Tractable Optimality in Episodic Latent MABs

Jeongyeol Kwon
University of Wisconsin-Madison
jeongyeol.kwon@wisc.edu

Yonathan Efroni
Meta, New York
jonathan.efroni@gmail.com

Constantine Caramanis
The University of Texas at Austin
constantine@utexas.edu

Shie Mannor
Technion, NVIDIA
shie@ee.technion.ac.il,
smannor@nvidia.com

Abstract

We consider a multi-armed bandit problem with A actions and M latent contexts, where an agent interacts with the environment for an episode of H time steps. Depending on the length of the episode, the learner may not be able to estimate accurately the latent context. The resulting partial observation of the environment makes the learning task significantly more challenging. Without any additional structural assumptions, existing techniques to tackle partially observed settings imply the decision maker can learn a near-optimal policy with $O(A)^H$ episodes, but do not promise more. In this work, we show that learning with *polynomial* samples in A is possible. We achieve this by using techniques from experiment design. Then, through a method-of-moments approach, we design a procedure that provably learns a near-optimal policy with $O(\text{poly}(A) + \text{poly}(M, H)^{\min(M,H)})$ interactions. In practice, we show that we can formulate the moment-matching via maximum likelihood estimation. In our experiments, this significantly outperforms the worst-case guarantees, as well as existing practical methods.

1 Introduction

In Multi-Armed Bandits (MABs), an agent learns to act optimally by interacting with an unknown environment. In many applications, interaction sessions are short, relative to the complexity of the overall task. As a motivating example, we consider a large content website that interacts with users arriving to the site, by making sequential recommendations. In each such *episode*, the system has only a few chances to recommend items before the user leaves the website – an interaction time typically much smaller than the number of actions or the types of users. In such settings, agents often have access to short horizon episodes, and have to learn how to process observations from different episodes to learn the best adaptation strategy.

Motivated by such examples, we consider the problem of learning a near-optimal policy in Latent Multi-Armed Bandits (LMABs). In each episode with time-horizon H, the agent interacts with one of M possible MAB environments (e.g., type of user) randomly chosen by nature. Without knowing the identity of the environment (we call this the *latent context*), an agent aims to maximize the expected cumulative reward per episode (see Definition 2.1 for a formal description). The LMAB framework is different from the setting considered in [31] where multiple episodes proceed in parallel without limiting the horizon of an episode H. For long horizons H, we show that it is possible to find a near-optimal policy and to determine near-optimal actions for each episode, as if we knew the hidden context. If $H \ll A$ where A is the total number of actions, however, this is no longer possible. Instead, we aim to learn the best *history-dependent* policy for H time steps.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

1.1 Our Results and Contributions

In this work, we study the problem of learning a near optimal policy of an LMABs for both long and short horizon H. In the long horizon case, we show the problem's latent parameters are learnable. However, for short horizon, the latent parameters may not be identifiable and a different perspective is required.

A naive approach to learn a near optimal policy of an LMAB starts by casting the problem as a Markov Decision Process (MDP). Assume that the reward values are discrete and defined over a finite alphabet of cardinality Z. By defining the state space as the set of all sequences of history observations an LMAB can be formulated as an MDP with $O((AZ)^H)$ states. Appealing to standard reinforcement learning algorithms (*e.g.,* [18, 6]) we can learn an ϵ-optimal policy with $O(AZ)^H/\epsilon^2$ samples.

A natural way to improve upon the naive approach is through using the unique structure of the LMAB setting. In this work we focus on the following question: *can we learn a near-optimal policy with fewer than $O(AZ)^H/\epsilon^2$ samples as the naive approach?*

Our work answers the above question affirmatively. Specifically, our main contributions are as follows. We show that the dependence of our algorithm is $\mathrm{poly}(A) + \mathrm{poly}(H, M)^{\min(H,M)}$, and thus tractable when either H or M is small, even for very large A. That is, we are particularly focused on the setting with a few contexts or relatively short episodes (in comparison to A), *i.e.,* $M = O(1)$ or $H = O(1)$, where a natural objective is to learn a near-optimal history-dependent policy for H time steps (see also Figure 3 in Appendix B.1).

1.2 Related Work

Due to space constraints, we discuss only the most closely related work here, and defer a lengthier discussion on related work to Appendix B.1.

Learning priors in multi-armed bandit problems Several recent works have considered the Bayesian learning framework with short time-horizon [29, 35, 21]. The focus in this line of work is on the design of algorithms that learn the prior, while acting with a fixed Bayesian algorithm (e.g., Thompson sampling). While Bayesian learning with short time-horizon may be viewed as a special case of LMABs, the baseline policy we compare ourselves to is the optimal H-step policy, which is a harder baseline than considering a fixed Bayesian algorithm.

Latent MDPs Some prior work considers the framework of Latent MDPs (LMDPs), which is the MDP generalization of LMAB [16, 37, 24, 23]. In particular, [24] has shown the information-theoretic limit of learning in LMDPs with no assumptions on MDPs, *i.e.,* an exponential number of sample episodes $\Omega(A^M)$ is necessary to learn a near-optimal policy in LMDPs. In contrast, we show that in LMABs, the required number of episodes can be polynomial in A. This does not contradict the result in [24], since their lower bound construction comes from the challenge in state-exploration with latent contexts. In contrast, there is no state-exploration issue for bandits, which enables the polynomial sample complexity in A. Furthermore, our upper bound does not require any assumptions or additional information such as good initialization or separations as in [24]. To the best of our knowledge, no existing results are known for learning a near optimal policy of LMAB instances for $M \geq 3$ without further assumptions.

Learning in POMDPs with full-rank observations One popular learning approach in partially observable systems is the tensor-decomposition method, which extracts the realization of model parameters from third-order tensors [1, 7, 15]. However, the recovery of model parameters require specific geometric assumptions on the full-rankness of a certain test-observation matrix. Furthermore, most prior work requires a uniform reachability assumption, *i.e.,* all latent spaces should be reached with non-negligible probabilities by any exploration policy for the parameter recovery. Recent results in [19, 30] have shown that the uniform reachability assumption can be dropped with the optimism principle. However, they still require the full-rankness of a test-observation matrix to keep the volume of a confidence set explicitly bounded. Since LMAB instances do not necessarily satisfy the full-rankness assumption, their results do not imply an upper bound for learning LMABs.

2 Preliminaries

We define the problem of **episodic** latent multi-armed bandits with time-horizon $H \geq 2$ as follows:

Definition 2.1 (Latent Multi-Armed Bandit (LMAB)) *Let LMAB be a tuple* $\mathcal{B} = (\mathcal{A}, \{w_m\}_{m=1}^M, \{\mu_m\}_{m=1}^M)$*, where* $\mathcal{A}$ *is a set of actions,* $\{w_m\}_{m=1}^M$ *are the mixing weights such that a latent context* m *is randomly chosen with probability* w_m*, and* μ_m *is the* model parameter *that describes a reward distribution, i.e.,* $\mathbb{P}_{\mu_m}(r \mid a) := \mathbb{P}(r \mid m, a)$*, according to an action* $a \in \mathcal{A}$ *conditioning on a latent context* m *(each* μ_m *parameterizes the probability model, and is not necessarily a mean reward vector).*

We do not assume a priori knowledge of mixing weights. The bulk of this paper considers discrete reward realizations, when the support of the reward distribution is finite and bounded. In Appendix D, we also show that our results can be adapted to Gaussian reward distributions.

Assumption 2.2 (Discrete Rewards) *The reward distribution has finite and bounded support. The reward attains a value in the set* $\mathcal{Z}$*. We assume that for all* $z \in \mathcal{Z}$ *we have* $|z| \leq 1$*. We denote the cardinality of* $\mathcal{Z}$ *as* Z *and assume that* $Z = O(1)$*.*

As an example, Bernoulli distribution satisfies Assumption 2.2 with $\mathcal{Z} = \{0, 1\}$ and $Z = 2$. We denote the probability of observing a reward value z by playing an action a as $\mu_m(a, z) := \mathbb{P}(r = z \mid m, a)$ in a context m. We often use μ_m as a reward-probability vector in $\mathbb{R}^{AZ}$ indexed by a tuple $(a, z) \in \mathcal{A} \times \mathcal{Z}$.

At the beginning of every episode, a latent context $m \in [M]$ is sampled from a mixing distribution $\{w_m\}_{m=1}^M$ and fixed for H time steps, however we cannot observe m directly. We consider a policy class Π which contains all history-dependent policies $\pi : (\mathcal{A} \times \mathcal{Z})^* \to \mathcal{A}$. Our goal is to find a near optimal policy $\pi \in \Pi$ that maximizes the expected cumulative reward V for each episode $V^\star = \max_{\pi \in \Pi} V(\pi) := \mathbb{E}^\pi \left[\sum_{t=1}^H r_t\right]$, where the expectation is taken over latent contexts and rewards generated by an LMAB instance, and actions following a policy π.

Definition 2.3 (Approximate Planning Oracle) *A planning oracle receives an LMAB instance* $\mathcal{B}$ *and returns an* ϵ*-approximate policy* π *such that* $V^\star - V(\pi) \leq \epsilon$*.*

Concretely, the point-based value-iteration (PBVI) algorithm [33] is an ϵ-approximate planning algorithm which runs in time $O(HMAZ(H^2/\epsilon)^{O(M)})$.

Additional notation. For any quantity q with respect to the true LMAB, $\mathcal{B}$, we use $\hat{q}$ to refer to its corresponding empirical estimate. We use $w_{\min} := \min_{m \in [M]} w_m$ for the minimum mixing weight. For any l^{th}-order tensor $T_l \in \mathbb{R}^{n \times \cdots \times n}$ (n repeated l times), we denote $\|T_l\|_\infty$ for the element-wise largest absolute value. For any vector in $v \in \mathbb{R}^n$ in dimension $n \in \mathbb{N}_+$, we use $v^{\bigotimes l}$ to denote a degree l tensorization of v. We often use $\|\mu_m(a, \cdot) - \hat{\mu}_m(a, \cdot)\|_1$ to mean the l_1 statistical distance between reward distributions: $\sum_{z \in \mathcal{Z}} |\mu_m(a, z) - \hat{\mu}_m(a, z)|$. Lastly, we denote by a_t and r_t as the action and reward realizations observed at time step $t \in [H]$.

3 Sample-Complexity of Learning LMABs

In this section, we develop our main algorithm that learns a near-optimal policy of an LMAB instance with $O(\text{poly}(A) + \text{poly}(M, H)^{\min(M,H)})$ samples. Towards achieving this goal, we first elaborate on a low rank representation of the LMAB problem. Then, we show how experimental design techniques can assist in utilizing this structure to improve over the naive upper bound to the problem.

3.1 Dimensionality Reduction via Experimental Design

When $H = 1$, estimating the latent context is not possible, and executing a standard MAB strategy (e.g., UCB) is optimal. However, to obtain a near-optimal history-dependent policy with longer time-horizons $H > 1$, tracking the mean reward of an action is not enough, since rewards from previous time steps are correlated with current and future rewards.

We consider a model-based learning approach from the lth-order moments of reward observations. Since there are A actions, this implies that we have $O(A^l)$ quantities to estimate, which would incur $O(A^l)$ sample complexity if we separately measure every correlation between l pairs of actions. On the other hand, when $M \ll A$, the distributions μ_m (recall these are reward probabilities for each action, in each context) occupy only a M-dimensional subspace of AZ-dimensional space: $\mathbf{U} := \text{span}\{\mu_1, \mu_2, \ldots, \mu_M\}$. Thus if we could estimate $\mathbf{U}$, and project all observations to $\mathbf{U}$, we could remove the unfavorable dependence on A.

However, even when the subspace $\mathbf{U}$ is known a priori, dimensionality reduction with bandit feedback is non-trivial. We need to compute directly the projection onto $\mathbf{U}$ of estimates of the reward probability vectors. To see the challenge, let $\{\beta_j\}_{j=1}^M \subset \mathbb{R}^{AZ}$ be an orthonormal basis of $\mathbf{U}$. To compute the projected estimate $\mu_m^\top \beta_j$, we need a sampling policy π such that $\pi(a) \propto \sum_z |\beta_j(a,z)|$. The variance of this sampling policy can be as much as $O(\|\beta_j\|_1^2)$, which in general scales with $\|\beta_j\|_1^2 = O(A)$. Therefore, reliable estimation for any statistics in the reduced dimension would still have a dependence on A. In particular, since our approach is based on higher-order method-of-moments, an estimation of lth-order statistics for $l \geq 2$ would require $\Omega(A^{l/2})$ samples.

The general idea of dimensionality reduction is not fatally flawed. Instead, we need to avoid the pitfalls of the approach outlined above. First, the calculation above shows that we need to control the $\|\cdot\|_1$-norm of the vectors β_j, ideally, $\|\beta_j\|_1 = O(1)$. Second, we only need a good estimate for each μ_m such that $\max_{a\in\mathcal{A}} \|\mu_m(a,\cdot) - \hat{\mu}_m(a,\cdot)\|_1 \leq \epsilon$ to compute an ϵ-optimal policy.

The key is to show that we can choose β_j's to be a subset of the standard basis in $\mathbb{R}^{AZ}$ (hence they will have $\|\beta_j\|_1 = 1$), in such a way that guarantees the approximation quality for μ_m from estimating $\mu_m^\top \beta_j$'s. In terms of the original problem, the existence of such a subset of the standard basis is equivalent to the existence of a small set of informative action-value pairs, that are sufficiently correlated with all other action-value pairs. This is called a *core set* in the experimental design literature, and its existence in our context, is an important consequence of the Kiefer–Wolfowitz theorem. Specifically, we can select a core set of coordinates with the following crucial lemma:

Lemma 3.1 *Let $\mathbf{U}$ be a given k-dimensional linear subspace in $\mathbb{R}^d$ where $d \gg k$ and let $u, \bar{u} \in \mathbf{U}$. There exists an algorithm that runs in time $\tilde{O}(dk^2)$ and returns the following.*

1. *A core set of at most $n = 4k \log\log k + 16$ coordinates $\{i_j\}_{j=1}^n \subseteq [d]$ such that $\|u - \bar{u}\|_\infty \leq \sqrt{2k} \max_{j\in[n]} |u(i_j) - \bar{u}(i_j)|$.*
2. *A linear transformation that maps $[\bar{u}(i_1), \ldots, \bar{u}(i_n)]$ to its corresponding $\bar{u} \in \mathbf{U}$.*

Note that in our setting, $d = AZ$ and $k = M$. We prove Lemma 3.1 in Section C.1, essentially as a corollary of the Kiefer–Wolfowitz theorem (and its geometric interpretation) for (near)-optimal experimental design [20, 39].

In the context of bandits and RL, the Kiefer–Wolfowitz theorem has been used to study how the misspecification in linear representation changes the problem landscape (*e.g.,* [27, 32]). However, experimental design has not been previously used for dimensionality reduction for problems with bandit feedback (as far as we know). We believe it is a powerful tool. We review the basics of experimental design in Appendix B.2.

A direct consequence of Lemma 3.1 is an algorithm to find a set of coordinates of size $\tilde{O}(M)$ that are sufficient to reconstruct the latent reward model $\{\mu_m\}_{m=1}^M$ where $\mu_m \in \mathbb{R}^{AZ}$. We refer to this set of coordinates as the set of core action-value pairs.

Corollary 3.2 *Suppose the subspace $\mathbf{U} = \text{span}\{\mu_1, \ldots, \mu_M\}$ is given. Then for any $\hat{\mu} \in \mathbf{U}$, there exists an algorithm that runs in time $\tilde{O}(ZAM^2)$ and returns the following.*

1. *A set of core action-value pairs $\{(a_j, z_j)\}_{j=1}^n \subseteq \mathcal{A} \times \mathcal{Z}$ of size at most $n = 4M \log\log M + 16$, such that for all $m \in [M]$*

$$\max_{a\in\mathcal{A}} \|\mu_m(a,\cdot) - \hat{\mu}_m(a,\cdot)\|_1 \leq 2Z\sqrt{2M} \cdot \max_{j\in[n]} |\nu_m(j) - \hat{\nu}_m(j)|,$$

where $\nu_m, \hat{\nu}_m \in \mathbb{R}^n$ such that $\nu_m(j) := \mu_m(a_j, z_j)$ and $\hat{\nu}_m(j) := \hat{\mu}_m(a_j, z_j)$.

2. *A linear transformation that maps* $[\hat{\nu}(i_1), \ldots, \hat{\nu}(i_n)]$ *to its corresponding* $\hat{\mu} \in \mathbf{U}$.

Assuming access to the subspace $\mathbf{U}$ (we remove this assumption in Section 3.4), Corollary 3.2 implies a strategy for estimating the latent model parameters $\{\mu_m\}_{m=1}^M$: obtain core action-value pairs $\{(a_j, z_j)\}_{j=1}^n$, estimate $\{\nu_m\}_{m=1}^M$, and construct latent model parameters via a linear transformation of $\{\nu_m\}_{m=1}^M$. In the following sections, we show that by matching higher-order moments we can estimate $\{\nu_m\}_{m=1}^M$ to sufficiently good accuracy with only polynomial dependence in A.

3.2 $H \geq 2M - 1$: Identifiable Regime in Wasserstein Metric

In the regime where that $H \geq 2M - 1$ we can measure moments up to order $(2M-1)^{\text{th}}$. Then, we can leverage recent advances in learning parameters of mixture distributions from higher-order moments [40, 12]. That is, we can recover $\{\nu_m\}_{m=1}^M$ by estimating $2M - 1$ higher-order moments and find $\{\hat{\nu}_m\}_{m=1}^M$ that matches them. This further emphasizes the importance of the dimensionality reduction we take to obtain the set of core action-values pairs; otherwise, a moment-based approach to recover $\{\mu_m\}_{m=1}^M$ would have an exponential dependence in A.

We now elaborate on the estimation procedure of $\{\nu_m\}_{m=1}^M$ from higher-order moments. We define the l-order tensor as $T_l := \sum_{m=1}^M w_m \nu_m^{\otimes l}$. For an LMAB instance, we can access the tensor T_l using observational data by simply estimating correlations between l core action-value pairs within each episode. Specifically, for any $\mathcal{I} = (i_1, i_2, ..., i_l) \in [n]^l$ the tensor $T_l(i_1, i_2, ..., i_l)$ is also given by

$$T_l(i_1, i_2, ..., i_l) = \mathbb{E}^{\pi_{\mathcal{I}}}\left[\Pi_{t=1}^l \mathbb{1}\{r_t = z_{i_t}\}\right], \tag{1}$$

where $\pi_{\mathcal{I}}$ is a policy that performs the sequence of actions $(a_{i_1}, a_{i_2}, ..., a_{i_l})$ for $t = 1, \ldots, l$.

Suppose we find estimators $\{(\hat{w}_m, \hat{\nu}_m)\}_{m=1}^M$ such that moments match T_l up to error δ for all $l \in [2M-1]$. The results in [40, 12] imply that the closeness in moments of up to $(2M-1)^{\text{th}}$ degree implies the closeness in model parameters:

Lemma 3.3 *Suppose* $\|T_l - \sum_{m=1}^M \hat{w}_m \hat{\nu}_m^{\otimes l}\|_\infty < \delta$ *for all* $l = 1, 2, ..., 2M-1$ *for some sufficiently small* $\delta > 0$. *Then,*

$$\inf_{\Gamma} \sum_{(m,m') \in [M]^2} \Gamma(m, m') \cdot \|\nu_m - \hat{\nu}_{m'}\|_\infty \leq O\left(M^3 n \cdot \delta^{-1/(2M-1)}\right), \tag{2}$$

where Γ *is a joint distribution over* $(m, m') \in [M]^2$ *satisfying*

$$\Gamma(m, m') \in \mathbb{R}_+^{M \times M} : \textstyle\sum_{m'=1}^M \Gamma(m, m') = w_m, \sum_{m=1}^M \Gamma(m, m') = \hat{w}_m. \tag{3}$$

The form of guarantee given for $\{\hat{w}_m, \hat{\nu}_m\}_{m=1}^M$ is in the *Wasserstein distance* between two latent model parameters. A useful property of the Wasserstein metric is that the distance measure is invariant to permutation of individual components, and flexible with arbitrarily small mixing probabilities or arbitrarily close components [40] (see also Appendix B.3 for the review on Wasserstein distance).

Once we obtain estimates of $\{\nu_m\}_{m=1}^M$, we can estimate $\{\hat{\mu}_m\}_{m=1}^M$ as implied by Corollary 3.2 (see Appendix C.8 for further details). We then show that for any history-dependent policy π, the expected cumulative rewards of π are approximately the same for any close LMAB instances.

Proposition 3.4 *Let* $\mathcal{B} = (\mathcal{A}, \{w_m\}_{m=1}^M, \{\mu_m\}_{m=1}^M)$ *and* $\hat{\mathcal{B}} = (\mathcal{A}, \{\hat{w}_m\}_{m=1}^M, \{\hat{\mu}_m\}_{m=1}^M)$ *be any two LMABs. Then, for any history-dependent policy* $\pi : (\mathcal{A} \times \mathcal{Z})^* \to \mathcal{A}$, *we have*

$$|V(\pi) - \hat{V}(\pi)| \leq H^2 \cdot \inf_{\Gamma} \sum_{(m,m') \in [M]^2} \left(\Gamma(m, m') \cdot \max_{a \in \mathcal{A}} \|\mu_m(a, \cdot) - \hat{\mu}_{m'}(a, \cdot)\|_1\right), \tag{4}$$

where the infimum over Γ *is taken over joint distributions over* (m, m') *satisfying* (3).

From Corollary 3.2 we have $\max_{a \in \mathcal{A}} \|\mu_m(a, \cdot) - \hat{\mu}_{m'}(a, \cdot)\|_1 \leq 2Z\sqrt{2M}\|\nu_m - \hat{\nu}_{m'}\|_\infty$ for any $m, m' \in [M]$. Plugging this into Proposition 3.4, we have

$$|V(\pi) - \hat{V}(\pi)| \leq \text{poly}(H, Z, M, n) \cdot \delta^{-1/(2M-1)}.$$

Thus, using Lemma 3.3 with $\delta < (\text{poly}(H, Z, M, n)/\epsilon)^{2M-1}$, we can conclude that any ϵ-optimal policy for $\hat{\mathcal{B}}$ is $O(\epsilon)$-optimal for the underlying LMAB $\mathcal{B}$.

3.3 $H < 2M - 1$: Unidentifiable Regime with Short Time-Horizon

A more interesting regime is the one in which the time-horizon H is smaller than the required degree of moments $2M - 1$. For such a setting *we cannot measure moments of degree higher than* H. Therefore, if $H < 2M - 1$, we cannot rely on the identifiablility of the underlying LMAB model.

