb{E}_{Q_1}[h_{\mathbf{u}}]| + |\mathbb{E}_{\chi_2}[g_{\mathbf{v}}] - \mathbb{E}_{Q_2}[g_{\mathbf{v}}]|
\end{aligned}$$

Since the single step sub-blocks are separated by at least $\frac{T_n}{8G}$ timesteps, we can apply Lemma 2 with $C = 1$ and $n = G$ to get bounds on both terms here, since $Q_i = \prod_g \chi_{i,g}$. Also remember that $\mathbb{P}(\mathcal{B}_n) \geq \frac{d_{min}(s,a)}{2}$ from equation 5.

$$\frac{1}{\mathbb{P}(\mathcal{B}_n)} \sup_{\mathbf{u},\mathbf{v}\in\mathbb{S}^{S-1}} |\mathbb{E}_{\chi_1}[h_{\mathbf{u}}]\mathbb{E}_{\chi_2}[g_{\mathbf{v}}] - \mathbb{E}_{Q_1}[h_{\mathbf{u}}]\mathbb{E}_{Q_2}[g_{\mathbf{v}}]| \leq \frac{1}{\mathbb{P}(\mathcal{B}_n)} \left(4G\lambda^{\frac{T_n}{8G}} + 4G\lambda^{\frac{T_n}{8G}} \right)$$
$$\leq \frac{16G\lambda^{\frac{T_n}{8G}}}{d_{min}(s,a)}$$

Bounding the third term (sub-block-level mixing)

Again, note that the third term is given by the following expression.

$$\frac{1}{\mathbb{P}(\mathcal{B}_n)} \left| \mathbb{E}_{Q_1}[\mathbb{1}_{c_{n,1}\neq 0}]\mathbb{E}_{Q_2}[\mathbb{1}_{c_{n,2}\neq 0}] - \mathbb{E}_{\chi_1}[\mathbb{1}_{c_{n,1}\neq 0}]\mathbb{E}_{\chi_2}[\mathbb{1}_{c_{n,2}\neq 0}] \right|$$

We can bound this above using the fact that $|ab - cd| \leq |b||a - c| + |c||b - d|$, to get the following upper bound.

$$\mathbb{E}_{Q_2}[\mathbb{1}_{c_{n,2}\neq 0}] \left|\mathbb{E}_{Q_1}[\mathbb{1}_{c_{n,1}\neq 0}] - \mathbb{E}_{\chi_1}[\mathbb{1}_{c_{n,1}\neq 0}]\right| + \mathbb{E}_{\chi_1}[\mathbb{1}_{c_{n,1}\neq 0}] \left|\mathbb{E}_{Q_2}[\mathbb{1}_{c_{n,2}\neq 0}] - \mathbb{E}_{\chi_2}[\mathbb{1}_{c_{n,2}\neq 0}]\right|$$

This in turn is bounded above by the expression below.

$$|\mathbb{E}_{Q_1}[\mathbb{1}_{c_{n,1}\neq 0}] - \mathbb{E}_{\chi_1}[\mathbb{1}_{c_{n,1}\neq 0}]| + |\mathbb{E}_{Q_2}[\mathbb{1}_{c_{n,2}\neq 0}] - \mathbb{E}_{\chi_2}[\mathbb{1}_{c_{n,2}\neq 0}]|$$

Since indicator functions are bounded above by 1, we can apply Lemma 2 as in the second term ($C = 1$, $n = G$) to bound both the differences above. Skipping the routine details, we finally get the following inequality, analogous to the second term.

$$\frac{1}{\mathbb{P}(\mathcal{B}_n)} \left|\mathbb{E}_{Q_1}[\mathbb{1}_{c_{n,1}\neq 0}]\mathbb{E}_{Q_2}[\mathbb{1}_{c_{n,2}\neq 0}] - \mathbb{E}_{\chi_1}[\mathbb{1}_{c_{n,1}\neq 0}]\mathbb{E}_{\chi_2}[\mathbb{1}_{c_{n,2}\neq 0}]\right| \leq \frac{16G\lambda^{\frac{T_n}{8G}}}{d_{min}(s,a)}$$

Bounding the fourth term (segment-level mixing)

Now note that the fourth term is the same as the expression below.

$$\frac{1}{\mathbb{P}(\mathcal{B}_n)}|\mathbb{E}_{\chi_1}[\mathbb{1}_{c_{n,1}\neq 0}]\mathbb{E}_{\chi_2}[\mathbb{1}_{c_{n,2}\neq 0}] - \mathbb{E}_{\chi}[\mathbb{1}_{c_{n,1}\neq 0}\mathbb{1}_{c_{n,2}\neq 0}]| = \frac{1}{\mathbb{P}(\mathcal{B}_n)}|\mathbb{E}_{\chi_1\times\chi_2}[\mathbb{1}_{c_{n,1}\neq 0}\mathbb{1}_{c_{n,2}\neq 0}] - \mathbb{E}_{\chi}[\mathbb{1}_{c_{n,1}\neq 0}\mathbb{1}_{c_{n,2}\neq 0}]|$$

We can now apply Lemma 2 with $C = 1$ and $n = 2$. The segments are separated by T and $\mathbb{P}(\mathcal{B}_n) \geq \frac{d_{min}(s,a)}{2}$, giving us the following bound.

$$\frac{1}{\mathbb{P}(\mathcal{B}_n)} |\mathbb{P}_{\chi_1}(c_{n,1} \neq 0)\mathbb{P}_{\chi_2}(c_{n,2} \neq 0) - \mathbb{P}(\mathcal{B}_n)| \leq \frac{8\lambda^{\frac{T_n}{4}}}{d_{min}(s,a)}$$

Combining all these, we get that

$$\begin{aligned}
\|\hat{\mathbf{M}}_{s,a} - \mathbf{M}_{s,a}\| &< \sqrt{\frac{32G}{N_{traj}(s,a)}(2S\log(12) + \log(\frac{2}{\delta}))} + \frac{16}{d_{min}(s,a)}\left(\frac{1}{4}\right)^{\frac{T_n}{4t_{mix}}} + \frac{32G}{d_{min}(s,a)}\left(\frac{1}{4}\right)^{\frac{T_n}{8Gt_{mix}}} \\
&\leq \sqrt{\frac{32}{N_{traj}(s,a)}(2S\log(12) + \log(\frac{2}{\delta}))} + \frac{48G}{d_{min}(s,a)}\left(\frac{1}{4}\right)^{\frac{T_n}{8Gt_{mix}}} \qquad (13)
\end{aligned}$$

as desired.

□

F. Proof of Theorem 3

Theorem 3 (Exact Clustering Guarantee)**.** *Pick any pair of trajectories n, m. Then for Freq_β so that it contains (s,a) with $d_{min}(s,a) \geq \Omega(\alpha)$, $T_n = \Omega(t_{mix}\log^4(1/\delta)/\alpha^3)$, with probability at least $1-\delta$,*

$$\left|\mathrm{dist}_1(m,n) - \|\Delta_{m,n}\|_2^2\right|$$

is

$$O\left(\sqrt{\frac{K\log(1/\delta)}{\alpha}}\left(\frac{t_{mix}}{T_n}\right)^{\frac{1}{3}}\right) + 4\epsilon_{sub}(\delta/2)$$

This means that if we choose $\lambda = 1$, then if $\epsilon_{sub}(\delta) \leq \Delta^2/32$ and $T_n = \Omega\left(K^{3/2}t_{mix}\frac{\log^4(N_{clust}/(\alpha\delta))}{\Delta^6\alpha^3}\right)$, no distance estimate attains a value between $\Delta^2/4$ and $\Delta^2/2$. So, Algorithm 2 attains exact clustering using a threshold of say $\Delta^2/3$ with probability at least $1-\delta$.

Proof. Consider the testing of trajectories m and n. Recall that we defined

$$\mathrm{dist}_1(m,n) := \max_{(s,a)\in SA_\alpha}\left[\left(\left(\hat{\mathbb{P}}_{n,1}(\cdot \mid s,a) - \hat{\mathbb{P}}_{m,1}(\cdot \mid s,a)\right)^T \mathbf{V}_{s,a}\right)\left(\left(\hat{\mathbb{P}}_{n,2}(\cdot \mid s,a) - \hat{\mathbb{P}}_{m,2}(\cdot \mid s,a)\right)^T \mathbf{V}_{s,a}\right)^T\right]$$

Let k_m be the label of trajectory m and k_n the label of trajectory n. According to our assumptions, if $k_m \neq k_n$, then we have an s,a so that $d_{k_m}(s,a), d_{k_n}(s,a) \geq \alpha$ and $\|\mathbb{P}_{k_m}(\cdot \mid s,a) - \mathbb{P}_{k_n}(\cdot \mid s,a)\|_2 \geq \Delta$. We will make s,a implicit in our notation except in $\mathbb{P}_j(\cdot \mid s,a)$. Let $c_{n,i} := N(n,i,s,a)$, $\mathbf{w}_{n,i} := \mathbf{N}(n,i,s,a,\cdot)$. Recall that we have two nested partitions: (1) of the entire trajectory into the two Ω_i and (2) of each segment Ω_i into G blocks. Finally, define $\mathrm{dist}_{1,(s,a)}$ as below, suppressing m and n. Note that $\mathrm{dist}_1(m,n)$ is the maximum of $\mathrm{dist}_{1,(s,a)}$ over all $(s,a) \in \mathrm{Freq}_\beta$, for the given two trajectories m and n.

$$\mathrm{dist}_{1,(s,a)} := \left[\left((\hat{\mathbb{P}}_{n,1}(\cdot \mid s,a) - \hat{\mathbb{P}}_{m,1}(\cdot \mid s,a))^T\mathbf{V}_{s,a}\right)\left((\hat{\mathbb{P}}_{n,2}(\cdot \mid s,a) - \hat{\mathbb{P}}_{m,2}(\cdot \mid s,a))^T V_{s,a}\right)^T\right]$$

We want to show that this is close to $\|\Delta_{m,n}(s,a)\|_2^2$ for the (s,a) pairs that we search over, where

$$\Delta_{m,n}(s,a) = \mathbb{P}_{k_m}(\cdot \mid s,a) - \mathbb{P}_{k_n}(\cdot \mid s,a)$$

Assume the lemma below for now, we prove it in the next subsection.

Lemma 4. *We claim that there is a universal constant C_1 so that for any (s,a) with $d_{min}(s,a) \geq \alpha/3$, with probability at least $1-\delta$,*

$$\left|\mathrm{dist}_{1,(s,a)} - \|\Delta_{m,n}(s,a)\|_2^2\right| \leq C_1\left(\sqrt{\frac{K+\log(1/\delta)}{G\alpha}}\right) + 4\epsilon_{sub}(\delta/2)$$

whenever $T_n \geq \Omega\left(Gt_{mix}\log(G/\delta)\log(1/\alpha)\right)$ and $G \geq \Omega\left(\frac{\log(1/\delta)}{\alpha^2}\right)$. Here, $\epsilon_{sub}(\delta)$ is the high probability bound on $\|\mathbb{P}_j(\cdot \mid s,a) - \boldsymbol{V}_{s,a}\boldsymbol{V}_{s,a}^T\mathbb{P}_j(\cdot \mid s,a)\|_2$ with $j = k_n, k_m$, from Theorem 2 (satisfied with probability $> 1-\delta$).

We now set $G = \left(\frac{T_n}{t_{mix}}\right)^{\frac{2}{3}}$. Then a sufficient condition on T_n to meet the conditions of the lemma is $T_n = \Omega(t_{mix}\log^4(1/\delta)/\alpha^3)$, under which, with probability at lest $1-\delta$, we have the following bound for (s,a) with $d_{min}(s,a) \geq \alpha/3$.

$$\left|\mathrm{dist}_{1,(s,a)} - \|\Delta_{m,n}(s,a)\|_2^2\right| \leq O\left(\sqrt{\frac{K\log(1/\delta)}{\alpha}}\left(\frac{t_{mix}}{T_n}\right)^{\frac{1}{3}}\right) + 4\epsilon_{sub}(\delta/2) \tag{14}$$

It is now easy to see that the first term on the right-hand side is less than $\Delta^2/8$ when $T_n = \Omega\left(K^{3/2} t_{mix} \frac{\log^{3/2}(1/\delta)}{\Delta^6 \alpha^{3/2}}\right)$ and $T_n = \Omega(t_{mix} \log^4(1/\delta)/\alpha^3)$. We can combine these to have the guarantee that the first term on the right-hand side is less $\Delta^2/8$ with probability at least $1-\delta$ when $T_n = \Omega\left(K^{3/2} t_{mix} \frac{\log^4(1/\delta)}{\Delta^6 \alpha^3}\right)$.

