Proceedings of Machine Learning Research vol 247:1–73, 2024 37th Annual Conference on Learning Theory

Near-Optimal Learning and Planning in Separated Latent MDPs

Fan Chen FANCHEN@MIT.EDU
Constantinos Daskalakis COSTIS@CSAIL.MIT.EDU
Noah Golowich NZG@MIT.EDU
Alexander Rakhlin RAKHLIN@MIT.EDU
Massachusetts Institute of Technology

Editors: Shipra Agrawal and Aaron Roth

Abstract

We study computational and statistical aspects of learning Latent Markov Decision Processes (LMDPs). In this model, the learner interacts with an MDP drawn at the beginning of each epoch from an unknown mixture of MDPs. To sidestep known impossibility results, we consider several notions of δ-separation of the constituent MDPs. The main thrust of this paper is in establishing a nearly-sharp *statistical threshold* for the horizon length necessary for efficient learning. On the computational side, we show that under a weaker assumption of separability under the optimal policy, there is a quasi-polynomial algorithm with time complexity scaling in terms of the statistical threshold. We further show a near-matching time complexity lower bound under the exponential time hypothesis.

Keywords: Partially observable reinforcement learning

1. Introduction

Reinforcement Learning (Kaelbling et al., 1996; Sutton and Barto, 2018) captures the common challenge of learning a good policy for an agent taking a sequence of actions in an unknown, dynamic environment, whose state transitions and reward emissions are influenced by the actions taken by the agent. Reinforcement learning has recently contributed to several headline results in Deep Learning, including Atari (Mnih et al., 2013), Go (Silver et al., 2016), and the development of Large Language Models (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022). This practical success has also sparked a burst of recent work on expanding its algorithmic, statistical and learning-theoretic foundations, towards bridging the gap between theoretical understanding and practical success.

In general, the agent might not fully observe the state of the environment, instead having imperfect observations of its state. Such a setting is captured by the general framework of Partially Observable Markov Decision Processes (POMDPs) (Smallwood and Sondik, 1973). In contrast to the fully-observable special case of Markov Decision Processes (MDPs) (Bellman, 1957), the setting of POMDPs is rife with statistical and computational barriers. In particular, there are exponential sample lower bounds for learning an approximately optimal policy (Krishnamurthy et al., 2016; Jin et al., 2020), and it is PSPACE-hard to compute an approximately optimal policy even when the transition dynamics and reward function are known to the agent (Papadimitriou and Tsitsiklis, 1987; Littman, 1994; Burago et al., 1996; Lusena et al., 2001). In view of these intractability results, a fruitful research avenue has been to identify conditions under which statistical and/or computational tractability can be resurrected. This is the avenue taken in this paper.

In particular, we study *Latent Markov Decision Processes (LMDPs)*, a learning setting wherein, as its name suggests, prior to the agent's interaction with the environment over an episode of H steps, nature samples an MDP, i.e. the state transition dynamics and the reward function, from a distribution ρ over MDPs, which share the same state and action sets. The learner can fully observe the state, but cannot observe which MDP was sampled, and she also does not know the distribution ρ. However, she can interact with the environment over several episodes for which, at the beginning of each episode, a fresh MDP is independently sampled from ρ. The learner's goal is to learn a policy that optimizes her reward in expectation when this policy is used on a random MDP sampled from ρ.

LMDPs are a special case of (overcomplete) POMDPs,[1] which capture many natural scenarios. For example, learning in an LMDP can model the task facing a robot that is moving around in a city but has no sensors to observe the weather conditions each day, which affect the pavement conditions and therefore the dynamics. Other examples include optimizing the experience of users drawn from some population in a web platform (Hallak et al., 2015), optimizing the outcomes of patients drawn from some population in healthcare provision (Steimle et al., 2021), and developing an optimal strategy against a population of possible opponents in a dynamic strategic interaction (Wurman et al., 2022). More broadly, LMDPs and the challenge of learning in LMDPs have been studied in a variety of settings under various names, including hidden-model MDPs (Chades et al., 2012), multi-task RL (Brunskill and Li, 2013; Liu et al., 2016), contextual MDPs (Hallak et al., 2015), hidden-parameter MDPs (Doshi-Velez and Konidaris, 2016), concurrent MDPs (Buchholz and Scheftelowitsch, 2019), multi-model MDPs (Steimle et al., 2021), and latent MDPs (Kwon et al., 2021b; Zhan et al., 2022; Chen et al., 2022a; Zhou et al., 2023).

Despite this work, we lack a complete understanding of what conditions enable computationally and/or sample efficient learning of optimal policies in LMDPs. We *do* know that some conditions must be placed, as in general, the problem is both computationally and statistically intractable. Indeed, it is known that an exponential number of episodes in the size L of the support of ρ, is necessary to learn an approximately optimal policy (Kwon et al., 2021b), and even when the LMDP is known, computing an optimal policy is PSPACE-hard (Steimle et al., 2021).

A commonly studied and intuitively simpler setting, which is a main focus of this paper, is that of *δ-strongly separated LMDPs*, where every pair of MDPs in the support of ρ are δ-separated in the sense that for every state-action pair their transition distributions differ by at least δ in total variation distance. Even in this setting, however, we lack a sharp characterization of the horizon length that is necessary and sufficient for sample-efficient learning. Previous works either require a very long horizon[2] (i.e. $H \gg SA$, Brunskill and Li (2013); Hallak et al. (2015); Liu et al. (2016)) or

1. Indeed, if $\mathcal{S}$ is the state space shared by all MDPs in the support $\mathcal{M}$ of the distribution ρ over MDPs, we may view this LMDP as a POMDP with state space $\mathcal{S} \times \mathcal{M}$. The state transition dynamics of this POMDP only allow transitions from state (s, m) to state (s', m') when $m = m'$, and the transition probability from (s, m) to (s', m) on action a is determined by the transition probability from s to s' on action a in MDP m. The observation model of this POMPD drops m when observing the state (s, m), and the initial state (s_0, m) is sampled by first sampling $m \sim \rho$, and then sampling s_0 from the initialization distribution of MDP m.
2. Even under such a long horizon, Brunskill and Li (2013); Hallak et al. (2015); Liu et al. (2016) have to require additional restrictive assumptions, e.g. the diameter of each MDP instance is bounded.

impose extra assumptions on the predictive state representation of the underlying LMDP (Kwon et al., 2021b).Other simplifying assumptions that have been studied include hindsight observability, i.e. observability of the index of the sampled MDP at the end of each episode, under which near-optimal regret guarantees have been obtained in certain parameter regimes (Kwon et al., 2021b; Zhou et al., 2023), as well as test-sufficiency (Zhan et al., 2022; Chen et al., 2022a) and decodability (Efroni et al., 2022), but here the known sample complexity bounds scale exponentially with the test-sufficiency/decodability window.

Our Contributions. In this paper, we nearly settle the challenge of learning in δ-strongly separated LMDPs, by providing a near-sharp characterization of the horizon length necessary for efficient learnability.

Our lower bound (Theorem 3.1) shows that, for there to be an algorithm that learns an ε-optimal policy in a δ-strongly separated LMDP from a polynomial number of samples, it must be that the horizon scales as

$$H \gtrsim \frac{\log(L/\varepsilon)}{\delta^2},$$

where L is the number of MDPs in the mixture. The threshold $H_\star \asymp \frac{\log(L/\varepsilon)}{\delta^2}$ has a fairly intuitive interpretation: when $H \geqslant H_\star$, we can use the history up to step $H_\star$ to recover the unobservable index of the underlying MDP instance with error probability at most ε (Proposition 4.1).

We complement our lower bound by proposing a sample-efficient algorithm (Algorithm 1) for learning an ε-optimal policy in a δ-strongly separated LMDP when

$$H \gtrsim \frac{\log(LS/\varepsilon\delta)}{\delta^2}.$$

Our sample complexity guarantees also hold beyond the strong separation condition. We study the setting where the MDP instances are separated under *every policy* (Section 4), a condition that is comparably less restrictive than the strong separation condition. We relax this separation assumption even further to separation under an *optimal policy*, although we need to make some extra assumptions in this case to preserve sample-efficiency (Section 4.1).

As a further application, we consider learning N-step decodable LMDPs, which is a natural class of structured LMDPs where strong separation does not hold. For such a class of LMDPs, we provide a sample-efficiency guarantee when $H \geqslant 2N$, and we also provide a lower bound which shows that this threshold is *sharp*.

Finally, we study the computational complexity of computing an optimal policy in a known separated LMDP, i.e. the problem of planning. We show that the threshold $H_\star$ tightly captures the time complexity of planning: it gives rise to a natural planning algorithm (Algorithm 2) with near-optimal time complexity under the exponential time hypothesis (ETH).

1.1. Related works

Planning in partially observable environment. Planning in a known POMDP has long been known to be PSPACE-compete (Papadimitriou and Tsitsiklis, 1987; Littman, 1994; Burago et al., 1996; Lusena et al., 2001), and planning in LMDP inherits such hardness (Chades et al., 2012;

Steimle et al., 2021). The recent work of Golowich et al. (2022b,a) established a property called "belief contraction" in POMDPs under an observability condition (Even-Dar et al., 2007), which leads to algorithms with quasi-polynomial statistical and computational efficiency.

Learning in partially observable environment. It is well-known that learning a near-optimal policy in an unknown POMDP is statistically hard in the worst-case: in particular, the sample complexity must scale at least exponentially in the horizon (Liu et al., 2022a; Krishnamurthy et al., 2016). Algorithms achieving such upper bounds are developed in (Kearns et al., 1999; Even-Dar et al., 2005). Under strong assumptions, such as full-rankness of the transition and observation matrices or availability of exploratory data, several algorithms based on spectral methods (Hsu et al., 2012; Azizzadenesheli et al., 2016; Guo et al., 2016; Xiong et al., 2021) and posterior sampling (Jahromi et al., 2022) have also been proven to be sample-efficient. However, due to the nature of their strong assumptions, these works fall short of addressing the challenge of exploration in an unknown partially observable environment.

Towards addressing this challenge, a line of recent works proposed various structural problem classes that can be learned sample-efficiently, including reactive POMDPs (Jiang et al., 2017), revealing POMDPs (Jin et al., 2020; Liu et al., 2022a,c), low-rank POMDPs with invertible emission operators (Cai et al., 2022; Wang et al., 2022), decodable POMDPs (Efroni et al., 2022), regular PSRs (Zhan et al., 2022), reward-mixing MDPs (Kwon et al., 2021a, 2023), PO-bilinear classes (Uehara et al., 2022b), POMDPs with deterministic latent transition (Uehara et al., 2022a), and POMDPs with hindsight observability (Lee et al., 2023). Based on the formulation of predictive state representation (PSR), Chen et al. (2022a); Liu et al. (2022b) proposed (similar) unified structural conditions which encompass most of these conditions, with a unified sample-efficient algorithm Optimistic Maximum Likelihood Estimation (OMLE). As LMDPs are a subclass of POMDPs, all of these results can be applied to LMDPs to provide structural conditions that enable learnability. However, when instantiated to LMDPs, these structural conditions are less intuitive, and in general they are incomparable to our separability assumptions and do not capture the full generality of the latter.

RL with function approximation. RL with general function approximation in fully observable environment has been extensively investigated in a recent line of work (Jiang et al., 2017; Sun et al., 2019; Du et al., 2021; Jin et al., 2021; Foster et al., 2021; Agarwal and Zhang, 2022; Chen et al., 2022b; Xie et al., 2022; Liu et al., 2023, etc.), and some of the proposed complexity measures and algorithms (e.g. Model-based Optimistic Posterior Sampling (Agarwal and Zhang, 2022; Chen et al., 2022b), and Estimation-to-Decision (Foster et al., 2021)) also apply to partially observable RL. In this work, our analysis of OMLE utilizes several tools developed in Liu et al. (2022a); Chen et al. (2022b,a); Xie et al. (2022).

2. Preliminaries

Latent Markov Decision Process. An LMDP M is specified by a tuple $\left\{\mathcal{S}, \mathcal{A}, (M_m)_{m=1}^{L}, H, \rho, R\right\}$, where $M_1, \cdots, M_L$ are L MDP instances with joint state space $\mathcal{S}$, joint action space $\mathcal{A}$, horizon H, and $\rho \in \Delta([L])$ is the mixing distribution over $M_1, \cdots, M_L$, and $R = (R_h : \mathcal{S} \times \mathcal{A} \to [0,1])_{h=1}^{H}$ is the reward function. For $m \in [L]$, the MDP M_m is specified by $\mathbb{T}_m : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ along with the initial state distribution $\nu_m \in \Delta(\mathcal{S})$. In what follows, we will parametrize each LMDP by a parameter θ (Section 2.2), but for now we provide a few definitions without overburdening the notation.

In an LMDP, the latent index of the current MDP is hidden from the agent: the agent can only see the resulting transition trajectory. Formally speaking, at the start of each episode, the environment randomly draws a latent index $m^\star \sim \rho$ (which is unobservable) and an initial state $s_1 \sim \nu_{m^\star}$, and then at each step h, after the agent takes action a_h, the environment generates the next state $s_{h+1} \sim \mathbb{T}_{m^\star}(\cdot|s_h, a_h)$ following the dynamics of MDP $M_{m^\star}$. The episode terminates immediately after a_H is taken.

Policies. A policy $\pi = \{\pi_h : (\mathcal{S} \times \mathcal{A})^{h-1} \times \mathcal{S} \to \Delta(\mathcal{A})\}_{h\in[H]}$ is a collection of H functions. At step $h \in [H]$, an agent running policy π observes the current state s_h and takes action $a_h \sim \pi_h(\cdot|\overline{\tau}_h) \in \Delta(\mathcal{A})$ based on the whole history $\overline{\tau}_h = (\tau_{h-1}, s_h) = (s_1, a_1, \ldots, s_{h-1}, a_{h-1}, s_h)$. (In particular, we have written $\tau_{h-1} = (s_1, a_1, \ldots, s_{h-1}, a_{h-1})$.) The policy class Π_{RND} is the set of all such history-dependent policies, and Π_{DM} is the set of all deterministic Markov policies, namely tuples $\pi = \{\pi_h : \mathcal{S} \to \mathcal{A}\}_{h\in[H]}$.

For any policy $\pi \in \Pi_{\mathrm{RND}}$, the interaction between π and the LMDP M induces a distribution $\mathbb{P}^\pi$ of the whole trajectory $\tau_H = (s_1, a_1, \cdots, s_H, a_H)$. The value of π is defined as

$$V(\pi) = \mathbb{E}^\pi\left[\sum_{h=1}^{H} R_h(s_h, a_h)\right].$$

We also use $\widetilde{\mathbb{P}}^\pi$ to denote the joint probability distribution of the latent index $m^\star$ and trajectory τ_H under policy π.

Miscellaneous notations For probability distributions p, q on a discrete measure space $\mathcal{X}$, the Hellinger distance and Bhattacharyya divergence are defined as

$$D_{\mathrm{H}}^2(p,q) := \tfrac{1}{2}\sum_{x\in\mathcal{X}}(\sqrt{p(x)} - \sqrt{q(x)})^2, \qquad D_{\mathrm{B}}(p,q) = -\log \sum_{x\in\mathcal{X}} \sqrt{p(x)q(x)}.$$

For expression f, g, we write $f \lesssim g$ if there is an absolute constant C such that $f \leqslant Cg$. We also use $f = \mathcal{O}(g)$ to signify the same thing.

2.1. Strong separation and separation under policies

In this section we introduce the various notions of separability we consider in this paper.

Definition 2.1 (Strong separation, Kwon et al. (2021b)) *An LMDP is δ-strongly separated if for all $m, l \in \mathrm{supp}(\rho)$ such that $m \neq l$,*

$$D_{\mathrm{TV}}\left(\mathbb{T}_m(\cdot|s,a), \mathbb{T}_l(\cdot|s,a)\right) \geqslant \delta, \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}.$$

Definition 2.2 (Decodability, Efroni et al. (2022)) *An LMDP M is N-step decodable if for any trajectory $\overline{\tau}_N = (s_1, a_1, \cdots, s_N)$, there is at most one latent index $m \in \mathrm{supp}(\rho)$ such that $\overline{\tau}_N$ is reachable starting from s_1 in the MDP instance M_m (i.e., the probability of observing $s_2, \cdots, s_N$ in M_m starting at s_1 and taking actions $a_1, \cdots, a_{N-1}$ is non-zero). In other words, there exists a* decoding function *ϕ_M that maps any reachable trajectory $\overline{\tau}_N$ to the latent index m.*

More generally, we can consider separability under the induced distributions over a trajectory. For any policy π, we define

$$\mathbb{M}_{m,h}(\pi, s) := [\mathbb{T}_m^\pi((a_1, s_2, \cdots, a_{h-1}, s_h) = \cdot | s_1 = s)] \in \Delta((\mathcal{A} \times \mathcal{S})^{h-1}), \tag{1}$$

where $\mathbb{T}_m^\pi$ is the probability distribution of the trajectory in the MDP instance M_m and under policy π.

For any increasing function $\varpi : \mathbb{N} \to \mathbb{R}$, we can define ϖ-separation as follows, which requires that the separation between any two MDP instances grow as ϖ.

Definition 2.3 (Separation with respect to a policy) *An LMDP is ϖ-separated under π if for all $m, l \in \mathrm{supp}(\rho)$ such that $m \neq l$,*

$$D_{\mathrm{B}}(\mathbb{M}_{m,h}(\pi, s), \mathbb{M}_{l,h}(\pi, s)) \geqslant \varpi(h), \qquad \forall h \geqslant 1, s \in \mathcal{S}.$$

We also define $\varpi^{-1}(x) := \min\{h \geqslant 1 : \varpi(h) \geqslant x\}$. In Section 4, we show that if the LMDP is ϖ-separated under *all policies* and $H \gtrsim \varpi^{-1}(\log(\text{problem parameters}))$, then a near-optimal policy can be learned sample-efficiently.

In particular, strong separation indeed implies separation under all policies.

Proposition 2.4 *If the LMDP M is δ-strongly separated, then it is ϖ_δ-separated under any policy $\pi \in \Pi_{\mathrm{RND}}$, where $\varpi_\delta(h) = \frac{\delta^2}{2}(h-1)$.*

Proposition 2.5 *The LMDP M is N-step decodable if and only if it is ϖ_N-separated under all policy $\pi \in \Pi_{\mathrm{RND}}$, where $\varpi_N(h) = \begin{cases} 0, & h < N, \\ \infty, & h \geqslant N. \end{cases}$*

The proof of Proposition 2.5 is provided in Appendix C.1. More generally, the following lemma gives a simple criteria for all-policy separation.

Lemma 2.6 *If an LMDP is ϖ-separated under any Markov policy $\pi \in \Pi_{\mathrm{DM}}$, then it is ϖ-separated under any general policy $\pi \in \Pi_{\mathrm{RND}}$.*

2.2. Model-based function approximation

In this paper, we consider the standard model-based learning setting, where we are given an LMDP model class Θ and a policy class $\Pi \subseteq \Pi_{\mathrm{RND}}$. Each $\theta \in \Theta$ parameterizes an LMDP $M_\theta = \{\mathcal{S}, \mathcal{A}, (M_{\theta,m})_{m=1}^L, H, \rho_\theta, R\}$, where the state space $\mathcal{S}$, action space $\mathcal{A}$, horizon H, integer L representing the number of MDPs, and reward function R are shared across all models, ρ_θ specifies the mixing weights for the L MDP instances under θ, and the MDP instance $M_{\theta,m}$ is specified by $(\mathbb{T}_{\theta,m}, \nu_{\theta,m})$ for each $m \in [L]$. For each model $\theta \in \Theta$ and policy $\pi \in \Pi_{\mathrm{RND}}$, we denote $\mathbb{P}_\theta^\pi$ to be the distribution of τ_H in M_θ under policy π, and let $V_\theta(\pi)$ be the value of π under M_θ.

We further assume that (a) the ground truth LMDP is parameterized by a model $\theta^\star \in \Theta$ (*realizability*); (b) the model class Θ admits a bounded log covering number $\log N_\Theta(\cdot)$ (Definition A.1); (c) the reward function R is known and bounded, $\sum_{h=1}^H \sup_{s,a} R_h(s,a) \leqslant 1$. [3]

3. For simplicity, we only consider deterministic known reward in this paper. For random reward $r_h \in \{0,1\}$ that possibly depends on the latent index m, we can consider the "augmented" LMDP with the augmented state $\tilde{s}_{h+1} = (s_{h+1}, r_h)$ similar to Kwon et al. (2021b).

In addition to the assumptions stated above, we also introduce the following assumption that the ground truth LMDP admits certain low-rank structure, which is a common assumption for sample-efficient partially observable RL (Wang et al., 2022; Chen et al., 2022a; Liu et al., 2022b).

Assumption 2.7 (Rank) *The rank of an LMDP M_θ is defined as $d_\theta := \max_{m\in[L]} \mathrm{rank}(\mathbb{T}_{\theta,m})$. We assume that the ground truth model $\theta^\star$ has rank $d < \infty$.*

Learning goal. The learner's goal is to output an ε-optimal policy $\widehat{\pi}$, i.e. a policy with sub-optimality $V_\star - V_{\theta^\star}(\widehat{\pi}) \leqslant \varepsilon$, where $V_\star = \max_{\pi\in\Pi} V_{\theta^\star}(\pi)$ is the optimal value of the ground truth LMDP.

3. Intractability of separated LMDP with horizon below threshold

Given the exponential hardness of learning general LMDPs, Kwon et al. (2021b) explore several structural conditions under which a near-optimal policy can be learned sample-efficiently. The core assumptions there include a strong separation condition (Definition 2.1) together with the bound

$$H \geqslant \delta^{-4} \log^2(S/\delta) \log(LSA\varepsilon^{-1}\delta^{-1}). \tag{2}$$

A natural question is whether such an assumption on the horizon is necessary. The main result of this section demonstrates the necessity of a moderately long horizon, i.e. in order to learn a δ-strongly separated LMDP in polynomial samples, it is necessary to have a horizon length that (asymptotically) exceeds $\frac{\log(L/\varepsilon)}{\delta^2}$.

Theorem 3.1 (Corollary of Theorems D.1 and D.2) *Suppose that there exists an integer $\mathsf{d} \geqslant 1$ and an algorithm $\mathfrak{A}$ with sample complexity $\max\{S, A, H, L, \varepsilon^{-1}, \delta^{-1}\}^{\mathsf{d}}$ that learns an ε-optimal policy with probability at least $3/4$ in any δ-strongly separated LMDP with $H \geqslant H_{\rm thre}(L, \varepsilon, \delta)$, for some function $H_{\rm thre}(L, \varepsilon, \delta)$. Then there exists constants $c_{\mathsf{d}}, \varepsilon_{\mathsf{d}}, L_{\mathsf{d}}$ (depending on d) and an absolute constant δ_0 such that*

$$H_{\rm thre}(L, \varepsilon, \delta) \geqslant \frac{c_{\mathsf{d}} \log(L/\varepsilon)}{\delta^2}, \qquad \forall \delta \leqslant \delta_0, \varepsilon \leqslant \varepsilon_{\mathsf{d}}, L \geqslant \max(L_{\mathsf{d}}, \delta^{-1}).$$

The proof of Theorem 3.1 is presented in Appendix D, where we also provide a more precise characterization of the sample complexity lower bounds in terms of H (Theorems D.1 and D.2). The lower bound of the threshold $H_{\rm thre}$ is nearly optimal, in the sense that it almost matches the learnable range (as per Corollary 4.4 below).

The following theorem provides a simpler lower bound for horizon length $H = \tilde{\Theta}\big(\delta^{-1}\log L\big)$. For such a short horizon, we show that we can recover the exponential lower bound developed in Kwon et al. (2021b) for learning non-separated LMDPs.

Theorem 3.2 *Suppose that $\delta \in (0, \frac{1}{4e^2}]$, $H \geqslant 3$, $A \geqslant 2$, $L \geqslant 2^{C\log^2(1/\delta)}$ are given such that*

$$CH \log H \log(1/\delta) \leqslant \frac{\log L}{\delta}. \tag{3}$$

Then there exists a class of δ-strongly separated LMDPs, each LMDP has L MDP instances, $S = (\log L)^{\mathcal{O}(\log H)}$ states, A actions, and horizon H, such that any algorithm requires $\Omega\left(A^{H-2}\right)$ samples to learn an $\frac{1}{4H}$-optimal policy with probability at least $\frac{3}{4}$.

Proof idea for Theorem 3.1. Theorem 3.1 is proved by transforming the known hard instances of general LMDPs (Appendix D.1) to hard instances of δ-strong separated LMDPs. In particular, given a LMDP M, we transform it to a δ-strongly separated LMDP M', so that each MDP instance M_m of M is transformed to a mixture of MDPs $\{M_{m,j}\}$, where each $M_{m,j} = M_i \otimes \mu_{m,j}$ is an MDP obtained by augmenting M_i with a distribution $\mu_{m,j}$ of the auxiliary observation (this operation $\otimes$ is formally defined in Definition D.6). The δ-strongly separated property of M' is ensured as long as $D_{\mathrm{TV}}\left(\mu_{m,j}, \mu_{m',j'}\right) \geqslant \delta$ for different pairs of $(m,j) \neq (m',j')$, and intuitively, M' is still a hard instance if the auxiliary observation does not reveal much information of the latent index.

Such a transformation is possible as long as $H = \frac{o(\log L)}{\delta^2}$. Here, we briefly illustrate how the transformation works for LMDP M consisted of only 2 MDP instances M_1, M_2. Using Proposition 3.3, we define the augmented MDPs $M_{1,j} = M_1 \otimes \mu_j$ for $j \in \mathrm{supp}(\nu_1)$ and $M_{2,j} = M_2 \otimes \mu_j$ for $j \in \mathrm{supp}(\nu_2)$, and assigning the mixing weights based on ν_1, ν_2. Then, result (1) ensures the transformed LMDP is δ-strongly separated, and result (2) ensures the auxiliary observation does not reveal much information of the latent index. The details of our transformation for general LMDPs is presented in Appendix D.2.

Proposition 3.3 (Simplified version of Proposition D.8) *Suppose that parameter $\delta, c > 0$ and integer $n \geqslant 2$ satisfy $Cn\log^2 n \leqslant \min\left\{c^{-1}, \delta^{-1}\right\}$. Then for $L \geqslant n^2$, $H \leqslant \frac{c\log L}{\delta^2}$, there exists $L' \leqslant L$ distributions $\mu_1, \cdots, \mu_{L'}$ over a set $\mathcal{O}$ satisfying $|\mathcal{O}| \leqslant O(\log L)$, such that:*

(1) $D_{\mathrm{TV}}\left(\mu_i, \mu_j\right) \geqslant \delta$ for $i \neq j$.

(2) There exists $\nu_1, \nu_2 \in \Delta([L'])$ such that $\mathrm{supp}(\nu_1)$ and $\mathrm{supp}(\nu_2)$ are disjoint, and

$$D_{\mathrm{TV}}\left(\mathbb{E}_{i\sim\nu_1}\mu_i^{\otimes H}, \mathbb{E}_{j\sim\nu_2}\mu_j^{\otimes H}\right) \leqslant L^{-n},$$

where for any distribution μ, $\mu^{\otimes H}$ is the distribution of $(o_1, \cdots, o_H)$ where $o_h \sim \mu$ independently.

Tighter threshold for decodable LMDPs For δ-strongly separated LMDP, Theorem 3.1 gives a lower bound of H_{thre} that scales as $\frac{\log(L/\varepsilon)}{\delta^2}$ and nearly matches the upper bounds (Corollary 4.4). The following result shows that, for N-step decodable LMDPs, we can identify the even tighter threshold of H: when $H \leqslant 2N - \omega(1)$, there is no sample-efficient algorithm; by contrast, when $H \geqslant 2N$, OMLE is sample-efficient (Corollary 4.5).

Theorem 3.4 *Suppose that integers $N \geqslant n \geqslant 2$, $A \geqslant 2$ are given. Then for $H = 2N - n$, there exists a class of N-step decodable LMDPs with $L = n$, $S = 3N - 1$ states, A actions, and horizon H, such that any algorithm requires $\Omega\left(A^{n-1}\right)$ samples to learn an $\frac{1}{4n}$-optimal policy with probability at least $\frac{3}{4}$.*

4. Learning separated LMDPs with horizon above threshold

In this section, we show that δ-strongly separated LMDP, or more generally, any LMDP under suitable policy separation assumptions, can be learned sample-efficiently, as long as the horizon H exceeds a threshold that depends on the separation condition and the logarithm of other problem parameters.

Algorithm 1 OPTIMISTIC MAXIMUM LIKELIHOOD ESTIMATION (OMLE)

Input: Model class Θ, policy class Π, exploration strategy $\mathsf{p}(\cdot) : \Pi \to \Pi_{\mathrm{RND}}$, parameter $\beta > 0, \varepsilon_{\mathrm{s}} \in (0,1], W \geqslant 1$.

Initialize: $\Theta^1 = \Theta, \mathcal{D} = \{\}$.

for $k = 1, \ldots, K$ **do**

Set `// See (5) for definition of` $e_{\theta,W}$.

$$(\theta^k, \pi^k) = \arg\max_{(\theta,\pi)} V_\theta(\pi), \qquad \text{s.t. } \theta \in \Theta^k, e_{\theta,W}(\pi) \leqslant \varepsilon_{\mathrm{s}}.$$

Execute $\pi^k_{\mathrm{sep}} = \mathsf{p}(\pi^k)$ to collect a trajectory τ^k_H, and add $(\pi^k_{\mathrm{sep}}, \tau^k_H)$ into $\mathcal{D}$.

Update confidence set

$$\Theta^{k+1} = \left\{ \widehat{\theta} \in \Theta : \textstyle\sum_{(\pi,\tau)\in\mathcal{D}} \log \mathbb{P}^\pi_{\widehat{\theta}}(\tau) \geqslant \max_{\theta\in\Theta} \sum_{(\pi,\tau)\in\mathcal{D}} \log \mathbb{P}^\pi_\theta(\tau) - \beta \right\}.$$

end

Output: $\widehat{\pi} := \mathrm{Unif}(\{\pi^1, \cdots, \pi^K\})$.

A crucial observation is that if that an LMDP M_θ is ϖ-separated under policy π, then the agent can "decode" the latent index from the trajectory $\overline{\tau}_h$, with error probability decaying exponentially in $\varpi(h)$.

Proposition 4.1 *Given an LMDP M_θ and parameter $W \geqslant 1$, for any trajectory $\overline{\tau}_W = (s_1, a_1, \cdots, s_W)$, we consider the latent index with maximum likelihood under $\overline{\tau}_W$:*

$$m_\theta(\overline{\tau}_W) := \arg\max_{m\in\mathrm{supp}(\rho_\theta)} \log \rho_\theta(m) + \log \nu_{\theta,m}(s_1) + \sum_{h=1}^{W-1} \log \mathbb{T}_{\theta,m}(s_{h+1}|s_h, a_h). \tag{4}$$

Then as long as M_θ is ϖ-separated under π, the decoding error *can be bounded as*

$$e_{\theta,W}(\pi) := \widetilde{\mathbb{P}}^\pi_\theta(m_\theta(\overline{\tau}_W) \neq m^\star) \leqslant L \exp(-\varpi(W)), \tag{5}$$

where we recall that $\widetilde{\mathbb{P}}^\pi_\theta$ is the joint probability distribution of the latent index $m^\star$ and trajectory τ_H in the LMDP M_θ under policy π.

The OMLE algorithm was originally proposed by Liu et al. (2022a) for learning revealing POMDPs, and it was later adapted for a broad class of model-based RL problems (Zhan et al., 2022; Chen et al., 2022b,a; Liu et al., 2023). Based on the observation above, we propose a variant of the OMLE algorithm for learning separated LMDPs.

Algorithm. On a given class Θ of LMDPs, the OMLE algorithm (Algorithm 1) iteratively performs the following steps while building up a dataset $\mathcal{D}$ consisting of trajectories drawn from the unknown LMDP:

1. (Optimism) Construct a confidence set $\Theta^k \subseteq \Theta$ based on the log-likelihood of all trajectories within dataset $\mathcal{D}$. The optimistic (model, policy) pair (θ^k, π^k) is then chosen greedily while ensuring that the decoding error $e_{\theta^k,W}(\pi^k)$ is small.

2. (Data collection) For an appropriate choice of *exploration strategy* $\mathsf{p}(\cdot)$ (described in Definition E.1), execute the explorative policy $\pi^k_{\mathrm{sep}} = \mathsf{p}(\pi^k)$, and then collect the trajectory into $\mathcal{D}$.