Instead, we make the following claim: to compute an optimal policy only for time horizon $H < 2M-1$, we only need to match the Hth order moments. This is formalized in the next lemma.

Lemma 3.5 *Suppose that* $\| \sum_{m=1}^{M} w_m \mu_m^{\otimes H} - \sum_{m=1}^{M} \hat{w}_m \hat{\mu}_m^{\otimes H} \|_\infty \leq \delta$*, then for any history dependent policy* $\pi \in \Pi$*, we have*

$$|V(\pi) - \hat{V}(\pi)| \leq H Z^H \cdot \delta.$$

Therefore, it is sufficient to find estimates of mixing weights and latent models that match the measured H^{th}-order moment, and then compute an optimal policy for the estimated model. Lemma 3.5 is natural for discrete reward distributions with bounded support. Interestingly, we extend this result to Gaussian rewards, which are continuous and unbounded, in Appendix D.

Hence, a natural idea is to estimate parameters $\{\hat{w}_m\}_{m=1}^M, \{\hat{\mu}_m\}_{m=1}^M$ that satisfy the condition in Lemma 3.5 without incurring $O(A^H)$ sample complexity. This is possible if we have good estimates of moment-matching parameters for a set of core action-values.

Proposition 3.6 *For any given* $l \geq 1$*, if* $\| \sum_{m=1}^{M} w_m \nu_m^{\otimes l} - \sum_{m=1}^{M} \hat{w}_m \hat{\nu}_m^{\otimes l} \|_\infty \leq \delta$*, then*

$$\left\| \sum_{m=1}^{M} w_m \mu_m^{\otimes l} - \sum_{m=1}^{M} \hat{w}_m \hat{\mu}_m^{\otimes l} \right\|_\infty \leq (2M)^{l/2} \cdot \delta.$$

By Proposition 3.6, we can conclude that it is sufficient to estimate $T_H := \sum_{m=1}^{M} w_m \nu_m^{\otimes H}$ and find $\{(\hat{w}_m, \hat{\nu}_m)\}_{m=1}^M$ that matches T_H element-wise up to accuracy $\delta := (\epsilon/H)/(Z\sqrt{2M})^H$.

3.4 Main Result

The sections above have outlined the key ideas we need *assuming knowledge of* $\mathbf{U}$. In this section, we show how we can estimate $\mathbf{U}$, and we describe the complete procedure that learns a near-optimal policy of an LMAB instance. We give the details in Algorithm 1. Our algorithm is divided to three steps as detailed below.

Step 1: Estimating $\mathbf{U}$. Algorithm 1 first estimates the second-order moments $M_2 = \sum_{m=1}^{M} w_m \mu_m \mu_m^\top$. Let $\widehat{\mathbf{U}}$ be the top-M eigenvectors of an empirical estimate of M_2, where $\hat{M}_2$ is constructed by collecting samples of reward correlations by taking random actions at first two time steps. Note that $\widehat{\mathbf{U}}$ may not be a good proxy for some μ_m with small mixing probability $w_m \approx 0$, and thus all elements in $\mathbf{U}$ are not necessarily close to $\widehat{\mathbf{U}}$. We defer the details of subspace recovery procedure to the proof of the following lemma in Appendix C.7.

Lemma 3.7 *Let* $\widehat{\mathbf{U}}$ *be a subspace spanned by top-*M *eigenvectors of* $\hat{M}_2$*. After we estimate* $\hat{M}_2$ *using* $N_0 = O(A^4 Z^2 \log(ZA/\eta)/\delta_{\text{sub}}^4)$ *episodes, with probability at least* $1 - \eta$*, for all* $m \in [M]$*, there exists* $\Delta_m : \|\Delta_m\|_\infty \leq \delta_{\text{sub}}/w_m^{1/2}$ *such that* $\mu_m + \Delta_m \in \widehat{\mathbf{U}}$*.*

Our choice of δ_{sub} differs in two regimes as the following:

$$\delta_{\text{sub}} = \frac{\epsilon}{2ZMH^2}, \qquad \text{if: } H \geq 2M - 1,$$

$$\delta_{\text{sub}} = \frac{\min\left(\sqrt{w_{\min} + \epsilon/(MH^2(Z\sqrt{2M})^H)},\ \epsilon/(H\sqrt{M})\right)}{2Z\sqrt{M}H}, \qquad \text{else: } H < 2M - 1, \tag{5}$$

where $w_{\min} = \min_{m \in [M]} w_m$. The main difference in two regimes is that for $H \geq 2M - 1$, when the parameters are identifiable, latent contexts with small mixing probabilities can be ignored

Algorithm 1

Input: Accuracy level $\epsilon, \delta_{\text{sub}}, \delta_{\text{tsr}} > 0$, model parameters M, A, Z, H.
// **Step 1**: Estimate subspace $\mathbf{U}$.
Construct $\hat{M}_2$, an estimate of $M_2 = \sum_{m=1}^{M} w_m \mu_m \mu_m^\top$ using $N_0 = \tilde{O}(A^4 Z^2/\delta_{\text{sub}}^4)$ episodes.
Calculate $\widehat{\mathbf{U}}$, the span of top-M eigenvectors of $\hat{M}_2$.
// **Step 2**: Estimate $\{(w_m, \nu_m)\}_{m=1}^M$.
Get core action-value pairs $\{(a_j, z_j)\}_{j=1}^n$ by calling Corollary 3.2.
Construct $\{\hat{T}_l\}_{l=1}^{\min(H,2M-1)}$ using $N = \tilde{O}(n^{\min(2M-1,H)} \cdot M^2/\delta_{\text{tsr}}^2)$ episodes.
Find valid empirical parameters $\{(\hat{w}_m, \hat{\nu}_m)\}_{m=1}^M$ satisfying (7).
// **Step 3**: Use the core-action value pairs to construct an empirical LMAB.
Construct empirical model $\hat{\mathcal{B}} = (\mathcal{A}, \{\hat{w}_m\}_{m=1}^M, \{\hat{\mu}_m\}_{m=1}^M)$.
Output: optimal policy of $\hat{\mathcal{B}}$ by calling a planning oracle (Definition 2.3) for $\hat{\mathcal{B}}$.

once $w_m = o(\epsilon/M)$ to guarantee the closeness in distributions of observations. However, in the parameter unidentifiable regime, total variation distance is bounded only through errors in the moment space. Therefore, the estimated subspace needs to be accurate even for contexts with small mixing probabilities to keep the higher-order moments well approximated. Note that for instances with well-balanced mixing probabilities, *i.e.,* if $w_{\min} = \Omega(1/M)$, the order of δ_{sub} remains the same as in the $H \geq 2M - 1$ case.

Step 2: Moment Matching. Given $\widehat{\mathbf{U}}$, the subspace spanned by top-M eigenvectors of $\hat{M}_2$, Algorithm 1, follows the procedure described in Sections 3.1-3.3. It constructs the set of (approximate) core action-values pairs $\{(a_j, z_j)\}_{j=1}^n$ (see Appendix C.8.1 for the detailed algorithm). Then, we construct higher-order moments of the core action-value pairs. For every multi-index $(i_1, i_2, ..., i_l) \in [n]^l$, using $N_1 = O(\log(ln^l/\eta)/\delta_{\text{tsr}}^2)$ episodes, we execute $a_t^k = a_{i_t}$ for $t = 1, ..., l$ and estimate higher-order moments (as also described in equation (1))

$$\hat{T}_l(i_1, i_2, ..., i_l) = \frac{1}{N_1} \sum_{k=1}^{N_1} \Pi_{t=1}^l \mathbb{1}\left\{r_t^k = z_{i_t}\right\}. \tag{6}$$

Using standard concentration inequalities and applying the union bounds over all elements in tensors, we get $\|\hat{T}_l - T_l\|_\infty < \delta_{\text{tsr}}$ with probability at least $1 - \eta$. Then we find empirical parameters $\{(\hat{w}_m, \hat{\nu}_m)\}_{m=1}^M$ that satisfy

$$\left\|\sum_{m=1}^M \hat{w}_m \hat{\nu}_m^{\bigotimes l} - \hat{T}_l\right\|_\infty < \delta_{\text{tsr}}, \qquad \forall l \in [\min(H, 2M-1)]. \tag{7}$$

We set $\delta_{\text{tsr}} = O(\epsilon/(ZH^2M^{3.5}n))^{2M-1}$ when $H \geq 2M - 1$. In the parameter unidentifiable regime $H < 2M - 1$, we set $\delta_{\text{tsr}} = O(\epsilon/H)/(Z\sqrt{2M})^H$.

Step 3: Constructing Empirical LMAB. Finally, Algorithm 1 uses the estimates $\{(\hat{w}_m, \hat{\nu}_m)\}_{m=1}^M$ to construct an empirical model $\hat{\mathcal{B}} = (\mathcal{A}, \{\hat{w}_m\}_{m=1}^M, \{\hat{\mu}_m)\}_{m=1}^M)$ after proper clipping and normalization. For this step to succeed, we require $\hat{w}_m$ and $\hat{\nu}_m$ to be valid parameters for the reconstruction of a valid empirical model. We state details on the recovery procedure in Appendix C.8.1.

Once a valid empirical model $\hat{\mathcal{B}}$ is obtained, the remaining step is to call the planning oracle that gives an ϵ-approximate optimal policy for $\hat{\mathcal{B}}$. We conclude this section with an end-to-end guarantee.

Theorem 3.8 *For any LMAB instance with M latent contexts, with probability at least $1 - \eta$, Algorithm 1 returns an ϵ-optimal policy given total number of episodes of at most*

$$\begin{aligned} &\text{poly}(H, M, Z, A, 1/\epsilon) \cdot \log(AZ/\eta) + \text{poly}(H, Z, M)^{2M-1} \cdot \log(M/\eta)/\epsilon^{4M-2}, \quad \textit{if } H \geq 2M-1, \\ &\text{poly}(H, w_{\min}^{-1}, Z, A, 1/\epsilon) \cdot \log(AZ/\eta) + H^2 \log(M/\eta) \cdot \left(2MZ^2\right)^H/\epsilon^2, \quad \textit{otherwise}. \end{aligned}$$

Note that the dependency on A is polynomial. This polynomial term is needed to control the subspace estimation error. The exponential term is derived from the closeness in moments as discussed earlier.

Remark 3.9 (Note on $w_{\min}$) *Ideally, mixing weights are well balanced such that $w_{\min} = \Omega(1/M)$. However, in general we may have arbitrarily small $w_{\min}$. Note that when $H \geq 2M - 1$, the upper bound does not depend on $w_{\min}$. However when $H < 2M - 1$ with arbitrarily small $w_{\min}$, effectively we have $w_{\min} \propto \epsilon/(Z\sqrt{M})^H$ as can be seen in equation (5). This is mainly because in the parameter unidentifiable regime, we give guarantees from moment-closeness, and thus the subspace has to be very accurate to approximate higher-order moments well. Nevertheless, we are not aware whether this is tight, and believe this to be an interesting question for future research.*

Continuous Rewards – Gaussian Case. So far we have focused on rewards with finite support $Z = O(1)$. However, some steps in the algorithm cannot be straightforwardly extended to continuous reward distributions. In Appendix D, we show a similar upper bound that is at most

$$\mathrm{poly}(H, M, A, 1/\epsilon) \cdot \log(A/\eta) + \mathrm{poly}\left(H, M, \log(1/(\eta\epsilon))\right)^{\min(H,M)} / \epsilon^{\min(2H+2, 4M-2)},$$

assuming the rewards are Gaussian with unknown mean (see Theorem D.2 for the exact upper bound).

4 Maximum Likelihood Implementation

In the previous section, we derived a procedure that learns a near-optimal policy of an LMAB instance. However, our procedure relies on matching higher-order moments, which, in general, is not suitable for practical implementation. In this section, we describe a maximum likelihood (MLE) method, motivated by our previous results. Importantly, we can use the Expectation-Maximization (EM) [11] heuristic to find an approximate solution to the MLE optimization problem.

We can start from the set of core action-value pairs $\{(a_j, z_j)\}_{j=1}^n$ given by Corollary 3.2. Now for every time step t, we choose i_t randomly from a uniform distribution over $[n]$, play action a_{i_t}, and observe $b_t := \mathbb{1}\{r_t = z_{i_t}\}$. After repeating this for N episodes, we formulate the log-likelihood function with parameterization $\theta = \{(w_m, \nu_m)\}_{m=1}^M$ as follows:

$$l_N(\theta) := \frac{1}{N} \sum_{k=1}^{N} \log \left(\sum_{m=1}^{M} w_m \Pi_{t=1}^{H} (b_t \nu_m(i_t) + (1 - b_t)(1 - \nu_m(i_t))) \right) \tag{8}$$

To prevent the confusion with searching parameters, we use q^* to denote any quantity q constructed with ground truth parameters, *e.g.,* $\theta^* := \{(w_m^*, \nu_m^*)\}_{m=1}^M$. Let θ_N be the maximum likelihood estimator, *i.e.,* $\theta_N = \arg\max_{\theta \in \Theta} l_N(\theta)$ in some valid parameter set Θ.

We can recover the sample-complexity guaranteed by moment-matching methods studied in the previous section. Specifically, we show that the maximum likelihood estimator with sufficiently many samples have nearly matching moments:

Lemma 4.1 *Consider the maximum likelihood estimator $\theta_N = \{(\hat{w}_m, \hat{\nu}_m)\}_{m=1}^M$ with N episodes for large enough N. If $N \geq C \cdot n^{\min(2H+1, 4M-1)} \log(N/\eta)/\delta^2$ for some sufficiently large constant $C > 0$, then with probability at least $1 - \eta$,*

$$\|T_l^* - \hat{T}_l\|_\infty \leq \delta, \qquad \forall l \in [\min(H, 2M - 1)]. \tag{9}$$

Therefore, by setting the same accuracy parameter δ used in Algorithm 1, we can obtain the required $\theta_N = \{(\hat{w}_m, \hat{\nu}_m)\}_{m=1}^M$ for matching moments. We only replace the moment-matching step (Step 2 in Algorithm 1) with solving the MLE optimization problem (8).

Implementation Details While the log-likelihood formulation (8) is still non-convex and thus intractable, we can rely on powerful heuristics: we initialize the parameters with clustering methods [2] or spectral methods [1, 7], and run the EM algorithm to improve the accuracy [11] (regardless of conditions or guarantees). Therefore, while the sample-complexity upper bound for both methods remains the same, we benefit from the log-likelihood formulation despite the non-convexity.

Specifically, we follow the same steps in Algorithm 1. We first find the core action-event pairs from second-order moments as described (**Step 1**). Then we form the maximum log-likelihood objective (8) and find an approximate MLE solution with the EM algorithm (**Step 2**). After obtaining θ_N, we recover an empirical model which we use as an input to an approximate planning oracle (**Step 3**). We

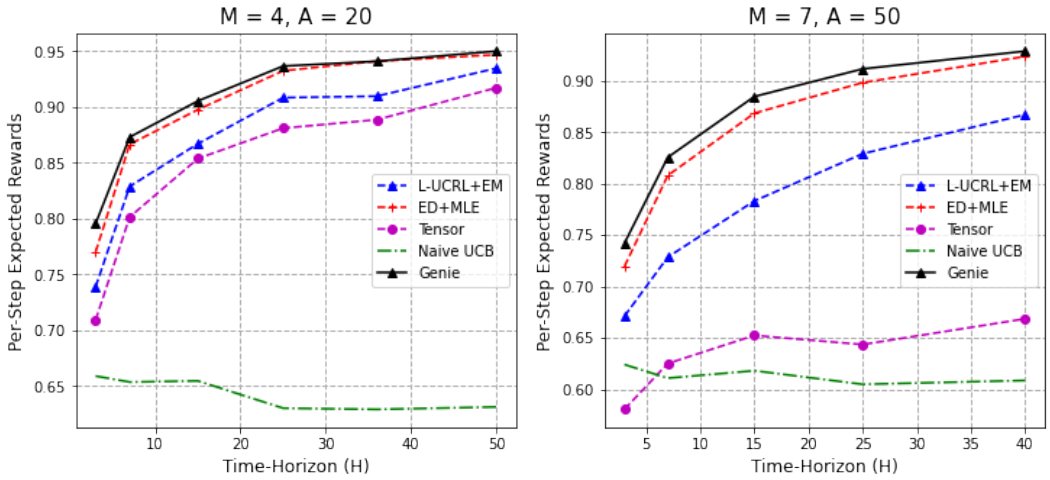

Figure 1: Per time-step rewards for increasing lengths of episodes with history-dependent policies returned after the exploration phase

compare our MLE-based implementation with experimental design (ED + MLE) to other baselines applicable to learning in LMABs (*e.g.,* L-UCRL + EM [24], tensor-decomposition methods [1, 7, 41], naive-UCB [6]). In our experimental results, we can observe that in practice, our method (ED+MLE) outperforms other baselines and the worst case guarantee given in Theorem 3.8.

Provable Benefits of MLE In Appendix E.2, we also show that MLE solutions can automatically adapt to mild separation conditions. Specifically, we show that if $H = \tilde{O}(M^2/\gamma^2)$ for some separation parameter $\gamma > 0$ (see Assumption E.1), then we can achieve the polynomial sample complexity for learning a near-optimal policy with the MLE solution θ_N.

In Figure 1, we compare the averaged *per time-step* rewards obtained with returned policies from different algorithms after some exploration episodes. Note that with increasing H, it becomes easier to identify contexts, *i.e.*, we have bigger separations, and thus MLE based approaches can benefit from larger separations. Due to space constraints, we defer the remaining details to Appendix A.

5 Conclusion and Future Work

In this work, we designed an algorithm that learns a near-optimal policy for an LMAB instance that requires only $\mathrm{poly}(A) + \mathrm{poly}(M, H)^{\min(H, 2M-1)}$ number of samples. We achieved this with the experimental design and method-of-moments, and the maximum likelihood estimation which is more suitable in practice. We discuss a few limitations of this work and future directions below.

Lower Bounds The question of a lower bound on the sample complexity of the LMAB problem remains unresolved. For LMDP, an MDP extension of LMAB, a lower bound of $\Omega(A^M)$ is known due to [24]. While we conjecture that some exponential dependence in M is unavoidable, characterizing the minimax dependence of the exponent is left as an interesting future research question.

Latent MDPs We believe that the moment-matching based approach we took in this work offers a promising way for designing RL algorithms in the presence of latent contexts. Specifically, we believe these techniques can be used to design algorithm that finds a near-optimal policy of LMDPs [24] or reward-mixing MDPs [23] with $M = O(1)$ and without further separation assumptions.

Linear Bandits / Continuous Rewards Our work has focused on the tabular setting, where all arms are independent. It will be an interesting future work to consider the same objective in linear bandit settings. Also, while we only considered Gaussian rewards, we believe our approach can also be extended to a broader class of parametric distributions. Investigating more general classes of rewards is an important future research direction from practical perspective.

Acknowledgement

This research was partially funded by the NSF IFML Institute (NSF 2019844), the NSF AI-EDGE Institute (NSF 2112471), and NSF 1704778, and was also supported by the Israel Science Foundation (grant No. 2199/20).

References

[1] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

[2] D. Arthur and S. Vassilvitskii. k-means++ the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, 2007.

[3] J.-Y. Audibert, S. Bubeck, et al. Minimax policies for adversarial and stochastic bandits. In *COLT*, volume 7, pages 1–122, 2009.

[4] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

[5] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[6] M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. *arXiv preprint arXiv:1703.05449*, 2017.

[7] K. Azizzadenesheli, A. Lazaric, and A. Anandkumar. Reinforcement learning of pomdps using spectral methods. In *Conference on Learning Theory*, pages 193–256, 2016.

[8] A. Bhaskara, M. Charikar, A. Moitra, and A. Vijayaraghavan. Smoothed analysis of tensor decompositions. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 594–603. ACM, 2014.

[9] E. Brunskill and L. Li. Sample complexity of multi-task reinforcement learning. In *Uncertainty in Artificial Intelligence*, page 122. Citeseer, 2013.

[10] R. Chawla, A. Sankararaman, A. Ganesh, and S. Shakkottai. The gossiping insert-eliminate algorithm for multi-agent bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3471–3481. PMLR, 2020.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[12] N. Doss, Y. Wu, P. Yang, and H. H. Zhou. Optimal estimation of high-dimensional gaussian mixtures. *arXiv preprint arXiv:2002.05818*, 2020.

[13] A. Garivier and E. Moulines. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pages 174–188. Springer, 2011.

[14] C. Gentile, S. Li, and G. Zappella. Online clustering of bandits. In *International Conference on Machine Learning*, pages 757–765, 2014.

[15] A. Gopalan, O.-A. Maillard, and M. Zaki. Low-rank bandits with latent mixtures. *arXiv preprint arXiv:1609.01508*, 2016.

[16] A. Hallak, D. Di Castro, and S. Mannor. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015.

[17] J. Hu, X. Chen, C. Jin, L. Li, and L. Wang. Near-optimal representation learning for linear bandits and linear RL. In *International Conference on Machine Learning*, pages 4349–4358. PMLR, 2021.

[18] T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

[19] C. Jin, S. Kakade, A. Krishnamurthy, and Q. Liu. Sample-efficient reinforcement learning of undercomplete pomdps. *Advances in Neural Information Processing Systems*, 33:18530–18539, 2020.

[20] J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

[21] B. Kveton, M. Konobeev, M. Zaheer, C.-w. Hsu, M. Mladenov, C. Boutilier, and C. Szepesvari. Meta-thompson sampling. In *International Conference on Machine Learning*, pages 5884–5893. PMLR, 2021.

[22] J. Kwon and C. Caramanis. The EM algorithm gives sample-optimality for learning mixtures of well-separated gaussians. In *Conference on Learning Theory*, pages 2425–2487, 2020.

[23] J. Kwon, Y. Efroni, C. Caramanis, and S. Mannor. Reinforcement learning in reward-mixing mdps. *Advances in Neural Information Processing Systems*, 34, 2021.

[24] J. Kwon, Y. Efroni, C. Caramanis, and S. Mannor. RL for latent mdps: Regret guarantees and a lower bound. *Advances in Neural Information Processing Systems*, 34, 2021.

[25] J. Kwon, Y. Efroni, C. Caramanis, and S. Mannor. Coordinated attacks against contextual bandits: Fundamental limits and defense mechanisms. In *Proceedings of the 39th International Conference on Machine Learning*, pages 11772–11789. PMLR, 2022.

[26] T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[27] T. Lattimore, C. Szepesvari, and G. Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670. PMLR, 2020.