Now note that if $\beta \geq \alpha/3$, then a separating state action pair always lies in Freq_β and thus, the maximum over the $\|\Delta_{m,n}(s,a)\|_2^2$ values corresponding to Freq_β is in fact either 0 if $k_m = k_n$ or larger than Δ^2 if $k_m \neq k_n$. So, if $\epsilon_{sub}(\delta/2) \leq \Delta^2/32$ and for each of the (s,a) pairs, the first term on the right-hand side of inequality 20 is less than $\Delta^2/8$, then our distance estimate $\text{dist}_1(m,n)$ is on the right side of any threshold as long as $\Delta^2/4 \leq \tau \leq \Delta^2/2$. That is, the distance estimate is then less than the threshold if $k_m = k_n$, and larger than it if $k_m \neq k_n$.

Note that upon choosing an occurrence threshold of order α, we will have at most $O(1/\alpha)$ many (s,a) pairs in Freq_β to maximize $\text{dist}_{1,(s,a)}$ over to get $\text{dist}_1(m,n)$. By applying a union bound over all (s,a) pairs in Freq_β and using the conclusion of the previous paragraph, we correctly determine if $k_m = k_n$ with probability $1-\delta$ for $T_n = \Omega\left(K^{3/2} t_{mix} \frac{\log^4(1/(\alpha\delta))}{\Delta^6 \alpha^3}\right)$, as long as $\epsilon_{sub}(\delta/2) \leq \Delta^2/32$ and $\Delta^2/4 \leq \tau \leq \Delta^2/2$.

By applying a union bound over incorrectly deciding whether or not $k_m = k_n$ for any of the $N_{clust}(N_{clust}-1)/2$ pairs, we get that we can recover the true clusters with probability at least $1-\delta$ for $T_n = \Omega\left(K^{3/2} t_{mix} \frac{\log^4(N_{clust}/(\alpha\delta))}{\Delta^6 \alpha^3}\right)$, whenever $\epsilon_{sub} \leq \Delta^2/32$ and as long as $\epsilon_{sub}(\delta/2) \leq \Delta^2/32$ and $\Delta^2/4 \leq \tau \leq \Delta^2/2$. □

F.1. Proof of Lemma 4

We recall the statement of the lemma.

Lemma 4. *We claim that there is a universal constant C_1 so that for any (s,a) with $d_{min}(s,a) \geq \alpha/3$, with probability at least $1-\delta$,*

$$\left|\text{dist}_{1,(s,a)} - \|\Delta_{m,n}(s,a)\|_2^2\right| \leq C_1 \left(\sqrt{\frac{K + \log(1/\delta)}{G\alpha}}\right) + 4\epsilon_{sub}(\delta/2)$$

whenever $T_n \geq \Omega\left(G t_{mix} \log(G/\delta) \log(1/\alpha)\right)$ and $G \geq \Omega\left(\frac{\log(1/\delta)}{\alpha^2}\right)$. Here, $\epsilon_{sub}(\delta)$ is the high probability bound on $\|\mathbb{P}_j(\cdot \mid s,a) - \boldsymbol{V}_{s,a}\boldsymbol{V}_{s,a}^T \mathbb{P}_j(\cdot \mid s,a)\|_2$ with $j = k_n, k_m$, from Theorem 2 (satisfied with probability $> 1-\delta$).

Notation: We say $c_{n,i} = N(n,i,s,a)$ as in the statement of the lemma and $\mathbf{w}_{n,i} = \mathbf{N}(n,i,s,a,\cdot)$. Let the joint distribution of the observations over the pair of trajectories (m,n) be χ. This means that χ is the product of the joint distribution of the observations over the trajectory m and that of the observations over the trajectory n, since trajectories are generated independently. Let its marginals on the segments Ω_i be χ_i. Let the marginals on each of the G single-step sub-blocks along with their next states be $\chi_{i,g}$. Denote the product distribution $\prod_g \chi_{i,g}$ by Q_i. Let $\mathcal{G}(s,a)$ denote the two sets of indices where the state-action pair (s,a) is observed in trajectories n and m. For brevity, we will abbreviate $\mathcal{G}(s,a)$ to $\mathcal{G}$. Note that the sizes of these two sets are exactly $c_{n,i}$ and $c_{m,i}$ respectively.

We first prove some preliminary lemmas.

F.1.1. Decomposition of $|\,\text{dist}_{1,(s,a)} - \|\Delta_{m,n}(s,a)\|_2^2\,|$

Lemma 5. *We claim that for each fixed value of $\mathcal{G}(s,a)$ (abbreviated to $\mathcal{G}$), with probability at least $1-\delta$, the following bound holds.*

$$\left|\text{dist}_{1,(s,a)} - \|\Delta_{m,n}(s,a)\|_2^2\right| \leq \sum_{i=1}^{2} 2\,\|\boldsymbol{\Delta}_i - \mathbb{E}_{Q_i}[\boldsymbol{\Delta}_i \mid \mathcal{G}]\|_2 + 4\epsilon_{sub}(\delta) + 4\left(\max_i \mathbb{1}_{c_{n,i}=0} + \max_i \mathbb{1}_{c_{m,i}=0}\right) \tag{15}$$

Here $c_{n,i} = N(n,i,s,a)$, $\epsilon_{sub}(\delta)$ is the high probability bound on $\|\mathbb{P}_{j_l}(\cdot \mid s,a) - \boldsymbol{V}_{s,a}\boldsymbol{V}_{s,a}^T \mathbb{P}_{j_l}(\cdot \mid s,a)\|_2$ from Theorem 2 (satisfied with probability $> 1-\delta$), and

$$\boldsymbol{\Delta}_i^T = (\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) - \hat{\mathbb{P}}_{m,i}(\cdot \mid s,a))^T \boldsymbol{V}_{s,a}$$

Remark 9. In the inequality,

- The first term is a concentration-type term, which will be broken into an "independent concentration" error and a mixing error to account for the low but non-zero dependence across blocks.
- The second term accounts for subspace estimation error.
- The third term accounts for actually observing s, a in our blocks.

Proof. We first establish a simple inequality.

$$\begin{aligned}
|\operatorname{dist}_{1,(s,a)} - \mathbb{E}_{Q_1}[\mathbf{\Delta}_1^T \mid \mathcal{G}]\mathbb{E}_{Q_2}[\mathbf{\Delta}_2 \mid \mathcal{G}]| &= |\mathbf{\Delta}_1^T \mathbf{\Delta}_2 - \mathbb{E}_{Q_1}[\mathbf{\Delta}_1^T \mid \mathcal{G}]\mathbb{E}_{Q_2}[\mathbf{\Delta}_2| \mid \mathcal{G}] \\
&\leq |(\mathbf{\Delta}_1^T - \mathbb{E}_{Q_1}[\mathbf{\Delta}_1^T \mid \mathcal{G}])\mathbb{E}_{Q_2}[\mathbf{\Delta}_2 \mid \mathcal{G}]| + |\mathbf{\Delta}_1^T(\mathbf{\Delta}_2 - \mathbb{E}_{Q_2}[\mathbf{\Delta}_2 \mid \mathcal{G}])| \\
&\leq \|\mathbf{\Delta}_1 - \mathbb{E}_{Q_1}[\mathbf{\Delta}_1 \mid \mathcal{G}]\|_2 \|\mathbb{E}_{Q_2}[\mathbf{\Delta}_2 \mid \mathcal{G}]\|_2 + \|\mathbf{\Delta}_1\|_2 \|\mathbf{\Delta}_2 - \mathbb{E}_{Q_2}[\mathbf{\Delta}_2 \mid \mathcal{G}]\|_2 \\
&\leq 2\|\mathbf{\Delta}_1 - \mathbb{E}_{Q_1}[\mathbf{\Delta}_1 \mid \mathcal{G}]\|_2 + 2\|\mathbf{\Delta}_2 - \mathbb{E}_{Q_2}[\mathbf{\Delta}_2 \mid \mathcal{G}]\|_2 \qquad (16)
\end{aligned}$$

Remark 10. Notice that because of this inequality, the double estimator does not impact any theoretical guarantees for exact clustering w.h.p, which is the form of the guarantees in both Kong et al. (2020) and Chen & Poor (2022). However, we find that using a double estimator allows for better performance in real life. This makes sense because while exact clustering doesn't need a double estimator, approximate clustering w.h.p. does depend on the expectation of the distances across pairs of trajectories. This expectation is controlled by the covariance of $\mathbf{\Delta}_1$ and $\mathbf{\Delta}_2$.

We define the following quantity.

$$\mathbf{diff}_i = \left(\mathbb{1}_{c_{n,i}\neq 0}\mathbb{P}_{k_m}(\cdot \mid s,a) - \mathbb{1}_{c_{m,i}\neq 0}\mathbb{P}_{k_n}(\cdot \mid s,a)\right)$$

Note that $\|\mathbf{diff}_i\|_2 \leq 2$. Note the following expectation, which uses the dieas from equation 12.

$$\begin{aligned}
\mathbb{E}_{Q_i}[\Delta_i \mid \mathcal{G}] &= \mathbb{E}_{Q_i}\left[\mathbf{V}_{s,a}^T(\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) - \hat{\mathbb{P}}_{m,i}(\cdot \mid s,a))\right] \\
&= \mathbf{V}_{s,a}^T\left(\mathbb{E}_{Q_i}[\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) \mid \mathcal{G}] - \mathbb{E}_{Q_i}[\hat{\mathbb{P}}_{m,i}(\cdot \mid s,a) \mid \mathcal{G}]\right) \\
&= \mathbf{V}_{s,a}^T\left(\mathbb{E}_{Q_i}\left[\frac{\mathbf{w}_{n,i}}{c_{n,i}}\mathbb{1}_{c_{n,i}\neq 0} \mid \mathcal{G}\right] - \mathbb{E}_{Q_i}\left[\frac{\mathbf{w}_{m,i}}{c_{m,i}}\mathbb{1}_{c_{m,i}\neq 0} \mid \mathcal{G}\right]\right) \\
&= \mathbf{V}_{s,a}^T\left(\frac{\mathbb{E}_{Q_i}[\mathbf{w}_{n,i} \mid \mathcal{G}]}{c_{n,i}}\mathbb{1}_{c_{n,i}\neq 0} - \frac{\mathbb{E}_{Q_i}[\mathbf{w}_{m,i} \mid \mathcal{G}]}{c_{m,i}}\mathbb{1}_{c_{m,i}\neq 0}\right) \\
&= \mathbf{V}_{s,a}^T\left(\frac{\mathbb{P}_{k_n}(\cdot \mid s,a)c_{n,i}}{c_{n,i}}\mathbb{1}_{c_{n,i}\neq 0} - \frac{\mathbb{P}_{k_m}(\cdot \mid s,a)c_{m,i}}{c_{m,i}}\mathbb{1}_{c_{m,i}\neq 0}\right) \\
&= \mathbf{V}_{s,a}^T\left(\mathbb{P}_{k_n}(\cdot \mid s,a)\mathbb{1}_{c_{n,i}\neq 0} - \mathbb{P}_{k_m}(\cdot \mid s,a)\mathbb{1}_{c_{m,i}\neq 0}\right) \\
&= \mathbf{V}_{s,a}^T\mathbf{diff}_i
\end{aligned}$$

We recall the following definition before proceeding to show the main inequality.