Guarantees. Under the following assumption on all-policy separation with a specific growth function ϖ, the OMLE algorithm can learn a near-optimal policy sample efficiently. In particular, when Θ is the class of all δ-strongly separated LMDPs, then Assumption 4.2 is fulfilled automatically with $\Pi = \Pi_{\mathrm{RND}}$ and $\varpi(h) = \frac{\delta^2}{2}(h-1)$ (Proposition 2.4).

Assumption 4.2 (Separation under all policies) *For any $\theta \in \Theta$ and any $\pi \in \Pi$, θ is ϖ-separated under π.*

Theorem 4.3 *Suppose that Assumption 2.7 and Assumption 4.2 hold. We fix any $\pi_{\mathrm{sep}} \in \Pi$, set $\mathsf{p}(\cdot)$ as in Definition E.1, and choose the parameters of Algorithm 1 so that*

$$\beta \geqslant 2\log N_\Theta(1/T) + 2\log(1/p) + 2, \qquad K = C_0\frac{Ld^2AH^2\iota\beta}{\varepsilon^2}, \qquad \varepsilon_{\mathrm{s}} = \frac{\varepsilon^2}{C_0Ld^2H^2\iota},$$

where $\iota = \log(LdH/\varepsilon)$ is a log factor, C_0 is a large absolute constant. Then, as long as W is suitably chosen so that

$$W \geqslant \varpi^{-1}(\log(L/\varepsilon_{\mathrm{s}})), \qquad H - W \geqslant \varpi^{-1}(\log(2L)), \tag{6}$$

Algorithm 1 outputs an ε-optimal policy $\widehat{\pi}$ with probability at least $1-p$ after observing K trajectories.

Note that the parameter W can always be found satisfying the conditions of Theorem 4.3 as long as $H \geqslant \varpi^{-1}(\log(2L)) + \varpi^{-1}(\log(L/\varepsilon_{\mathrm{s}}))$. In particular, OMLE is sample-efficient for learning δ-strongly separated LMDPs with a moderate requirement on the horizon H (which nearly matches the lower bound of Theorem 3.1).

Corollary 4.4 *Suppose that $|\mathcal{S}| = S$ and Θ is the class of all δ-strongly separated LMDPs. Then as long as*

$$H \geqslant \frac{10\log(LS\varepsilon^{-1}\delta^{-1}) + C}{\delta^2} \tag{7}$$

for some absolute constant C, we can suitably instantiate Algorithm 1 so that it outputs an ε-optimal policy $\widehat{\pi}$ with high probability using $K = \widetilde{\mathcal{O}}(\frac{L^2S^4A^2H^4}{\varepsilon^2})$ episodes.

Compared to the results of Kwon et al. (2021b), Corollary 4.4 requires neither a good initialization that is close to the ground truth model, nor does it require additional assumptions, e.g. test-sufficiency, which is also needed in Zhan et al. (2022); Chen et al. (2022a). Furthermore, Kwon et al. (2021b) also requires (2), while the range of tractable horizon (7) here is wider, and it nearly matches the threshold in Theorem 3.1. A more detailed discussion is deferred to Appendix B.

Furthermore, OMLE is also sample-efficient for learning N-step decodable LMDPs, as long as $H \geqslant 2N$.

Corollary 4.5 (Learning decodable LMDPs) *Suppose that Θ is a class of N-step decodable LMDPs with horizon length $H \geqslant 2N$. Then we can suitably instantiate Algorithm 1 so that it outputs an ε-optimal policy $\widehat{\pi}$ with high probability using $K = \widetilde{\mathcal{O}}(\frac{Ld^2AH^2\log N_\Theta}{\varepsilon^2})$ episodes.*

In Efroni et al. (2022), a sample complexity that scales with A^N is established for learning general N-step decodable POMDPs. By contrast, Corollary 4.5 demonstrates that for N-step decodable LMDPs, a horizon length of $H \geqslant 2N$ suffices to ensure polynomial learnability. As Theorem 3.4 indicates, requiring $H \geqslant 2N - \mathcal{O}(1)$ is also necessary for polynomial sample complexity, and hence the threshold $H \geqslant 2N$ is nearly *sharp* for N-step decodable LMDPs. This result also demonstrates that the condition (6) (and our two-phase analysis; see Appendix E.2) is generally necessary for Theorem 4.3.

4.1. Sample-efficient learning with two-policy separation

In general, requiring separation under all policies is a relatively restrictive assumption, because it is possible that the LMDP is well-behaved under only a small subset of policies that contains the optimal policy. In this section, we discuss the sample-efficiency of OMLE under the following assumption of separation under an optimal policy.

Assumption 4.6 (Separation under an optimal policy) *There exists an optimal policy $\pi_\star$ of the LMDP $M_{\theta^\star}$, such that $M_{\theta^\star}$ is ϖ-separated under $\pi_\star$.*

In order to obtain sample-efficiency guarantee, we also need the following technical assumption on a prior-known *separating policy* π_{sep}. Basically, we assume that in each LMDP, the MDP instances are sufficient "diverse" under π_{sep}, so that any mixture of them is qualitatively different from any MDP model.

Assumption 4.7 (Prior knowledge of a suitable policy for exploration) *There exists a known policy π_{sep} and parameters $(W_{\mathrm{exp}}, \alpha)$ such that for any model $\theta \in \Theta$, the following holds:*

(a) M_θ is ϖ-separated under π_{sep}.

(b) For any MDP model $\mathbb{T}_{\mathrm{ref}}$ and state $s \in \mathcal{S}$, it holds that for any $\lambda \in \Delta(\mathrm{supp}(\rho_\theta))$,

$$D_{\mathrm{TV}}\left(\mathbb{E}_{m\sim\lambda}\Big[\mathbb{M}^{\theta}_{m,W_{\mathrm{exp}}}(\pi_{\mathrm{sep}}, s)\Big], \mathbb{M}_{\mathrm{ref},W_{\mathrm{exp}}}(\pi_{\mathrm{sep}}, s)\right) \geqslant \alpha(1 - \max_{m} \lambda_m), \tag{8}$$

where

$$\mathbb{M}_{\mathrm{ref},h}(\pi_{\mathrm{sep}}, s) = \left[\mathbb{T}_{\mathrm{ref}}^{\pi_{\mathrm{sep}}}((a_1, s_2, \cdots, s_h) = \cdot | s_1 = s)\right] \in \Delta((\mathcal{S} \times \mathcal{A})^{h-1})$$

is the distribution of trajectory induced by running π_{sep} on the MDP with transition $\mathbb{T}_{\mathrm{ref}}$.

Theorem 4.8 *Suppose that Assumption 2.7, Assumption 4.6, and Assumption 4.7 hold. We set $\mathsf{p}(\cdot)$ based on π_{sep} as in Definition E.1, and choose the parameters of Algorithm 1 so that*

$$\beta \geqslant 2\log N_\Theta(1/T) + 2\log(1/p) + 2, \qquad K = C_0 \frac{L^3 d^5 A H^6 \iota^3 \beta}{\alpha^2 \varepsilon^4}, \qquad \varepsilon_{\mathrm{s}} = \frac{\alpha \varepsilon^2}{C_0 L d^2 H^2 \iota},$$

where $\iota = \log(LdH\alpha^{-1}\varepsilon^{-1})$ is a log factor, C_0 is a large absolute constant. Then, as long as W is suitably chosen so that

$$W \geqslant \varpi^{-1}(\log(L/\varepsilon_{\mathrm{s}})), \qquad H - W \geqslant W_{\mathrm{exp}},$$

Algorithm 1 outputs an ε-optimal policy $\widehat{\pi}$ with probability at least $1 - p$.

In Appendix E.7, we also provide a sufficient condition of Assumption 4.7, which is more intuitive.

5. Computation complexity of separated LMDPs

In this section, we investigate the computational complexity of planning in a *given* LMDP, i.e. a description of the ground truth model $\theta^\star$ is provided to the learner.[4] For planning, a longer horizon does not reduce the time complexity (in contrast to learning, where a longer horizon does help).

In general, we cannot expect a polynomial time planning algorithm for δ-strongly separated LMDP, because even the problem of computing an approximate optimal value in any given δ-strongly separated LMDP is NP-hard.

Proposition 5.1 *If there is an algorithm that computes the ε-approximate optimal value of any given δ-strongly separated LMDP in $\mathrm{poly}(L, S, A, H, \varepsilon^{-1}, \delta^{-1})$ time, then P=NP.*

On the other hand, utilizing the Proposition 4.1, we propose a simple planning algorithm (Algorithm 2) for any LMDP that is separated under its optimal policy. The algorithm design is inspired by the Short Memory Planning algorithm proposed by Golowich et al. (2022b).

Theorem 5.2 *Suppose that in the LMDP M, there exists an optimal policy $\pi_\star$ such that M is ϖ-separated under $\pi_\star$. Then Algorithm 2 with $W \geqslant \varpi^{-1}(\log(L/\varepsilon))$ outputs an ε-optimal policy $\widehat{\pi}$ in time*

$$(SA)^W \times \mathrm{poly}(S, A, H, L).$$

As a corollary, Algorithm 2 can output an ε-optimal policy (along with an ε-approximate optimal value) of any given δ-strongly separated LMDP in time

$$(SA)^{2\delta^{-2}\log(L/\varepsilon)} \times \mathrm{poly}(L, S, A, H).$$

In the following, we demonstrate such a time complexity is nearly optimal for planning in δ-strongly separated LMDP, under the Exponential Time Hypothesis (ETH):

Conjecture 5.3 (ETH, Impagliazzo and Paturi (2001)) *There is no $2^{o(n)}$-time algorithm which can determine whether a given 3SAT formula on n variables is satisfiable.*

In the following theorems, we provide quasi-polynomial time lower bounds for planning in δ-strongly separated LMDP, assuming ETH. In order to provide a more precise characterization of the time complexity lower bound in terms of all the parameters $(L, \varepsilon, \delta, A)$, we state our hardness results in with varying $(L, \varepsilon, \delta, A)$ pair, with mild assumptions of their growth. To this end, we consider $\mathcal{F} = \{(b_t)_{t\geqslant 1}, b_t \leqslant b_{t+1} \leqslant 2b_t\}$, the set of all increasing sequences with moderate growth.

Theorem 5.4 *Suppose that we are given a sequence of parameters $\mathcal{C} = \{(\varepsilon_t, A_t, \delta_t)\}_{t\geqslant 1}$, such that the sequences $(\log \varepsilon_t^{-1})_{t\geqslant 1}$, $(\delta_t^{-1})_{t\geqslant 1}$, $(\log A_t)_{t\geqslant 1} \in \mathcal{F}$, and*

$$\varepsilon_t \leqslant \frac{\delta_t^{10}}{(\log A_t)^5}, \qquad \varepsilon_t \leqslant \frac{1}{t}, \qquad\qquad \forall t \geqslant 1. \tag{9}$$

Then, under Exponential Time Hypothesis (Conjecture 5.3), no $A^{o(\delta^{-2}\log(1/\varepsilon))}$-time algorithm can determine the ε-optimal value of any given δ-strongly separated LMDP with $(\varepsilon, \delta, A) \in \mathcal{C}$ whose parameters H, L, S satisfy $H \leqslant \frac{\log(1/\varepsilon)}{\delta^2}$ and $\max\{L, S\} = \mathrm{poly}(\log(1/\varepsilon), \log A, \delta^{-1})$.

4. In this section, we omit the subscript of $\theta^\star$ for notational simplicity, because the LMDP $M = M_{\theta^\star}$ is given and fixed.

Algorithm 2 Short Memory Planning with Context Inference

Data: $W \geqslant 1$, LMDP model $M = M_{\theta^\star}$

Set $\widehat{V}_{m,H+1}(\varnothing) = 0$ for all $m \in [L]$

for $h = H, H-1, \cdots, W$ **do**

For each pair $(s_h, a_h, m) \in \mathcal{S} \times \mathcal{A} \times [L]$, update

$$\widehat{Q}_{m,h}(s_h, a_h) = \mathbb{E}_{s_{h+1} \sim \mathbb{T}_m(\cdot|s_h,a_h)}\left[\widehat{V}_{m,h+1}(s_{h+1})\right] + R_h(s_h, a_h).$$

Set $\widehat{V}_{m,h}(s_h) = \max_{a_h} \widehat{Q}_{m,h}(s_h, a_h)$ and store $\pi_{m,h}(s_h) = \arg\max_{a_h} \widehat{Q}_{m,h}(s_h, a_h)$.

end

for *each* $\overline{\tau}_W = (s_1, a_1, \cdots, s_W)$ **do**

Compute $m = m(\overline{\tau}_W)$ and set

$$\widehat{V}(\overline{\tau}_W) = \mathbb{P}(m|\overline{\tau}_W) \cdot \widehat{V}_{m,W}(s_W).$$

end

for $h = W-1, \cdots, 1$ **do**

For each $(\overline{\tau}_h, a_h) \in (\mathcal{S} \times \mathcal{A})^h$, update

$$\widehat{Q}(\overline{\tau}_h, a_h) = \mathbb{E}_{s_{h+1}|\overline{\tau}_h, a_h}\left[\widehat{V}(\overline{\tau}_h, a_h, s_{h+1})\right] + R_h(s_h, a_h), \qquad \forall \overline{\tau}_h, a_h$$

Set $\widehat{V}(\overline{\tau}_h) = \max_{a_h} \widehat{Q}(\overline{\tau}_h, a_h)$ and store $\pi_h(\overline{\tau}_h) = \arg\max_{a_h} \widehat{Q}(\overline{\tau}_h, a_h)$.

end

Result: description of the determinstic policy $\widehat{\pi}$ given by

$$\widehat{\pi}(\overline{\tau}_h) = \begin{cases} \pi_h(\overline{\tau}_h), & h < W, \\ \pi_{h,m(\overline{\tau}_W)}(s_h), & h \geqslant W. \end{cases}$$

Theorem 5.5 *Suppose that we are given a sequence of parameters* $\mathcal{C} = \{(L_t, A_t, \delta_t)\}_{t \geqslant 1}$*, such that the sequences* $(\log L_t)_{t \geqslant 1}, (\delta_t^{-1})_{t \geqslant 1}, (\log A_t)_{t \geqslant 1} \in \mathcal{F}$*,* $(L_t)_{t \geqslant 1}$ *is strictly increasing, and*

$$\log\log L_t \ll \frac{\log A_t}{\delta_t^2} \leqslant \operatorname{poly}\log L_t, \qquad \forall t \geqslant 1. \tag{10}$$

Then, under Exponential Time Hypothesis (Conjecture 5.3), no $A^{o\left(\delta^{-2}\frac{\log L}{\log\log L}\right)}$*-time algorithm can determine the* ε*-optimal value of any given* δ*-strongly separated LMDP with* $(L, A, \delta) \in \mathcal{C}$ *whose parameters* H, L, S *satisfy* $H \leqslant \frac{\log L}{\delta^2}$*, and* $\varepsilon = \frac{1}{\operatorname{poly}(\log L)}$*,* $S = \exp\left(\mathcal{O}(\log^2 \log L)\right)$*.*

In particular, the results above show that under ETH, a time complexity that scales with $A^{\delta^{-2}\log(L/\varepsilon)}$ is hard to avoid for planning in δ-strongly separated LMDP, in the sense that our iteration complexity lower bounds apply to any planning algorithm that works for general parameters $(L, A, \delta, \varepsilon)$. Therefore, the threshold $H_\star \asymp \frac{\log(L/\varepsilon)}{\delta^2}$ indeed also captures the computational complexity of planning.

Acknowledgements

CD is supported by NSF Awards CCF-1901292, DMS-2022448, and DMS2134108, a Simons Investigator Award, and the Simons Collaboration on the Theory of Algorithmic Fairness. NG is supported by a Fannie & John Hertz Foundation Fellowship and an NSF Graduate Fellowship. FC and AR acknowledge support from ARO through award W911NF-21-1-0328, DOE through award DE-SC0022199, and the Simons Foundation and the NSF through award DMS-2031883.

References

Alekh Agarwal and Tong Zhang. Model-based rl with optimistic posterior sampling: Structural conditions and sample complexity. *arXiv preprint arXiv:2206.07659*, 2022.

Baruch Awerbuch and Robert Kleinberg. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74(1):97–114, 2008.

Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar. Reinforcement learning of pomdps using spectral methods. In *Conference on Learning Theory*, pages 193–256. PMLR, 2016.

Richard Bellman. A markovian decision process. *Journal of mathematics and mechanics*, pages 679–684, 1957.

Emma Brunskill and Lihong Li. Sample complexity of multi-task reinforcement learning. *arXiv preprint arXiv:1309.6821*, 2013.

Peter Buchholz and Dimitri Scheftelowitsch. Computation of weighted sums of rewards for concurrent mdps. *Mathematical Methods of Operations Research*, 89:1–42, 2019.

Dima Burago, Michel De Rougemont, and Anatol Slissenko. On the complexity of partially observed markov decision processes. *Theoretical Computer Science*, 157(2):161–183, 1996.

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

Iadine Chades, Josie Carwardine, Tara Martin, Samuel Nicol, Régis Sabbadin, and Olivier Buffet. Momdps: a solution for modelling adaptive management problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 267–273, 2012.

Fan Chen, Yu Bai, and Song Mei. Partially observable rl with b-stability: Unified structural condition and sharp sample-efficient algorithms. *arXiv preprint arXiv:2209.14990*, 2022a.

Fan Chen, Song Mei, and Yu Bai. Unified algorithms for rl with decision-estimation coefficients: No-regret, pac, and reward-free learning. *arXiv preprint arXiv:2209.11745*, 2022b.

Fan Chen, Huan Wang, Caiming Xiong, Song Mei, and Yu Bai. Lower bounds for learning in revealing pomdps. *arXiv preprint arXiv:2302.01333*, 2023.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Finale Doshi-Velez and George Konidaris. Hidden parameter markov decision processes: A semiparametric regression approach for discovering latent task parametrizations. In *IJCAI: proceedings of the conference*, volume 2016, page 1432. NIH Public Access, 2016.

Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

Yonathan Efroni, Chi Jin, Akshay Krishnamurthy, and Sobhan Miryoosefi. Provable reinforcement learning with a short-term memory. *arXiv preprint arXiv:2202.03983*, 2022.

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Reinforcement learning in pomdps without resets. 2005.

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. The value of observation for monitoring dynamic systems. In *IJCAI*, pages 2474–2479, 2007.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Learning in observable pomdps, without computationally intractable oracles. *arXiv preprint arXiv:2206.03446*, 2022a.

Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Planning in observable pomdps in quasipolynomial time. *arXiv preprint arXiv:2201.04735*, 2022b.

Zhaohan Daniel Guo, Shayan Doroudi, and Emma Brunskill. A pac rl algorithm for episodic pomdps. In *Artificial Intelligence and Statistics*, pages 510–518. PMLR, 2016.

Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015.

Yanjun Han. Mixture vs mixture and moment matching, 2019. URL https://theinformaticists.com/2019/08/28/lecture-7-mixture-vs-mixture-and-moment-matching/.

Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

Russell Impagliazzo and Ramamohan Paturi. On the complexity of k-sat. *Journal of Computer and System Sciences*, 62(2):367–375, 2001.

Mehdi Jafarnia Jahromi, Rahul Jain, and Ashutosh Nayyar. Online learning for unknown partially observable mdps. In *International Conference on Artificial Intelligence and Statistics*, pages 1712–1732. PMLR, 2022.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

Chi Jin, Sham Kakade, Akshay Krishnamurthy, and Qinghua Liu. Sample-efficient reinforcement learning of undercomplete pomdps. *Advances in Neural Information Processing Systems*, 33: 18530–18539, 2020.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

Michael Kearns, Yishay Mansour, and Andrew Ng. Approximate planning in large pomdps via reusable trajectories. *Advances in Neural Information Processing Systems*, 12, 1999.

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Pac reinforcement learning with rich observations. *Advances in Neural Information Processing Systems*, 29, 2016.

Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, and Shie Mannor. Reinforcement learning in reward-mixing mdps. *Advances in Neural Information Processing Systems*, 34:2253–2264, 2021a.

Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, and Shie Mannor. Rl for latent mdps: Regret guarantees and a lower bound. *Advances in Neural Information Processing Systems*, 34: 24523–24534, 2021b.

Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, and Shie Mannor. Reward-mixing mdps with few latent contexts are learnable. In *International Conference on Machine Learning*, pages 18057–18082. PMLR, 2023.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Jonathan Lee, Alekh Agarwal, Christoph Dann, and Tong Zhang. Learning in pomdps is sample-efficient with hindsight observability. In *International Conference on Machine Learning*, pages 18733–18773. PMLR, 2023.

Michael L Littman. Memoryless policies: Theoretical limitations and practical results. In *From Animals to Animats 3: Proceedings of the third international conference on simulation of adaptive behavior*, volume 3, page 238. MIT Press Cambridge, MA, USA, 1994.

Qinghua Liu, Alan Chung, Csaba Szepesvári, and Chi Jin. When is partially observable reinforcement learning not scary? *arXiv preprint arXiv:2204.08967*, 2022a.

Qinghua Liu, Praneeth Netrapalli, Csaba Szepesvari, and Chi Jin. Optimistic mle–a generic model-based algorithm for partially observable sequential decision making. *arXiv preprint arXiv:2209.14997*, 2022b.

Qinghua Liu, Csaba Szepesvári, and Chi Jin. Sample-efficient reinforcement learning of partially observable markov games. *arXiv preprint arXiv:2206.01315*, 2022c.

Qinghua Liu, Praneeth Netrapalli, Csaba Szepesvari, and Chi Jin. Optimistic mle: A generic model-based algorithm for partially observable sequential decision making. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 363–376, 2023.

Yao Liu, Zhaohan Guo, and Emma Brunskill. Pac continuous state online multitask reinforcement learning with identification. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 438–446, 2016.

Christopher Lusena, Judy Goldsmith, and Martin Mundhenk. Nonapproximability results for partially observable markov decision processes. *Journal of artificial intelligence research*, 14:83–103, 2001.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Christos H Papadimitriou and John N Tsitsiklis. The complexity of markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.

Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6(2012-2016):7, 2014.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

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

Lauren N Steimle, David L Kaufman, and Brian T Denton. Multi-model markov decision processes. *IISE Transactions*, 53(10):1124–1139, 2021.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pages 2898–2933. PMLR, 2019.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. 2018.

Masatoshi Uehara, Ayush Sekhari, Jason D Lee, Nathan Kallus, and Wen Sun. Computationally efficient pac rl in pomdps with latent determinism and conditional embeddings. *arXiv preprint arXiv:2206.12081*, 2022a.

Masatoshi Uehara, Ayush Sekhari, Jason D Lee, Nathan Kallus, and Wen Sun. Provably efficient reinforcement learning in partially observable dynamical systems. *arXiv preprint arXiv:2206.12020*, 2022b.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Embed to control partially observed systems: Representation learning with provable sample efficiency. *arXiv preprint arXiv:2205.13476*, 2022.

Peter R. Wurman, Samuel Barrett, Kenta Kawamoto, James MacGlashan, Kaushik Subramanian, Thomas J. Walsh, Roberto Capobianco, Alisa Devlic, Franziska Eckert, Florian Fuchs, Leilani Gilpin, Piyush Khandelwal, Varun Kompella, HaoChih Lin, Patrick MacAlpine, Declan Oller, Takuma Seno, Craig Sherstan, Michael D. Thomure, Houmehr Aghabozorgi, Leon Barrett, Rory Douglas, Dion Whitehead, Peter Dürr, Peter Stone, Michael Spranger, and Hiroaki Kitano. Outracing champion gran turismo drivers with deep reinforcement learning. *Nature*, 602(7896): 223–228, 2022. doi: 10.1038/s41586-021-04357-7. URL https://doi.org/10.1038/s41586-021-04357-7.

Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M Kakade. The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157*, 2022.

Yi Xiong, Ningyuan Chen, Xuefeng Gao, and Xiang Zhou. Sublinear regret for learning pomdps. *arXiv preprint arXiv:2107.03635*, 2021.

Wenhao Zhan, Masatoshi Uehara, Wen Sun, and Jason D Lee. Pac reinforcement learning for predictive state representations. *arXiv preprint arXiv:2207.05738*, 2022.

Runlong Zhou, Ruosong Wang, and Simon Shaolei Du. Horizon-free and variance-dependent reinforcement learning for latent markov decision processes. In *International Conference on Machine Learning*, pages 42698–42723. PMLR, 2023.

Appendix A. Technical tools

A.1. Covering number

Definition A.1 (Covering) *A ρ-cover of the LMDP model class Θ is a tuple $(\widehat{\mathbb{P}}, \Theta_0)$, where $\Theta_0 \subset \Theta$ is a finite set, and for each $\theta_0 \in \Theta_0$, $\pi \in \Pi_{\mathrm{RND}}$, $\widehat{\mathbb{P}}^{\pi}_{\theta_0}(\cdot) \in \mathbb{R}^{\mathcal{T}}_{\geqslant 0}$ specifies an* optimistic likelihood function *such that the following holds:*

(1) For $\theta \in \Theta$, there exists a $\theta_0 \in \Theta_0$ satisfying: for all $\tau \in \mathcal{T}^H$ and $\pi \in \Pi_{\mathrm{RND}}$, it holds that $\widehat{\mathbb{P}}^{\pi}_{\theta_0}(\tau) \geqslant \mathbb{P}^{\pi}_{\theta}(\tau)$.

(2) For $\theta \in \Theta_0$, $\pi \in \Pi_{\mathrm{RND}}$, it holds $\left\| \mathbb{P}^{\pi}_{\theta}(\tau_H = \cdot) - \widehat{\mathbb{P}}^{\pi}_{\theta}(\tau_H = \cdot) \right\|_1 \leqslant \rho^2$.

The optimistic covering number $N_\Theta(\rho)$ is defined as the minimal cardinality of Θ_0 such that there exists $\widetilde{\mathbb{P}}$ such that $(\widetilde{\mathbb{P}}, \Theta_0)$ is an optimistic ρ-cover of Θ.

The above definition of covering is taken from Chen et al. (2022b). It is known that the covering number defined above can be upper bounded by the bracket number adopted in Zhan et al. (2022); Liu et al. (2022b). In particular, when Θ is a class of LMDPs with $|\mathcal{S}| = S, |\mathcal{A}| = A$, horizon H, and with L latent contexts, we have

$$\log N_\Theta(\rho) \leqslant CLS^2 A \log(CLSAH/\rho),$$

where C is an absolute constant (see e.g. Chen et al. (2022a); Liu et al. (2022a)).

A.2. Information theory

In this section, we summarize several basic inequalities related to TV distance, Hellinger distance and Bhattacharyya divergence.

Lemma A.2 *For any two distribution $\mathbb{P}, \mathbb{Q}$ over $\mathcal{X}$, it holds that $D_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) \leqslant \sqrt{2} D_{\mathrm{H}}(\mathbb{P}, \mathbb{Q})$, and*

$$D_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) \geqslant D^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}) = 1 - \exp\left(-D_{\mathrm{B}}(\mathbb{P}, \mathbb{Q})\right). \tag{11}$$

Conversely, we also have (Pinsker inequality)

$$D_{\mathrm{B}}(\mathbb{P}, \mathbb{Q}) \geqslant -\frac{1}{2} \log(1 - D^2_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})) \geqslant \frac{1}{2} D^2_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}). \tag{12}$$

Lemma A.3 (Foster et al. (2021, Lemma A.11)) *For distributions $\mathbb{P}, \mathbb{Q}$ defined on $\mathcal{X}$ and function $h : \mathcal{X} \to [0, R]$, we have*

$$\mathbb{E}_{\mathbb{P}}[h(X)] \leqslant 3\mathbb{E}_{\mathbb{Q}}[h(X)] + 2RD^2_{\mathrm{H}}(\mathbb{P}, \mathbb{Q}).$$

Lemma A.4 *For any pair of random variable (X, Y), it holds that*

$$\mathbb{E}_{X \sim \mathbb{P}_X}\left[D_{\mathrm{TV}}\left(\mathbb{P}_{Y|X}, \mathbb{Q}_{Y|X}\right)\right] \leqslant 2D_{\mathrm{TV}}\left(\mathbb{P}_{X,Y}, \mathbb{Q}_{X,Y}\right).$$

Conversely, it holds that

$$D_{\mathrm{TV}}\left(\mathbb{P}_{X,Y}, \mathbb{Q}_{X,Y}\right) \leqslant D_{\mathrm{TV}}\left(\mathbb{P}_X, \mathbb{Q}_X\right) + \mathbb{E}_{X \sim \mathbb{P}_X}\left[D_{\mathrm{TV}}\left(\mathbb{P}_{Y|X}, \mathbb{Q}_{Y|X}\right)\right].$$

Lemma A.5 (Chen et al. (2022b, Lemma A.4)) *For any pair of random variable (X, Y), it holds that*

$$\mathbb{E}_{X\sim\mathbb{P}_X}\left[D_{\mathrm{H}}^2\left(\mathbb{P}_{Y|X}, \mathbb{Q}_{Y|X}\right)\right] \leqslant 2D_{\mathrm{H}}^2\left(\mathbb{P}_{X,Y}, \mathbb{Q}_{X,Y}\right).$$

Conversely, it holds that

$$D_{\mathrm{H}}^2\left(\mathbb{P}_{X,Y}, \mathbb{Q}_{X,Y}\right) \leqslant 3D_{\mathrm{H}}^2\left(\mathbb{P}_{X}, \mathbb{Q}_{X}\right) + 2\mathbb{E}_{X\sim\mathbb{P}_X}\left[D_{\mathrm{H}}^2\left(\mathbb{P}_{Y|X}, \mathbb{Q}_{Y|X}\right)\right].$$

A.3. Technical inequalities

Lemma A.6 *For distributions $\mathbb{P}_1, \cdots, \mathbb{P}_L \in \Delta(\mathcal{O})$ and $\mu, \nu \in \Delta([L])$ so that $\mathrm{supp}(\mu)\cap\mathrm{supp}(\nu) = \varnothing$, we have*

$$D_{\mathrm{B}}(\mathbb{E}_{i\sim\mu}[\mathbb{P}_i], \mathbb{E}_{j\sim\nu}[\mathbb{P}_j]) \geqslant \min_{i\neq j} D_{\mathrm{B}}(\mathbb{P}_i, \mathbb{P}_j) - \log(L/2).$$

As a corollary, if $D_{\mathrm{B}}(\mathbb{P}_i, \mathbb{P}_j) \geqslant \log L$ for all $i \neq j$, then for any $\mu, \nu \in \Delta([L])$, we have

$$D_{\mathrm{TV}}\left(\mathbb{E}_{i\sim\mu}[\mathbb{P}_i], \mathbb{E}_{j\sim\nu}[\mathbb{P}_j]\right) \geqslant \frac{1}{2}D_{\mathrm{TV}}\left(\mu, \nu\right).$$

Proof. By definition,

$$\begin{aligned}
\exp\left(-D_{\mathrm{B}}(\mathbb{E}_{i\sim\mu}[\mathbb{P}_i], \mathbb{E}_{j\sim\nu}[\mathbb{P}_j])\right) &= \sum_x \sqrt{\mathbb{E}_{i\sim\mu}[\mathbb{P}_i(x)]\mathbb{E}_{j\sim\nu}[\mathbb{P}_j(x)]} \\
&\leqslant \sum_x \sum_{i,j} \sqrt{\mu(i)\nu(j)\mathbb{P}_i(x)\mathbb{P}_j(x)} \\
&= \sum_{i,j} \sqrt{\mu(i)\nu(j)} \exp\left(-D_{\mathrm{B}}(\mathbb{P}_i, \mathbb{P}_j)\right) \\
&\leqslant \left(\sum_i \sqrt{\mu(i)}\right)\left(\sum_j \sqrt{\nu(j)}\right) \max_{i\neq j} \exp\left(-D_{\mathrm{B}}(\mathbb{P}_i, \mathbb{P}_j)\right) \\
&\leqslant \frac{L}{2} \exp\left(-\min_{i\neq j} D_{\mathrm{B}}(\mathbb{P}_i, \mathbb{P}_j)\right),
\end{aligned}$$

where the last inequality follows from the fact that $\sum_i \sqrt{\mu(i)} \leqslant \sqrt{\#\mathrm{supp}(\mu)}$ and $\sum_j \sqrt{\nu(j)} \leqslant \sqrt{\#\mathrm{supp}(\nu)}$. Taking $-\log$ on both sides completes the proof. □

Lemma A.7 *Suppose that for distributions $\mathbb{P}_1, \cdots, \mathbb{P}_L \in \Delta(\mathcal{O})$, we have $D_{\mathrm{B}}(\mathbb{P}_i, \mathbb{P}_j) \geqslant \log(2L)$ for all $i \neq j$. Then for the matrix $\mathbb{M} = [\mathbb{P}_1, \cdots, \mathbb{P}_L] \in \mathbb{R}^{\mathcal{O}\times L}$, there exists $\mathbb{M}^+ \in \mathbb{R}^{L\times\mathcal{O}}$ such that $\|\mathbb{M}^+\|_1 \leqslant 2$ and $\mathbb{M}^+\mathbb{M} = I_L$.*

Proof. We construct $\mathbb{M}^+$ explicitly. Consider the matrix $Z \in \mathbb{R}^{L\times\mathcal{O}}$ given by

$$[Z]_{m,o} = \frac{\mathbb{P}_m(o)}{\sum_{i\in[L]} \mathbb{P}_i(o)}.$$

Then clearly $\|Z\|_1 \leqslant 1$, and the matrix $Y = Z\mathbb{M}$ is given by

$$[Y]_{l,m} = \sum_{o\in\mathcal{O}} \frac{\mathbb{P}_l(o)\mathbb{P}_m(o)}{\sum_{i\in[L]} \mathbb{P}_i(o)}.$$

For $l \neq m$, we know

$$0 \leqslant [Y]_{l,m} \leqslant \sum_{o\in\mathcal{O}} \frac{\mathbb{P}_l(o)\mathbb{P}_m(o)}{2\sqrt{\mathbb{P}_l(o)\mathbb{P}_m(o)}} = \frac{1}{2}\sum_{o\in\mathcal{O}} \sqrt{\mathbb{P}_l(o)\mathbb{P}_m(o)} = \frac{1}{2}\exp\left(-D_{\mathrm{B}}(\mathbb{P}_l, \mathbb{P}_m)\right) \leqslant \frac{1}{4L}.$$

Furthermore,

$$0 \leqslant 1 - [Y]_{m,m} = \sum_{o\in\mathcal{O}} \sum_{l\neq m} \frac{\mathbb{P}_l(o)\mathbb{P}_m(o)}{\sum_{i\in[L]} \mathbb{P}_i(o)} = \sum_{l\neq m} [Y]_{l,m} \leqslant \frac{1}{4}.$$

Combining these two inequalities, we know $\|I_L - Y\|_1 \leqslant \frac{1}{2}$, and hence $\left\|Y^{-1}\right\|_1 \leqslant 2$. Therefore, we can take $\mathbb{M}^+ = Y^{-1}Z$ so that $\|\mathbb{M}^+\|_1 \leqslant \left\|Y^{-1}\right\|_1 \|Z\|_1 \leqslant 2$ and $\mathbb{M}^+\mathbb{M} = I_L$. □

A.4. Eluder arguments

In this section, we present the eluder arguments that are necessary for our analysis in Appendix E. The following proposition is from Chen et al. (2022a, Corollary E.2) (with suitable rescaling).