[28] M. L. Littman, A. R. Cassandra, and L. P. Kaelbling. Learning policies for partially observable environments: Scaling up. In *Machine Learning Proceedings 1995*, pages 362–370. Elsevier, 1995.

[29] C.-Y. Liu and L. Li. On the prior sensitivity of thompson sampling. In *International Conference on Algorithmic Learning Theory*, pages 321–336. Springer, 2016.

[30] Q. Liu, A. Chung, C. Szepesvári, and C. Jin. When is partially observable reinforcement learning not scary? *arXiv preprint arXiv:2204.08967*, 2022.

[31] O.-A. Maillard and S. Mannor. Latent bandits. In *International Conference on Machine Learning*, pages 136–144, 2014.

[32] A. Modi, N. Jiang, A. Tewari, and S. Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR, 2020.

[33] J. Pineau, G. Gordon, and S. Thrun. Anytime point-based approximations for large pomdps. *Journal of Artificial Intelligence Research*, 27:335–380, 2006.

[34] W. Schudy and M. Sviridenko. Concentration and moment inequalities for polynomials of independent random variables. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 437–446. SIAM, 2012.

[35] M. Simchowitz, C. Tosh, A. Krishnamurthy, D. J. Hsu, T. Lykouris, M. Dudik, and R. E. Schapire. Bayesian decision-making under misspecified priors with applications to meta-learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[36] A. Slivkins and E. Upfal. Adapting to a changing environment: the brownian restless bandits. In *COLT*, pages 343–354, 2008.

[37] L. N. Steimle, D. L. Kaufman, and B. T. Denton. Multi-model markov decision processes. *Optimization Online URL http://www. optimization-online. org/DB_FILE/2018/01/6434. pdf*, 2018.

[38] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[39] M. J. Todd. *Minimum-volume ellipsoids: Theory and algorithms*. SIAM, 2016.

[40] Y. Wu, P. Yang, et al. Optimal estimation of gaussian mixtures via denoised method of moments. *Annals of Statistics*, 48(4):1981–2007, 2020.

[41] X. Zhou, Y. Xiong, N. Chen, and X. Gao. Regime switching bandits. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

6 Checklist

1. For all authors...
   - (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   - (b) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
   - (c) Did you discuss any potential negative societal impacts of your work? [N/A] : This work is primarily focused on the theoretical aspect and its purpose is too general for us to speculate about broader societal impacts.
   - (d) Did you describe the limitations of your work? [Yes]
2. If you are including theoretical results...
   - (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   - (b) Did you include complete proofs of all theoretical results? [Yes] : in Supplementary Materials
3. If you ran experiments...
   - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] : in Supplementary Materials
   - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] : We did average errors over multiple independent experiments, but error bars are not reported due to computational burden.
   - (d) Did you include the amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

The remaining issues are not applicable to us.

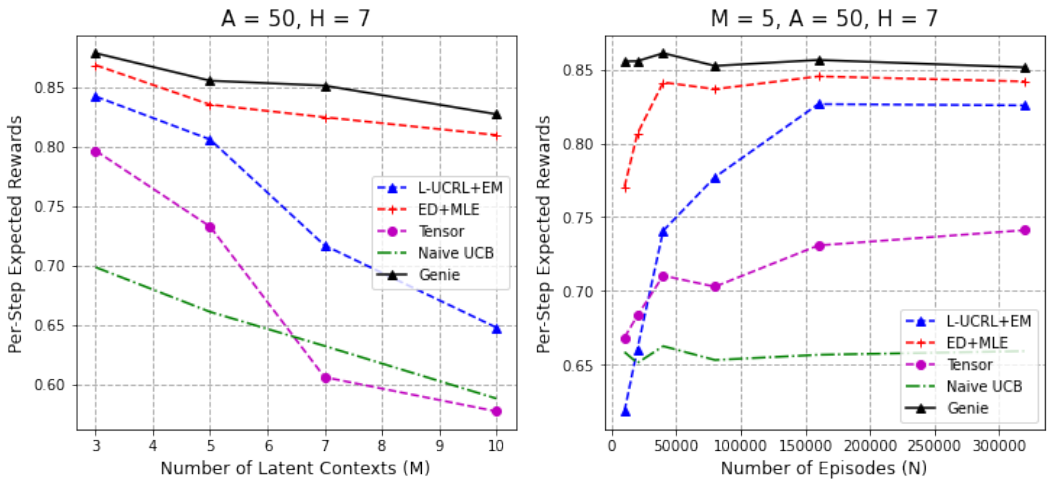

Figure 2: Left: Performance with increasing M, Right: Performance with increasing exploration episodes

Appendix A Experiments

In this section, we demonstrate our maximum likelihood estimation (8) on synthetic examples. We generate random LMAB instances with Bernoulli rewards, where the mean-reward vectors $\{\mu_m\}_{m=1}^M$ lie in a subspace of dimension roughly 4, *i.e.,* rank $\left(\sum_{m=1}^M w_m \mu_m \mu_m^\top\right) \approx 4$. We compare our MLE with experimental design method (8) (ED+MLE) to three benchmarks:

1. Naive UCB [4] without considering contexts (thus a returned policy is stationary).
2. Tensor-decomposition methods [1, 41].
3. L-UCRL with spectral initialization [24].

Even when the theoretical conditions required for the success of tensor-decomposition or spectral methods do not, in practice, standard tensor-decomposition technique by [1] serves as good initialization for the EM or L-UCRL algorithms. After estimating the LMAB $\hat{\mathcal{B}}$, we compute a heuristic policy using Q-MDP [28] for $\hat{\mathcal{B}}$. We refer to the computed policy by Q-MDP [28] using the true LMAB model $\mathcal{B}$ as the *genie policy*.

In the first experiment, we compare the performance of the aforementioned four alternatives. We draw a random LMAB instance with $M = 4, A = 20$ using $N = 5 \cdot 10^4$ sampled episodes, and $M = 7, A = 50$ using $N = 100000$ episodes (Figure 1). We compare the averaged *per time-step* rewards obtained with each policy with increasing length of episodes H.

When $M = 4$, all methods (except naive UCB) exhibit the same pattern of improved performance as the algorithms access more data. As we generated instances to satisfy the full-rank condition, *i.e.,* rank $\left(\sum_{m=1}^M w_m \mu_m \mu_m^\top\right) \approx 4$, even pure tensor-decomposition method works well in practice in this setting. However, when $M = 7$, pure tensor-decomposition method significantly under-performs L-UCRL or ED+MLE. This demonstrates that for LMAB instances with rank degeneracy, additional iteration steps with EM are necessary to get a good solution. Furthermore, the performance of ED+MLE and of L-UCRL does not significantly drop in $M = 7$. This demonstrated that practically it works much better than what can be guaranteed in theory in the worst case. We conjecture that this is because the EM iteration converges to the MLE (even when converging to local optimums), and MLE solutions in general show much better performance under some mild conditions. (*e.g.,* with Assumption E.1, or if random perturbations are applied to the underlying LMAB model [8]).

In our second experiment, we test the performance of the different methods while scanning different control parameters. We first fix $A = 50$, $H = 7$, and change the number of contexts (left of Figure 2). Since we keep generating instances satisfying rank $\left(\sum_{m=1}^M w_m \mu_m \mu_m^\top\right) \approx 4$, tensor-based method performs significantly worse as M increases. We then fix $M = 5$ and observe the performance gain by

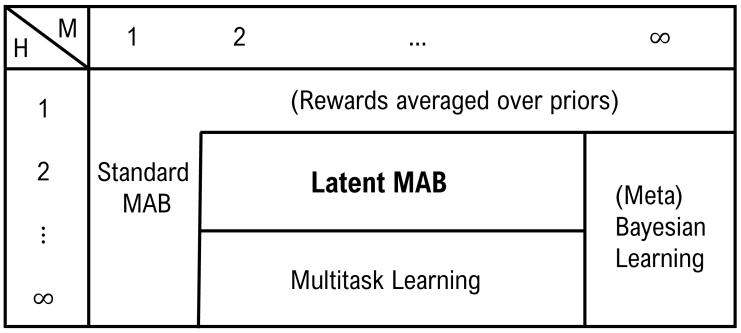

Figure 3: Nature of the problem depends on the length of episodes (H) and number of different contexts (M)

increasing the sample episodes (right of Figure 2). We see that the performance of both L-UCRL+EM and our method (ED+MLE) approaches to that of genie as we increase N.

From both experiments, we conclude that, in practice, MLE based methods perform much better than the worst case guarantee shown via method of moments. We believe it will be an interesting future direction whether we can derive much better guarantees for MLE based solutions.

Appendix B Additional Preliminaries

B.1 Additional Related Work

Comparison of Regimes Nature of learning a near-optimal policy in LMABs might change depending on the length of episodes H and the number of contexts M (recall Figure 3). For instance, when $H = 1$ or $M = 1$, the problem of learning in LMAB is essentially equivalent to the classical Multi-Armed Bandits problem that has been extensively studied in literature (*e.g.,* see [26] and references therein). In another well-studied literature of Bayesian learning, each episode is assumed to have a structured prior over contexts ($M \to \infty$, *e.g.,* all expected rewards of arms are independently sampled from beta conjugate priors). Depending on the length of time-horizon H, the problem can be solved with Thompson Sampling (TS) [38] $H \to \infty$, or the problem is reduced to learning the prior [21]. When the time-horizon is sufficiently long but finite *e.g.,* if $H = \Omega(A/\gamma^2)$ for some separation parameters between a finite number of contexts $M < \infty$, then it is possible cluster observations from every episode. This setting has been studied in the literature of multitask RL which we describe below.

In this work, we are particularly focused on the setting with a few contexts and relatively short episodes (in comparison to A), *i.e.,* $M = O(1)$ and $H = O(1)$, where the most natural objective is to learn a near-optimal history-dependent policy.

Reward-Mixing MDPs Another closely related work is to learn an LMDP (MDP extension of LMAB) with common state transition probabilities, and thus only the reward function changes depending on latent contexts [23]. The authors in [23] have developed an efficient moment-matching based algorithm to learn a near-optimal policy without any assumptions on separations or system dynamics. However, it can only handle the $M = 2$ case with balanced mixing weights $w_1 = w_2 = 1/2$. It is currently not obvious how to extend their result to $M \geq 3$ cases without incurring $O(A^M)$ sample complexity in LMABs.

Regime switching bandits LMAB may be also seen as a special type of adversarial or non-stationary bandits (*e.g.,* [5, 3, 13, 36]) with time steps being specified for when the underlying reward distributions (*i.e.,* latent contexts) may change. The standard objective in non-stationary bandits is to find the best stationary policy in hindsight with unlimited possible contexts. Recently, [41] considered a non-stationary bandit with a finite number of contexts $M = O(1)$ and the objective of finding the optimal *history-dependent* policy. Their setting and goal subsume our goal of learning the optimal policy in LMABs; however, results in [41] require linear independence between reward probability vectors, and thus their setting essentially falls into the category of tractable POMDPs with full-rank observations.

Multitask RL with explicit clustering We can cluster observations from each episode if we are given sufficiently long time horizons $H = \tilde{\Omega}(A/\gamma^2)$ in each episode [9, 14, 16]. Here, $\gamma > 0$ is the amount of 'separation' between contexts such that for all $m \neq m'$, $\max_{a \in \mathcal{A}} \|\mu_m(a, \cdot) - \mu_{m'}(a, \cdot)\|_2 \geq \gamma$ where μ_m is a mean-reward vector of actions in the mth context. We focus on significantly more general cases where there is no obvious way of clustering observations, *e.g.,* when $H \ll A$ or μ_m can be arbitrarily close to some other $\mu_{m'}$ for $m \neq m'$. If we are given a similar separation condition, we also show that a polynomial sample complexity is achievable as long as $H = \tilde{O}(M^2/\gamma^2)$ (see also section E.2). Note that this could still be in $H \ll A$ regime with large number of actions A.

Miscellaneous There are other modeling approaches where multiple episodes proceed in parallel without limits on the time-horizon [31, 14, 10, 17, 25]. In such problems, the goal is to quickly adapt policies for individual episodes assuming certain similarities between tasks. In contrast, in episodic LMAB settings, we assume that every episode starts in a sequential order, and our goal is to learn an optimal history-dependent policy that can maximize rewards for a single episode with limited time horizon.

B.2 Experimental Design

We now give a high-level overview on experimental design techniques used in this work. Suppose we are given a matrix $\Phi \in \mathbb{R}^{d \times k}$ where $d \gg k$. Define a distribution over the rows of Φ, $\rho \in \Delta^d$ an element in the d-dimensional simplex. We want to select a small subset of rows of Φ which minimizes $g(\rho)$ defined below:

$$G(\rho) = \sum_{i \in [d]} \rho(i) \Phi_{i,:} \Phi_{i,:}^\top, \qquad g(\rho) = \max_{i \in [d]} \|\Phi_{i,:}\|^2_{G(\rho)^{-1}}, \tag{10}$$

where $\Phi_{i,:} \in \mathbb{R}^k$ be the ith row of Φ and $G(\rho) \in \mathbb{R}^{k \times k}$. To achieve this task we use results from the experimental design literature. The following theorem shows the existence of ρ that minimizes $g(\rho)$ with *a small support* over the row indices of Φ:

Theorem B.1 (Theorem 4.4 in [27]) *There exists a probability distribution ρ such that $g(\rho) \leq 2k$ and $|supp(\rho)| \leq 4k \log\log k + 16$. Furthermore, we can compute such ρ in time $\tilde{O}(dk^2)$.*

As noted in [27], Theorem B.1 can be obtained from results of Chapter 3 in [39]. Using this fundamental theorem, [27] showed the following proposition:

Proposition B.2 (Proposition 4.5 in [27]) *Let ρ be a distribution over the rows of Φ that satisfies the condition of Theorem B.1. Suppose a vector $\mu \in \mathbb{R}^d$ can be represented as a sum $\mu = v + \Delta$ where $v \in V$, $\|\Delta\|_\infty \leq \epsilon_0$. Let η be any small noise with $\eta \in [-\epsilon_1, \epsilon_1]^d$. Then $\|\Phi\hat{\theta} - \mu\|_\infty \leq \epsilon_0 + (\epsilon_0 + \epsilon_1)\sqrt{2k}$ where*

$$\hat{\theta} = G(\rho)^{-1} \sum_{i \in [d]} \rho(i)(\mu(i) + \eta(i))\Phi_{i,:}. \tag{11}$$

Crucially, we use Proposition B.2 to reduce the sample complexity in A.

B.3 Wasserstein Distance

We now give a brief overview on the Wasserstein distance and its applications in latent mixture models. Wasserstein distance is a convenient error metric to measure the parameter distance between two latent models $\{(w_m, \nu_m)\}_{m=1}^M$ and $\{(\hat{w}_m, \hat{\nu}_m)\}_{m=1}^M$, where $w_m, \hat{w}_m$ are mixing probabilities and $\nu_m, \hat{\nu}_m$ are some parameters for individual contexts.

Wasserstein distance is defined as follows. Let ν be a finite-support distribution over $\{\nu_m\}_{m=1}^M$ with probabilities $\{w_m\}_{m=1}^M$, *i.e.,* $\nu = \sum_{m=1}^M w_m \delta_{\nu_m}$ where δ_v is a Direc-delta distribution with a single mass on $v \in \mathbb{R}^n$. Similarly with parameters $\{(\hat{w}_m, \hat{\nu}_m)\}_{m=1}^M$, define an atomic distribution

$\hat{\nu} = \sum_m \hat{w}_m \delta_{\hat{\nu}_m}$. We define a Wasserstein distance between ν and $\hat{\nu}$ with respect to l_∞ norm as the following:

$$W(\nu, \hat{\nu}) := \inf_{\Gamma} \mathbb{E}_{(m,m')\sim\Gamma} \left[\|\nu_m - \hat{\nu}_{m'}\|_\infty\right] = \inf_{\Gamma} \sum_{(m,m')\in[M]^2} \Gamma(m, m') \cdot \|\nu_m - \hat{\nu}_{m'}\|_\infty,$$

where the infimum is taken over all couplings over joint distributions ν and $\hat{\nu}$ which are marginally distributed as ν and $\hat{\nu}$ respectively (identical to (3)), *i.e.,*

$$\Gamma(m, m') \in \mathbb{R}_+^{M\times M} : \sum_{m'=1}^{M} \Gamma(m, m') = w_m, \sum_{m=1}^{M} \Gamma(m, m') = \hat{w}_m.$$

One nice property of Wasserstein metric is that the distance measure is invariant to permutation of individual components, and flexible with arbitrarily small mixing probabilities or close parameters for different contexts [40, 12].

Appendix C Proofs for Section 3

C.1 Proof of Lemma 3.1

Let $\{\beta_j\}_{j=1}^k \subseteq \mathbb{R}^d$ be orthonormal basis of a k-dimensional subspace $\mathbf{U}$. Let $\Phi \in \mathbb{R}^{d\times k}$ be a matrix of form $[\beta_1 \quad \beta_2 \quad \dots \quad \beta_k]$, *i.e.,* the jth column of Φ is β_j. We need to show the existence of a small set of core coordinates, from which we can reconstruct $\mu \in \mathbf{U}$.

Proposition B.2 implies that we can find a set of coordinates $\{i_j\}_{j=1}^n$ with cardinality at most $n = 4d \log\log d + 16$, such that if $\mu \in \mathbf{U}$ and we can access the vector $[\mu(i_1), \dots, \mu(i_n)]$ up to accuracy ϵ_1, then we can reconstruct μ up to $\epsilon_1\sqrt{2k}$ error. We can also infer that for any $\mu \in \mathbf{U}$, we have $\Phi\theta = \mu$ where θ is given by $G(\rho)^{-1} \sum_{i\in[d]} \rho(i)\mu(i)a_i$.

To conclude Lemma 3.1, suppose we find $\hat{\mu} \in \mathbf{U}$ such that $|\mu(i_j) - \hat{\mu}(i_j)| \le \epsilon_0$ for all $j \in [n]$. Let η be such that $\eta(i) = \mu(i) - \hat{\mu}(i)$ if $i \in \{i_j\}_{j=1}^n$, and $\eta(i) = 0$ otherwise. Applying Proposition B.2 with $\Delta = 0$ and η be as defined above, with $\hat{\mu} = \Phi\hat{\theta}$, we have $\|\hat{\mu} - \mu\|_\infty \le \epsilon_0\sqrt{2k}$.

C.2 Proof of Corollary 3.2

Let $\{\beta_m\}_{m=1}^M \subseteq \mathbb{R}^{AZ}$ be orthonormal basis of a M-dimensional subspace that includes $\mathbf{U} = span(\{\mu_m\}_{m=1}^M)$. Then by Lemma 3.1, we have

$$\begin{aligned}\max_{a\in\mathcal{A}} \|\mu_m(a, \cdot) - \hat{\mu}_m(a, \cdot)\|_1 &= \max_{a\in\mathcal{A}} \sum_{z\in\mathcal{Z}} |\mu_m(a, z) - \hat{\mu}_m(a, z)| \\ &\le 2Z\sqrt{2M} \max_{j\in[n]} |\mu_m(a_j, z_j) - \hat{\mu}_m(a_j, z_j)|,\end{aligned}$$

with a set of core action-value pairs $\{(a_j, z_j)\}_{j=1}^n$. Plugging $v_m(j) = \mu_m(a_j, z_j)$ and $\hat{\nu}_m(j) = \hat{\mu}_m(a_j, z_j)$ gives Corollary 3.2.

Remark C.1 (Eliminating factor Z for $H \ge M$) *Instead of core action-value pairs, we can also define core action-event pairs to save a factor of Z. For instance, define a basis $\{\tilde{\beta}_m\}_{m=1}^M$ in a lifted space $\mathbb{R}^{A\times 2^Z}$ defined as the following:*

$$\tilde{\beta}_m(a, S) := \sum_{z\in S} \beta_m(a, z), \qquad \forall a \in \mathcal{A},\ S \subseteq \mathcal{Z}.$$

For each μ_m, define ϕ_m:

$$\phi_m(a, S) := \sum_{z\in S} \mu_m(a, z), \qquad \forall a \in \mathcal{A}, S \subseteq \mathcal{Z}. \tag{12}$$

By definition, span$(\{\phi_m\}_{m=1}^M) \subseteq$ *span*$(\{\tilde{\beta}_m\}_{m=1}^M)$. *Since* $\frac{1}{2}\|\mu_m(a,\cdot) - \hat{\mu}_m(a,\cdot)\|_1 = \max_{S\subseteq\mathcal{Z}} |\phi_m(a,S) - \hat{\phi}_m(a,S)|$, *it follows from Lemma 3.1 that we can find a set of core action-events* $\{(a_j, \mathcal{Z}_j)\}_{j=1}^n$ *of size at most* $n = 4M\log\log M + 16$ *such that for any* $\phi_m, \hat{\phi}_m \in V$, *we have*

$$\|\phi_m - \hat{\phi}_m\|_\infty \le \sqrt{2M} \cdot \max_{j\in[n]} |\phi_m(a_j, \mathcal{Z}_j) - \hat{\phi}_m(a_j, \mathcal{Z}_j)|.$$

This approach enables to approximate reward distributions in l_1 *statistical distance, and thus we can save a factor of* Z *in subsequent analysis. However, it comes with exponentially more expensive (in* Z*) computations for finding the core set.*

C.3 Proof of Lemma 3.3

To prove this result, we use the recent results on converting closeness in higher-order moments to closeness in Wasserstein distances for atomic distributions [40, 12]. A key result in [40] states a connection between moments and Wasserstein distance in one-dimensional case:

Theorem C.2 (Proposition 1 in [40]) *If* $n = 1$ *and* $\left|\sum_{m=1}^M w_m \nu_m^d - \sum_{m=1}^M \hat{w}_m \hat{\nu}_m^d\right| < \delta$ *for all* $d = 1, 2, ..., 2M-1$, *then*

$$W(\gamma, \hat{\gamma}) \le O\left(M\delta^{1/(2M-1)}\right).$$

Theorem C.2, which holds for 1-dimensional mixtures–can be generalized to the high-dimensional case, as the following theorem shows.

