Reward-Mixing MDPs with Few Latent Contexts are Learnable

Jeongyeol Kwon [1] **Yonathan Efroni** [2] **Constantine Caramanis** [3] **Shie Mannor** [4 5]

Abstract

We consider episodic reinforcement learning in reward-mixing Markov decision processes (RMMDPs): at the beginning of every episode nature randomly picks a latent reward model among M candidates and an agent interacts with the MDP throughout the episode for H time steps. Our goal is to learn a near-optimal policy that nearly maximizes the H time-step cumulative rewards in such a model. Prior work (Kwon et al., 2021a) established an upper bound for RMMDPs with $M = 2$. In this work, we resolve several open questions for the general RMMDP setting. We consider an arbitrary $M \geq 2$ and provide a sample-efficient algorithm–$\texttt{EM}^2$–that outputs an ϵ-optimal policy using $O\left(\epsilon^{-2} \cdot S^d A^d \cdot \text{poly}(H, Z)^d\right)$ episodes, where S, A are the number of states and actions respectively, H is the time-horizon, Z is the support size of reward distributions and $d = O(\min(M, H))$. We also provide a $(SA)^{\Omega(\sqrt{M})}/\epsilon^2$ lower bound, supporting that super-polynomial sample complexity in M is necessary.

1. Introduction

We consider the framework of Latent MDPs (LMDPs), which has been studied in several prior works (Chadès et al., 2012; Brunskill & Li, 2013; Hallak et al., 2015; Steimle et al., 2018; Kwon et al., 2021b) and can be understood as an extension of probabilistic mixture models to the sequential decision making setting. In LMDPs, one MDP is randomly chosen from M possible candidate models at the beginning of every episode, and an agent interacts with the chosen MDP for H time steps of an episode. However, the identity of the chosen MDP is unknown to the agent. We call this the *latent context*. This models the setting where the decision-maker is unable to measure or perhaps even estimate an important identifying feature of the environment.

This problem falls into the general partially observable Markov decision process (POMDP) framework. While versatile, POMDPs are generally hard to learn, primarily because the optimal policy depends on the entire history of the process (Smallwood & Sondik, 1973; Krishnamurthy et al., 2016). To learn near-optimal policies with latent contexts, existing POMDP solutions would require strong assumptions on reachability of the system (*e.g.,* Azizzadenesheli et al. (2016); Guo et al. (2016)) or certain separability assumptions (*e.g.,* see conditions proposed in Liu et al. (2022); Golowich et al. (2022); Kwon et al. (2021b)). However, these assumptions do not necessarily align with the applications (*e.g.,* dynamic web application (Hallak et al., 2015), medical treatment (Steimle et al., 2018), transfer learning (Brunskill & Li, 2013)).

In Kwon et al. (2021a) the authors developed a sample-efficient algorithm in the special case of two reward-mixing MDPs (RMMDPs): when the state transition models are shared across different MDPs, the number of latent contexts is $M = 2$.While that work requires no additional assumptions (notably, reachability and separability), the techniques are specific to $M = 2$. The more general $M \geq 2$ RMMDP setting is yet to be studied and no provable guarantees are known to date.

1.1. Our Contributions

In this work we develop new techniques to resolve several open questions for learning near-optimal policies in RMMDPs with $M \geq 2$. We summarize our main results:

1. We design an algorithm that learns an ϵ-optimal policy for an RMMDP with $M \geq 2$ that interacts with the environment at most $\tilde{O}(\text{poly}(M, H) \cdot SA)^{\min(2M-1, H)}/\epsilon^2$ episodes.
2. For the special case that all the probability distribution of the reward is a strict integral of the base probability, we show that the exponent of S and A can be improved from $O(M)$ to $O(\log M)$. Examples of such cases include the case when all rewards are deterministic

[1] Wisconsin Institute for Discovery, University of Wisconsin-Madison, USA [2] Meta, New York [3] Department of Electrical and Computer Engineering, University of Texas at Austin, USA [4] Technion, Israel [5] Nvidia Research. Correspondence to: Jeongyeol Kwon <kwonchungli@utexas.edu>, Yonathan Efroni <jonathan.efroni@gmail.com>.

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

conditioned on latent contexts.

3. For general instances of RMMDPs, we establish a lower bound of $(SA)^{\Omega(\sqrt{M})}/\epsilon^2$, justifying that a super-polynomial sample complexity in M is necessary. This is the first lower bound for the general RMMDP setting.

Our approach is based on constructing the latent reward model from the $O(\min(M, H))$ first moments of the reward function in different state-action pairs, while estimating the shared transition model using reward-free exploration techniques (Kaufmann et al., 2021). By estimating these quantities we construct an RMMDP which approximates the underlying and unknown RMMDP sufficiently well. Further, we show that the optimal policy of the approximate RMMDP is near optimal for the unknown RMMDP.

1.2. Key Challenge: Circumventing Unidentifiability

Moment matching based algorithms are well known in their ability to learn latent mixture models (see *e.g.,* Moitra & Valiant (2010); Doss et al. (2020) and references therein). Towards applying this technique to the RMMDP setting, suppose we can estimate a correlation of rewards at q-different state-actions $\boldsymbol{x} = (s_i, a_i)_{i=1}^q$, which we refer as a reward moment. If we can estimate reward-moments for all $\boldsymbol{x}$ up to some degree $d \in \mathbb{N}_+$, then we can recover the latent reward model. This is the well-known moment-matching technique in literature on learning finite mixture models.

The RMMDP setting has a fundamental difference: the agent can only access the environment by sampling trajectories. In this case there are many trajectories that simply cannot be realized and have zero probability to be observed under any sampling policy, and the reward moments along this trajectory cannot be estimated. For example, in a loop-free system, any state-action (s_i, a_i) cannot be visited more than once in the same episode, in which case we cannot get any samples of the higher-order moment that repeats the same state more than once. In such a case, the true latent reward model is not identifiable.

This *model unidentifiability* issue can also be found in – seemingly unrelated – literature of learning mixtures of discrete product distributions (Freund & Mansour, 1999; Feldman et al., 2008; Chen & Moitra, 2019). There, the task of learning latent mixture parameters is also challenging due to the model unidentifiability issue, since higher-order statistics with multiplicity cannot be estimated. Hence, most work in this direction focused on the density estimation which minimizes the *statistical distance* between observations, rather than insisting on recovering latent parameters (there are a few exceptions, *e.g.,* Gordon et al. (2021)). In the RMMDP setting, instead of focusing on model identifiablity, we cast the following fundamental question:

How can we efficiently learn a near optimal policy of the RMMDP model?

or, analogously, we ask whether identifiability of the latent model is truly necessary if our ultimate goal is only finding a good policy.

In this work, we answer this question affirmatively. We design a model-based approach that recovers a latent reward model that matches the measurable higher order reward moments. We show this is sufficient to recover an RMMDP model that approximates the *trajectory distributions* for all policies with a requirement to recover the true underlying latent reward model. With this in hand, it is then straightforward to find a near optimal policy for the true underlying RMMDP.

1.3. Related Work

Recent years have witnessed a substantial progress in developing efficient RL algorithms for a number of challenging tasks arising from both theory and practice (*e.g.,* Jaksch et al. (2010); Mnih et al. (2013); Silver et al. (2018); Kober et al. (2013); Bellemare et al. (2016); Azar et al. (2017); Tang et al. (2017)). The standard framework for RL assumes a Markovian environment, where full information on the current state is provided, and the optimal policy depends only on the current observation. In contrast, little is understood on partially observed systems where the underlying state cannot be directly decoded from current observations. Due to the vast volume of literature, we only discuss a few works that are most relevant to us.

Prior work on RMMDPs. The most relevant work to ours is Kwon et al. (2021a). There, the authors considered the RMMDP setting with $M = 2$ and with uniform priors, and designed an algorithm that learns a near optimal policy for this setting. Their approach relies heavily on partial parameter recovery guarantees, which is not possible for $M \geq 3$. In this work, we develop new techniques to tackle RMMDPs with $M \geq 3$ problems, that avoid any parameters estimation. As a further benefit to our different approach that does not attempt to partially estimate parameters, we also improve upon their results for $M = 2$, improving the sample complexity from $O(\epsilon^{-4})$ to the optimal $O(\epsilon^{-2})$ dependence.

Solutions for general POMDPs. As a special case of POMDPs, we may consider applying existing algorithms that learn a near optimal policy of a generic POMDP to the RMMDPs. There is a growing body of work that focuses on the case when single or multiple-step observations from test action sequences are *sufficient statistics* of the environment (*e.g.,* Boots et al. (2011); Krishnamurthy et al. (2016); Azizzadenesheli et al. (2016); Golowich et al. (2022); Efroni et al. (2022); Liu et al. (2022); Zhan et al. (2022)). In such a scenario, latent model parameters can be learned up to some parameter transformations when the system is irreducible or optimistically explored. This approach has been applied to function approximation settings in some recent work under similar sufficient statistic assumptions (*e.g.,* Cai et al. (2022); Zhan et al. (2022); Uehara et al. (2022)). However, RMMDP instances do not necessarily satisfy the statistical sufficiency of test-observation sequences: the latent context cannot necessarily be decoded even in hindsight. Thus, their results do not apply for RMMDPs.

Multitask RL RMMDP can be considered as a special case of multitask reinforcement learning in MDP environments (Taylor & Stone, 2009; Brunskill & Li, 2013; Liu et al., 2016; Hallak et al., 2015) with a different reward function to each task. If we are given a sufficiently long time horizon (and some separation between contexts) for an individual task to identify the context, then we can efficiently learn the latent model by clustering the trajectories. Then, if we can learn the latent model, we can easily learn a near-optimal policy from an estimated model. However, for such condition to hold, we need very long time horizon $H \gg SA$. In many scenarios such as dynamic web applications or medical treatments (Hallak et al., 2015; Steimle et al., 2018), we have a relatively short time-horizon $H = O(1)$ for each task and thus cannot identify the latent context or the latent model. We do not make any assumptions about the length of time-horizon within an episode, or about seperability.

Miscellaneous While we assume that episodes start in a sequential order, in other applications such as in recommendation systems, episodes can proceed in parallel without limit on the time-horizon (Maillard & Mannor, 2014; Gentile et al., 2014; Hu et al., 2021; Kwon et al., 2022). In such problems, the goal is to learn an optimal policy for each episode (or task) as quickly as possible exploiting the similarity between tasks. In contrast, the goal in RMMDP is to learn an optimal adaptive, *i.e.,* history-dependent policy within the limited time horizon.

2. Preliminaries

The reward-mixing Markov decision processes is defined as follows.

Definition 2.1 (Reward-Mixing Markov Decision Process (RMMDP))**.** *An RMMDP $\mathcal{M}$ consists of a tuple $(\mathcal{S}, \mathcal{A}, T, \nu, \{w_m, \mu_m\}_{m=1}^M)$ with a state space $\mathcal{S}$; action space $\mathcal{A}$; a shared transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ that maps a state-action pair and a next state to a probability; ν is a common initial state distribution; $\{w_i\}_{i=1}^M$ are the mixing weights such that at the beginning of every episode a reward model μ_m is chosen with probability w_m; 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.*

We further assume the reward values are finite and bounded.

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.*

As an example, the 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 executing an action a at a state s, as $\mu_m(s, a, z) := \mathbb{P}(r = z \mid m, s, a)$ in a context m. We consider a policy class Π which contains all history-dependent policies $\pi : (\mathcal{S}, \mathcal{A}, \mathcal{Z})^* \times \mathcal{S} \to \mathcal{A}$. We are interested in finding a near-optimal policy $\pi \in \Pi$ that is ϵ-optimal with respect to the optimal value: $V_{\mathcal{M}}^* := \max_{\pi \in \Pi} \mathbb{E}_\pi \left[\sum_{t=1}^H r_t\right]$, where $\mathbb{E}_\pi[\cdot]$ is expectation taken over the model $\mathcal{M}$ with a policy π.

Notation We use $[d] := \{1, \ldots, d\}$ and $[d]_+ := \{0\} \cup [d]$. We often denote a state-action pair (s, a) as one symbol $x = (s, a) \in \mathcal{S} \times \mathcal{A}$. For any length t part of a trajectory $(y_1, y_2, ..., y_t)$, we often simplify the notation as $y_{1:t}$. We use $|\boldsymbol{x}|$ for the length of sequence $\boldsymbol{x}$. For a subset of indices $\mathcal{I} \subseteq [d]$, we write $\boldsymbol{x}_\mathcal{I} := (x_i)_{i \in \mathcal{I}}$ to refer to a subsequence of $\boldsymbol{x}$ at positions $\mathcal{I}$. We define $V_{\mathcal{M}}^\pi$ as an expected cumulative reward for model $\mathcal{M}$ with policy π. Lastly, we denote the cardinality of the state and action space as $S := |\mathcal{S}|$ and $A := |\mathcal{A}|$.

3. Algorithm

The idea for learning a near-optimal policy of an RMMDP for the special case of $M = 2$ with equal mixing weights was developed in Kwon et al. (2021a). With techniques that seem specialized for $M = 2$, the authors show that estimates of the second-order correlation of rewards, measured between different time-steps, is sufficient to find a near-optimal policy. We develop new techniques to address this problem, and extend it beyond the $M = 2$ case.

3.1. Recovering Latent Model from Higher Order Moments

For a general RMMDPs, we define the moment of rewards of degree $q \leq d$ for some $d \in \mathbb{N}_+$ as follows:

$$\mathbf{M}(\boldsymbol{x}, \boldsymbol{z}) := \textstyle\sum_{m=1}^M w_m \Pi_{i=1}^q \mu_m(x_i, z_i), \qquad (1)$$

for every $\boldsymbol{x} = (x_i)_{i=1}^q \in (\mathcal{S} \times \mathcal{A})^{\bigotimes q}$ and $\boldsymbol{z} = (z_i)_{i=1}^q \in \mathcal{Z}^{\bigotimes q}$. To empirically estimate these higher order moments we observe they can be cast as a conditional expectation. Let π be a policy that does not depend on the reward observation.

Algorithm 1 Estimate and Match Moments ($\texttt{EM}^2$)

```
    Input: $d \in \mathbb{N}, \epsilon, \eta \in (0,1), \iota_c > 0$
    // Estimate transition model, initial distribution, moments of latent reward and their uncertainty by pure exploration
    $(\hat{T}, \hat{\nu}, \mathbf{M}_n(\cdot,\cdot), n(\cdot)) \leftarrow \texttt{EstimateMoments}(d, \epsilon, \eta)$ (see Appendix C, Algorithm 2).
    // Construct reward latent model with matching moments
    Find $\{\hat{w}_m, \hat{\mu}_m\}_{m=1}^{M}$ such that
        $|\hat{\mathbf{M}}(\boldsymbol{x}, \boldsymbol{z}) - \mathbf{M}_n(\boldsymbol{x}, \boldsymbol{z})| \le \sqrt{\iota_c / n(\boldsymbol{x})}$
    for all $q \in [d]$, $\boldsymbol{x} \in (\mathcal{S} \times \mathcal{A})^{\bigotimes q}, \boldsymbol{z} \in \mathcal{Z}^{\bigotimes q}$.
    Empirical RMMDP: $\hat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \hat{T}, \hat{\nu}, \{\hat{w}_m, \hat{\mu}_m\}_{m=1}^{M})$.
    Return an optimal policy of $\hat{\mathcal{M}}$.
```

Assume that $\mathbb{P}_\pi(\boldsymbol{x} \text{ observed}) > 0$ where $\boldsymbol{x}$ observed is the event that there exists some $1 \le t_1 < ... < t_q \le H$ such that $x_{t_i} = x_i$, *i.e.,* all state-actions in $\boldsymbol{x}$ are visited in the same episode. Then $\mathbf{M}(\boldsymbol{x}, \boldsymbol{z})$ can be estimated from the conditional expectation:

$$\mathbf{M}(\boldsymbol{x}, \boldsymbol{z}) = \mathbb{E}_\pi[\Pi_{i=1}^{q} \mathbb{1}\{r_{t_i} = z_i\} \,|\boldsymbol{x} \text{ observed}]. \tag{2}$$

Since this is an expectation of observable quantities, we can estimate them from sample trajectories if a roll-in policy can visit $\boldsymbol{x}$ with good probability. We use $n(\boldsymbol{x})$ to denote the number of samples (trajectories) used to estimate $\mathbf{M}(\boldsymbol{x}, \cdot)$.

The key challenge arises when not all $\boldsymbol{x}$ can be visited in the same trajectory, or they can be visited with significantly different probabilities. To quantify this challenge, let $\xi(\boldsymbol{x})$ be the maximum visitation probability of $\boldsymbol{x}$:

$$\xi(\boldsymbol{x}) := \max\nolimits_{\pi \in \Pi} \mathbb{P}_\pi(\boldsymbol{x} \text{ observed}). \tag{3}$$

In the extreme, this quantity can be zero for many $\boldsymbol{x}$'s. For such $\boldsymbol{x}$, we cannot expect to estimate $\mathbf{M}(\boldsymbol{x}, \cdot)$ even with infinite sample trajectories. This results in the unidentifiability of the latent reward model.

It turns out that we only need an estimated moment $\hat{\mathbf{M}}(\boldsymbol{x}, \cdot)$ to be accurate proportionally to $\xi(\boldsymbol{x})$. Intuitively, the smaller $\xi(\boldsymbol{x})$ is the less accurate estimate of $\mathbf{M}(\boldsymbol{x}, \cdot)$ is required. In our analysis, we show that if we can explore the environment to collect samples of higher-order moments in such a way, then trajectory distributions of *all* policies are uniformly close to the true model, even though higher-order moment estimates are non-uniformly accurate.

3.2. Algorithm Overview

The algorithm we introduce and analyze is the Estimate and Match Moments ($\texttt{EM}^2$) procedure, depicted in Algorithm 1. $\texttt{EM}^2$ consists of two stages: (i) collect samples to estimate the transition model, initial distribution and the d-order moments of the reward models for all sequences $\boldsymbol{x} \in (\mathcal{S} \times \mathcal{A})^{\bigotimes d}$, i.e., estimate $\mathbf{M}_n(\boldsymbol{x}, \boldsymbol{z})$ for all $\boldsymbol{x} \in (\mathcal{S} \times \mathcal{A})^{\bigotimes d}$ and $\boldsymbol{z} \in \mathcal{Z}^{\bigotimes d}$ (see definition in equation (1)), (ii) find an approximate RMMDP model $\hat{\mathcal{M}}$ whose reward moments up to degree d match within confidence intervals of $\propto \sqrt{1/n(\boldsymbol{x})}$. One such model is assured to exist since the underlying true model satisfies these constraints.