Proposition A.8 (Chen et al. (2022a)) *Suppose we have a sequence of functions $\{f_k : \mathbb{R}^n \to \mathbb{R}\}_{k\in[K]}$:*

$$f_k(x) := \max_{r\in\mathcal{R}} \sum_{j=1}^{J} |\langle x, y_{k,j,r}\rangle| ,$$

which is given by the family of vectors $\{y_{k,j,r}\}_{(k,j,r)\in[K]\times[J]\times\mathcal{R}} \subset \mathbb{R}^n$. Further assume that there exists $L_1 > 0$ such that $f_k(x) \leqslant L_1 \|x\|_1$.

Consider further a sequence of vectors $(x_i)_{i\in\mathcal{I}} \subset \mathbb{R}^n$ such that the subspace spanned by $(x_i)_{i\in\mathcal{I}}$ has dimension at most d. Then for any sequence of $p_1, \cdots, p_K \in \Delta(\mathcal{I})$ and constant $M > 0$, it holds that

$$\sum_{k=1}^{K} M \wedge \mathbb{E}_{i\sim p_k}[f_k(x_i)] \leqslant \sqrt{4d\log\left(1 + \frac{KdL_1 \max_i \|x_i\|_1}{M}\right)\left[KM + \sum_{k=1}^{K}\sum_{t<k} \mathbb{E}_{i\sim p_t}[f_k(x_i)^2]\right]}.$$

The following proposition is an generalized version of the results in Xie et al. (2022, Appendix D). We provide a proof for the sake of completeness.

Proposition A.9 (Xie et al. (2022)) *Suppose that $p_1, \cdots, p_K$ is a sequence of distributions over $\mathcal{X}$, and there exists $\mu \in \Delta(\mathcal{X})$ such that $p_k(x)/\mu(x) \leqslant C_{\mathrm{cov}}$ for all $x \in \mathcal{X}$, $k \in [K]$. Then for any sequence $f_1, \cdots, f_K$ of functions $\mathcal{X} \to [0, 1]$ and constant $M \geqslant 1$, it holds that*

$$\sum_{k=1}^{K} \mathbb{E}_{x\sim p_k} f_k(x) \leqslant \sqrt{2C_{\mathrm{cov}}\log\left(1 + \frac{C_{\mathrm{cov}}K}{M}\right)\left[2KM + \sum_{k=1}^{K}\sum_{t<k} \mathbb{E}_{x\sim p_t} f_k(x)^2\right]}$$

Proof. For any $x \in \mathcal{X}$, define

$$\tilde{p}_k(x) = M\mu(x) + \sum_{t \leqslant k} p_t(x).$$

Then by Cauchy inequality,

$$\mathbb{E}_{x \sim p_k} f_k(x) = \sum_{x \in \mathcal{X}} p_k(x) f_k(x) \leqslant \sqrt{\sum_{x \in \mathcal{X}} \frac{p_k(x)^2}{\tilde{p}_k(x)} \sum_{x \in \mathcal{X}} \tilde{p}_k(x) f_k(x)^2}.$$

Applying Cauchy inequality again, we obtain

$$\sum_{k=1}^{K} \mathbb{E}_{x \sim p_k} f_k(x) \leqslant \sqrt{\sum_{k=1}^{K} \sum_{x \in \mathcal{X}} \frac{p_k(x)^2}{\tilde{p}_k(x)}} \cdot \sqrt{\sum_{k=1}^{K} \sum_{x \in \mathcal{X}} \tilde{p}_k(x) f_k(x)^2}$$

Notice that

$$\sum_{x \in \mathcal{X}} \tilde{p}_k(x) f_k(x)^2 \leqslant M + 1 + \sum_{t < k} \mathbb{E}_{x \sim p_t} f_k(x)^2,$$

and hence it remains to bound

$$\sum_{k=1}^{K} \sum_{x \in \mathcal{X}} \frac{p_k(x)^2}{\tilde{p}_k(x)} \leqslant \sum_{x \in \mathcal{X}} C_{\mathrm{cov}} \mu(x) \cdot \sum_{k=1}^{K} \frac{p_k(x)}{\tilde{p}_k(x)}.$$

Using the fact that $u \leqslant 2\log(1+u) \forall u \in [0,1]$, we have

$$\begin{aligned}
\sum_{k=1}^{K} \frac{p_k(x)}{\tilde{p}_k(x)} &\leqslant 2 \sum_{k=1}^{K} \log\left(1 + \frac{p_k(x)}{\tilde{p}_k(x)}\right) \\
&\leqslant 2 \sum_{k=1}^{K} \log\left(1 + \frac{p_k(x)}{M\mu(x) + \sum_{t<k} p_t(x)}\right) \\
&= 2\log\left(\frac{M\mu(x) + \sum_{t \leqslant K} p_t(x)}{M\mu(x)}\right) \\
&\leqslant 2\log\left(1 + \frac{C_{\mathrm{cov}} K}{M}\right)
\end{aligned}$$

Combining the inequalities above completes the proof. $\square$

Proposition A.10 *Suppose that* $\mathbb{T} \in \mathbb{R}^{\mathcal{S} \times (\mathcal{S} \times \mathcal{A})}$ *is a transition matrix such that* $\mathrm{rank}(\mathbb{T}) = d$*. Then there exists a distribution* $\nu \in \Delta(\mathcal{S})$ *such that* $\mathbb{T}(s'|s,a) \leqslant d \cdot \nu(s')$ $\forall (s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$.

Proof. Consider the set

$$\mathcal{P} = \{\mathbb{T}(\cdot|s,a) : s \in \mathcal{S}, a \in \mathcal{A}\} \subset \mathbb{R}^{\mathcal{S}}.$$

Then $\mathrm{rank}(\mathbb{T}) = d$ implies that $\mathcal{P}$ spans a d-dimensional subspace of $\mathbb{R}^{\mathcal{S}}$. Clearly, $\mathcal{P}$ is compact, and hence it has a barycentric spanner (Awerbuch and Kleinberg, 2008), i.e. there exists $\{\nu_1, \cdots, \nu_d\} \subseteq \mathcal{P}$, such that for any $\mu \in \mathcal{P}$, there are $\lambda_1, \cdots, \lambda_d \in [-1,1]$ such that

$$\mu = \lambda_1 \nu_1 + \cdots + \lambda_d \nu_d.$$

Therefore, we can take $\nu = \frac{1}{d} \sum_{i=1}^{d} \nu_i$. $\square$

Appendix B. Further comparison with related work

In Kwon et al. (2021b), to learn a δ-strongly separated LMDP, the proposed algorithms require a horizon $H \gtrsim \delta^{-4} \log^2(S/\delta) \log(LSA\varepsilon^{-1}\delta^{-1})$, and also one of the following assumptions:

- a good initialization, i.e. an initial approximation of the latent dynamics of the ground truth model, with error bounded by $o(\delta^2)$ (Kwon et al., 2021b, Theorem 3.4).
- The so-called sufficient-test condition and sufficient-history condition, along with the reachability of states (Kwon et al., 2021b, Theorem 3.5).

Chen et al. (2022a) further show that, for general LMDPs (not necessarily δ-strongly separated), the sufficient-test condition itself implies that the OMLE algorithm is sample-efficient. More concretely, their result applies to any *W-step revealing LMDP*. A LMDP is W-step α-revealing if the W-step emission matrix

$$\mathbb{K}(s) := [\mathbb{T}_m(s_{2:W} = \mathbf{s}|s_1 = s, a_{1:W-1} = \mathbf{a})]_{(\mathbf{s},\mathbf{a}),m} \in \mathbb{R}^{(\mathcal{A}\times\mathcal{S})^{W-1}\times[L]}$$

admits a left inverse $\mathbb{K}^+(s)$ for all $s \in \mathcal{S}$ such that $\|\mathbb{K}^+(s)\|_1 \leqslant \alpha^{-1}$. This condition implies the standard W-step revealing condition of POMDPs (Liu et al., 2022a; Chen et al., 2022a) because the state s is observable in LMDPs[5]. In particular, the following theorem now follows from Chen et al. (2022a, Theorem 9).

Theorem B.1 *The class of W-step α-revealing LMDPs can be learning using* $\mathrm{poly}(A^W, \alpha^{-1}, L, S, H, \varepsilon^{-1})$ *samples.*

Without additional assumption, it is only known that a δ-strongly separated LMDP is W-step α-revealing with $W = \lceil \frac{2\log(2L)}{\delta^2} \rceil$ and $\alpha = 2$. [6] Therefore, when applied to δ-strongly separated LMDPs, Theorem B.1 gives a sample complexity bound that scales with $A^{\delta^{-2}\log L}$, which is quasi-polynomial in (A, L). Further, as Theorem 3.2 indicates, such a quasi-polynomial sample complexity is also unavoidable if the analysis only relies on the revealing structure of δ-strongly separated LMDP and does not take the horizon length H into account.

On the other hand, our analysis in Appendix E is indeed built upon the revealing structure of δ-strongly separated LMDP. However, we also leverage the special structure of separated LMDP, so that we can avoid using the brute-force exploration strategy that essentially samples $a_{H-W+1:H-1} \sim \mathrm{Unif}(\mathcal{A}^{W-1})$ in the course of the algorithm. Such a uniform-sampling exploration approach for learning the system dynamics of the last W steps is generally necessary in learning revealing POMDPs, as the lower bounds of Chen et al. (2023) indicate. It turns out to be unnecessary for separated LMDP. Appendix E.2 provides a technical overview with more details.

Appendix C. Proofs for Section 2

C.1. Proof of Proposition 2.4

Fix $m, l \in \mathrm{supp}(\rho)$, $m \neq l$. By definition,

$$D_{\mathrm{B}}(\mathbb{M}_{m,h+1}(\pi, s), \mathbb{M}_{l,h+1}(\pi, s))$$

5. see, e.g. Chen et al. (2022a, Proposition B.10) or the proof of Theorem E.5 in Appendix E.6.
6. This result can be obtained by applying Lemma A.7 to the distributions of trajectories induced by policy $\mathrm{Unif}(\mathcal{A}^{W-1})$.

$$
\begin{aligned}
&= -\log \sum_{a_{1:h}, s_{2:h+1}} \sqrt{\mathbb{T}_m^\pi(a_1, s_2, \cdots, s_{h+1}|s_1 = s)\mathbb{T}_l^\pi(a_1, s_2, \cdots, s_{h+1}|s_1 = s)} \\
&= -\log \sum_{a_{1:h}, s_{2:h}} \sqrt{\mathbb{T}_m^\pi(a_1, s_2, \cdots, s_h, a_h|s_1 = s)\mathbb{T}_l^\pi(a_1, s_2, \cdots, s_h, a_h|s_1 = s)} \cdot \sum_{s_h} \sqrt{\mathbb{T}_m(s_{h+1}|s_h, a_h)\mathbb{T}_l(s_{h+1}|s_h, a_h)} \\
&= -\log \sum_{a_{1:h}, s_{2:h}} \sqrt{\mathbb{T}_m^\pi(a_1, s_2, \cdots, s_h, a_h|s_1 = s)\mathbb{T}_l^\pi(a_1, s_2, \cdots, s_h, a_h|s_1 = s)} \cdot \exp\left(-D_{\mathrm{B}}(\mathbb{T}_m(\cdot|s_h, a_h), \mathbb{T}_l(\cdot|s_h, a_h))\right).
\end{aligned}
$$

Because M is a δ-strongly separated LMDP, using (12), we know

$$
D_{\mathrm{B}}(\mathbb{T}_m(\cdot|s, a), \mathbb{T}_l(\cdot|s, a)) \geqslant \frac{1}{2} D_{\mathrm{TV}}^2\left(\mathbb{T}_m(\cdot|s, a), \mathbb{T}_l(\cdot|s, a)\right) \geqslant \frac{\delta^2}{2}, \qquad \forall (s, a) \in \mathcal{S} \times \mathcal{A}.
$$

Therefore, we can proceed to bound

$$
\begin{aligned}
& D_{\mathrm{B}}(\mathbb{M}_{m,h+1}(\pi, s), \mathbb{M}_{l,h+1}(\pi, s)) \\
\geqslant\ & \frac{\delta^2}{2} - \log \sum_{a_{1:h}, s_{2:h}} \sqrt{\mathbb{T}_m^\pi(a_1, s_2, \cdots, s_h, a_h|s_1 = s)\mathbb{T}_l^\pi(a_1, s_2, \cdots, s_h, a_h|s_1 = s)} \\
=\ & \frac{\delta^2}{2} - \log \sum_{a_{1:h-1}, s_{2:h}} \sqrt{\mathbb{T}_m^\pi(a_1, s_2, \cdots, s_h|s_1 = s)\mathbb{T}_l^\pi(a_1, s_2, \cdots, s_h|s_1 = s)} \cdot \sum_{a_h} \pi(a_h|s, a_1, s_2, \cdots, s_h) \\
=\ & \frac{\delta^2}{2} - \log \sum_{a_{1:h-1}, s_{2:h}} \sqrt{\mathbb{T}_m^\pi(a_1, s_2, \cdots, s_h|s_1 = s)\mathbb{T}_l^\pi(a_1, s_2, \cdots, s_h|s_1 = s)} \\
=\ & \frac{\delta^2}{2} + D_{\mathrm{B}}(\mathbb{M}_{m,h}(\pi, s), \mathbb{M}_{l,h}(\pi, s)).
\end{aligned}
$$

Applying the inequality above recursively, we obtain $D_{\mathrm{B}}(\mathbb{M}_{m,h+1}(\pi, s), \mathbb{M}_{l,h+1}(\pi, s)) \geqslant \frac{\delta^2}{2} h$, the desired result. $\square$

C.2. Proof of Proposition 2.5

Suppose that M is a N-step decodable LMDP. By definition of ϖ_N-separation, we only need to show that for any $m, l \in \mathrm{supp}(\rho)$, $m \neq l$ and policy $\pi \in \Pi_{\mathrm{RND}}$, it holds that

$$
\mathrm{supp}(\mathbb{M}_{m,h}(\pi, s)) \cap \mathrm{supp}(\mathbb{M}_{l,h}(\pi, s)) = \varnothing, \qquad \forall h \geqslant N, s \in \mathcal{S},
$$

or equivalently,

$$
\mathbb{T}_m^\pi(a_1, s_2, \cdots, s_h|s_1 = s)\mathbb{T}_l^\pi(a_1, s_2, \cdots, s_h|s_1 = s) = 0, \qquad \forall h \geqslant N, \forall \overline{\tau}_h = (s_1, a_1, \cdots, s_h).
$$

This is because the N-step decoability of M implies that for any $\overline{\tau}_h = (s_1, a_1, \cdots, s_h)$, there exists at most one $m^\star \in \mathrm{supp}(\rho)$ such that

$$
\mathbb{T}_{m^\star}(s_2|s_1, a_1) \cdots \mathbb{T}_{m^\star}(s_h|s_{h-1}, a_{h-1}) > 0.
$$

The desired result follows immediately. $\square$

C.3. Proof of Lemma 2.6

For notational simplicity, we denote

$$\mathrm{BC}(\mathbb{P}, \mathbb{Q}) = \exp(-D_{\mathrm{B}}(\mathbb{P}, \mathbb{Q})).$$

Fix $h \geqslant 1$ and $m, l \in \mathrm{supp}(\rho)$, $m \neq l$. We only need to show that the following policy optimization problem

$$\max_{\pi \in \Pi_{\mathrm{RND}}} \mathrm{BC}(\mathbb{M}_{m,h+1}(\pi, s), \mathbb{M}_{l,h+1}(\pi, s)) \tag{13}$$

is attained at a deterministic Markov policy. Recall that

$$\begin{aligned}&\mathrm{BC}(\mathbb{M}_{m,h+1}(\pi, s), \mathbb{M}_{l,h+1}(\pi, s))\\ &= \sum_{a_{1:h}, s_{2:h}} \sqrt{\mathbb{T}^{\pi}_{m}(a_1, s_2, \cdots, s_h, a_h | s_1 = s)\mathbb{T}^{\pi}_{l}(a_1, s_2, \cdots, s_h, a_h | s_1 = s)} \cdot \mathrm{BC}(\mathbb{T}_m(\cdot|s_h, a_h), \mathbb{T}_l(\cdot|s_h, a_h)).\end{aligned}$$

Therefore, (13) is attained at a policy π with

$$\pi_h(s_h) = \arg\max_{a \in \mathcal{A}} \mathrm{BC}(\mathbb{T}_m(\cdot|s_h, a)\mathbb{T}_l(\cdot|s_h, a)).$$

Inductively repeating the argument above for $h' = h, h-1, \cdots, 1$ completes the proof. $\square$

C.4. Proof of Proposition 4.1

Notice that $m_\theta(\overline{\tau}_W) = \arg\max_{m \in \mathrm{supp}(\rho)} \widetilde{\mathbb{P}}_\theta(m|\overline{\tau}_W)$. Therefore,

$$\begin{aligned}\widetilde{\mathbb{P}}^{\pi}_{\theta}(m^\star \neq m_\theta(\overline{\tau}_W)) &= \sum_{\overline{\tau}_W} \widetilde{\mathbb{P}}_\theta(m^\star \neq m_\theta(\overline{\tau}_W)|\overline{\tau}_W) \cdot \widetilde{\mathbb{P}}^{\pi}_{\theta}(\overline{\tau}_W)\\ &= \sum_{\overline{\tau}_W} \sum_{m \neq m_\theta(\overline{\tau}_W)} \widetilde{\mathbb{P}}_\theta(m|\overline{\tau}_W) \cdot \widetilde{\mathbb{P}}^{\pi}_{\theta}(\overline{\tau}_W)\\ &= \sum_{m^\star, \overline{\tau}} \sum_{m \neq m_\theta(\overline{\tau}_W)} \widetilde{\mathbb{P}}_\theta(m|\overline{\tau}_W) \cdot \widetilde{\mathbb{P}}^{\pi}_{\theta}(m^\star, \overline{\tau}_W)\\ &\leqslant \sum_{m^\star, \overline{\tau}} \sum_{m \neq m^\star} \widetilde{\mathbb{P}}_\theta(m|\overline{\tau}_W) \cdot \widetilde{\mathbb{P}}^{\pi}_{\theta}(m^\star, \overline{\tau}_W)\\ &= \sum_{m \neq l} \sum_{\overline{\tau}_W} \frac{\widetilde{\mathbb{P}}^{\pi}_{\theta}(m, \overline{\tau}_W)\widetilde{\mathbb{P}}^{\pi}_{\theta}(l, \overline{\tau}_W)}{\widetilde{\mathbb{P}}^{\pi}_{\theta}(\overline{\tau}_W)}\\ &= \sum_{m \neq l} \sum_{\overline{\tau}_W} \frac{\widetilde{\mathbb{P}}^{\pi}_{\theta}(m, \overline{\tau}_W|s_1)\widetilde{\mathbb{P}}^{\pi}_{\theta}(l, \overline{\tau}_W|s_1)}{\widetilde{\mathbb{P}}^{\pi}_{\theta}(\overline{\tau}_W|s_1)} \widetilde{\mathbb{P}}_\theta(s_1).\end{aligned}$$

For any $s \in \mathcal{S}$ and $m \in [L]$, we denote $\rho_{m|s} = \widetilde{\mathbb{P}}_\theta(m|s_1 = s)$, and then

$$\widetilde{\mathbb{P}}^{\pi}_{\theta}(m, \overline{\tau}_W|s_1 = s) = \rho_{m|s}\mathbb{T}^{\pi}_{\theta,m}(\overline{\tau}_W|s_1 = s), \qquad \widetilde{\mathbb{P}}^{\pi}_{\theta}(\overline{\tau}_W|s_1 = s) = \sum_{m} \rho_{m|s}\mathbb{T}^{\pi}_{\theta,m}(\overline{\tau}_W|s_1 = s),$$

Therefore, using the fact that

$$\widetilde{\mathbb{P}}_{\theta}^{\pi}(\overline{\tau}_W|s_1=s) \geqslant 2\sqrt{\rho_{m|s}\rho_{l|s}}\cdot\sqrt{\mathbb{T}_{\theta,m}^{\pi}(\overline{\tau}_W|s_1=s)\mathbb{T}_{\theta,l}^{\pi}(\overline{\tau}_W|s_1=s)},$$

we have

$$\begin{aligned}\sum_{\overline{\tau}_W}\frac{\widetilde{\mathbb{P}}_{\theta}^{\pi}(m,\overline{\tau}_W|s_1)\widetilde{\mathbb{P}}_{\theta}^{\pi}(l,\overline{\tau}_W|s_1)}{\widetilde{\mathbb{P}}_{\theta}^{\pi}(\overline{\tau}_W|s_1)} &\leqslant \frac{\sqrt{\rho_{m|s}\rho_{l|s}}}{2}\sum_{\overline{\tau}_W}\sqrt{\mathbb{T}_{\theta,m}^{\pi}(\overline{\tau}_W|s_1)\mathbb{T}_{\theta,l}^{\pi}(\overline{\tau}_W|s_1)}\\ &= \frac{\sqrt{\rho_{m|s}\rho_{l|s}}}{2}\exp\left(-D_{\mathrm{B}}\left(\mathbb{M}_{m,W}^{\theta}(\pi,s_1),\mathbb{M}_{l,W}^{\theta}(\pi,s_1)\right)\right).\end{aligned}$$

Thus, taking summation over $m\neq l$ and using $\sum_{m\neq l}\sqrt{\rho_{m|s}\rho_{l|s}}\leqslant L-1$ gives

$$\widetilde{\mathbb{P}}_{\theta}^{\pi}(m_{\theta}(\overline{\tau}_W)\neq m^{\star})\leqslant L\exp(-\varpi(W)).$$

□

Appendix D. Proofs for Section 3

We first present two theorems that provide a more precise statement of our sample complexity lower bounds.

Theorem D.1 *There are constants c, C so that for any $H\geqslant 1$, $\delta\in(0,\frac{1}{4e^2}]$, $L\geqslant 2$ and integer $2\leqslant n\leqslant H-1$ satisfying*

$$Cn\log^4 n\leqslant\min\left\{\frac{\log L}{H\delta^2},\delta^{-1},2^{c\sqrt{\log L}}\right\},\tag{14}$$

there exists a class of δ-strongly separated LMDPs with L hidden MDPs, $S=(\log L)^{\mathcal{O}(\log n)}$ states, A actions, and horizon H, so that any algorithm requires $\Omega\left(\min\{A,L\}^{n-1}\right)$ samples to learn an $\frac{1}{4n}$-optimal policy.

Theorem D.2 *For any $\delta\in(0,\frac{1}{4e^2}]$ and integer $n\geqslant 2$, there is $N_{n,\delta}\leqslant 2^{\mathcal{O}((1+\delta n)\log^2 n)}$ so that for any $\varepsilon>0$, integer $H, A\geqslant 2$ satisfying*

$$n<H\leqslant\frac{\log(1/\varepsilon)}{40\delta^2}+n,\qquad\varepsilon\leqslant\frac{1}{N_{n,\delta}},\tag{15}$$

there exists a class of δ-strongly separated LMDPs with parameters (L,S,A,H), where

$$L\leqslant N_{n,\delta},\qquad S\leqslant H^{\mathcal{O}((1+\delta n)\log^2 n)},$$

such that any algorithm requires $\Omega\left(A^{n-1}\right)$ samples to learn an ε-optimal policy.

We also present a slightly more general version of Theorem 3.2, as follows.

Theorem D.3 *Suppose that $\delta\in(0,\frac{1}{4e^2}]$, $H\geqslant n+1\geqslant 3$, $A\geqslant 2$, $L\geqslant 2^{C\log n\log(1/\delta)}$ are given such that*

$$CH\log n\log(1/\delta)\leqslant\frac{\log L}{\delta}.\tag{16}$$

Then there exists a class of δ-strongly separated LMDP with L hidden MDPs, $S = (\log L)^{\mathcal{O}(\log n)}$ states, A actions, horizon H, such that any algorithm requires $\Omega\left(A^{n-1}\right)$ samples to learn an $\frac{1}{4n}$-optimal policy with probability at least $\frac{3}{4}$.

Based on the results above, we can now provide a direct proof of Theorem 3.1. In our proof, it turns out that we can take $c_{\mathsf{d}} = \frac{1}{\tilde{\Theta}(\mathsf{d})}$.

Proof of Theorem 3.1. Fix $n = 3\mathsf{d} + 1$, $\delta_0 = \frac{1}{4e^2}$. We proceed to prove Theorem 3.1 by decomposing

$$\log(L/\varepsilon) = \log(L) + \log(1/\varepsilon) \leqslant \frac{1}{2}\max\{\log L, \log(1/\varepsilon)\},$$

and then show that $H_{\rm thre}(L, \varepsilon, \delta)$ must be greater than each of the terms in the maximum above, by applying Theorem D.1, Theorem D.3, and Theorem D.2 separately.

Let $n_1 = \mathcal{O}(n\log^4 n)$ be the LHS of (14), and $N = N_{n,\delta_0} \leqslant 2^{\mathcal{O}(n\log^2 n)}$ be given by Theorem D.2. We choose $L_{\mathsf{d}} := 2^{C_1 n_1 \log^2 n_1}$ for some large absolute constant C_1 so that $L_{\mathsf{d}} \geqslant N$, and set $\varepsilon_{\mathsf{d}} = \frac{1}{N}$, $c_{\mathsf{d}} = \frac{1}{C_1 n_1 \log^2 n}$. In the following, we work with $L \geqslant \max(L_{\mathsf{d}}, \delta^{-1})$, $\varepsilon \leqslant \varepsilon_{\mathsf{d}}$.

Part 1. In this part, we prove the lower bound involving the term $\log L$. We separately consider the case $\delta \leqslant \frac{1}{n_1}$ (Theorem D.1) and $\delta > \frac{1}{n_1}$ (using Theorem D.3).

Case 1: $\delta \leqslant \frac{1}{n_1}$. In this case, we take $H_L = \max\left(\lfloor\frac{\log L}{n_1\delta^2}\rfloor, n_1\right)$. For $H = H_L$ and any $A \geqslant 2$, applying Theorem D.1 gives a class of δ-strongly separated LMDPs with parameters (L, S_1, A, H) where $S_1 \leqslant (\log L)^{\mathcal{O}(\log n)}$, so that any algorithm requires $\Omega\left((A \wedge L)^{n-1}\right)$ samples for learning ε_{d}-optimal policy (because $\varepsilon_{\mathsf{d}} \leqslant \frac{1}{4n}$). However, for $A = L$, we have assumed that $\mathfrak{A}$ succeeds with $\max\left\{S_1, L, H_L, \varepsilon_{\mathsf{d}}^{-1}, \delta^{-1}\right\}^{\mathsf{d}} \leqslant L^{n-1}$ samples. Therefore, since we have assumed that $\mathfrak{A}$ outputs an ε-optimal policy if $H \geqslant H_{\rm thre}(L, \varepsilon, \delta)$, we must have $H_L < H_{\rm thre}(L, \varepsilon, \delta)$.

Case 2: $\delta > \frac{1}{n_1}$. In this case, we take $H_L = \lfloor\frac{\log L}{C_1\log^2(n)\delta}\rfloor$. By definition, $H_L > n$. Hence, for $H = H_L$ and any $A \geqslant 2$, applying Theorem D.3 gives a class of δ-strongly separated LMDPs with parameters (L, S_2, A, H) where $S_2 \leqslant (\log L)^{\mathcal{O}(\log n)}$, so that any algorithm requires $\Omega\left(A^{n-1}\right) = \Omega\left(A^{\mathsf{d}+1}\right)$ samples for learning ε-optimal policy. However, for $A \geqslant \max\left\{L, S_2, H, \varepsilon^{-1}, \delta^{-1}\right\}$, we have assumed that $\mathfrak{A}$ succeeds with A^{d} samples, as long as $H \geqslant H_{\rm thre}(L, \varepsilon, \delta)$. Therefore, we must have $H_L < H_{\rm thre}(L, \varepsilon, \delta)$.

Therefore, in both cases, we have $H_L < H_{\rm thre}(L, \varepsilon, \delta)$. By definition, it always holds that $H_L \geqslant \frac{1}{C_1 n_1 \log^2 n} \cdot \frac{\log L}{\delta^2}$, and the desired result of this part follows.

Part 2. We take $H_\varepsilon = \lfloor\frac{\log(1/\varepsilon)}{9\delta^2}\rfloor + n$. For any $H \leqslant H_\varepsilon, A \geqslant 2$, Theorem D.2 provides a class of δ-strongly separated LMDPs with parameters (L_3, S_3, A, H) with $L_3 = N$ and $S_3 \leqslant H^{\mathcal{O}((1+\delta n)\log^2 n)}$, so that any algorithm requires $\Omega\left(A^{n-1}\right) = \Omega\left(A^{\mathsf{d}+1}\right)$ samples for learning ε-optimal policy. However, for values $A \geqslant \max\left\{N, S_3, H, \varepsilon^{-1}, \delta^{-1}\right\}$, we have assumed that $\mathfrak{A}$ succeeds with A^{d} samples. Therefore, since we have assumed that $\mathfrak{A}$ outputs an ε-optimal policy if $H \geqslant H_{\rm thre}(L, \varepsilon, \delta)$, we must have $H_\varepsilon < H_{\rm thre}(L, \varepsilon, \delta)$.

Combining the two parts above completes the proof of Theorem 3.1. □

In the remaining part of this section, we present the proof of Theorem D.1, Theorem D.2 and Theorem D.3.

Organization In Appendix D.1, we present the hard instances of general (non-separated) LMDP (Kwon et al., 2021b). Then we present our tools of transforming LMDP into separated LMDP in Appendix D.2. The proofs of Theorem 3.2, Theorem D.1 and Theorem D.2 then follow.