Theorem C.3 *For* $n \ge 2$, *if* $\left\|\sum_{m=1}^M w_m \nu_m^d - \sum_{m=1}^M \hat{w}_m \hat{\nu}_m^d\right\|_\infty < \delta$ *for all* $d = 1, 2, ..., 2M-1$, *then*

$$W(\gamma, \hat{\gamma}) \le O\left(M^3 n \cdot \delta^{1/(2M-1)}\right).$$

Proof. The idea follows the proof of Lemma 3.1 in [12]. Suppose a standard Gaussian random variable $\theta \sim \mathcal{N}(0, I)$ in $\mathbb{R}^n$. For any $x \in \mathbb{R}^n$, anti-concentration of Normal distribution says

$$\mathbb{P}(|\theta^\top x| \le \tau\|x\|_2) \le \tau,$$

for any $\tau > 0$. By union bound and the fact that $\|x\|_2 \ge \|x\|_\infty$, if we define $\mathcal{X} = \{v - \hat{v} | v \in \{\nu_m\}_{m=1}^M, \hat{v} \in \{\hat{\nu}_m\}_{m=1}^M\}$, then we have

$$\mathbb{P}(|\theta^\top x| \le \tau\|x\|_\infty) \le M^2\tau, \qquad \forall x \in \mathcal{X}.$$

Also with high probability, we have $\mathbb{P}(\|\theta\|_1 \ge 2n) \le \frac{n\mathbb{E}_{t\sim\mathcal{N}(0,1)}[|t|]}{2n} < 1/2$ by Markov inequality. Thus,

$$\mathbb{P}\left(\frac{\theta^\top x}{\|\theta\|_1} > \frac{\tau}{2n}\|x\|_\infty\right) > 1/2 - M^2\tau.$$

By setting $\tau = M^2/2$, this probabilistic argument implies that there exists $\theta \in \mathbb{R}^n$ with a unit L_1 norm, $\|\theta\|_1 = 1$, such that for all $x \in \mathcal{X}$, we have $\|x\|_\infty \le 4M^2 n|\theta^\top x|$. Using this θ, we have

$$W(\gamma, \hat{\gamma}) = \inf_\Gamma\{\mathbb{E}_{(m,m')\sim\Gamma}[\|\nu_m - \hat{\nu}_{m'}\|_\infty]\} \le 4M^2 n \cdot \inf_\Gamma\{\mathbb{E}_{(m,m')\sim\Gamma}[|\theta^\top(\nu_m - \hat{\nu}_{m'})|]\}. \quad (13)$$

On the other hand, consider 1-dimensional M-atomic distributions $\gamma_\theta := \sum_{m=1}^M w_m \delta_{\theta^\top\nu_m}$ and $\hat{\gamma}_\theta := \sum_{m=1}^M \hat{w}_m \delta_{\theta^\top\hat{\nu}_m}$. Then by Cauchy-Schwartz inequality

$$\left|\sum_{m=1}^M w_m(\theta^\top\nu_m)^d - \sum_{m=1}^M \hat{w}_m(\theta^\top\hat{\nu}_m)^d\right| \le \|\theta\|_1^d \left\|\sum_{m=1}^M w_m \nu_m^{\bigotimes d} - \sum_{m=1}^M \hat{w}_m \hat{\nu}_m^{\bigotimes d}\right\|_\infty \le \delta,$$

for all $d = 1, 2, ..., 2M-1$. Since $\gamma_\theta, \hat{\gamma}_\theta$ are 1-dimensional atomic distributions, by Theorem C.2,

$$W(\gamma_\theta, \hat{\gamma}_\theta) := \inf_\Gamma\{\mathbb{E}_{(m,m')\sim\Gamma}[|(\theta^\top\nu_m) - (\theta^\top\hat{\nu}_{m'})|]\} \le O\left(M\delta^{1/(2M-1)}\right).$$

Note that any coupling over $\nu_\theta, \hat{\nu}_\theta$ can be converted to a coupling over $\nu, \hat{\nu}$ and vice versa. Plugging the above result into (13), we have the theorem. □

With the above theorems, if we set $\delta < (\epsilon/(M^3 n))^{2M-1}$, we have $W(\gamma, \hat{\gamma}) \le O(\epsilon)$.

C.4 Proof of Proposition 3.4

First note that the difference in expected values of a policy is bounded by total variation distance between trajectory distribution:

$$
\begin{aligned}
&|V(\pi) - \hat{V}(\pi)| \\
&\leq H \cdot \sum_{\substack{r_{1:H} \in \mathcal{R}^H, \\ a_{1:H} \in \mathcal{A}^H}} |\mathbb{P}^\pi(r_{1:H}, a_{1:H}) - \hat{\mathbb{P}}^\pi(r_{1:H}, a_{1:H})| \\
&\leq H \sum_{r_{1:H}, a_{1:H}} \left| \sum_{m=1}^M w_m \Pi_{t=1}^H \mu_m(a_t, r_t) - \sum_{m=1}^M \hat{w}_m \Pi_{t=1}^H \hat{\mu}_m(a_t, r_t) \right| \cdot \Pi_{t=1}^H \pi(a_t | a_{1:t-1}, r_{1:t-1}).
\end{aligned}
$$

Now we only need to focusing on total variation distance bound. Fron this point, when we sum over sequences, we sum over all possible sequences if not specified. With a slight abuse in notation, we use a compact notation for probability of action sequences:

$$
\pi(a_{1:t} | r_{1:t-1}) \coloneqq \Pi_{t'=1}^t \pi(a_{t'} | r_{1:t'-1}, a_{1:t'-1}). \tag{14}
$$

First, suppose any coupling Γ between contexts m and m' in two systems, and we can write

$$
\begin{aligned}
&\sum_{r_{1:H}, a_{1:H}} \left| \sum_{m=1}^M w_m \Pi_{t=1}^H \mu_m(a_t, r_t) - \sum_{m=1}^M \hat{w}_m \Pi_{t=1}^H \hat{\mu}_m(a_t, r_t) \right| \cdot \pi(a_{1:H} | r_{1:H-1}) \\
&\qquad \leq \sum_{r_{1:H}, a_{1:H}} \pi(a_{1:H} | r_{1:H-1}) \sum_{(m,m')} \Gamma(m, m') \left| \Pi_{t=1}^H \mu_m(a_t, r_t) - \Pi_{t=1}^H \hat{\mu}_{m'}(a_t, r_t) \right|,
\end{aligned}
$$

where the inequality holds by the triangle inequality. Now we can proceed as

$$
\begin{aligned}
&\sum_{(m,m')} \Gamma(m, m') \sum_{r_{1:H}, a_{1:H}} \pi(a_{1:H} | r_{1:H-1}) \left| \Pi_{t=1}^H \mu_m(a_t, r_t) - \Pi_{t=1}^H \hat{\mu}_{m'}(a_t, r_t) \right| \\
&\leq \sum_{(m,m')} \Gamma(m, m') \cdot \Bigg(\sum_{r_{1:H-1}, a_{1:H}} \pi(a_{1:H} | r_{1:H-1}) \cdot \Pi_{t=1}^{H-1} \mu_m(a_t, r_t) \sum_{r_H} |\mu_m(r_H; a_H) - \hat{\mu}_{m'}(a_H, r_H)| \\
&\quad + \sum_{r_{1:H-1}, a_{1:H}} \pi(a_{1:H} | r_{1:H-1}) \cdot \left| \Pi_{t=1}^{H-1} \mu_m(a_t, r_t) - \Pi_{t=1}^{H-1} \hat{\mu}_{m'}(a_t, r_t) \right| \cdot \sum_{r_H} |\hat{\mu}_{m'}(a_H, r_H)| \Bigg),
\end{aligned}
$$

where we used triangle inequality at the $t = H$ time step. Note that $\sum_{r_H} |\mu_m(a_H, r_H) - \hat{\mu}_m(a_H, r_H)| = \|\mu(a_H, \cdot) - \hat{\mu}_m(a_H, \cdot)\|_1$. Thus, we can bound the first term by

$$
\begin{aligned}
&\sum_{r_{1:H-1}, a_{1:H}} \pi(a_{1:H} | r_{1:H-1}) \cdot \Pi_{t=1}^{H-1} \mu_m(a_t, r_t) \sum_{r_H} |\mu_m(a_H, r_H) - \hat{\mu}_{m'}(a_H, r_H)| \\
&\qquad \leq \max_{a \in \mathcal{A}} \|\mu_m(a, \cdot) - \hat{\mu}_{m'}(a, \cdot)\|_1.
\end{aligned}
$$

where we summed over all probabilities over sequences where we used the fact that

$$
\sum_{r_{1:H-1}, a_{1:H}} \pi(a_{1:H} | r_{1:H-1}) \cdot \Pi_{t=1}^{H-1} \mu_m(a_t, r_t) = \sum_{r_{1:H-1}, a_{1:H}} \mathbb{P}_m^\pi(r_{1:H-1}, a_{1:H}) = 1.
$$

For the second term, we observe that

$$
\begin{aligned}
&\sum_{r_{1:H-1}, a_{1:H}} \pi(a_{1:H} | r_{1:H-1}) \left| \Pi_{t=1}^{H-1} \mu_m(a_t, r_t) - \Pi_{t=1}^{H-1} \mu_{m'}(a_t, r_t) \right| \sum_{r_H} |\hat{\mu}_{m'}(a_H, r_H)| \\
&\qquad = \sum_{r_{1:H-1}, a_{1:H-1}} \pi(a_{1:H-1} | r_{1:H-2}) \left| \Pi_{t=1}^{H-1} \mu_m(a_t, r_t) - \Pi_{t=1}^{H-1} \hat{\mu}_{m'}(a_t, r_t) \right|.
\end{aligned}
$$

since $\sum_{r_H} \hat{\mu}_{m'}(a, r_H) = 1$ for any a. From here, we can apply the same decomposition which we used at $t = H$ level, and apply the same argument recursively until $t = 1$. That gives for any coupling Γ,

$$
\sum_{r_{1:H}, a_{1:H}} \left| \sum_{m=1}^M w_m \Pi_{t=1}^H \mu_m(a_t, r_t) - \sum_{m=1}^M \hat{w}_m \Pi_{t=1}^H \hat{\mu}_m(a_t, r_t) \right| \cdot \pi(a_{1:H} | r_{1:H-1})
$$

$$\leq H \cdot \sum_{(m,m')} \Gamma(m,m') \max_{a\in\mathcal{A}} \|\mu_m(a,\cdot) - \hat{\mu}_{m'}(a,\cdot)\|_1.$$

Since the inequality holds for all valid couplings $\Gamma(m,m')$ such that

$$\sum_m \Gamma(m,m') = \hat{w}_{m'}, \text{and, } \sum_{m'} \Gamma(m,m') = w_m,$$

we can take the infimum over Γ, and conclude that

$$|V(\pi) - \hat{V}(\pi)| \leq H^2 \cdot \inf_{\Gamma} \sum_{(m,m')} \Gamma(m,m') \max_{a\in\mathcal{A}} \|\mu_m(a,\cdot) - \hat{\mu}_{m'}(a,\cdot)\|_1.$$

C.5 Proof of Lemma 3.5

In order to bound difference between expected long-term rewards for a fixed history-dependent policy π, it is sufficient to bound the difference in distributions of observations following π. We first explicitly write down total variation distance between observations from $\{(w_m,\mu_m)\}_{m=1}^M$ and $\{(\hat{w}_m,\hat{\mu}_m)\}_{m=1}^M$:

$$\begin{aligned}
&\sum_{r_{1:H},a_{1:H}} |\mathbb{P}^\pi(r_{1:H},a_{1:H}) - \hat{\mathbb{P}}^\pi(r_{1:H},a_{1:H})| \\
&\quad= \sum_{r_{1:H},a_{1:H}} \left| \sum_{m=1}^M w_m \Pi_{t=1}^H \mu_m(a_t,r_t) - \sum_{m=1}^M \hat{w}_m \Pi_{t=1}^H \hat{\mu}_m(a_t,r_t) \right| \pi(a_{1:H}|r_{1:H-1}) \\
&\quad\leq \delta \sum_{r_{1:H},a_{1:H}} \Pi_{t=1}^H \pi(a_t|a_{1:t-1},r_{1:t-1}) = Z^H \cdot \delta,
\end{aligned}$$

where $\pi(a_{1:H}|r_{1:H-1})$ is as defined in (14), in the first inequality we used the condition $\|\sum_{m=1}^M w_m \mu_m^{\bigotimes H} - \sum_{m=1}^M \hat{w}_m \hat{\mu}_m^{\bigotimes H}\|_\infty \leq \delta$, and the last inequality follows from that:

$$\begin{aligned}
\sum_{r_{1:H},a_{1:H}} \pi(a_{1:H}|r_{1:H-1}) &= \sum_{r_{1:H}} \sum_{a_{1:H}} \Pi_{t=1}^H \pi(a_t|a_{1:t-1},r_{1:t-1}) \\
&= \sum_{r_{1:H}\in\mathcal{R}^H} \sum_{a_{1:H-1}\in\mathcal{A}^{H-1}} \Pi_{t=1}^{H-1} \pi(a_t|a_{1:t-1},r_{1:t-1}) \sum_{a_H\in\mathcal{A}} \pi(a_H|a_{1:H-1},r_{1:H-1}) \\
&= \sum_{r_{1:H}\in\mathcal{R}^H} \sum_{a_{1:H-1}\in\mathcal{A}^{H-1}} \Pi_{t=1}^{H-1} \pi(a_t|a_{1:t-1},r_{1:t-1}) = \ldots = \sum_{r_{1:H}\in\mathcal{R}^H} 1 = Z^H.
\end{aligned}$$

Now the difference in expected rewards follows as

$$|f(\pi) - \hat{f}(\pi)| \leq H \cdot \sum_{r_{1:H},a_{1:H}} |\mathbb{P}^\pi(r_{1:H},a_{1:H}) - \hat{\mathbb{P}}^\pi(r_{1:H},a_{1:H})| \leq HZ^H \cdot \delta.$$

C.6 Proof of Proposition 3.6

To show this result, we first express each coordinate of μ_m in terms of ν_m. That is, for all $(a,z) \in \mathcal{A}\times\mathcal{Z}$, using Proposition B.2, by setting $\eta = 0$ we can express μ_m as

$$\mu_m = T\nu_m,$$

for some linear mapping $T \in \mathbb{R}^{AZ\times n}$ (see Section C.8.1 for details on T). Furthermore, from the conclusion of Proposition B.2 which implies the robustness of μ_m against the perturbation of ν_m in l_∞ norm, we can infer that every row of T has l_1 norm bounded by $\sqrt{2M}$, *i.e.,* $\|T_{(a,z),:}\|_1 \leq \sqrt{2M}$ for all $(a,z) \in \mathcal{A}\times\mathcal{Z}$.

To bound $\|\sum_{m=1}^M w_m \mu_m^{\bigotimes l} - \sum_{m=1}^M \hat{w}_m \hat{\mu}_m^{\bigotimes l}\|_\infty$, we only need to check one entry of tensors at any position $((a_1,z_1),\ldots,(a_l,z_l))$ and all other entries are bounded in a similar fashion. We first check that

$$\sum_{m=1}^M w_m \cdot \Pi_{i=1}^l \mu_m(a_i,z_i) - \sum_{m=1}^M \hat{w}_m \cdot \Pi_{i=1}^l \hat{\mu}_m(a_i,z_i)$$

$$= \sum_{m=1}^{M} w_m \cdot \Pi_{i=1}^{l} \left(\sum_{j=1}^{n} \nu_m(j) T_{(a_i,z_i),j} \right) - \sum_{m=1}^{M} \hat{w}_m \cdot \Pi_{i=1}^{l} \left(\sum_{j=1}^{n} \hat{\nu}_m(j) T_{(a_i,z_i),j} \right),$$

Unfolding the product expression over i for the original parameter part ν,

$$\sum_{m=1}^{M} w_m \cdot \Pi_{i=1}^{l} \left(\sum_{j=1}^{n} \nu_m(j) T_{(a_i,z_i),j} \right) = \sum_{m=1}^{M} w_m \sum_{(j_1,j_2,\ldots,j_l)\in[n]^l} \Pi_{k=1}^{l} \nu_m(j_k) \cdot \Pi_{k=1}^{l} T_{(a_k,z_k),j_k}$$

$$= \sum_{(j_1,j_2,\ldots,j_l)\in[n]^l} \Pi_{k=1}^{l} T_{(a_k,z_k),j_k} \cdot \sum_{m=1}^{M} w_m \Pi_{k=1}^{l} \nu_m(j_k).$$

Plugging this expression, we conclude the proof:

$$\left| \sum_{m=1}^{M} w_m \cdot \Pi_{i=1}^{l} \mu_m(a_i,z_i) - \sum_{m=1}^{M} \hat{w}_m \cdot \Pi_{i=1}^{l} \hat{\mu}_m(a_i,z_i) \right|$$

$$= \left| \sum_{(j_1,j_2,\ldots,j_l)\in[n]^l} \Pi_{k=1}^{l} T_{(a_k,z_k),j_k} \cdot \left(\sum_{m=1}^{M} w_m \Pi_{k=1}^{l} \nu_m(j_k) - \sum_{m=1}^{M} \hat{w}_m \Pi_{k=1}^{l} \hat{\nu}_m(j_k) \right) \right|$$

$$\leq \delta \cdot \sum_{(j_1,j_2,\ldots,j_l)\in[n]^l} \left| \Pi_{k=1}^{l} T_{(a_k,z_k),j_k} \right| = \delta \cdot \Pi_{k=1}^{l} \left(\sum_{j=1}^{n} |T_{(a_k,z_k),j}| \right) \leq \delta (2M)^{l/2},$$

which is what we needed to show.

C.7 Proof of Lemma 3.7

For each episode k where $k \in [N_0]$, let the first and second actions be $a_1^k, a_2^k \sim \text{Unif}(\mathcal{A})$, and let r_1^k, r_2^k be the corresponding reward feedback. We construct an empirical second-order moments $\hat{M}_2$ such that $\hat{M}_2(i,j)$ is the mean of $r_1 \cdot r_2$ when $a_1^k = a_i, a_2^k = a_j$. Specifically, we construct $\hat{M}_2$ as the following:

$$\hat{M}_2 = \frac{1}{2N_0} \sum_{k=1}^{N_0} \boldsymbol{e}_{(a_1^k,r_1^k)} \cdot \boldsymbol{e}_{(a_2^k,r_2^k)}^{\top} + \boldsymbol{e}_{(a_2^k,r_2^k)} \cdot \boldsymbol{e}_{(a_1^k,r_1^k)}^{\top}, \tag{15}$$

where $\boldsymbol{e}_{(a_t^k,r_t^k)}$ is a standard basis vector in $\mathbb{R}^{AZ}$ with 1 at position (a_t^k, r_t^k). The argument follows from a rather standard concentration argument for dimensionality reduction (e.g., Lemma 3.5 in [12]). Let $\delta = \|M_2 - \hat{M}_2\|_2$. With standard measure of concentration arguments, we can show that $\|M_2 - \hat{M}_2\|_\infty < C \cdot \sqrt{A^2 \log(AZ/\eta)/N_0}$ with probability at least $1-\eta$, which is translated to $\delta \leq C \cdot \sqrt{\frac{A^4 Z^2 \log(AZ/\eta)}{N_0}}$ for some universal constant $C > 0$.

Let $P_{\mathbf{U}}$ be the orthogonal projector onto the top-M eigenspace of M_2 and $P_{\mathbf{U}}^{\perp} = I - P_{\mathbf{U}}$. We can also define similar quantities from the empirical estimate $\hat{M}_2$. Let $P_{\widehat{\mathbf{U}}}$ similarly be the orthogonal projector onto the top-M eigenspace of $\hat{M}_2$, and let $P_{\widehat{\mathbf{U}}^{\perp}} = I - P_{\widehat{\mathbf{U}}}$. By Weyl's theorem, we have $\lambda_{M+1}(\hat{M}_2) \leq \delta$ since $\text{rank}(M_2)$ is M. Our goal is to bound

$$\|\mu_m - P_{\widehat{\mathbf{U}}} \mu_m\|_2^2 = \|P_{\widehat{\mathbf{U}}^{\perp}} \mu_m\|_2^2 = \max_{y \in Range(\widehat{\mathbf{U}}^{\perp}) \cap \mathbb{S}^{AZ-1}} (y^{\top} \mu_m)^2.$$

for any $m \in [M]$. Let y be the vector maximizing the above. Since $w_m \mu_m \mu_m^{\top} \preceq M_2$ and since y belongs to the eigenspace of ranks lower than M,

$$w_m y^{\top} \mu_m \mu_m^{\top} y \leq y^{\top} M_2 y = y^{\top}(M_2 - \hat{M}_2) y + y^{\top} \hat{M}_2 y \leq \|M_2 - \hat{M}_2\|_2 + \delta \leq 2\delta.$$

We have shown that $\|\mu_m - \widehat{\mathbf{U}} \mu_m\|_2 \leq \sqrt{2\delta / w_m}$. Thus we need $\delta := O(\delta_{\text{sub}}^2)$ to bound the l_2 error by $\delta_{\text{sub}} / w_m^{1/2}$. Thus $N_0 = O\left(Z^2 A^4 \log(AZ/\eta) / \delta_{\text{sub}}^4\right)$ samples are sufficient for the estimation of the subspace $\mathbf{U}$. Note that we have not optimized for the polynomial factors which can be improved by more tightly bounding $\delta = \|M_2 - \hat{M}_2\|_2$ from N_0 samples.

C.8 Deferred Details of Theorem 3.8

The proof of Theorem 3.8 follows by collecting the results from the preceding lemmas for subspace estimation (Lemma 3.7) and results from Section 3.2 and 3.3. Before we get into the proof, let us describe some details about finding a set of core action-value pairs, and how to recover the original model parameters $\{\mu_m\}_{m=1}^M$ from $\{\nu_m\}_{m=1}^M$.

C.8.1 Detailed Procedures for Experimental Design

Step 1. Subspace estimation We first find the set of core action-value pairs $\{(a_j, z_j)\}_{j=1}^n$ following the same procedure in Corollary 3.2. By Lemma 3.7, for every μ_m, we have

$$\mu_m + \Delta_m \in \widehat{\mathbf{U}},$$

where $\|\Delta_m\|_\infty \le \delta_{\text{sub}}/w_m^{1/2}$.

Step 2. Pick core action-event pairs Let $\{\hat{\beta}_j\}_{j=1}^M$ be the orthonormal basis of $\widehat{\mathbf{U}}$. We can use this basis as input to Corollary 3.2 to get a set of (approximate) core action-event pairs. Specifically, let $\hat{\Phi} \in \mathbb{R}^{AZ\times M}$ be a feature matrix where each jth column $\hat{\Phi}_{:,j}$ is given as:

$$\hat{\Phi}_{:,j}(a,z) := \hat{\beta}_j(a,z), \qquad \forall a \in \mathcal{A}, z \in \mathcal{Z}. \tag{16}$$

After invoking Theorem B.1 with supplying $\hat{\Phi}$ as input, we use the support of ρ as the set of core action-value pairs $\{(a_j, z_j)\}_{j=1}^n$. Let $\hat{G}(\rho)$ be defined as in equation (10). Note that with too small mixing weights w_m, we can instead use $\Delta_m = -\mu_m$ with $\|\Delta_m\|_\infty \le 1$.