Remark 3.1 (Estimation of T and ν)**.** *The transition model and initial state distribution can be estimated with pure-exploration schemes e.g., (Kaufmann et al., 2021). They can be either separately estimated or estimated simultaneously with the reward moments.*

Next we elaborate on the estimation of higher-order moments in stage (i), and on finding a latent reward model with matching moments in stage (ii).

3.3. Pure Exploration of Higher-Order Moments

The estimation of $\mathbf{M}(\boldsymbol{x}, \cdot)$ can be carried out in multiple ways. A naive approach for doing that is to iterate over all moments up to degree $q \le d$, all $\boldsymbol{x} \in \cup_{q=1}^{d} (\mathcal{S} \times \mathcal{A})^{\bigotimes q}$ such that $\xi(\boldsymbol{x})$ is larger than some threshold, and estimate the conditional mean via equation (2): by executing roll-in policy that maximizes the probability of observing $\boldsymbol{x}$. Note that the roll-in policy and $\xi(\boldsymbol{x})$ can be approximately computed after estimating the transition dynamics. Although simple, this approach requires many trajectories for collecting samples of moments that are hard to reach, resulting in total sample-complexity of $O(SA)^{2d}$.

A more involved but more systematic way for estimating the higher-order moments $\mathbf{M}(\boldsymbol{x}, \cdot)$ is to employ a pure exploration scheme (Kaufmann et al., 2021) on a higher-order MDP, analogously to the idea developed in Kwon et al. (2021a) for $M = 2$. For completeness, we restate the formal definition of the higher order MDP and the pure-exploration mechanism we use in Appendix C. This procedure allows us to estimate the higher order moments in a more sample efficient manner, relatively to the naive algorithm; the total sample-complexity reduces from $O(SA)^{2d}$ to $O(SA)^d$.

Once the pure exploration phase ends, we have a collection of samples for all moments of degree at most d. Then, for any degree $q \le d$, moment $\boldsymbol{x} \in (\mathcal{S} \times \mathcal{A})^{\bigotimes q}$ with any paired sequence $\boldsymbol{z} \in \mathcal{Z}^{\bigotimes q}$, let the quantity $\mathbf{M}_n(\boldsymbol{x}, \boldsymbol{z})$ be the empirical estimate of $\mathbf{M}(\boldsymbol{x}, \boldsymbol{z})$ (see Algorithm 2 for more details). We remark that we choose our roll-in policy to be independent of past reward observations, and thus we can simply take the average of samples to get $\mathbf{M}_n(\boldsymbol{x}, \boldsymbol{z})$. Using a standard measure of concentration for martingales (Wainwright, 2019), we can show that

$$|\mathbf{M}(\boldsymbol{x}, \boldsymbol{z}) - \mathbf{M}_n(\boldsymbol{x}, \boldsymbol{z})| \le \sqrt{\iota_c / n(\boldsymbol{x})}.$$

The above holds for all combinations of $\boldsymbol{x}$ and $\boldsymbol{z}$ with probability at least $1 - \eta$ by an application of the union bound, when the logarithmic constant is given as $\iota_c = O(d \log(SAZ/\eta))$.

3.4. Finding Latent Model with Matched Moments

Once we obtain the empirical moments $\mathbf{M}_n(\cdot, \cdot)$ from the exploration phase, we can search over all the RMMDP models to find an empirical model $\hat{\mathcal{M}} := (\mathcal{S}, \mathcal{A}, \hat{T}, \hat{\nu}, \{\hat{w}_m, \hat{\mu}_m\}_{m=1}^M)$ that satisfies $\left|\hat{\mathbf{M}}(\boldsymbol{x}, \boldsymbol{z}) - \mathbf{M}_n(\boldsymbol{x}, \boldsymbol{z})\right| \le \sqrt{\iota_c / n(\boldsymbol{x})}$, then we are guaranteed that

$$\left|\mathbf{M}(\boldsymbol{x}, \boldsymbol{z}) - \hat{\mathbf{M}}(\boldsymbol{x}, \boldsymbol{z})\right| \le 2\sqrt{\iota_c / n(\boldsymbol{x})}, \\ \forall (\boldsymbol{x}, \boldsymbol{z}) \in \textstyle\bigcup_{q=1}^d (\mathcal{S} \times \mathcal{A})^{\bigotimes q} \times \mathcal{Z}^{\bigotimes q}. \qquad (4)$$

That is, we find an RMMDP model where its first d moments approximately match the ones of the true model.

Computational Challenges for the Model Recovery Solving equation (4) is a hard computational task. Brute-force approaches, which iterate over all possible candidates, may take time exponential in $O(SA)$. Even for a simpler setting of learning mixtures of discrete n product distributions, it is not obvious how to find the latent parameters that matches all multilinear moments (*i.e.,* moments without any multiplicity) (Feldman et al., 2008; Chen & Moitra, 2019). The best known computational complexity for that problem, with uniform uncertainties, is $O(n/\epsilon)^{O(M^2)}$ due to Chen & Moitra (2019). However, since we have non-uniform uncertainties across all moments, we expect that solving equation (4) is computationally harder problem. We leave these computational challenges as future work, and henceforth focus on the sample-complexity upper bound of learning near optimal policy of RMMDP.

4. Upper Bounds

In this section we highlight the key tools with which we establish a sample complexity guarantee for $\mathtt{EM}^2$. To simplify the discussion, we momentarily assume that transition models are known, *i.e.,* T and ν are given. This section focuses on analyzing the performance difference between two RMMDP models $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$, where $\mathcal{M}^{(1)}$ is the true RMMDP model, and $\mathcal{M}^{(2)}$ is an empirical RMMDP model with the same transition and initial state probabilities T and ν, but different latent reward model and mixing weights, *i.e.,* $T^{(2)} = T, \nu^{(2)} = \nu$, and $w_m^{(2)} = \hat{w}_m, \mu_m^{(2)} = \hat{\mu}_m$. We note that prior knowledge of T and ν is not required in our final result (see Appendix C).

As mentioned earlier, the difference between the value of any fixed policy $\pi \in \Pi$ measured on two RMMDPs can be bounded by the l_1-statistical distance in trajectory distributions. Consider the set of all possible trajectories $\mathcal{T} = (\mathcal{S} \times \mathcal{A} \times \mathcal{Z})^{\bigotimes H}$, that is, any state-action-reward sequence of length H. Then,

$$|V_{\mathcal{M}^{(1)}}^{\pi} - V_{\mathcal{M}^{(2)}}^{\pi}| \le H \cdot \|(\mathbb{P}_{\pi}^{(1)} - \mathbb{P}_{\pi}^{(2)})((x, r)_{1:H})\|_1 \\ = H \cdot \textstyle\sum_{\tau \in \mathcal{T}} |\mathbb{P}_{\pi}^{(1)}(\tau) - \mathbb{P}_{\pi}^{(2)}(\tau)|, \quad (5)$$

For any policy $\pi \in \Pi$, our goal is to show that $\sum_{\tau \in \mathcal{T}} |\mathbb{P}_{\pi}^{(1)}(\tau) - \mathbb{P}_{\pi}^{(2)}(\tau)| \le O(\epsilon/H)$, *i.e.,* the true and empirical models are close in l_1-statistical distance for all history-dependent policies.

In particular, we need to bound the l_1 distance of length H trajectories as a function of the distance between the first $O(\min(M, H))$ reward moments of $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$, and without exponential dependence on H.

The accuracy with which we estimate moment $\mathbf{M}(\boldsymbol{x}, \cdot)$ depends on $n(\boldsymbol{x})$, the number of trajectories that visit $\boldsymbol{x}$. We divide the level of uncertainties of trajectories based on the number of samples collected for each moment. We define the following sets:

$$\mathcal{X}_l = \{\boldsymbol{x} \in \textstyle\bigcup_{q=1}^d (\mathcal{S} \times \mathcal{A})^{\bigotimes q} \mid n(\boldsymbol{x}) \ge n_l\}, \\ \mathcal{E}_l = \{x_{1:H} \in (\mathcal{S} \times \mathcal{A})^{\bigotimes H} \mid \forall q \le d : \\ \forall\, 1 \le t_1 < \ldots < t_q \le H,\ (x_{t_i})_{i=1}^q \in \mathcal{X}_l\}, \quad (6)$$

for a decreasing sequence $(n_l)_{l=1}^L$ which we give in Lemma 4.3. Here, $\mathcal{E}_l$ is a set of length at most d state-actions in which every subsequence has been sampled at least n_l times. Further, observe that $\mathcal{E}_0 \subseteq \mathcal{E}_1 \subseteq \cdots \subseteq \mathcal{E}_L$. We split the set of trajectories into disjoint sets $\mathcal{E}_0' = \mathcal{E}_0$, $\mathcal{E}_{L+1}' = \mathcal{E}_L^c$ and $\mathcal{E}_l' = \mathcal{E}_{l-1}^c \cap \mathcal{E}_l$ and for $l \in [L]$, *i.e.,* a set of trajectories with all correlations of degree at most d sampled more than n_l and at least one set of correlation explored less than n_{l-1} times.

With this definition, we can rewrite the above bound on the l_1 statistical distance between all trajectories *for any policy*:

$$\|\mathbb{P}_{\pi}^{(1)} - \mathbb{P}_{\pi}^{(2)}\|_1 = \sum_{l=0}^{L+1} \sum_{\tau : x_{1:H} \in \mathcal{E}_l'} |\mathbb{P}_{\pi}^{(1)}(\tau) - \mathbb{P}_{\pi}^{(2)}(\tau)| \\ \le \sum_{l=0}^{L+1} \sup_{\pi \in \Pi} \mathbb{P}_{\pi}^{(1)}(x_{1:H} \in \mathcal{E}_l') \cdot O(\epsilon_l), \quad (7)$$

where $\mathbb{P}_{\pi}^{(1)}(x_{1:H} \in \mathcal{E}_l')$, is the probability that a random trajectory τ observed with a roll-in policy π belongs to $\mathcal{E}_l'$, and ϵ_l is the overall statistical distance of trajectory distributions (conditioned on $\mathcal{E}_l'$) at level l. The first relation in equation (7) holds since any trajectory τ belongs to one of the sets $\mathcal{E}_l'$. The main challenge of the analysis is to bound ϵ_l, as a function of the moment distance. We can define this more carefully using the sets above: let δ_l be the maximum error between the d^{th} order reachable moments in level l:

$$\delta_l := \max_{\boldsymbol{x}\in\mathcal{X}_l}\max_{\boldsymbol{z}\in\mathcal{Z}^{\otimes d}}\max_{\mathcal{I}\subseteq[d]}\left|\mathbf{M}^{(1)}(\boldsymbol{x}_\mathcal{I},\boldsymbol{z}_\mathcal{I})-\mathbf{M}^{(2)}(\boldsymbol{x}_\mathcal{I},\boldsymbol{z}_\mathcal{I})\right|. \tag{8}$$

The essential content of Lemma 4.1 and Theorem 4.2 is to show that ϵ_l and δ_l are linearly related, with a term that does not grow exponentially with H.

Lemma 4.1 (Eventwise Total Variance Discrepancy)**.** *Let δ_l be defined as in equation (8), i.e., the maximum mismatch in moments up to degree $d = \min(2M-1, H)$. For any $l \in [L]_+$ and any history-dependent policy $\pi \in \Pi$, we have:*

$$\begin{aligned}&\textstyle\sum_{\tau:x_{1:H}\in\mathcal{E}'_l}|\mathbb{P}^{(1)}_\pi(\tau)-\mathbb{P}^{(2)}_\pi(\tau)|\\&\quad\le\sup_{\pi\in\Pi}\mathbb{P}^{(1)}_\pi(x_{1:H}\in\mathcal{E}'_l)\cdot(4HZ)^d\cdot\delta_l.\end{aligned} \tag{9}$$

Thus, Lemma 4.1 implies the second relation in equation (7), setting $\epsilon_l = (4HZ)^{\min(2M-1,H)}\cdot\delta_l$. Moreover, it generalizes an analogous result that was proved for the $M = 2$ case (see Kwon et al. (2021a), Lemma 4.1). There are several notable differences between these results:

1. The threshold value n_l is set to the order of δ_l^{-2}. In contrast, in Kwon et al. (2021a), n_l was set to be of the order of δ_l^{-4}. Thanks to this improvement, we obtain the optimal dependence in ϵ in our final result, namely, $O(\epsilon^{-2})$ as opposed to $O(\epsilon^{-4})$.

2. We do not rely on a parameter recovery guarantee, which cannot be attained for an RMMDP, without strong identifiability assumptions. Instead, we directly convert the closeness in moments to closeness in total variation distance of the trajectory distributions for all history-dependent policies. We prove this result by an induction argument on the number of contexts and time-horizon. In Chen & Moitra (2019), a similar induction idea is used for showing the robust identifiability of mixtures of discrete product distributions directly from closeness in moments (see Lemma 5.5 in Chen & Moitra (2019)).

To prove this result, we in fact prove a more general result which may be of independent interest:

Theorem 4.2 (Bound on Total Variation from Moment Closeness)**.** *Let $\delta > 0$ and let $\mathcal{M}^{(1)}, \mathcal{M}^{(2)}$ be two RM-MDPs. Assume that $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ have the same transition kernel and initial state distribution, but have different latent reward models, and potentially different number M_1 and M_2 of latent contexts. Define $\mathcal{X}_d$ to be the set of length $d := \min(H, M_1 + M_2 - 1)$ state-action sequences that have nearly matched moments*

$$\mathcal{X}_d := \Big\{\boldsymbol{x}\in(\mathcal{S}\times\mathcal{A})^{\otimes d}\Big|\forall\boldsymbol{z}\in\mathcal{Z}^{\otimes d}: \max_{\mathcal{I}\subseteq[d]}\left|\mathbf{M}^{(1)}(\boldsymbol{x}_\mathcal{I},\boldsymbol{z}_\mathcal{I})-\mathbf{M}^{(2)}(\boldsymbol{x}_\mathcal{I},\boldsymbol{z}_\mathcal{I})\right|\le\delta\Big\}.$$

Let $\mathcal{E}_{\text{tot}}$ be the set of trajectories for which all subsequences of length d are in $\mathcal{X}_d$, i.e., $\mathcal{E}_{\text{tot}}$ is the set of all well-explored trajectories:

$$\mathcal{E}_{\text{tot}} := \left\{x_{1:H}|\forall\, t_1 < \ldots < t_d:\ (x_{t_q})_{q=1}^d \in \mathcal{X}_d\right\}.$$

Then for any subset of well-explored trajectories $\mathcal{E}\subseteq\mathcal{E}_{\text{tot}}$, for any history-dependent policy π, we have

$$\begin{aligned}&\textstyle\sum_{\tau:x_{1:H}\in\mathcal{E}}|\mathbb{P}^{(1)}_\pi(\tau)-\mathbb{P}^{(2)}_\pi(\tau)|\\&\qquad\le\sup_{\pi\in\Pi}\mathbb{P}^{(1)}_\pi(x_{1:H}\in\mathcal{E})\cdot(4HZ)^d\cdot\delta.\end{aligned}$$

This result then, bounds the total variation distance between two models with possibly different number of latent contexts, as long as their moments are close. The proof proceeds by induction on the total number of contexts M_1+M_2. Lemma 4.1 is a direct corollary of Theorem 4.2. We refer the reader to Appendix A for the details.

The results thus far translate bounds in moment distance, to bounds in total-variation distance. The next result bounds the sample complexity required to control the moment distance, i.e., given 4, the sample complexity required to control the probability that a given trajectory belongs to the sets $\mathcal{E}'_l$.

Lemma 4.3. *Consider the sets defined in (6) where we set $n_0 = K/(SA)^d, n_{l+1} = n_l/4$ for $l = 0, 1, \ldots, L$, and L such that $n_L > \iota_c$ and $n_{L+1} \le \iota_c$. Then, there exists a pure-exploration algorithm which takes $\epsilon_{\text{pe}} > 0$ as an input parameter, such that with probability (w.p.) at least $1-\eta$, using at most K episodes, for*

$$K \ge C\cdot(SA)^d\epsilon_{\text{pe}}^{-2}\log(K/\eta), \tag{10}$$

with some absolute constant $C > 0$, we have

$$\sup_{\pi\in\Pi}\mathbb{P}_\pi(x_{1:H}\in\mathcal{E}'_l)\le O\left(H^d\epsilon_{\text{pe}}\cdot\sqrt{n_l/\iota_c}\right). \tag{11}$$

The proof of Lemma 4.3 is given in Appendix C.1.

Now using the above Lemma, Equation (4) to translate it to a moment bound, and Equation (7) and Lemma 4.1, we can bound the difference in expected value between the true and empirical models for an arbitrary history-dependent policy π:

$$\begin{aligned}|V^\pi_{\mathcal{M}^{(1)}}-V^\pi_{\mathcal{M}^{(2)}}|&\le H\cdot\|(\mathbb{P}^{(1)}_\pi-\mathbb{P}^{(2)}_\pi)(\tau)\|_1\\&\le H\textstyle\sum_{l=0}^L H^d\cdot O(\epsilon_{\text{pe}})\cdot(4HZ)^d.\end{aligned}$$

Combining the above results, and taking $\epsilon_{\text{pe}} = \epsilon/(HL(4H^2Z)^d)$, $L = O(\log(n_0)) \le O(\log K)$, and $d = \min(2M-1, h)$, our main result follows, and establishing a sample complexity guarantee for $\texttt{EM}^2$ (see Appendix A for the full proof):

Theorem 4.4 (Sample Complexity of Learning RMMDPs with $M \geq 2$)**.** *Let $d = \min(2M-1, H)$. There exists a universal constant $C > 0$ such that there exists an algorithm using at most K episodes where,*

$$K \geq C \cdot \frac{(SA)^d}{\epsilon^2} \cdot \mathrm{poly}(d, H, Z)^d \cdot \mathrm{poly}\log(K/\eta),$$

and outputs an ϵ-optimal policy w.p. at least $1-\eta$.

4.1. Improved Upper Bound with Integral Probabilities

We have shown that for general instances of RMMDPs, we can learn an ϵ-optimal policy using $O(SA)^{2M-1}$ samples. This upper bound can be significantly improved if we make an additional assumption on latent reward models. Suppose that for any $m \in [M], x \in \mathcal{S}\times\mathcal{A}, z \in \mathcal{Z}$, $\mu_m(x,z)$ can take a value from a finitely discretized set $\mathcal{P} = \{0, 1/P, \ldots, 1\}$ for some positive integer $P \in \mathbb{N}_+$. Under this assumption, we show that estimating $d = \lceil 2P \log M \rceil$ are sufficient to learn a near optimal policy. This translates to an improved $(SA)^{O(\log M)}/\epsilon^2$ sample complexity for under this assumption. An interesting special case of such scenario is when the reward is deterministic conditioned on a context, *i.e.,* $\mu_m(x,z)$ takes value from $\mathcal{P} = \{0,1\}$ with $P = 1$.