Additional notations For any step h, we write $\tau_h = (s_1, a_1, \cdots, s_h, a_h)$ and $\tau_{h:h'} = (s_h, a_h, \cdots, s_{h'}, a_{h'})$. Denote

$$\mathbb{P}_\theta(\tau_h) = \mathbb{P}_\theta(s_{1:h}|\mathrm{do}(a_{1:h-1})), \tag{17}$$

i.e., the probability of observing $s_{1:h}$ if the agent deterministically executes actions $a_{1:h-1}$ in the LMDP M_θ. Also denote $\pi(\tau_h) := \prod_{h' \leqslant h} \pi_{h'}(a_{h'}|\tau_{h'-1}, s_{h'})$, and then $\mathbb{P}^\pi_\theta(\tau_h) = \mathbb{P}_\theta(\tau_h) \times \pi(\tau_h)$ gives the probability of observing τ_h for the first h steps when executing π in LMDP M_θ.

D.1. Lower bound constructions for non-separated LMDPs

In this section, we review a lower bound of Kwon et al. (2021b) on the sample complexity of learning latent MDPs *without* separation constraints; we state and prove some intermediate lemmas regarding this lower bound which are useful later on in our proofs.

Theorem D.4 (Kwon et al. (2021b)) *For $n \geqslant 1$, there exists a class of LMDP with $L = n$, $S = n + 1$, $H = n + 1$, such that any algorithm requires $\Omega\left(A^{n-1}\right)$ samples to learn an $\frac{1}{2n}$-optimal policy.*

In the following, we present the construction in Kwon et al. (2021b) of a family of LMDPs

$$\mathcal{M} = \left\{M_\theta : \theta \in \mathcal{A}^{n-1}\right\} \cup \{M_\varnothing\}. \tag{18}$$

For any $\theta = \mathbf{a} \in \mathcal{A}^{n-1}$, we construct a LMDP M_θ as follows.

- The state space is

$$\mathcal{S}_0 = \{s_\ominus, s_{\oplus,1}, \cdots, s_{\oplus,n}\}.$$

- The action space is $\mathcal{A}$ and the horizon is $H \geqslant n + 1$.
- $L = n$, and for each $m \in [n]$, the MDP $M_{\theta,m}$ has mixing weight $\frac{1}{n}$.
- In the MDP $M_{\theta,m}$, the initial state is $s_{\oplus,1}$, and the state $s_\ominus$ is an absorbing state.

  For $m > 1$, the transition dynamics of $M_{\theta,m}$ is given as follows.

  - At state $s_{\oplus,h}$ with $h < m - 1$, taking any action leads to $s_{\oplus,h+1}$.
  - At state $s_{\oplus,m-1}$, taking action $a \neq \mathbf{a}_{m-1}$ leads to $s_{\oplus,m}$, and taking action $\mathbf{a}_{m-1}$ leads to $s_\ominus$.
  - At state $s_{\oplus,h}$ with $m \leqslant h < n$, taking action $a \neq \mathbf{a}_h$ leads to $s_\ominus$, and taking action $\mathbf{a}_h$ leads to $s_{\oplus,h+1}$.
  - At state $s_{\oplus,n}$, taking any action leads to $s_\ominus$.

The transition dynamics of $M_{\theta,1}$ is given as follows.

- At state $s_{\oplus,h}$ with $h < n$, taking action $a \neq \mathbf{a}_h$ leads to $s_{\ominus}$, and taking action $\mathbf{a}_h$ leads to $s_{\oplus,h+1}$.
- The state $s_{\oplus,n}$ is an absorbing state.

- The reward function is given by $R_h(s,a) = \mathbf{1}\{s = s_{\oplus,n}, h = n+1\}$.

Construction of the reference LMDP For $\bar{\theta} = \varnothing$, we construct a LMDP with state space $\mathcal{S}_0$ and MDP instances $M_{\bar{\theta},1} = \cdots = M_{\bar{\theta},n}$ with mixing weights $\rho = \mathrm{Unif}([n])$, where the initial state is always $s_{\oplus,1}$ and the transition is given by

$$\mathbb{T}_{\bar{\theta},m}(s_{\oplus,h+1}|s_{\oplus,h},a) = \frac{n-h}{n-h+1}, \qquad \mathbb{T}_{\bar{\theta},m}(s_{\ominus}|s_{\oplus,h},a) = \frac{1}{n-h+1}, \qquad \forall h \in [n],$$

and $s_{\ominus}$ is an absorbing state.

Define $\Theta = \mathcal{A}^{n-1} \sqcup \{\bar{\theta}\}$. An important observation is that for any $\theta \in \Theta$, in the LMDP M_θ, any reachable trajectory τ_H must have $s_{1:H}$ belonged to one of the following sequences

$$\begin{aligned}\mathbf{s}_h &= (s_{\oplus,1}, \cdots, s_{\oplus,h}, \underbrace{s_{\ominus}, \cdots, s_{\ominus}}_{H-h}), \quad \text{for some } h \in [n],\\ \text{or } \mathbf{s}_{n,+} &= (s_{\oplus,1}, \cdots, s_{\oplus,n}, \underbrace{s_{\oplus,n}, \cdots, s_{\oplus,n}}_{H-n}).\end{aligned}$$

In particular, for any action sequence $a_{1:H}$, we have

$$\mathbb{P}_{\bar{\theta}}(s_{1:H} = \mathbf{s}_h | a_{1:H}) = \frac{1}{n}, \qquad \forall h \in [n]. \tag{19}$$

We summarize the crucial property of the LMDP class $\{M_\theta\}_{\theta\in\Theta}$ in the following lemma.

Lemma D.5 *For each $\theta = \mathbf{a} \in \mathcal{A}^{n-1}$, the following holds.*

(a) For any action sequence $a_{1:H}$ such that $a_{1:n-1} \neq \mathbf{a}$, it holds

$$\mathbb{P}_\theta(s_{1:H} = \mathbf{s}_h | a_{1:H}) = \frac{1}{n}, \qquad \forall h \in [n]. \tag{20}$$

On the other hand, for the action sequence $a_{1:H}$ such that $a_{1:n-1} = \mathbf{a}$,

$$\mathbb{P}_\theta(s_{1:H} = \mathbf{s}_{n,+} | a_{1:H}) = \frac{1}{n}, \qquad \mathbb{P}_\theta(s_{1:H} = \mathbf{s}_h | a_{1:H}) = \frac{1}{n}, \qquad \forall h \in [n-1]. \tag{21}$$

(b) For any policy π, define

$$w_\theta(\pi) = \prod_{h=1}^{n} \pi(a_h = \mathbf{a}_h | s_{\oplus,1}, \mathbf{a}_1, \cdots, s_{\oplus,h}). \tag{22}$$

Then $\sum_{\theta\in\mathcal{A}^{n-1}} w_\theta(\pi) = 1$, and it also holds that

$$V_\theta(\pi) = \frac{1}{n} w_\theta(\pi), \qquad D_{\mathrm{TV}}\left(\mathbb{P}_\theta^\pi, \mathbb{P}_{\bar{\theta}}^\pi\right) = \frac{1}{n} w_\theta(\pi).$$

In particular, the optimal value in θ is $V_\theta^\star = \frac{1}{n}$, attained by taking $\mathbf{a}$ in the first $n-1$ steps.

Proof. We first prove (a). We inductively prove the following fact.

Fact: For $1 \leqslant h < n$ and any action sequence $a_{1:h}$, there is a unique index $m \in [h]$ such that in the MDP $M_{\theta,m}$, taking action sequence $a_{1:h}$ leads to the trajectory $s_{\oplus,1} \to \cdots \to s_{\oplus,h} \to s_{\ominus}$.

The base case $h = 1$ is obvious. Suppose that the statement holds for all $h' < h$. Then in the MDP $M_{\theta,1}, \cdots, M_{\theta,h}$, there are $h-1$ many MDPs such that taking $a_{1:h-1}$ leads to $s_{\ominus}$ at some step $< h$, and hence there is exactly one index m' such that in $M_{\theta,m'}$, taking $a_{1:h-1}$ leads to the state $s_{\oplus,h}$. Therefore, if $a_h \neq \mathbf{a}_h$, then taking $a_{1:h}$ in $M_{\theta,m'}$ leads to $s_{\oplus,1} \to \cdots \to s_{\oplus,h} \to s_{\ominus}$. Otherwise, we have $a_h = \mathbf{a}_h$, and $a_{1:h}$ in $M_{\theta,h}$ leads to $s_{\oplus,1} \to \cdots \to s_{\oplus,h} \to s_{\ominus}$. The uniqueness is also clear, because for $l > h$, taking $a_{1:h}$ always lead to $s_{\oplus,h+1}$. This completes the proof of the case h.

Now, we consider any given action sequence $a_{1:H}$. For any step $h < n$, there exists a unique index $m(h)$ such that in the MDP $M_{\theta,m(h)}$, taking action sequence $a_{1:n}$ leads to the trajectory $s_{\oplus,1} \to \cdots \to s_{\oplus,h} \to s_{\ominus} \to \cdots$. Thus, there is also a unique index $m(n)$ such that in the MDP $M_{\theta,m(n)}$, taking action sequence $a_{1:n-1}$ leads to the trajectory $s_{\oplus,1} \to \cdots \to s_{\oplus,n}$. Then there are two cases: (1) $a_{1:n-1} \neq \mathbf{a}$, then $m(n) \neq 1$, and hence taking $a_{1:H}$ leads to the trajectory $s_{\oplus,1} \to \cdots \to s_{\oplus,n} \to s_{\ominus} \to \cdots$ in $M_{\theta,m(n)}$. (2) $a_{1:n-1} = \mathbf{a}$, which implies $m(n) = 1$, and hence taking $a_{1:H}$ in $M_{\theta,m(n)}$ leads to the trajectory $s_{\oplus,1} \to \cdots \to s_{\oplus,n} \to s_{\oplus,n} \to \cdots$. This completes the proof of (a).

We next prove (b) using (a). Notice that $V_\theta(\pi) = \mathbb{P}_\theta^\pi(s_{n+1} = s_{\oplus,n})$. By definition, $s_{h+1} = s_{\oplus,n}$ can only happen when the agent is in the MDP $M_{\theta,1}$ and takes actions $a_{1:n} = \mathbf{a}$, and hence

$$\begin{aligned}
\mathbb{P}_\theta^\pi(s_{n+1} = s_{\oplus,n}) &= \mathbb{P}_\theta^\pi(s_1 = s_{\oplus,1}, a_1 = \mathbf{a}_1, \cdots, s_n = s_{\oplus,n}, a_n = \mathbf{a}_n) \\
&= \frac{1}{n}\mathbb{T}_{\theta,1}^\pi(s_1 = s_{\oplus,1}, a_1 = \mathbf{a}_1, \cdots, s_n = s_{\oplus,n}, a_n = \mathbf{a}_n) \\
&= \frac{1}{n}\prod_{h=1}^n \pi(a_h = \mathbf{a}_h | s_{\oplus,1}, \mathbf{a}_1, \cdots, s_{\oplus,h}) = \frac{1}{n} w_\theta(\pi).
\end{aligned}$$

More generally, we have

$$\begin{aligned}
2D_{\mathrm{TV}}\left(\mathbb{P}_\theta^\pi, \mathbb{P}_{\bar\theta}^\pi\right) &= \sum_{\tau_H} \pi(\tau_H) \times |\mathbb{P}_\theta(\tau_H) - \mathbb{P}_{\bar\theta}(\tau_H)| \\
&= \sum_{\tau_H : s_{1:H} = \mathbf{s}_{n,+}, a_{1:n-1} = \mathbf{a}} \pi(\tau_H) \times \left|\frac{1}{n} - 0\right| + \sum_{\tau_H : s_{1:H} = \mathbf{s}_n, a_{1:n-1} = \mathbf{a}} \pi(\tau_H) \times \left|0 - \frac{1}{n}\right| \\
&= \frac{2}{n}\pi(s_{\oplus,1}, \mathbf{a}_1, \cdots, s_{\oplus,n-1}, \mathbf{a}_{n-1}),
\end{aligned}$$

where the second equality is because $\mathbb{P}_\theta(\tau_H) \neq \mathbb{P}_{\bar\theta}(\tau_H)$ only when $s_{1:H} \in \{\mathbf{s}_n, \mathbf{s}_{n,+}\}$ and $a_{1:n-1} = \mathbf{a}$, and the last line follows from recursively applying $\sum_{a_h} \pi(a_h | \tau_{h-1}, s_h) = 1$. This completes the proof of (b). □

D.2. Tools

Definition D.6 *Suppose that $M = (\mathcal{S}, \mathcal{A}, \mathbb{T}, \mu, H)$ is a MDP instance, $\mathcal{O}$ is a finite set, and $\mu \in \Delta(\mathcal{O})$ is a distribution. Then we define $M \otimes \mu$ to be the MDP instance given by $(\mathcal{S} \times \mathcal{O}, \mathcal{A}, \mathbb{T} \otimes \mu, \rho \otimes \mu, H)$, where we define*

$$[\mathbb{T} \otimes \mu]((s', o') | (s, o), a) = \mathbb{T}(s' | s, a) \cdot \mu(o').$$

Given a finite set $\mathcal{O}$, Definition D.7 introduces a property of a collection of distributions $\mu_1, \ldots, \mu_{L'} \in \Delta(\mathcal{O})$ which, roughly speaking, states that the distributions μ_i are *separated* in total variation distance but that certain mixtures of H-wise tensorizations of the distributions μ_i are *close* in total variation distance. Given that such collections of distributions exist, we will "augment" the hard instance of (non-separated) LMDPs from Appendix D.1 with the μ_i (per Definition D.6) to create hard instances of separated LMDPs.

Definition D.7 *A $(L, H, \delta, \gamma, L')$-family over a space $\mathcal{O}$ is a collection of distributions $\{\mu_i\}_{i\in[L']} \subset \Delta(\mathcal{O})$ and $\xi_1, \cdots, \xi_L \in \Delta([L'])$ such that the following holds:*

(1) $\mathrm{supp}(\xi_k) \cap \mathrm{supp}(\xi_l) = \varnothing$ *for all* $k, l \in [L]$ *with* $k \neq l$.

(2) The distribution $\mathbf{Q}_k := \mathbb{E}_{i\sim\xi_k}\left[\mu_i^{\otimes H}\right] \in \Delta(\mathcal{O}^H)$ *satisfies* $D_{\mathrm{TV}}(\mathbf{Q}_k, \mathbf{Q}_1) \leqslant \gamma$ *for all* $k \in [L]$.

(3) $D_{\mathrm{TV}}(\mu_i, \mu_j) \geqslant \delta$ *for all* $i \neq j$, $i, j \in \cup_k \mathrm{supp}(\xi_k)$.

Proposition D.8 and Lemma D.9 state that $(L, H, \delta, \gamma, L')$-families exist, for appropriate settings of the parameters.

Proposition D.8 *Suppose that* $H \geqslant 1$, $\delta \in (0, \frac{1}{4e^2}]$. *Then the following holds:*

(a) Let $d = \lceil 4e^2\delta H \rceil$. *Then there exists a* $(2, H, \delta, 0, N)$*-family over* $[2d]$ *with* $N \leqslant \min\left(\frac{1}{2e\delta}, 2H\right)^d$.

(b) Suppose $\lambda \in [1, \frac{1}{4e^2\delta}]$ *is a real number and* $d \geqslant \lambda\cdot 4e^7\delta^2 H$. *Then there exists is a* $(2, H, \delta, \gamma, N)$*-family over* $[2d]$ *with* $\gamma \leqslant 4e^{-\lambda d}$ *and* $N \leqslant (2e(\lambda+1))^d$.

Lemma D.9 *Suppose that $\mathcal{Q}$ is a $(2, H, \delta, \gamma, L)$-family over a space $\mathcal{O}$. Then there exists a $(2^r, H, \delta, r\gamma, L^r)$ family over space $\mathcal{O}^r$.*

Proofs of the two results above are deferred to Appendices D.6 and D.7.

Definition D.10 (Augmenting an MDP with a family) *Suppose that $M = (\mathcal{S}, \mathcal{A}, (M_m)_{m=1}^L, H, \rho, R)$ is a LMDP instance and $\mathcal{Q} = (\{\mu_i\}_{i\in[L']}, \{\xi_m\}_{m\in[L]})$ is a $(L, H, \delta, \gamma, L')$-family over $\mathcal{O}$. Then $M \otimes \mathcal{Q} = (\mathcal{S}\times\mathcal{O}, \mathcal{A}, (M'_i)_{i=1}^{L'}, H, \rho', \tilde{R})$ is defined to be the following δ-strongly separated LMDP instance:*

- *For each $i \in \cup_{m\in[L]}\mathrm{supp}(\xi_m) \subset [L']$, there is a unique index $m(i) \in [L]$ such that $i \in \mathrm{supp}(\xi_{m(i)})$; we define $M'_i := M_{m(i)} \otimes \mu_i$, with mixing weight $\rho'(i) := \rho_{m(i)} \cdot \xi_{m(i)}(i)$.*
- *The reward function $\tilde{R}$ is given by $\tilde{R}_h((s, o), a) = R_h(s, a)$.*

Proposition D.11 *Suppose that $M_\theta = (\mathcal{S}, \mathcal{A}, (M_{\theta,m})_{m=1}^L, H, \rho, R)$ is a LMDP instance, $\mathcal{Q}$ is a $(L, H, \delta, \gamma, L')$-family over $\mathcal{O}$, so that $M_\theta \otimes \mathcal{Q}$ is a LMDP with state space $\tilde{\mathcal{S}} = \mathcal{S}\times\mathcal{O}$. Let $\Pi_{\mathcal{S}}$ be the set of all H-step policies operating over $\mathcal{S}$, and $\Pi_{\tilde{\mathcal{S}}}$ be the set of all H-step policies operating over $\tilde{\mathcal{S}}$.*

For any policy $\pi \in \Pi_{\mathcal{S}}$, we let $\mathbb{P}^\pi_{\theta,\mathcal{Q}}$ denote the distribution of trajectory under π in the LMDP $M_\theta \otimes \mathcal{Q}$, and we let $V_{\theta,\mathcal{Q}}(\pi)$ denote the value function of π. Then the following statements hold:

(a) We can regard $\Pi_{\mathcal{S}}$ as a subset of $\Pi_{\tilde{\mathcal{S}}}$ naturally, because any policy $\pi \in \Pi_{\mathcal{S}}$ can operate over state space $\tilde{\mathcal{S}} = \mathcal{S}\times\mathcal{O}$ by ignoring the second component of the state $\tilde{s} \in \tilde{\mathcal{S}}$. Then, for any policy $\pi \in \Pi_{\mathcal{S}}$, $V_\theta(\pi) = V_{\theta,\mathcal{Q}}(\pi)$. In particular, $V^\star_\theta \leqslant V^\star_{\theta,\mathcal{Q}}$.

(b) For any policy $\pi \in \Pi_{\tilde{\mathcal{S}}}$, we define $\pi_{\mathcal{Q}} = \mathbb{E}_{o_{1:H}\sim \mathbf{Q}_1}[\pi(\cdot|o_{1:H})] \in \Pi_{\mathcal{S}}$, i.e. $\pi_{\mathcal{Q}}$ is the policy that executes π over state space $\mathcal{S}$ by randomly drawing a sequence $o_{1:H} \sim \mathbf{Q}_1$ at the beginning of each episode. Then we have $|V_{\theta,\mathcal{Q}}(\pi) - V_\theta(\pi_{\mathcal{Q}})| \leqslant \gamma$.

(c) For LMDPs with parameters $\theta, \bar{\theta}$ and any policy $\pi \in \Pi_{\tilde{\mathcal{S}}}$, it holds

$$D_{\mathrm{TV}}\left(\mathbb{P}^\pi_{\theta,\mathcal{Q}}, \mathbb{P}^\pi_{\bar{\theta},\mathcal{Q}}\right) \leqslant 2\gamma + D_{\mathrm{TV}}\left(\mathbb{P}^{\pi_{\mathcal{Q}}}_{\theta}, \mathbb{P}^{\pi_{\mathcal{Q}}}_{\bar{\theta}}\right).$$

Proof. For any $\tilde{s} = (s, o) \in \tilde{\mathcal{S}} = \mathcal{S} \times \mathcal{O}$, we denote $\tilde{s}[1] = s$. Fact (a) follows directly from the definition: for any policy $\pi \in \Pi_{\mathcal{S}}$,

$$V_{\theta,\mathcal{Q}}(\pi) = \mathbb{E}^\pi_{\theta,\mathcal{Q}}\left[\sum_{h=1}^H \tilde{R}_h(\tilde{s}_h, a_h)\right] = \mathbb{E}^\pi_{\theta,\mathcal{Q}}\left[\sum_{h=1}^H R_h(\tilde{s}_h[1], a_h)\right] = \mathbb{E}^\pi_\theta[R_h(s_h, a_h)],$$

where the last equality is because the marginal distribution $\mathbb{P}^\pi_{\theta,\mathcal{Q}}$ over $(\mathcal{S}\times\mathcal{A})^H$ agrees with $\mathbb{P}^\pi_\theta$ by our construction. This completes the proof of (a).

We next prove (b) and (c). In the following, we fix any policy $\pi \in \Pi_{\tilde{\mathcal{S}}}$.

By definition, for any $\tau_H = (\tilde{s}_1, a_1, \cdots, \tilde{s}_H, a_H) \in (\tilde{\mathcal{S}} \times \mathcal{A})^H$, we have $\tilde{s}_h = (s_h, o_h) \in \mathcal{S} \times \mathcal{O}$, and

$$\begin{aligned}
\mathbb{P}^\pi_{\theta,\mathcal{Q}}(\tau_H) &= \sum_{i\in[L']} \rho'(i) \times \mathbb{P}^\pi_{M'_{\theta,i}}(\tau_H) \\
&= \sum_{m\in[L]} \rho(m) \sum_i \xi_m(i) \mathbb{P}^\pi_{M_{\theta,m}\otimes\mu_i}(\tau_H) \\
&= \sum_{m\in[L]} \rho(m) \sum_i \xi_m(i) \times \pi(\tau_H) \times \mathbb{P}_{\theta,m}(s_{1:H}|a_{1:H}) \times \mu_i(o_1)\cdots\mu_i(o_H) \\
&= \sum_{m\in[L]} \rho(m) \times \pi(\tau_H) \times \mathbb{P}_{\theta,m}(s_{1:H}|a_{1:H}) \times \mathbf{Q}_m(o_{1:H}).
\end{aligned}$$

Consider the distribution $\widehat{\mathbb{P}}^\pi_{\theta,\mathcal{Q}} \in \Delta((\tilde{\mathcal{S}} \times \mathcal{A})^H)$ given as follows:

$$\begin{aligned}
\widehat{\mathbb{P}}^\pi_{\theta,\mathcal{Q}}(\tau_H) &= \pi(\tau_H) \times \mathbf{Q}_1(o_{1:H}) \times \mathbb{P}_\theta(s_{1:H}|a_{1:H}) \\
&= \pi(\tau_H) \times \mathbf{Q}_1(o_{1:H}) \times \sum_{m\in[L]} \rho(m)\mathbb{P}_{\theta,m}(s_{1:H}|a_{1:H}).
\end{aligned}$$

Then, by definition,

$$\mathbb{P}^\pi_{\theta,\mathcal{Q}}(\tau_H) - \widehat{\mathbb{P}}^\pi_{\theta,\mathcal{Q}}(\tau_H) = \pi(\tau_H) \times \sum_{m\in[L]} \rho(m)\mathbb{P}_m(s_{1:H}|a_{1:H}) \cdot (\mathbf{Q}_m(o_{1:H}) - \mathbf{Q}_1(o_{1:H})),$$

and hence

$$\begin{aligned}
&D_{\mathrm{TV}}\left(\mathbb{P}^\pi_{\theta,\mathcal{Q}}, \widehat{\mathbb{P}}^\pi_{\theta,\mathcal{Q}}\right) \\
&= \frac{1}{2}\sum_{\tau_H}\left|\mathbb{P}^\pi_{\theta,\mathcal{Q}}(\tau_H) - \widehat{\mathbb{P}}^\pi_{\theta,\mathcal{Q}}(\tau_H)\right|
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \frac{1}{2}\sum_{\tau_H}\pi(\tau_H)\times\sum_{m\in[L]}\rho(m)\mathbb{P}_m(s_{1:H}|a_{1:H})\cdot|\mathbf{Q}_m(o_{1:H})-\mathbf{Q}_1(o_{1:H})| \\
&=\frac{1}{2}\sum_{m\in[L]}\rho(m)\sum_{o_{1:H}}|\mathbf{Q}_m(o_{1:H})-\mathbf{Q}_1(o_{1:H})|\sum_{s_{1:H},a_{1:H}}\pi((s,o)_{1:H},a_{1:H})\times\mathbb{P}_m(s_{1:H}|a_{1:H}) \\
&=\frac{1}{2}\sum_{m\in[L]}\rho(m)\sum_{o_{1:H}}|\mathbf{Q}_m(o_{1:H})-\mathbf{Q}_1(o_{1:H})|\leqslant\gamma,
\end{aligned}
$$

where the last line follows from the fact that for any fixed $o_{1:H}$, $\pi((s,o)_{1:H},a_{1:H})\times\mathbb{P}_m(s_{1:H}|a_{1:H})$ gives a probability distribution over $(s_{1:H},a_{1:H})$.

Let $\widehat{\mathbb{E}}^{\pi}_{\theta,\mathcal{Q}}$ be the expectation taken over $\widehat{\mathbb{P}}^{\pi}_{\theta,\mathcal{Q}}$. Then it holds that

$$
\begin{aligned}
&\widehat{\mathbb{E}}^{\pi}_{\theta,\mathcal{Q}}\left[\sum_{h=1}^{H}\tilde{R}_h(\tilde{s}_h,a_h)\right] \\
&=\sum_{\tau_H}\pi(\tau_H)\times\mathbf{Q}_1(o_{1:H})\times\mathbb{P}_\theta(s_{1:H}|a_{1:H})\times\left(\sum_{h=1}^{H}R_h(s_h,a_h)\right) \\
&=\sum_{s_{1:H},a_{1:H}}\left(\sum_{o_{1:H}}\mathbf{Q}_1(o_{1:H})\cdot\pi(a_{1:H}|s_{1:H},o_{1:H})\right)\times\mathbb{P}_\theta(s_{1:H}|a_{1:H})\times\left(\sum_{h=1}^{H}R_h(s_h,a_h)\right) \\
&=\sum_{s_{1:H},a_{1:H}}\pi_{\mathcal{Q}}(a_{1:H}|s_{1:H})\times\mathbb{P}_\theta(s_{1:H}|a_{1:H})\times\left(\sum_{h=1}^{H}R_h(s_h,a_h)\right)=V_\theta(\pi_{\mathcal{Q}}),
\end{aligned}
$$

where the last line follows from our definition of $\pi_{\mathcal{Q}}$, which is a policy given by

$$
\pi_{\mathcal{Q}}(\cdot)=\mathbb{E}_{o_{1:H}\sim\mathbf{Q}_1}[\pi(\cdot|o_{1:H})].
$$

Therefore, we can bound

$$
|V_{\theta,\mathcal{Q}}(\pi)-V_\theta(\pi_{\mathcal{Q}})|=\left|\mathbb{E}^{\pi}_{\theta,\mathcal{Q}}\left[\sum_{h=1}^{H}\tilde{R}_h(\tilde{s}_h,a_h)\right]-\widehat{\mathbb{E}}^{\pi}_{\theta,\mathcal{Q}}\left[\sum_{h=1}^{H}\tilde{R}_h(\tilde{s}_h,a_h)\right]\right|\leqslant D_{\mathrm{TV}}\left(\mathbb{P}^{\pi}_{\theta,\mathcal{Q}},\widehat{\mathbb{P}}^{\pi}_{\theta,\mathcal{Q}}\right)\leqslant\gamma,
$$

and hence complete the proof of (b).

Similarly, using the fact that $D_{\mathrm{TV}}\left(\mathbb{P}^{\pi}_{\theta,\mathcal{Q}},\widehat{\mathbb{P}}^{\pi}_{\theta,\mathcal{Q}}\right)\leqslant\gamma$ and $D_{\mathrm{TV}}\left(\mathbb{P}^{\pi}_{\bar{\theta},\mathcal{Q}},\widehat{\mathbb{P}}^{\pi}_{\bar{\theta},\mathcal{Q}}\right)\leqslant\gamma$, we have

$$
D_{\mathrm{TV}}\left(\mathbb{P}^{\pi}_{\theta,\mathcal{Q}},\mathbb{P}^{\pi}_{\bar{\theta},\mathcal{Q}}\right)\leqslant 2\gamma+D_{\mathrm{TV}}\left(\widehat{\mathbb{P}}^{\pi}_{\theta,\mathcal{Q}},\widehat{\mathbb{P}}^{\pi}_{\bar{\theta},\mathcal{Q}}\right).
$$

Further, by definition,

$$
\begin{aligned}
D_{\mathrm{TV}}\left(\widehat{\mathbb{P}}^{\pi}_{\theta,\mathcal{Q}},\widehat{\mathbb{P}}^{\pi}_{\bar{\theta},\mathcal{Q}}\right)&=\frac{1}{2}\sum_{\tau_H}\pi(\tau_H)\times\mathbf{Q}_1(o_{1:H})\times|\mathbb{P}_\theta(s_{1:H}|a_{1:H})-\mathbb{P}_{\bar{\theta}}(s_{1:H}|a_{1:H})| \\
&=\frac{1}{2}\sum_{s_{1:H},a_{1:H}}\left(\sum_{o_{1:H}}\mathbf{Q}_1(o_{1:H})\cdot\pi(a_{1:H}|s_{1:H},o_{1:H})\right)\times|\mathbb{P}_\theta(s_{1:H}|a_{1:H})-\mathbb{P}_{\bar{\theta}}(s_{1:H}|a_{1:H})|
\end{aligned}
$$

$$= \frac{1}{2} \sum_{s_{1:H}, a_{1:H}} \pi_{\mathcal{Q}}(a_{1:H} | s_{1:H}) \times |\mathbb{P}_\theta(s_{1:H} | a_{1:H}) - \mathbb{P}_{\bar{\theta}}(s_{1:H} | a_{1:H})|$$
$$= D_{\mathrm{TV}}\left(\mathbb{P}_\theta^{\pi_{\mathcal{Q}}}, \mathbb{P}_{\bar{\theta}}^{\pi_{\mathcal{Q}}}\right).$$

Combining the above two equations completes the proof of (c). □

Fix an action set $\mathcal{A}$ and $n \in \mathbb{N}$. Recall the MDPs M_θ, indexed by $\theta \in \mathcal{A}^{n-1} \cup \{\varnothing\}$, introduced in (18). Proposition D.12 below uses Lemma D.5 to show that when these MDPs are augmented with a $(n, H, \delta, \gamma, L)$-family per Definition D.6, then the resulting family of LMDPs also requires many samples to learn.

Proposition D.12 *Suppose that $n \geqslant 2$, $A \geqslant 2$, $H \geqslant n+1$, $\gamma \in [0, \frac{1}{4n})$, and $\mathcal{Q}$ is a $(n, H, \delta, \gamma, L)$-family over $\mathcal{O}$. Consider*

$$\widetilde{\mathcal{M}} = \left\{M_\theta \otimes \mathcal{Q} : \theta \in \mathcal{A}^{n-1}\right\} \cup \{M_\varnothing \otimes \mathcal{Q}\},$$

which is a class of δ-strongly separated LMDPs with parameters (L, S, A, H), where $S = (n+1)\,|\mathcal{O}|$. Suppose $\mathfrak{A}$ is an algorithm such that for any $M \in \widetilde{\mathcal{M}}$, $\mathfrak{A}$ interacts with M for T episodes and outputs an $\frac{1}{4n}$-optimal policy $\widehat{\pi}$ for M with probability at least $\frac{3}{4}$. Then it holds that

$$T \geqslant \frac{1}{8} \min\left\{\frac{1}{2\gamma}, A^{n-1} - 2\right\}.$$

Proof. In the following, we denote $\bar{\theta} = \varnothing$, consistently with the notations in Appendix D.1.