Step 3. Search constraints for moment-matching After finding a core action-event pairs $\{(a_j, z_j)\}_{j=1}^n$ as in Corollary 3.2, we estimate ν_m from higher-order tensors $\{\hat{T}_l\}_{l=1}^{\min(H,2M-1)}$. When searching parameters for $\{(\hat{w}_m, \hat{\nu}_m)\}_{m=1}^M$, we can put constraints to ensure that $\hat{w}_m$ and $\hat{\nu}_m$ belong to a set of valid parameters for all $m \in [M]$:

$$\begin{aligned}
&\sum_{m=1}^M \hat{w}_m = 1, \quad w_{\min} \le \hat{w}_m, \quad 0 \preceq \hat{\nu}_m \preceq 1, \\
&\left|\sum_{z\in\mathcal{Z}} (\hat{T}\hat{\nu}_m)(a,z) - 1\right| \le -2Z\sqrt{M}\delta_{\text{sub}}/\hat{w}_m^{1/2}, \\
&-2\sqrt{M}\delta_{\text{sub}}/\hat{w}_m^{1/2} \preceq \hat{T}\hat{\nu}_m \preceq 1 + 2\sqrt{M}\delta_{\text{sub}}/\hat{w}_m^{1/2}, \qquad \forall m \in [M], a \in \mathcal{A},
\end{aligned} \tag{17}$$

where $\preceq$ is an element-wise inequality. That is, we want that $\hat{v}_m = \hat{T}\hat{\nu}_m$ is not too far from $\hat{\mu}_m$ after clipping and normalization. Without loss of generality, we assume that a rough estimate of $w_{\min}$ is known in advance (otherwise, we can repeat the same procedure with geometrically decreasing estimates of $w_{\min}$, *e.g.,* $1/M, 1/2M, 1/4M, ..., 1/M^{O(\min(M,H))}$, and pick the best returned policy). A solution satisfying all constraints is guaranteed to exist since the true model $\{(w_m, \nu_m)\}_{m=1}^M$ also satisfies constraints.

Step 4. Recovery of parameters Let $\hat{\mu}_m$ be computed by (11) using $\hat{\nu}_m$ and $\hat{\Phi}$, *i.e.,* let $\hat{T} \in \mathbb{R}^{AZ\times n}$ be defined as

$$\hat{T}_{:,j} := \rho(a_j, z_j)\hat{\Phi}\hat{G}(\rho)^{-1}\hat{\Phi}_{(a_j,z_j),:} \qquad \forall j \in [n]. \tag{18}$$

We let $v_m = \hat{T}\nu_m, \hat{v}_m = \hat{T}\hat{\nu}_m$. We recover $\hat{\mu}_m$ from $\hat{v}_m = \hat{T}\hat{\nu}_m$ as

$$\begin{aligned}
\tilde{\mu}_m &:= \text{clip}(\hat{v}_m, 0, 1), \\
\hat{\mu}_m(a,z) &:= \frac{\tilde{\mu}_m(a,z)}{\sum_{z'\in\mathcal{Z}} \tilde{\mu}_m(a,z')}, \qquad \forall a \in \mathcal{A}, z \in \mathcal{Z}.
\end{aligned} \tag{19}$$

A simple algebra can show that the normalized estimates of $\hat{\mu}_m$ are close to μ_m. Specifically, we show the following with Proposition B.2:

$$\begin{aligned}
\|\mu_m - \hat{v}_m\|_\infty &\le \|\Delta_m\|_\infty + (\|\Delta_m\|_\infty + \max_{j\in[n]} |\mu_m(a_j,z_j) - \hat{\mu}_m(a_j,z_j)|) \cdot \sqrt{2M} \\
&\le \|\Delta_m\|_\infty + (\|\Delta_m\|_\infty + \|\nu_m - \hat{\nu}_m\|_\infty) \cdot \sqrt{2M}.
\end{aligned}$$

C.8.2 Proof of Theorem 3.8, Case I: $H \geq 2M - 1$

We search the empirical parameters $\{(\hat{w}_m, \hat{\nu}_m)\}_{m=1}^M)$ over the set (17). After finding $\hat{w}_m, \hat{\nu}_m$ that satisfy the moment matching condition (7), we recover $\hat{\mu}_m$ from $\hat{v}_m = \hat{T}\hat{\nu}_m$ after clipping and normalization (19). Then we first observe that for any $a \in \mathcal{A}$,

$$\|\mu_m(a,\cdot) - \tilde{\mu}_m(a,\cdot)\|_1 \leq \sum_{z\in\mathcal{Z}} |\mu_m(a,z) - \hat{v}_m(a,z)| \leq Z \max_{z\in\mathcal{Z}} |\mu_m(a,z) - \hat{v}_m(a,z)|,$$

where in the first inequality, we used the fact that clipping can only improve the l_1 error. Then, errors from the normalization can be bounded as

$$\begin{aligned}
\|\mu_m(a,\cdot) - \hat{\mu}_m(a,\cdot)\|_1 &\leq \|\mu_m(a,\cdot) - \tilde{\mu}_m(a,\cdot)\|_1 + \|\tilde{\mu}_m(a,\cdot) - \hat{\mu}_m(a,\cdot)\|_1 \\
&= \|\mu_m(a,\cdot) - \tilde{\mu}_m(a,\cdot)\|_1 + |\|\tilde{\mu}_m(a,\cdot)\|_1 - 1| \\
&\leq 2\|\mu_m(a,\cdot) - \tilde{\mu}_m(a,\cdot)\|_1 \\
&\leq 2Z \max_{z\in\mathcal{Z}} |\mu_m(a,z) - \hat{v}_m(a,z)| \\
&\leq 2Z \left(\|\Delta_m\|_\infty + (\|\Delta_m\|_\infty + \|\nu_m - \hat{\nu}_m\|_\infty)\sqrt{2M} \right).
\end{aligned}$$

The second line holds since

$$\begin{aligned}
\|\tilde{\mu}_m(a,\cdot) - \hat{\mu}_m(a,\cdot)\|_1 &= \sum_{z\in\mathcal{Z}} \left| \tilde{\mu}_m(a,z) - \frac{\tilde{\mu}_m(a,z)}{\sum_{z\in\mathcal{Z}} \tilde{\mu}_m(a,z)} \right| \\
&= \left| 1 - \frac{1}{\sum_{z\in\mathcal{Z}} \tilde{\mu}_m(a,z)} \right| \cdot \sum_{z\in\mathcal{Z}} |\tilde{\mu}_m(a,z)| \\
&= \left| \sum_{z\in\mathcal{Z}} \tilde{\mu}_m(a,z) - \frac{\sum_{z\in\mathcal{Z}} \tilde{\mu}_m(a,z)}{\sum_{z\in\mathcal{Z}} \tilde{\mu}_m(a,z)} \right| = |\|\tilde{\mu}_m(a,z)\|_1 - 1|.
\end{aligned}$$

where the third relation holds since $\sum_{z\in\mathcal{Z}} |\tilde{\mu}_m(a,z)| \,/\, \left|\sum_{z\in\mathcal{Z}} \tilde{\mu}_m(a,z)\right| = 1$ since $\tilde{\mu}_m(a,z) \geq 0$.

By the choice of δ_{sub}, we have $\|\Delta_m\|_\infty \leq \delta_{\text{sub}}/w_m^{1/2} \leq \epsilon/(2ZMH^2 w_m^{1/2})$. Now we can call Proposition 3.4, and proceed as

$$\begin{aligned}
|V(\pi) - \hat{V}(\pi)| &\leq H^2 \cdot \inf_{\Gamma} \sum_{(m,m')\in[M]^2} \left(\Gamma(m,m') \cdot \max_{a\in\mathcal{A}} \|\mu_m(a,\cdot) - \hat{\mu}_{m'}(a,\cdot)\|_1 \right) \\
&\leq 2ZH^2 \cdot \inf_{\Gamma} \sum_{(m,m')\in[M]^2} \Gamma(m,m') \cdot \left(\sqrt{M/w_m} \cdot \epsilon/(2ZMH^2) + \sqrt{2M} \cdot \|\nu_m - \hat{\nu}_{m'}\|_\infty \right) \\
&\leq 2ZH^2 \cdot \left(\sum_{m\in[M]} w_m \cdot \sqrt{M/w_m} \cdot \epsilon/(2ZMH^2) + \sqrt{2M} \cdot W(\gamma,\hat{\gamma}) \right) \\
&\leq 2ZH^2 \cdot \left(\epsilon/(2H^2) + 2\sqrt{2M} \cdot W(\gamma,\hat{\gamma}) \right),
\end{aligned}$$

where in the last inequality, we used Cauchy-Schwartz inequality $\sum_{m=1}^M \sqrt{w_m} \leq \sqrt{M}$. Hence if we have $2ZH^2\sqrt{2M}W(\gamma,\hat{\gamma}) \leq \epsilon$, which is given by the choice of δ_{sub} and Lemma 3.3, we have $|V(\pi) - \hat{V}(\pi)| \leq O(\epsilon)$.

C.8.3 Proof of Theorem 3.8, Case II: $H < 2M - 1$

If $H < 2M - 1$, we start by observing that for any $0 \preceq \hat{\nu}_m \preceq 1$, using Proposition B.2 and setting $\mu = 0$, $\epsilon_0 = 0$, $\eta = \nu_m$ and $\epsilon_1 = 1$, we have $-\sqrt{2M} \preceq \hat{T}\hat{\nu}_m \preceq \sqrt{2M}$.

To exploit the moment-closeness property, we define an auxiliary model $\{(w_m, v_m)\}_{m=1}^M$ where $v_m := \hat{T}\nu_m$. Similarly to $H \geq 2M - 1$ case, for any $a \in \mathcal{A}$, we have that

$$\|v_m(a,\cdot) - \mu_m(a,\cdot)\|_1 \leq 2Z\|\Delta_m\|_\infty(1 + \sqrt{2M}) \leq Z\sqrt{2M}\delta_{\text{sub}}/w_m^{1/2}. \tag{20}$$

We also have that

$$\|v_m(a,\cdot)\|_1 \le \min\left(Z\sqrt{2M}\delta_{\text{sub}}/w_m^{1/2}, Z\sqrt{2M}\right). \tag{21}$$

Let $\hat{v}_m = \hat{T}\hat{\nu}_m$ and $\hat{\mu}_m$ be defined as in (19). Recall our goal to bound

$$V(\pi) - \hat{V}(\pi) = \sum_{a_{1:H}, r_{1:H}} \left(\sum_{t=1}^{H} r_t\right) \left(\sum_{m=1}^{M} w_m \Pi_{t=1}^{H} \mu_m(a_t, r_t) - \sum_{m=1}^{M} \hat{w}_m \Pi_{t=1}^{H} \hat{\mu}_m(a_t, r_t)\right) \pi(a_{1:H}|r_{1:H-1}).$$

Define auxiliary value functions $V_{\text{aux}}(\pi)$ and $\hat{V}_{\text{aux}}(\pi)$ as

$$V_{\text{aux}}(\pi) := \sum_{a_{1:H}, r_{1:H}} \left(\sum_{t=1}^{H} r_t\right) \sum_{m=1}^{M} w_m \Pi_{t=1}^{H} v_m(a_t, r_t) \pi(a_{1:H}|r_{1:H-1}),$$

$$\hat{V}_{\text{aux}}(\pi) := \sum_{a_{1:H}, r_{1:H}} \left(\sum_{t=1}^{H} r_t\right) \sum_{m=1}^{M} \hat{w}_m \Pi_{t=1}^{H} \hat{v}_m(a_t, r_t) \pi(a_{1:H}|r_{1:H-1}),$$

Then $|V(\pi) - \hat{V}(\pi)| \le |V(\pi) - V_{\text{aux}}(\pi)| + |V_{\text{aux}}(\pi) - \hat{V}_{\text{aux}}(\pi)| + |\hat{V}_{\text{aux}}(\pi) - \hat{V}(\pi)|$. We bound each term separately.

Term I. $|V(\pi) - V_{\text{aux}}(\pi)|$: This is less than

$$\begin{aligned}
|V(\pi) - V_{\text{aux}}(\pi)| &= \left| \sum_{a_{1:H}, r_{1:H}} \left(\sum_{t=1}^{H} r_t\right) \sum_{m=1}^{M} w_m \left(\Pi_{t=1}^{H} \mu_m(a_t, r_t) - \Pi_{t=1}^{H} v_m(a_t, r_t)\right) \pi(a_{1:H}|r_{1:H-1}) \right| \\
&\le H \cdot \sum_{a_{1:H}, r_{1:H}} \left| \sum_{m=1}^{M} w_m \left(\Pi_{t=1}^{H} \mu_m(a_t, r_t) - \Pi_{t=1}^{H} v_m(a_t, r_t)\right) \pi(a_{1:H}|r_{1:H-1}) \right| \\
&\le H \cdot \sum_{a_{1:H-1}, r_{1:H-1}} \sum_{m=1}^{M} w_m \left|\Pi_{t=1}^{H-1} \mu_m(a_t, r_t) - \Pi_{t=1}^{H-1} v_m(a_t, r_t)\right| \sum_{a_H, r_H} |v_m(a_H, r_H)| \pi(a_{1:H}|r_{1:H-1}) \\
&\quad + H \cdot \sum_{a_{1:H-1}, r_{1:H-1}} \sum_{m=1}^{M} w_m \Pi_{t=1}^{H-1} \mu_m(a_t, r_t) \sum_{a_H, r_H} |\mu_m(a_H, r_H) - v_m(a_H, r_H)| \, \pi(a_{1:H}|r_{1:H-1}) \\
&\le H \sum_{m=1}^{M} w_m \max_{a \in \mathcal{A}} \|v_m(a,\cdot)\|_1 \cdot \sum_{a_{1:H-1}, r_{1:H-1}} \left|\Pi_{t=1}^{H-1} \mu_m(a_t, r_t) - \Pi_{t=1}^{H-1} v_m(a_t, r_t)\right| \pi(a_{1:H-1}|r_{1:H-2}) \\
&\quad + H \sum_{m=1}^{M} w_m \max_{a \in \mathcal{A}} \|\mu_m(a,\cdot) - v_m(a,\cdot)\|_1 \cdot \sum_{a_{1:H-1}, r_{1:H-1}} \Pi_{t=1}^{H-1} \mu_m(a_t, r_t) \pi(a_{1:H}|r_{1:H-1}).
\end{aligned}$$

For the second term, we have

$$\left| \sum_{a_{1:H-1}, r_{1:H-1}} \Pi_{t=1}^{H-1} \mu_m(a_t, r_t) \pi(a_{1:H}|r_{1:H-1}) \right| = \left| \sum_{a_{1:H-1}, r_{1:H-1}} \mathbb{P}_m(a_{1:H-1}, r_{1:H-1}) \right| = 1$$

for all $m \in [M]$. For the first term, we can recursively apply the same inequality until the time step reaches to $t = 1$. Applying this recursively,

$$|V(\pi) - V_{\text{aux}}(\pi)| \le H^2 \cdot \sum_{m=1}^{M} w_m \left(\max_{a \in \mathcal{A}} \|v_m(a,\cdot)\|_1\right)^{H-1} \cdot \max_{a \in \mathcal{A}} \|\mu_m(a,\cdot) - v_m(a,\cdot)\|_1.$$

To bound the above, we first consider the case when $w_m \ge \frac{\epsilon}{H^2 M (Z\sqrt{2M})^H}$. Note that if $w_{\min} \ge \frac{\epsilon}{H^2 M (Z\sqrt{2M})^H}$, then this is always the case. In this case, for every $a \in \mathcal{A}$, we have (recall (20))

$$\max_{a \in \mathcal{A}} \|\mu_m(a,\cdot) - v_m(a,\cdot)\|_1 \le 2\sqrt{M} Z \delta_{\text{sub}}/w_m^{1/2} \le \frac{\epsilon}{H^2 \sqrt{M w_m}},$$

$$\max_{a\in\mathcal{A}} \|v_m(a,\cdot)\|_1 \le 1 + \max_{a\in\mathcal{A}} \|v_m(a,\cdot) - \mu_m(a,\cdot)\|_1$$
$$\le 1 + 2\sqrt{M} Z \delta_{\text{sub}}/w_m^{1/2} \le 1 + 1/H,$$

where we use our choice of δ_{sub} in (5). By the second condition,

$$(\max_{a\in\mathcal{A}} \|v_m(a,\cdot)\|_1)^{H-1} \le (1+1/H)^{H-1} \le e.$$

The first condition can be combined with Proposition B.2 similarly to the $H \ge 2M-1$ case to get

$$H^2 \sum_{m: w_m \ge \frac{\epsilon}{H^2 M (Z\sqrt{2M})^H}} w_m \left(\max_{a\in\mathcal{A}} \|v_m(a,\cdot)\|_1\right)^{H-1} \max_{a\in\mathcal{A}} \|\mu_m(a,\cdot) - v_m(a,\cdot)\|_1$$
$$\le e \sum_{m=1}^{M} \sqrt{w_m/M}\,\epsilon \le O(\epsilon).$$

On the other hand, if $w_m < \frac{\epsilon}{H^2 M (Z\sqrt{2M})^H}$, then we can directly bound as

$$H^2 \sum_{m: w_m < \frac{\epsilon}{H^2 M (Z\sqrt{2M})^H}} w_m \left(\max_{a\in\mathcal{A}} \|v_m(a,\cdot)\|_1\right)^{H-1} \max_{a\in\mathcal{A}} \|\mu_m(a,\cdot) - v_m(a,\cdot)\|_1$$
$$\le \frac{\epsilon}{M} \sum_{m=1}^{M} \frac{1}{(Z\sqrt{2M})^H} (Z\sqrt{2M})^H \le O(\epsilon).$$

Thus, we have $|V(\pi) - V_{\text{aux}}(\pi)| \le O(\epsilon)$.

Term II. $|V_{\text{aux}}(\pi) - \hat{V}_{\text{aux}}(\pi)|$: We use the moment-closeness properties between v_m and $\hat{v}_m$ given similarly to Lemma 3.6.

Lemma C.4 *For any given degree $l \ge 1$, if $\|\sum_{m=1}^M w_m \nu_m^{\otimes l} - \sum_{m=1}^M \hat{w}_m \hat{\nu}_m^{\otimes l}\|_\infty \le \delta$, then v_m and $\hat{v}_m$ satisfy*

$$\left\| \sum_{m=1}^{M} w_m v_m^{\otimes l} - \sum_{m=1}^{M} \hat{w}_m \hat{v}_m^{\otimes l} \right\|_\infty \le (2M)^{l/2} \cdot \delta.$$

Proof. Note that $\|\sum_{m=1}^M w_m (\hat{T}\nu_m)^{\otimes l} - \sum_{m=1}^M \hat{w}_m (\hat{T}\hat{\nu}_m)^{\otimes l}\|_\infty \le (2M)^{l/2}\delta$, following the same argument in Appendix C.6: the conclusion of Proposition B.2 also implies that the l_1 norm of every row in $\hat{T}$ is less than $\sqrt{2M}$, *i.e.,*

$$\|\hat{T}_{(a,z),:}\|_1 \le \sqrt{2M}, \qquad \forall (a,z) \in \mathcal{A}\times\mathcal{Z}, \tag{22}$$

Lemma follows since v_m is a vector consisting of partial coordinates of $\hat{T}\nu_m$. □

Now we proceed as

$$|V_{\text{aux}}(\pi) - \hat{V}_{\text{aux}}(\pi)| = \left| \sum_{a_{1:H}, r_{1:H}} \left(\sum_{t=1}^{H} r_t\right) \left(\sum_{m=1}^{M} w_m \Pi_{t=1}^{H} v_m(a_t, r_t) - \sum_{m=1}^{M} \hat{w}_m \Pi_{t=1}^{H} \hat{v}_m(a_t, r_t)\right) \pi(a_{1:H}|r_{1:H-1}) \right|$$
$$\le H \cdot \sum_{a_{1:H}, r_{1:H}} \left| \sum_{m=1}^{M} w_m \Pi_{t=1}^{H} v_m(a_t, r_t) - \sum_{m=1}^{M} \hat{w}_m \Pi_{t=1}^{H} \hat{v}_m(a_t, r_t) \right| \pi(a_{1:H}|r_{1:H-1})$$
$$\le H Z^H (2M)^{H/2} \delta_{\text{tsr}}.$$

Choice of $\delta_{\text{tsr}} = (\epsilon/H)/(\sqrt{2M}Z)^H$ for $H < 2M-1$ gives $|V_{\text{aux}}(\pi) - \hat{V}_{\text{aux}}(\pi)| = O(\epsilon)$.