This is a reminiscent of quasi-polynomial sample-complexity for learning a mixture of subcubes (Chen & Moitra, 2019), *i.e.,* learning a mixture of binary product distributions $\mathcal{Z} = \{0,1\}$ when the latent model parameters can only take values from $\mathcal{P} = \{0, 1/2, 1\}$. While not used for a more general setting, we show that their main identifiability (of distribution from moments) results can be similarly applied to RMMDP problems with general observation support $\mathcal{Z}$ and integral probability set $\mathcal{P}$.

Lemma 4.5 (Modified Lemma 4.1 for Integral Probabilities)**.** *Suppose $\mu_m(x,z)$ takes values only from $\mathcal{P} = \{0, 1/P, \ldots, 1\}$ for all $m \in [M], x \in \mathcal{X}$ and $z \in \mathcal{Z}$. Let δ_l be defined as in equation (8) for the maximum mismatch in moments up to degree $d = \min(\lceil 2P\log M\rceil, H)$. For any $l \in \{0, 1, ..., L+1\}$ and any history-dependent policy $\pi \in \Pi$, we have*

$$\sum_{\tau : x_{1:H} \in \mathcal{E}_l'} |\mathbb{P}_\pi^{(1)}(\tau) - \mathbb{P}_\pi^{(2)}(\tau)| \leq \sup_{\pi\in\Pi} \mathbb{P}_\pi^{(1)}(x_{1:H} \in \mathcal{E}_l') \cdot M^{O(MP\log P)} \cdot \delta_l. \qquad (12)$$

See Appendix B.4 for the proof. Combining Lemma 4.5 with Lemma 4.3, we get the following quasi-polynomial sample-complexity result for integral reward probabilities:

Theorem 4.6. *Suppose $\mu_m(x,z)$ takes values only from $\mathcal{P} = \{0, 1/P, \ldots, 1\}$ for all $m \in [M], x \in \mathcal{X}$ and $z \in \mathcal{Z}$ where $P \in \mathbb{N}_+$ is an absolute constant. If $H > 2P\log M$, then there exists a universal constant $C > 0$ such that there exists an algorithm using at most K episodes where,*

$$K \geq C \cdot \frac{(SA)^{2P\log M}}{\epsilon^2} \cdot M^{O(M)} \cdot \mathrm{poly}(H, \log(K/\eta)),$$

and outputs an ϵ-optimal policy w.p. at least $1-\eta$.

Note that for the case of deterministic rewards, we can instantiate the above theorem with $P = 1$.

5. Lower Bound

In previous sections, we designed an algorithm that learns a near-optimal policy for general instances of RMMDPs with discrete rewards given $O\left((SA)^{O(M)}\right)$ samples. In this section, we complement this upper bound by showing that a super polynomial dependence on S and A is necessary for $M = \omega(1)$ from information-theoretic standpoint. Specifically, we show that there exists a class of instances which cannot avoid $(SA)^{\Omega(\sqrt{M})}$ sample complexity.

To show this lower bound, we construct the family of hard instances where each instance $\mathcal{M}$ is defined as follows. All the instances share the same dynamics: at every time step $t \in [H]$, the environment visits a unique state s_t^*. At every state $s_t = s_t^*$ (or time step t), all actions except one *correct* action $a_t^* \in \mathcal{A}$ returns a reward sampled from a uniform distribution over a binary alphabet $\{0,1\}$. In this section, we only consider binary rewards, and we omit the z-part for indexing μ_m with (x,z), *i.e.,* use $\mu_m(x)$ to denote $\mu_m(x,1) = \mathbb{E}[\mathbb{1}\{r=1\} | x]$. We also denote $\mathbf{M}(\boldsymbol{x})$ as

$$\mathbf{M}(\boldsymbol{x}) := \textstyle\sum_{m=1}^M w_m \Pi_{i=1}^d \mu_m(x_i).$$

We want to construct an example such that for all but the correct sequence of actions $a_{1:H} = a_{1:H}^*$, distributions of observed reward sequences are not statistically distinguishable from playing uniform actions. Such an example can be constructed by finding a moment-matching correct actions. Specifically, let $d = H = \Omega(\sqrt{M})$ be the desired degree of matching moments that we need for the construction of hard instances. For simplicity, let $\mu_m^* \in \mathbb{R}^d$ be the restriction of μ_m to correct actions, *i.e.,* $\mu_m^*(t) = \mu_m(s_t^*, a_t^*)$ for all $t \in [d]$. We set the family of hard instances of the latent reward model by borrowing a construction from Chen & Moitra (2019).

Lemma 5.1 (Result of Section 4.3 in Chen & Moitra (2019))**.** *There exists some $d = \Omega(\sqrt{M})$ such that for any $\epsilon \leq (2d)^{-2d}$, there exists a realization $\{\mu_m^*\}_{m=1}^M$ and mixing weights $\{w_m\}_{m=1}^M$, such that all degree $q < d$ multilinear moments of μ_m^* is equal to $(1/2)^q$:*

$$\textstyle\sum_{m=1}^M w_m \Pi_{t\in\mathcal{I}} \mu_m^*(t) = (1/2)^q, \qquad \forall \mathcal{I} \subsetneq [d] : |\mathcal{I}| = q.$$

Furthermore, the degree-d moment is ϵ-away from the uniform distribution:

$$\textstyle\sum_{m=1}^M w_m \Pi_{t=1}^d \mu_m^*(t) \geq (1/2)^d + \epsilon.$$

Intuitively, the moment-matching example (up to degree $d-1$) would require to explore almost all possible length $d=H$ sequence of actions, since there would be no information gain if a wrong sequence of actions $a_{1:H} \neq a^*_{1:H}$ is played. We show that any ϵ-optimal policy for any $\epsilon \le (2d)^{-2d}$ needs to play the correct sequence with non-negligible probability:

Lemma 5.2. *Let $\mathcal{M}$ be the lower-bound instance constructed with Lemma 5.1 with $\epsilon \le (2d)^{-2d}$ and $d > 4$. The optimal cumulative rewards for $\mathcal{M}$ is at least $(d/2)+\epsilon \cdot 2^{d-2}$. Furthermore, let π_ϵ be an ϵ-optimal policy for $\mathcal{M}$ with , then we have $\mathbb{P}_{\pi_\epsilon}(a_{1:H} = a^*_{1:H}) \ge 1/4$.*

Note that $(2d)^{-2d} = M^{-O(\sqrt{M})}$ can still be significantly larger than $(SA)^{-\Omega(\sqrt{M})}$. To formalize the lower bound argument, we can use the fundamental equality on information gain with bandit feedback (Garivier et al., 2019) modified to the RMMDP setting :

Lemma 5.3. *Let ψ be any exploration strategy in RMMDPs for K episodes. Let $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ be two RMMDPs with the same transition and initial state probabilities. Let $N_{\psi,x_{1:H}}(K)$ be the number of times that a trajectory τ ends up with a sequence of state-actions $x_{1:H}$ for K episodes. Then,*

$$\sum_{x_{1:H}} \mathbb{E}^{(1)}\left[N_{\psi,x_{1:H}}(K)\right] \cdot KL\left(\mathbb{P}^{(1)}(\cdot|x_{1:H}), \mathbb{P}^{(2)}(\cdot|x_{1:H})\right) = KL\left(\mathbb{P}^{(1)}_\psi(\tau^{1:K}), \mathbb{P}^{(2)}_\psi(\tau^{1:K})\right), \tag{13}$$

where $\mathbb{P}_\psi(\tau^{1:K})$ is a distribution of K trajectories obtained with ψ, and $\mathbb{P}(\cdot|x_{1:H})$ is a marginal probability of a reward sequence $r_{1:H}$ obtained from a fixed test $x_{1:H}$.

We convert this result to an information-theoretic lower bound for learning RMMDPs. Specifically, let $\mathcal{M}^{(1)}$ be the base system where rewards are always uniformly distributed over $\{0,1\}$, and $\mathcal{M}^{(2)} = \mathcal{M}$ be the moment-matching system from Lemma 5.1. If we can give an upper bound to equation (13), then by Pinsker's inequality and Le Cam's two-point method (LeCam, 1973), we can argue that two systems from trajectory observations are not distinguishable, and thus we cannot learn the optimal policy for $\mathcal{M}^{(2)}$.

To see this, note that the left hand side of equation (13) is 0 except for the correct state-action sequence $x^*_{1:H}$. On the other hand, all information from the first model is symmetric over all sequences of (state)-actions, and thus for any exploration strategy ψ, there must exist at least one $x^\psi_{1:H}$ sequence such that $\mathbb{E}^{(1)}\left[N_{\psi,x^\psi_{1:H}}(K)\right] \le A^{-H} \cdot K$. Thus for the moment-matching instance $\mathcal{M}^{(2)}$ with $x^*_{1:H} = x^\psi_{1:H}$, at least $K = \Omega(A^H)$ episodes are necessary to distinguish the two systems from trajectory observations $\tau^{1:K}$. We can translate this argument into an $\Omega(A^H)$ lower bound for learning general RMMDPs, and using the action-amplification argument used in Kwon et al. (2021b), we can obtain an $(SA)^{\Omega(\sqrt{M})}$ lower bound with $d = H = \Omega(\sqrt{M})$.

Theorem 5.4 (Lower Bound for RMMDPs)**.** *There exists a universal constant $C > 0$ and a class of RMMDPs such that to obtain an ϵ-optimal policy for $\epsilon < M^{-C\sqrt{M}}$, we need at least $(SA)^{\Omega(\sqrt{M})}/\epsilon^2$ episodes.*

The proof of Theorem 5.4 follows from Lemma 5.1-5.3, and the full proof can be found in Appendix D.4.

6. Conclusion

In this work, we resolve several major open questions concerning the learnability of the RMMDP setting. We design the $\mathtt{EM}^2$ algorithm and establish an $O(SA)^{O(M)}/\epsilon^2$ upper bound for learning an ϵ-optimal policy of a general RMMDP. Hence, a near optimal policy of an RMMDP can be efficiently learned for $M = O(1)$. We compliment our upper bound with $(SA)^{\Omega(\sqrt{M})}/\epsilon^2$ lower bound.

One natural question is whether our results can be extended to a more general framework of Latent MDP (Kwon et al., 2021b), where the transition dynamics can also depend on latent contexts. We note that dealing with non-identical transitions imposes additional significant challenges. The main bottleneck is that the target higher-order statistics are not easily accessible anymore, since now the way to explore higher order reward (and transition) moments must be learned as well. We believe this is an important question to be addressed in future.

Other future research questions include investigating the gap between the upper and lower bounds, suggesting natural assumptions that can assist in reducing the sample complexity further, and considering the case where M is unknown. Finally, designing a practical algorithm that can operate in large-scale RMMDP problems is an interesting next step to take.

Acknowledgement

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

References

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

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

Bellemare, M. G., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. Unifying count-based exploration and intrinsic motivation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 1479–1487, 2016.

Boots, B., Siddiqi, S. M., and Gordon, G. J. Closing the learning-planning loop with predictive state representations. *The International Journal of Robotics Research*, 30(7):954–966, 2011.

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

Cai, Q., Yang, Z., and Wang, Z. Reinforcement learning from partial observation: Linear function approximation with provable sample efficiency. In *International Conference on Machine Learning*, pp. 2485–2522. PMLR, 2022.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Chadès, I., Carwardine, J., Martin, T., Nicol, S., Sabbadin, R., and Buffet, O. MOMDPs: a solution for modelling adaptive management problems. In *Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI-12)*, 2012.

Chen, S. and Moitra, A. Beyond the low-degree algorithm: mixtures of subcubes and their applications. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pp. 869–880, 2019.

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

Efroni, Y., Jin, C., Krishnamurthy, A., and Miryoosefi, S. Provable reinforcement learning with a short-term memory. *arXiv preprint arXiv:2202.03983*, 2022.

Feldman, J., O'Donnell, R., and Servedio, R. A. Learning mixtures of product distributions over discrete domains. *SIAM Journal on Computing*, 37(5):1536–1564, 2008.

Freund, Y. and Mansour, Y. Estimating a mixture of two product distributions. In *Proceedings of the twelfth annual conference on Computational learning theory*, pp. 53–62, 1999.

Garivier, A., Ménard, P., and Stoltz, G. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

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

Golowich, N., Moitra, A., and Rohatgi, D. Learning in observable pomdps, without computationally intractable oracles. *arXiv preprint arXiv:2206.03446*, 2022.

Gordon, S., Mazaheri, B. H., Rabani, Y., and Schulman, L. Source identification for mixtures of product distributions. In *Conference on Learning Theory*, pp. 2193–2216. PMLR, 2021.

Guo, Z. D., Doroudi, S., and Brunskill, E. A pac rl algorithm for episodic pomdps. In *Artificial Intelligence and Statistics*, pp. 510–518, 2016.

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

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

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

Kaufmann, E., Ménard, P., Domingues, O. D., Jonsson, A., Leurent, E., and Valko, M. Adaptive reward-free exploration. In *Algorithmic Learning Theory*, pp. 865–891. PMLR, 2021.

Kober, J., Bagnell, J. A., and Peters, J. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Krishnamurthy, A., Agarwal, A., and Langford, J. PAC reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, pp. 1840–1848, 2016.

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

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

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

LeCam, L. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pp. 38–53, 1973.

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

Liu, Y., Guo, Z., and Brunskill, E. PAC continuous state online multitask reinforcement learning with identification. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pp. 438–446, 2016.

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

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Moitra, A. and Valiant, G. Settling the polynomial learnability of mixtures of gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 93–102. IEEE, 2010.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Smallwood, R. D. and Sondik, E. J. The optimal control of partially observable markov processes over a finite horizon. *Operations research*, 21(5):1071–1088, 1973.

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

Tang, H., Houthooft, R., Foote, D., Stooke, A., Chen, X., Duan, Y., Schulman, J., De Turck, F., and Abbeel, P. # exploration: A study of count-based exploration for deep reinforcement learning. In *31st Conference on Neural Information Processing Systems (NIPS)*, volume 30, pp. 1–18, 2017.

Taylor, M. E. and Stone, P. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(7), 2009.

Uehara, M., Sekhari, A., Lee, J. D., Kallus, N., and Sun, W. Provably efficient reinforcement learning in partially observable dynamical systems. *arXiv preprint arXiv:2206.12020*, 2022.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Zhan, W., Uehara, M., Sun, W., and Lee, J. D. Pac reinforcement learning for predictive state representations. *arXiv preprint arXiv:2207.05738*, 2022.

A. Proof of Main Theorem 4.4

A.1. Proof of Lemma 4.1

Note that invoking Theorem 4.2, it directly follows that $\mathcal{E}_l' \subseteq \mathcal{E}_{tot}$ with setting $\delta := \delta_l$ for $\mathcal{X}_d$. Lemma 4.1 follows from the conclusion of Theorem 4.2 with $M_1 = M_2 = M$ and $d = \min(2M-1, H)$.

A.2. Proof of Theorem 4.2

We start by unfolding the expression of statistical distance:

$$\sum_{\tau: x_{1:H} \in \mathcal{E}} |\mathbb{P}_\pi^{(1)}(\tau) - \mathbb{P}_\pi^{(2)}(\tau)| = \sum_{x_{1:H}} \sum_{r_{1:H}} \mathbb{1}\left\{x_{1:H} \in \mathcal{E}\right\} \nu(s_1) \Pi_{t=1}^{H-1} T(s_{t+1}|s_t, a_t) \cdot \Pi_{t=1}^{H} \pi(a_t|h_t)$$
$$\times \left| \sum_{m=1}^{M_1} w_m^{(1)} \Pi_{t=1}^{H} \mu_m^{(1)}(x_t, r_t) - \sum_{m=1}^{M_2} w_m^{(2)} \Pi_{t=1}^{H} \mu_m^{(2)}(x_t, r_t) \right|.$$

This can be established by directly using the RMMDP model assumption, the Markovian underlying dynamics and the fact that $T(s_{t+1} \mid s_t, a_t)$ and $\pi(a_t \mid h_t)$ are positive.

With a slight abuse of notation, we compactly define $T_{1:t} := v(s_1) \cdot \Pi_{t'=1}^{t-1} T(s_{t'}+1|s_{t'}, a_{t'})$ and $\pi_{1:t} := \Pi_{t'=1}^{t} \pi(a_{t'}|h_{t'})$. Let $h_t := ((s,a,r)_{1:t-1}, s_t)$ be a history before taking an action at the t^{th} time step. The above can be rewritten as:

$$\sum_{\tau: x_{1:H} \in \mathcal{E}} |\mathbb{P}_\pi^{(1)}(\tau) - \mathbb{P}_\pi^{(2)}(\tau)| = \sum_{x_{1:H}} \sum_{r_{1:H}} \mathbb{1}\left\{\mathcal{E}\right\} T_{1:H} \cdot \pi_{1:H} \cdot \left| \mathbf{M}^{(1)}(x_{1:H}, r_{1:H}) - \mathbf{M}^{(2)}(x_{1:H}, r_{1:H}) \right|,$$

where the equality holds by the definition of the higher-order moments (equation (1)).

The proof proceeds by induction on the number of latent contexts. A similar idea was employed in Chen & Moitra (2019) for a related problem of learning mixtures of product distributions. Specifically, the authors in Chen & Moitra (2019) have shown the statistical closeness between mixtures of product distributions from matching higher-order multilinear moments. The key to proceed with the mathematical induction is to reduce the number of contexts *or* the length of sequence at least by one whenever we process one time-step event.

To apply induction, we first need to check the base case. The base case for Theorem 4.2 is when $M_1 = M_2 = 1$ or $H < M_1 + M_2$. Before we proceed, we define a few definitions on the probability of encountering trajectories of interest.