Notice that by Proposition D.11 (a), for any $\theta \in \mathcal{A}^{n-1}$, we have $V_{\theta,\mathcal{Q}}^\star \geqslant \frac{1}{n}$. Furthermore, for any $\pi \in \Pi_{\tilde{\mathcal{S}}}$,

$$V_{\theta,\mathcal{Q}}(\pi) \leqslant V_\theta(\pi_{\mathcal{Q}}) + \gamma = \frac{1}{n} w_\theta(\pi_{\mathcal{Q}}) + \gamma.$$

In the following, for each $\theta \in \mathcal{A}^{n-1}$, we denote $\widetilde{M}_\theta := M_\theta \otimes \mathcal{Q}$ and $\tilde{w}_\theta(\pi) = w_\theta(\pi_{\mathcal{Q}})$ for any policy $\pi \in \Pi_{\tilde{\mathcal{S}}}$ (recall the definition of $w_\theta(\cdot)$ in (22)). Therefore, using item (b) of Lemma D.5, if π is $\frac{1}{4n}$-optimal in $\widetilde{M}_\theta$, then we have $\tilde{w}_\theta(\pi) \geqslant \frac{3}{4} - n\gamma > \frac{1}{2}$. Also notice that by Proposition D.11 (c) and Lemma D.5 (b),

$$D_{\mathrm{TV}}\left(\mathbb{P}_{\theta,\mathcal{Q}}^{\pi}, \mathbb{P}_{\bar{\theta},\mathcal{Q}}^{\pi}\right) \leqslant 2\gamma + D_{\mathrm{TV}}\left(\mathbb{P}_\theta^{\pi_{\mathcal{Q}}}, \mathbb{P}_{\bar{\theta}}^{\pi_{\mathcal{Q}}}\right) = 2\gamma + \tilde{w}_\theta(\pi). \tag{23}$$

Consider the following set of near-optimal policies in $\widetilde{M}_\theta$:

$$\Pi_\theta^\star := \left\{\pi \in \Pi_{\tilde{\mathcal{S}}} : V_{\theta,\mathcal{Q}}^\star - V_{\theta,\mathcal{Q}}(\pi) \leqslant \frac{1}{4n}\right\} \subseteq \left\{\pi \in \Pi_{\tilde{\mathcal{S}}} : \tilde{w}_\theta(\pi) > \frac{1}{2}\right\}. \tag{24}$$

We know $\mathbb{P}_{\theta,\mathcal{Q}}^{\mathfrak{A}}(\widehat{\pi} \in \Pi_\theta^\star) \geqslant \frac{3}{4}$, where we use $\mathbb{P}_{\theta,\mathcal{Q}}^{\mathfrak{A}}$ to denote the probability distribution induced by executing $\mathfrak{A}$ in the LMDP $\widetilde{M}_\theta$. Using the fact (from Lemma D.5) that $\sum_{\theta \in \mathcal{A}^{n-1}} \tilde{w}_\theta(\pi) = 1$, we also know that $\Pi_\theta^\star \cap \Pi_{\theta'}^\star = \varnothing$ for any $\theta \neq \theta' \in \mathcal{A}^{n-1}$. Therefore,

$$\sum_{\theta \in \mathcal{A}^{n-1}} \mathbb{P}_{\bar{\theta},\mathcal{Q}}^{\mathfrak{A}}(\widehat{\pi} \in \Pi_\theta^\star) \leqslant 1.$$

Hence, there is a set $\Theta_0 \subset \mathcal{A}^{n-1}$ such that $|\Theta_0| \geqslant A^{n-1} - 2$, and for each $\theta \in \Theta_0$, $\mathbb{P}^{\mathfrak{A}}_{\bar{\theta},\mathcal{Q}}(\hat{\pi} \in \Pi^\star_\theta) \leqslant \frac{1}{2}$, which implies that

$$D_{\mathrm{TV}}\left(\mathbb{P}^{\mathfrak{A}}_{\theta,\mathcal{Q}}, \mathbb{P}^{\mathfrak{A}}_{\bar{\theta},\mathcal{Q}}\right) \geqslant \frac{1}{4}, \qquad \forall \theta \in \Theta_0.$$

Now we proceed to upper bound the quantity $D_{\mathrm{TV}}\left(\mathbb{P}^{\mathfrak{A}}_{\theta,\mathcal{Q}}, \mathbb{P}^{\mathfrak{A}}_{\bar{\theta},\mathcal{Q}}\right)$. Notice that the algorithm $\mathfrak{A}$ can be described by interaction rules $\left\{\pi^{(t)}\right\}_{t\in[T]}$, where $\pi^{(t)}$ is a function that maps the history $(\tau^{(1)}, \cdots, \tau^{(t-1)})$ to a policy in Π_{RND} to be executed in the t-th episode. Then, by Lemma A.4, it holds that

$$D_{\mathrm{TV}}\left(\mathbb{P}^{\mathfrak{A}}_{\theta,\mathcal{Q}}, \mathbb{P}^{\mathfrak{A}}_{\bar{\theta},\mathcal{Q}}\right) \leqslant \sum_{t=1}^{T} \mathbb{E}^{\mathfrak{A}}_{\bar{\theta}}\left[D_{\mathrm{TV}}\left(\mathbb{P}^{\pi^{(t)}}_{\theta,\mathcal{Q}}, \mathbb{P}^{\pi^{(t)}}_{\bar{\theta},\mathcal{Q}}\right)\right] = T \cdot \mathbb{E}_{\pi\sim q_{\mathfrak{A}}}\left[D_{\mathrm{TV}}\left(\mathbb{P}^{\pi}_{\theta,\mathcal{Q}}, \mathbb{P}^{\pi}_{\bar{\theta},\mathcal{Q}}\right)\right],$$

where $q_{\mathfrak{A}} \in \Delta(\Pi_{\mathrm{RND}})$ is the distribution of $\pi = \pi^{(t)}$ with $t \in \mathrm{Unif}([T])$ and $(\pi^{(1)}, \cdots, \pi^{(T)}) \sim \mathbb{P}^{\mathfrak{A}}_{\bar{\theta}}$. Therefore, using (23), we know

$$D_{\mathrm{TV}}\left(\mathbb{P}^{\mathfrak{A}}_{\theta,\mathcal{Q}}, \mathbb{P}^{\mathfrak{A}}_{\bar{\theta},\mathcal{Q}}\right) \leqslant 2T\gamma + T \cdot \mathbb{E}_{\pi\sim q_{\mathfrak{A}}} \tilde{w}_\theta(\pi),$$

where the last equality follows from Lemma D.5 (b). Taking summation over $\theta \in \Theta_0$, we obtain

$$|\Theta_0| \cdot 2T\gamma + T \geqslant \sum_{\theta\in\Theta_0} \left(2T\gamma + T \cdot \mathbb{E}_{\pi\sim q_{\mathfrak{A}}} \tilde{w}_\theta(\pi)\right) \geqslant \frac{1}{4} |\Theta_0|.$$

The desired result follows immediately. □

D.3. Proof of Theorem 3.2 and Theorem D.3

Proof of Theorem D.3 Fix a given $n \leqslant H - 1$, we set $r = \lceil \log_2 n \rceil$. By Proposition D.8 (a) and Lemma D.9, there exists a $(n, H, \delta, 0, L_0)$-family over $[2d]^r$, where $d = \lceil 4e^2\delta H \rceil$ and $L_0 \leqslant \left(\frac{1}{2e\delta}\right)^{dr}$. Notice that (3) and $\log L \gtrsim \log n \log(1/\delta)$ together ensure that $L_0 \leqslant L$. Hence, applying Proposition D.12 completes the proof. □

Proof of Theorem 3.2 Notice that for sufficiently large constant C, the presumptions of Theorem 3.2 that $\log L \geqslant C \log^2(1/\delta)$ and (3) together ensure we can apply Theorem D.3 with $n = H - 1$, and hence the proof is completed. □

D.4. Proof of Theorem D.1

Set $\lambda = 2n \log^2 n$. Also set

$$d = \max\left\{\lceil 2\lambda^{-1} n \log L \rceil, \lceil \lambda \cdot 4e^7 H \delta^2 \rceil\right\}. \tag{25}$$

Notice that we have $1 \leqslant \lambda \leqslant \frac{1}{4e^2\delta}$ as long as we choose the absolute constant $C \geqslant 8e^2$ in (14). Then, applying Proposition D.8 (b), there exists a $(2, H, \delta, \gamma, N)$-family over $[2d]$ with

$$N \leqslant (e(\lambda+1))^d, \qquad \gamma \leqslant 4e^{-d\lambda}.$$

Denote $r = \lceil \log_2 n \rceil$. By our assumption (14), we have $\log L \geqslant (c^{-1} \log n)^2$, and hence choosing c sufficiently small and C sufficiently large ensures that we have $N^r \leqslant L$. Further, by our choice of d in (25), we have $r\gamma \leqslant L^{-n}$.

Hence, by Lemma D.9, there exists a $(n, H, \delta, L^{-n}, L)$-family over $[2d]^r$, and we denote it as $\mathcal{Q}$. Applying Proposition D.12 to $\mathcal{Q}$, we obtain a family $\widetilde{\mathcal{M}}$ of δ-strongly separated LMDPs, with state space $\tilde{\mathcal{S}} = \mathcal{S} \times [2d]^r$, and any algorithm requires $\Omega\left(A^n \wedge L^n\right)$ samples to learn $\widetilde{\mathcal{M}}$. Noticing that $|\tilde{\mathcal{S}}| \leqslant (n+1)(2d)^r = (\log L)^{\mathcal{O}(\log n)}$ completes the proof. $\square$

D.5. Proof of Theorem D.2

Let $d_0 = \lceil 4e^2\delta(n+1) \rceil$, $r = \lceil \log_2 n \rceil$, and $\bar{H} = H - n - 1$. By Proposition D.8 and Lemma D.9, there exists a $(n, n+1, \delta, 0, N)$-family over $[2d_0]^r$ with $N \leqslant \min\left(\frac{1}{2e\delta}, 2n\right)^{d_0 r}$. In particular, we choose $N_{n,\delta} = (4nN)^2$, and then it holds that $N_{n,\delta} = 2^{\mathcal{O}((1+\delta n)\log^2 n)}$.

Applying Proposition D.12 to this family, we obtain $\widetilde{\mathcal{M}}$ a class of δ-strongly separated LMDP with state space $\tilde{\mathcal{S}} = \mathcal{S}_0 \times [2d_0]^r$, action space $\mathcal{A}$, horizon $n+1$. Recall that by our construction in Proposition D.12 (and Appendix D.1), for each $\theta \in \mathcal{A}^{n-1} \cup \{\bar{\theta}\}$,$\widetilde{M}_\theta$ is given by $(\tilde{\mathcal{S}}, \mathcal{A}, (\widetilde{M}_{\theta,m})_{m=1}^N, n+1, \rho_\theta, \tilde{R})$, and the mixing weight $\rho_\theta \in \Delta([N])$ of the MDPs $\widetilde{M}_{\theta,1}, \cdots, \widetilde{M}_{\theta,N}$ does not depend on θ, i.e. $\rho_\theta = \rho$ for a fixed $\rho \in \Delta([N])$. Furthermore, for each $m \in [N]$, the initial distribution $\nu_{\theta,m}$ of $\widetilde{M}_{\theta,m}$ is also independent of θ, i.e. $\nu_{\theta,m} = \nu_m$ for a fixed $\nu_m \in \Delta(\tilde{\mathcal{S}})$. We also know that $\tilde{R} = (\tilde{R}_h : \tilde{\mathcal{S}} \times \mathcal{A} \to [0,1])_{h=1}^{n+1}$ is the reward function.

For each θ, we construct an augmented δ-strongly separated LMDP $\widetilde{M}_\theta^+$ with horizon H, as follows.

Fix $d = 2\lceil C_1 \log N \rceil$ for a large absolute constant C_1 so that there exists $\mu_1, \cdots, \mu_N \in \{-1,1\}^d$ such that $\langle \mu_i, \mathbb{1} \rangle = 0 \forall i \in [N]$ and $\|\mu_i - \mu_j\|_1 \geqslant d/2$ (see e.g. Lemma F.4). Denote $\bar{\delta} = 4\delta$ and set $\eta = \frac{1}{2}$.

- The state space is $\tilde{\mathcal{S}}^+ = \tilde{\mathcal{S}} \sqcup \mathcal{S}^+ \sqcup \{\mathsf{terminal}_1, \cdots, \mathsf{terminal}_N\}$, where

$$\mathcal{S}^+ = \left\{(k_1, \cdots, k_d) \in \mathbb{N}^d : k_1 + \cdots + k_d \leqslant \bar{H} - 1\right\}.$$

  We will construct the transition so that at the state outside $\tilde{\mathcal{S}}$, the transition does not depend on θ. We also write $\partial\mathcal{S}^+ = \left\{(k_1, \cdots, k_d) \in \mathbb{N}^d : k_1 + \cdots + k_d = \bar{H} - 1\right\}$.

- The initial state is always $(0, \cdots, 0) \in \mathcal{S}^+$.

- For $s \in \mathcal{S}^+ \backslash \partial\mathcal{S}^+$, we set

$$\mathbb{T}_m(s + \mathbf{e}_i | s, a) = \frac{1 + \bar{\delta}\mu_m[i]}{d}.$$

- For $s \in \partial\mathcal{S}^+$, we define

$$p_m(s) = \prod_{i=1}^d (1 + \bar{\delta}\mu_m[i])^{s[i]},$$

and we set $\bar{p}(s) = \min_{l\in[N]} p_l(s)$,

$$\mathbb{T}_m(s'|s,a) = \eta \frac{\bar{p}(s)}{p_m(s)} \cdot \nu_m(s'), \qquad s' \in \tilde{\mathcal{S}},$$

and $\mathbb{T}_m(\mathsf{terminal}_m|s,a) = 1 - \eta\frac{\bar{p}(s)}{p_m(s)}$.

- For state $s \in \{\mathsf{terminal}_1, \cdots, \mathsf{terminal}_N\}$, we set $\mathbb{T}_m(\mathsf{terminal}_m|s,a) = 1$.
- The reward function is given by $\tilde{R}_h^+ = 0$ for all $h \in [\bar{H}]$, and $\tilde{R}_{\bar{H}+h}^+ = \tilde{R}_h$ for $h \in [n+1]$.

By our construction, it is clear that $\widetilde{M}_\theta^+$ is δ-strongly separated, and $|\tilde{\mathcal{S}}^+| \leqslant n + N + 2 + H^d$.

Furthermore, we can also notice that for any trajectory $\tau_H = (s_{1:H}, a_{1:H})$ such that $s_{\bar{H}+1} \notin \tilde{\mathcal{S}}$, the probability $\mathbb{P}_{\theta,+}(\tau_H) = \mathbb{P}_+(\tau_H)$ does not depend on θ. Furthermore, for any trajectory $\tau_{\bar{H}}$, the probability $\mathbb{P}_{\theta,+}(\tau_H) = \mathbb{P}_+(\tau_H)$ is also independent of θ.

Now, we consider the event $E = \left\{ s_{\bar{H}+1} \in \tilde{\mathcal{S}} \right\}$. Notice that the probability $\mathbb{P}_{\theta,+}(E) = p$ also does not depend on θ.

Lemma D.13 *For any trajectory $\tau_{\bar{H}} = (s_{1:\bar{H}}, a_{1:\bar{H}})$, we have*

$$\mathbb{P}_{\theta,+}(\tau_{\bar{H}+1:H} = \cdot|E, \tau_{\bar{H}}) = \mathbb{P}_{\theta,\mathcal{Q}}(\tau_{1:n+1} = \cdot),$$

which does not depend on τ.

Proof. For any reachable trajectory $\tau_{\bar{H}} = (s_{1:\bar{H}}, a_{1:\bar{H}})$, we have $s_{h+1} = s_h + \mathbf{e}_{i_h}$ for all $h < \bar{H}$. Hence, for $m \in [N]$ and $s \in \tilde{\mathcal{S}}$,

$$\begin{aligned}
\mathbb{P}_{\widetilde{M}_{\theta,m}^+}(\tau_{\bar{H}}, s_{\bar{H}+1} = s) &= \prod_{h=1}^{\bar{H}} \mathbb{T}_m(s_{h+1}|s_h, a_h) \\
&= \mathbb{T}_m(s|s_{\bar{H}}, a_{\bar{H}}) \times \prod_{h=1}^{\bar{H}-1} \frac{1 + \bar{\delta}\mu_m[i_h]}{d} \\
&= \mathbb{T}_m(s|s_{\bar{H}}, a_{\bar{H}}) \times \frac{1}{d^{\bar{H}-1}} \prod_{i=1}^{d} (1 + \bar{\delta}\mu_m[i])^{s_{\bar{H}}[i]} \\
&= \nu_m(s) \times \eta \frac{\bar{p}(s_{\bar{H}})}{p_m(s_{\bar{H}})} \times \frac{p_m(s_{\bar{H}})}{d^{\bar{H}-1}} \\
&= \eta\nu_m(s) \times \frac{\bar{p}(s_{\bar{H}})}{d^{\bar{H}-1}},
\end{aligned}$$

which is independent of θ. Hence, for any $\theta \in \Theta$, we have

$$\widetilde{\mathbb{P}}_{\theta,+}(m^\star = m, s_{\bar{H}+1} = s|E, \tau_{\bar{H}}) = \frac{\rho(m)\mathbb{P}_{\widetilde{M}_{\theta,m}^+}(\tau_{\bar{H}}, s_{\bar{H}+1} = s)}{\sum_{l\in[N]} \sum_{s\in\tilde{\mathcal{S}}} \rho(m)\mathbb{P}_{\widetilde{M}_{\theta,l}^+}(\tau_{\bar{H}}, s_{\bar{H}+1} = s)} = \rho(m)\nu_m(s).$$

In other words, conditional on the event E and any reachable trajectory $\tau_{\bar{H}}$, the posterior distributions of $(m^\star, s_{\bar{H}+1})$ in $\widetilde{M}_\theta^+$ is the same as the distribution of $(m^\star, s_1)$ in $\widetilde{M}_\theta$. Hence, for any trajectory $\tau \in (\tilde{\mathcal{S}} \times \mathcal{A})^{H-\bar{H}}$ that starts with $s \in \tilde{\mathcal{S}}$, we have

$$
\begin{aligned}
&\mathbb{P}_{\theta,+}(\tau_{\bar{H}+1:H} = \tau | E, \tau_{\bar{H}}) \\
=& \sum_{m\in[N]} \widetilde{\mathbb{P}}_{\theta,+}(\tau_{\bar{H}+1:H} = \tau | m^\star = m, s_{\bar{H}+1} = s) \cdot \widetilde{\mathbb{P}}_{\theta,+}(m^\star = m, s_{\bar{H}+1} = s | E, \tau_{\bar{H}}) \\
=& \sum_{m\in[N]} \rho(m)\nu_m(s)\mathbb{P}_{\widetilde{M}_{\theta,m}}(\tau_{\bar{H}+1:H} = \tau | m^\star = m, s_{\bar{H}+1} = s) \\
=& \mathbb{P}_{\theta,\mathcal{Q}}(\tau_{1:n+1} = \tau),
\end{aligned}
$$

where in the second equality we also use the fact that in the MDP $\widetilde{M}_{\theta,m}^+$ and starting at state $s \in \tilde{\mathcal{S}}$, the agent will stay in $\tilde{\mathcal{S}}$, and the transition dynamics of $\widetilde{M}_{\theta,m}^+$ over $\tilde{\mathcal{S}}$ agrees with $\widetilde{M}_{\theta,m}$. This completes the proof of Lemma D.13. □

Using the observations above and Lemma D.13, we know that for any policy $\pi \in \Pi_{\tilde{\mathcal{S}}^+}$, we have

$$
V_{\theta,+}(\pi) = p \cdot \mathbb{E}_{\tau_{\bar{H}-1}|E}\big[V_{\theta,\mathcal{Q}}(\pi(\cdot|\tau_{\bar{H}-1}))\big],
$$

where $\mathbb{P}_{\theta,+}(E) = p$, the expectation is taken over distribution of $\tau_{\bar{H}-1}$ conditional on the event E, and $\pi(\cdot|\tau_{\bar{H}-1})$ is regarded as a policy for the LMDP $\widetilde{M}_\theta$ by conditional on the trajectory $\tau_{\bar{H}-1}$ and restricting to $\tilde{\mathcal{S}}$.

Therefore, for each $\pi \in \Pi_{\tilde{\mathcal{S}}^+}$, there is a corresponding policy $\pi_+ = \mathbb{E}_{\tau_{\bar{H}-1}|E}\big[\pi(\cdot|\tau_{\bar{H}-1})\big] \in \Pi_{\tilde{\mathcal{S}}}$, such that $V_{\theta,+}(\pi) = p \cdot V_{\theta,\mathcal{Q}}(\pi_+) = p\tilde{w}_\theta(\pi)$. Similarly, we can also show that (using (23))

$$
D_{\mathrm{TV}}\left(\mathbb{P}_{\theta,+}^{\pi}, \mathbb{P}_{\bar{\theta},+}^{\pi}\right) = pD_{\mathrm{TV}}\left(\mathbb{P}_{\theta,\mathcal{Q}}^{\pi_+}, \mathbb{P}_{\bar{\theta},\mathcal{Q}}^{\pi_+}\right) \leqslant p\tilde{w}_\theta(\pi_+).
$$

The following lemma provides a lower bound of p (the proof of Lemma D.14 is deferred to the end of this section).

Lemma D.14 *It holds that*

$$
\mathbb{P}_{\theta,+}(E) = p \geqslant \frac{\eta}{N}(1 - \bar{\delta}^2)^{\bar{H}-1}.
$$

In particular, $p > 2n\varepsilon$.

With the preparations above, we now provide the proof of Theorem D.2, whose argument is analogous to the proof of Proposition D.12.

Proof of Theorem D.2 Suppose that $\mathfrak{A}$ is an algorithm such that for any $M \in \widetilde{\mathcal{M}}$, $\mathfrak{A}$ interacts with M for T episodes and outputs an $\frac{1}{4n}$-optimal policy $\widehat{\pi}$ for M with probability at least $\frac{3}{4}$.

Notice that $V_{\theta,\mathcal{Q}}^\star = \frac{p}{n}$, and $\varepsilon < \frac{p}{2n}$. Thus, if $\widehat{\pi}$ is $\frac{p}{4n}$-optimal in $\widetilde{M}_\theta^+$, then $\tilde{w}_\theta(\pi) > \frac{1}{2}$. Now, consider the following set of near-optimal policies in $\widetilde{M}_\theta^+$:

$$
\Pi_{\theta,+}^\star := \left\{\pi \in \Pi_{\tilde{\mathcal{S}}^+} : \pi \text{ is } \varepsilon\text{-optimal in } \widetilde{M}_\theta^+\right\}. \tag{26}
$$

Then $\Pi^\star_{\theta,+}$ are mutually disjoint for $\theta \in \mathcal{A}^{n-1}$. We then have

$$\mathbb{P}^{\mathfrak{A}}_{\theta,+}(\widehat{\pi} \in \Pi^\star_{\theta,+}) \geqslant \frac{3}{4}, \qquad \sum_{\theta \in \mathcal{A}^{n-1}} \mathbb{P}^{\mathfrak{A}}_{\bar{\theta},+}(\widehat{\pi} \in \Pi^\star_{\theta}) \leqslant 1.$$

Repeating the argument as in the proof of Proposition D.12 gives $T \geqslant \frac{1}{4p}(A^{n-1}-2)$, and the desired result follows. □

Proof of Lemma D.14. We next lower bound the probability p. By definition,

$$\begin{aligned}
\mathbb{P}_{\theta,+}(s_{\bar{H}+1} \in \tilde{\mathcal{S}}) &= \sum_{\tau_{\bar{H}} \text{ reachable}, s_{\bar{H}+1} \in \tilde{\mathcal{S}}} \mathbb{P}_{\theta,+}(\tau_{\bar{H}}, s_{\bar{H}+1}) \\
&= \sum_{\tau_{\bar{H}} \text{ reachable}, s_{\bar{H}+1} \in \tilde{\mathcal{S}}} \sum_{m \in [N]} \rho(m) \mathbb{P}_{\widetilde{M}^+_{\theta,m}}(\tau_{\bar{H}}, s_{\bar{H}+1} = s) \\
&= \sum_{\tau_{\bar{H}} \text{ reachable}} \eta \cdot \frac{\bar{p}(s_{\bar{H}})}{d^{\bar{H}-1}} \\
&= \sum_{i_1,\cdots,i_{\bar{H}-1} \in [d]} \frac{\eta}{d^{\bar{H}-1}} \cdot \bar{p}\left(\mathbf{e}_{i_1} + \cdots + \mathbf{e}_{i_{\bar{H}-1}}\right) \\
&\geqslant \frac{\eta}{d^{\bar{H}-1}} \left(\sum_{i_1,\cdots,i_{\bar{H}-1} \in [d]} \frac{1}{\bar{p}\left(\mathbf{e}_{i_1} + \cdots + \mathbf{e}_{i_{\bar{H}-1}}\right)} \right)^{-1},
\end{aligned}$$

where in the last line we apply Cauchy inequality. Notice that for any $s \in \partial\mathcal{S}^+$,

$$\frac{1}{\bar{p}(s)} = \max_{l \in [N]} \frac{1}{p_l(s)} \leqslant \sum_{l \in [N]} \frac{1}{p_l(s)},$$

and we also have

$$\begin{aligned}
\sum_{i_1,\cdots,i_{\bar{H}-1} \in [d]} \frac{1}{p_m\left(\mathbf{e}_{i_1} + \cdots + \mathbf{e}_{i_{\bar{H}-1}}\right)} &= \sum_{i_1,\cdots,i_{\bar{H}-1} \in [d]} \frac{1}{\prod_{h=1}^{\bar{H}-1}(1+\bar{\delta}\mu_m[i_h])} = \left(\sum_i \frac{1}{1+\bar{\delta}\mu_m[i]} \right)^{\bar{H}-1} \\
&= \left(\frac{d}{2} \times \frac{1}{1+\bar{\delta}} + \frac{d}{2} \times \frac{1}{1-\bar{\delta}} \right)^{\bar{H}-1} = \frac{d^{\bar{H}-1}}{(1-\bar{\delta}^2)^{\bar{H}-1}},
\end{aligned}$$

where the second line follows from the fact that $\mu_m \in \{-1,1\}^d$ and $\langle \mu_m, \mathbb{1} \rangle = 0$. Combining the inequalities above gives $p \geqslant \frac{\eta}{N}(1-\bar{\delta}^2)^{\bar{H}-1}$.

In particular, to prove $p > 2n\varepsilon$, we only need to prove $(\bar{H}-1)\log\frac{1}{1-\bar{\delta}^2} \leqslant \log(1/(4Nn\varepsilon))$. Notice that $\log\frac{1}{1-\bar{\delta}^2} \leqslant \frac{\bar{\delta}^2}{1-\bar{\delta}^2}$, $\bar{\delta} = 4\delta$, and we also have $\frac{1}{4nN\varepsilon} \geqslant \frac{1}{\sqrt{\varepsilon}}$ using $\varepsilon \leqslant \frac{1}{N_{n,\delta}} = \frac{1}{(4nN)^2}$. Combining these completes the proof. □

D.6. Proof of Proposition D.8

Towards proving Proposition D.8, we first prove the following proposition, which provides a simple approach of bounding TV distance between mixtures of distributions of a special form.

Proposition D.15 *Let $n, d \in \mathbb{N}$ be given. For $\mathbf{x} \in [-1, 1]^d$, we consider the distribution*

$$\mathbb{Q}_{\mathbf{x}} = \left[\frac{1+\mathbf{x}[1]}{2d}; \frac{1-\mathbf{x}[1]}{2d}; \cdots ; \frac{1+\mathbf{x}[d]}{2d}; \frac{1-\mathbf{x}[d]}{2d}\right] \in \Delta([2d]). \tag{27}$$

Then, for distributions μ, ν over $[-1, 1]^d$, it holds that

$$D_{\mathrm{TV}}\left(\mathbb{E}_{\mathbf{x}\sim\mu}[\mathbb{Q}_{\mathbf{x}}^{\otimes n}], \mathbb{E}_{\mathbf{y}\sim\nu}[\mathbb{Q}_{\mathbf{y}}^{\otimes n}]\right)^2 \leqslant \frac{1}{4}\sum_{\ell=0}^{n}\binom{n}{\ell}\cdot\frac{1}{d^\ell}\left\|\mathbf{\Delta}_\ell\right\|_2^2,$$

where we denote

$$\mathbf{\Delta}_\ell := \mathbb{E}_{\mathbf{x}\sim\mu}\left[\mathbf{x}^{\otimes\ell}\right] - \mathbb{E}_{\mathbf{y}\sim\nu}\left[\mathbf{y}^{\otimes\ell}\right] \in \mathbb{R}^{d^\ell}.$$

Proof. We utilize the idea of the orthogonal polynomials (see e.g. Han (2019)) to simplify our calculation. For simplicity, we denote $\mathcal{O} = [2d]$. By definition, for any $\mathbf{o} = (o_1, \cdots, o_n) \in \mathcal{O}^n$, we have

$$\frac{\mathbb{Q}_{\mathbf{x}}^{\otimes n}(\mathbf{o})}{\mathbb{Q}_{\mathbf{0}}^{\otimes n}(\mathbf{o})} = \prod_{j=1}^{n}\frac{\mathbb{Q}_{\mathbf{x}}(o_j)}{\mathbb{Q}_{\mathbf{0}}(o_j)} = \sum_{\boldsymbol{k}\in\mathbb{N}^d} c_{n,\boldsymbol{k}}(\mathbf{o})\mathbf{x}^{\boldsymbol{k}},$$

where for $\boldsymbol{k} = (k_1, \cdots, k_d) \in \mathbb{N}^d$ we denote $|\boldsymbol{k}| = k_1 + \cdots + k_d$, $\mathbf{x}^{\boldsymbol{k}} = \mathbf{x}[1]^{k_1}\cdots\mathbf{x}[d]^{k_d}$, and $c_{n,\boldsymbol{k}} : \mathcal{O}^n \to \mathbb{R}$ are coefficients satisfying $c_{n,\boldsymbol{k}}(\mathbf{o}) = 0$ for all $|\boldsymbol{k}| > n$. Notice that for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\sum_{\mathbf{o}\in\mathcal{O}^n}\frac{\mathbb{Q}_{\mathbf{x}}^{\otimes n}(\mathbf{o})\mathbb{Q}_{\mathbf{y}}^{\otimes n}(\mathbf{o})}{\mathbb{Q}_{\mathbf{0}}^{\otimes n}(\mathbf{o})} = \sum_{o_1,\cdots,o_n\in\mathcal{O}}\prod_{j=1}^{n}\frac{\mathbb{Q}_{\mathbf{x}}(o_j)\mathbb{Q}_{\mathbf{y}}(o_j)}{\mathbb{Q}_{\mathbf{0}}(o_j)} = \left(\sum_{o\in\mathcal{O}}\frac{\mathbb{Q}_{\mathbf{x}}(o)\mathbb{Q}_{\mathbf{y}}(o)}{\mathbb{Q}_{\mathbf{0}}(o)}\right)^n = \left(1 + \frac{\langle x, y\rangle}{d}\right)^n.$$

On the other hand, it also holds (where the expectation $\mathbb{E}_{\mathbf{0}}$ is taken over $\mathbf{o} \sim \mathbb{Q}_{\mathbf{0}}$)

$$\begin{aligned}
\sum_{\mathbf{o}\in\mathcal{O}^n}\frac{\mathbb{Q}_{\mathbf{x}}^{\otimes n}(\mathbf{o})\mathbb{Q}_{\mathbf{y}}^{\otimes n}(\mathbf{o})}{\mathbb{Q}_{\mathbf{0}}^{\otimes n}(\mathbf{o})} &= \mathbb{E}_{\mathbf{0}}\left[\frac{\mathbb{Q}_{\mathbf{x}}^{\otimes n}(\mathbf{o})}{\mathbb{Q}_{\mathbf{0}}^{\otimes n}(\mathbf{o})}\cdot\frac{\mathbb{Q}_{\mathbf{y}}^{\otimes n}(\mathbf{o})}{\mathbb{Q}_{\mathbf{0}}^{\otimes n}(\mathbf{o})}\right] \\
&= \mathbb{E}_{\mathbf{0}}\left[\sum_{\boldsymbol{k}\in\mathbb{N}^d} c_{n,\boldsymbol{k}}(\mathbf{o})\mathbf{x}^{\boldsymbol{k}}\sum_{\boldsymbol{j}\in\mathbb{N}^d} c_{n,\boldsymbol{j}}(\mathbf{o})\mathbf{y}^{\boldsymbol{j}}\right] \\
&= \sum_{\boldsymbol{k},\boldsymbol{j}\in\mathbb{N}^d}\mathbb{E}_{\mathbf{0}}[c_{n,\boldsymbol{k}}(\mathbf{o})c_{n,\boldsymbol{j}}(\mathbf{o})]\cdot\mathbf{x}^{\boldsymbol{k}}\mathbf{y}^{\boldsymbol{j}}.
\end{aligned}$$