Term III. $|\hat{V}_{\text{aux}}(\pi) - \hat{V}(\pi)|$: This case is almost similar to the case $|V(\pi) - V_{\text{aux}}(\pi)|$.

$$
\begin{aligned}
|\hat{V}(\pi) - \hat{V}_{\text{aux}}(\pi)| &= \left| \sum_{a_{1:H}, r_{1:H}} \left(\sum_{t=1}^{H} r_t \right) \sum_{m=1}^{M} \hat{v}_m \left(\Pi_{t=1}^{H} \hat{\mu}_m(a_t, r_t) - \Pi_{t=1}^{H} \hat{\mu}_m(a_t, r_t) \right) \pi(a_{1:H} | r_{1:H-1}) \right| \\
&\leq H \cdot \sum_{a_{1:H}, r_{1:H}} \left| \sum_{m=1}^{M} \hat{w}_m \left(\Pi_{t=1}^{H} \hat{v}_m(a_t, r_t) - \Pi_{t=1}^{H} \hat{\mu}_m(a_t, r_t) \right) \pi(a_{1:H} | r_{1:H-1}) \right| \\
&\leq H \cdot \sum_{a_{1:H-1}, r_{1:H-1}} \sum_{m=1}^{M} \hat{w}_m \left| \Pi_{t=1}^{H-1} \hat{v}_m(a_t, r_t) - \Pi_{t=1}^{H-1} \hat{\mu}_m(a_t, r_t) \right| \sum_{a_H, r_H} |\hat{v}_m(a_H, r_H)| \pi(a_{1:H} | r_{1:H-1}) \\
&\quad + H \cdot \sum_{a_{1:H-1}, r_{1:H-1}} \sum_{m=1}^{M} \hat{w}_m \Pi_{t=1}^{H-1} \hat{\mu}_m(a_t, r_t) \sum_{a_H, r_H} |\hat{\mu}_m(a_H, r_H) - \hat{v}_m(a_H, r_H)| \, \pi(a_{1:H} | r_{1:H-1}) \\
&\leq H \cdot \sum_{a_{1:H-1}, r_{1:H-1}} \sum_{m=1}^{M} \hat{w}_m \max_{a \in \mathcal{A}} \|\hat{v}_m(a, \cdot)\|_1 \left| \Pi_{t=1}^{H-1} \hat{\mu}_m(a_t, r_t) - \Pi_{t=1}^{H-1} \hat{v}_m(a_t, r_t) \right| \pi(a_{1:H-1} | r_{1:H-2}) \\
&\quad + H \sum_{m=1}^{M} \hat{w}_m \max_{a \in \mathcal{A}} \|\hat{\mu}_m(a, \cdot) - \hat{v}_m(a, \cdot)\|_1 \sum_{a_{1:H-1}, r_{1:H-1}} \Pi_{t=1}^{H-1} \hat{\mu}_m(a_t, r_t) \pi(a_{1:H} | r_{1:H-1}) \\
&\leq H^2 \cdot \left(\sum_{m=1}^{M} \hat{w}_m \left(\max_{a \in \mathcal{A}} \|\hat{v}_m(a, \cdot)\|_1 \right)^{H-1} \cdot \max_{a \in \mathcal{A}} \|\hat{\mu}_m(a, \cdot) - \hat{v}_m(a, \cdot)\|_1 \right).
\end{aligned}
$$

For each $m \in [M]$, if $\hat{w}_m \geq \frac{\epsilon}{H^2 M (Z\sqrt{2M})^H}$, then

$$
\begin{aligned}
\|\hat{v}_m(a, \cdot)\|_1 &\leq \left| \sum_{z : \hat{v}_m(a,z) < 0} \hat{v}_m(a, z) \right| + \left| \sum_{z : \hat{v}_m(a,z) > 0} \hat{v}_m(a, z) \right| \\
&\leq 1 + 2Z\sqrt{M} \delta_{\text{sub}} / \hat{w}_m^{1/2} \leq 1 + 1/H,
\end{aligned}
$$

where we used the choice of δ_{sub} in (5). We also need to show that $\|\hat{v}_m(a, \cdot) - \hat{\mu}_m(a, \cdot)\|_1$ is bounded. Let $\tilde{\mu}_m$ be the intermediate step after clipping $\hat{v}_m$ before normalization as in 19. Due to the third condition of (17), clipped amount can be at most

$$
\|\hat{v}_m(a, \cdot) - \tilde{\mu}_m(a, \cdot)\|_1 \leq 2Z\sqrt{M} \delta_{\text{sub}} / \hat{w}_m^{1/2}.
$$

With the second condition of (17), we have

$$
\begin{aligned}
\|\tilde{\mu}_m(a, \cdot)\|_1 = \sum_{z \in \mathcal{Z}} \tilde{\mu}_m(a, z) &\leq \left| \sum_{z \in \mathcal{Z}} \tilde{\mu}_m(a, z) - \hat{\mu}_m(a, z) \right| + \left| \sum_{z \in \mathcal{Z}} \hat{\mu}_m(a, z) \right| \\
&\leq \|\hat{v}_m(a, \cdot) - \tilde{\mu}_m(a, \cdot)\|_1 + \|\hat{\mu}_m(a, \cdot)\|_1 \leq 1 + 3Z\sqrt{M} \delta_{\text{sub}} / \hat{w}_m^{1/2}.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\|\tilde{v}_m(a, \cdot)\|_1 = \sum_{z \in \mathcal{R}} \tilde{v}_m(a, z) &\geq \left| \sum_{z \in \mathcal{R}} v_m(a, z) \right| - \left| \sum_{z \in \mathcal{R}} \tilde{v}_m(a, z) - v_m(a, z) \right| \\
&\geq |\hat{\phi}_m(a, \mathcal{R})| - \|\hat{v}_m(a, \cdot) - \tilde{v}_m(a, \cdot)\|_1 \geq 1 - 3Z\sqrt{M} \delta_{\text{sub}} / \hat{w}_m^{1/2}.
\end{aligned}
$$

Therefore, we can show that

$$
\begin{aligned}
\|\hat{\mu}_m(a, \cdot) - \hat{v}_m(a, \cdot)\|_1 &\leq \|\hat{\mu}_m(a, \cdot) - \tilde{v}_m(a, \cdot)\|_1 + \|\tilde{v}_m(a, \cdot) - \hat{v}_m(a, \cdot)\|_1 \\
&\leq \frac{|\|\tilde{v}_m(a, \cdot)\|_1 - 1|}{\|\tilde{v}_m(a, \cdot)\|_1} + 2Z\sqrt{M} \delta_a / \hat{w}_m^{1/2} \\
&\leq 8Z\sqrt{M} \delta_{\text{sub}} / \hat{w}_m^{1/2},
\end{aligned}
$$

where we used $1 - 3Z\delta_{\text{sub}} / \hat{w}_m^{1/2} \geq 1/2$ due to the first constraint and the choice of δ_{sub}.

If $\hat{w}_m < \frac{\epsilon}{H^2 M (Z\sqrt{2M})^H}$, then we can use the fact that all l_1-norm of rows of $\hat{T}$ are less than $\sqrt{2M}$ (equation (22)), and thus $\|\hat{v}_m(a,\cdot)\|_1 \leq \sqrt{2M}$ for all $a \in \mathcal{A}$. We also have that $\|\hat{\mu}_m(a,\cdot) - \hat{v}_m(a,\cdot)\|_1 \leq 1 + \sqrt{2M}$. Now we plug all things together, and proceed as

$$
\begin{aligned}
|\hat{V}(\pi) - \hat{V}_{\text{aux}}(\pi)| &\leq H^2 \cdot \left(\sum_{m=1}^{M} \hat{w}_m \left(\max_{a\in\mathcal{A}} \|\hat{v}_m(a,\cdot)\|_1 \right)^{H-1} \cdot \max_{a\in\mathcal{A}} \|\hat{\mu}_m(a,\cdot) - \hat{v}_m(a,\cdot)\|_1 \right) \\
&\leq eH^2 \cdot \left(\sum_{m:\hat{w}_m \geq \frac{\epsilon}{H^2 M (Z\sqrt{2M})^H}} \hat{w}_m \cdot \max_{a\in\mathcal{A}} \|\hat{\mu}_m(a,\cdot) - \hat{v}_m(a,\cdot)\|_1 \right) \\
&\quad + H^2 \cdot \left(\sum_{m:\hat{w}_m < \frac{\epsilon}{H^2 M (Z\sqrt{2M})^H}} \hat{w}_m \cdot (Z\sqrt{2M})^H \right) \\
&\leq 4eH^2 \sum_{m=1}^{M} \epsilon \sqrt{\hat{w}_m / M} / H^2 + \epsilon \leq O(\epsilon).
\end{aligned}
$$

Collecting all three terms, we have $|V(\pi) - \hat{V}(\pi)| \leq O(\epsilon)$. This concludes the proof of Theorem 3.8.

Appendix D LMAB with Gaussian Rewards

So far we have focused on rewards with finite support $Z = O(1)$. In this section we consider and LMAB setting with Gaussian rewards and generalize Algorithm 1. Indeed, some steps in the algorithm cannot be straightforwardly extended if $Z = \infty$. In this subsection, we consider a standard Gaussian reward distribution–a special case of continuous rewards–and generalize Algorithm 1 to this setting. We make the following assumption.

Assumption D.1 (Gaussian Rewards) *The reward distribution conditioning on an action $a \in \mathcal{A}$ in a context $m \in [M]$ is $\mathcal{N}(\mu_m(a), 1)$ for some $|\mu_m(a)| \leq 1$.*

Even though the rewards have infinite support, we show that the same conclusion holds as in finite-support case, *i.e.,* the sample-complexity is upper bounded by $O((MH/\epsilon)^{O(\min(H,M))} + \text{poly}(A, H, M))$. Algorithms for the Gaussian case differ significantly in identifiable $H \geq 2M - 1$ and unidentifiable regimes $H < 2M - 1$. When $H \geq 2M - 1$, there are only minor changes in the algorithm design for defining core actions and how tensors are constructed. We handle this case in Appendix D.1. A more interesting case is the parameter unidentifiable regime, where we follow an alternative approach and discretize the support of rewards by $O(\epsilon/H^2)$-level. This approach is described in Appendix D.2.

We can reach similar conclusions for the Gaussian rewards:

Theorem D.2 *Consider any LMAB with M contexts under Gaussian reward Assumption D.1. There exists an algorithm such that with probability at least $1 - \eta$, it returns an ϵ-optimal policy using a number of episodes at most*

$$
\begin{aligned}
&\text{poly}(H, M, A, 1/\epsilon, \log(A/\eta)) + \text{poly}(\log(M/\eta), H, M)^{2M-1} \cdot \epsilon^{-(4M-2)}, \quad \textit{if } H \geq 2M-1, \\
&\text{poly}(H, w_{\min}^{-1}, A, 1/\epsilon, \log(A/\eta)) + \text{poly}(\log(MH/(\eta\epsilon)), H, M)^{H} \cdot \epsilon^{-(2H+2)}, \quad \textit{otherwise}.
\end{aligned}
$$

Note that the dependency on ϵ is at most $\epsilon^{-(4M-2)}$ and smaller when $H < 2M - 2$. In the parameter identifiable regime, similarly to discrete reward cases, near-optimality of returned policy comes from the closeness of latent model parameters in the Wasserstein metric. In the parameter unidentifiable regime, we first discretize the support of rewards in $O(\epsilon)$ level. Then we can apply the same procedures for handling discrete rewards as in Section 3.3. We provide the full details in Appendix D.

D.1 Algorithm for Identifiable Regime $H \geq 2M - 1$

Let $\hat{M}_2 \in \mathbb{R}^{A \times A}$ be the empirical second-order moments as

$$\hat{M}_2 = \frac{1}{2N_0} \sum_{k=1}^{N_0} r_1^k r_2^k \cdot \boldsymbol{e}_{a_1^k} \boldsymbol{e}_{a_2^k}^\top.$$

Then let $\widehat{\mathbf{U}}$ be the subspace spanned by top-M eigenvectors of $\hat{M}_2$. A similar conclusion to Lemma 3.7 holds:

Lemma D.3 *Let $\widehat{\mathbf{U}}$ be a subspace spanned by top-M eigenvectors of $\hat{M}_2$. After we estimate $\hat{M}_2$ using $N_0 = O(A^4 \log(A/\eta)/\delta_{\text{sub}}^4)$ episodes, with probability at least $1 - \eta$, for all $m \in [M]$, there exists $\Delta_m : \|\Delta_m\|_\infty \leq \delta_{\text{sub}}/w_m^{1/2}$ such that $\mu_m + \Delta_m \in \widehat{\mathbf{U}}$.*

Proof of Lemma D.3 is identical to the proof of Lemma 3.7. Let $\delta_{\text{sub}} = \epsilon/(2MH^2)$.

Similarly to finite-support reward distributions, let $\{\hat{\beta}_j\}_{j=1}^M$ be the orthonormal basis of $\widehat{\mathbf{U}}$ and construct $\hat{\Phi} \in \mathbb{R}^{A \times M}$ such that the jth column of $\hat{\Phi}$ is $\hat{\beta}_j$, *i.e.,* $\hat{\Phi}_{:,j} = \hat{\beta}_j$. We invoke Theorem B.1 to get a set of core actions $\{a_j\}_{j=1}^n$. The main difference to the finite-support case is that we do not need to specify a corresponding event of rewards. Instead, we measure the correlation in terms of actual reward values. Specifically, for every multi-index $(i_1, i_2, ..., i_l) \in [n]^l$, using N_1 episodes where

$$N_1 = O\left(\log(ln^l/\eta)\right)^l / \delta_{\text{tsr}}^2.$$

We play $a_t^k = a_{i_t}$ for $t = 1, ..., l$ and $k \in [N_b]$, and estimate higher-order moments:

$$\hat{T}_l(i_1, ..., i_l) = \frac{1}{N_1} \sum_{k=1}^{N_1} \Pi_{t=1}^d r_t^k.$$

We can easily verify that

$$T_l = \mathbb{E}[\hat{T}_l] = \sum_{m=1}^M w_m \nu_m^{\bigotimes l}.$$

We can apply the concentration of higher-order polynomials of sub-Gaussian random variables element-wise, given from the following lemma on hypercontractivity inequality:

Lemma D.4 (Hypercontractivity Inequality (Theorem 1.9 in [34])) *Consider a degree-l polynomial f defined over a set of N independent samples of zero-mean unit-variance Gaussians, such that $f(X_{1:N}) := f(X_1, ..., X_N)$. Then,*

$$\mathbb{P}\left(|f(X_{1:N}) - \mathbb{E}[f(X_{1:N})]| \geq \lambda\right) \leq e^2 \exp\left(-\left(\frac{\lambda^2}{C \cdot var(f(X_{1:N}))}\right)^{1/l}\right),$$

for some absolute constant $C > 0$.

To show the concentration of $\hat{T}_l(i_1, ..., i_l)$ around $T_l(i_1, ..., i_l)$, we can apply Lemma D.4 with plugging $\lambda = O(\delta_{\text{tsr}})$ and $var(f) \leq 2^l \cdot var(X^l)/N_1$, where $X \sim \mathcal{N}(0, 1)$. Here, f can be viewed as a degree-l polynomial of $X_t^k := r_t^k - \mu_{m^k}(a_t^k)$ for $t = 1, \ldots, l$ and $k = 1, ..., N_1$, where m^k is a latent context for the episode k.

Since $var(X^l) \leq O(l^l)$, we need $N_1 = O(l^l \log^l\left(ln^l/\eta\right)/\delta_{\text{sub}}^2)$ to make the exponent less than $\eta/(ln^l)$. Take union bound over all elements in $\hat{T}_l$ ensures that $\|T_l - \hat{T}_l\|_\infty \leq \delta_{\text{tsr}}$ with probability at least $1 - \eta$.

Now we find a set of parameters $\{(\hat{w}_m, \hat{\nu}_m)\}_{m=1}^M$ with the only constraint:

$$\hat{w}_m \in \mathbb{R}_+, \ \sum_{m=1}^M \hat{w}_m = 1.$$

We aim to find

$$\left\| \sum_{m=1}^{M} \hat{w}_m \hat{\nu}_m^{\bigotimes l} - \hat{T}_l \right\|_\infty \leq \delta_{\text{tsr}}, \qquad \forall l \in [2M-1].$$

Now we can construct an empirical model by recovering $\hat{\mu}_m = \hat{T}\hat{\nu}_m$ where $\hat{T} \in \mathbb{R}^{A \times M}$ is defined similarly to (18) as

$$\hat{T}_{:,j} := \rho(a_j)\hat{\Phi}\hat{G}(\rho)^{-1}\hat{\Phi}_{a_j,:}, \qquad \forall j \in [n],$$

where ρ is a distribution over rows of $\hat{\Phi}$ found by Theorem B.1 and $\hat{G}(\rho)$ is defined as in (10). We do not need extra clipping and normalization steps here since any $\hat{\mu}_m$ is a valid model parameter. Now with $\{(\hat{w}_m, \hat{\mu}_m)\}$, we call the planning oracle 2.3 and obtain an ϵ-optimal policy.

D.2 Algorithm with Short Time-Horizon $H < 2M - 1$

We first discretize possible reward values: let $\mathcal{Z} = \{z_1, z_2, ..., z_Z\}$ where $z_i = -4\sqrt{\log(H/\epsilon)} + (i-1) \cdot \epsilon/H^2$ and $Z = \lfloor 8H^2\sqrt{\log(H/\epsilon)}/\epsilon \rfloor$. We define an auxiliary reward (pseudo) p.d.f $p_m(a, \cdot)$ for each a and m as the following: for all $s \in [Z-1]$,

$$p_m(a, r) = \frac{H^2}{\epsilon} \int_{z_s}^{z_{s+1}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - \mu_m(a))^2}{2}\right) dx, \qquad \forall r \in [z_s, z_{s+1}), \tag{23}$$

and $p_m(a, r) = 0$ for all $r \in (-\infty, z_1) \cup [z_Z, \infty)$. Define $\tilde{V}(\cdot)$ be the policy evaluation function in a (pseudo) LMAB model $\tilde{\mathcal{B}} = (\mathcal{A}, \{w_m\}_{m=1}^M, \{p_m\}_{m=1}^M)$:

$$\tilde{V}(\pi) := \sum_{m=1}^{M} w_m \cdot \sum_{a_{1:H}} \int_{r_{1:H}} \left(\sum_{t=1}^{H} r_t\right) \Pi_{t=1}^{H} p_m(a_t, r_t)\pi(a_{1:H}|r_{1:H-1}) d(r_{1:H}).$$

We first show that $p_m(a, \cdot)$ is good approximation of true reward distributions:

Lemma D.5 *Let $\mathcal{B}$ and $\tilde{\mathcal{B}}$ defined as above. Then for any history-dependent policy π, $|V(\pi) - \tilde{V}(\pi)| \leq 10\epsilon$.*

Given Lemma D.5, we will discretize reward values are so that we can leverage the result of Section 3.3 for short time-horizon. Specifically, let $\overline{\mathcal{B}}$ be a model $\{(w_m, q_m)\}_{m=1}^M$ with discrete reward distributions taking values in $\mathcal{Z} \cup \{0\}$, and let $q_m(a, z) = \frac{\epsilon}{H^2} \cdot p_m(a, z)$ for $z \in \mathcal{Z}$ and $q_m(a, 0) = \mathbb{P}_m(r \notin [z_1, z_Z]|a)$. As if the underlying model is $\overline{\mathcal{B}}$, we run Algorithm 1 with manually modifying the observed rewards $r_t \to \overline{r}_t$:

$$\begin{aligned} \overline{r}_t &= 0, && \text{if } r_t < z_1 \text{ or } r_t \geq z_Z, \\ \overline{r}_t &= z_s, && \text{for some } s \in [Z-1], \text{ s.t. } r_t \in [z_s, z_{s+1}). \end{aligned}$$

From actions $(a_1, ..., a_H)$ and reward observations $(\overline{r}_1, ..., \overline{r}_H)$, we can now apply Algorithm 1 for the parameter unidentifiable case $H < 2M - 1$.

D.3 Proof of Theorem D.2

Now we are ready to prove the Theorem D.2 for Gaussian rewards.

D.3.1 Identifiable Regime $H \geq 2M - 1$:

By Lemma 3.3, we know that $W(\gamma, \hat{\gamma}) \leq O\left(M^3 n \delta_{\text{tsr}}^{-1/(2M-1)}\right)$ where $\nu = \sum_{m=1}^M w_m \delta_{\nu_m}$ and $\hat{\nu} = \sum_{m=1}^M \hat{w}_m \delta_{\hat{\nu}_m}$. With experimental design, by Corollary 3.2, we can ensure that

$$\max_{a \in \mathcal{A}} |\mu_m(a) - \hat{\mu}_{m'}(a)| \leq \sqrt{2M} \|\nu_m - \hat{\nu}_{m'}\|_\infty.$$

Observe that for any $a \in \mathcal{A}$, total variation distance between standard Gaussians is bounded by the distance between centers of Gaussians, *i.e.,* $d_{TV}(\mathcal{N}(\mu_m(a), 1), \mathcal{N}(\hat{\mu}_{m'}(a), 1)) \leq |\mu_m(a) - \hat{\mu}_{m'}(a)|$. Using Proposition 3.4, we can show that

$$\begin{aligned}
|V(\pi) - \hat{V}(\pi)| &\leq 2H^2 \cdot \inf_{\Gamma} \sum_{(m,m')} \Gamma(m, m') \max_{a \in \mathcal{A}} d_{TV}(\mathcal{N}(\mu_m(a), 1), \mathcal{N}(\hat{\mu}_{m'}(a), 1)) \\
&\leq 2H^2 \cdot \inf_{\Gamma} \sum_{(m,m')} \Gamma(m, m') \max_{a \in \mathcal{A}} |\mu_m(a) - \hat{\mu}_{m'}(a)| \\
&\leq 2\sqrt{2M} H^2 \cdot \inf_{\Gamma} \sum_{(m,m')} \Gamma(m, m') \|\nu_m - \hat{\nu}_{m'}\|_\infty \\
&\leq 2\sqrt{2M} H^2 W(\gamma, \hat{\gamma}).
\end{aligned}$$

Plugging the choice of $\delta_{\text{tsr}} = O(\epsilon/(H^2 M^{3.5} n))^{2M-1}$, this is less than ϵ.

D.3.2 Unidentifiable Regime $H < 2M - 1$:

Let us first compare the expected rewards from $\tilde{\mathcal{B}}$ and $\overline{\mathcal{B}}$ with any fixed policy π.

$$\begin{aligned}
\tilde{V}(\pi) &= \sum_{m=1}^{M} w_m \cdot \sum_{a_{1:H}} \int_{r_{1:H}} \left(\sum_{t=1}^{H} r_t \right) \Pi_{t=1}^{H} p_m(a_t, r_t) \pi(a_{1:H} | r_{1:H-1}) d(r_{1:H}), \\
\overline{V}(\pi) &= \sum_{m=1}^{M} w_m \cdot \sum_{a_{1:H}} \sum_{\overline{r}_{1:H}} \left(\sum_{t=1}^{H} \overline{r}_t \right) \Pi_{t=1}^{H} q_m(a_t, \overline{r}_t) \pi(a_{1:H} | \overline{r}_{1:H-1}).
\end{aligned}$$

With slight abuse in notation, let $\overline{r}_t$ be a quantized value of r_t. Then we can show that

$$\begin{aligned}
\left| \tilde{V}(\pi) - \overline{V}(\pi) \right| &\leq \sum_{t=1}^{H} \int_{r_{1:H}: r_t \in [z_1, z_Z), \forall t \in [H]} \left| \sum_{m=1}^{M} w_m \sum_{a_{1:H}} (r_t - \overline{r}_t) \Pi_{t=1}^{H} p_m(a_t, r_t) \pi(a_{1:H} | r_{1:H-1}) d(r_{1:H}) \right| \\
&\quad + \sum_{t=1}^{H} \int_{r_{1:H}: r_t \notin [z_1, z_Z), \exists t \in [H]} \left| \sum_{m=1}^{M} w_m \sum_{a_{1:H}} \overline{r}_t \Pi_{t=1}^{H} q_m(a_t, r_t) \pi(a_{1:H} | r_{1:H-1}) d(r_{1:H}) \right| \\
&\leq \sum_{t=1}^{H} \int_{r_{1:H}: r_t \in [z_1, z_Z), \forall t \in [H]} \sum_{m=1}^{M} w_m \sum_{a_{1:H}} (\epsilon/H^2) \Pi_{t=1}^{H} p_m(a_t, r_t) \pi(a_{1:H} | r_{1:H-1}) d(r_{1:H}) \\
&\quad + |H z_Z| \cdot \mathbb{P}(\exists t \in [H],\ s.t.\ r_t \notin [z_1, z_Z)) \\
&\leq \epsilon/H + |H^2 z_Z| \cdot \mathbb{P}_{X \sim \mathcal{N}(0,1)}(X \geq z_Z - 1),
\end{aligned}$$

where in the first inequality, we used $p_m(a, r) = 0$ for $r \notin [z_1, z_Z)$. Note that $\mathbb{P}_{X \sim \mathcal{N}(0,1)}(|X| \geq z_Z - 1) \leq (H/\epsilon)^4$ with $z_Z = 4\sqrt{\log(H/\epsilon)}$.