Additional Notation Let us denote the maximum probability of ending up with a trajectory τ conditioned on a history h, such that the $x_{1:H}$ part belongs to $\mathcal{E}$ as:

$$\mathbb{P}^*(\mathcal{E}|h) := \sup_{\pi \in \Pi} \mathbb{P}_\pi(\mathcal{E}|h). \tag{14}$$

By definition, we have the following inequalities, for history $h_t = ((s,a,r)_{1:t-1}, s_t)$ at time t, action a_t and any history-dependent policy π, we have

$$\mathbb{P}^*(\mathcal{E}|h_H) \geq \sum_{a_H} \mathbb{1}\left\{x_{1:H} \in \mathcal{E}\right\} \pi(a_H|h_H), \quad t = H, \tag{15}$$

$$\mathbb{P}^*(\mathcal{E}|h_t) \geq \sum_{a_t} \mathbb{P}^*(\mathcal{E}|h_t, a_t) \pi(a_t|h_t), \quad t < H, \tag{16}$$

$$\mathbb{P}^*(\mathcal{E}|h_t, a_t) \geq \sum_{s_{t+1}} \mathbb{P}^*(\mathcal{E}|h_{t+1}) T(s_{t+1}|s_t, a_t), \quad t < H. \tag{17}$$

Also, since $\mathbb{P}_{\mathcal{E}}^*(\cdot)$ only depends on the occurance of $x_{1:H}$, any two RMMDP models with the same transition and initial distribution have the same value for $\mathbb{P}_{\mathcal{E}}^*$:

$$\mathbb{P}_{\mathcal{E}}^*(h) = \sup_{\pi \in \Pi} \mathbb{P}_\pi^{(1)}(\mathcal{E}|h) = \sup_{\pi \in \Pi} \mathbb{P}_\pi^{(2)}(\mathcal{E}|h).$$

Hence when we consider the same transition model, we often omit (1) and (2) in superscript from $\mathbb{P}_\pi^{(1)}(\mathcal{E}|h)$ or $\mathbb{P}_\pi^{(2)}(\mathcal{E}|h)$. Also note that $\mathbb{P}_{\mathcal{E}}^*(h_t) = \mathbb{P}_{\mathcal{E}}^*(x_{1:t-1}, s_t)$ and $\mathbb{P}_{\mathcal{E}}^*(h_t, a_t) = \mathbb{P}_{\mathcal{E}}^*(x_{1:t})$ since $\mathbb{P}_{\mathcal{E}}^*$ only depends on the state-action part of the history.

Base Case I: When $M_1 = M_2 = 1$, note that

$$\mathbf{M}(x_{1:H}, r_{1:H}) = \Pi_{t=1}^{H} \mu(x_t, r_t),$$

where we can omit the subscript m for reward models μ since there is only one latent context. Then,

$$\begin{aligned}
\sum_{\tau : x_{1:H} \in \mathcal{E}} |\mathbb{P}_{\pi}^{(1)}(\tau) - \mathbb{P}_{\pi}^{(2)}(\tau)| &= \sum_{x_{1:H}} \sum_{r_{1:H}} \mathbb{1}\left\{\mathcal{E}\right\} \pi_{1:H} T_{1:H} \left| \Pi_{t=1}^{H} \mu^{(1)}(x_t, r_t) - \Pi_{t=1}^{H} \mu^{(2)}(x_t, r_t) \right| \\
&\le \sum_{x_{1:H}} \sum_{r_{1:H-1}} \mathbb{1}\left\{\mathcal{E}\right\} \pi_{1:H} T_{1:H} \left| \Pi_{t=1}^{H-1} \mu^{(1)}(x_t, r_t) - \Pi_{t=1}^{H-1} \mu^{(2)}(x_t, r_t) \right| \cdot \left(\sum_{r_H} \mu^{(1)}(x_H, r_H) \right) \\
&\quad + \sum_{x_{1:H}} \sum_{r_{1:H-1}} \mathbb{1}\left\{\mathcal{E}\right\} \pi_{1:H} T_{1:H} \left(\Pi_{t=1}^{H-1} \mu^{(2)}(x_t, r_t) \right) \cdot \sum_{r_H} \left| \mu^{(1)}(x_H, r_H) - \mu^{(2)}(x_H, r_H) \right|.
\end{aligned}$$

By the moment-closeness condition, note that we have $\left|\mu^{(1)}(x_H, r_H) - \mu^{(2)}(x_H, r_H)\right| \le \delta$. We also know that for any $x_H \in \mathcal{S} \times \mathcal{A}$, we have $\sum_{r_H} \mu^{(1)}(x_H, r_H) = 1$. On one hand, it is easy to verify that

$$\begin{aligned}
&\sum_{x_{1:H}} \sum_{r_{1:H-1}} \mathbb{1}\left\{\mathcal{E}\right\} \pi_{1:H} T_{1:H} \Pi_{t=1}^{H-1} \mu^{(2)}(x_t, r_t) \\
&\quad = \sum_{x_{1:H-1}} \sum_{r_{1:H-1}} \pi_{1:H-1} T_{1:H-1} \Pi_{t=1}^{H-1} \mu^{(2)}(x_t, r_t) \sum_{s_H} T(s_H | x_{H-1}) \sum_{a_H} \mathbb{1}\left\{\mathcal{E}\right\} \pi(a_H | h_H) \\
&\quad \le \sum_{x_{1:H-1}} \sum_{r_{1:H-1}} \pi_{1:H-1} T_{1:H-1} \Pi_{t=1}^{H-1} \mu^{(2)}(x_t, r_t) \sum_{s_H} T(s_H | x_{H-1}) \mathbb{P}_{\mathcal{E}}^{*}(h_H) \\
&\quad \le \sum_{x_{1:H-1}} \sum_{r_{1:H-2}} \pi_{1:H-1} T_{1:H-1} \left(\Pi_{t=1}^{H-2} \mu^{(2)}(x_t, r_t) \right) \mathbb{P}_{\mathcal{E}}^{*}(h_{H-1}, a_{H-1}) \sum_{r_{H-1}} \mu^{(2)}(x_{H-1}, r_{H-1}) \\
&\quad = \sum_{x_{1:H-1}} \sum_{r_{1:H-2}} \pi_{1:H-1} T_{1:H-1} \left(\Pi_{t=1}^{H-2} \mu^{(2)}(x_t, r_t) \right) \mathbb{P}_{\mathcal{E}}^{*}(h_{H-1}, a_{H-1}) \le ... \le \mathbb{P}_{\mathcal{E}}^{*}(\emptyset) = \sup_{\pi \in \Pi} \mathbb{P}_{\pi}(\mathcal{E}),
\end{aligned}$$

where the inequalities come from equation (17) and equation (15). We also have

$$\begin{aligned}
&\sum_{x_{1:H}} \sum_{r_{1:H-1}} \mathbb{1}\left\{\mathcal{E}\right\} \pi_{1:H} T_{1:H} \left| \Pi_{t=1}^{H-1} \mu^{(1)}(x_t, r_t) - \Pi_{t=1}^{H-1} \mu^{(2)}(x_t, r_t) \right| \\
&\quad \le \sum_{x_{1:H-1}} \sum_{r_{1:H-1}} \mathbb{P}_{\mathcal{E}}^{*}(h_{H-1}, a_{H-1}) \pi_{1:H-1} T_{1:H-1} \left| \Pi_{t=1}^{H-1} \mu^{(1)}(x_t, r_t) - \Pi_{t=1}^{H-1} \mu^{(2)}(x_t, r_t) \right|.
\end{aligned}$$

Applying the same argument recursively, we can proceed from $t = H$ to $t = 1$ and get:

$$\sum_{x_{1:H}} \sum_{r_{1:H}} \mathbb{1}\left\{\mathcal{E}\right\} \pi_{1:H-1} T_{1:H} \left| \Pi_{t=1}^{H} \mu^{(1)}(x_t, r_t) - \Pi_{t=1}^{H} \mu^{(2)}(x_t, r_t) \right| \le \mathbb{P}_{\mathcal{E}}^{*}(\emptyset) H Z \delta.$$

Base Case II: If $H \le M_1 + M_2 - 1$, then by the moment-closeness condition,

$$\begin{aligned}
\sum_{\tau : x_{1:H} \in \mathcal{E}} |\mathbb{P}^{1,\pi}(\tau) - \mathbb{P}^{2,\pi}(\tau)| &= \sum_{x_{1:H}} \sum_{r_{1:H}} \mathbb{1}\left\{\mathcal{E}\right\} \pi_{1:H} T_{1:H} \left| \mathbf{M}^{(1)}(x_{1:H}, r_{1:H}) - \mathbf{M}^{(2)}(x_{1:H}, r_{1:H}) \right| \\
&\le \delta \cdot \sum_{x_{1:H}} \sum_{r_{1:H}} \mathbb{1}\left\{\mathcal{E}\right\} \pi_{1:H} T_{1:H} \\
&= \delta \cdot \sum_{x_{1:H-1}} \sum_{r_{1:H-1}} \pi_{1:H-1} T_{1:H-1} \sum_{x_H} \mathbb{1}\left\{\mathcal{E}\right\} \pi(a_H | h_H) T(s_H | s_{H-1}, a_{H-1}) \sum_{r_H} 1 \\
&\le Z\delta \cdot \sum_{x_{1:H-1}} \sum_{r_{1:H-1}} \pi_{1:H-1} T_{1:H-1} \mathbb{P}_{\mathcal{E}}^{*}(h_{H-1}, a_{H-1}) \le \ldots \le \mathbb{P}_{\mathcal{E}}^{*}(\emptyset) \cdot Z^H \delta = \sup_{\pi \in \Pi} \mathbb{P}_{\pi}(\mathcal{E}) \cdot Z^H \delta.
\end{aligned}$$

where inequalities hold due to the moment matching condition and inequalities for $\mathbb{P}_{\mathcal{E}}^{*}(\cdot)$.

Induction on H and $M_1 + M_2$. Suppose that the inductive assumption is true for all two RMMDP models when the total number of latent contexts is less than $M_1 + M_2$, or when the length of episode is less than H. Let $l(x, z)$ be the smallest probability among all latent contexts, *i.e.,*

$$l(x, z) := \min \left(\min_{m \in [M_1]} \mu_m^{(1)}(x, z), \min_{m \in [M_2]} \mu_m^{(2)}(x, z) \right).$$

Note that the moment-closeness condition says that for any $1 \le t_1 < t_2 < \ldots < t_d \le H$,

$$\left| \sum_{m=1}^{M_1} w_m^{(1)} \Pi_{q=1}^d \mu_m^{(1)}(x_{t_q}, z_q) - \sum_{m=1}^{M_2} w_m^{(2)} \Pi_{q=1}^d \mu_m^{(2)}(x_{t_q}, z_q) \right| \le \delta, \qquad \forall x_{1:H} \in \mathcal{E}, \{z_q\}_{q=1}^d \in \mathcal{Z}^{\bigotimes d},$$

and similarly for degree $d-1$ moments of any parts of trajectories in $\mathcal{E}$. Call the event that occurred at $t = 1$ (x_1, r_1). Without loss of generality, suppose that the minimum for $l(x_1, r_1)$ is achieved from the first RMMDP model $\mathcal{M}^{(1)}$. Define $p^{(1)} := \sum_{m=1}^{M_1} w_m^{(1)} \left(\mu_m^{(1)}(x_1, r_1) - l(x_1, r_1) \right)$ and $p^{(2)} := \sum_{m=1}^{M_1} w_m^{(2)} \left(\mu_m^{(2)}(x_1, r_1) - l(x_1, r_1) \right)$. By the moment closeness condition, $\left| p^{(1)} - p^{(2)} \right| \le \delta$. Note that in each model, we can decompose the probability of each trajectory $\tau = (x, r)_{1:H}$ as

$$\begin{aligned}
&\pi_{1:H} T_{1:H} \cdot \sum_{m=1}^{M_1} w_m^{(1)} \Pi_{t=1}^H \mu_m^{(1)}(x_t, r_t) \\
&\qquad = \mathbb{P}_\pi^{(1)}(x_1, r_1) l(x_1, r_1) \pi_{2:H} T_{1:H-1} \cdot \sum_{m=1}^{M_1} w_m^{(1)} \Pi_{t=2}^H \mu_m^{(1)}(x_t, r_t) \\
&\qquad + \mathbb{P}_\pi^{(1)}(x_1, r_1) \pi_{2:H} T_{1:H-1} \cdot \sum_{m=1}^{M_1} w_m^{(1)} (\mu_m^{(1)}(x_1, r_1) - l(x_1, r_1)) \Pi_{t=2}^H \mu_m^{(1)}(x_t, r_t).
\end{aligned}$$

Let us define two auxiliary models $\mathcal{M}^{(3)}$ and $\mathcal{M}^{(4)}$ as follows:

1. $\mathcal{M}^{(3)}$ has the transition model $T^{(3)}(\cdot) := T^{(1)}(\cdot)$, initial state distribution $\nu^{(3)}(\cdot) := T^{(1)}(\cdot | s_1, r_1)$, latent reward models $\mu_m^{(3)}(\cdot) := \mu_m^{(1)}(\cdot)$, and mixing weights $w_m^{(3)} := \frac{1}{p^{(1)}} w_m^{(1)} (\mu_m^{(1)}(x_1, r_1) - l(x_1, r_1))$.
2. $\mathcal{M}^{(4)}$ is defined similarly as $\mathcal{M}^{(3)}$ from $\mathcal{M}^{(2)}$, except for the mixing weights $w_m^{(4)} := \frac{1}{p^{(2)}} w_m^{(2)} (\mu_m^{(2)}(x_1, r_1) - l(x_1, r_1))$.

Note that $\mathcal{M}^{(3)}$ has at most $M_1 - 1$ non-zero mixing weights, since $l(x_1, r_1)$ must match at least one of reward probabilities $\{\mu_m^{(1)}(x_1, r_1)\}_{m=1}^{M_1}$. Hence we can leave out only non-zero mixing weights in $\{w_m^{(3)}\}_{m=1}^{M_1}$ and consider as if there are only $M_1 - 1$ latent contexts in $\mathcal{M}^{(3)}$.

Let us define $\mathcal{E}_{x_1} := \{x_{2:H} | (x_1, x_{2:H}) \in \mathcal{E}\}$, a subset of trajectories in $\mathcal{E}$ starting from x_1. Note that $\mathbb{P}^\pi(x_{2:H} \in \mathcal{E}_{x_1} | x_1) \le \mathbb{P}_\mathcal{E}^*(x_1)$. We can decompose the statistical distance of trajectories as the following:

$$\begin{aligned}
&\sum_{x_{1:H}} \sum_{r_{1:H}} \mathbb{1}\{\mathcal{E}\} \pi_{1:H} T_{1:H} \left| \sum_{m=1}^{M_1} w_m^{(1)} \Pi_{t=1}^H \mu_m^{(1)}(x_t, r_t) - \sum_{m=1}^{M_2} w_m^{(2)} \Pi_{t=2}^H \mu_m^{(2)}(x_t, r_t) \right| \\
&\le \sum_{x_1, r_1} \mathbb{P}^{1,\pi}(x_1) l(x_1, r_1) \cdot \sum_{x_{2:H}, r_{2:H}} \mathbb{1}\{\mathcal{E}_{x_1}\} \pi_{2:H} T_{2:H} \left| \mathbf{M}^{(1)}(x_{2:H}, r_{2:H}) - \mathbf{M}^{(2)}(x_{2:H}, r_{2:H}) \right| \\
&\ + \sum_{x_1, r_1} \mathbb{P}^{1,\pi}(x_1) \cdot \sum_{x_{2:H}, r_{2:H}} \mathbb{1}\{\mathcal{E}_{x_1}\} \pi_{2:H} T_{2:H} \cdot \left| p^{(1)} \mathbf{M}^{(3)}(x_{2:H}, r_{2:H}) - p^{(2)} \mathbf{M}^{(4)}(x_{2:H}, r_{2:H}) \right|.
\end{aligned}$$

We observe that in the first term, the summation starting from $t = 2$ to H can be considered as a statistical difference between two RMMDP models $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ with a new common initial state distribution $\nu'(\cdot) := T(\cdot | s_1, a_1)$ in a shorter time-horizon of length $H - 1$. A new policy $\pi'(\cdot | h)$ is $\pi(\cdot | h, r_1, x_1)$ in this setup. Note that the moment-closeness condition remains the same, and therefore by inductive assumption on H, we have

$$\sum_{x_{2:H}, r_{2:H}} \mathbb{1}\{\mathcal{E}_{x_1}\} \pi_{2:H} T_{2:H} \left| \mathbf{M}^{(1)}(x_{2:H}, r_{2:H}) - \mathbf{M}^{(2)}(x_{2:H}, r_{2:H}) \right| \le \mathbb{P}_\mathcal{E}^*(x_1) \cdot (4(H-1)Z)^d \delta.$$

For the second term, we show in Section A.2.1 that

$$\begin{aligned}\sum_{x_{2:H},r_{2:H}} \mathbb{1}\left\{\mathcal{E}_{x_1}\right\} \pi_{2:H} T_{2:H} \cdot \left| p^{(1)} \mathbf{M}^{(3)}(x_{2:H}, r_{2:H}) - p^{(2)} \mathbf{M}^{(4)}(x_{2:H}, r_{2:H}) \right| \\ \leq \mathbb{P}^*_{\mathcal{E}}(x_1) \cdot 4 \cdot (4(H-1)Z)^{d-1} \delta. \end{aligned} \tag{18}$$

Assuming this, we have

$$\begin{aligned} &\sum_{x_{1:H}} \sum_{r_{1:H}} \mathbb{1}\left\{\mathcal{E}\right\} \pi_{1:H} T_{1:H} \left| \sum_{m=1}^{M_1} w_m^{(1)} \Pi_{t=1}^H \mu_m^{(1)}(x_t, r_t) - \sum_{m=1}^{M_2} w_m^{(2)} \Pi_{t=2}^H \mu_m^{(2)}(x_t, r_t) \right| \\ &\leq \sum_{x_1} \mathbb{P}^{(1)}_{\pi}(x_1) \mathbb{P}^*_{\mathcal{E}}(x_1) \cdot \sum_{r_1} (4(H-1)Z)^d \delta \cdot l(x_1, r_1) \\ &+ \sum_{x_1} \mathbb{P}^{(1)}_{\pi}(x_1) \mathbb{P}^*_{\mathcal{E}}(x_1) \cdot \sum_{r_1} 4 \cdot (4(H-1)Z)^{d-1} \delta. \end{aligned}$$

Note that $\sum_{r_1} l(x_1, r_1) \leq \sum_{r_1} \mu_m^{(1)}(x_1, r_1) \leq 1$ for any fixed $m \in [M]$, and thus the above can be further bounded by

$$\delta (4(H-1))^{d-1} Z^d \cdot \left(\sum_{x_1} \mathbb{P}^{(1)}_{\pi}(x_1) \mathbb{P}^*_{\mathcal{E}}(x_1) \right) \cdot (4(H-1) + 4) \leq \mathbb{P}^*_{\mathcal{E}}(\emptyset) \cdot (4HZ)^d \delta,$$

which proves the result for M_1, M_2, H. Applying the same argument inductively for all increasing M_1, M_2 and H, the above also holds for $M_1 = M_2 = M$ and $d = \min(2M-1, H)$.