Therefore, by comparing the coefficients between the two sides of

$$\left(1 + \frac{\langle x, y\rangle}{d}\right)^n = \sum_{\boldsymbol{k},\boldsymbol{j}\in\mathbb{N}^d}\mathbb{E}_{\mathbf{0}}[c_{n,\boldsymbol{k}}(\mathbf{o})c_{n,\boldsymbol{j}}(\mathbf{o})]\cdot\mathbf{x}^{\boldsymbol{k}}\mathbf{y}^{\boldsymbol{j}},$$

we have

$$\mathbb{E}_{\mathbf{0}}[c_{n,\boldsymbol{k}}(\mathbf{o})c_{n,\boldsymbol{j}}(\mathbf{o})] = \begin{cases} 0, & \boldsymbol{k} \neq \boldsymbol{j}, \\ \binom{n}{|\boldsymbol{k}|}\frac{N_{\boldsymbol{k}}}{d^{|\boldsymbol{k}|}}, & \boldsymbol{k} = \boldsymbol{j}, \end{cases}$$

where for $\boldsymbol{k} = (k_1, \cdots, k_d)$ such that $|\boldsymbol{k}| = \ell$, $N_{\boldsymbol{k}} = \binom{\ell}{k_1, \cdots, k_d}$. Now, we can express

$$\begin{aligned}
2D_{\mathrm{TV}}\left(\mathbb{E}_{\mathbf{x}\sim\mu}\big[\mathbb{Q}_{\mathbf{x}}^{\otimes n}\big], \mathbb{E}_{\mathbf{y}\sim\nu}\big[\mathbb{Q}_{\mathbf{y}}^{\otimes n}\big]\right) &= \mathbb{E}_{\mathbf{0}}\left|\mathbb{E}_{\mathbf{x}\sim\mu}\left[\frac{\mathbb{Q}_{\mathbf{x}}^{\otimes n}(\mathbf{o})}{\mathbb{Q}_{\mathbf{0}}^{\otimes n}(\mathbf{o})}\right] - \mathbb{E}_{\mathbf{y}\sim\mu}\left[\frac{\mathbb{Q}_{\mathbf{y}}^{\otimes n}(\mathbf{o})}{\mathbb{Q}_{\mathbf{0}}^{\otimes n}(\mathbf{o})}\right]\right| \\
&= \mathbb{E}_{\mathbf{0}}\left|\mathbb{E}_{\mathbf{x}\sim\mu}\left[\sum_{\boldsymbol{k}\in\mathbb{N}^d} c_{n,\boldsymbol{k}}(\mathbf{o})\mathbf{x}^{\boldsymbol{k}}\right] - \mathbb{E}_{\mathbf{y}\sim\nu}\left[\sum_{\boldsymbol{k}\in\mathbb{N}^d} c_{n,\boldsymbol{k}}(\mathbf{o})\mathbf{y}^{\boldsymbol{k}}\right]\right| \\
&= \mathbb{E}_{\mathbf{0}}\left|\sum_{\boldsymbol{k}\in\mathbb{N}^d} c_{n,\boldsymbol{k}}(\mathbf{o})\Delta_{\boldsymbol{k}}\right|,
\end{aligned}$$

where in the last line we abbreviate $\Delta_{\boldsymbol{k}} = \mathbb{E}_{\mathbf{x}\sim\mu}\big[\mathbf{x}^{\boldsymbol{k}}\big] - \mathbb{E}_{\mathbf{y}\sim\nu}\big[\mathbf{y}^{\boldsymbol{k}}\big]$ for $\boldsymbol{k} \in \mathbb{N}^d$. By Jensen inequality,

$$\begin{aligned}
4D_{\mathrm{TV}}\left(\mathbb{E}_{\mathbf{x}\sim\mu}\big[\mathbb{Q}_{\mathbf{x}}^{\otimes n}\big], \mathbb{E}_{\mathbf{y}\sim\nu}\big[\mathbb{Q}_{\mathbf{y}}^{\otimes n}\big]\right)^2 &\leqslant \mathbb{E}_{\mathbf{0}}\left|\sum_{\boldsymbol{k}\in\mathbb{N}^d} c_{n,\boldsymbol{k}}(\mathbf{o})\Delta_{\boldsymbol{k}}\right|^2 \\
&= \mathbb{E}_{\mathbf{0}}\left[\sum_{\boldsymbol{k}\in\mathbb{N}^d} c_{n,\boldsymbol{k}}(\mathbf{o})\Delta_{\boldsymbol{k}} \sum_{\boldsymbol{j}\in\mathbb{N}^d} c_{n,\boldsymbol{j}}(\mathbf{o})\Delta_{\boldsymbol{j}}\right] \\
&= \sum_{\boldsymbol{k},\boldsymbol{j}\in\mathbb{N}^d} \mathbb{E}_{\mathbf{0}}[c_{n,\boldsymbol{k}}(\mathbf{o})c_{n,\boldsymbol{j}}(\mathbf{o})] \cdot \Delta_{\boldsymbol{k}}\Delta_{\boldsymbol{j}} \\
&= \sum_{\boldsymbol{k}\in\mathbb{N}^d} \binom{n}{|\boldsymbol{k}|}\frac{N_{\boldsymbol{k}}}{d^{|\boldsymbol{k}|}}\Delta_{\boldsymbol{k}}^2 \\
&= \sum_{\ell=0}^{n} \binom{n}{\ell}\frac{1}{d^\ell} \sum_{\boldsymbol{k}\in\mathbb{N}^d:|\boldsymbol{k}|=\ell} N_{\boldsymbol{k}}\Delta_{\boldsymbol{k}}^2 \\
&= \sum_{\ell=0}^{n} \binom{n}{\ell}\frac{1}{d^\ell}\left\|\boldsymbol{\Delta}_\ell\right\|_2^2,
\end{aligned}$$

where the last equality follows directly from definition:

$$\begin{aligned}
\sum_{\boldsymbol{k}\in\mathbb{N}^d:|\boldsymbol{k}|=\ell} N_{\boldsymbol{k}}\Delta_{\boldsymbol{k}}^2 &= \sum_{\boldsymbol{k}\in\mathbb{N}^d:|\boldsymbol{k}|=\ell} N_{\boldsymbol{k}}\left|\mathbb{E}_{\mathbf{x}\sim\mu}\Big[\mathbf{x}^{\boldsymbol{k}}\Big] - \mathbb{E}_{\mathbf{y}\sim\nu}\Big[\mathbf{y}^{\boldsymbol{k}}\Big]\right|^2 \\
&= \sum_{i_1,\cdots,i_\ell\in[d]^\ell} \left|\mathbb{E}_{\mathbf{x}\sim\mu}[\mathbf{x}[i_1]\cdots\mathbf{x}[i_\ell]] - \mathbb{E}_{\mathbf{y}\sim\nu}[\mathbf{y}[i_1]\cdots\mathbf{y}[i_\ell]]\right|^2 \\
&= \left\|\mathbb{E}_{\mathbf{x}\sim\mu}\Big[\mathbf{x}^{\otimes\ell}\Big] - \mathbb{E}_{\mathbf{y}\sim\nu}\Big[\mathbf{y}^{\otimes\ell}\Big]\right\|_2^2.
\end{aligned}$$

$\square$

Corollary D.16 *Let $d, N, K, H \in \mathbb{N}$ and $\delta \in (0, 1]$ be given so that $N \geqslant \binom{K+d-1}{d} + 1$. Suppose $\mathbf{x}_1, \cdots, \mathbf{x}_N \in [-\delta, \delta]^d$. Then there exist two distributions $\xi_0, \xi_1 \in \Delta([N])$, such that* $\mathrm{supp}(\xi_0) \cap \mathrm{supp}(\xi_1) = \varnothing$ *and*

$$D_{\mathrm{TV}}^2\left(\mathbb{E}_{i\sim\xi_0}\big[\mathbb{Q}_{\mathbf{x}_i}^{\otimes H}\big], \mathbb{E}_{i\sim\xi_1}\big[\mathbb{Q}_{\mathbf{x}_i}^{\otimes H}\big]\right) \leqslant \sum_{k=K}^{H}\left(\frac{eH\delta^2}{K}\right)^k.$$

Proof. Consider the following system of equations:

$$\sum_{i=1}^{N} v_i \mathbf{x}_i[1]^{k_1} \cdots \mathbf{x}_i[d]^{k_d} = 0, \qquad \forall k_j \geqslant 0, k_1 + \cdots + k_d \leqslant K - 1.$$

There are exactly $\binom{K+d-1}{d}$ equations, and hence such a system must have a non-zero solution $v^\star \in \mathbb{R}^N$. Notice that $\sum_{i=1}^N v_i^\star = 0$, and we then take $\xi_0 = [v^\star]_+/V$, $\xi_1 = [-v^\star]_+/V \in \Delta([N])$, where $V = \|[v^\star]_+\|_1 = \|[-v^\star]_+\|_1$ is the normalizing factor. Clearly, $\mathrm{supp}(\xi_0) \cap \mathrm{supp}(\xi_1) = \varnothing$, and we also have

$$\mathbb{E}_{i\sim\xi_0} \mathbf{x}_i^{\otimes \ell} = \mathbb{E}_{i\sim\xi_1} \mathbf{x}_i^{\otimes \ell}, \qquad \forall \ell = 0, \cdots K-1.$$

Consider $\mathbf{\Delta}_\ell := \mathbb{E}_{i\sim\xi_0} \mathbf{x}_i^{\otimes \ell} - \mathbb{E}_{i\sim\xi_1} \mathbf{x}_i^{\otimes \ell}$; then we have $\mathbf{\Delta}_\ell = 0$ for $\ell < K$, and we also have

$$\|\mathbf{\Delta}_\ell\|_2 \leqslant 2 \max_i \left\| \mathbf{x}_i^{\otimes \ell} \right\|_2 \leqslant 2 \|\mathbf{x}_i\|_2^\ell \leqslant 2(\sqrt{d}\delta)^\ell, \qquad \forall \ell \geqslant 0.$$

This implies that $\frac{1}{d^\ell} \|\mathbf{\Delta}_\ell\|_2^2 \leqslant 4\delta^{2\ell}$ always holds. Therefore, applying Proposition D.15 with $n = H$ and using the fact that $\binom{H}{k} \leqslant \left(\frac{eH}{k}\right)^k$, we obtain

$$D_{\mathrm{TV}}^2 \left(\mathbb{E}_{i\sim\xi_0} [\mathbb{Q}_{\mathbf{x}_i}^{\otimes H}], \mathbb{E}_{i\sim\xi_1} [\mathbb{Q}_{\mathbf{x}_i}^{\otimes H}] \right) \leqslant \sum_{k=K}^{H} \left(\frac{eH}{k} \right)^k \cdot (\delta)^{2k} \leqslant \sum_{k=K}^{H} \left(\frac{eH\delta^2}{K} \right)^k.$$

□

Proof of Proposition D.8 Choose $\delta_\infty > 0$, $d \geqslant 1$, and an integer $K \leqslant \left(\frac{\delta_\infty}{2e^2\delta} - 1\right)d + 1$ (to be specified later in the proof). For the ℓ_∞-ball $\mathbf{B} := [-\delta_\infty, \delta_\infty]^d$, we consider its packing number under the ℓ_1-norm, denoted $M(\cdot; \mathbf{B}, \|\cdot\|_1)$. Using Wainwright (2019, Lemma 5.5 & 5.7), we have

$$M(\delta_1; \mathbf{B}, \|\cdot\|_1) \geqslant \left(\frac{1}{\delta_1}\right)^d \frac{\mathrm{vol}(\mathbf{B})}{\mathrm{vol}(\mathbf{B}')}, \qquad \forall \delta_1 > 0,$$

where $\mathbf{B}' = \left\{ x \in \mathbb{R}^d : \|x\|_1 \leqslant 1 \right\}$ is the ℓ_1 unit ball. Notice that $\mathrm{vol}(\mathbf{B}) = (2\delta_\infty)^d$, $\mathrm{vol}(\mathbf{B}') = \frac{2^d}{d!}$. Thus, using the fact $d! > (d/e)^d$, we have

$$M(\delta_1; \mathbf{B}, \|\cdot\|_1) \geqslant d! \left(\frac{\delta_\infty}{\delta_1}\right)^d > \left(\frac{d\delta_\infty}{e\delta_1}\right)^d$$

In particular, $M := M(2d\delta; \mathbf{B}, \|\cdot\|_1) > \left(\frac{\delta_\infty}{2e\delta}\right)^d$. Notice that our choice of K ensures that for $N = \binom{K+d-1}{d} + 1$, it holds that $N \leqslant M$. Therefore, we can pick N vectors $\mathbf{x}_1, \cdots, \mathbf{x}_N \in \mathbf{B}$ such that $\|\mathbf{x}_i - \mathbf{x}_j\|_1 \geqslant 2d\delta$.

Consider the distributions $\mu_i = \mathbb{Q}_{\mathbf{x}_i} \in \Delta([2d])$ for each $i \in [N]$. Clearly, we have $D_{\mathrm{TV}}(\mu_i, \mu_j) \geqslant \delta$ for $i \neq j$. Also, by Corollary D.16, there exists $\xi_0, \xi_1 \in \Delta([N])$ such that $\mathrm{supp}(\xi_0) \cap \mathrm{supp}(\xi_1) = \varnothing$,

$$D_{\mathrm{TV}}^2 \left(\mathbb{E}_{i\sim\xi_0} \left[\mu_i^{\otimes H} \right], \mathbb{E}_{i\sim\xi_0} \left[\mu_i^{\otimes H} \right] \right) \leqslant \sum_{k=K}^{H} \left(\frac{eH\delta_\infty^2}{K} \right)^k.$$

Consider $\mathcal{Q} = \{(\mu_1, \cdots, \mu_N), (\xi_0, \xi_1)\}$.

Proof of Proposition D.8 (a). In this case, we pick $\delta_\infty = 1$, $K = H + 1$, $d = \lceil 4e^2\delta H \rceil$. Then $\mathcal{Q}$ is a $(2, H, \delta, 0, N)$-family over $[2d]$, with $N \leqslant \min\left(\frac{1}{2e\delta}, 2H\right)^d$.

Proof of Proposition D.8 (b). In this case, we take $K = \lceil \lambda d \rceil$, $\delta_\infty = 2e^2\delta(\lambda + 1)$, so $\frac{eH\delta_\infty^2}{K} \leqslant e^{-2}$ and hence $\mathcal{Q}$ is a $(2, H, \delta, \gamma, N)$-family over $[2d]$ with $\gamma \leqslant 2e^{-\lambda d}$ and $N \leqslant (2e(\lambda + 1))^d$. □

D.7. Proof of Lemma D.9

Suppose that $\mathcal{Q} = \{(\mu_1, \cdots, \mu_N), (\xi_0, \xi_1)\}$ is a $(2, H, \delta, \gamma, N)$-family over $\mathcal{O}$. Then, for each integer $m \in \{0, 1, \cdots, 2^r - 1\}$, we consider its binary representation $m = (m_r \cdots m_1)_2$, and define

$$\tilde{\xi}_m = \xi_{m_r} \otimes \cdots \otimes \xi_{m_1} \in [N]^r.$$

Further, for each $\boldsymbol{k} = (k_1, \cdots, k_r) \in [N]^r$, we define

$$\tilde{\mu}_{\boldsymbol{k}} = \mu_{k_1} \otimes \cdots \otimes \mu_{k_r} \in \mathcal{O}^r.$$

Under the definitions above, we know

$$\mathbb{E}_{\boldsymbol{k}\sim\tilde{\xi}_m}\left[\tilde{\mu}_{\boldsymbol{k}}^{\otimes H}\right] = \mathbb{E}_{k_1\sim\xi_{m_1}}\left[\mu_{k_1}^{\otimes H}\right] \otimes \cdots \otimes \mathbb{E}_{k_r\sim\xi_{m_r}}\left[\mu_{k_r}^{\otimes H}\right],$$

and hence for $0 \leqslant m, l \leqslant 2^r - 1$, it holds that

$$D_{\mathrm{TV}}\left(\mathbb{E}_{\boldsymbol{k}\sim\tilde{\xi}_m}\left[\tilde{\mu}_{\boldsymbol{k}}^{\otimes H}\right], \mathbb{E}_{\boldsymbol{k}\sim\tilde{\xi}_l}\left[\tilde{\mu}_{\boldsymbol{k}}^{\otimes H}\right]\right) \leqslant \sum_{i=1}^{r} D_{\mathrm{TV}}\left(\mathbb{E}_{k\sim\xi_{m_i}}\left[\mu_k^{\otimes H}\right], \mathbb{E}_{k\sim\xi_{l_i}}\left[\mu_k^{\otimes H}\right]\right) \leqslant r\gamma.$$

We also know that $\mathrm{supp}(\tilde{\xi}_m) \cap \mathrm{supp}(\tilde{\xi}_l) = \varnothing$ as long as $m \neq l$. For $\boldsymbol{k}, \boldsymbol{j} \in \cup_m \mathrm{supp}(\tilde{\xi}_m)$ such that $\boldsymbol{k} \neq \boldsymbol{j}$, it also holds that

$$D_{\mathrm{TV}}\left(\tilde{\mu}_{\boldsymbol{k}}, \tilde{\mu}_{\boldsymbol{j}}\right) \geqslant \max_{1\leqslant i\leqslant r} D_{\mathrm{TV}}\left(\mu_{k_i}, \mu_{j_i}\right) \geqslant \delta.$$

Therefore, $\mathbf{Q}' = \left\{(\tilde{\mu}_{\boldsymbol{k}})_{\boldsymbol{k}\in[N]^r}, (\tilde{\xi}_0, \cdots, \tilde{\xi}_{2^r-1})\right\}$ is indeed a $(2^r, H, \delta, r\gamma, N^r)$-family over $\mathcal{O}^r$. □

D.8. Proof of Theorem 3.4

In this section, we modify the constructions in Appendix D.1 to obtain a class of hard instances of N-step decodable LMDPs

$$\mathcal{M}^+ = \left\{M_\theta^+ : \theta \in \mathcal{A}^{n-1}\right\} \cup \left\{M_\varnothing^+\right\}, \tag{28}$$

and then sketch the proof of Theorem 3.4 (as most parts of the proof follow immediately from Appendix D.1 and Proposition D.12).

For any given integer N, n, A, we set $k = N - n$ so that $H = n + 2k$, and we take $\mathcal{A} = [A]$. We specify the state space, action space and reward function (which are shared across all LMDP instances) as follows.

- The state space is

$$\mathcal{S} = \{s_{\oplus,i} : -k+1 \leqslant i \leqslant n+k\} \bigsqcup \{s_{\ominus,i} : 2 \leqslant i \leqslant n+k\} \bigsqcup \{\mathsf{terminal}_1, \cdots, \mathsf{terminal}_n\}.$$

- The action space is $\mathcal{A}$.
- The reward function is given by $R_h(s,a) = \mathbf{1}\{s = s_{\oplus,n}, h = n+k+1\}$.

We remark that, our below construction has (essentially) the same LMDP dynamics at the state $s \in \mathcal{S}_+ := \{s_{\oplus,1}, \cdots, s_{\oplus,n}\}$, as the construction in Appendix D.1. The auxiliary states $s_{\ominus,2}, \cdots, s_{\ominus,n+k}, \mathsf{terminal}_1, \cdots, \mathsf{tern}$ are introduced so that we can ensure N-step decodability, while the auxiliary states $s_{\oplus,-k+1}, \cdots, s_{\oplus,0}$ are introduced to so that we can take the horizon H to equal $N+k$.

Construction of the LMDP M_θ^+ For any $\theta = \mathbf{a} \in \mathcal{A}^{n-1}$, we construct a LMDP M_θ^+ as follows.

- $L = n$, the MDP instances of M_θ^+ is given by $M_{\theta,1}^+, \cdots, M_{\theta,n}^+$ with mixing weight $\rho = \mathrm{Unif}([n])$.
- For each $m \in [n]$, in the MDP $M_{\theta,m}^+$, the initial state is $s_{\oplus,-k+1}$, and the transition dynamics at state $s \notin \mathcal{S}_+ = \{s_{\oplus,1}, \cdots, s_{\oplus,n}\}$ is specified as follows and does not depend on θ:
  - At state $s_{\oplus,h}$ with $h \leqslant 0$, taking any action leads to $s_{\oplus,h+1}$.
  - At state $s_{\ominus,h}$ with $h < n+k$, taking any action leads to $s_{\ominus,h+1}$.
  - At state $s \in \{s_{\ominus,n+k}, \mathsf{terminal}_1, \cdots, \mathsf{terminal}_n\}$, taking any action leads to $\mathsf{terminal}_m$.

  For $m > 1$, the transition dynamics of $M_{\theta,m}^+$ at state $s \in \mathcal{S}_+$ is given as follows (similar to Appendix D.1).
  - At state $s_{\oplus,h}$ with $h < m$, taking any action leads to $s_{\oplus,h+1}$.
  - At state $s_{\oplus,m-1}$, taking action $a \neq \mathbf{a}_{m-1}$ leads to $s_{\oplus,m}$, and taking action $\mathbf{a}_{m-1}$ leads to $s_{\ominus,m}$.
  - At state $s_{\oplus,h}$ with $m \leqslant h < n$, taking action $a \neq \mathbf{a}_h$ leads to $s_\ominus$, and taking action $\mathbf{a}_h$ leads to $s_{\oplus,h+1}$.
  - At state $s_{\oplus,n}$, taking any action leads to $s_{\ominus,n+1}$.

  The transition dynamics of $M_{\theta,1}^+$ at state $s \in \mathcal{S}_+$ is given as follows.
  - At state $s_{\oplus,h}$ with $h < n$, taking action $a \neq \mathbf{a}_h$ leads to $s_\ominus$, and taking action $\mathbf{a}_h$ leads to $s_{\oplus,h+1}$.
  - The state $s_{\oplus,n}$ is an absorbing state.

Construction of the reference LMDP For $\bar{\theta} = \varnothing$, we construct the LMDP $M_{\bar{\theta}}$ with state space $\mathcal{S}$, MDP instances $M_{\bar{\theta},1}, \cdots, M_{\bar{\theta},n}$, mixing weights $\rho = \mathrm{Unif}([n])$, where for each $m \in [n]$, the transition dynamics of $M_{\bar{\theta},m}$ is specified as follows: (1) the initial state is always $s_{\oplus,-k+1}$, (2) the transition dynamics at state $s \notin \mathcal{S}_+$ agrees with the transition dynamics of $M_{\theta,m}$ described as above, (3) at state $s_{\oplus,h}$ with $h < m$, taking any action leads to $s_{\oplus,h+1}$, and (4) at state $s_{\oplus,h}$ with $h \geqslant m$, taking any action leads to $s_{\ominus,h+1}$.

Sketch of proof The following are several key observations for the LMDP M_θ ($\theta \in \mathcal{A}^{n-1} \sqcup \{\bar{\theta}\}$).

(1) At state $s \in \mathcal{S}_+$, the transition dynamics of $M^+_{\theta,m}$ agrees with the transition dynamics of $M_{\theta,m}$ (defined in Appendix D.1), in the sense that we identify the state $s_\ominus$ there as the set of $\{s_{\ominus,2}, \cdots, s_{\ominus,n+k}\}$.

(2) With horizon $H = n+2k$, we always have $s_H \in \{s_{\oplus,n}, s_{\ominus,n+k}\}$, and all the states in $\{\mathsf{terminal}_1, \cdots, \mathsf{terminal}_n\}$ are not reachable. In other words, the auxiliary states $\mathsf{terminal}_1, \cdots, \mathsf{terminal}_n$ (introduced for ensuring N-step decodability) do not reveal information of the latent index because they are never reached.

(3) M_θ is N-step decodable, because:

(3a) M_θ is N-step decodable when we start at $s \in \{s_{\ominus,2}, \cdots, s_{\ominus,n+k}, \mathsf{terminal}_1, \cdots, \mathsf{terminal}_n\}$. This follows immediately from definition, because in M_θ, any reachable trajectory $\overline{\tau}_N$ starting at such state s must end with $s_N = \mathsf{terminal}_m$, where m is the index of the MDP instance $M_{\theta,m}$. Similar argument also shows that M_θ is N-step decodable when we start at $s \in \{s_{\oplus,2}, \cdots, s_{\oplus,n}\}$.

(3b) M_θ is n-step decodable when we start at $s_{\oplus,1}$. This follows immediately from our proof of Lemma D.5 (a), which shows that for any reachable trajectory $\overline{\tau}_n$, there is a unique latent index m such that $\overline{\tau}_n$ is reachable under $M_{\theta,n}$. Therefore, we also know that M_θ is N-step decodable when we start at $s \in \{s_{\oplus,-k+1}, \cdots, s_{\oplus,0}\}$.

Given the above observations, we also know that our argument in the proof of Proposition D.12 indeed applies to $\mathcal{M}^+$, which concludes that the class $\mathcal{M}^+$ of N-step decodable LMDPs requires $\Omega\left(A^{n-1}\right)$ samples to learn. $\square$

Appendix E. Proofs for Section 4

Miscellaneous notations We identify $\Pi_{\mathrm{RND}} = \Delta(\Pi_{\mathrm{RND}})$ as both the set of all policies and all distributions over policies interchangeably.

Also, recall that for any step h, we write $\tau_h = (s_1, a_1, \cdots, s_h, a_h)$, and $\tau_{h:h'} = (s_h, a_h, \cdots, s_{h'}, a_{h'})$ compactly. Also recall that

$$\mathbb{P}_\theta(\tau_h) = \mathbb{P}_\theta(s_{1:h}|\mathrm{do}(a_{1:h-1})),$$

i.e., $\mathbb{P}_\theta(\tau_h)$ is the probability of observing $s_{1:h}$ if the agent deterministically executes actions $a_{1:h-1}$ in the LMDP M_θ. Also denote $\pi(\tau_h) := \prod_{h' \leqslant h} \pi_{h'}(a_{h'}|\tau_{h'-1}, s_{h'})$, and then $\mathbb{P}^\pi_\theta(\tau_h) = \mathbb{P}_\theta(\tau_h) \times \pi(\tau_h)$ gives the probability of observing τ_h for the first h steps when executing π in LMDP M_θ.

For any policy $\pi, \pi' \in \Pi$ and step $h \in [H]$, we define $\pi \circ_h \pi'$ to be the policy that executes π for the first $h-1$ steps, and then starts executing π_{sep} at step h (i.e. discarding the history τ_{h-1}).

To avoid confusion, we define $\mathbb{P}_\theta(\tau_{h:H}|\tau_{h-1}, \pi)$ to be the probability of observing $\tau_{h:H}$ conditional on the history τ_{h-1} if we start executing π at the step h (i.e. π does not use the history data τ_{h-1}). By contrast, consistently with the standard notation of conditional probability, $\mathbb{P}^\pi_\theta(\tau_{h:H}|\tau_{h-1})$ is the conditional probability of the model $\mathbb{P}^\pi_\theta$, i.e. the probability of observing $\tau_{h:H}$ conditional on the history τ_{h-1} under policy π. Therefore, we have

$$\mathbb{P}^\pi_\theta(\tau_{h:H}|\tau_{h-1}) = \mathbb{P}_\theta(\tau_{h:H}|\tau_{h-1}, \pi(\cdot|\tau_{h-1})). \tag{29}$$

E.1. Details of Algorithm OMLE

Given a *separating policy* π_{sep}, we can construct a corresponding map $\mathsf{p}(\cdot) : \Pi_{\mathrm{RND}} \to \Pi_{\mathrm{RND}}$, that transforms any policy π to an explorative version of it. The definition of $\mathsf{p}(\cdot)$ below is similar to the choice of the explorative policies for learning PSRs in Zhan et al. (2022); Chen et al. (2022a); Liu et al. (2022b).

Definition E.1 *Suppose that $\pi_{\mathrm{sep}} \in \Pi_{\mathrm{RND}}$ is a given policy and $1 \leqslant W \leqslant H$. For any step $1 \leqslant h \leqslant H$, we define $\varphi_h : \Pi_{\mathrm{RND}} \to \Pi_{\mathrm{RND}}$ to be a policy modification given by*

$$\varphi_h(\pi) = \pi \circ_h \mathrm{Unif}(\mathcal{A}) \circ_{h+1} \pi_{\mathrm{sep}}, \qquad \pi \in \Pi_{\mathrm{RND}},$$

i.e. $\varphi_h(\pi)$ means that we follow π for the first $h-1$ steps, take $\mathrm{Unif}(\mathcal{A})$ at step h, and start executing π_{sep} afterwards.

Further, we define $\phi(\cdot)$, $\mathsf{p}(\cdot)$ as follows:

$$\phi(\pi) = \pi \circ_W \pi_{\mathrm{sep}}, \qquad \mathsf{p}(\pi) = \frac{1}{2}\phi(\pi) + \frac{1}{2H}\sum_{h=0}^{H-1} \varphi_h(\pi).$$

The following guarantee pertaining to the confidence set maintained in OMLE is taken from Chen et al. (2022a, Proposition E.2). There is a slight difference in the policy modification applied to π^t, which does not affect the argument in Chen et al. (2022a, Appendix E.1).

Proposition E.2 (Confidence set guarantee) *Suppose that we choose $\beta \geqslant 2\log N_\Theta(1/T) + 2\log(1/p) + 2$ in Algorithm 1. Then with probability at least $1-p$, the following holds:*

(a) For all $k \in [K]$, $\theta^\star \in \Theta^k$;

(b) For all $k \in [K]$ and any $\theta \in \Theta^k$, it holds that

$$\sum_{t=1}^{k-1} D_{\mathrm{H}}^2\left(\mathbb{P}_\theta^{\mathsf{p}(\pi^t)}, \mathbb{P}_{\theta^\star}^{\mathsf{p}(\pi^t)}\right) \leqslant 2\beta. \tag{30}$$

Let E_0 be the event that both (a) and (b) of Proposition E.2 above hold true. In the following, we will analyze the performance of Algorithm 1 conditional on the suceess event E_0.

The following proposition relates the sub-optimality of the output policy $\widehat{\pi}$ of Algorithm 1 to the error of estimation.

Proposition E.3 *Suppose that Assumption 4.6 holds, and $W \geqslant \varpi^{-1}(\log(L/\varepsilon_{\mathrm{s}}))$. Conditional on the success event E_0, we have*

$$V_\star - V_{\theta^\star}(\widehat{\pi}) \leqslant \frac{1}{K}\sum_{k=1}^{K} D_{\mathrm{TV}}\left(\mathbb{P}_{\theta^k}^{\pi^k}, \mathbb{P}_{\theta^\star}^{\pi^k}\right).$$

Proof. Under the given condition on W, it holds $e_{\theta^\star,W}(\pi_\star) \leqslant \varepsilon_{\mathrm{s}}$ (Proposition 4.1). By Proposition E.2 (a), we also have $\theta^\star \in \Theta^k$ for each $k \in [K]$. Therefore, by the choice of (θ^k, π^k) in Algorithm 1, it holds that $V_\star = V_{\theta^\star}(\pi_\star) \leqslant V_{\theta^k}(\pi^k)$. Hence,

$$V_\star - V_{\theta^\star}(\pi^k) \leqslant V_{\theta^k}(\pi^k) - V_{\theta^\star}(\pi^k) \leqslant D_{\mathrm{TV}}\left(\mathbb{P}_{\theta^k}^{\pi^k}, \mathbb{P}_{\theta^\star}^{\pi^k}\right),$$

where the last inequality follows from the definition of TV distance and the fact that $\sum_{h=1}^{H} R_h(s_h, a_h) \in [0,1]$ for any trajectory. Taking average over $k \in [K]$ completes the proof. □

E.2. Proof overview

Given Proposition E.2 and Proposition E.3, upper bounding the sub-optimality of the output $\widehat{\pi}$ reduces to the following task.

$$\text{Task: upper bound } \sum_{k=1}^{K} D_{\mathrm{TV}}\left(\mathbb{P}_{\theta^k}^{\pi^k}, \mathbb{P}_{\theta^\star}^{\pi^k}\right), \text{ given that } \forall k \in [K], \sum_{t=1}^{k-1} D_{\mathrm{H}}^2\left(\mathbb{P}_{\theta^k}^{\mathsf{p}(\pi^t)}, \mathbb{P}_{\theta^\star}^{\mathsf{p}(\pi^t)}\right) \leqslant 2\beta.$$

A typical strategy, used in Liu et al. (2022a); Chen et al. (2022b,a); Liu et al. (2023), of relating these two terms is three-fold: (1) find a decomposition of the TV distance, i.e. an upper bound of $D_{\mathrm{TV}}\left(\mathbb{P}_\theta^\pi, \mathbb{P}_{\theta^\star}^\pi\right)$; (2) show that the decomposition can be upper bounded by the squared Hellinger distance $D_{\mathrm{H}}^2\left(\mathbb{P}_\theta^\pi, \mathbb{P}_{\theta^\star}^\pi\right)$; (3) apply an eluder argument on the decomposition to complete the proof.