$$\Delta_{m,n}(s,a) = \mathbb{P}_{k_m}(\cdot \mid s,a) - \mathbb{P}_{k_n}(\cdot \mid s,a)$$

$$\begin{aligned}
\left|\mathbb{E}_{Q_1}[\mathbf{\Delta}_1^T \mid \mathcal{G}]\mathbb{E}_{Q_2}[\mathbf{\Delta}_2 \mid \mathcal{G}] - \|\Delta_{m,n}(s,a)\|_2^2\right| &= \left|\mathbf{diff}_1^T\mathbf{V}_{s,a}\mathbf{V}_{s,a}^T\mathbf{diff}_2 - \mathbf{diff}_1^T\mathbf{diff}_2\right| + \left|\mathbf{diff}_1^T\mathbf{diff}_2 - \|\Delta_{m,n}(s,a)\|_2^2\right| \\
&\leq \|\mathbf{diff}_1\|_2 \left\|\mathbf{diff}_2 - \mathbf{V}_{s,a}\mathbf{V}_{s,a}^T\mathbf{diff}_2\right\|_2 + \|\mathbf{diff}_1 - \Delta_{m,n}(s,a)\|_2 \|\mathbf{diff}_2\|_2 \\
&\quad + \|\mathbf{diff}_1\|_2 \|\mathbf{diff}_2 - \Delta_{m,n}(s,a)\|_2 \\
&\leq \|\mathbf{diff}_1\|_1 \left\|\mathbf{diff}_2 - \mathbf{V}_{s,a}\mathbf{V}_{s,a}^T\mathbf{diff}_2\right\|_2 + \|\mathbf{diff}_1 - \Delta_{m,n}(s,a)\|_2 \|\mathbf{diff}_2\|_1 \\
&\quad + \|\mathbf{diff}_1\|_1 \|\mathbf{diff}_2 - \Delta_{m,n}(s,a)\|_2
\end{aligned}$$

$$
\begin{aligned}
&\leq 2\left\|\mathbf{diff}_2 - \mathbf{V}_{s,a}\mathbf{V}_{s,a}^T\mathbf{diff}_2\right\|_2 + 2\left\|\mathbf{diff}_1 - \Delta_{m,n}(s,a)\right\|_2 \\
&\quad + 2\left\|\mathbf{diff}_2 - \Delta_{m,n}(s,a)\right\|_2 \\
&\leq 2\left\|\mathbb{P}_{k_m}(\cdot \mid s,a) - \mathbf{V}_{s,a}\mathbf{V}_{s,a}^T\mathbb{P}_{k_m}(\cdot \mid s,a)\right\|_2 \\
&\quad + 2\left\|\mathbb{P}_{k_n}(\cdot \mid s,a) - \mathbf{V}_{s,a}\mathbf{V}_{s,a}^T\mathbb{P}_{k_n}(\cdot \mid s,a)\right\|_2 \\
&\quad + 2\left\|\mathbb{1}_{c_{m,1}=0}\mathbb{P}_{k_m}(\cdot \mid s,a) - \mathbb{1}_{c_{n,1}=0}\mathbb{P}_{k_n}(\cdot \mid s,a)\right\|_2 \\
&\quad + 2\left\|\mathbb{1}_{c_{m,2}=0}\mathbb{P}_{k_m}(\cdot \mid s,a) - \mathbb{1}_{c_{n,2}=0}\mathbb{P}_{k_n}(\cdot \mid s,a)\right\|_2 \\
&\leq 4\epsilon_{sub}(\delta) + 2\left(\mathbb{1}_{c_{m,1}=0}\left\|\mathbb{P}_{k_m}(\cdot \mid s,a)\right\|_2 + \mathbb{1}_{c_{n,1}=0}\left\|\mathbb{P}_{k_n}(\cdot \mid s,a)\right\|_2\right) \\
&\quad + 2\left(\mathbb{1}_{c_{m,2}=0}\left\|\mathbb{P}_{k_m}(\cdot \mid s,a)\right\|_2 + \mathbb{1}_{c_{n,2}=0}\left\|\mathbb{P}_{k_n}(\cdot \mid s,a)\right\|_2\right) \\
&\leq 4\epsilon_{sub}(\delta) + 4\left(\max_i \mathbb{1}_{c_{n,i}=0} + \max_i \mathbb{1}_{c_{m,i}=0}\right)
\end{aligned}
$$

Combining this with inequality 16, we have the following final bound.

$$
\left|\mathrm{dist}_{1,(s,a)} - \left\|\Delta_{m,n}(s,a)\right\|_2^2\right| \leq \sum_{i=1}^{2} 2\left\|\mathbf{\Delta}_i - \mathbb{E}_{Q_i}[\mathbf{\Delta}_i \mid \mathcal{G}]\right\|_2 + 4\epsilon_{sub}(\delta) + 4\left(\max_i \mathbb{1}_{c_{n,i}=0} + \max_i \mathbb{1}_{c_{m,i}=0}\right) \tag{17}
$$

where we remind the reader that $c_{n,i} = N(n,i,s,a)$ and recall the definition of Δ_i.

$$
\mathbf{\Delta}_i^T = (\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) - \hat{\mathbb{P}}_{m,i}(\cdot \mid s,a))^T\mathbf{V}_{s,a}
$$

□

F.1.2. Bounding the concentration-type term

We bound the first term in the decomposition lemma (Lemma 5) with high probability.

Lemma 6. *With probability at least $1-\delta$, when $T_n \geq \Omega\left(Gt_{mix}\log\left(\frac{G}{\delta}\log(1/\alpha)\right)\right)$ and $G \geq \Omega\left(\frac{\log(1/\delta)}{\alpha^2}\right)$, we have the following bound.*

$$
\left\|\Delta_i - \mathbb{E}_{Q_i}[\Delta_i \mid \mathcal{G}]\right\|_2 \leq O\left(\sqrt{\frac{K+\log(1/\delta)}{G\alpha}}\right)
$$

Proof. Recall that the joint distribution of the observations over the pair of trajectories (m,n) is χ. Its marginals on the segments Ω_i are χ_i. The marginals on each of the G single-step sub-blocks is $\chi_{i,g}$. The product distribution $\prod_g \chi_{i,g}$ is Q_i. Recall that $\mathcal{G}(n,s,a)$ denotes the two sets of indices where (s,a) is observed in trajectory n and m respectively, and the sets have sizes $c_{n,i}$ and $c_{m,i}$ respectively.

Let $\mathbf{w}_{n,i,g}$ be the one hot vector of the next state if the (i,g) sub-block witnesses (s,a), and the zero vector otherwise. Let $c_{n,i,g}$ be the indicator of (s,a) in the (i,g) sub-block. Then $\mathbf{w}_{n,i} = \sum_g \mathbf{w}_{n,i,g}$ and $c_{n,i} = \sum_g c_{n,i,g}$.

1. Covering argument for the product distribution

Pick a unit vector $\mathbf{u} \in \mathcal{R}^K$ and consider the following inequality. Remember that we abbreviate $\mathcal{G}(n,s,a)$ to $\mathcal{G}$.

$$
\begin{aligned}
|\mathbf{u}^T(\Delta_i - \mathbb{E}_{Q_i}[\Delta_i \mid \mathcal{G}])| &\leq |\mathbf{u}^T\mathbf{V}_{s,a}(\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) - \mathbb{E}_{Q_i}[\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) \mid \mathcal{G}])| \\
&\quad + |\mathbf{u}^T\mathbf{V}_{s,a}(\hat{\mathbb{P}}_{m,i}(\cdot \mid s,a) - \mathbb{E}_{Q_i}[\hat{\mathbb{P}}_{m,i}(\cdot \mid s,a) \mid \mathcal{G}])|
\end{aligned}
$$

We work with the term for trajectory n, WLOG. Any bounds thus obtained will also apply to trajectory m. Notice the following equation.

$$
|\mathbf{u}^T\mathbf{V}_{s,a}^T(\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) - \mathbb{E}_{Q_i}[\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) \mid \mathcal{G}])| = \left|\frac{1}{c_{n,i}}\sum_{g\in\mathcal{G}(n,s,a)}\left(\mathbf{u}^T\mathbf{V}_{s,a}^T\mathbf{w}_{n,i,g} - \mathbb{E}_{Q_i}[\mathbf{u}^T\mathbf{V}_{s,a}^T\mathbf{w}_{n,i,g} \mid \mathcal{G}]\right)\right|
$$

Note that $|\mathbf{u}^T\mathbf{V}_{s,a}^T\mathbf{w}_{n,i,g}| \leq \|\mathbf{u}\|_2\|\mathbf{V}_{s,a}^T\mathbf{w}_{n,i,g}\|_2 \leq 1$. Note that conditioned on the set of (s,a) observations in trajectory n, the next states are independent under the product distribution Q_i (but not under χ_i, of course). Now, using the conditional version of Hoeffding's inequality from Lemma 3, we get the following bound.

$$\mathbb{P}_{Q_i}\left(\left|\frac{1}{c_{n,i}}\sum_{g\in\mathcal{G}(n,s,a)}\left(\mathbf{u}^T\mathbf{V}_{s,a}^T\mathbf{w}_{n,i,g} - \mathbb{E}_{Q_i}[\mathbf{u}^T\mathbf{V}_{s,a}^T\mathbf{w}_{n,i,g} \mid \mathcal{G}]\right)\right| > \frac{\epsilon}{8}\middle|\mathcal{G}\right) \leq 2e^{-\frac{\epsilon^2 c_{n,i}}{32}}$$

Note that if $X \leq Y + Z$, then $\mathbb{P}(X > \frac{\epsilon}{4}) \leq \mathbb{P}(Y > \frac{\epsilon}{8}) + \mathbb{P}(Z > \frac{\epsilon}{8})$ by a union bound. We apply this to the inequalities above with $X = |\mathbf{u}^T(\Delta_i - \mathbb{E}_{Q_i}[\Delta_i])|$ to get the following concentration inequality.

$$\mathbb{P}_{Q_i}\left(|\mathbf{u}^T(\Delta_i - \mathbb{E}_{Q_i}[\Delta_i \mid \mathcal{G}])| > \frac{\epsilon}{4} \mid \mathcal{G}\right) \leq 2e^{-\frac{\epsilon^2 c_{n,i}}{32}} + 2e^{-\frac{\epsilon^2 c_{n,i}}{32}} = 4e^{-\frac{\epsilon^2 c_{n,i}}{32}}$$

Consider a covering of $\mathbb{S}^{K-1}$ by balls of radius $1/4$. We will need at most 12^K such balls. Call the set of their centers C. We know that for any vector $\mathbf{v}$, the following holds.

$$\sup_{\|\mathbf{u}\|_2\leq 1}\mathbf{u}^T\mathbf{v} = \|\mathbf{v}\|_2 \leq 2\sup_{\mathbf{u}\in C}\mathbf{u}^T\mathbf{v}$$

We use this to arrive at the concentration inequality below.

$$\begin{aligned}
\mathbb{P}_{Q_i}\left(\|\Delta_i - \mathbb{E}_{Q_i}[\Delta_i \mid \mathcal{G}])\|_2 > \frac{\epsilon}{2} \mid \mathcal{G}\right) &\leq \mathbb{P}_{Q_i}\left(\exists \mathbf{u}\in C; |\mathbf{u}^T(\Delta_i - \mathbb{E}_{Q_i}[\Delta_i \mid \mathcal{G}])| > \frac{\epsilon}{4} \mid \mathcal{G}\right)\\
&\leq \sum_{\mathbf{u}\in C}\mathbb{P}_{Q_i}\left(|\mathbf{u}^T(\Delta_i - \mathbb{E}_{Q_i}[\Delta_i \mid \mathcal{G}])| > \frac{\epsilon}{4} \mid \mathcal{G}\right)\\
&< 4 * 12^K * e^{-\frac{\epsilon^2 c_{n,i}}{32}}
\end{aligned}$$

2. Bounding $c_{n,i}$

We bound $c_{n,i}$ with high probability under the distribution Q_i, using the regular Hoeffding's inequality, noting that $\mathbb{E}_{Q_i}[c_{n,i}] = \sum_g \mathbb{P}_{Q_i}(c_{n,i,g} \neq 0) = \sum_g \mathbb{P}_{\chi_i}(c_{n,i,g} \neq 0)$. We can show that $\mathbb{P}_{\chi_i}(c_{n,i,g} \neq 0) \geq \frac{d_{min}(s,a)}{2}$ for $T_n \geq \Omega(Gt_{mix}\log(1/\alpha))$ by using the same kind of computation as in equation 4.