A.2.1. Proof of Equation (18)

We first separate a subtle issue of mismatch between $p^{(1)}$ and $p^{(2)}$:

$$\begin{aligned} &\sum_{x_{2:H},r_{2:H}} \mathbb{1}\left\{\mathcal{E}_{x_1}\right\} \pi_{2:H} T_{2:H} \cdot \left| p^{(1)} \mathbf{M}^{(3)}(x_{2:H}, r_{2:H}) - p^{(2)} \mathbf{M}^{(4)}(x_{2:H}, r_{2:H}) \right| \\ &\quad \leq p^{(1)} \cdot \sum_{x_{2:H},r_{2:H}} \mathbb{1}\left\{\mathcal{E}_{x_1}\right\} \pi_{2:H} T_{2:H} \left| \mathbf{M}^{(3)}(x_{2:H}, r_{2:H}) - \mathbf{M}^{(4)}(x_{2:H}, r_{2:H}) \right| \\ &\quad + |p^{(1)} - p^{(2)}| \cdot \sum_{x_{2:H},r_{2:H}} \mathbb{1}\left\{\mathcal{E}_{x_1}\right\} \pi_{2:H} T_{2:H} \mathbf{M}^{(4)}(x_{2:H}, r_{2:H}). \end{aligned}$$

Since $|p^{(1)} - p^{(2)}| \leq \delta$, we have

$$|p^{(1)} - p^{(2)}| \cdot \sum_{x_{2:H},r_{2:H}} \mathbb{1}\left\{\mathcal{E}_{x_1}\right\} \pi_{2:H} T_{2:H} \mathbf{M}^{(4)}(x_{2:H}, r_{2:H}) \leq \mathbb{P}^*_{\mathcal{E}}(x_1) \delta.$$

For the remaining term, we examine the moment-closeness condition for the auxiliary model $\mathcal{M}^{(3)}$ and $\mathcal{M}^{(4)}$. If $M_1 = 1$, then we must have $p^{(1)} = 0$ and thus the remaining term is 0. Hence we focus on the case that $M_1 > 1$. We can consider two cases: if $p^{(1)} \leq 4\delta$, then instead of using the moment-closeness condition, we apply

$$\begin{aligned} &p^{(1)} \cdot \sum_{x_{2:H},r_{2:H}} \mathbb{1}\left\{\mathcal{E}_{x_1}\right\} \pi_{2:H} T_{2:H} \left| \mathbf{M}^{(3)}(x_{2:H}, r_{2:H}) - \mathbf{M}^{(4)}(x_{2:H}, r_{2:H}) \right| \\ &\quad \leq 4\delta \cdot \sum_{x_{2:H},r_{2:H}} \mathbb{1}\left\{\mathcal{E}_{x_1}\right\} \pi_{2:H} T_{2:H} \left(\mathbf{M}^{(3)}(x_{2:H}, r_{2:H}) + \mathbf{M}^{(4)}(x_{2:H}, r_{2:H}) \right) \\ &\quad \leq \mathbb{P}^*_{\mathcal{E}}(x_1) \cdot 8\delta, \end{aligned}$$

and we are done as long as $(4(H-1)Z)^{d-1} > 8$. Otherwise, let us compare the moments of degree up to $d-1$ in $\mathcal{M}^{(3)}$ and $\mathcal{M}^{(4)}$. Consider any moments consisting of $q \leq d-1$ pairs of state-actions in any trajectory $x_{2:H} \in \mathcal{E}_{x_1}$ at non-overlapping time-steps $2 \leq t_1 < \ldots < t_q \leq H$. For any $\boldsymbol{z} = z_{1:q} \in \mathcal{Z}^{\bigotimes q}$ with $\boldsymbol{x} = (x_{t_{q'}})_{q'=1}^q$, we can check that

$$\left| \mathbf{M}^{(3)}(\boldsymbol{x}, \boldsymbol{z}) - \mathbf{M}^{(4)}(\boldsymbol{x}, \boldsymbol{z}) \right| = \left| \sum_{m=1}^{M_1-1} w_m^{(3)} \Pi_{q'=1}^q \mu_m^{(3)}(x_{t_{q'}}, z_{q'}) - \sum_{m=1}^{M_2} w_m^{(4)} \Pi_{q'=1}^q \mu_m^{(4)}(x_{t_{q'}}, z_{q'}) \right|.$$

Recall that

$$\sum_{m=1}^{M_1-1} w_m^{(3)} \Pi_{q'=1}^{q} \mu_m^{(3)}(x_{t_{q'}}, z_{q'}) = \frac{1}{p^{(1)}} \sum_{m=1}^{M_1} w_m^{(1)} (\mu_m^{(1)}(x_1, r_1) - l(x_1, r_1)) \Pi_{q'=1}^{q} \mu_m^{(1)}(x_{t_{q'}}, z_{q'}),$$

and similarly for the moments in $\mathcal{M}^{(4)}$. Hence we can decompose the moment difference as the following:

$$\begin{aligned}
&\left| \mathbf{M}^{(3)}(\boldsymbol{x}, \boldsymbol{z}) - \mathbf{M}^{(4)}(\boldsymbol{x}, \boldsymbol{z}) \right| \\
&\leq \frac{1}{p^{(1)}} \left| \sum_{m=1}^{M_1} w_m^{(1)} \mu_m^{(1)}(x_1, r_1) \Pi_{q'=1}^{q} \mu_m^{(1)}(x_{t_{q'}}, z_{q'}) - \sum_{m=1}^{M_2} w_m^{(2)} \mu_m^{(2)}(x_1, r_1) \Pi_{q'=1}^{q} \mu_m^{(2)}(x_{t_{q'}}, z_{q'}) \right| \\
&+ \frac{l(x_1, r_1)}{p^{(1)}} \left| \sum_{m=1}^{M_1} w_m^{(1)} \Pi_{q'=1}^{q} \mu_m^{(1)}(x_{t_{q'}}, z_{q'}) - \sum_{m=1}^{M_2} w_m^{(2)} \Pi_{q'=1}^{q} \mu_m^{(2)}(x_{t_{q'}}, z_{q'}) \right| \\
&+ \left| \frac{1}{p^{(1)}} - \frac{1}{p^{(2)}} \right| \cdot \left| \sum_{m=1}^{M_1} w_m^{(2)} (\mu_m^{(2)}(x_1, r_1) - l(x_1, r_1)) \Pi_{q'=1}^{q} \mu_m^{(2)}(x_{t_{q'}}, z_{q'}) \right|.
\end{aligned}$$

By the moment-closeness condition for $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ up to degree d, the first two terms can be easily bounded by $\frac{2\delta}{p^{(1)}}$. For the last term, note that

$$\sum_{m=1}^{M_1} w_m^{(2)} (\mu_m^{(2)}(x_1, r_1) - l(x_1, r_1)) \Pi_{q'=1}^{q} \mu_m^{(2)}(x_{t_{q'}}, z_{q'}) \leq \sum_{m=1}^{M_1} w_m^{(2)} (\mu_m^{(2)}(x_1, r_1) - l(x_1, r_1)) \leq p^{(2)},$$

and also $|p^{(1)} - p^{(2)}| \leq \delta$, and thus the last term is bounded by $\delta / p^{(1)}$. Therefore, we can conclude that $\mathcal{M}^{(3)}$ and $\mathcal{M}^{(4)}$ satisfies the moment-closeness condition (regarding trajectories in $\mathcal{E}_{x_1}$) with $\delta' = 3\delta / p^{(1)}$. Applying the inductive assumption for $M_1 - 1, M_2$ and $H - 1$, we have

$$\begin{aligned}
p^{(1)} \cdot &\sum_{x_{2:H}, r_{2:H}} \mathbb{1}\{\mathcal{E}_{x_1}\} \pi_{2:H} T_{1:H-1} \left| \mathbf{M}^{(3)}(x_{2:H}, r_{2:H}) - \mathbf{M}^{(4)}(x_{2:H}, r_{2:H}) \right| \\
&\leq \mathbb{P}_{\mathcal{E}}^*(x_1) \cdot 3 \cdot (4(H-1)Z)^{d-1} \cdot \delta.
\end{aligned}$$

Finally, we can apply the results to

$$\begin{aligned}
&\sum_{x_{2:H}, r_{2:H}} \mathbb{1}\{\mathcal{E}_{x_1}\} \pi_{2:H} T_{1:H-1} \cdot p^{(1)} \left| p^{(1)} \mathbf{M}^{(3)}(x_{2:H}, r_{2:H}) - p^{(2)} \mathbf{M}^{(4)}(x_{2:H}, r_{2:H}) \right| \\
&\qquad \leq \mathbb{P}_{\mathcal{E}}^*(x_1)(\delta + 3 \cdot (4(H-1)Z)^{d-1} \cdot \delta) \leq \mathbb{P}_{\mathcal{E}}^*(x_1) \cdot 4 \cdot (4(H-1)Z)^{d-1} \cdot \delta.
\end{aligned}$$

This proves (18), and thus completes Lemma 4.1.

A.3. Proof of Theorem 4.4

We first note that

$$(\mathcal{S} \times \mathcal{A} \times \mathcal{Z})^{\bigotimes H} = \bigcup_{l=0}^{L+1} \mathcal{E}_l'.$$

Thus, we can split the sum over all trajectories into $L + 2$ levels, to bound the statistical distance between trajectory distributions (reiterating equation (7)):

$$\|(\mathbb{P}_\pi^{(1)} - \mathbb{P}_\pi^{(2)})(\tau)\|_1 = \sum_{l=0}^{L+1} \sum_{\tau: x_{1:H} \in \mathcal{E}_l'} |\mathbb{P}_\pi^{(1)}(\tau) - \mathbb{P}_\pi^{(2)}(\tau)| \leq \sum_{l=0}^{L+1} \sup_{\pi \in \Pi} \mathbb{P}_\pi^{(1)}(x_{1:H} \in \mathcal{E}_l') \cdot O(\epsilon_l).$$

Note that this holds for all history-dependent policies. Then we apply the results from Lemma 4.1 and 4.3, which yields

$$\begin{aligned}\|(\mathbb{P}^{(1)}_\pi - \mathbb{P}^{(2)}_\pi)(\tau)\|_1 &\le \sum_{l=0}^{L+1} \sup_{\pi\in\Pi} \mathbb{P}^{(1)}_\pi(x_{1:H}\in\mathcal{E}'_l)\cdot O(\epsilon_l) \\ &\le \sum_{l=0}^{L+1} H^d \epsilon_{\text{pe}}\sqrt{n_l/\iota_c}\cdot(4HZ)^d\cdot O\left(\sqrt{\iota_c/n_l}\right) \\ &\le O(L)H^d\epsilon_{\text{pe}}(4HZ)^d.\end{aligned}$$

Now using $\epsilon_{\text{pe}} = \epsilon/(HL(4H^2Z)^d)$, we get $\|(\mathbb{P}^{(1)}_\pi - \mathbb{P}^{(2)}_\pi)(\tau)\|_1 \le O(\epsilon/H)$, which in turn gives

$$|V^\pi_{\mathcal{M}^{(1)}} - V^\pi_{\mathcal{M}^{(2)}}| \le H\cdot\|(\mathbb{P}^{(1)}_\pi - \mathbb{P}^{(2)}_\pi)((x,r)_{1:H})\|_1 \le O(\epsilon),$$

as desired.

B. Additional Theoretical Results

B.1. Improved Results for $M = 2$

The work of Kwon et al. (2021a) considers the problem of learning RMMDPs for $M = 2$. There, the authors analyze the case in which the mixing weights are balanced, *i.e.*, $w_1 = w_2 = 1/2$. They design an algorithm with sample complexity of $O(poly(H,Z)(SA)^2/\epsilon^4)$. We now show that for the special setting considered in Kwon et al. (2021a) Theorem 4.4 can be improved to yield an upper bound of $O(poly(H,Z)(SA)^2/\epsilon^2)$: strictly improving the dependence from their ϵ^{-4} dependence without resulting in any degradation in the polynomial dependence of (SA).

The following lemma is key to the improved result.

Lemma B.1. *For any RMMDP with $M = 2$ and $w_1 = w_2 = 1/2$, the following holds: for any length three sequences of state-action $\boldsymbol{x} = (x_i)_{i=1}^3$ and rewards $\boldsymbol{z} = (z_i)_{i=1}^3$,*

$$\begin{aligned}\mathbf{M}(\boldsymbol{x},\boldsymbol{z}) = &-2\mathbf{M}(\boldsymbol{x}_{\{1\}},\boldsymbol{z}_{\{1\}})\cdot\mathbf{M}(\boldsymbol{x}_{\{2\}},\boldsymbol{z}_{\{2\}})\cdot\mathbf{M}(\boldsymbol{x}_{\{3\}},\boldsymbol{z}_{\{3\}}) + \mathbf{M}(\boldsymbol{x}_{\{1\}},\boldsymbol{z}_{\{1\}})\cdot\mathbf{M}(\boldsymbol{x}_{\{2,3\}},\boldsymbol{z}_{\{2,3\}}) \\ &+ \mathbf{M}(\boldsymbol{x}_{\{2\}},\boldsymbol{z}_{\{2\}})\cdot\mathbf{M}(\boldsymbol{x}_{\{1,3\}},\boldsymbol{z}_{\{1,3\}}) + \mathbf{M}(\boldsymbol{x}_{\{3\}},\boldsymbol{z}_{\{3\}})\cdot\mathbf{M}(\boldsymbol{x}_{\{1,2\}},\boldsymbol{z}_{\{1,2\}}).\end{aligned}$$

That is, for this special case, if the first and second moments nearly match, then the third moments are also guaranteed to match. Equipped with the above lemma along with Theorem 4.4, we can get a corollary that strictly improves the result of Kwon et al. (2021a):

Corollary B.2 (Improved Sample Complexity for Balanced 2-RMMDPs)**.** *There exists a universal constant $C > 0$ such that if $M = 2$ and $w_1 = w_2 = 1/2$, then there exists an algorithm that produces an ϵ-optimal policy with probability at least $1-\eta$, using at most K episodes, for*

$$K \ge C\cdot\frac{(SA)^2}{\epsilon^2}\cdot\text{poly}(H,Z)\cdot\text{poly}\log(K/\eta).$$

We believe the idea of expressing the third-order moment using lower-order moments can also be applied when the prior is unknown with extra exploration procedures (see, *e.g.*, Appendix E in Kwon et al. (2021a)). We leave this as future work.

B.2. Proof of Lemma B.1

This equation directly comes from unfolding the expression:

$$\begin{aligned}-2\cdot\frac{1}{8}&(\mu_1(x_1,z_1)+\mu_2(x_1,z_1))(\mu_1(x_2,z_2)+\mu_2(x_2,z_2))(\mu_1(x_3,z_3)+\mu_2(x_3,z_3)) \\ &+\frac{1}{4}(\mu_1(x_1,z_1)+\mu_2(x_1,z_1))(\mu_1(x_2,z_2)\mu_1(x_3,z_3)+\mu_2(x_2,z_2)\mu_2(x_3,z_3)) + \dots \\ &= \frac{1}{2}\left(\mu_1(x_1,z_1)\mu_1(x_2,z_2)\mu_1(x_3,z_3)+\mu_2(x_1,z_1)\mu_2(x_2,z_2)\mu_2(x_3,z_3)\right),\end{aligned}$$

after canceling out cross-context multiplied terms.

B.3. Proof of Corollary B.2

With Lemma B.1, we can directly verify that all trajectories in $\mathcal{E}_l$ defined with $d = 2$ satisfies the moment closeness condition (8) up to degree 3 with $\delta_l = O\left(\sqrt{\iota_c/n_l}\right)$. That is, we only need to explore up to second-order moments of the system, but the guarantee on the moment-closeness can be given up to the third-order degree. Thus, we can invoke Lemma 4.1 with $d = 3$, and combine that with Lemma 4.3 with $d = 2$, which gives

$$\begin{aligned}
\|(\mathbb{P}_\pi^{(1)} - \mathbb{P}_\pi^{(2)})(\tau)\|_1 &\le \sum_{l=0}^{L+1} \sup_{\pi \in \Pi} \mathbb{P}_\pi^{(1)}(x_{1:H} \in \mathcal{E}'_l) \cdot O(\epsilon_l) \\
&\le \sum_{l=0}^{L+1} H^2 \epsilon_{\text{pe}} \sqrt{n_l/\iota_c} \cdot (4HZ)^3 \cdot O\left(\sqrt{\iota_c/n_l}\right) \\
&\le O(L) H^2 \epsilon_{\text{pe}} (4HZ)^3,
\end{aligned}$$

where $\epsilon_{\text{pe}} = \epsilon/(H^6 L (4Z)^3)$. Plugging this to the first part of Lemma 4.3, after K exploration episodes where

$$K \ge C \cdot \frac{(SA)^2}{\epsilon_{\text{pe}}^2} \log(K/\eta) = C \cdot \frac{(SA)^2}{\epsilon^2} \cdot \text{poly}(H, Z) \cdot \text{poly}\log(K/\eta),$$

we obtain an $O(\epsilon)$-optimal policy.

B.4. Proof of Lemma 4.5

The proofs here are largely adapted from Chen & Moitra (2019) (see their Lemma 3.1 and 3.8 for the proof of distributional identifiability from low-degree moments). We first define some notation.

We often use a single letter y to denote a pair of state-action and reward (x, z), and thus we use $\mu_m(y) = \mu_m(x, z)$. $\mathcal{Y}_q$ be a q power set of state-action-rewards:

$$\mathcal{Y}_q := \{\boldsymbol{y} = (y_1, \dots, y_q) | (y_1, \dots, y_q) \in (\mathcal{S} \times \mathcal{A} \times \mathcal{Z})^{\bigotimes q}\}.$$

Let $\mathcal{Y}_0 = \{\emptyset\}$ be a null sequence set, and let $\mathcal{Y} := \cup_{q=1}^H \mathcal{Y}_q$ be a set of at most length H sequence (with possible repetitions) of state-action-rewards. Then we define a latent moment matrix $\mathbf{M} \in \mathbb{R}^{|\mathcal{Y}| \times M}$ whose rows are indexed by $\boldsymbol{y} \in \mathcal{Y}$ such that

$$\mathbf{M}(\boldsymbol{y}, m) := \Pi_{q=1}^{|\boldsymbol{y}|} \mu_m(y_q).$$

By convention, $\mathbf{M}(\emptyset, m) = 1$. For any $Y \subseteq \mathcal{Y}$, let $\mathbf{M}_Y$ be a row restriction of $\mathbf{M}$ to Y. We also denote a single row vector $\mathbf{M}_{\boldsymbol{y}}$ indexed by $\boldsymbol{y}$. We denote a length of sequence $\boldsymbol{y}$ as $|\boldsymbol{y}|$. For any $J \subseteq [|\boldsymbol{y}|]$, $\boldsymbol{y}_J$ is a subsequence of $\boldsymbol{y}$ restricted to J. If $J = \emptyset$, then $\boldsymbol{y}_J$ means $\emptyset$.