Note that a system with manually discretized rewards can be described by the model $\overline{\mathcal{B}}$. Assuming the returned policy $\hat{\pi}$ from Algorithm 1 is $O(\epsilon)$-optimal for $\overline{\mathcal{B}}$, by triangle inequality for the policy evaluation for any policy π,

$$|\overline{V}(\pi) - V(\pi)| \leq |\overline{V}(\pi) - \tilde{V}(\pi)| + |\tilde{V}(\pi) - V(\pi)|,$$

we conclude that $\hat{\pi}$ is $O(\epsilon)$-optimal for $\mathcal{B}$ with Gaussian rewards.

D.4 Proof of Lemma D.5

We start by unfolding the expression for policy value differences.

$$|f(\pi) - \tilde{f}(\pi)| \leq \sum_{m=1}^{M} w_m \cdot \sum_{a_{1:H}} \int_{r_{1:H}} \left(\sum_{t=1}^{H} |r_t| \right) \left| \Pi_{t=1}^{H} p_m(a_t, r_t) - \Pi_{t=1}^{H} g_m(a_t, r_t) \right| \pi(a_{1:H} | r_{1:H-1}) d(r_{1:H}),$$

where $g_m(a_t, r_t) := \frac{1}{\sqrt{2\pi}} \exp\left(-(r_t - \mu_m(a_t))^2/2\right)$. We first rule out reward values greater than $z_Z = 4\sqrt{\log(H/\epsilon)}$. Define a set of bad reward sequences $\mathcal{E}_b = \{r_{1:H} | \exists t \in [H], \textit{s.t.}, |r_t| > z_Z\}$. Then for any $t_0 \in [H]$,

$$\begin{aligned}
&\sum_{a_{1:H}} \int_{r_{1:H} \in \mathcal{E}_b} |r_{t_0}| \left| \Pi_{t=1}^H p_m(a_t, r_t) - \Pi_{t=1}^H g_m(a_t, r_t) \right| \pi(a_{1:H} | r_{1:H-1}) d(r_{1:H}) \\
&\quad = \sum_{a_{1:H}} \int_{r_{1:H} \in \mathcal{E}_b} |r_{t_0}| \cdot \Pi_{t=1}^H g_m(a_t, r_t) \pi(a_{1:H} | r_{1:H-1}) d(r_{1:H}) \\
&\quad \leq \sum_{a_{1:H}} \int_{r_{1:H} \in \mathcal{E}_b \cap \{|r_{t_0}| \leq z_Z\}} |r_{t_0}| \cdot \Pi_{t=1}^H g_m(a_t, r_t) \pi(a_{1:H} | r_{1:H-1}) d(r_{1:H}) \\
&\quad + \sum_{a_{1:H}} \int_{r_{1:H} \in \mathcal{E}_b \cap \{|r_{t_0}| > z_Z\}} |r_{t_0}| \cdot \Pi_{t=1}^H g_m(a_t, r_t) \pi(a_{1:H} | r_{1:H-1}) d(r_{1:H}) \\
&\quad \leq z_Z \cdot \mathbb{P}_m(\mathcal{E}_b) + \sum_{a_{1:t_0}} \int_{r_{1:t_0} \in \{|r_{t_0}| > z_Z\}} |r_{t_0}| \cdot \Pi_{t=1}^{t_0} g_m(a_t, r_t) \pi(a_{1:t_0} | r_{1:t_0-1}) d(r_{1:t_0}),
\end{aligned}$$

where the last inequality results from integrating over probabilities for time steps $t_0 + 1, ..., H$. The last summation term can be further bounded by integrating out t_0^{th} time step since

$$\begin{aligned}
\int_{\{|r_{t_0}| > z_Z\}} |r_{t_0}| g_m(a_{t_0}, r_{t_0}) d(r_{t_0}) &= \int_{\{|x| > 4\sqrt{\log(H/\epsilon)}\}} \frac{|x|}{\sqrt{2\pi}} \exp\left(-\frac{(x - \mu_m(a_t))^2}{2}\right) dx \\
&\leq 2\mathbb{E}_{X \sim \mathcal{N}(0,1)} \left[|X| \cdot \mathbb{1}\{|X| > 3\sqrt{\log(H/\epsilon)}\} \right] \leq 2\epsilon^2/H^2,
\end{aligned}$$

where in the first inequality we used $|\mu_m(a_t)| \leq 1 \leq \sqrt{\log(H/\epsilon)}$. In last inequality we used $\mathbb{E}[|X| \cdot \mathbb{1}\{|X| \geq t\}] \leq \sqrt{\mathbb{P}(|X| \geq t)}$ by Cauchy-Schwartz inequality, and then used the Gaussian tail bound $\mathbb{P}(|X| \geq t) \leq \exp(-t^2/2)$ for $t = 3\sqrt{\log(H/\epsilon)}$. Therefore, we now get

$$\begin{aligned}
&\sum_{a_{1:H}} \int_{r_{1:H} \in \mathcal{E}_b} |r_{t_0}| \left| \Pi_{t=1}^H p_m(a_t, r_t) - \Pi_{t=1}^H g_m(a_t, r_t) \right| \pi(a_{1:H} | r_{1:H-1}) d(r_{1:H}) \\
&\quad \leq 4\sqrt{\log(H/\epsilon)} \cdot \epsilon^4/H^3 + 2\epsilon^2/H \leq 4\epsilon^2/H,
\end{aligned}$$

where we used $\mathbb{P}_m(\mathcal{E}_b) \leq H \cdot \mathbb{P}(|X| \geq z_Z - 1) \leq \epsilon^4/H^3$ with sufficiently small $\epsilon > 0$. This can be similarly done for all t_0, and thus

$$\sum_{a_{1:H}} \int_{r_{1:H} \in \mathcal{E}_b} \left(\sum_{t=1}^H |r_t| \right) \left| \Pi_{t=1}^H p_m(a_t, r_t) - \Pi_{t=1}^H g_m(a_t, r_t) \right| \pi(a_{1:H} | r_{1:H-1}) d(r_{1:H}) \leq 4\epsilon^2.$$

We remain to bound

$$\sum_{a_{1:H}} \int_{r_{1:H} \in \mathcal{E}_b^c} \left(\sum_{t=1}^H |r_t| \right) \left| \Pi_{t=1}^H p_m(a_t, r_t) - \Pi_{t=1}^H g_m(a_t, r_t) \right| \pi(a_{1:H} | r_{1:H-1}) d(r_{1:H}).$$

Note that $|r_t| < z_Z$ for all $t \in [H]$ when $r_{1:H} \in \mathcal{E}_b^c$. For each $t \in [H]$, we aim to bound

$$\sum_{a_{1:H}} \int_{r_{1:H} \in \mathcal{E}_b^c} |r_t| \left| \Pi_{t=1}^H p_m(a_t, r_t) - \Pi_{t=1}^H g_m(a_t, r_t) \right| \pi(a_{1:H} | r_{1:H-1}) d(r_{1:H}).$$

Next, for any $r_t \in [z_s, z_{s+1}]$ for $s \in [L-1]$, we observe that

$$\begin{aligned}
|p_m(a_t, r_t) - g_m(a_t, r_t)| &= \frac{1}{z_{s+1} - z_s} \int_{z_s}^{z_{s+1}} |g_m(a_t, x) - g_m(a_t, r_t)| \, dx \\
&\leq \frac{1}{z_{s+1} - z_s} \int_{z_s}^{z_{s+1}} |g_m(a_t, x)| + \frac{1}{z_{s+1} - z_s} \int_{z_s}^{z_{s+1}} |g_m(a_t, x) - g_m(a_t, r_t)| \, dx.
\end{aligned}$$

Then we observe that

$$|g_m(a_t, x) - g_m(a_t, r_t)| = \left| \frac{d}{dx} g_m(a_t, x')(r_t - x') \right|$$
$$\leq \frac{1}{\sqrt{2\pi}} |x' \exp(-x'^2/2)||z_{s+1} - z_s|,$$

for some $x' \in [z_s, z_{s+1}]$ where we used the mean-value theorem. A simple algebra shows that for any $x, x' \in [z_s, z_{s+1}]$,

$$|x' \exp(-x'^2/2) - x \exp(-x^2/2)| \leq |r_t \exp(-r_t^2/2)| + 2|z_{s+1} - z_s|,$$

where we used the second derivative of $x \exp(-x^2/2)$, which is $(x^2 - 1) \exp(-x^2/2)$, is always less than 1 in absolute value.

Plugging above relations into bounding the difference between p_m and g_m yields

$$|p_m(a_t, x) - g_m(a_t, r_t)| \leq \frac{1}{z_{s+1} - z_s} \left(\frac{\epsilon^8}{H^8} \int_{z_s}^{z_{s+1}} g_m(a_t, x) dx + \int_{z_s}^{z_{s+1}} |g_m(a_t, x) - g_m(a_t, r_t)| dx \right)$$
$$\leq \frac{\epsilon^8}{H^8} + \frac{1}{\sqrt{2\pi}} \int_{z_s}^{z_{s+1}} |r_t \exp(-r_t^2/2)| + 2|z_{s+1} - z_s| dx$$
$$\leq \frac{3\epsilon^2}{H^4} + \frac{2\epsilon}{H^2\sqrt{2\pi}} |r_t| \cdot \exp(-r_t^2/2).$$

Using this the above, we bound

$$\sum_{a_{1:H}} \int_{r_{1:H} \in \mathcal{E}_b^c} |r_t| \left| \Pi_{t=1}^H p_m(a_t, r_t) - \Pi_{t=1}^H g_m(a_t, r_t) \right| \pi(a_{1:H}|r_{1:H-1}) d(r_{1:H}). \tag{24}$$

If $t = H$, then

$$(24) \leq \sum_{a_{1:H}} \int_{r_{1:H-1}} \Pi_{t=1}^{H-1} g_m(a_t, r_t) \pi(a_{1:H}|r_{1:H-1}) d(r_{1:H-1}) \int_{r_H \in \mathcal{E}_b^c} |r_H||p_m(a_H, r_H) - g_m(a_H, r_H)| \cdot d(r_H)$$
$$+ \sum_{a_{1:H}} \int_{r_{1:H-1} \in \mathcal{E}_b^c} \left| \Pi_{t=1}^H p_m(a_t, r_t) - \Pi_{t=1}^H g_m(a_t, r_t) \right| \pi(a_{1:H}|r_{1:H-1}) d(r_{1:H-1}) \int_{r_H \in \mathcal{E}_b^c} |r_H| p_m(a_H, r_H) \cdot d(r_H).$$

For the first term,

$$\int_{r_H \in \mathcal{E}_b^c} |r_H||p_m(a_H, r_H) - g_m(a_H, r_H)| \cdot d(r_H) \leq z_Z \cdot (z_Z - z_1) \cdot \frac{3\epsilon^2}{H^4} + \frac{4\epsilon}{H^2} \leq \frac{8\epsilon}{H^2},$$

where we used $\int_{r_H} \frac{1}{\sqrt{2\pi}} r_H^2 \exp(-r_H^2/2) \leq 1$ and $32(\epsilon/H) \cdot \log(H/\epsilon) < 1$ for sufficiently small ϵ. Note that we also have

$$\sum_{a_{1:H}} \int_{r_{1:H-1}} \Pi_{t=1}^{H-1} g_m(a_t, r_t) \pi(a_{1:H}|r_{1:H-1}) d(r_{1:H-1}) \leq 1.$$

For the second term, we first have

$$\int_{r_H \in \mathcal{E}_b^c} |r_H| p_m(a_H, r_H) \cdot d(r_H) \leq 1 + \mu_m(a_H) \leq 2.$$

Furthermore, we can show that

$$\sum_{a_{1:H}} \int_{r_{1:H-1} \in \mathcal{E}_b^c} \left| \Pi_{t=1}^H p_m(a_t, r_t) - \Pi_{t=1}^H g_m(a_t, r_t) \right| \pi(a_{1:H}|r_{1:H-1}) d(r_{1:H-1})$$
$$= \sum_{a_{1:H-1}} \int_{r_{1:H-1} \in \mathcal{E}_b^c} \left| \Pi_{t=1}^H p_m(a_t, r_t) - \Pi_{t=1}^H g_m(a_t, r_t) \right| \pi(a_{1:H-1}|r_{1:H-2}) \sum_{a_H} \pi(a_H|a_{1:H-1}, r_{1:H-1}) d(r_{1:H-1})$$
$$= \sum_{a_{1:H-1}} \int_{r_{1:H-1} \in \mathcal{E}_b^c} \left| \Pi_{t=1}^H p_m(a_t, r_t) - \Pi_{t=1}^H g_m(a_t, r_t) \right| \pi(a_{1:H-1}|r_{1:H-2}) d(r_{1:H-1}),$$

from which we recursively apply similar arguments. Thus we can conclude that

$$\sum_{a_{1:H}} \int_{r_{1:H} \in \mathcal{E}_b^c} |r_H| \left| \Pi_{t=1}^H p_m(a_t, r_t) - \Pi_{t=1}^H g_m(a_t, r_t) \right| \pi(a_{1:H} | r_{1:H-1}) d(r_{1:H}) \leq O(\epsilon/H).$$

$|r_t|$ with other time steps can also be similarly bounded. Thus, we can conclude that

$$\sum_{a_{1:H}} \int_{r_{1:H}} \left(\sum_{t=1}^H |r_t| \right) \left| \Pi_{t=1}^H p_m(a_t, r_t) - \Pi_{t=1}^H g_m(a_t, r_t) \right| \pi(a_{1:H} | r_{1:H-1}) d(r_{1:H}) \leq O(\epsilon),$$

and therefore $|V(\pi) - \tilde{V}(\pi)| \leq O(\epsilon)$ since $\sum_{m=1}^M w_m = 1$.

Appendix E Deferred Details in Section 4

E.1 Additional Definitions

Let us define a few notation and interaction protocol. Suppose at the beginning of episode, a latent context $m_0 \in [M]$ is chosen, and and at each time step $t \in [H]$, we play a_{i_t} where i_t is sampled from Unif($[n]$). Let $\boldsymbol{i} = (i_1, i_2, ..., i_H)$ be the sequence of indices of played core actions, and $\boldsymbol{b} = (b_1, b_2, ...b_H)$ be the event-observation sequence in the episode, where $b_t := \mathbb{1}\{r_t = Z_{i_t}\}$. Let the parameter space Θ be the set of valid parameters:

$$\Theta = \{\theta = \{(w_m, \nu_m)\}_{m=1}^M | \forall j \in [n], m \in [M] \text{ s.t. } w_m, \nu_m(i) \in \mathbb{R}_+, \sum_{m=1}^M w_m = 1, \nu_m(j) \leq 1\}. \tag{25}$$

We use superscript k to denote quantities observed in the kth episode. The probability of a trajectory under a model $\theta \in \Theta$ in the kth episode is defined by

$$\mathbb{P}_\theta(\boldsymbol{b}^k, \boldsymbol{i}^k) := (1/n)^H \cdot \sum_{m=1}^M w_m \Pi_{t=1}^H (b_t^k \nu_m(i_t) + (1 - b_t^k)(1 - \nu_m(i_t))).$$

E.2 Polynomial Upper Bounds with Separation

In this subsection, we specify the details on separation conditions that make the polynomial sample complexity possible with MLE solutions. Suppose that there exists a context revealing action for any $m \neq m' \in [M]$, *i.e.,* we are given the following assumption:

Assumption E.1 (Separated Bandit Instances) *For any $m \neq m' \in [M]$, there exists some (unknown) $a \in \mathcal{A}$ such that $\|\mu_m^*(a, \cdot) - \mu_{m'}^*(a, \cdot)\|_1 \geq \gamma$ for some known $\gamma > 0$.*

Under Assumption E.1, if the time horizon $H = \tilde{O}(Z^2 M^2/\gamma^2)$ is given enough to identify the context within each episode, then we can significantly improve the sample complexity for learning LMAB, from exponential to polynomial. Note that the time-horizon H can be still much smaller than A and thus we cannot explore all actions within a single episode, which is in contrast to explicit clustering based approaches studied in [9, 16].

Maximum likelihood estimator for LMABs with separation can guarantee the following:

Lemma E.2 *Consider the maximum likelihood estimator $\theta_N = \{(\hat{w}_m, \hat{\nu}_m)\}_{m=1}^M$ under Assumption E.1 with time-horizon $H \geq C_1 \cdot nMZ^2 \log(1/(\epsilon w_{\min}))/\gamma^2$ for some universal constant $C_1 > 0$. If $N = C_2 \cdot w_{\min}^{-2} n \cdot \log(N/\eta)/\epsilon^2$ for some large constant $C_2 > 0$, then with probability at least $1 - \eta$, we have (up to some permutations in θ_N)*

$$|w_m^* - \hat{w}_m| \leq \epsilon w_{\min}, \ \|\nu_m^* - \hat{\nu}_m\|_\infty \leq 2\epsilon, \qquad \forall m \in [M].$$

To get the above result, we first observe a consequence due to experimental design: the converse of Corollary 3.2 implies that if Assumption E.1 holds for all $m \neq m'$, then we have $\|\nu_m^* - \nu_{m'}^*\|_\infty \geq$

$\gamma/(Z\sqrt{2M})$. Thus, if $H/n = \tilde{O}(MZ^2/\gamma^2)$, then we can play each core action $O(H/n)$-times and get an $O(\gamma/(Z\sqrt{M}))$-accurate estimator $\hat{\nu}_m$ for one of $\{\nu^*_{m'}\}_{m'=1}^M$. Since we have good separation between samples from different contexts, by proper clustering arguments, the sample complexity of recovering θ^* can be polynomial. With Lemma E.2, we can use equation (4) to connect the near-optimality of returned policy computed with $\{(\hat{w}_m, \hat{\mu}_m)\}_{m=1}^M$ and the closeness in Wasserstein metric. We mention that for H, the dependence on Z can be removed with more computationally expensive experimental design (see also Remark C.1).

E.3 Proof of Lemma 4.1

We connect the maximum likelihood estimator to total variation distance between observations from θ^* and θ_N. The connection between MLE θ_N and closeness in distributions of observations $\boldsymbol{b}$ can be established by the following lemma.

Lemma E.3 *There exists a universal constant $C > 0$ such that with probability at least $1-\eta$,*

$$\sum_{\boldsymbol{b}\in\{0,1\}^H} \sum_{\boldsymbol{i}\in[n]^H} |\mathbb{P}_{\theta_N}(\boldsymbol{b},\boldsymbol{i}) - \mathbb{P}_{\theta^*}(\boldsymbol{b},\boldsymbol{i})| \le C\sqrt{\frac{n\log(nHN)+\log(1/\eta)}{N}}.$$

That is, total variation distance between two observation distributions is bounded by $\tilde{O}\left(\sqrt{n/N}\right)$. On the other hand, for any $l \in [\min(H, 2M-1)]$ and any multi-index $(i_1, i_2, ..., i_l) \in [n]^l$, we have

$$\begin{aligned}
C\sqrt{\frac{n\log(nHN)+\log(1/\eta)}{N}} &\ge \sum_{\boldsymbol{b},\boldsymbol{i}} |\mathbb{P}_{\theta_N}(\boldsymbol{b},\boldsymbol{i}) - \mathbb{P}_{\theta^*}(\boldsymbol{b},\boldsymbol{i})| \\
&\ge \sum_{\boldsymbol{b}} |\mathbb{P}_{\theta_N} - \mathbb{P}_{\theta^*}|\left(b_{1:l}|a_{1:l} = (a_{i_1}, a_{i_2}, ..., a_{i_l})\right)\cdot \mathbb{P}(a_{1:d} = (a_{i_1}, a_{i_2}, ..., a_{i_l})) \\
&= \sum_{b_{1:l}} |\mathbb{P}_{\theta_N} - \mathbb{P}_{\theta^*}|\left(b_{1:l}|a_{1:l} = (a_{i_1}, a_{i_2}, ..., a_{i_l})\right)\cdot n^{-l} \ge n^{-l}\cdot \|\hat{T}_l - T_l\|_\infty.
\end{aligned}$$

This implies that

$$\|\hat{T}_l - T_l\|_\infty \le Cn^l \cdot \sqrt{\frac{n\log(HNn)+\log(1/\eta)}{N}}.$$

Applying this to all $l \in [\min(H, 2M-1)]$, we get Lemma 4.1.

E.4 Proof of Lemma E.2

From Lemma E.3, without loss of generality, we assume that the total variation distance between observations from θ_N and θ^* is bounded by $\epsilon w_{\min}/2$ since $N = \tilde{O}(w_{\min}^{-2} n/\epsilon^2)$:

$$\sum_{\boldsymbol{b},\boldsymbol{i}} |\mathbb{P}_{\theta_N}(\boldsymbol{b},\boldsymbol{i}) - \mathbb{P}_{\theta^*}(\boldsymbol{b},\boldsymbol{i})| \le \epsilon w_{\min}/2.$$

We will verify that for every $m \in [M]$, there exists $m' \in [M]$ such that $\|\hat{\nu}_m - \nu^*_{m'}\|_\infty \le 2\epsilon$ and $|\hat{w}_m - w^*_{m'}| \le \epsilon w_{\min}$.