$$\begin{aligned}
\mathbb{P}_{Q_i}\left(c_{n,i} < G\frac{d_{min}(s,a)}{4}\right) &= \mathbb{P}_{Q_i}\left(c_{n,i} < G\frac{d_{min}(s,a)}{2} - G\frac{d_{min}(s,a)}{4}\right)\\
&\leq \mathbb{P}_{Q_i}\left(c_{n,i} < \mathbb{E}_{Q_i}[c_{n,i}] - G\frac{d_{min}(s,a)}{4}\right)\\
&= \mathbb{P}\left(\sum_g \mathbb{1}_{c_{n,i,g}\neq 0} < \sum_g \mathbb{E}_{Q_i}[\mathbb{1}_{c_{n,i,g}\neq 0}] - G\frac{d_{min}(s,a)}{4}\right)\\
&\leq \exp\left(-\frac{d_{min}(s,a)^2 G}{32}\right)
\end{aligned}$$

This is less than $\delta/2$ for $G \geq \Omega\left(\frac{\log(2/\delta)}{\alpha^2}\right) \geq \Omega\left(\frac{\log(2/\delta)}{d_{min}(s,a)^2}\right)$. So for such G, remembering that $\mathcal{G}$ was an abbreviation for the random set $\mathcal{G}(n,s,a)$,

$$\mathbb{P}_{Q_i}\left(\|\Delta_i - \mathbb{E}_{Q_i}[\Delta_i \mid \mathcal{G}])\|_2 > \frac{\epsilon}{2} \mid \mathcal{G}(n,s,a)\right) \leq 4 * 12^K * e^{-\frac{\epsilon^2 G d_{min}(s,a)}{128}} + \frac{\delta}{2}$$

Since this holds for all possible $\mathcal{G}(n,s,a)$ values and the right hand side doesn't depend on $\mathcal{G}(n,s,a)$, we can take the expectation over the random set $\mathcal{G}(n,s,a)$ to get the following inequality.

$$\mathbb{P}_{Q_i}\left(\|\Delta_i - \mathbb{E}_{Q_i}[\Delta_i \mid \mathcal{G}])\|_2 > \frac{\epsilon}{2}\right) \leq 4 * 12^K * e^{-\frac{\epsilon^2 G d_{min}(s,a)}{128}} + \frac{\delta}{2}$$

3. Accounting for non-independence (mixing error)

We know that we can bound the difference in the probability of any event E between χ_i and Q_i by applying Lemma 2 to the function $h = \mathbb{1}_E$ with $n = G$ and $C = 1$ as we have before, giving us the following inequality.

$$\begin{aligned}\mathbb{P}_{\chi_i}\left(\|\Delta_i - \mathbb{E}_{Q_i}[\Delta_i \mid \mathcal{G}])\|_2 > \frac{\epsilon}{2}\right) &\leq \mathbb{P}_{Q_i}\left(\|\Delta_i - \mathbb{E}_{Q_i}[\Delta_i \mid \mathcal{G}])\|_2 > \frac{\epsilon}{2}\right) + \frac{\delta}{2} + 4G\left(\frac{1}{4}\right)^{\frac{T_n}{8Gt_{mix}}} \\ &\leq 4 * 12^K * e^{-\frac{\epsilon^2 G d_{min}(s,a)}{128}} + \frac{\delta}{2} + 4G\left(\frac{1}{4}\right)^{\frac{T_n}{8Gt_{mix}}}\end{aligned}$$

We know that both terms are less than $\frac{\delta}{4}$ when $T_n \geq \Omega\left(Gt_{mix}\log\left(\frac{G}{\delta}\right)\right)$ and $G \geq \Omega\left(\frac{K+\log(1/\delta)}{\epsilon^2\alpha}\right)$, since $d_{min}(s,a) \geq \alpha/3$. We thus have the following bound with probability at least $1-\delta$, when $T_n \geq \Omega\left(Gt_{mix}\log\left(\frac{G}{\delta}\right)\log(1/\alpha)\right)$ and $G \geq \Omega\left(\frac{\log(1/\delta)}{\alpha^2}\right)$.

$$\|\Delta_i - \mathbb{E}_{Q_i}[\Delta_i \mid \mathcal{G}])\|_2 \leq O\left(\sqrt{\frac{K+\log(1/\delta)}{G\alpha}}\right)$$

□

F.1.3. Bounding the probability of not observing s, a

We bound the third term in the decomposition lemma (Lemma 5) with high probability. We first need an auxiliary lemma for this.

Lemma 7. *For $T_n \geq \Omega\left(Gt_{mix}\log(1/\alpha)\right)$, we have the following bound.*

$$\mathbb{P}(c_{n,i} = 0) \leq \left(1 - \frac{d_{min}(s,a)}{2}\right)^G + 4G\left(\frac{1}{4}\right)^{\frac{T_n}{8Gt_{mix}}}$$

Remark 11. Again, we can think of this sum as a bound on the probability of not observing s, a in the blocks if they were independent (the first term) versus a mixing error between blocks to account for their non-independence (the second term).

Proof. Recall that the joint distribution of the observations over the pair of trajectories (m, n) is χ. Its marginals on the segments Ω_i are χ_i. The marginals on each of the G single-step sub-blocks is $\chi_{i,g}$. The product distribution $\prod_g \chi_{i,g}$ is Q_i. Recall that $\mathcal{G}(n, s, a)$ denotes the two sets of indices where (s, a) is observed in trajectory n and m respectively, and the sets have sizes $c_{n,i}$ and $c_{m,i}$ respectively.

Remember that $\mathbf{w}_{n,i,g}$ is the one hot vector of the next state if the (i, g) sub-block witnesses (s, a), and the zero vector otherwise, and that $c_{n,i,g}$ is the indicator of (s, a) in the (i, g) sub-block. Also recall that then $\mathbf{w}_{n,i} = \sum_g \mathbf{w}_{n,i,g}$ and $c_{n,i} = \sum_g c_{n,i,g}$.

Define $h := \prod_{g=1}^{G}(1 - c_{n,i,g})$. Under any distribution Q over these sub-blocks, $\mathbb{E}_Q h$ is the probability of not observing s, a in any of them. Let $d_{i,g,n}$ be the distribution of state-action pairs at the first observation of sub-block (i, g). Let $d_{k_n}(\cdot,\cdot)$ be the stationary distribution under label k_n for state-action pairs. We use Lemma 2 with h as above, $C = 1$, $n = G$ and $a_n = \frac{T_n}{8G}$ to note the following chain of inequalities.

$$\mathbb{P}(c_{n,i} = 0) = \mathbb{E}_{\chi_i} h$$

$$
\begin{aligned}
&\leq \mathbb{E}_{Q_i} h + |\mathbb{E}_{Q_i} h - \mathbb{E}_{\chi_i} h| \\
&\leq \left(\prod_{g=1}^{G} \mathbb{E}_{Q_i}(1 - c_{n,i,g}) \right) + 4G\lambda^{\frac{T_n}{8G}} \\
&\leq \left(\prod_{g=1}^{G} (1 - d_{k_n}(s,a) + TV(d_{i,g,n}, d_{k_n}) \right) + 4G\lambda^{\frac{T_n}{8G}} \\
&\leq \left(\prod_{g=1}^{G} (1 - d_{k_n}(s,a) + 4\lambda^{\frac{T_n}{8G}}) \right) + 4G\lambda^{\frac{T_n}{8G}} \\
&= \left(1 - d_{k_n}(s,a) + 4\lambda^{\frac{T_n}{8G}} \right)^G + 4G\lambda^{\frac{T_n}{8G}} \\
&\leq \left(1 - \frac{d_{k_n}(s,a)}{2} \right)^G + 4G\lambda^{\frac{T_n}{8G}} \\
&\leq \left(1 - \frac{d_{min}(s,a)}{2} \right)^G + 4G\lambda^{\frac{T_n}{8G}}
\end{aligned}
$$

where the inequality in the second to last line holds for $T_n \geq \Omega\left(Gt_{mix}\log(1/\alpha)\right) \geq \Omega\left(Gt_{mix}\log(1/d_{min}(s,a))\right)$.

□

From the above lemma, the following corollary immediately follows by getting conditions to bound each term on the right hand side by $\delta/2$, upon also noting that $-\log(1-x) \geq x$, so $\log\left(\frac{1}{1-\alpha/2}\right) \geq \alpha/2$.

Corollary 1. *For $T_n \geq \Omega\left(Gt_{mix}\log(G/\delta)\log(1/\alpha)\right)$ and $G \geq \Omega\left(\frac{\log(1/\delta)}{\alpha}\right)$, we have with probability at least $1-\delta$ that*

$$
4\left(\max_i \mathbb{1}_{c_{n,i}=0} + \max_i \mathbb{1}_{c_{m,i}=0}\right) = 0
$$

F.1.4. COMBINING THE BOUNDS

We finally combine these lemmas to prove Lemma 4 – the lemma that this section was dedicated to. The conditions of the lemmas combine to ask that $T_n \geq \Omega\left(Gt_{mix}\log(G/\delta)\log(1/\alpha)\right)$ and $G \geq \Omega\left(\frac{\log(1/\delta)}{\alpha^2}\right)$.

Proof of Lemma 4. Combining the decomposition from Lemma 5 with the bounds in Lemma 6 and Corollary 1, we conclude using union bounds on the low probability events that we are excluding that there is a universal constant C_1 so that with probability at least $1-\delta$,

$$
\left|\mathrm{dist}_{1,(s,a)} - \|\Delta_{m,n}(s,a)\|_2^2\right| \leq C_1\left(\sqrt{\frac{K+\log(1/\delta)}{G\alpha}}\right) + 4\epsilon_{sub}(\delta/2)
$$

whenever $T_n \geq \Omega\left(Gt_{mix}\log(G/\delta)\log(1/\alpha)\right)$ and $G \geq \Omega\left(\frac{\log(1/\delta)}{\alpha^2}\right)$.

□

G. Guarantees for one step of the EM Algorithm for mixtures of MDPs

Remember that the M-step is just the model estimation step, so Theorem 4 provides guarantees for that. We also have the following guarantees for the E-step of hard EM.

Theorem 6. *Consider any (s, a) with $d_{min}(s, a) \geq \alpha/3$ where model estimation accuracy is ϵ with $\epsilon \leq \min(\Delta/4, \Delta^2 g_{min}/64)$ where g_{min} is the least non-zero value of $\mathbb{P}_k(s' \mid s, a)$ across k, s'. Using log-likelihood ratios of transitions of all such (s, a) pairs, we can classify any set of N new trajectories with probability $1 - \delta$ if it has length $T_n = \Omega(t_{mix} \log^4(N/\delta) \log^3(1/f_{min})/\alpha^3\Delta^3)$.*

Remark 12. The dependence on g_{min} is unavoidable. For example, if the estimate for the models was only off at the value of k, s' attaining g_{min} and our estimate for g_{min} was $\hat{\mathbb{P}}_k(s' \mid s, a) = 0$, then no trajectory from label k witnessing s' will get correctly classified. This event will happen roughly with probability g_{min}, up to a mixing error, and g_{min} cannot be made less than some arbitrary δ chosen to bound the probability of all undesirable events.

Proof. We are inspired by the lower bound obtained in Lemma 1 of Wong & Shen (1995) for obtaining our sample complexity bounds. Consider a separating state-action pair s, a. We first establish Hellinger distance lower bounds between the distributions $\hat{\mathbb{P}}_k(\cdot \mid s, a)$ and $\hat{\mathbb{P}}_l(\cdot \mid s, a)$. Notice that

$$TV(\hat{\mathbb{P}}_k(\cdot \mid s, a), \mathbb{P}_k(\cdot \mid s, a)) = \frac{1}{2}\|\hat{\mathbb{P}}_k(\cdot \mid s, a) - \mathbb{P}_k(\cdot \mid s, a)\|_1 \leq \epsilon/2 \leq \Delta/4$$

The same holds for l as well. Combining the latter with $\|\mathbb{P}_k(\cdot \mid s, a) - \mathbb{P}_l(\cdot \mid s, a)\|_1 \geq \|\mathbb{P}_k(\cdot \mid s, a) - \mathbb{P}_l(\cdot \mid s, a)\|_2 \geq \Delta$ and using the inequality $H(P, Q) \geq TV(P, Q)/\sqrt{2}$, we get the following bound.

$$H(\mathbb{P}_k(\cdot \mid s, a), \hat{\mathbb{P}}_l(\cdot \mid s, a)) \geq \frac{1}{\sqrt{2}} TV(\hat{\mathbb{P}}_k(\cdot \mid s, a), \hat{\mathbb{P}}_l(\cdot \mid s, a)) \geq \frac{\Delta}{4\sqrt{2}}$$

We now recall notation from the previous section. Again, we modify notation slightly, in a natural way. Let χ_n be the joint distribution of observations recorded in trajectory n, with their marginals on each single-element sub-block being $\chi_{n,g}$. Let Q_n be the product distribution $Q_n = \prod_{n,g} \chi_{n,g}$. Let $\mathcal{G}(n, s, a)$ be the set of sub-blocks (n, g) in which (s, a) is observed in trajectory n. Let c_n be the size of this set. We have the following lemma.