Now for two RMMDP models $\mathcal{M}^{(1)}$ and $\mathcal{M}(2)$ with the same transition and initial state probabilities, let $\mathbf{M}^{(1)}$ and $\mathbf{M}^{(2)}$ be latent moment matrices respectively, and let $\overline{\mathbf{M}} \in \mathbb{R}^{|\mathcal{Y}| \times 2M}$ be a column-concatenation of two matrices $\overline{\mathbf{M}} = \left[\mathbf{M}^{(1)} | \mathbf{M}^{(2)}\right]$. We first show that for any row of $\overline{\mathbf{M}}$ corresponding to a sequence $\boldsymbol{y}$ of length larger than $d = O(\log M)$, $\overline{\mathbf{M}}_{\boldsymbol{y}}$ is in the row span of $\overline{\mathbf{M}}_{Y(\boldsymbol{y})}$ where $Y(\boldsymbol{y})$ is a set of at most d pairs in $\boldsymbol{y}$

$$Y(\boldsymbol{y}) := \{\boldsymbol{y}_J | \forall J \subseteq [|\boldsymbol{y}|] : |J| \le d\}.$$

Formally, we show the following lemma:

Lemma B.3. *For any $\boldsymbol{y} \in \mathcal{Y}$ with $|\boldsymbol{y}| > d = \lceil 2P \log M \rceil$, the rows of $\overline{M}_{Y(\boldsymbol{y})}$ span all rows in $\overline{M}_{\boldsymbol{y}}$.*

The proof of Lemma B.3 is deferred to Section B.4.1. The implication of Lemma B.3 is crucial: it implies that if we can match up to all degree $d = O(\log M)$ moments exactly, then we can predict probabilities of arbitrary length of trajectories exactly. This means two RMMDP models are identical in terms of trajectory distributions. Of course, we always have a sampling noise in our estimates, and the main challenge is to understand how much the overall statistical error is amplified.

We first observe that the statistical distance between two RMMDP models for any history-dependent policy π can be represented as the following:

$$\sum_{\tau : x_{1:H} \in \mathcal{E}'_l} |\mathbb{P}_\pi^{(1)}(\tau) - \mathbb{P}_\pi^{(2)}(\tau)| = \left\| \mathbf{D} \overline{\mathbf{M}}_{\mathcal{Y}_H} w \right\|_1,$$

where $\mathbf{D}$ is a diagonal matrix whose diagonal element is defined as:

$$\mathbf{D}\left(\boldsymbol{y} = (s,a,r)_{1:H}\right) = \mathbb{1}\left\{\mathcal{E}_l'\right\} \pi_{1:H} T_{1:H},$$

and $w \in \mathbb{R}^{2M}$ is a vector concatenating $w^{(1)}$ and $-w^{(2)}$ such that $w := \left[w^{(1)} | - w^{(2)}\right]^\top$. Let $\mathbf{N}$ be a row restriction of $\overline{\mathbf{M}}$ to pairs of degree $\leq d$ that are explored in $\mathcal{E}_l$, *i.e.,* $\mathbf{N} := \overline{\mathbf{M}}_{\overline{\mathcal{Y}}_l}$ where

$$\overline{\mathcal{Y}}_l := \cup_{\tau \in \mathcal{E}_l} Y(\tau).$$

Here we consider τ in the form of $\boldsymbol{y}$ of length H. By the moment closeness condition, note that

$$\|\mathbf{N}w\|_\infty \leq \delta_l.$$

The remaining steps follow the proof of Lemma 3.8 in Chen & Moitra (2019), and again we rewrite the major procedures for the completeness of the paper. We show by contradiction that if $\left\|\mathbf{D}\overline{\mathbf{M}}_{\mathcal{Y}_H} w\right\|_1 > \mathbb{P}^*_{\mathcal{E}_l'}(\emptyset) \cdot \epsilon$, then it must hold that $\|\mathbf{N}w\|_\infty > M^{-O(M)} \cdot \epsilon$. This concludes Lemma 4.5 by plugging $\epsilon = M^{O(M)} \cdot \delta_l$.

To show this, let $r := \text{rank}(\mathbf{N})$ be the rank of $\mathbf{N}$, and let $\mathbf{N}_r$ be the column restriction of $\mathbf{N}$ to r linearly independent columns. Since the columns of $\mathbf{N}_r$ span all columns of $\mathbf{N}$, we can find $w_r := w + v$ such that $v \in \ker(\mathbf{N})$ and w_r is only supported on the r coordinates corresponding to columns selected by $\mathbf{N}_r$. Since $\mathbf{N}_r$ is full rank, $\sigma_{min}^\infty(\mathbf{N}_r) > 0$ where $\sigma_{min}^\infty(A) := \min_u \|Au\|_\infty / \|u\|_\infty$ for a matrix A. If we can give proper lower bounds for $\sigma_{min}^\infty(\mathbf{N}_r)$ and $\|w_r\|_\infty$, then we can bound $\|\mathbf{N}w\|_\infty \geq \sigma_{min}^\infty(\mathbf{N}_r) \cdot \|w_r\|_\infty$. Now this follows from the two following lemmas.

Lemma B.4. *If a matrix $A \in \mathbb{R}^{n \times k}$ is a full column rank with $n \gg k$, and if all elements of A are integral multiples of some $p > 0$, then $\sigma_{min}^\infty(A) \geq p^k \cdot k^{-O(k)}$.*

Note that all entries of $\mathbf{N}_r$ are integral multiples of $1/P^d$, and thus we have that

$$\sigma_{min}^\infty(\mathbf{N}_r) \geq P^{-dr} r^{-O(r)} \geq M^{-O(MP \log P)}.$$

Since $P = O(1)$, this is bounded below by $M^{-O(M)}$. The proof of Lemma B.4 is given in Section B.4.2. On the other hand, we can show that $\|w_r\|_\infty > \epsilon/M$.

Lemma B.5. *If $\left\|\mathbf{D}\overline{\mathbf{M}}_{\mathcal{Y}_H} w\right\|_1 > \mathbb{P}^*_{\mathcal{E}_l'}(\emptyset) \cdot \epsilon$, then for any $v \in \ker(\mathbf{N})$, $\|w + v\|_\infty > \epsilon/(2M)$.*

Proof. Let $w_r = w + v$. Note that by Lemma B.3, all rows of $\mathbf{D}\overline{\mathbf{M}}_{\mathcal{Y}_H}$ are spanned by the rows of $\mathbf{N}$: for any $\tau \in \mathcal{E}_l'$, then $\overline{\mathbf{M}}_\tau$ is in the span of $\overline{\mathbf{M}}_{Y(\tau)}$ and thus spanned by the rows of $\mathbf{N}$, and otherwise $\mathbf{D}(\tau) = 0$ by definition and thus the row of $\mathbf{D}\overline{\mathbf{M}}_{\mathcal{Y}_H}$ corresponding to τ is 0. Obviously, 0 vector is in the span of the rows of $\mathbf{N}$. Now since $v \in \ker(\mathbf{N})$, $\mathbf{D}\overline{\mathbf{M}}_{\mathcal{Y}_H} v = 0$ for any $v \in \ker(\mathbf{N})$. Therefore we have

$$\left\|\mathbf{D}\overline{\mathbf{M}}_{\mathcal{Y}_H} w\right\|_1 = \left\|\mathbf{D}\overline{\mathbf{M}}_{\mathcal{Y}_H} w_r\right\|_1 \leq \left\|\mathbf{D}\overline{\mathbf{M}}_{\mathcal{Y}_H}\right\|_{1,1} \|w_r\|_\infty,$$

where $\left\|\mathbf{D}\overline{\mathbf{M}}_{\mathcal{Y}_H}\right\|_{1,1}$ is the absolute sum of all elements in $\mathbf{D}\overline{\mathbf{M}}_{\mathcal{Y}_H}$. Note that the sum of the m^{th} column of $\mathbf{D}\overline{\mathbf{M}}_{\mathcal{Y}_H}$ is equal to

$$\sum_{(x,r)_{1:H}} \mathbb{1}\left\{\mathcal{E}_l'\right\} \pi_{1:H} T_{1:H} \Pi_{t=1}^H \mu_m^{(1)}(x_t, r_t) \leq \mathbb{P}^*_{\mathcal{E}_l'}(\emptyset),$$

and similar inequalities hold for the $(M+m)^{th}$ column for $m \in [M]$. Since there are $2M$ columns, $\left\|\mathbf{D}\overline{\mathbf{M}}_{\mathcal{Y}_H}\right\|_{1,1}$ is further bounded by $2M\mathbb{P}^*_{\mathcal{E}_l'}(\emptyset)$. Finally, by a contradicting assumption, we have

$$\epsilon \mathbb{P}^*_{\mathcal{E}_l'}(\emptyset) < 2M\mathbb{P}^*_{\mathcal{E}_l'}(\emptyset) \|w_r\|_\infty,$$

which proves the lemma. □

Combining Lemma B.4 and B.5, we obtain the desired contradiction that $\|\mathbf{N}w\|_\infty > M^{-O(M)}\epsilon$. By letting $\epsilon = M^{O(M)}\delta_l$, we can conclude that $\left\|\mathbf{D}\overline{\mathbf{M}}_{\mathcal{Y}_H} w\right\|_1 \leq \mathbb{P}^*_{\mathcal{E}_l'}(\emptyset) \cdot M^{O(M)}\delta_l$, and we can conclude that

$$\sum_{\tau : x_{1:H} \in \mathcal{E}_l'} |\mathbb{P}_\pi^{(1)}(\tau) - \mathbb{P}_\pi^{(2)}(\tau)| = \left\|\mathbf{D}\overline{\mathbf{M}}_{\mathcal{Y}_H} w\right\|_1 \leq \mathbb{P}^*_{\mathcal{E}_l'}(\emptyset) \cdot M^{O(M)}\delta_l.$$

B.4.1. Proof of Lemma B.3

This largely follows from the proof of Lemma 3.1 in Chen & Moitra (2019), and we rewrite the major procedures in there for the completeness of the paper. We show this lemma by mathematical induction on the length of sequence. For convenience, let $n = |\boldsymbol{y}|$ be the length of target sequence. We show that there exists non-trivial coefficients $\{\alpha_J\}_{J \subseteq [n]}$ such that

$$\sum_{J \subseteq [n]} \alpha_J \overline{\mathbf{M}}_{\boldsymbol{y}_J} = 0,$$

and that $\alpha_{[n]}$ is nonzero. If we can do this inductively from $n = d+1$, then we are done by mathematical induction. We construct an auxiliary polynomial function f of n variables $x = \{x_j\}_{j=1}^n$ such that:

$$f(x) = \Pi_{j=1}^n (x_j - \lambda_j) = \sum_{J \subseteq [n]} \alpha_J \Pi_{j \in J} x_j,$$

for some $\{\lambda_j\}_{j=1}^n$. Note that the coefficient $\alpha_{[n]}$ is always 1. The strategy is to construct a polynomial f such that $f(x) = 0$ at all $x = \{\mu_m^{(b)}(y_j)\}_{j=1}^n$ for all $m \in [M]$ and $b = 1, 2$. Note that any column of $\sum_{J \subseteq [n]} \alpha_J \overline{\mathbf{M}}_{\boldsymbol{y}_J}$ corresponds to one of $\{\mu_m^{(b)}(y_j)\}_{j=1}^n$. The existence of such polynomial f guarantees that $\overline{\mathbf{M}}_{\boldsymbol{y}}$ is in the span of the rows of lower degree pairs in the same sequence, which inductively implies the lemma.

To construct f, we start with $f_0(x) = 1$ at $t = 0$ and inductively construct f_{t+1} from f_t where $f_t(x) = \Pi_{j=1}^t (x_j - \lambda_j)$. At any time step t, define a set of surviving columns $R_t = \left\{(b, m) \middle| f_t\left(\{\mu_m^{(b)}(y_j)\}_{j=1}^n\right) \neq 0\right\}$. Since $\mu_m^{(b)}(\cdot)$ can take values only from the candidate probability set $\mathcal{P}$, by the pigeonhole principle, we can choose $\lambda_{t+1} \in \mathcal{P}$ such that $|R_{t+1}| \le \left\lfloor \left(1 - \frac{1}{P+1}\right) |R_t| \right\rfloor$. Since $|R_0| = 2M$, once t reach $(P+1)\log(2M)$, there will be no surviving columns and we find the desired polynomial $f = f_t(x) \cdot \Pi_{j=t+1}^n x_j$.

B.4.2. Proof of Lemma B.4

This is reminiscent of Lemma 3.7 in Chen & Moitra (2019). We can pick k rows of A such that a row restriction of A to the selected rows, which we denote as A_k, is full rank. By definition, $\sigma_{min}^{\infty,1}(A) \ge \sigma_{min}^{\infty}(A_k)$. Now A_k is a $k \times k$ square matrix and $\det(A_k) > 0$, and thus we can equivalently say

$$\sigma_{min}^{\infty}(A_k) = \min_u \frac{\|A_k u\|_\infty}{\|u\|_\infty} = \min_{u'} \frac{\|u'\|_\infty}{\|A_k^{-1} u'\|_\infty} \ge \min_{u'} \frac{\|u'\|_\infty}{\|A_k^{-1}\|_{\infty,\infty} \|u'\|_1} \ge \frac{1}{k} \frac{1}{\|A_k^{-1}\|_{\infty,\infty}},$$

where $\|A_k^{-1}\|_{\infty,\infty}$ is the largest element of A_k^{-1}. The determinant of any $(k-1) \times (k-1)$ minor is at most $(k-1)! = k^{O(k)}$, and $\det(A_k)$ is some nonzero integral multiple of p^k. Using the Cramer's matrix inversion formula, we can conclude that $\|A_k^{-1}\|_{\infty,\infty} \le p^{-k} k^{O(k)}$.

B.5. Proof of Theorem 4.6

Similarly to the proof of Theorem 4.4 (which can be found in Section A.3), we can show that

$$\begin{aligned}
\|(\mathbb{P}_\pi^{(1)} - \mathbb{P}_\pi^{(2)})(\tau)\|_1 &= \sum_{l=0}^{L+1} \sum_{\tau : x_{1:H} \in \mathcal{E}_l'} |\mathbb{P}_\pi^{(1)}(\tau) - \mathbb{P}_\pi^{(2)}(\tau)| \\
&\le \sum_{l=0}^{L+1} H^d \epsilon_{\text{pe}} \cdot \sqrt{n_l / \iota_c} \cdot M^{O(MP\log P)} \cdot O\left(\sqrt{\iota_c / n_l}\right) \\
&\le L H^d \epsilon_{\text{pe}} M^{O(M)},
\end{aligned}$$

where we used Lemma 4.5 and Lemma 4.3 with $d = 2P\log M$ since $H > 2P\log M$. Plugging $\epsilon_{\text{pe}} = \epsilon / (H^{d+1} L M^{O(M)})$, we have

$$|V_{\mathcal{M}^{(1)}}^\pi - V_{\mathcal{M}^{(2)}}^\pi| \le H \cdot \|(\mathbb{P}_\pi^{(1)} - \mathbb{P}_\pi^{(2)})(\tau)\|_1 \le O(\epsilon).$$

Algorithm 2

```
**Function:** EstimateMoments$(d, \epsilon, \eta)$
Let $\epsilon_{\text{pe}} := \epsilon / (HL(4H^2 Z)^d))$
Initialize $\widetilde{Q}_{(\cdot)}(\cdot) = \widetilde{V}_0 = 1$, $n(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \bigcup_{q=1}^{d} (\mathcal{S} \times \mathcal{A})^{\bigotimes q}$.
Initialize $\hat{T}(\cdot) = \hat{\nu}(\cdot) = 0$, $n_T(\cdot) = 0$.
**while** $\widetilde{V}_0 > \epsilon_{pe}$ **do**
  Get an initial state $s_1$ for the $k^{th}$ episode. Let $\boldsymbol{v}_1 = (\emptyset, \dots, \emptyset)$, $i = 1$, $r_c = 1$
  **for** $t = 1, 2, ..., H$ **do**
    Pick $(a_t, b_t) = \arg\max_{(a,b) \in \widetilde{\mathcal{A}}} \widetilde{Q}_t((i_t, \boldsymbol{v}_t, s_t), (a, b))$.
    Play action $a_t$, observe next state $s_{t+1}$ and reward $r_t$.
    **if** $i_t \le d$ **and** $b_t \neq 0$ **then**
      Record $z_{i_t} = r_t$
    **end if**
    Update $(i_{t+1}, \boldsymbol{v}_{t+1}, s_{t+1})$ according to the choice of $a_t$ and $b_t$ following the rule in (19)
    **if** $i_t \le d$ **and** $i_{t+1} = d + 1$ **then**
      $\boldsymbol{x}_c := (v_{t+1}^j)_{j=1}^{i_t}$, $\boldsymbol{z}_c = (z_j)_{j=1}^{i_t}$
    **end if**
  **end for**
  **if** $i_{H+1} = d + 1$ **then**
    $n(\boldsymbol{x}_c) \leftarrow n(\boldsymbol{x}_c) + 1$
    $\mathbf{M}_n(\boldsymbol{x}_c, \boldsymbol{z}) \leftarrow (1 - 1/n(\boldsymbol{x}_c))\, \mathbf{M}_n(\boldsymbol{x}_c, \boldsymbol{z}) + \mathbb{1}\{\boldsymbol{z} = \boldsymbol{z}_c\} / n(\boldsymbol{x}_c)$ for all $\boldsymbol{z} \in \mathcal{Z}^{\bigotimes \text{length}(\boldsymbol{x})}$
  **end if**
  Update $\hat{T}, \hat{\nu}$ and $n_T$ from a trajectory $(s_1, a_1, s_2, a_2, ..., s_H, a_H)$
  Update $\widetilde{Q}_{(\cdot)}$ and $\widetilde{V}$ using (20), (21)
**end while**
**Return** $\hat{T}, \hat{\nu}, \mathbf{M}_n, n$
```

C. Reward Free Pure-Exploration for Higher-Order Moments

Recall that our goal in the exploration phase is to obtain a collection of samples to estimate reward-moments. In this section, we describe a systematic way of collecting samples using the reward-free exploration scheme. The underlying idea is simple: for each $\boldsymbol{x} = (x_i)_{i=1}^q$ for some $q \le d$, we want to visit all state-actions in $\boldsymbol{x}$ in order. We construct a d^{th}-order MDP where each state is a combination of some state-action sequence and the current state. In this system, if we have visited up to i^{th} elements in $\boldsymbol{x}$ at time-step t, and if the $(i+1)^{th}$ element in $\boldsymbol{x}$ is reached at time $t+1$, then we can advance a state from $((x_1, ..., x_i), s_t)$ to $((x_1, ..., x_i, x_{i+1}), s_{t+1})$. Then we can simplify the goal to reach $(\boldsymbol{x}, \cdot)$ at the end of the episode as much as possible in this system.