For example, we describe this strategy for the special case of MDPs.

Example E.4 *Suppose that Θ is instead a class of MDPs and $\mathsf{p}(\pi) = \pi$, then we can decompose*

$$D_{\mathrm{TV}}\left(\mathbb{P}_\theta^\pi, \mathbb{P}_{\theta^\star}^\pi\right) \leqslant \underbrace{\sum_{h=1}^{H-1} \mathbb{E}_{\theta^\star}^\pi D_{\mathrm{TV}}\left(\mathbb{T}_\theta(\cdot|s_h, a_h), \mathbb{T}_{\theta^\star}(\cdot|s_h, a_h)\right)}_{=:G_{\theta^\star}(\pi,\theta)} \leqslant 2H D_{\mathrm{TV}}\left(\mathbb{P}_\theta^\pi, \mathbb{P}_{\theta^\star}^\pi\right). \tag{31}$$

In tabular case, the decomposition $G_{\theta^\star}(\cdot,\cdot)$ can be written as an inner product over $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, i.e. $G_{\theta^\star}(\pi,\theta) = \langle X(\theta), W(\pi)\rangle$ for appropriate embeddings $X(\theta), W(\pi) \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. Then, using the eluder argument for linear functionals (i.e. the "elliptical potential lemma", Lattimore and Szepesvári (2020)), we can prove that under (30), it holds that $\sum_k D_{\mathrm{TV}}\left(\mathbb{P}_{\theta^k}^{\pi^k}, \mathbb{P}_{\theta^\star}^{\pi^k}\right) \leqslant \widetilde{O}(\sqrt{SA \cdot KH^2\beta})$.

More generally, beyond the tabular case, we can also apply a coverability argument (see e.g. Xie et al. (2022) and also Proposition A.9) as follows. Suppose that $\mathrm{rank}(\mathbb{T}_{\theta^\star}) \leqslant d$. We can then invoke Proposition A.10 to show that $G_{\theta^\star}$ admits the following representation:

$$G_{\theta^\star}(\pi,\theta) = \mathbb{E}_{x\sim p(\pi)} f_\theta(x),$$

where $p : \Pi \to \Delta(\mathcal{S}\times\mathcal{A})$ is such that there exists $\mu \in \Delta(\mathcal{S}\times\mathcal{A})$, $\|p(\pi)/\mu\|_\infty \leqslant d \cdot A$ for all π. Hence, Proposition A.9 implies that $\sum_k D_{\mathrm{TV}}\left(\mathbb{P}_{\theta^k}^{\pi^k}, \mathbb{P}_{\theta^\star}^{\pi^k}\right) \leqslant \widetilde{O}(\sqrt{dA \cdot KH^2\beta})$. ◊

Analyzing the separated LMDPs In our analysis, we first decompose the TV distance between LMDPs into two parts:

$$\begin{aligned} D_{\mathrm{TV}}\left(\mathbb{P}_\theta^\pi, \mathbb{P}_{\theta^\star}^\pi\right) \leqslant{} & D_{\mathrm{TV}}\left(\mathbb{P}_\theta^\pi(\overline{\tau}_W = \cdot), \mathbb{P}_{\theta^\star}^\pi(\overline{\tau}_W = \cdot)\right) \\ & + \mathbb{E}_{\theta^\star}^\pi\left[D_{\mathrm{TV}}\left(\mathbb{P}_\theta^\pi(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W), \mathbb{P}_{\theta^\star}^\pi(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W)\right)\right] \end{aligned} \tag{32}$$

where the part (a) is the TV distance between the distribution of trajectory up to step W, and part (b) is the TV distance between the conditional distribution of the last $H - W + 1$ steps trajectory. We analyze part (a) and part (b) separately.

Part (a) Under the assumption of ϖ-separation under $\pi_{\rm sep}$ and $H - W \geqslant \varpi^{-1}(\log(2L))$, we can show that a variant of the revealing condition (Liu et al., 2022a; Chen et al., 2022a; Liu et al., 2023) holds (Lemma A.7). Therefore, restricting to dynamics of the first W steps, we can regard Θ as a class of revealing POMDPs, and then apply the eluder argument developed in Chen et al. (2022a). More specifically, our analysis of part (a) relies on the following result, which is almost an immediately corollary of the analysis in Chen et al. (2022a, Appendix D & E).

Theorem E.5 *Suppose that for all $\theta \in \Theta$, θ is ϖ-separated under $\pi_{\rm sep}$, and $H-W \geqslant \varpi^{-1}(\log(2L))$. Then conditional on the success event E_0,*

$$\sum_{k=1}^{K} D_{\rm TV}\left(\mathbb{P}_{\theta^k}^{\phi(\pi^k)}, \mathbb{P}_{\theta^\star}^{\phi(\pi^k)}\right) \lesssim \sqrt{LdAH^2\iota_K \cdot K\beta},$$

where $\iota_K = \log(LdH \cdot K/(A\beta))$ is a logarithmic factor.

We provide a more detailed discussion of Theorem E.5 and a simplified proof in Appendix E.6. Notice that, although the statement of Theorem E.5 bounds the total variation distance between the entire (H-step) trajectories $\mathbb{P}_{\theta^k}^{\phi(\pi^k)}$ and $\mathbb{P}_{\theta^\star}^{\phi(\pi^k)}$, the policies $\phi(\pi^k)$ act according to the *fixed* policy $\pi_{\rm sep}$ on steps $h \geqslant W$. Thus, Theorem E.5 is not establishing that the model $\theta^\star$ is being learned in any meaningful way after step W (indeed, it cannot since we may not have $H - h \geqslant \varpi^{-1}(\log(2L))$ for $h > W$). To learn the true model $\theta^\star$ at steps $h \geqslant W$, we need to analyze part (b) of (32).

Part (b) The main idea for analyzing the steps $h \geqslant W$ is that, given $e_\theta(\pi)$ is small, we can regard

$$\mathbb{P}_\theta^\pi(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W) \approx \mathbb{M}^\theta_{m_\theta(\overline{\tau}_W), H-W+1}(\pi(\cdot|\tau_{W-1}), s_W). \tag{33}$$

In other words, conditional on the first W steps, the dynamics of the trajectory $\overline{\tau}_{W:H}$ is close to the dynamics of the MDP $M_{\theta,m_\theta(\overline{\tau}_W)}$. Therefore, we can decompose part (b) in a fashion similar to the decomposition (31) for MDP (Proposition E.7), and then apply the eluder argument of Proposition A.9 (see Corollary E.9).

E.3. Structural properties of separated LMDP

In this section, we formalize the idea described in the part (b) of our proof overview.

For each $h \in [H]$ and trajectory $\overline{\tau}_h$, we define the belief state of the trajectory $\overline{\tau}_h$ under model θ as

$$\mathbf{b}_\theta(\overline{\tau}_h) = \left[\widetilde{\mathbb{P}}_\theta(m|\overline{\tau}_h)\right]_{m\in[L]} \in \Delta([L]). \tag{34}$$

Recall the definition of $\mathbb{M}_{m,h}(\cdot) \in \Delta((\mathcal{A}\times\mathcal{S})^{h-1})$ in (1). Then, conditional on the trajectory $\overline{\tau}_W$, the distribution of $\overline{\tau}_{W:H} = (a_W, \cdots, a_{H-1}, s_H)$ under policy π can be written as

$$\begin{aligned}\mathbb{P}_\theta^\pi(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W) &= \mathbb{E}_{m\sim\mathbf{b}_\theta(\overline{\tau}_W)}\left[\mathbb{T}^\pi_{\theta,m}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W)\right] \\ &= \mathbb{E}_{m\sim\mathbf{b}_\theta(\overline{\tau}_W)}\left[\mathbb{M}^\theta_{m,H-W+1}(\pi|_{\tau_{W-1}}, s_W)\right]\end{aligned} \tag{35}$$

where $\pi|_{\tau_{W-1}} = \pi(\cdot|\tau_{W-1})$ is the policy obtained from π by conditional on $\overline{\tau}_W$. In particular,

$$D_{\rm TV}\left(\mathbb{P}_\theta^\pi(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W), \mathbb{M}^\theta_{m_\theta(\overline{\tau}_W),H-W+1}(\pi|_{\tau_{W-1}}, s_W)\right) \leqslant \sum_{m\neq m_\theta(\overline{\tau}_W)} \mathbf{b}_\theta(\overline{\tau}_W)[m]. \tag{36}$$

We denote

$$e_\theta(\overline{\tau}_W) := \sum_{m \neq m_\theta(\overline{\tau}_W)} \mathbf{b}_\theta(\overline{\tau}_W)[m]. \tag{37}$$

Notice that by the definition of $\mathbf{b}_\theta(\overline{\tau}_W)$,

$$e_\theta(\overline{\tau}_W) = \sum_{m \neq m_\theta(\overline{\tau}_W)} \mathbf{b}_\theta(\overline{\tau}_W)[m] = 1 - \max_m \mathbf{b}_\theta(\overline{\tau}_W)[m] = \widetilde{\mathbb{P}}_\theta(m \neq m_\theta(\overline{\tau}_W)|\overline{\tau}_W), \tag{38}$$

and hence $e_{\theta,W}(\pi) = \mathbb{E}_{\bar{\theta}}^{\pi}[e_\theta(\overline{\tau}_W)]$.

In the following, we denote $\bar{W} := H - W + 1$, and we will use the inequality

$$D_{\mathrm{TV}}\left(\mathbb{P}_{\theta}^{\pi}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W), \mathbb{M}^{\theta}_{m_\theta(\overline{\tau}_W),\bar{W}}(\pi|_{\tau_{W-1}}, s_W)\right) \leqslant e_\theta(\overline{\tau}_W), \tag{39}$$

(which follows from Eqs. (36) and (38)) and the fact that $e_{\theta,W}(\pi) = \mathbb{E}_{\bar{\theta}}^{\pi}[e_\theta(\overline{\tau}_W)]$ repeatedly. This formalizes the idea of (33). Also notice that $\phi(\pi) = \pi \circ_W \pi_{\mathrm{sep}}$, and hence we also have

$$D_{\mathrm{TV}}\left(\mathbb{P}_{\theta}^{\phi(\pi)}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W), \mathbb{M}^{\theta}_{m_\theta(\overline{\tau}_W),\bar{W}}(\pi_{\mathrm{sep}}, s_W)\right) \leqslant e_\theta(\overline{\tau}_W). \tag{40}$$

The following proposition shows that, as long as the model θ is close to $\bar{\theta}$, there is a correspondence between the maps m_θ and $m_{\bar{\theta}}$.

Proposition E.6 *Suppose that θ and $\bar{\theta}$ are ϖ-separated under π_{sep} and $\bar{W} = H - W + 1 \geqslant \varpi^{-1}(1)$. Then there exists a map $\sigma = \sigma_{\theta;\bar{\theta}} : [L] \times \mathcal{S} \to [L]$ such that for any $(W-1)$-step policy π,*

$$\mathbb{P}_{\bar{\theta}}^{\pi}(m_\theta(\overline{\tau}_W) \neq \sigma(m_{\bar{\theta}}(\overline{\tau}_W), s_W)) \leqslant 288 D_{\mathrm{H}}^2\left(\mathbb{P}_{\theta}^{\phi(\pi)}, \mathbb{P}_{\bar{\theta}}^{\phi(\pi)}\right) + 144 e_{\theta,W}(\pi) + 144 e_{\bar{\theta},W}(\pi), \tag{41}$$

where $\phi(\pi) = \pi \circ_W \pi_{\mathrm{sep}}$ is defined in Definition E.1.

Proof. In the following proof, we abbreviate $\varepsilon = D_{\mathrm{H}}^2\left(\mathbb{P}_{\theta}^{\phi(\pi)}, \mathbb{P}_{\bar{\theta}}^{\phi(\pi)}\right)$. By Lemma A.5,

$$\mathbb{E}_{\bar{\theta}}^{\pi}\left[D_{\mathrm{TV}}^2\left(\mathbb{P}_{\theta}^{\phi(\pi)}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W), \mathbb{P}_{\bar{\theta}}^{\phi(\pi)}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W)\right)\right] \leqslant 4\varepsilon. \tag{42}$$

Using (40) and the triangle inequality of TV distance, we have

$$\begin{aligned}
&D_{\mathrm{TV}}\left(\mathbb{M}^{\theta}_{m_\theta(\overline{\tau}_W),\bar{W}}(\pi_{\mathrm{sep}}, s_W), \mathbb{M}^{\bar{\theta}}_{m_{\bar{\theta}}(\overline{\tau}_W),\bar{W}}(\pi_{\mathrm{sep}}, s_W)\right) \\
\leqslant\ & D_{\mathrm{TV}}\left(\mathbb{P}_{\theta}^{\phi(\pi)}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W), \mathbb{P}_{\bar{\theta}}^{\phi(\pi)}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W)\right) + e_\theta(\overline{\tau}_W) + e_{\bar{\theta}}(\overline{\tau}_W),
\end{aligned}$$

and hence

$$\begin{aligned}
&\mathbb{E}_{\bar{\theta}}^{\pi}\left[D_{\mathrm{TV}}^2\left(\mathbb{M}^{\theta}_{m_\theta(\overline{\tau}_W),\bar{W}}(\pi_{\mathrm{sep}}, s_W), \mathbb{M}^{\bar{\theta}}_{m_{\bar{\theta}}(\overline{\tau}_W),\bar{W}}(\pi_{\mathrm{sep}}, s_W)\right)\right] \\
\leqslant\ & 3\mathbb{E}_{\bar{\theta}}^{\pi}\left[D_{\mathrm{TV}}^2\left(\mathbb{P}_{\theta}^{\phi(\pi)}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W), \mathbb{P}_{\bar{\theta}}^{\phi(\pi)}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W)\right)\right] + 3\mathbb{E}_{\bar{\theta}}^{\pi}[e_\theta(\overline{\tau}_W)] + 3\mathbb{E}_{\bar{\theta}}^{\pi}[e_{\bar{\theta}}(\overline{\tau}_W)].
\end{aligned} \tag{43}$$

By definition, we know $\mathbb{E}^{\pi}_{\bar\theta}[e_{\bar\theta}(\overline{\tau}_W)] = e_{\bar\theta,W}(\pi)$, and by Lemma A.3, we also have

$$\begin{aligned}\mathbb{E}^{\pi}_{\bar\theta}[e_{\theta}(\overline{\tau}_W)] &\leqslant 3\mathbb{E}^{\pi}_{\theta}[e_{\theta}(\overline{\tau}_W)] + 2D^2_{\mathrm{H}}\left(\mathbb{P}^{\pi}_{\theta}(\overline{\tau}_W = \cdot), \mathbb{P}^{\pi}_{\bar\theta}(\overline{\tau}_W = \cdot)\right) \\ &= 3e_{\theta,W}(\pi) + 2D^2_{\mathrm{H}}\left(\mathbb{P}^{\pi}_{\theta}(\overline{\tau}_W = \cdot), \mathbb{P}^{\pi}_{\bar\theta}(\overline{\tau}_W = \cdot)\right).\end{aligned} \tag{44}$$

Plugging the inequalities (42) and (44) into (43), we have

$$\mathbb{E}^{\pi}_{\bar\theta}\Big[D^2_{\mathrm{TV}}\left(\mathbb{M}^{\theta}_{m_\theta(\overline{\tau}_W),\bar W}(\pi_{\mathrm{sep}}, s_W), \mathbb{M}^{\bar\theta}_{m_{\bar\theta}(\overline{\tau}_W),\bar W}(\pi_{\mathrm{sep}}, s_W)\right)\Big] \leqslant 18\varepsilon + 9e_{\theta,W}(\pi) + 9e_{\bar\theta,W}(\pi) =: \varepsilon'.$$

In other words, it holds that

$$\sum_{l,\bar l,s}\mathbb{P}^{\pi}_{\bar\theta}\big(s_W = s, m_\theta(\overline{\tau}_W) = l, m_{\bar\theta}(\overline{\tau}_W) = \bar l\big)\cdot D^2_{\mathrm{TV}}\left(\mathbb{M}^{\theta}_{l,\bar W}(\pi_{\mathrm{sep}}, s), \mathbb{M}^{\bar\theta}_{\bar l,\bar W}(\pi_{\mathrm{sep}}, s)\right) \leqslant \varepsilon'. \tag{45}$$

Notice that $\bar W \geqslant \varpi^{-1}(1)$. Thus, using (11), for any $m, l \in \mathrm{supp}(\rho_\theta)$ such that $m \neq l$, we have

$$D_{\mathrm{TV}}\left(\mathbb{M}^{\theta}_{l,\bar W}(\pi_{\mathrm{sep}}, s), \mathbb{M}^{\theta}_{m,\bar W}(\pi_{\mathrm{sep}}, s)\right) \geqslant \frac{1}{2}.$$

Hence, we choose $\sigma = \sigma_{\theta;\bar\theta}$ as

$$\sigma_{\theta;\bar\theta}(\bar l, s) \in \underset{l\in\mathrm{supp}(\rho_\theta)}{\arg\min}\ D_{\mathrm{TV}}\left(\mathbb{M}^{\theta}_{l,\bar W}(\pi_{\mathrm{sep}}, s), \mathbb{M}^{\bar\theta}_{\bar l,\bar W}(\pi_{\mathrm{sep}}, s)\right). \tag{46}$$

Then for any $l \in \mathrm{supp}(\rho_\theta)$ such that $l \neq \sigma(\bar l, s)$, it holds that

$$\begin{aligned}&2D_{\mathrm{TV}}\left(\mathbb{M}^{\theta}_{l,\bar W}(\pi_{\mathrm{sep}}, s), \mathbb{M}^{\bar\theta}_{\bar l,\bar W}(\pi_{\mathrm{sep}}, s)\right) \\ \geqslant\ & D_{\mathrm{TV}}\left(\mathbb{M}^{\theta}_{l,\bar W}(\pi_{\mathrm{sep}}, s), \mathbb{M}^{\bar\theta}_{\bar l,\bar W}(\pi_{\mathrm{sep}}, s)\right) + D_{\mathrm{TV}}\left(\mathbb{M}^{\theta}_{\sigma(\bar l,s),\bar W}(\pi_{\mathrm{sep}}, s), \mathbb{M}^{\bar\theta}_{\bar l,\bar W}(\pi_{\mathrm{sep}}, s)\right) \\ \geqslant\ & D_{\mathrm{TV}}\left(\mathbb{M}^{\theta}_{l,\bar W}(\pi_{\mathrm{sep}}, s), \mathbb{M}^{\theta}_{\sigma(\bar l,s),\bar W}(\pi_{\mathrm{sep}}, s)\right) \geqslant \frac{1}{2},\end{aligned}$$

and hence $D_{\mathrm{TV}}\left(\mathbb{M}^{\theta}_{l,\bar W}(\pi_{\mathrm{sep}}, s), \mathbb{M}^{\bar\theta}_{\bar l,\bar W}(\pi_{\mathrm{sep}}, s)\right) \geqslant \frac{1}{4}$. Therefore,

$$\begin{aligned}\varepsilon' &\geqslant \sum_{l,\bar l,s}\mathbb{P}^{\pi}_{\bar\theta}\big(s_W = s, m_\theta(\overline{\tau}_W) = l, m_{\bar\theta}(\overline{\tau}_W) = \bar l\big)\cdot D^2_{\mathrm{TV}}\left(\mathbb{M}^{\theta}_{l,\bar W}(\pi_{\mathrm{sep}}, s), \mathbb{M}^{\bar\theta}_{\bar l,\bar W}(\pi_{\mathrm{sep}}, s)\right) \\ &\geqslant \sum_{\bar l,s}\sum_{l\neq\sigma(\bar l,s)}\mathbb{P}^{\pi}_{\bar\theta}\big(s_W = s, m_\theta(\overline{\tau}_W) = l, m_{\bar\theta}(\overline{\tau}_W) = \bar l\big)\cdot\frac{1}{16} \\ &= \frac{1}{16}\cdot\mathbb{P}^{\pi}_{\bar\theta}(m_\theta(\overline{\tau}_W) \neq \sigma(m_{\bar\theta}(\overline{\tau}_W), s_W)).\end{aligned}$$

The proof is hence completed. □

Proposition E.7 (Performance decomposition) *Given LMDP model θ and reference LMDP $\bar\theta$, for any trajectory $\overline{\tau}_h$ with step $W \leqslant h < H$, we define*

$$\mathcal{E}^{\theta;\bar\theta}(\overline{\tau}_h) = \max_{a\in\mathcal{A}} D_{\mathrm{TV}}\left(\mathbb{T}^{\theta}_{\sigma(m_{\bar\theta}(\overline{\tau}_W),s_W)}(\cdot|s_h, a), \mathbb{T}^{\bar\theta}_{m_{\bar\theta}(\overline{\tau}_W)}(\cdot|s_h, a)\right), \tag{47}$$

where $\sigma = \sigma_{\theta;\bar\theta} : [L] \times \mathcal{S} \to [L]$ is the function defined in (46). Then it holds that

$$D_{\mathrm{TV}}\left(\mathbb{P}^{\pi}_{\theta}, \mathbb{P}^{\pi}_{\bar\theta}\right) \leqslant 300 D_{\mathrm{TV}}\left(\mathbb{P}^{\phi(\pi)}_{\theta}, \mathbb{P}^{\phi(\pi)}_{\bar\theta}\right) + 150 e_{\theta,W}(\pi) + 150 e_{\bar\theta,W}(\pi) + \sum_{h=W}^{H-1} \mathbb{E}^{\pi}_{\bar\theta} \mathcal{E}^{\theta;\bar\theta}(\overline{\tau}_h). \quad (48)$$

Conversely, for any step $W \leqslant h < H$,

$$\begin{aligned}\mathbb{E}^{\pi}_{\bar\theta} \mathcal{E}^{\theta;\bar\theta}(\overline{\tau}_h)^2 \leqslant 18 A D^2_{\mathrm{H}}\left(\mathbb{P}^{\varphi_h(\pi)}_{\theta}, \mathbb{P}^{\varphi_h(\pi)}_{\bar\theta}\right) + 300 D^2_{\mathrm{H}}\left(\mathbb{P}^{\phi(\pi)}_{\theta}, \mathbb{P}^{\phi(\pi)}_{\bar\theta}\right) \\ + 200 e_{\theta,W}(\pi) + 200 e_{\bar\theta,W}(\pi).\end{aligned} \quad (49)$$

Proof. We first prove (48). Notice that, by Lemma A.4,

$$\begin{aligned}D_{\mathrm{TV}}\left(\mathbb{P}^{\pi}_{\theta}, \mathbb{P}^{\pi}_{\bar\theta}\right) \leqslant\ & D_{\mathrm{TV}}\left(\mathbb{P}^{\pi}_{\theta}(\overline{\tau}_W = \cdot), \mathbb{P}^{\pi}_{\bar\theta}(\overline{\tau}_W = \cdot)\right) \\ & + \mathbb{E}^{\pi}_{\bar\theta}\big[D_{\mathrm{TV}}\left(\mathbb{P}^{\pi}_{\theta}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W), \mathbb{P}^{\pi}_{\bar\theta}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W)\right)\big].\end{aligned} \quad (50)$$

Using (39) and the triangle inequality of TV distance, we have

$$\begin{aligned}& D_{\mathrm{TV}}\left(\mathbb{P}^{\pi}_{\theta}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W), \mathbb{P}^{\pi}_{\bar\theta}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W)\right) \\ \leqslant\ & D_{\mathrm{TV}}\left(\mathbb{M}^{\theta}_{m_\theta(\overline{\tau}_W),\bar W}(\pi|_{\tau_{W-1}}, s_W), \mathbb{M}^{\bar\theta}_{m_{\bar\theta}(\overline{\tau}_W),\bar W}(\pi|_{\tau_{W-1}}, s_W)\right) + e_\theta(\overline{\tau}_W) + e_{\bar\theta}(\overline{\tau}_W),\end{aligned}$$

and taking expectation over $\overline{\tau}_W \sim \mathbb{P}^{\pi}_{\bar\theta}$, we obtain

$$\begin{aligned}& \mathbb{E}^{\pi}_{\bar\theta}\big[D_{\mathrm{TV}}\left(\mathbb{P}^{\pi}_{\theta}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W), \mathbb{P}^{\pi}_{\bar\theta}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W)\right)\big] \\ \leqslant\ & \mathbb{E}^{\pi}_{\bar\theta}\Big[D_{\mathrm{TV}}\left(\mathbb{M}^{\theta}_{m_\theta(\overline{\tau}_W),\bar W}(\pi|_{\tau_{W-1}}, s_W), \mathbb{M}^{\bar\theta}_{m_{\bar\theta}(\overline{\tau}_W),\bar W}(\pi|_{\tau_{W-1}}, s_W)\right)\Big] + \mathbb{E}^{\pi}_{\bar\theta}[e_\theta(\overline{\tau}_W)] + \mathbb{E}^{\pi}_{\bar\theta}[e_{\bar\theta}(\overline{\tau}_W)].\end{aligned} \quad (51)$$

For the last two term in the RHS of (51), we have $\mathbb{E}^{\pi}_{\bar\theta}[e_{\bar\theta}(\overline{\tau}_W)] = e_{\bar\theta,W}(\pi)$ and

$$\mathbb{E}^{\pi}_{\bar\theta}[e_\theta(\overline{\tau}_W)] \leqslant \mathbb{E}^{\pi}_{\theta}[e_\theta(\overline{\tau}_W)] + D_{\mathrm{TV}}\left(\mathbb{P}^{\pi}_{\theta}(\overline{\tau}_W = \cdot), \mathbb{P}^{\pi}_{\bar\theta}(\overline{\tau}_W = \cdot)\right). \quad (52)$$

To bound the first term in the RHS of (51), we consider the event $E_{\theta;\bar\theta} := \{m_\theta(\overline{\tau}_W) = \sigma(m_{\bar\theta}(\overline{\tau}_W), s_W)\}$. Under event $E_{\theta;\bar\theta}$, by Lemma A.4 we have

$$\begin{aligned}& D_{\mathrm{TV}}\left(\mathbb{M}^{\theta}_{m_\theta(\overline{\tau}_W),\bar W}(\pi|_{\tau_{W-1}}, s_W), \mathbb{M}^{\bar\theta}_{m_{\bar\theta}(\overline{\tau}_W),\bar W}(\pi|_{\tau_{W-1}}, s_W)\right) \\ \leqslant\ & \sum_{h=W}^{H-1} \mathbb{E}\left[D_{\mathrm{TV}}\left(\mathbb{T}^{\theta}_{m_\theta(\overline{\tau}_W)}(\cdot|s_h, a_h), \mathbb{T}^{\bar\theta}_{m_{\bar\theta}(\overline{\tau}_W)}(\cdot|s_h, a_h)\right)\Big|\, \tau_h \sim \mathbb{P}^{\pi}_{\bar\theta}(\cdot|\overline{\tau}_W)\right] \\ \leqslant\ & \sum_{h=W}^{H-1} \mathbb{E}\left[\max_a D_{\mathrm{TV}}\left(\mathbb{T}^{\theta}_{m_\theta(\overline{\tau}_W)}(\cdot|s_h, a), \mathbb{T}^{\bar\theta}_{m_{\bar\theta}(\overline{\tau}_W)}(\cdot|s_h, a)\right)\Big|\, \overline{\tau}_h \sim \mathbb{P}^{\pi}_{\bar\theta}(\cdot|\overline{\tau}_W)\right] \\ \overset{E_{\theta;\bar\theta}}{=}\ & \sum_{h=W}^{H-1} \mathbb{E}\left[\mathcal{E}^{\theta;\bar\theta}(\overline{\tau}_h)\Big|\, \tau_h \sim \mathbb{P}^{\pi}_{\bar\theta}(\cdot|\overline{\tau}_W)\right] = \sum_{h=W}^{H-1} \mathbb{E}^{\pi}_{\bar\theta}\left[\mathcal{E}^{\theta;\bar\theta}(\overline{\tau}_h)\Big|\,\overline{\tau}_W\right].\end{aligned}$$

Taking expectation over $\overline{\tau}_W \sim \mathbb{P}^{\pi}_{\bar{\theta}}$, it holds

$$\mathbb{E}^{\pi}_{\bar{\theta}}\Big[D_{\mathrm{TV}}\left(\mathbb{M}^{\theta}_{m_\theta(\overline{\tau}_W),\bar{W}}(\pi|_{\tau_{W-1}}, s_W), \mathbb{M}^{\bar{\theta}}_{m_{\bar{\theta}}(\overline{\tau}_W),\bar{W}}(\pi|_{\tau_{W-1}}, s_W)\right)\Big] \leqslant \mathbb{P}(E^c_{\theta;\bar{\theta}}) + \sum_{h=W}^{H-1} \mathbb{E}^{\pi}_{\bar{\theta}}\mathcal{E}^{\theta;\bar{\theta}}(\overline{\tau}_h). \tag{53}$$

Combining (51) with (52), (53) and (41) (Proposition E.6), the proof of (48) is completed.