Lemma 8. *Let the random variables for the next states following each (s, a) observation given by $S_1, S_2, \ldots S_{c_n}$ and let the true label be $k_n = k$. Then for any $l \neq k$, consider the likelihood ratio over next state transitions from (s, a).*

$$LR_n(s, a) = \prod_{i=1}^{c_n} \frac{\hat{\mathbb{P}}_k(S_i \mid s, a)}{\hat{\mathbb{P}}_l(S_i \mid s, a)}$$

We claim that $LR_n(s, a) > 0$ with probability at least $1 - \delta$ for $T_n \geq \Omega\left(G t_{mix} \log\left(\frac{G}{\delta}\right) \log(1/\alpha)\right)$ and $G \geq \Omega\left(\frac{\log(1/f_{min}) \log(1/\delta)}{\alpha^2 \Delta^2}\right)$.

Just like in the proof of Theorem 3, now set $G = \left(\frac{T_n}{t_{mix}}\right)^{\frac{2}{3}}$. Then a sufficient condition on T_n to meet the conditions of the lemma is $T_n = \Omega(t_{mix} \log^4(1/\delta) \log^3(1/f_{min})/\alpha^3\Delta^3)$.

Now remember that upon choosing an occurrence threshold β of order α, we will have at most $O(1/\alpha)$ many (s, a) pairs in Freq_β. By applying a union bound over all (s, a) pairs in Freq_β, we get that with probability $1 - \delta$, we get that the sum of the log-likelihood ratios of next-state transitions starting in Freq_β between the true label's model estimate and any other label's model estimate is positive whenever $T_n = \Omega(t_{mix} \log^4(1/\delta) \log^3(1/f_{min})/\alpha^3\Delta^3)$.

We now take another union bound over the N new trajectories to get that we can exactly classify all of them with probability at least $1 - \delta$ whenever $T_n \geq \Omega(t_{mix} \log^4(N/\delta) \log^3(1/f_{min})/\alpha^3\Delta^3)$.

G.1. Proof of Lemma 8

We first perform a computation analogous to Lemma 1 in Wong & Shen (1995). Let $D_1 = \mathbb{P}_k(\cdot \mid s,a)$, $D_2 = \mathbb{P}_l(\cdot \mid s,a)$, $\hat{D}_1 = \hat{\mathbb{P}}_k(\cdot \mid s,a)$, $\hat{D}_2 = \hat{\mathbb{P}}_l(\cdot \mid s,a)$. Fix $b > 0$. We use the conditional Markov inequality and the fact that conditioned on $\mathcal{G}(n,s,a)$ and under the product distribution $\hat{Q}_n$, the Hellinger distance between the next-state distributions at any (s,a) observation is $H(\hat{D}_1, \hat{D}_2)$, which satisfies $H(\hat{D}_1, \hat{D}_2) \geq \Delta/4\sqrt{2}$. This is crucially due to the independence and the fact that we are fixing $\mathcal{G}(n,s,a)$ by conditioning on it. As usual, abbreviate $\mathcal{G}(n,s,a)$ to $\mathcal{G}$ for brevity.

$$
\begin{aligned}
\mathbb{P}_{Q_n}(LR_n(s,a) \leq e^{c_n b/2} \mid \mathcal{G}) &= \mathbb{P}_{Q_n}\left(\prod_{i=1}^{c_n}\left(\frac{\hat{D}_2(S_i)}{\hat{D}_1(S_i)}\right)^{1/2} \geq e^{-c_n b/2}\middle|\mathcal{G}\right) \\
&\leq e^{c_n b/2}\left(\mathbb{E}_{Q_n}\left[\left(\frac{\hat{D}_2(S_i)}{\hat{D}_1(S_i)}\right)^{1/2}\middle|\mathcal{G}\right]\right)^{c_n} \\
&= e^{c_n b/2}\left(\mathbb{E}_{D_1}\left[\left(\frac{\hat{D}_2(S_i)}{\hat{D}_1(S_i)}\right)^{1/2}\right]\right)^{c_n} \\
&= e^{c_n b/2}\left(\mathbb{E}_{D_1}\left[\left(\frac{D_1(S_i)}{\hat{D}_1(S_i)}\right)^{1/2}\left(\frac{\hat{D}_2(S_i)}{D_1(S_i)}\right)^{1/2}\right]\right)^{c_n} \\
&\leq e^{c_n b/2}\left(\mathbb{E}_{D_1}\left[\left(1+\Delta^2/64\right)^{1/2}\left(\frac{\hat{D}_2(S_i)}{D_1(S_i)}\right)^{1/2}\right]\right)^{c_n} \\
&= e^{c_n b/2}\left(1+\Delta^2/64\right)^{c_n/2}\left(1-\frac{H(D_1,D_2)^2}{2}\right)^{c_n} \\
&\leq e^{c_n b/2}\left(1-\Delta^2/128\right)^{c_n/2} \\
&\leq e^{c_n b/2} e^{-c_n \Delta^2/128}
\end{aligned}
$$

Setting $b = \Delta^2/256$, we get that $\mathbb{P}_{Q_n}(LR_n(s,a) \leq e^{c_n\Delta^2/256} \mid \mathcal{G}) \leq e^{-c_n\Delta^2/256}$. Now by following a very similar computation to that in point 2 in section F.1.2, we get that for $T_n \geq \Omega(Gt_{mix}\log(1/\alpha))$ and $G \geq \Omega\left(\frac{\log(1/\delta)}{\alpha^2}\right)$, $c_n \geq Gd_{min}(s,a)/4$ with probability at least $1-\delta/2$. That is, for such T_n and G,

$$
\mathbb{P}_{Q_n}(LR_n(s,a) \leq e^{Gd_{min}(s,a)\Delta^2/512} \mid \mathcal{G}) \leq \mathbb{P}_{Q_n}(LR_n(s,a) \leq e^{c_n\Delta^2/128} \mid \mathcal{G}) \leq e^{-Gd_{min}(s,a)\Delta^2/512} + \frac{\delta}{2}
$$

Since this holds for any value of $\mathcal{G} = \mathcal{G}(n,s,a)$, we can say that with probability at least $1-\delta$, for $T_n \geq \Omega(Gt_{mix}\log(1/\alpha))$ and $G \geq \Omega\left(\frac{\log(1/\delta)}{\alpha^2}\right)$, $c_n \geq Gd_{min}(s,a)/4$, we have the following bound.

$$
\mathbb{P}_{Q_n}(LR_n(s,a) \leq e^{Gd_{min}(s,a)\Delta^2/512}) \leq e^{-Gd_{min}(s,a)\Delta^2/512} + \frac{\delta}{2}
$$

After following a computation very similar to that in point 3 of section F.1.2, we get that for $T_n \geq \Omega\left(Gt_{mix}\log\left(\frac{G}{\delta}\right)\log(1/\alpha)\right)$ and $G \geq \Omega\left(\frac{\log(1/\delta)}{\alpha^2\Delta^2}\right)$,

$$
\mathbb{P}_{\chi}(LR_n(s,a) \leq e^{Gd_{min}(s,a)\Delta^2/512}) \leq \delta
$$

Note that we want $e^{Gd_{min}(s,a)\Delta^2/512} \geq f_l/f_k$, in which case it suffices to ask $e^{Gd_{min}(s,a)\Delta^2/512} \geq 1/f_{min}$. Combining this with earlier conditions, for $G \geq \Omega\left(\frac{\log(1/\delta)\log(1/f_{min})}{\alpha^2\Delta^2}\right)$ and $T_n \geq \Omega\left(Gt_{mix}\log\left(\frac{G}{\delta}\right)\log(1/\alpha)\right)$,

$$\mathbb{P}_{\chi}\left(\frac{f_k}{f_l} LR_n(s,a) \leq 1\right) \leq \delta$$

□

H. Proof of Theorem 4

Theorem 4 (Model Estimation Guarantee)**.** *For any state action pair (s,a) with $d_{min}(s,a) \geq \alpha/3$, and for $GN_{clust} \geq \Omega\left(\frac{\log(1/\delta)}{f_{min}^2 \alpha^2}\right)$ and $T_n \geq \Omega(Gt_{mix}\log(G/\delta))$, with probability greater than $1-\delta$,*

$$\|\hat{\mathbb{P}}_k(\cdot \mid s,a) - \mathbb{P}_k(\cdot \mid s,a)\|_1$$

is bounded above by

$$O\left(\left(\frac{t_{mix}}{T_n}\right)^{1/3}\sqrt{\frac{1}{N_{clust} f_{min}\alpha}(S+\log(\frac{1}{\delta}))}\right)$$

Proof. The proof is quite straightforward and employs the techniques used so far, especially those used in section F.1.2. Let k be the (now known) label that we're working with.

We modify previous notation a bit for this proof. For brevity of notation, we denote by $c_{n,g}$ the indicator variable for observing (s,a) in the gth single-step sub-block of the trajectory n. Denote by $\mathbf{w}_{n,g}$ one-hot vector of the next state observed if the currect state-action pair is (s,a), and set it to the zero-vector otherwise. Note that $\sum_g c_{n,g} = N(n,s,a)$ and $\sum_g \mathbf{w}_{n,g} = \mathbf{N}(n,s,a,\cdot)$. We denote the set of indices (n,g) of all s,a observations that come from label k (across the GN_{clust} observations recorded) by $\mathcal{N}(s,a,k)$. Let the size of this set be $N(s,a,k)$. Note that $N(s,a,k) = \sum_{n\in\mathcal{C}_k} N(n,s,a) = \sum_{n,g} c_{n,g}$. Also note the following alternate expression for $\hat{\mathbb{P}}_k(\cdot \mid s,a)$.

$$\hat{\mathbb{P}}_k(\cdot \mid s,a) := \frac{\sum_{(n,g)\in\mathcal{N}(s,a,k)} \mathbf{w}_{n,g}}{\sum_{(n,g)\in\mathcal{N}(s,a,k)} c_{n,g}}\mathbb{1}_{N(s,a,k)\neq 0} = \frac{\sum_{(n,g)\in\mathcal{N}(s,a,k)} \mathbf{w}_{n,g}}{N(s,a,k)}\mathbb{1}_{N(s,a,k)\neq 0} \tag{18}$$

Let χ_n be the joint distribution of observations recorded in trajectory n, with their marginals on each single-element sub-block being $\chi_{n,g}$. Let χ be the joint distribution of all observations recorded across all trajectories. Since the trajectories are independent, we know that $\chi = \prod_n \chi_n$. Let Q_g be the joint distribution of the observations at the gth sub-block. Note that this is also the marginal of the joint distribution χ on the gth sub-block, and since the trajectories are independent, $Q_g = \prod_n \chi_{g,n}$. Finally, denote by Q the product distribution $\prod_g Q_g = \prod_g \prod_n \chi_{g,n}$. This would be the distribution if all observations recorded were independent (across sub-blocks).