First note that for all $m \ne m' \in [M]$, we have $\|\nu^*_m - \nu^*_{m'}\|_\infty \ge \lambda := \gamma/(Z\sqrt{2M})$. If it is not, then the model does not satisfy Assumption E.1. We start with the following lemma:

Lemma E.4 *Suppose that there exists $m \in [M]$ such that $\|\nu^*_m - \hat{\nu}_{m'}\|_\infty \ge \lambda/4$ for all $m' \in [M]$. For every $j \in [n]$, define $E_{m,j}$ an event defined as:*

$$E_{m,j} := \left\{ \left| \frac{\sum_{t=1}^H \mathbb{1}\{i_t = j\} b_t}{\sum_{t=1}^H \mathbb{1}\{i_t = j\}} - \nu^*_m(j) \right| < \lambda/8 \right\},$$

and let $E_m = \cap_{j=1}^n E_{m,j}$. Then

$$\mathbb{P}_{\theta^*}(E_m) \ge w_m/2,\ \mathbb{P}_{\theta_N}(E_m) \le \epsilon w_{\min}. \tag{26}$$

Proof. Let us first check that $\mathbb{P}_{\theta^*}(E) \geq w_m/2$. Let $n_j = \sum_{t=1}^{H} \mathbb{1}\{i_t = j\}$. Since,

$$\mathbb{P}_{\theta^*}(E) \geq w_m \cdot \mathbb{P}_{\theta^*}(E|m_0 = m) = w_m \cdot \cap_{j=1}^{n} \mathbb{P}_{\theta^*}(E_j|m_0 = m),$$

it suffices to show that

$$\mathbb{P}_{\theta^*}(E_j|m_0 = m) \geq \mathbb{P}_{\theta^*}\left(\left|\sum_{t=1}^{H} \mathbb{1}\{i_t = j\} b_t - n_j \nu_m^*(j)\right| \leq n_j \lambda/8 \,\middle|\, m_0 = m, n_j \geq H/(2n)\right) \mathbb{P}_{\theta^*}\left(n_j \geq H/(2n)\right)$$
$$\geq \left(1 - \exp(-\lambda^2 \cdot H/(64n))\right)\left(1 - \exp\left(\frac{-(1/2)(H/2n)^2}{H(1/n)(1-1/n) + (1/3)(H/2n)}\right)\right).$$

where in the last inequality we applied Hoeffeding's concentration inequality for first term, and then used Bernstein's inequality. Plugging $H > C \cdot nMZ^2 \log(1/(\epsilon w_{\min}))/\gamma^2$ for some sufficiently large constant $C > 0$, we get

$$\mathbb{P}_{\theta^*}(E_j|m_0 = m) \geq \left(1 - \exp(-2(\gamma^2/(128Z^2M)) \cdot H/(2n))\right) \cdot \left(1 - \exp\left(-H/(16n)\right)\right) \geq 1 - (\epsilon w_{\min})^2.$$

By union bound, we have $\mathbb{P}_{\theta^*}(E) \geq w_m \cdot (1 - n\epsilon^2 w_{\min}^2) \geq w_m/2$.

Now we check that $\mathbb{P}_{\theta_N}(E) \leq \epsilon w_{\min}$. Starting from

$$\mathbb{P}_{\theta_N}(E) = \sum_{m'=1}^{M} w_{m'} \mathbb{P}_{\theta_N}(E|m_0 = m'),$$

it suffices to show that $\mathbb{P}_{\theta_N}(E|m_0 = m') \leq \epsilon w_{\min}$ for all $m' \in [M]$. Let us fix m' and define

$$E_{m',j} := \left\{\left|\frac{\sum_{t=1}^{H} \mathbb{1}\{i_t = j\} b_t}{\sum_{t=1}^{H} \mathbb{1}\{i_t = j\}} - \hat{\nu}_m(j)\right| < \lambda/8\right\},$$

and let $E_{m'} = \cap_{j=1}^{n} E_{m',j}$. Following the same argument for $\mathbb{P}_{\theta^*}(E_j|m_0 = m)$, we can show that

$$\mathbb{P}_{\theta_N}(E_{m'}|m_0 = m') \geq 1 - (\epsilon w_{\min})^2.$$

If this happens, then it implies E^c since $\|\nu_m^* - \hat{\nu}_{m'}\|_\infty \geq \lambda/2$. Thus,

$$\sum_{m'=1}^{M} \hat{w}_{m'} \mathbb{P}_{\theta_N}(E_{m'}|m_0 = m') \leq \sum_{m'=1}^{M} \hat{w}_{m'} \mathbb{P}_{\theta_N}(E^c|m_0 = m') = \mathbb{P}_{\theta_N}(E^c).$$

In other words, we have

$$\mathbb{P}_{\theta_N}(E) \leq 1 - \sum_{m'=1}^{M} \hat{w}_{m'} \mathbb{P}_{\theta_N}(E_{m'}|m_0 = m') \leq \epsilon^2 w_{\min}^2.$$

□

The conclusion of Lemma E.4 contradict that $|\mathbb{P}_{\theta^*}(E_m) - \mathbb{P}_{\theta_N}(E_m)| \leq \epsilon w_{\min}/2$ due to the total variation distance bound from Lemma E.3. Thus, we ensure that for all $m \in [M]$, there exists $m' \in [M]$ such that $\|\nu_m^* - \hat{\nu}_{m'}\|_\infty \leq \lambda/4$.

Now without loss of generality, we can ignore the permutation invariance of models and assume that

$$\|\nu_m^* - \hat{\nu}_m\|_\infty \leq \lambda/4, \qquad \forall m \in [M].$$

Then we now show that for all $m \in [M]$, it holds that

$$|w_m^* - \hat{w}_m^*| \leq \epsilon w_{\min}, \quad \|\nu_m^* - \hat{\nu}_m^*\|_\infty \leq \epsilon.$$

For every m, let us define E_m similarly to Lemma E.4:

$$E_{m,j} := \left\{\left|\frac{\sum_{t=1}^{H} \mathbb{1}\{i_t = j\} b_t}{\sum_{t=1}^{H} \mathbb{1}\{i_t = j\}} - \nu_m^*(j)\right| < \lambda/2\right\},$$

and $E_m := \cap_{j=1}^{n} E_{m,j}$. Then we proceed as the following:

$$\sum_{\boldsymbol{b},\boldsymbol{i}} |\mathbb{P}_{\theta_N}(\boldsymbol{b},\boldsymbol{i}) - \mathbb{P}_{\theta^*}(\boldsymbol{b},\boldsymbol{i})| \geq \sum_{(\boldsymbol{i},\boldsymbol{b}) \in E_m} |\mathbb{P}_{\theta_N}(\boldsymbol{b},\boldsymbol{i}) - \mathbb{P}_{\theta^*}(\boldsymbol{b},\boldsymbol{i})|$$

$$
\begin{aligned}
&= \sum_{(\boldsymbol{i},\boldsymbol{b})\in E_m} \left| \sum_{m'=1}^{M} \hat{w}_{m'} \mathbb{P}_{\theta_N}(\boldsymbol{b},\boldsymbol{i}|m_0=m') - \sum_{m'=1}^{M} w_{m'} \mathbb{P}_{\theta^*}(\boldsymbol{b},\boldsymbol{i}|m_0=m') \right| \\
&\geq \sum_{\boldsymbol{i}} \sum_{\boldsymbol{b}} |\hat{w}_m \mathbb{P}_{\theta_N}(\boldsymbol{b},\boldsymbol{i}|m_0=m) - w_m \mathbb{P}_{\theta^*}(\boldsymbol{b},\boldsymbol{i}|m_0=m)| \\
&\quad - (\hat{w}_m \mathbb{P}_{\theta_N}(E_m^c|m_0=m) + w_m \mathbb{P}_{\theta^*}(E_m^c|m_0=m)) \\
&\quad - \sum_{m'\neq m} (\hat{w}_{m'} \mathbb{P}_{\theta_N}(E_m|m_0=m') + w_{m'} \mathbb{P}_{\theta^*}(E_m|m_0=m')) .
\end{aligned}
$$

Recall that for any $m' \neq m$, the event E_m implies $E_{m'}^c$ since every ν_m^* and $\nu_{m'}^*$ are separated by at least $\lambda = \gamma/(Z\sqrt{2M})$, which implies that ν_m^* and $\hat{\nu}_{m'}$ are separated by $(3/4)\lambda$. Following the same argument as in the proof for Lemma E.4,

$$
\begin{aligned}
\mathbb{P}_{\theta_N}(E_m|m_0=m') &\leq \mathbb{P}_{\theta_N}(E_{m'}^c|m_0=m') \leq \epsilon^2 w_{\min}^2, \\
\mathbb{P}_{\theta^*}(E_m|m_0=m') &\leq \mathbb{P}_{\theta^*}(E_{m'}^c|m_0=m') \leq \epsilon^2 w_{\min}^2.
\end{aligned}
$$

Similarly, we can also check that

$$
\begin{aligned}
\mathbb{P}_{\theta_N}(E_m^c|m_0=m) &\leq \epsilon^2 w_{\min}^2, \\
\mathbb{P}_{\theta^*}(E_m^c|m_0=m) &\leq \epsilon^2 w_{\min}^2.
\end{aligned}
$$

Thus, we have

$$
\sum_{\boldsymbol{i},\boldsymbol{b}} |\mathbb{P}_{\theta_N}(\boldsymbol{b},\boldsymbol{i}) - \mathbb{P}_{\theta^*}(\boldsymbol{b},\boldsymbol{i})| \geq \sum_{\boldsymbol{b},\boldsymbol{i}} |\hat{w}_m \mathbb{P}_{\theta_N}(\boldsymbol{b},\boldsymbol{i}|m_0=m) - w_m \mathbb{P}_{\theta^*}(\boldsymbol{b},\boldsymbol{i}|m_0=m)| - 2\epsilon^2 w_{\min}^2.
$$

We remain to lower bound $\sum_{\boldsymbol{b},\boldsymbol{i}} |\hat{w}_m \mathbb{P}_{\theta_N}(\boldsymbol{b},\boldsymbol{i}|m_0=m) - w_m^* \mathbb{P}_{\theta^*}(\boldsymbol{b},\boldsymbol{i}|m_0=m)|$. First note that

$$
\epsilon w_{\min}/2 \geq \sum_{\boldsymbol{b},\boldsymbol{i}} |\hat{w}_m \mathbb{P}_{\theta_N}(\boldsymbol{b},\boldsymbol{i}|m_0=m) - w_m^* \mathbb{P}_{\theta^*}(\boldsymbol{b},\boldsymbol{i}|m_0=m)| \geq |\hat{w}_m - w_m^*|.
$$

Thus we have $|\hat{w}_m - w_m^*| \leq \epsilon w_{\min}$. Now given this, we can proceed as

$$
\begin{aligned}
\epsilon w_{\min}/2 &\geq \sum_{\boldsymbol{b},\boldsymbol{i}} |\hat{w}_m \mathbb{P}_{\theta_N}(\boldsymbol{b},\boldsymbol{i}|m_0=m) - w_m^* \mathbb{P}_{\theta^*}(\boldsymbol{b},\boldsymbol{i}|m_0=m)| \\
&\geq w_m^* \cdot \sum_{\boldsymbol{b},\boldsymbol{i}} |\mathbb{P}_{\theta_N}(\boldsymbol{b},\boldsymbol{i}|m_0=m) - \mathbb{P}_{\theta^*}(\boldsymbol{b},\boldsymbol{i}|m_0=m)| - |\hat{w}_m - w_m^*|.
\end{aligned}
$$

Now let $i_m := arg\max_{j\in[n]} |\nu_m^*(j) - \hat{\nu}_m(j)|$. Define an event b_t being 1 at the first time a_{i_m} is played:

$$
F := \{b_t = 1, t = arg \min_{t'\in[H]} i_{t'} = i_m\}.
$$

Note that since $H \gg n\log(1/(\epsilon w_{\min}))$, i_m is played at least once with probability at least $1 - (\epsilon w_{\min})^2$. Then we can lower bound the total variation distance as

$$
\begin{aligned}
\sum_{\boldsymbol{b},\boldsymbol{i}} |\mathbb{P}_{\theta_N}(\boldsymbol{b},\boldsymbol{i}|m_0=m) - \mathbb{P}_{\theta^*}(\boldsymbol{b},\boldsymbol{i}|m_0=m)| &\geq |\mathbb{P}_{\theta_N}(F|m_0=m) - \mathbb{P}_{\theta^*}(F|m_0=m)| \\
&\geq |\nu_m^*(i_m) - \hat{\nu}_m(i_m)| - 2(\epsilon w_{\min})^2.
\end{aligned}
$$

Hence we can conclude that

$$
\epsilon w_{\min}/2 \geq w_m |\nu_m^*(i_m) - \hat{\nu}_m(i_m)| - 2(\epsilon w_{\min})^2 - \epsilon w_{\min},
$$

which implies $\|\nu_m^* - \hat{\nu}_m\|_\infty \leq 2\epsilon$.

Note that the Wasserstein distance can thus be bounded as

$$
W(\gamma^*, \hat{\gamma}) \leq \sum_{m=1}^{M} |w_m^* - \hat{w}_m| + \sum_{m=1}^{M} w_m^* \|\nu_m^* - \hat{\nu}_m\|_\infty \leq \epsilon M w_{\min} + 2\epsilon \leq 3\epsilon.
$$

Remark E.5 *In our guarantee, we assumed that $H \gg \log(1/\epsilon)$, i.e., H should be increased logarithmically with the final accuracy. As one might imagine, this is not the optimal condition for separations between individual models. A more delicate and technically involved analysis might reveal that the sub-optimal dependency on $\log(1/\epsilon)$ can be dropped with the EM algorithm as in learning Gaussian mixture models (e.g., [22]). Since it is technically much more complicated, we leave the task of verifying more tight separation conditions as future work.*

E.5 Proof of Lemma E.3

This is a rather standard consequence of MLE for parameterized distributions. Let us define well-conditioned parameters $\Theta' \subseteq \Theta$ defined as follows:

$$\Theta' = \Big\{\theta = \{(w_m, \nu_m)\}_{m=1}^M | \forall j \in [n], m \in [M] \text{ s.t. } w_m,\ \nu_m(j) \in \mathbb{R}_+, \\ \sum_{m=1}^M w_m = 1, \\ \epsilon \le w_m,\ \epsilon \le \nu_m(j) \le 1-\epsilon, \quad \forall m \in [M], j \in [n]\Big\}.$$

Let Θ_ϵ be ϵ^2-covering of Θ'. Since there are $n+M$ free parameters, log-cardinality of Θ_ϵ is at most $O(1/\epsilon)^{2(n+M)}$. Note that Θ_ϵ is a (not necessarily minimal) ϵ-cover for Θ as well.

To simplify the notation, we often use $X := (\boldsymbol{b}, \boldsymbol{i})$ (and $X^k = (\boldsymbol{b}^k, \boldsymbol{i}^k)$) to replace a sample trajectory. Our goal is to bound

$$TV(\theta_N, \theta^*) := \sum_{X=(\boldsymbol{b},\boldsymbol{i}):\boldsymbol{b}\in\{0,1\}^H, \boldsymbol{i}\in[n]^H} |\mathbb{P}_{\theta_N}(X) - \mathbb{P}_{\theta^*}(X)|,$$

where $TV(\theta_1, \theta_2)$ is a total variation distance between $\mathbb{P}_{\theta_1}$ and $\mathbb{P}_{\theta_2}$ for any $\theta_1, \theta_2 \in \Theta$.

Let $\overline{\theta} = \{(\overline{w}_m, \overline{\nu}_m\}_{m=1}^M \in \Theta_\epsilon$ such that

$$\begin{aligned} \frac{w_m^*}{\overline{w}_m} &\le 1+2\epsilon, && \forall m \in [M], \\ \frac{\nu_m^*(i)}{\overline{\nu}_m(i)} &\le 1+2\epsilon, && \forall m \in [M], i \in [n], \\ \frac{1-\nu_m^*(i)}{1-\overline{\nu}_m(i)} &\le 1+2\epsilon, && \forall m \in [M], i \in [n]. \end{aligned}$$

Such $\overline{\theta}$ is guaranteed to exist in Θ_ϵ by construction. A simple algebra shows that for any trajectory X, we have

$$\frac{\mathbb{P}_{\theta^*}(X)}{\mathbb{P}_{\overline{\theta}}(X)} \le (1+2\epsilon)^H.$$

As long as $\epsilon < 1/H^2$, this is bounded by constant.

Now, fix any $\theta \in \Theta_\epsilon$ and let $l(X) := \frac{1}{2}\log\left(\frac{\mathbb{P}_\theta(X)}{\mathbb{P}_{\overline{\theta}}(X)}\right)$. Then using Chernoff's method, we get

$$\begin{aligned} &\mathbb{P}_{\theta^*}\left(\sum_{k=1}^N l(X^k) - \log\left(\mathbb{E}_{\theta^*}\left[\exp\left(\sum_{k=1}^N l(X^k)\right)\right]\right) > \lambda\right) \\ &= \mathbb{P}_{\theta^*}\left(\exp\left(\sum_{k=1}^N l(X^k) - \log\left(\mathbb{E}_{\theta^*}\left[\sum_{k=1}^N \exp(l(X^k))\right]\right)\right) > \exp(\lambda)\right) \\ &\le \mathbb{E}_{\theta^*}\left[\exp\left(\sum_{k=1}^N l(X^k) - \log\left(\mathbb{E}_{\theta^*}\left[\exp\left(\sum_{k=1}^N l(X^k)\right)\right]\right)\right)\right] \cdot \exp(-\lambda) \\ &= \exp(-\lambda), \end{aligned}$$

where we used $\mathbb{E}[\exp(S - \log(\mathbb{E}[\exp(S)]))] = \frac{\mathbb{E}[\exp(S)]}{\mathbb{E}[\exp(S)]} = 1$ for any S and distribution. Taking union bound over Θ_ϵ, we can conclude that with probability at least $1-\eta$,

$$-\log\left(\mathbb{E}_{\theta^*}\left[\Pi_{k=1}^N \sqrt{\frac{\mathbb{P}_\theta(X^k)}{\mathbb{P}_{\overline{\theta}}(X^k)}}\right]\right) \le -\frac{1}{2}\sum_{k=1}^N \log\left(\frac{\mathbb{P}_\theta(X^k)}{\mathbb{P}_{\overline{\theta}}(X^k)}\right) + \log(|\Theta_\epsilon|/\eta).$$

Now, since $X^1, ..., X^N$ are independent, we have

$$
\begin{aligned}
-\log\left(\mathbb{E}_{\theta^*}\left[\Pi_{k=1}^N \sqrt{\frac{\mathbb{P}_\theta(X^k)}{\mathbb{P}_{\overline{\theta}}(X^k)}}\right]\right) &= -\sum_{k=1}^N \log\left(\mathbb{E}_{\theta^*}\left[\sqrt{\frac{\mathbb{P}_\theta(X)}{\mathbb{P}_{\overline{\theta}}(X)}}\right]\right) \\
&= -\sum_{k=1}^N \log\left(\mathbb{E}_{\theta^*}\left[\sqrt{\frac{\mathbb{P}_\theta(X)}{\mathbb{P}_{\theta^*}(X)}} \cdot \sqrt{\frac{\mathbb{P}_{\theta^*}(X)}{\mathbb{P}_{\overline{\theta}}(X)}}\right]\right) \\
&\geq -\sum_{k=1}^N \log\left((1+2\epsilon)^{H/2} \cdot \mathbb{E}_{\theta^*}\left[\sqrt{\frac{\mathbb{P}_\theta(X)}{\mathbb{P}_{\theta^*}(X)}}\right]\right) \\
&= -\frac{HN}{2}\log(1+2\epsilon) - \sum_{k=1}^N \log\left(\mathbb{E}_{\theta^*}\left[\sqrt{\frac{\mathbb{P}_\theta(X)}{\mathbb{P}_{\theta^*}(X)}}\right]\right).
\end{aligned}
$$

Using $-\ln(x) \geq 1 - x$, we get

$$
\begin{aligned}
-\sum_{k=1}^N \log\left(\mathbb{E}_{\theta^*}\left[\sqrt{\frac{\mathbb{P}_\theta(X)}{\mathbb{P}_{\theta^*}(X)}}\right]\right) &\geq \sum_{k=1}^N \left(1 - \mathbb{E}_{\theta^*}\left[\sqrt{\frac{\mathbb{P}_\theta(X)}{\mathbb{P}_{\theta^*}(X)}}\right]\right) \\
&= \sum_{k=1}^N \left(1 - \sum_X \sqrt{\mathbb{P}_{\theta^*}(X)\mathbb{P}_\theta(X)}\right) = N\mathcal{H}^2(\theta^*, \theta),
\end{aligned}
$$

where $\mathcal{H}(\theta_1, \theta_2)$ is a Hellinger distance between $\mathbb{P}_{\theta_1}$ and $\mathbb{P}_{\theta_2}$. Also note that $TV(\theta^*, \theta) \leq \mathcal{H}(\theta^*, \theta)$. Collecting all, we now can say that

$$
N \cdot TV^2(\theta^*, \theta) \leq -\frac{1}{2}\sum_{k=1}^N \log\left(\frac{\mathbb{P}_\theta(X^k)}{\mathbb{P}_{\overline{\theta}}(X^k)}\right) + \log(\Theta_\epsilon/\eta) + \frac{HN}{2}\log(1+2\epsilon),
$$

with probability at least $1 - \eta$. Finally, let $\overline{\theta}_N \in \Theta_\epsilon$ be the one close to θ_N such that

$$
\frac{\mathbb{P}_{\theta_N}(X)}{\mathbb{P}_{\overline{\theta}_N}(X)} \leq (1+2\epsilon)^H,
$$

similarly to when defining $\overline{\theta}$. Then,

$$
\begin{aligned}
N \cdot TV^2(\theta^*, \overline{\theta}_N) &\leq -\frac{1}{2}\sum_{k=1}^N \log\left(\frac{\mathbb{P}_{\theta_N}(X^k)}{\mathbb{P}_{\overline{\theta}}(X^k)}\right) + \frac{1}{2}\sum_{k=1}^N \log\left(\frac{\mathbb{P}_{\theta_N}(X^k)}{\mathbb{P}_{\overline{\theta}_N}(X^k)}\right) + \log(\Theta_\epsilon/\eta) + \frac{HN}{2}\log(1+2\epsilon) \\
&\leq \log(\Theta_\epsilon/\eta) + HN\log(1+2\epsilon),
\end{aligned}
$$

where we used the fact that θ_N is the maximum likelihood estimator. With a proper scaling of $\epsilon \ll 1/(nHN)^4$, we get

$$
TV(\theta^*, \overline{\theta}_N) \leq O\left(\sqrt{\frac{(n+M)\log(nHN) + \log(1/\eta)}{N}}\right).
$$

Finally, it is not hard to show that $TV(\theta_N, \overline{\theta}_N) \leq 2\epsilon H \leq 1/N$. We can conclude that

$$
TV(\theta^*, \theta_N) \leq C \cdot \sqrt{\frac{n\log(nHN) + \log(1/\eta)}{N}},
$$

for some sufficiently large constant $C > 0$.