We can formalize the above idea. This part mostly follows Kwon et al. (2021a), and we restate most of the definitions for completeness. We employ the reward-free exploration idea for d^{th}-order MDPs, which is defined as the following:

Definition C.1 (d^{th}-Order MDPs)**.** *A d^{th}-order MDP $\widetilde{\mathcal{M}}$ is defined on a state-space $\widetilde{\mathcal{S}}$ and action-space $\widetilde{\mathcal{A}}$ where*

$$\widetilde{\mathcal{S}} = \left\{ (i, \boldsymbol{v}, s) |\; i \in [d+1], \boldsymbol{v} : (v^i)_{i=1}^d \in ((\mathcal{S} \times \mathcal{A}) \cup \{\emptyset\})^{\bigotimes d}, s \in \mathcal{S} \right\},$$
$$\widetilde{\mathcal{A}} = \{(a, b) |\; a \in \mathcal{A}, b \in \{0, 1, -1\}\}.$$

In $\widetilde{\mathcal{M}}$, an augmented state $(i_t, \boldsymbol{v}_t, s_t)$ evolves under an action (a_t, z_t) as follows:

$$i_1 = 1, \boldsymbol{v}_1 = (\emptyset, \dots, \emptyset), s_1 \sim \nu(\cdot), \quad s_{t+1} \sim T(\cdot | s_t, a_t), \quad v_{t+1}^j = v_t^j \;\; \forall j \in [d]/\{i_t\},$$
$$i_{t+1} = \begin{cases} d+1 & \text{if } b_t = -1 \\ i_t + 1 & \text{else if } b_t = 1 \text{ and } i_t \le d \\ i_t & \text{else} \end{cases}, \quad v_{t+1}^{i_t} = \begin{cases} (s_t, a_t) & \text{if } b_t \neq 0 \\ v_t^{i_t} & \text{else} \end{cases} \text{when } i_t \le d. \tag{19}$$

In short, additional state variables i and v select which state-actions to include in a moment to estimate in the current episode.

Additional action variable b selects whether to include or skip the current state-action, or decide a moment to sample with currently saved state-actions in v.

Let us define the upper confidence action-value function $\widetilde{Q}$ and value function $\widetilde{V}$ that is defined as in the form of Bellman-equation w.r.t. $\widetilde{\mathcal{M}}$ with pure-exploration bonus:

$$
\begin{aligned}
&q_c((i,\boldsymbol{v},s),(a,b)) = \mathbb{1}\left\{(i \le d) \cap (i' = d+1)\right\} \cdot \left(1 \wedge \sqrt{\frac{\iota_c}{n(\boldsymbol{x}_c)}}\right), q_T(s,a) = \left(1 \wedge \sqrt{\frac{\iota_T}{n_T(s,a)}}\right) \\
&\widetilde{Q}_t((i,\boldsymbol{v},s),(a,b)) = 1 \wedge \left(q_c((i,\boldsymbol{v},s),(a,b)) + \mathbb{E}_{s' \sim \hat{T}(\cdot|s,a)}\left[\widetilde{V}_{t+1}(i',\boldsymbol{v}',s')\right] + q_T(s,a)\right), \qquad (20)
\end{aligned}
$$

and

$$
\widetilde{V}_t(i,\boldsymbol{v},s) = \max_{(a',b') \in \widetilde{\mathcal{A}}} \widetilde{Q}_t((i,\boldsymbol{v},s),(a',b')), \quad \widetilde{V}_0 = \sqrt{\iota_\nu/k} + \textstyle\sum_s \hat{\nu}(s) \cdot \widetilde{V}_1(1,\boldsymbol{v}_1,s), \qquad (21)
$$

Here, i' and $\boldsymbol{v}'$ are the first and second coordinates of the next state following the (deterministic) transition rule (19) for i and $\boldsymbol{v}$. $\mathbb{1}\left\{(i \le d) \cap (i' = d+1)\right\}$ is an indicator of whether to finish and collect samples for correlations stored in v'. By convention, we let $\widetilde{Q}_{H+1}(\cdot) = 0$. K is the total number of episodes to be explored. The logarithmic factor $\iota_c = O(d\log(SAZ/\eta))$ is properly set confidence interval parameters. The pure-exploration bonus q_c encourages to collect samples for the moments that have not been sufficiently explored yet. This is controlled by the number $n(\boldsymbol{x}_c)$ of collected samples for $\boldsymbol{x}_c = ((v^j)_{j=1}^{i-1}, (s,a))$. Variables n_T, q_T are defined for estimating transition models which we describe below, where $n_T(s,a)$ is the number of total times that (s,a) has been visited, and $q_T(s,a)$ is pure-exploration bonus for visiting (s,a).

Exploration for Estimating Moments In every episode, we take a greedy augmented action (a_t, b_t) that maximizes $\widetilde{Q}_t$ at every time step $t \in [H]$. We continue this pure-exploration process for K episodes until $\widetilde{V}_0 \le \epsilon_{\text{pe}}$ with a threshold parameter ϵ_{pe} for the pure exploration. The pure-exploration procedure is summarized in Algorithm 2. The main purpose of Algorithm 2 is to auto-balance the amount of samples for moments proportional to each moment's reachability.

Estimate Transition Models The transition models and initial state distributions can be easily estimated in the pure-exploration phase. In equation (20), $q_T(\cdot)$ is an exploration bonus term for the uncertainty in transition probabilities, and $\iota_T = O(S\log(K/\eta))$ and $\iota_\nu = O(S\log(K/\eta))$ are properly set confidence constants. Specifically, we can add bonus terms $q_T(\cdot)$ and $\sqrt{\iota_\nu/k}$ to upper-confidence functions $\widetilde{Q}$ and $\widetilde{V}_0$ to encourage the exploration to estimate transition model $\hat{T}$ and initial state distribution $\hat{\nu}$ simultaneously with higher-order moments of latent reward models. The update step (line 2) can be implemented in a straight-forward manner.

Additional Notation We denote $\mathbb{1}_k\{\boldsymbol{x}\}$ as a random variable indicating whether $\boldsymbol{x}$ is visited at the k^{th} episode. Let $\widetilde{\Pi} : \widetilde{\mathcal{S}} \to \widetilde{\mathcal{A}}$ be the class of *stationary* policies in d^{th}-order MDPs. $\widetilde{\pi}_k$ be the policy (greedy with respect to $\widetilde{Q}$) executed in the k^{th} episode. Let $n_k(\boldsymbol{x}) := \sum_{k'=1}^{k-1} \mathbb{1}_{k'}\{\boldsymbol{x}\}$ and the expected quantities $\bar{n}_k(\boldsymbol{x}) := \sum_{k'=1}^{k-1} \mathbb{E}^{\widetilde{\pi}_{k'}}[\mathbb{1}_{k'}\{\boldsymbol{x}\}]$. We define a desired high probability event $\mathcal{E}_{pe}$ for martingale sums:

$$
n_k(\boldsymbol{x}) \ge \frac{1}{2}\bar{n}_k(\boldsymbol{x}) - c_l \cdot d\log(SAK/\eta), \qquad \forall k \in [K], \boldsymbol{x} \in \textstyle\bigcup_{q=1}^{d} (\mathcal{S} \times \mathcal{A})^{\bigotimes q}, \qquad (22)
$$

for some absolute constant $c_l > 0$. With a standard measure of concentration argument for martingale sums (Wainwright, 2019), and taking union bound on all k and $\boldsymbol{x}$, we can show that $\mathbb{P}(\mathcal{E}_{pe}) \ge 1 - \eta$.

We denote $\hat{T}_k, \hat{\nu}_k$ for the empirically estimated transition and initial distribution models at the beginning of k^{th} episode. Similarly to $n_k(\boldsymbol{x})$, let $n_k(s,a), \bar{n}_k(s,a)$ be the actual and expected visit count for a single state-action $(s,a) \in (\mathcal{S} \times \mathcal{A})$, and let $\#_k(s,a)$ be the random variable that the number of times that (s,a) is visited at the k^{th} episode. This is for tracking the uncertainties in $\hat{T}_k$. Similarly to equation (22), it holds that

$$
n_k(s,a) \ge \frac{1}{2}\bar{n}_k(s,a) - c_l \cdot \log(SAK/\eta), \qquad \forall k \in [K], (s,a) \in (\mathcal{S} \times \mathcal{A}),
$$

with probability at least $1 - \eta$.

C.1. Proof of Lemma 4.3

Proof of equation (10): We first show that Algorithm 2 terminates after at most K episodes with probability at least $1-\eta$ where

$$K \geq C \cdot (SA)^d \epsilon_{pe}^{-2} \cdot \log(K/\eta),$$

for some absolute constant $C > 0$. Let us examine $\widetilde{V}_0$ at the k^{th} episode. This can be decomposed as

$$\begin{aligned} \widetilde{V}_0 &= \sqrt{\iota_\nu / k} + \sum_s \hat{\nu}_k(s) \cdot \widetilde{V}_1(i_1, \boldsymbol{v}_1, s) \\ &\leq \sqrt{\iota_\nu / k} + \|\hat{\nu}_k(s) - \nu(s)\|_1 + \sum_s \nu(s) \cdot \max_{(a,b) \in \widetilde{\mathcal{A}}} \widetilde{Q}_1((i_1, \boldsymbol{v}_1, s), (a,b)) \\ &\leq 2\sqrt{\iota_\nu / k} + \mathbb{E}_{\widetilde{\pi}_k} \left[\widetilde{Q}_1((i_1, \boldsymbol{v}_1, s_1), \widetilde{\pi}_k(i_1, \boldsymbol{v}_1, s_1)) \right], \end{aligned}$$

where in the last inequality we used $\|\hat{\nu}_k - \nu(s)\|_1 \leq \sqrt{\iota_\nu / k}$ by standard martingale inequalities. Then, we can recursively bound expectation of $\widetilde{Q}_t$ for $t \geq 1$. For convenience, let us denote $\boldsymbol{x}_t = ((v_t^j)_{j=1}^{i_t - 1}, (s_t, a_t))$ be the moment that can be sampled at the current time step, and

$$\begin{aligned} &\mathbb{E}_{\widetilde{\pi}_k} \left[\widetilde{Q}_t((i_t, \boldsymbol{v}_t, s_t), (a_t, b_t)) \right] \\ &\quad = \mathbb{E}_{\widetilde{\pi}_k} \left[q_r((i_t, \boldsymbol{v}_t, s_t), (a_t, b_t)) + q_T(s_t, a_t) \right] + \mathbb{E}_{\widetilde{\pi}_k} \left[\textstyle\sum_{s_{t+1}} \hat{T}_k(s_{t+1} | s_t, a_t) \cdot \widetilde{V}_{t+1}(i_{t+1}, \boldsymbol{v}_{t+1}, s_{t+1}) \right] \\ &\quad \leq \mathbb{E}_{\widetilde{\pi}_k} \left[\widetilde{Q}_{t+1}((i_{t+1}, \boldsymbol{v}_{t+1}, s_{t+1}), (a_{t+1}, b_{t+1})) \right] \\ &\qquad + 2\mathbb{E}_{\widetilde{\pi}_k} \left[\mathbb{1}\left\{ i_t \leq d \cap i_{t+1} = d+1 \right\} \left(1 \wedge \sqrt{\frac{\iota_c}{n_k(\boldsymbol{x}_t)}} \right) \right] + 2\mathbb{E}_{\widetilde{\pi}_k} \left[1 \wedge \sqrt{\frac{\iota_T}{n_k(s_t, a_t)}} \right], \end{aligned}$$

where in the last inequality, we used that $\|T(\cdot | s_t, a_t) - \hat{T}(\cdot | s_t, a_t)\|_1 \leq \sqrt{\iota_T / n_k(s_t, a_t)}$ by martingale concentration, and $|\widetilde{Q}_{t+1}(\cdot)| \leq 1$. Note that the indicator $\mathbb{1}\left\{ i_t \leq d \cap i_{t+1} = d+1 \right\}$ means whether we collect the sample at the t^{th} time step, *i.e.,* $\mathbb{1}\left\{ \text{collect at } t \right\}$. Putting together, at the k^{th} episode, we have

$$\begin{aligned} \widetilde{V}_0 &\leq 2\sqrt{\iota_\nu / k} + 2 \sum_{t=1}^H \mathbb{E}_{\widetilde{\pi}_k} \left[\left(1 \wedge \sqrt{\frac{\iota_T}{n_k(s_t, a_t)}} \right) + \mathbb{1}\left\{ \text{collect at } t \right\} \left(1 \wedge \sqrt{\frac{\iota_c}{n_k(\boldsymbol{x}_t)}} \right) \right] \\ &= 2\sqrt{\iota_\nu / k} + 2 \sum_{(s,a)} \left(1 \wedge \sqrt{\frac{\iota_T}{n_k(s,a)}} \right) \cdot \mathbb{E}_{\widetilde{\pi}_k} \left[\#_k(s,a) \right] + 2 \sum_{\boldsymbol{x}} \left(1 \wedge \sqrt{\frac{\iota_c}{n_k(\boldsymbol{x})}} \right) \cdot \mathbb{E}_{\widetilde{\pi}_k} \left[\mathbb{1}_k\{\boldsymbol{x}\} \right]. \end{aligned}$$

From equation (22), we have that

$$\begin{aligned} \sum_{\boldsymbol{x}} \left(1 \wedge \sqrt{\frac{\iota_c}{n_k(\boldsymbol{x})}} \right) \cdot \mathbb{E}_{\widetilde{\pi}_k} \left[\mathbb{1}_k\{\boldsymbol{x}\} \right] &\leq 2 \sum_{\boldsymbol{x}} \sqrt{\frac{\iota_c}{1 + \bar{n}_k(\boldsymbol{x})}} \cdot (\bar{n}_{k+1}(\boldsymbol{x}) - \bar{n}_k(\boldsymbol{x})) \\ &\quad + \sum_{\boldsymbol{x}} \mathbb{1}\left\{ \bar{n}_k(\boldsymbol{x}) < 4 \cdot c_l d \log(SAK/\eta) \right\} (\bar{n}_{k+1}(\boldsymbol{x}) - \bar{n}_k(\boldsymbol{x})), \end{aligned}$$

where we used by definition that $\mathbb{E}_{\widetilde{\pi}_k} \left[\mathbb{1}_k\{\boldsymbol{x}\} \right] = (\bar{n}_{k+1}(\boldsymbol{x}) - \bar{n}_k(\boldsymbol{x}))$. Similarly, we also have

$$\begin{aligned} \sum_{(s,a)} \left(1 \wedge \sqrt{\frac{\iota_T}{n_k(s,a)}} \right) \cdot \mathbb{E}_{\widetilde{\pi}_k} \left[\#_k(s,a) \right] &\leq 2 \sum_{s,a} \sqrt{\frac{\iota_T}{1 + \bar{n}_k(s,a)}} \cdot (\bar{n}_{k+1}(s,a) - \bar{n}_k(s,a)) \\ &\quad + H \sum_{(s,a)} \mathbb{1}\left\{ \bar{n}_k(s,a) < 4 \cdot c_l \log(SAK/\eta) \right\} (\bar{n}_{k+1}(s,a) - \bar{n}_k(s,a)). \end{aligned}$$

using an integral inequality $\sum_{k=1}^K \sqrt{1/(1+n_k)}(n_{k+1} - n_k) \leq \int_1^{n_K} 1/x dx$ for any non-decreasing sequence $(n_k)_{k=1}^K$, we can sum over all K episodes until $\widetilde{V}_0 > \epsilon_{\text{pe}}$ and thus, we have

$$K \epsilon_{pe} \leq 4\sqrt{\iota_\nu K} + O(c_l d H (SA)^d \log(KSA/\eta)) + 8 \sum_{(s,a)} \sqrt{\iota_T \bar{n}_{K+1}(s,a)} + 8 \sum_{\boldsymbol{x}} \sqrt{\iota_c \bar{n}_{K+1}(\boldsymbol{x})}.$$

We now note that $\sum_{s,a} \bar{n}_{K+1}(s,a) = HK$ and $\sum_{\boldsymbol{x}} \bar{n}_{K+1}(\boldsymbol{x}) \le K$. Using a Cauchy-Schwartz inequality, we get

$$K\epsilon_{pe} \le O\left(\sqrt{\iota_\nu K} + dH(SA)^d \log(KSA/\eta) + \sqrt{\iota_T HSAK} + \sqrt{\iota_c (SA)^d K} \right).$$

The bound on K is concluded by plugging the confidence parameters, which ensures that K should satisfy

$$K \le O\left(Hd(SA)^d \epsilon_{pe}^{-2} \log(KSA/\eta) \right),$$

until we terminate Algorithm 2 after at most K episodes with probability at least $1-\eta$.