1. Concentration under the product distribution

We have the following computation.

$$\begin{aligned}
\mathbb{E}_Q[\hat{\mathbb{P}}_k(\cdot \mid s,a) \mid \mathcal{N}(s,a,k)] &= \mathbb{E}_Q\left[\frac{\sum_{n\in\mathcal{N}_{clust}} \mathbf{w}_n}{N(s,a,k)}\mathbb{1}_{N(s,a,k)\neq 0}\middle| N(s,a,k)\right] \\
&= \mathbb{E}\left[\frac{\sum_{n\in\mathcal{N}(s,a,k)} \mathbf{w}_n}{N(s,a,k)}\middle| N(s,a,k)\right]\mathbb{1}_{N(s,a,k)\neq 0} \\
&= \frac{\sum_n \mathbb{E}_Q[\mathbf{w}_n \mid \mathcal{N}(s,a,k)]}{N(s,a,k)}\mathbb{1}_{N(s,a,k)\neq 0} \\
&= \frac{\sum_n \mathbb{P}_k(\cdot \mid s,a)c_n}{N(s,a,k)}\mathbb{1}_{N(s,a,k)\neq 0} \\
&= \frac{\mathbb{P}_k(\cdot \mid s,a)(\sum_n c_n)}{N(s,a,k)}\mathbb{1}_{N(s,a,k)\neq 0} \\
&= \frac{\mathbb{P}_k(\cdot \mid s,a)N(s,a,k)}{N(s,a,k)}\mathbb{1}_{N(s,a,k)\neq 0} \\
&= \mathbb{P}_k(\cdot \mid s,a)\mathbb{1}_{N(s,a,k)\neq 0}
\end{aligned}$$

Now we set up our covering argument. Remember that $[-1,1]^S$ is the set of all vectors $\mathbf{u} \in \mathcal{R}^S$ with $\|u\|_\infty \leq 1$. Consider a covering of $[-1,1]^S$ by boxes of side length $\frac{1}{4}$ and centers lying in $[-1,1]^S$. We will need at most 12^S such boxes and if C is the set of their centers, then for any vector $\mathbf{v}$

$$\|\mathbf{v}\|_1 = \sup_{\mathbf{u}\in[-1,1]^{S-1}} |\mathbf{u}^T\mathbf{v}| \leq 2\max_{\mathbf{u}\in C}|\mathbf{u}^T\mathbf{v}| \leq 2\|\mathbf{v}\|_1$$

Also, for any $\mathbf{u} \in C$, note that

$$\begin{aligned}|\mathbf{u}^T \hat{\mathbb{P}}_{n,i}(\cdot \mid s,a)| &\leq \|\mathbf{u}\|_\infty \left\| \frac{\mathbf{w}_{n,1}}{c_{n,1}} \right\|_1 \\ &\leq \left\| \frac{\mathbf{w}_{n,1}}{c_{n,1}} \right\|_1 \\ &= 1\end{aligned}$$

and so $|\mathbf{u}^T \mathbb{E}_Q[\hat{\mathbb{P}}_k(\cdot \mid s,a) \mid \mathcal{N}(s,a,k)]| \leq \mathbb{E}[|\mathbf{u}^T \hat{\mathbb{P}}_k(\cdot \mid s,a)| \mid \mathcal{N}(s,a,k)] \leq 1$. Again, note that conditioned on the set of all (s,a) observations recorded, the next states $\mathbf{w}_{n,g}$ are all independent under the product distribution Q (but not under χ, of course). Recalling the expression for $\hat{\mathbb{P}}_k(\cdot \mid s,a)$ from equation 18, this means that we can use the conditional version of Hoeffding's inequality, giving us the following bound.

$$\mathbb{P}_Q\left(\left|\mathbf{u}^T(\hat{\mathbb{P}}_k(\cdot \mid s,a) - \mathbb{E}_Q[\hat{\mathbb{P}}_k(\cdot \mid s,a) \mid \mathcal{N}(s,a,k)])\right| > \frac{\epsilon}{4} \middle| \mathcal{N}(s,a,k)\right) < 2e^{-\frac{\epsilon^2 N(s,a,k)}{8}}$$

Doing this for all 12^S vectors $\mathbf{u} \in C$, we get the following inequality.

$$\mathbb{P}_Q\left(\left\|(\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) - \mathbb{E}_Q[\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) \mid \mathcal{N}(s,a,k)])\right\|_1 > \frac{\epsilon}{2} \middle| \mathcal{N}(s,a,k)\right)$$

is bounded above by

$$\begin{aligned}&\mathbb{P}_Q\left(\exists \mathbf{u} \in C; \left|\mathbf{u}^T(\hat{\mathbb{P}}_k(\cdot \mid s,a) - \mathbb{E}_Q[\hat{\mathbb{P}}_k(\cdot \mid s,a) \mid \mathcal{N}(s,a,k)])\right| > \frac{\epsilon}{4} \middle| \mathcal{N}(s,a,k)\right) \\ &\leq \sum_{\mathbf{u} \in C} \mathbb{P}_Q\left(\left|\mathbf{u}^T(\hat{\mathbb{P}}_k(\cdot \mid s,a) - \mathbb{E}_Q[\hat{\mathbb{P}}_k(\cdot \mid s,a) \mid \mathcal{N}(s,a,k)])\right| > \frac{\epsilon}{4} \middle| \mathcal{N}(s,a,k)\right) \\ &< 12^S * e^{-\frac{\epsilon^2 N(s,a,a,k)}{8}}\end{aligned}$$

2. Bounding $N(s,a,k)$ under the product distribution

Now note that $N(s,a,k) = \sum_{(n,g) \in \mathcal{N}_{clust} \times [G]} c_{n,g}$. So,

$$\mathbb{E}_Q[N(s,a,k)] = \sum_{(n,g) \in \mathcal{N}_{clust} \times [G]} \mathbb{E}_Q[c_{n,g}] = \sum_{(n,g) \in \mathcal{N}_{clust} \times [G]} \mathbb{P}_\chi(c_{n,g} \neq 0)$$

We can show the following inequality.

$$\mathbb{P}_\chi(c_{n,g} \neq 0) = \mathbb{P}_\chi(c_{n,g} \neq 0 \mid k_n = k)\mathbb{P}(k_n = k) \geq \frac{d_{min}(s,a)}{2} f_{min}$$

for $T_n \geq \Omega(G t_{mix} \log(1/\alpha))$, getting the last inequality by using a computation very similar to the one in equation 4, along with the fact that $\mathbb{P}(k_n = k) = f_k$. So, $\mathbb{E}_Q[N(s,a,k)] \geq \frac{G N_{clust} f_{min} d_{min}(s,a)}{2}$.

$$\begin{aligned}\mathbb{P}_Q\left(N(s,a,k) < G N_{clust} \frac{f_{min} d_{min}(s,a)}{4}\right) &= \mathbb{P}_Q\left(N(s,a,k) < G N_{clust} \frac{f_{min} d_{min}(s,a)}{2} - G N_{clust} \frac{f_{min} d_{min}(s,a)}{4}\right) \\ &\leq \mathbb{P}_Q\left(N(s,a,k) < \mathbb{E}[N(s,a,k)] - G N_{clust} \frac{f_{min} d_{min}(s,a)}{4}\right) \\ &= \mathbb{P}_Q\left(\sum_{(n,g) \in \mathcal{N}_{clust} \times [G]} c_{n,g} < \mathbb{E}[N(s,a,k)] - G N_{clust} \frac{f_{min} d_{min}(s,a)}{4}\right) \\ &\leq \exp\left(-\frac{f_{min}^2 d_{min}(s,a)^2 G N_{clust}}{8}\right)\end{aligned}$$

This is less than $\delta/2$ for $GN_{clust} \geq \Omega\left(\frac{\log(1/\delta)}{f_{min}^2\alpha^2}\right)$. So, with probability at least $1-\delta/2$, for $GN_{clust} \geq \Omega\left(\frac{\log(1/\delta)}{f_{min}^2\alpha^2}\right)$ and $T_n \geq \Omega(Gt_{mix}\log(1/\alpha))$, we have the following bound.

$$\mathbb{P}_Q\left(\left\|(\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) - \mathbb{E}_Q[\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) \mid \mathcal{N}(s,a,k)])\right\|_1 > \frac{\epsilon}{2}\right) \leq 12^S e^{-\frac{\epsilon^2 GN_{clust} f_{min} d_{min}(s,a)}{128}}$$

3. Mixing error to account for non-independence in the true joint distribution

Note that we can think of the combined dataset as a Markov chain over the tuple of n observations, with a joint distribution χ over observations. Its marginal over the gth single-step sub-blocks is Q_g and $Q = \prod_g Q_g$. We now want to apply Lemma 2, noting that the relevant function of this Markov chain is $\mathbb{1}_E$ where E is the event $\|\hat{\mathbb{P}}_k(\cdot \mid s,a) - \mathbb{P}_k(\cdot \mid s,a)\|_1 < \frac{\epsilon}{2}$. Clearly, in this case, n from the lemma is G and C from the lemma is 1. We use this to get the following bound.

$$\mathbb{P}_\chi\left(\left\|(\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) - \mathbb{E}_Q[\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) \mid \mathcal{N}(s,a,k)])\right\|_1 > \frac{\epsilon}{2}\right)$$

is bounded above by

$$\mathbb{P}_Q\left(\left\|(\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) - \mathbb{E}_Q[\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) \mid \mathcal{N}(s,a,k)])\right\|_1 > \frac{\epsilon}{2}\right) + 4G\left(\frac{1}{4}\right)^{\frac{T_n}{8Gt_{mix}}}$$
$$\leq 12^S e^{-\frac{\epsilon^2 GN_{clust} f_{min} d_{min}(s,a)}{128}} + 4G\left(\frac{1}{4}\right)^{\frac{T_n}{8Gt_{mix}}}$$

Each term is less than $\delta/4$ for $GN_{clust} \geq \Omega\left(\frac{1}{\epsilon^2 f_{min}\alpha}(S+\log(\frac{1}{\delta}))\right)$ and $T_n \geq \Omega(Gt_{mix}\log(G/\delta))$. So for such G, N_{clust}, T_n, with probability greater than $1-\delta$,

$$\|\hat{\mathbb{P}}_k(\cdot \mid s,a) - \mathbb{P}_k(\cdot \mid s,a)\|_1 < \epsilon$$

Alternatively, for $GN_{clust} \geq \Omega\left(\frac{\log(1/\delta)}{f_{min}^2\alpha^2}\right)$ and $T_n \geq \Omega(Gt_{mix}\log(G/\delta)\log(1/\alpha))$, with probability greater than $1-\delta$,

$$\|\hat{\mathbb{P}}_k(\cdot \mid s,a) - \mathbb{P}_k(\cdot \mid s,a)\|_1 \leq O\left(\sqrt{\frac{1}{GN_{clust} f_{min}\alpha}(S+\log(\frac{1}{\delta}))}\right)$$

Letting $G = \left(\frac{T_n}{t_{mix}}\right)^{2/3}$, for $\left(\frac{T_n}{t_{mix}}\right)^{2/3} N_{clust} \geq \Omega\left(\frac{\log(1/\delta)}{f_{min}^2\alpha^2}\right)$ and $T_n \geq \Omega(t_{mix}\log^4(1/\delta)\log^4(1/\alpha))$, with probability greater than $1-\delta$,

$$\|\hat{\mathbb{P}}_k(\cdot \mid s,a) - \mathbb{P}_k(\cdot \mid s,a)\|_1 \leq O\left(\left(\frac{t_{mix}}{T_n}\right)^{1/3}\sqrt{\frac{1}{N_{clust} f_{min}\alpha}(S+\log(\frac{1}{\delta}))}\right)$$

□

I. Proof of Theorem 5

We recall the theorem here.

Theorem 5 (Classification Guarantee). *Let $\epsilon_{mod}(\delta)$ be a high probability bound on the model estimation error $\|\hat{\mathbb{P}}_k(\cdot \mid s,a) - \mathbb{P}_k(\cdot \mid s,a)\|_2$. Then there is a universal constant C_3 so that Algorithm 3 can identify the true labels for trajectories in $\mathcal{N}_{class}$ with probability at least $1-\delta$ for $T_n = \Omega\left(K^{3/2} t_{mix} \frac{\log^4(N_{class}/(\alpha\delta))}{\Delta^6 \alpha^3}\right)$, whenever $\epsilon_{mod}(\delta/2) \leq \frac{C_3 \Delta^4 f_{min}\alpha}{K}$ and $N_{clust} \geq \Omega\left(\frac{\log(1/\delta)}{f_{min}^2 \alpha^2}\right)$.*

Proof. The proof is very similar to the proof of theorem 3. Consider the testing of trajectory n. Recall that in algorithm 3, we defined

$$\text{dist}_1(n,k) := \max_{(s,a)\in SA_\alpha} \left[\left(\left(\hat{\mathbb{P}}_{n,1}(\cdot \mid s,a) - \hat{\mathbb{P}}_k(\cdot \mid s,a) \right)^T \tilde{\mathbf{V}}_{s,a} \right) \left(\left(\hat{\mathbb{P}}_{n,2}(\cdot \mid s,a) - \hat{\mathbb{P}}_k(\cdot \mid s,a) \right)^T \tilde{\mathbf{V}}_{s,a} \right)^T \right]$$

Let k_n the label of trajectory n. According to our assumptions, if $k_n \neq k$, then we have an s,a so that $d_{k_n}(s,a) \geq \alpha$ and $\|\mathbb{P}_{k_n}(\cdot \mid s,a) - \mathbb{P}_k(\cdot \mid s,a)\|_2 \geq \Delta$. Again, we will make s,a implicit in our notation except in $\mathbb{P}_j(\cdot \mid s,a)$. Let $c_{n,i} := N(n,i,s,a)$, $\mathbf{w}_{n,i} := \mathbf{N}(n,i,s,a,\cdot)$. Recall that we have two nested partitions: (1) of the entire trajectory into the two Ω_i and (2) of each segment Ω_i into G blocks. Finally, define $\text{dist}_{1,(s,a)}$ as below, suppressing n and k. Note that $\text{dist}_1(n,k)$ is the maximum of $\text{dist}_{1,(s,a)}$ over all $(s,a) \in \text{Freq}_\beta$, for the given trajectory n and label k.

$$\text{dist}_{1,(s,a)} := \left[\left((\hat{\mathbb{P}}_{n,1}(\cdot \mid s,a) - \hat{\mathbb{P}}_k(\cdot \mid s,a))^T \tilde{\mathbf{V}}_{s,a} \right) \left((\hat{\mathbb{P}}_{n,2}(\cdot \mid s,a) - \hat{\mathbb{P}}_k(\cdot \mid s,a))^T \tilde{\mathbf{V}}_{s,a} \right)^T \right]$$

We want to show that this is close to $\|\Delta_{n,k}(s,a)\|_2^2$ for the (s,a) pairs that we search over, where

$$\Delta_{n,k}(s,a) = \mathbb{P}_{k_n}(\cdot \mid s,a) - \mathbb{P}_k(\cdot \mid s,a)$$

Recall that $\|\hat{\mathbb{P}}_k(\cdot \mid s,a) - \mathbb{P}_k(\cdot \mid s,a)\|_2 \leq \epsilon_{mod}(\delta)$ for any $1 \leq k \leq K$. Let $\mathbf{M}_{s,a}^{true} = \sum_{1\leq k\leq K} \hat{f}_{k,s,a} \mathbb{P}_k(\cdot \mid s,a)\mathbb{P}_k(\cdot \mid s,a)^T$. We use the fact that $\|aa^T - bb^T\| \leq (\|a\|_2 + \|b\|_2)\|a-b\|_2$ in the bound below.

$$\begin{aligned}
\|\mathbf{M}_{s,a}^{true} - \tilde{\mathbf{M}}_{s,a}\| &\leq \sum_{1\leq k\leq K} \hat{f}_{k,s,a} \|\mathbb{P}_k(\cdot \mid s,a)\mathbb{P}_k(\cdot \mid s,a)^T - \hat{\mathbb{P}}_k(\cdot \mid s,a)\hat{\mathbb{P}}_k(\cdot \mid s,a)^T\| \\
&\leq \sum_{1\leq k\leq K} \hat{f}_{k,s,a} (\|\hat{\mathbb{P}}_k(\cdot \mid s,a)\|_2 + \|\mathbb{P}_k(\cdot \mid s,a)\|_2)\|\hat{\mathbb{P}}_k(\cdot \mid s,a) - \mathbb{P}_k(\cdot \mid s,a)\|_2 \\
&\leq \sum_{1\leq k\leq K} 2\hat{f}_{k,s,a} \|\hat{\mathbb{P}}_k(\cdot \mid s,a) - \mathbb{P}_k(\cdot \mid s,a)\|_2 \\
&\leq 2\epsilon_{mod}(\delta)
\end{aligned}$$

Also note that if we redefine $\mathcal{B}_n$ to be the event of observing (s,a) in a trajectory (instead of in both segments as in the notation in previous proofs), then $\hat{f}_{k,s,a} = \frac{\sum_n \mathbb{1}_{k_n=k}\mathbb{1}_{\mathcal{B}_n}}{\sum_n \mathbb{1}_{\mathcal{B}_n}} \geq \frac{\sum_n \mathbb{1}_{k_n=k}\mathbb{1}_{\mathcal{B}_n}}{N_{clust}}$. So, $\mathbb{E}[\hat{f}_{k,s,a}] \geq \mathbb{P}(k_n = k \cap \mathcal{B}_n) = \mathbb{P}(\mathcal{B}_n \mid k_n = k)\mathbb{P}(k_n = k) \geq f_{min}\mathbb{P}(\mathcal{B}_n \mid k_n = k)$. Using a computation very similar to the one leading up to inequality 5, we note that $\mathbb{P}(\mathcal{B}_n \mid k_n = k) \geq d_{min}(s,a)/2$ for $T_n \geq \Omega(t_{mix}\log(1/\alpha))$. In that case, $\mathbb{E}[\hat{f}_{k,s,a}] \geq f_{min} d_{min}(s,a)/2 \geq f_{min}\alpha/2$. Additionally, using a standard concentration argument, $\hat{f}_{k,s,a} \geq \mathbb{E}[\hat{f}_{k,s,a}]/2 \geq f_{min}\alpha/4$ for $N_{clust} \geq \Omega\left(\frac{\log(1/\delta)}{f_{min}^2\alpha^2}\right) \geq \Omega\left(\frac{\log(1/\delta)}{\mathbb{E}[\hat{f}_{k,s,a}]^2}\right)$

We now apply Lemma 3 of Chen & Poor (2022), with $p^{(k)} = \hat{f}_{k,s,a}$, $\mathbf{y}^{(k)} = \mathbb{P}_k(\cdot \mid s,a)$, $\mathbf{M} = \mathbf{M}_{s,a}^{true}$ and $\mathbf{M}_* = \mathbf{M}_{s,a}$. We use the right-hand side of the bound in the lemma to get the bound below for all $1 \leq k \leq K$, which holds for a universal constant C_2 with probability at least $1-\delta$ whenever $N_{clust} \geq \Omega\left(\frac{\log(1/\delta)}{f_{min}^2\alpha^2}\right)$ and $T_n \geq \Omega(t_{mix}\log(1/\alpha))$.

$$\|\mathbb{P}_k(\cdot \mid s,a) - \tilde{\mathbf{V}}_{s,a}\tilde{\mathbf{V}}_{s,a}^T \mathbb{P}_k(\cdot \mid s,a)\|_2 \leq \sqrt{\frac{2K\epsilon_{mod}(\delta)}{\hat{f}_{k,s,a}}} \leq C_2\sqrt{\frac{K\epsilon_{mod}(\delta)}{f_{min}\alpha}} \tag{19}$$

Assume the lemma below for now, we prove it in the next subsection.

Lemma 9. *We claim that there is a universal constant C_1 so that for any (s,a) with $d_{min}(s,a) \geq \alpha/3$, with probability at least $1-\delta$,*

$$\left|\text{dist}_{1,(s,a)} - \|\Delta_{n,k}(s,a)\|_2^2\right| \leq O\left(\sqrt{\frac{K+\log(1/\delta)}{G\alpha}}\right) + 8C_2\sqrt{\frac{K\epsilon_{mod}(\delta/2)}{f_{min}\alpha}}$$

whenever $T_n \geq \Omega\left(Gt_{mix}\log(G/\delta)\log(1/\alpha)\right)$ and $G \geq \Omega\left(\frac{\log(1/\delta)}{\alpha^2}\right)$. Here, $\epsilon_{mod}(\delta)$ is a high probability bound on $\|\mathbb{P}_k(\cdot \mid s,a) - \tilde{\boldsymbol{V}}_{s,a}\tilde{\boldsymbol{V}}_{s,a}^T\mathbb{P}_k(\cdot \mid s,a)\|_2$ for all $1 \leq k \leq K$ (which holds with probability at least $1-\delta$).

We now set $G = \left(\frac{T_n}{t_{mix}}\right)^{\frac{2}{3}}$. Then a sufficient condition on T_n to meet the conditions of the lemma is $T_n = \Omega(t_{mix}\log^4(1/\delta)/\alpha^3)$, under which, with probability at lest $1-\delta$, we have the following bound for (s,a) with $d_{min}(s,a) \geq \alpha/3$.

$$\left|\text{dist}_{1,(s,a)} - \|\Delta_{n,k}(s,a)\|_2^2\right| \leq O\left(\sqrt{\frac{K\log(1/\delta)}{\alpha}}\left(\frac{t_{mix}}{T_n}\right)^{\frac{1}{3}}\right) + 8C_2\sqrt{\frac{K\epsilon_{mod}(\delta/2)}{f_{min}\alpha}} \tag{20}$$

It is now easy to see that the first term on the right-hand side is less than $\Delta^2/8$ when $T_n = \Omega\left(K^{3/2}t_{mix}\frac{\log^{3/2}(1/\delta)}{\Delta^6\alpha^{3/2}}\right)$ and $T_n = \Omega(t_{mix}\log^4(1/\delta)/\alpha^3)$. We can combine these to have the guarantee that the first term on the right-hand side is less $\Delta^2/8$ with probability at least $1-\delta$ when $T_n = \Omega\left(K^{3/2}t_{mix}\frac{\log^4(1/\delta)}{\Delta^6\alpha^3}\right)$.

Now note that if $\beta \geq \alpha/3$, then a separating state action pair always lies in Freq_β and thus, the maximum over the $\|\Delta_{n,k}(s,a)\|_2^2$ values corresponding to Freq_β is in fact either 0 if $k = k_n$ or larger than Δ^2 if $k \neq k_n$. So, if $8C_2\sqrt{\frac{K\epsilon_{mod}(\delta/2)}{f_{min}\alpha}} \leq \Delta^2/32$ and for each of the (s,a) pairs, the first term on the right-hand side of inequality 20 is less than $\Delta^2/8$, then our distance estimate $\text{dist}_1(n,k)$ is on the right side of $\Delta^2/3$. That is, the distance estimate is then less than $\Delta^2/4$ if $k = k_n$, and larger than it if $k \neq k_n$. As a consequence, the output of the argmin in algorithm 3 is k_n in this situation.

Note that upon choosing an occurrence threshold of order α, we will have at most $O(1/\alpha)$ many (s,a) pairs in Freq_β to maximize $\text{dist}_{1,(s,a)}$ over to get $\text{dist}_1(n,k)$. By applying a union bound over all (s,a) pairs in Freq_β and using the conclusion of the previous paragraph, algorithm 3 correctly predicts the label k_n for trajectory n with probability $1-\delta$ whenever $T_n = \Omega\left(K^{3/2}t_{mix}\frac{\log^4(1/(\alpha\delta))}{\Delta^6\alpha^3}\right)$ and $8C_2\sqrt{\frac{K\epsilon_{mod}(\delta/2)}{f_{min}\alpha}} \leq \Delta^2/32$.

By applying a union bound over incorrectly predicting k_n for any of the $N_{class}(N_{class}-1)/2$ pairs, we get that algorithm 3 can recover the true labels with probability at least $1-\delta$ for $T_n = \Omega\left(K^{3/2}t_{mix}\frac{\log^4(N_{class}/(\alpha\delta))}{\Delta^6\alpha^3}\right)$, whenever $8C_2\sqrt{\frac{K\epsilon_{mod}(\delta/2)}{f_{min}\alpha}} \leq \Delta^2/32$.

Finally note that due to inequality 19, we get that algorithm 3 can recover the true labels with probability at least $1-\delta$ for $T_n = \Omega\left(K^{3/2}t_{mix}\frac{\log^4(N_{class}/(\alpha\delta))}{\Delta^6\alpha^3}\right)$, whenever $\epsilon_{mod}(\delta/2) \leq \frac{C_3\Delta^4 f_{min}\alpha}{K}$.

□

I.1. Proof of Lemma 9

We recall the lemma here.