Proof of equation (11): To prove this part, we first note that by union bound, we have

$$\begin{aligned}
\sup_{\pi \in \Pi} \mathbb{P}_\pi(x_{1:H} \in \mathcal{E}_l') &= \sup_{\pi\in\Pi} \mathbb{P}_\pi \left(\bigcup_{q=1}^{d} \bigcup_{1 \le t_1 < \ldots < t_q \le H} (x_{t_i})_{i=1}^q \in \mathcal{X}_l \cap \mathcal{X}_{l-1}^c \right) \\
&\le \sum_{q=1}^{d} \sum_{1 \le t_1 < \ldots < t_q \le H} \sup_{\pi\in\Pi} \mathbb{P}_\pi \left((x_{t_i})_{i=1}^q \in \mathcal{X}_{l-1}^c \right).
\end{aligned}$$

For each fixed q and $\boldsymbol{t} = (t_i)_{i=1}^q$, we consider a sub-class of pure-exploration policies $\widetilde{\Pi}(\boldsymbol{t}) : (\widetilde{\mathcal{S}} \times [H]) \to \widetilde{\mathcal{A}}$ such that each $\widetilde{\pi} \in \widetilde{\Pi}(\boldsymbol{t})$ takes $b_t = 1$ when $t = t_i$ for some $i < q$, $b_t = -1$ when $t = t_q$, and otherwise takes $b_t = 0$. Within this policy class, define the value function $\widetilde{V}^{\boldsymbol{t}}$ and action-value function $\widetilde{Q}^{\boldsymbol{t}}$ with respect to $\widetilde{\Pi}(\boldsymbol{t})$ as the following:

$$\begin{aligned}
& q_T(s,a) = \left(1 \wedge \sqrt{\frac{\iota_T}{n_{K+1}(s,a)}} \right), \\
& q_{t,c}((i,\boldsymbol{v},s),(a,b)) = \mathbb{1}\left\{ t = q \cap \boldsymbol{x}_c \in \mathcal{X}_{l-1}^c \right\} \cdot \left(1 \wedge \sqrt{\frac{\iota_c}{n_{K+1}(\boldsymbol{x}_c)}} \right), \\
& \widetilde{Q}_t^{\boldsymbol{t}}((i,\boldsymbol{v},s),(a,b)) = 1 \wedge \left(q_{t,c}((i,\boldsymbol{v},s),(a,b)) + \mathbb{E}_{s' \sim \hat{T}(\cdot|s,a)} \left[\widetilde{V}_{t+1}^{\boldsymbol{t}}(i', \boldsymbol{v}', s') \right] + q_T(s,a) \right),
\end{aligned}$$

and

$$\widetilde{V}_t^{\boldsymbol{t}}(i,\boldsymbol{v},s) = \max_{a \in \mathcal{A}} \widetilde{Q}_t^{\boldsymbol{t}}((i,\boldsymbol{v},s),(a,b_t)), \quad \widetilde{V}_0^{\boldsymbol{t}} = \sqrt{\iota_\nu / K} + \textstyle\sum_s \hat{\nu}(s) \cdot \widetilde{V}_1^{\boldsymbol{t}}(1, \boldsymbol{v}_1, s).$$

with $\widetilde{Q}_{H+1}^{\boldsymbol{t}} = 0$. By construction, $\widetilde{V}_0$ is an upper confidence bound of $\widetilde{V}_0^{\boldsymbol{t}}$:

$$\epsilon_{\text{pe}} \ge \widetilde{V}_0 \ge \widetilde{V}_0^{\boldsymbol{t}},$$

since $\widetilde{V}_0^{\boldsymbol{t}}$ is computed with more restriction on policies. Note that the exploration-bonus from collecting a sample of moments $q_{t,c}$ is always larger than $\sqrt{\iota_c / n_{l-1}}$. On the other hand, $\sup_\pi \mathbb{P}_\pi \left((x_{t_i})_{i=1}^q \in \mathcal{X}_{l-1}^c \right)$ can be computed through the same dynamic programming on $\widetilde{Q}^*$ with slight changes of exploration bonus:

$$\begin{aligned}
& q_{t,c}((i,\boldsymbol{v},s),(a,b)) = \mathbb{1}\left\{ t = q \cap \boldsymbol{x}_c \in \mathcal{X}_{l-1}^c \right\}, \\
& \widetilde{Q}_t^*((i,\boldsymbol{v},s),(a,b)) = q_{t,c} + \sum_{s'} T(s'|s,a) \cdot \widetilde{V}_{t+1}^*(i', \boldsymbol{v}', s),
\end{aligned}$$

and

$$\widetilde{V}_t^*(i,\boldsymbol{v},s) = \max_{a\in\mathcal{A}} \widetilde{Q}_t^*((i,\boldsymbol{v},s),(a,b_t)).$$

Then,

$$\widetilde{V}_0^* = \sum_s \nu(s) \cdot \widetilde{V}_1^*(1, \boldsymbol{v}_1, s) = \sup_{\pi\in\Pi} \mathbb{P}_\pi \left((x_{t_i})_{i=1}^q \in \mathcal{X}_{l-1}^c \right).$$

Finally, with the setting of confidence interval parameters ι_T for transition errors, we can inductively show that

$$\widetilde{Q}_t^{\boldsymbol{t}} \geq \widetilde{Q}_t^* \cdot \sqrt{\iota_c / n_{l-1}}, \quad \widetilde{V}_t^{\boldsymbol{t}} \geq \widetilde{V}_t^* \cdot \sqrt{\iota_c / n_{l-1}}.$$

This implies that

$$\epsilon_{\text{pe}} \geq \widetilde{V}_0^{\boldsymbol{t}} \geq \sqrt{\iota_c / n_{l-1}} \cdot \sup_{\pi \in \Pi} \mathbb{P}_\pi \left((x_{t_i})_{i=1}^q \in \mathcal{X}_{l-1}^c \right).$$

We conclude the final result (equation (11)):

$$\begin{aligned} \sup_{\pi \in \Pi} \mathbb{P}_\pi (x_{1:H} \in \mathcal{E}_l') &\leq \sum_{q=1}^{d} \sum_{1 \leq t_1 < \ldots < t_q \leq H} \sup_{\pi \in \Pi} \mathbb{P}_\pi \left((x_{t_i})_{i=1}^q \in \mathcal{X}_{l-1}^c \right) \\ &\leq O\left(H^d \epsilon_{\text{pe}} \cdot \sqrt{n_{l-1} / \iota_c} \right). \end{aligned}$$

D. Proofs for the Lower Bound

D.1. Proof of Lemma 5.1

This construction follows from the result in Section 4.3 in Chen & Moitra (2019), and in particular, their Lemma 4.8 with a slight change in constants (*e.g.,* let $\lambda_2 = -\lambda_1 \propto -\epsilon \cdot 2^{-d}$). We refer the readers to Chen & Moitra (2019) for detailed constructions.

D.2. Proof of Lemma 5.2

The optimal policy π^* is the one which always plays optimal actions up to time step $d-1$, and select the last action depending on the conditional expectation of the last reward. Specifically, suppose we played a sequence of actions $(a_t^*)_{t=1}^{d-1}$ and the received a reward sequence $(r_t)_{t=1}^{d-1}$. It is not difficult to verify that the conditional probability of last reward according to a_H^* is given as follows:

$$\mathbb{E}\left[r_d | (a_t^*)_{t=1}^{d-1}, (r_t)_{t=1}^{d-1}, a_d^* \right] = \begin{cases} 1/2 + \epsilon \cdot 2^{d-1} & \text{if } \sum_{t=1}^{d-1} (1 - r_t) \text{ is even} \\ 1/2 - \epsilon \cdot 2^{d-1} & \text{otherwise} \end{cases}, \tag{23}$$

That is, the number of 0 in a sequence $(r_t)_{t=1}^{d-1}$ is even, then the probability of getting 1 is larger, and otherwise the probability of getting 0 is larger. Thus, the optimal policy can play $a_H = a_H^*$ if the number of 0 is even, and play anything else otherwise. Cumulative rewards of the optimal policy is given as follows:

$$\begin{aligned} \mathbb{E}_{\pi^*} \left[\sum_{t=1}^d r_t \right] &= \mathbb{E}_{\pi^*} \left[\sum_{t=1}^{d-1} r_t \right] + \mathbb{E}\left[\mathbb{E}\left[r_d \middle| (a_t)_{t=1}^{d-1} = (a_t^*)_{t=1}^{d-1}, (r_t)_{t=1}^{d-1}, a_d \sim \pi^* \right] \right] \\ &\geq (d-1)/2 + \frac{1}{2}(1/2 + \epsilon \cdot 2^{d-1}) + \frac{1}{2}(1/2) \\ &= d/2 + \epsilon \cdot 2^{d-2}, \end{aligned}$$

where in the second equality, we used the fact that all reward sequences of length $d-1$ has the same marginal probability.

Now for any history-dependent policy π, we note that

$$\mathbb{E}_\pi \left[\sum_{t=1}^d r_t \right] = (d-1)/2 + \mathbb{E}_\pi [r_d] \leq d/2 + \epsilon \cdot 2^{d-1} \cdot \mathbb{P}_\pi (a_{1:d} = a_{1:d}^*).$$

Thus, for any ϵ-optimal policy π_ϵ with $\epsilon < (2d)^{-2d}$, we have

$$d/2 + \epsilon \cdot 2^{d-2} - \epsilon < d/2 + \epsilon \cdot 2^{d-1} \cdot \mathbb{P}_\pi (a_{1:d} = a_{1:d}^*),$$

which in turn implies $\mathbb{P}_\pi (a_{1:d} = a_{1:d}^*) \geq 1/2 - 1/2^{d-1} \geq 1/4$.

D.3. Proof of Lemma 5.3

This is a fundamental equality whose bandit version can be found in *e.g.,* Cesa-Bianchi & Lugosi (2006), Garivier et al. (2019). We start by unfolding the expression for KL-divergence:

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{P}_{\psi}^{(1)}(\tau^{1:K}), \mathbb{P}_{\psi}^{(2)}(\tau^{1:K})\right) &= \mathbb{E}_{\psi}^{(1)}\left[\log\left(\frac{\mathbb{P}_{\psi}^{(1)}(\tau^{1:K})}{\mathbb{P}_{\psi}^{(2)}(\tau^{1:K})}\right)\right] \\
&= \mathbb{E}_{\psi}^{(1)}\left[\log\left(\frac{\mathbb{P}_{\psi}^{(1)}(\tau^{1:K-1})}{\mathbb{P}_{\psi}^{(2)}(\tau^{1:K-1})}\right)\right] + \mathbb{E}_{\psi}^{(1)}\left[\log\left(\frac{\mathbb{P}_{\psi}^{(1)}(\tau^{K}|\tau^{1:K-1})}{\mathbb{P}_{\psi}^{(2)}(\tau^{K}|\tau^{1:K-1})}\right)\right].
\end{aligned}$$

Note that for any $\tau^K = (x_{1:H}^K, r_{1:H}^K)$,

$$\mathbb{P}_{\psi}^{(1)}(\tau^K|\tau^{1:K-1}) = \left(\Pi_{t=1}^{H} T(s_t^K|x_{t-1}^K)\psi(a_t^K|h_t^K, \tau^{1:K-1})\right) \cdot \mathbb{P}^{(1)}\left(r_{1:H}^K|x_{1:H}^K\right),$$

and similarly, for $\mathbb{P}_{\psi}^{(2)}$

$$\mathbb{P}_{\psi}^{(2)}(\tau^K|\tau^{1:K-1}) = \left(\Pi_{t=1}^{H} T(s_t^K|x_{t-1}^K)\psi(a_t^K|h_t^K, \tau^{1:K-1})\right) \cdot \mathbb{P}^{(2)}\left(r_{1:H}^K|x_{1:H}^K\right),$$

which implies

$$\begin{aligned}
\mathbb{E}_{\psi}^{(1)}\left[\log\left(\frac{\mathbb{P}_{\psi}^{(1)}(\tau^{K}|\tau^{1:K-1})}{\mathbb{P}_{\psi}^{(2)}(\tau^{K}|\tau^{1:K-1})}\right)\right] &= \mathbb{E}_{\psi}^{(1)}\left[\mathbb{E}_{\psi}^{(1)}\left[\sum_{x_{1:H}} \log\left(\frac{\mathbb{P}^{(1)}\left(r_{1:H}^K|x_{1:H}^K\right)}{\mathbb{P}^{(2)}\left(r_{1:H}^K|x_{1:H}^K\right)}\right) \mathbb{1}\left\{x_{1:H}^K = x_{1:H}\right\} \Big| \tau^{1:K-1}\right]\right] \\
&= \sum_{x_{1:H}} \mathbb{E}_{\psi}^{(1)}\left[\log\left(\frac{\mathbb{P}^{(1)}\left(r_{1:H}^K|x_{1:H}\right)}{\mathbb{P}^{(2)}\left(r_{1:H}^K|x_{1:H}\right)}\right) \mathbb{1}\left\{x_{1:H}^K = x_{1:H}\right\}\right] \\
&= \sum_{x_{1:H}} \mathrm{KL}\left(\mathbb{P}^{(1)}\left(\cdot|x_{1:H}\right), \mathbb{P}^{(2)}\left(\cdot|x_{1:H}\right)\right) \cdot \mathbb{E}_{\psi}^{(1)}\left[\mathbb{1}\left\{x_{1:H}^K = x_{1:H}\right\}\right],
\end{aligned}$$

where the second equality is an application of the tower rule. Applying this recursively in K, we can show that

$$\mathrm{KL}\left(\mathbb{P}_{\psi}^{(1)}(\tau^{1:K}), \mathbb{P}_{\psi}^{(2)}(\tau^{1:K})\right) = \sum_{x_{1:H}} \mathrm{KL}\left(\mathbb{P}^{(1)}\left(\cdot|x_{1:H}\right), \mathbb{P}^{(2)}\left(\cdot|x_{1:H}\right)\right) \cdot \mathbb{E}_{\psi}^{(1)}\left[\sum_{k=1}^{K} \mathbb{1}\left\{x_{1:H}^k = x_{1:H}\right\}\right].$$

By definition of $N_{\psi,x_{1:H}}(K)$, we have

$$\mathbb{E}_{\psi}^{(1)}\left[\sum_{k=1}^{K} \mathbb{1}\left\{x_{1:H}^k = x_{1:H}\right\}\right] = N_{\psi,x_{1:H}}(K).$$

Plugging the above, we get the desired result.

D.4. Proof of Theorem 5.4

Let $\{\mu_m^*\}_{m=1}^M$ be the specific set of vectors in $\mathbb{R}^d$ satisfying Lemma 5.1 with $d = \Omega(\sqrt{M}) \geq 5$ being an odd number satisfying the condition in Lemma 5.1. Suppose the transition model follows the construction in Section 5: at every time step $t \in [H]$, we deterministicially move to a unique state s_t^*, and the reward values are binary, *i.e.,* $\mathcal{Z} = \{0, 1\}$. At every state $s_t = s_t^*$ (or time step t), all actions except one *correct* action $a_t^* \in \mathcal{A}$ returns a reward sampled from a uniform distribution over $\{0, 1\}$. The correct actions a_t^* can be any action in $\mathcal{A}$.

Consider two base systems $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$: in $\mathcal{M}^{(1)}$, reward distributions from all state-actions are uniform over $\{0, 1\}$. In $\mathcal{M}^{(2)}$, $\mu_m(s_t^*, a_t^*) = \mu_m^*(t)$, and otherwise uniform over $\{0, 1\}$ similarly. As we can see in Lemma 5.2, the optimal expected cumulative reward in $\mathcal{M}^{(1)}$ is 1/2, whereas in $\mathcal{M}^{(2)}$ optimal value is greater than $1/2 + \epsilon \cdot 2^{d-2}$. Suppose there exists a PAC-algorithm ψ such that for any RMMDP instances, ψ can output an ϵ-optimal policy π_ϵ after K episodes with probability greater than $2/3$. Then, we can use ψ to test whether the system is $\mathcal{M}^{(1)}$ or $\mathcal{M}^{(2)}$, for any chosen optimal actions with probability greater than $2/3$.

However, note that for any state-action sequence $x_{1:H} \neq x^*_{1:H}$,

$$\mathrm{KL}\left(\mathbb{P}^{(1)}(\cdot|x_{1:d}), \mathbb{P}^{(2)}(\cdot|x_{1:d})\right) = 0,$$

and

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{P}^{(1)}(\cdot|x^*_{1:d}), \mathbb{P}^{(2)}(\cdot|x^*_{1:d})\right) &= \sum_{r_{1:d}} \mathbb{P}^{(1)}(r_{1:d}|x^*_{1:d}) \cdot \log\left(\frac{\mathbb{P}^{(1)}(r_{1:d}|x^*_{1:d})}{\mathbb{P}^{(2)}(r_{1:d}|x^*_{1:d})}\right) \\
&= \left(\frac{1}{2}\right)^d \cdot \sum_{r_{1:d}} \log\left(\frac{\mathbb{P}^{(1)}(r_d|x^*_{1:d}, r_{1:d-1})}{\mathbb{P}^{(2)}(r_d|x^*_{1:d}, r_{1:d-1})}\right) \\
&= \left(\frac{1}{2}\right)^{d+1} \cdot \sum_{r_{1:d}} \log\left(\frac{1/2}{1/2+\epsilon_0}\right) + \log\left(\frac{1/2}{1/2-\epsilon_0}\right) \\
&= \left(\frac{1}{2}\right)^{d+1} \cdot \sum_{r_{1:d}} O(\epsilon_0^2) = O(\epsilon_0^2),
\end{aligned}$$

where $\epsilon_0 = \epsilon \cdot 2^{d-1}$ due to (23). Let ψ' be an augmented exploration strategy that first runs ψ for K episodes and run the returned policy for $O(1/\epsilon_0^2)$ extra episodes. Let $K' = K + O(1/\epsilon_0^2)$ be the total number of episodes. We can apply Lemma 5.3 to obtain that after running an algorithm ψ' for K' episodes in both systems, we get

$$\mathbb{E}^{(1)}\left[N_{\psi', x^*_{1:H}}(K')\right] \cdot O(\epsilon_0^2) = \mathrm{KL}\left(\mathbb{P}^{(1)}_{\psi'}(\tau^{1:K'}), \mathbb{P}^{(2)}_{\psi'}(\tau^{1:K'})\right).$$

By Pinsker's inequality, it holds that

$$\mathrm{TV}\left(\mathbb{P}^{(1)}_{\psi'}(\tau^{1:K'}), \mathbb{P}^{(2)}_{\psi'}(\tau^{1:K'})\right) \le \frac{1}{2}\sqrt{\mathrm{KL}\left(\mathbb{P}^{(1)}_{\psi'}(\tau^{1:K'}), \mathbb{P}^{(2)}_{\psi'}(\tau^{1:K'})\right)}.$$

Note that since everything is symmetric in system $\mathcal{M}^{(1)}$, there exists at least one $x^*_{1:H}$ such that the expected number of the sequence being executed is small:

$$N_{\psi', x^*_{1:H}}(K') \le A^{-d} \cdot K'.$$

Therefore, due to LeCam's two point method (LeCam, 1973), K' must satisfy that

$$A^{-d} \cdot K' \cdot O(\epsilon_0^2) = \Omega(1).$$

This implies that $K' \ge \Omega(A^d/\epsilon_0^2) - O(1/\epsilon_0^2) = \Omega(A^d/\epsilon_0^2)$.

Using the action amplification argument in Kwon et al. (2021b) (see their lower bound construction in their Appendix), we can effectively construct the system with $O(SA/(H \log_A S))$-actions (in this system, each action selection happens through $O(\log_A S)$-steps). As long as $S = O(\mathrm{poly}(A))$, this gives a lower bound $\Omega\left(\left(\frac{SA}{d}\right)^d \cdot \frac{1}{\epsilon^2}\right)$. As we can take $d = \Omega(\sqrt{M})$ from Lemma 5.1, we are done.