Lemma 9. *We claim that there is a universal constant C_1 so that for any (s,a) with $d_{min}(s,a) \geq \alpha/3$, with probability at*

least $1-\delta$,

$$\left|\text{dist}_{1,(s,a)} - \|\Delta_{n,k}(s,a)\|_2^2\right| \leq O\left(\sqrt{\frac{K+\log(1/\delta)}{G\alpha}}\right) + 8C_2\sqrt{\frac{K\epsilon_{mod}(\delta/2)}{f_{min}\alpha}}$$

whenever $T_n \geq \Omega\left(Gt_{mix}\log(G/\delta)\log(1/\alpha)\right)$ *and* $G \geq \Omega\left(\frac{\log(1/\delta)}{\alpha^2}\right)$. *Here,* $\epsilon_{mod}(\delta)$ *is a high probability bound on* $\|\mathbb{P}_k(\cdot \mid s,a) - \tilde{\boldsymbol{V}}_{s,a}\tilde{\boldsymbol{V}}_{s,a}^T\mathbb{P}_k(\cdot \mid s,a)\|_2$ *for all* $1 \leq k \leq K$ *(which holds with probability at least* $1-\delta$*).*

Proof. The proof of this lemma is very similar to the proof of Lemma 4.

Notation: We say $c_{n,i} = N(n,i,s,a)$ as in the statement of the lemma and $\mathbf{w}_{n,i} = \mathbf{N}(n,i,s,a,\cdot)$. Let the joint distribution of the observations over trajectory n be χ. Let its marginals on the segments Ω_i be χ_i. Let the marginals on each of the G single-step sub-blocks along with their next states be $\chi_{i,g}$. Denote the product distribution $\prod_g \chi_{i,g}$ by Q_i. Let $\mathcal{G}(n,s,a)$ denote the set of indices where the state-action pair (s,a) is observed in trajectory n. For brevity, we will abbreviate $\mathcal{G}(n,s,a)$ to $\mathcal{G}$. Note that the size of this set is exactly $c_{n,i}$.

We first prove a preliminary lemma, similar to lemma 5.

I.1.1. Decomposition of $|\,\text{dist}_{1,(s,a)} - \|\Delta_{n,k}(s,a)\|_2^2\,|$

Lemma 10. *We claim that for each fixed value of* $\mathcal{G}(s,a)$ *(abbreviated to* $\mathcal{G}$*), with probability at least* $1-\delta$*, the following bound holds.*

$$\left|\text{dist}_{1,(s,a)} - \|\Delta_{n,k}(s,a)\|_2^2\right| \leq \sum_{i=1}^{2} 2\left\|\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) - \mathbb{E}_{Q_i}[\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) \mid \mathcal{G}]\right\|_2 + 8C_2\sqrt{\frac{K\epsilon_{mod}(\delta)}{f_{min}\alpha}} + 4\left(\max_i \mathbb{1}_{c_{n,i}=0}\right) \tag{21}$$

Here $c_{n,i} = N(n,i,s,a)$ *and* $\epsilon_{mod}(\delta)$ *is a high probability bound on* $\|\mathbb{P}_k(\cdot \mid s,a) - \tilde{\boldsymbol{V}}_{s,a}\tilde{\boldsymbol{V}}_{s,a}^T\mathbb{P}_k(\cdot \mid s,a)\|_2$ *(satisfied with probability* $> 1-\delta$*).*

Remark 13. In the inequality,

- The first term is a concentration-type term, which will be broken into an "independent concentration" error and a mixing error to account for the low but non-zero dependence across blocks.
- The second term accounts for subspace estimation error.
- The third term accounts for actually observing s,a in our blocks.

Proof. Define the following quantities.

$$\boldsymbol{\Delta}_i^T = (\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) - \hat{\mathbb{P}}_k(\cdot \mid s,a))^T\tilde{\mathbf{V}}_{s,a}$$
$$\bar{\boldsymbol{\Delta}}_i^T = (\mathbb{E}_{Q_i}[\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) \mid \mathcal{G}] - \mathbb{P}_k(\cdot \mid s,a))^T\tilde{\mathbf{V}}_{s,a}$$

We first establish a simple inequality, using the fact that $|a^Tb - c^Td| \leq \|b\|_2\|a-c\|_2 + \|c\|_2\|b-d\|_2$

$$\begin{aligned}
|\,\text{dist}_{1,(s,a)} - \bar{\boldsymbol{\Delta}}_1^T\bar{\boldsymbol{\Delta}}_2| &= |\boldsymbol{\Delta}_1^T\boldsymbol{\Delta}_2 - \bar{\boldsymbol{\Delta}}_1^T\bar{\boldsymbol{\Delta}}_2| \\
&\leq \|\boldsymbol{\Delta}_1 - \bar{\boldsymbol{\Delta}}_1\|_2\|\boldsymbol{\Delta}_2\|_2 + \|\bar{\boldsymbol{\Delta}}_1^T\|_2\|\boldsymbol{\Delta}_2 - \bar{\boldsymbol{\Delta}}_2\|_2 \\
&\leq 2\|\boldsymbol{\Delta}_1 - \bar{\boldsymbol{\Delta}}_1\|_2 + 2\|\boldsymbol{\Delta}_2 - \bar{\boldsymbol{\Delta}}_2\|_2 \\
&\leq \sum_{i=1}^{2} 2\|\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) - \mathbb{E}_{Q_i}[\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) \mid \mathcal{G}]\|_2 + \sum_{i=1}^{2} 2\|\hat{\mathbb{P}}_k(\cdot \mid s,a) - \mathbb{P}_k(\cdot \mid s,a)\|_2 \\
&\leq \sum_{i=1}^{2} 2\|\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) - \mathbb{E}_{Q_i}[\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) \mid \mathcal{G}]\|_2 + 4\epsilon_{mod}(\delta)
\end{aligned} \tag{22}$$

Also note the following computation.

$$\begin{aligned}
\mathbb{E}_{Q_i}[\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) \mid \mathcal{G}] &= \mathbb{E}_{Q_i}\left[\frac{\mathbf{w}_{n,i}}{c_{n,i}}\mathbb{1}_{c_{n,i}\neq 0} \mid \mathcal{G}\right] \\
&= \frac{\mathbb{E}_{Q_i}[\mathbf{w}_{n,i} \mid \mathcal{G}]}{c_{n,i}}\mathbb{1}_{c_{n,i}\neq 0} \\
&= \frac{\mathbb{P}_{k_n}(\cdot \mid s,a)c_{n,i}}{c_{n,i}}\mathbb{1}_{c_{n,i}\neq 0} \\
&= \mathbb{1}_{c_{n,i}\neq 0}\mathbb{P}_{k_n}(\cdot \mid s,a)
\end{aligned}$$

We define the following quantity, overloading notation from Lemma 5.

$$\mathbf{diff}_i = \mathbb{1}_{c_{n,i}\neq 0}\mathbb{P}_{k_n}(\cdot \mid s,a) - \mathbb{P}_k(\cdot \mid s,a)$$

Note that $\bar{\mathbf{\Delta}}_i = \mathbf{diff}_i^T\tilde{\mathbf{V}}_{s,a}$. We recall the following definition before proceeding to show the main inequality.

$$\Delta_{n,k}(s,a) = \mathbb{P}_{k_n}(\cdot \mid s,a) - \mathbb{P}_k(\cdot \mid s,a)$$

$$\begin{aligned}
\left|\bar{\mathbf{\Delta}}_1^T\bar{\mathbf{\Delta}}_2 - \|\Delta_{n,k}(s,a)\|_2^2\right| &= \left|\mathbf{diff}_1^T\tilde{\mathbf{V}}_{s,a}\tilde{\mathbf{V}}_{s,a}^T\mathbf{diff}_2 - \mathbf{diff}_1^T\mathbf{diff}_2\right| + \left|\mathbf{diff}_1^T\mathbf{diff}_2 - \|\Delta_{n,k}(s,a)\|_2^2\right| \\
&\leq \|\mathbf{diff}_1\|_2\left\|\mathbf{diff}_2 - \tilde{\mathbf{V}}_{s,a}\tilde{\mathbf{V}}_{s,a}^T\mathbf{diff}_2\right\|_2 + \|\mathbf{diff}_1 - \Delta_{n,k}(s,a)\|_2\|\mathbf{diff}_2\|_2 \\
&\quad + \|\mathbf{diff}_1\|_2\|\mathbf{diff}_2 - \Delta_{n,k}(s,a)\|_2 \\
&\leq \|\mathbf{diff}_1\|_1\left\|\mathbf{diff}_2 - \tilde{\mathbf{V}}_{s,a}\tilde{\mathbf{V}}_{s,a}^T\mathbf{diff}_2\right\|_2 + \|\mathbf{diff}_1 - \Delta_{n,k}(s,a)\|_2\|\mathbf{diff}_2\|_1 \\
&\quad + \|\mathbf{diff}_1\|_1\|\mathbf{diff}_2 - \Delta_{n,k}(s,a)\|_2 \\
&\leq 2\left\|\mathbf{diff}_2 - \tilde{\mathbf{V}}_{s,a}\tilde{\mathbf{V}}_{s,a}^T\mathbf{diff}_2\right\|_2 + 2\|\mathbf{diff}_1 - \Delta_{n,k}(s,a)\|_2 + 2\|\mathbf{diff}_2 - \Delta_{n,k}(s,a)\|_2 \\
&\leq 2\left\|\mathbb{P}_{k_n}(\cdot \mid s,a) - \tilde{\mathbf{V}}_{s,a}\tilde{\mathbf{V}}_{s,a}^T\mathbb{P}_{k_n}(\cdot \mid s,a)\right\|_2 \\
&\quad + 2\left\|\mathbb{P}_k(\cdot \mid s,a) - \tilde{\mathbf{V}}_{s,a}\tilde{\mathbf{V}}_{s,a}^T\mathbb{P}_k(\cdot \mid s,a)\right\|_2 \\
&\quad + 2\mathbb{1}_{c_{n,1}=0}\|\mathbb{P}_{k_n}(\cdot \mid s,a)\|_2 + 2\mathbb{1}_{c_{n,2}=0}\|\mathbb{P}_{k_n}(\cdot \mid s,a)\|_2 \\
&\leq 4C_2\sqrt{\frac{K\epsilon_{mod}(\delta)}{f_{min}\alpha}} + 4\left(\max_i \mathbb{1}_{c_{n,i}=0}\right)
\end{aligned}$$

Notice that $4C_2\sqrt{\frac{K\epsilon_{mod}(\delta)}{f_{min}\alpha}} \geq 4\epsilon_{mod}(\delta)$ since $\epsilon_{mod}(\delta) \leq 2$, $C_2 \geq 2$, $K \geq 1$, $f_{min}, \alpha \leq 1$. Combining this and the computation above with inequality 16, we have the following final bound.

$$\left|\mathrm{dist}_{1,(s,a)} - \|\Delta_{n,k}(s,a)\|_2^2\right| \leq \sum_{i=1}^{2} 2\left\|\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) - \mathbb{E}_{Q_i}[\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) \mid \mathcal{G}]\right\|_2 + 8C_2\sqrt{\frac{K\epsilon_{mod}(\delta)}{f_{min}\alpha}} + 4\left(\max_i \mathbb{1}_{c_{n,i}=0}\right) \tag{23}$$

where we remind the reader that $c_{n,i} = N(n,i,s,a)$. □

I.1.2. Bounding the concentration-type term

We bound the first term in the decomposition lemma (Lemma 10) with high probability.

Lemma 11. *With probability at least $1-\delta$, when $T_n \geq \Omega\left(Gt_{mix}\log\left(\frac{G}{\delta}\log(1/\alpha)\right)\right)$ and $G \geq \Omega\left(\frac{\log(1/\delta)}{\alpha^2}\right)$, we have the following bound.*

$$\left\|\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) - \mathbb{E}_{Q_i}[\hat{\mathbb{P}}_{n,i}(\cdot \mid s,a) \mid \mathcal{G}]\right\|_2 \leq O\left(\sqrt{\frac{K+\log(1/\delta)}{G\alpha}}\right)$$

Proof. The proof of this lemma is verbatim the proof of Lemma 6 after the first inequality. □

I.1.3. COMBINING THE BOUNDS

We reuse Corollary 1 along with Lemma 11 applied to Lemma 10 to get the following bound with probability at least $1-\delta$,

$$\left|\text{dist}_{1,(s,a)} - \|\Delta_{n,k}(s,a)\|_2^2\right| \leq O\left(\sqrt{\frac{K+\log(1/\delta)}{G\alpha}}\right) + 8C_2\sqrt{\frac{K\epsilon_{mod}(\delta)}{f_{min}\alpha}}$$

whenever $T_n \geq \Omega\left(Gt_{mix}\log(G/\delta)\log(1/\alpha)\right)$ and $G \geq \Omega\left(\frac{\log(1/\delta)}{\alpha^2}\right)$.

□