We proceed similarly to prove (49). Notice that for any trajectory τ_h,

$$\mathbb{P}_\theta(s_{h+1} = \cdot|\tau_h) = \mathbb{E}_{m\sim \mathbf{b}_\theta(\overline{\tau}_h)}[\mathbb{T}_{\theta,m}(\cdot|s_h, a_h)].$$

Therefore,

$$D_{\mathrm{TV}}\left(\mathbb{P}_\theta(s_{h+1} = \cdot|\tau_h), \mathbb{T}^{\theta}_{m_\theta(\overline{\tau}_W)}(\cdot|s_h, a_h)\right) \leqslant \sum_{m\neq m_\theta(\overline{\tau}_h)} \mathbf{b}_\theta(\overline{\tau}_h)[m] = e_\theta(\overline{\tau}_h),$$

and hence

$$\begin{aligned} D_{\mathrm{TV}}\left(\mathbb{T}^{\theta}_{m_\theta(\overline{\tau}_W)}(\cdot|s_h, a_h), \mathbb{T}^{\bar{\theta}}_{m_{\bar{\theta}}(\overline{\tau}_W)}(\cdot|s_h, a_h)\right) \leqslant{}& D_{\mathrm{TV}}\left(\mathbb{P}_\theta(s_{h+1} = \cdot|\tau_h), \mathbb{P}_{\bar{\theta}}(s_{h+1} = \cdot|\tau_h)\right) \\ &+ e_\theta(\overline{\tau}_h) + e_{\bar{\theta}}(\overline{\tau}_h). \end{aligned}$$

In particular, given $h \geqslant W$, for any trajectory $\overline{\tau}_h$ whose prefix $\overline{\tau}_W$ satisfies $\overline{\tau}_W \in E_{\theta;\bar{\theta}}$, we have

$$\mathcal{E}^{\theta;\bar{\theta}}(\overline{\tau}_h) \leqslant \max_a D_{\mathrm{TV}}\left(\mathbb{P}_\theta(s_{h+1} = \cdot|\overline{\tau}_h, a), \mathbb{P}_{\bar{\theta}}(s_{h+1} = \cdot|\overline{\tau}_h, a)\right) + e_\theta(\overline{\tau}_h) + e_{\bar{\theta}}(\overline{\tau}_h).$$

Thus,

$$\mathbf{1}\left\{E_{\theta;\bar{\theta}}\right\}\mathcal{E}^{\theta;\bar{\theta}}(\tau_h)^2 \leqslant 3\max_a D^2_{\mathrm{TV}}\left(\mathbb{P}_\theta(s_{h+1} = \cdot|\overline{\tau}_h, a), \mathbb{P}_{\bar{\theta}}(s_{h+1} = \cdot|\overline{\tau}_h, a)\right) + 3e_\theta(\overline{\tau}_h) + 3e_{\bar{\theta}}(\overline{\tau}_h).$$

Taking expectation over $\tau_h \sim \mathbb{P}^{\pi}_{\bar{\theta}}$, we have

$$\begin{aligned} \mathbb{E}^{\pi}_{\bar{\theta}}\mathcal{E}^{\theta;\bar{\theta}}(\tau_h)^2 \leqslant \mathbb{P}^{\pi}_{\bar{\theta}}\left(E^c_{\theta;\bar{\theta}}\right) &+ 3\mathbb{E}^{\pi}_{\bar{\theta}}\Big[\max_a D^2_{\mathrm{TV}}\left(\mathbb{P}_\theta(s_{h+1} = \cdot|\overline{\tau}_h, a), \mathbb{P}_{\bar{\theta}}(s_{h+1} = \cdot|\overline{\tau}_h, a)\right)\Big] \\ &+ 3\mathbb{E}^{\pi}_{\bar{\theta}}[e_\theta(\overline{\tau}_h)] + 3\mathbb{E}^{\pi}_{\bar{\theta}}[e_{\bar{\theta}}(\overline{\tau}_h)]. \end{aligned}$$

Notice that

$$\begin{aligned} &\mathbb{E}^{\pi}_{\bar{\theta}}\Big[\max_a D^2_{\mathrm{TV}}\left(\mathbb{P}_\theta(s_{h+1} = \cdot|\overline{\tau}_h, a), \mathbb{P}_{\bar{\theta}}(s_{h+1} = \cdot|\overline{\tau}_h, a)\right)\Big] \\ \leqslant{}& \mathbb{E}^{\pi}_{\bar{\theta}}\left[\sum_a D^2_{\mathrm{TV}}\left(\mathbb{P}_\theta(s_{h+1} = \cdot|\overline{\tau}_h, a), \mathbb{P}_{\bar{\theta}}(s_{h+1} = \cdot|\overline{\tau}_h, a)\right)\right] \\ \leqslant{}& 2\mathbb{E}^{\pi}_{\bar{\theta}}\left[\sum_a D^2_{\mathrm{H}}\left(\mathbb{P}_\theta(s_{h+1} = \cdot|\overline{\tau}_h, a), \mathbb{P}_{\bar{\theta}}(s_{h+1} = \cdot|\overline{\tau}_h, a)\right)\right] \\ ={}& 2\mathbb{E}^{\pi}_{\bar{\theta}}\left[A\cdot D^2_{\mathrm{H}}\left(\mathbb{P}_\theta(s_{h+1} = \cdot|\overline{\tau}_h, a_h \sim \mathrm{Unif}(\mathcal{A})), \mathbb{P}_{\bar{\theta}}(s_{h+1} = \cdot|\overline{\tau}_h, a_h \sim \mathrm{Unif}(\mathcal{A}))\right)\right] \\ \leqslant{}& 4AD^2_{\mathrm{H}}\left(\mathbb{P}^{\pi\circ_h \mathrm{Unif}(\mathcal{A})}_{\theta}(\overline{\tau}_{h+1} = \cdot), \mathbb{P}^{\pi\circ_h \mathrm{Unif}(\mathcal{A})}_{\bar{\theta}}(\overline{\tau}_{h+1} = \cdot)\right) \end{aligned}$$

$$\leqslant 4AD_{\rm H}^2\left(\mathbb{P}_{\theta}^{\varphi_h(\pi)}, \mathbb{P}_{\bar{\theta}}^{\varphi_h(\pi)}\right),$$

where the third inequality follows from Lemma A.5. By definition, we know $\mathbb{E}_{\bar{\theta}}^{\pi}[e_{\bar{\theta}}(\overline{\tau}_h)] = e_{\bar{\theta},h}(\pi) \leqslant e_{\theta,W}(\pi)$ (Lemma E.8), and using Lemma A.3, we also have

$$\begin{aligned}\mathbb{E}_{\bar{\theta}}^{\pi}[e_{\theta}(\overline{\tau}_h)] &\leqslant 3\mathbb{E}_{\theta}^{\pi}[e_{\theta}(\overline{\tau}_h)] + 2D_{\rm H}^2\left(\mathbb{P}_{\theta}^{\pi}(\overline{\tau}_h = \cdot), \mathbb{P}_{\bar{\theta}}^{\pi}(\overline{\tau}_h = \cdot)\right)\\ &\leqslant 3e_{\theta,W}(\pi) + 2D_{\rm H}^2\left(\mathbb{P}_{\theta}^{\varphi_h(\pi)}, \mathbb{P}_{\bar{\theta}}^{\varphi_h(\pi)}\right).\end{aligned}$$

Combining the inequalities above with (41) completes the proof. $\square$

Lemma E.8 *For $h \geqslant W$, it holds that $e_{\theta,h}(\pi) \leqslant e_{\theta,W}(\pi)$.*

Proof. By definition,

$$\begin{aligned}e_{\theta,h}(\pi) &= \mathbb{E}_{\theta}^{\pi}\left[1 - \max_{m} \widetilde{\mathbb{P}}_{\theta}(m^\star = m|\overline{\tau}_h)\right]\\ &\leqslant \mathbb{E}_{\theta}^{\pi}\left[1 - \widetilde{\mathbb{P}}_{\theta}(m^\star = m_{\theta}(\overline{\tau}_W)|\overline{\tau}_h)\right]\\ &= 1 - \widetilde{\mathbb{P}}(m^\star = m_{\theta}(\overline{\tau}_W))\\ &= e_{\theta,W}(\overline{\tau}_W).\end{aligned}$$

$\square$

E.4. Proof of Theorem 4.3

We first present and prove a more general result as follows; Theorem 4.3 is then a direct corollary.

Corollary E.9 *Under the success event E_0 of Proposition E.2, it holds that*

$$V_\star - V_{\theta^\star}(\widehat{\pi}) \lesssim \sqrt{Ld^2\iota_K\left(\frac{AH^2\beta}{K} + \frac{\bar{W}^2(U_+ + KU_\star)}{K^2}\right)} + \varepsilon_{\rm s},$$

where we denote $\iota_K = \log(LdH \cdot K/(A\beta))$, and

$$U_\star = \sum_{k=1}^{K} e_{\theta^\star,W}(\pi^k), \qquad U_+ = \sum_{1\leqslant t<k\leqslant K} e_{\theta^k,W}(\pi^t).$$

Proof. Recall that by Proposition E.3, we have that under E_0

$$V_\star - V_{\theta^\star}(\widehat{\pi}) \leqslant \frac{1}{K}\sum_{k=1}^{K} D_{\rm TV}\left(\mathbb{P}_{\theta^k}^{\pi^k}, \mathbb{P}_{\theta^\star}^{\pi^k}\right).$$

Taking summation of (48) over $(\theta^1, \pi^1), \cdots, (\theta^K, \pi^K)$, we have

$$\sum_{k=1}^{K} D_{\rm TV}\left(\mathbb{P}_{\theta^k}^{\pi^k}, \mathbb{P}_{\theta^\star}^{\pi^k}\right) \lesssim \sum_{k=1}^{K} D_{\rm TV}\left(\mathbb{P}_{\theta^k}^{\phi(\pi^k)}, \mathbb{P}_{\theta^\star}^{\phi(\pi^k)}\right) + \sum_{k=1}^{K}\left(e_{\theta^k,W}(\pi^k) + e_{\theta^\star,W}(\pi^k)\right)$$

$$+ \sum_{k=1}^{K} \sum_{h=W}^{H-1} \mathbb{E}_{\theta^\star}^{\pi^k} \mathcal{E}^{\theta^k;\theta^\star}(\overline{\tau}_h).$$

By Theorem E.5, we can bound the first term in the RHS above as

$$\sum_{k=1}^{K} D_{\mathrm{TV}}\left(\mathbb{P}_{\theta^k}^{\phi(\pi^k)}, \mathbb{P}_{\theta^\star}^{\phi(\pi^k)}\right) \lesssim \sqrt{LdAH^2\iota_K K\beta}.$$

Combining with the fact that $e_{\theta^k,W}(\pi^k) \leqslant \varepsilon_{\mathrm{s}}$, we obtain

$$\sum_{k=1}^{K} D_{\mathrm{TV}}\left(\mathbb{P}_{\theta^k}^{\pi^k}, \mathbb{P}_{\theta^\star}^{\pi^k}\right) \lesssim \sqrt{LdAH^2\iota_K K\beta} + K\varepsilon_{\mathrm{s}} + U_\star + \sum_{h=W}^{H-1} \sum_{k=1}^{K} \mathbb{E}_{\theta^\star}^{\pi^k} \mathcal{E}^{\theta^k;\theta^\star}(\overline{\tau}_h). \tag{54}$$

Using (49) and the definition of $\mathsf{p}(\cdot)$, we also know that for all $t, k \in [K]$,

$$\sum_{h=W}^{H-1} \mathbb{E}_{\theta^\star}^{\pi^t} \mathcal{E}^{\theta^k;\theta^\star}(\overline{\tau}_h)^2 \lesssim AHD_{\mathrm{H}}^2\left(\mathbb{P}_{\theta^k}^{\mathsf{p}(\pi^t)}, \mathbb{P}_{\theta^\star}^{\mathsf{p}(\pi^t)}\right) + \bar{W}e_{\theta^k,W}(\pi^t) + \bar{W}e_{\theta^\star,W}(\pi^t). \tag{55}$$

Therefore, using (30) and the fact that E_0 holds, we have

$$\sum_{t<k} \sum_{h=W}^{H-1} \mathbb{E}_{\theta^\star}^{\pi^t} \mathcal{E}^{\theta^k;\theta^\star}(\overline{\tau}_h)^2 \lesssim AH\beta + \bar{W}U_k, \tag{56}$$

where we denote $U_k := \sum_{t<k}\left(e_{\theta^k,W}(\pi^t) + e_{\theta^\star,W}(\pi^t)\right)$. Therefore, it remains to bridge between the inequalities in Eqs. (54) and (56) above using Proposition A.9.

Fix a $W \leqslant h \leqslant H-1$. Notice that $\mathcal{E}^{\theta^k;\theta^\star}(\overline{\tau}_h)$ only depends on $\overline{\tau}_h$ through the tuple

$$x_h = (m_{\theta^\star}(\overline{\tau}_W), s_W, s_h) \in \mathcal{X} := [L] \times \mathcal{S} \times \mathcal{S},$$

and hence we can consider the distribution $p_{t,h} = \mathbb{P}_{\theta^\star}^{\pi^t}(x_h = \cdot) \in \Delta(\mathcal{X})$. It remains to shows that there exists a distribution $\mu_h \in \Delta(\mathcal{X})$ such that $p_{t,h}(x)/\mu_h(x) \leqslant C_{\mathrm{cov}} \forall x \in \mathcal{X}$ for some parameter C_{cov}.

Under Assumption 2.7, by Proposition A.10, there exist distributions $\tilde{\mu}_m \in \Delta(\mathcal{S})$ for each $m \in [L]$ such that

$$\mathbb{T}_{\theta^\star,m}(s'|s,a) \leqslant d \cdot \tilde{\mu}_m(s'), \qquad \forall m \in [L], (s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}.$$

Therefore, in the case $h > W$, for any $x = (m, s, s') \in \mathcal{X}$, we have

$$\begin{aligned}
p_{t,h}(x) = \mathbb{P}_{\theta^\star}^{\pi^t}(x_h = x) \leqslant \mathbb{P}_{\theta^\star}^{\pi^t}(s_W = s, s_h = s') \\
&= \mathbb{E}_{(m^\star,\tau_{h-1},s_h)}\left[\mathbf{1}\left\{s_W = s, s_h = s'\right\}\right] \\
&= \mathbb{E}_{(m^\star,\tau_{h-1})}\left[\mathbf{1}\left\{s_W = s\right\}\mathbb{E}\left[\mathbf{1}\left\{s_h = s'\right\}\middle|\, s_h \sim \widetilde{\mathbb{P}}_{\theta^\star}(\cdot|\tau_{h-1}, m^\star)\right]\right] \\
&= \mathbb{E}_{(m^\star,\tau_{h-1})}\left[\mathbf{1}\left\{s_W = s\right\}\mathbb{T}_{m^\star}^{\theta^\star}(s'|s_{h-1}, a_{h-1})\right]
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \mathbb{E}_{(m^\star,\tau_{h-1})}\left[\mathbf{1}\{s_W=s\}\cdot d\cdot\tilde{\mu}_{m^\star}(s')\right] \\
&= \mathbb{E}_{(m^\star,\tau_{W-1})}\left[\mathbb{E}\left[\mathbf{1}\{s_W=s\} \middle| s_W\sim\widetilde{\mathbb{P}}_{\theta^\star}(\cdot|\tau_{W-1},m^\star)\right]\cdot d\cdot\tilde{\mu}_{m^\star}(s')\right] \\
&= \mathbb{E}_{(m^\star,\tau_{W-1})}\left[\mathbb{T}^{\theta^\star}_{m^\star}(s|s_{W-1},a_{W-1})\cdot d\cdot\tilde{\mu}_{m^\star}(s')\right] \\
&\leqslant \mathbb{E}_{m^\star}\left[d\cdot\tilde{\mu}_{m^\star}(s)\cdot d\cdot\tilde{\mu}_{m^\star}(s')\right] \\
&= d^2\sum_{m^\star\in[L]}\rho_{\theta^\star}(m^\star)\tilde{\mu}_{m^\star}(s)\tilde{\mu}_{m^\star}(s'),
\end{aligned}
$$

where the expectation is taken over $(m^\star,\tau_H)\sim\widetilde{\mathbb{P}}^{\pi^t}_{\theta^\star}$. Thus, we can choose $\mu_h\in\Delta(\mathcal{X})$ as

$$
\mu_h(m,s,s')=\frac{1}{L}\sum_{m^\star\in[L]}\rho_{\theta^\star}(m^\star)\tilde{\mu}_{m^\star}(s)\tilde{\mu}_{m^\star}(s'),\qquad\forall(m,s,s')\in\mathcal{X}.
$$

Then, for $h>W$, $t\in[T]$ and any $x\in\mathcal{X}$, we know $p_{t,h}(x)\leqslant Ld^2\cdot\mu_h(x)$. For the case $h=W$, an argument essentially the same as above also yields that there exists a $\mu_W\in\Delta(\mathcal{X})$ such that $p_{t,W}(x)\leqslant Ld\cdot\mu_W(x)$ for all $t\in[T]$, $x\in\mathcal{X}$.

We can now apply Proposition A.9 with $M=A\beta$ to obtain that for all $W\leqslant h\leqslant H-1$,

$$
\sum_{k=1}^K\mathbb{E}^{\pi^k}_{\theta^\star}\mathcal{E}^{\theta^k;\theta^\star}(\overline{\tau}_h)\lesssim\sqrt{Ld^2\log\left(1+\frac{Ld^2K}{A\beta}\right)\left[KA\beta+\sum_{k=1}^K\sum_{t<k}\mathbb{E}^{\pi^t}_{\theta^\star}\mathcal{E}^{\theta^k;\bar{\theta}}(\overline{\tau}_h)^2\right]}. \tag{57}
$$

Taking summation over $W\leqslant h\leqslant H-1$ and using (56), we have

$$
\sum_{h=W}^{H-1}\sum_{k=1}^K\mathbb{E}^{\pi^k}_{\theta^\star}\mathcal{E}^{\theta^k;\theta^\star}(\overline{\tau}_h)\lesssim\sqrt{Ld^2\iota_K\left[KAH^2\beta+\bar{W}^2\sum_{k=1}^K U_k\right]}. \tag{58}
$$

Combining (58) above with (54), we can conclude that

$$
\begin{aligned}
\sum_{k=1}^K D_{\mathrm{TV}}\left(\mathbb{P}^{\pi^k}_{\theta^k},\mathbb{P}^{\pi^k}_{\theta^\star}\right)&\lesssim\sqrt{LdAH^2\iota_K K\beta}+K\varepsilon_{\mathrm{s}}+U_\star+H\sqrt{Ld^2\iota_K\left[KAH^2\beta+\bar{W}^2\sum_{k=1}^K U_k\right]} \\
&\lesssim\sqrt{Ld^2\iota_K\left(KAH^2\beta+\bar{W}^2(KU_\star+U_+)\right)}+K\varepsilon_{\mathrm{s}}+U_\star \\
&\lesssim\sqrt{Ld^2\iota_K\left(KAH^2\beta+\bar{W}^2(KU_\star+U_+)\right)}+K\varepsilon_{\mathrm{s}},
\end{aligned}
$$

where the last inequality follows from $U_\star\leqslant K$ and hence $U_\star\leqslant\sqrt{KU_\star}$. Applying Proposition E.3 completes the proof. □

Proof of Theorem 4.3 Under Assumption 4.2, it holds that $e_{\theta,W}(\pi)\leqslant\varepsilon_{\mathrm{s}}$ for all $\theta\in\Theta$ and $\pi\in\Pi$ (Proposition 4.1). Therefore, $U_\star\leqslant K\varepsilon_{\mathrm{s}}$, $U_+\leqslant K^2\varepsilon_{\mathrm{s}}$, and Corollary E.9 implies that as long as

$$
K\gtrsim\frac{Ld^2AH^2\iota_K}{\varepsilon^2}\cdot\beta,\qquad\varepsilon_{\mathrm{s}}\lesssim\frac{\varepsilon^2}{Ld^2\bar{W}^2\iota_K},
$$

we have $V_\star-V_{\theta^\star}(\widehat{\pi})\leqslant\varepsilon$, which is fulfilled by the choice of parameters in Theorem 4.3. □

E.5. Proof of Theorem 4.8

According to Corollary E.9, we only need to upper bound the term $U_\star$ and U_+ under Assumption 4.7. The following proposition links these two quantities with the condition $e_{\theta^k,W}(\pi^k) \leqslant \varepsilon_{\mathrm{s}} \forall k \in [K]$.

Proposition E.10 *Suppose that Assumption 4.7 holds. Then for any policy π, LMDP model θ and reference LMDP model $\bar{\theta}$, it holds that*

$$e_{\theta,W}(\pi) \leqslant \frac{1}{\alpha}\Big[3D_{\mathrm{TV}}\left(\mathbb{P}_\theta^{\phi(\pi)}, \mathbb{P}_{\bar{\theta}}^{\phi(\pi)}\right) + e_{\bar{\theta},W}(\pi)\Big]$$

Proof. Using (40) and the triangle inequality, we have

$$\begin{aligned}
&D_{\mathrm{TV}}\left(\mathbb{P}_\theta^{\phi(\pi)}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W), \mathbb{M}^{\bar{\theta}}_{m_{\bar{\theta}}(\overline{\tau}_W),\bar{W}}(\pi_{\mathrm{sep}}, s_W)\right) \\
&\leqslant D_{\mathrm{TV}}\left(\mathbb{P}_\theta^{\phi(\pi)}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W), \mathbb{P}_{\bar{\theta}}^{\phi(\pi)}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W)\right) + e_{\bar{\theta}}(\overline{\tau}_W).
\end{aligned}$$

On the other hand,

$$\mathbb{P}_\theta^{\phi(\pi)}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W) = \mathbb{E}_{m\sim \mathbf{b}_\theta(\overline{\tau}_W)}\Big[\mathbb{M}^{\theta}_{m,\bar{W}}(\pi_{\mathrm{sep}}, s_W)\Big],$$

and hence by (8), it holds that

$$D_{\mathrm{TV}}\left(\mathbb{P}_\theta^{\phi(\pi)}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W), \mathbb{M}^{\bar{\theta}}_{m_{\bar{\theta}}(\overline{\tau}_W),\bar{W}}(\pi_{\mathrm{sep}}, s_W)\right) \geqslant \alpha\Big(1 - \max_m \mathbf{b}_\theta(\overline{\tau}_W)[m]\Big) = \alpha e_\theta(\overline{\tau}_W).$$

Taking expectation over $\overline{\tau}_W \sim \mathbb{P}^\pi_{\bar{\theta}}$, we obtain

$$\begin{aligned}
\alpha\mathbb{E}^\pi_{\bar{\theta}}[e_\theta(\overline{\tau}_W)] &\leqslant \mathbb{E}^\pi_{\bar{\theta}}\Big[D_{\mathrm{TV}}\left(\mathbb{P}_\theta^{\phi(\pi)}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W), \mathbb{M}^{\bar{\theta}}_{m_{\bar{\theta}}(\overline{\tau}_W),\bar{W}}(\pi_{\mathrm{sep}}, s_W)\right)\Big] \\
&\leqslant \mathbb{E}^\pi_{\bar{\theta}}\Big[D_{\mathrm{TV}}\left(\mathbb{P}_\theta^{\phi(\pi)}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W), \mathbb{P}_{\bar{\theta}}^{\phi(\pi)}(\overline{\tau}_{W:H} = \cdot|\overline{\tau}_W)\right)\Big] + \mathbb{E}^\pi_{\bar{\theta}}[e_{\bar{\theta}}(\overline{\tau}_W)] \\
&\leqslant 2D_{\mathrm{TV}}\left(\mathbb{P}_\theta^{\phi(\pi)}, \mathbb{P}_{\bar{\theta}}^{\phi(\pi)}\right) + e_{\bar{\theta},W}(\pi),
\end{aligned}$$

where the last inequality follows from Lemma A.4 and the fact that $\mathbb{E}^\pi_{\bar{\theta}}[e_{\bar{\theta}}(\overline{\tau}_W)] = e_{\bar{\theta},W}(\pi)$. Notice that we also have

$$\begin{aligned}
\mathbb{E}^\pi_{\bar{\theta}}[e_\theta(\overline{\tau}_W)] &\geqslant \mathbb{E}^\pi_{\theta}[e_\theta(\overline{\tau}_W)] - D_{\mathrm{TV}}\left(\mathbb{P}^\pi_\theta(\overline{\tau}_W = \cdot), \mathbb{P}^\pi_{\bar{\theta}}(\overline{\tau}_W = \cdot)\right) \\
&= e_{\theta,W}(\pi) - D_{\mathrm{TV}}\left(\mathbb{P}^\pi_\theta(\overline{\tau}_W = \cdot), \mathbb{P}^\pi_{\bar{\theta}}(\overline{\tau}_W = \cdot)\right).
\end{aligned}$$

Combining the inequalities above completes the proof. □

Proof of Theorem 4.8 According to our choice of (θ^k, π^k), we know that $e_{\theta^k,W}(\pi^k) \leqslant \varepsilon_{\mathrm{s}}$ always holds for $k \in [K]$. Hence, by Proposition E.10,

$$e_{\theta^\star,W}(\pi^k) \leqslant \frac{1}{\alpha}\Big[3D_{\mathrm{TV}}\left(\mathbb{P}_{\theta^k}^{\phi(\pi^k)}, \mathbb{P}_{\theta^\star}^{\phi(\pi^k)}\right) + \varepsilon_{\mathrm{s}}\Big].$$

Summing over $k \in [K]$, we obtain that

$$U_\star = \sum_{k=1}^K e_{\theta^\star,W}(\pi^k) \leqslant \frac{1}{\alpha}\left[3\sum_{k=1}^K D_{\mathrm{TV}}\left(\mathbb{P}_{\theta^k}^{\phi(\pi^k)}, \mathbb{P}_{\theta^k}^{\phi(\pi^k)}\right) + K\varepsilon_{\mathrm{s}}\right]$$

$$\lesssim \frac{1}{\alpha}\sqrt{LdAH^2\iota_K K\beta} + \frac{K\varepsilon_{\mathrm{s}}}{\alpha},$$

where the last inequality follows from Theorem E.5.

Similarly, by Proposition E.10, we can bound

$$\begin{aligned}
e_{\theta^k,W}(\pi^t) &\leqslant \frac{1}{\alpha}\Big[3D_{\mathrm{TV}}\left(\mathbb{P}_{\theta^k}^{\phi(\pi^t)}, \mathbb{P}_{\theta^t}^{\phi(\pi^t)}\right) + e_{\theta^t}(\pi^t)\Big] \\
&\leqslant \frac{1}{\alpha}\Big[3D_{\mathrm{TV}}\left(\mathbb{P}_{\theta^k}^{\phi(\pi^t)}, \mathbb{P}_{\theta^\star}^{\phi(\pi^t)}\right) + 3D_{\mathrm{TV}}\left(\mathbb{P}_{\theta^t}^{\phi(\pi^t)}, \mathbb{P}_{\theta^\star}^{\phi(\pi^t)}\right) + e_{\theta^t}(\pi^t)\Big].
\end{aligned}$$

Therefore, taking summation over $1 \leqslant t < k \leqslant K$, we have

$$U_+ = \sum_{1\leqslant t<k\leqslant K} e_{\theta^k,W}(\pi^t) \lesssim \frac{1}{\alpha}\left[\sum_{1\leqslant t<k\leqslant K} D_{\mathrm{TV}}\left(\mathbb{P}_{\theta^k}^{\phi(\pi^t)}, \mathbb{P}_{\theta^\star}^{\phi(\pi^t)}\right) + K\sum_{t=1}^{K} D_{\mathrm{TV}}\left(\mathbb{P}_{\theta^t}^{\phi(\pi^t)}, \mathbb{P}_{\theta^\star}^{\phi(\pi^t)}\right) + K^2\varepsilon_{\mathrm{s}}\right].$$

By Cauchy inequality, it holds

$$\sum_{1\leqslant t<k\leqslant K} D_{\mathrm{TV}}\left(\mathbb{P}_{\theta^k}^{\phi(\pi^t)}, \mathbb{P}_{\theta^\star}^{\phi(\pi^t)}\right) \leqslant \sqrt{K^2\cdot \sum_{1\leqslant t<k\leqslant K} D_{\mathrm{TV}}^2\left(\mathbb{P}_{\theta^k}^{\phi(\pi^t)}, \mathbb{P}_{\theta^\star}^{\phi(\pi^t)}\right)} \lesssim K\sqrt{K\beta},$$

where we use the fact that $D_{\mathrm{TV}} \leqslant \sqrt{2}D_{\mathrm{H}}$ and Proposition E.2. Combining Theorem E.5 with the above two inequalities, we can conclude that

$$U_+ = \sum_{1\leqslant t<k\leqslant K} e_{\theta^k,W}(\pi^t) \lesssim \frac{1}{\alpha}K\sqrt{LdAH^2\iota K\beta} + \frac{K^2\varepsilon_{\mathrm{s}}}{\alpha}.$$

Hence, Corollary E.9 implies that

$$V_\star - V_{\theta^\star}(\widehat{\pi}) \lesssim \sqrt{Ld^2\iota_K\left(\frac{AH^2\beta}{\alpha K} + \frac{\varepsilon_{\mathrm{s}}}{\alpha} + \frac{1}{\alpha}\sqrt{\frac{LdAH^2\iota_K\beta}{K}}\right)}.$$

Therefore, to ensure that $V_\star - V_{\theta^\star}(\widehat{\pi}) \leqslant \varepsilon$, we only need to ensure

$$K \gtrsim \frac{L^3d^5AH^6\iota_K^3}{\alpha^2\varepsilon^4}\cdot\beta, \qquad \varepsilon_{\mathrm{s}} \lesssim \frac{\alpha\varepsilon^2}{Ld^2\bar{W}^2\iota_K}.$$

In particular, the choice of parameters in Theorem 4.8 suffices. □

E.6. Proof of Theorem E.5

The proof of Theorem E.5 is (almost) a direct analog of the analysis in Chen et al. (2022a, Appendix D & G). However, we may not directly invoke the guarantees there for general PSR to obtain Theorem E.5 because PSR is formalized in terms of a set of *core action sequences*, so that the system dynamics is uniquely determined by the dynamics under these action sequences. However, for our setting, we are instead given an explorative policy π_{sep}, which is not necessary a mixture of action sequences.

Therefore, in the following, we present a minimal self-contained proof of Theorem E.5, which is in essence a slight modification of the original proof in Chen et al. (2022a). We refer the reader to Chen et al. (2022a) for more detailed analysis and proofs.

In the following, we first introduce the notations for POMDPs, which generalize LMDPs.

POMDPs A Partially Observable Markov Decision Process (POMDP) is a sequential decision process whose transition dynamics are governed by *latent states*. A POMDP is specified by a tuple $\{\mathcal{Z}, \mathcal{O}, \mathcal{A}, \mathbb{T}, \mathbb{O}, H, \mu_1\}$, where $\mathcal{Z}$ is the latent state space, $\mathbb{O}(\cdot|\cdot) : \mathcal{Z} \to \Delta(\mathcal{O})$ is the emission dynamics, $\mathbb{T}(\cdot|\cdot,\cdot) : \mathcal{Z} \times \mathcal{A} \to \Delta(\mathcal{Z})$ is the transition dynamics over the latent states, and $\mu_1 \in \Delta(\mathcal{Z})$ specifies the distribution of initial state z_1. At each step h, given the latent state z_h (which the agent cannot observe), the system emits observation $o_h \sim \mathbb{O}(\cdot|z_h)$, receives action $a_h \in \mathcal{A}$ from the agent, and then transits to the next latent state $z_{h+1} \sim \mathbb{T}(\cdot|z_h, a_h)$ in a Markov fashion. The episode terminates immediately after a_H is taken.

In a POMDP with observation space $\mathcal{O}$ and action space $\mathcal{A}$, a policy $\pi = \{\pi_h : (\mathcal{O} \times \mathcal{A})^{h-1} \times \mathcal{O} \to \Delta(\mathcal{A})\}_{h=1}^H$ is a collection of H functions. At step $h \in [H]$, an agent running policy π observes the observation o_h and takes action $a_h \sim \pi_h(\cdot|\tau_{h-1}, o_h) \in \Delta(\mathcal{A})$ based on the history $(\tau_{h-1}, o_h) = (o_1, a_1, \ldots, o_{h-1}, a_{h-1}, o_h)$. The environment then generates the next observation o_{h+1} based on $\tau_h = (o_1, a_1, \cdots, o_h, a_h)$ (according to the dynamics of the underlying POMDP).

Suppose that $\widetilde{\Theta}$ is a set of POMDP models with common action space $\mathcal{A}$ and observation space $\mathcal{O}$, such that each $\theta \in \widetilde{\Theta}$ specifies the tuple $(\mathbb{T}_\theta, \mathbb{O}_\theta, \mu_\theta)$ and hence the POMDP dynamics. [7]

Suppose that a step parameter $1 \leqslant W < H$ is given, along with a policy π_{sep}. Then, for each policy π, we define

$$\varphi(\pi) := \frac{1}{W} \sum_{h=0}^{W-1} \pi \circ_h \mathrm{Unif}(\mathcal{A}) \circ_{h+1} \pi_{\mathrm{sep}} \tag{59}$$

analogously to Definition E.1. We also consider the emission matrix induced by π_{sep}:

$$\mathbb{K}_\theta = \left[\mathbb{P}_\theta^{\pi_{\mathrm{sep}}}((o_1, a_1, \cdots, o_{\bar{W}}) = \overline{\tau} | s_1 = s)\right]_{(\overline{\tau}, s)} \in \mathbb{R}^{\mathcal{T} \times \mathcal{Z}}, \tag{60}$$

where $\bar{W} = H - W + 1$, $\mathcal{T} = (\mathcal{O} \times \mathcal{A})^{\bar{W}-1} \times \mathcal{O}$. Suppose that for each $\theta \in \Theta$, there exists $\mathbb{K}_\theta^+ \in \mathbb{R}^{\mathcal{Z} \times \mathcal{T}}$ such that $\mathbb{K}_\theta^+ \mathbb{K}_\theta = I_{\mathcal{Z}}$, and we write $\Lambda_{\exp} := \max_{\theta \in \Theta} \left\|\mathbb{K}_\theta^+\right\|_1$.

Operator representation of POMDP dynamics Define

$$\mathbf{B}_\theta(o, a) = \mathbb{K}_\theta \mathbb{T}_{\theta,a} \,\mathrm{diag}(\mathbb{O}_\theta(o|\cdot)) \mathbb{K}_\theta^+, \qquad \mathbf{q}_{\theta,0} = \mathbb{K}_\theta \mu_\theta. \tag{61}$$

where we denote $\mathbb{T}_{\theta,a} := \mathbb{T}_\theta(\cdot|\cdot, a) \in \mathbb{R}^{\mathcal{Z} \times \mathcal{Z}}$ for each $a \in \mathcal{A}$, and $\mathrm{diag}(\mathbb{O}_\theta(o|\cdot))\mathbb{R}^{\mathcal{Z} \times \mathcal{Z}}$ is the diagonal matrix with the (z, z)-entry being $\mathbb{O}(o|z)$ for each $z \in \mathcal{Z}$.

An important property of the definition (61) is that, for any trajectory $\overline{\tau}_{h+\bar{W}} = (\tau_h, o_{h+1}, a_{h+1}, \cdots, o_{h+\bar{W}})$, it holds that

$$\mathbf{e}_{(o_{h+1}, a_{h+1}, \cdots, o_{h+\bar{W}})}^\top \mathbf{B}_\theta(o_h, a_h) \cdots \mathbf{B}_\theta(o_1, a_1) \mathbf{q}_{\theta,0} = \mathbb{P}_\theta(o_{h+1}, a_{h+1}, \cdots, o_{h+\bar{W}} | \tau_h, \pi_{\mathrm{sep}}) \times \mathbb{P}_\theta(o_{1:h} | \mathrm{do}(a_{1:h})),$$

where we recall that $\mathbb{P}_\theta(o_{h+1}, a_{h+1}, \cdots, o_{h+\bar{W}} | \tau_h, \pi_{\mathrm{sep}})$ is the probability of observing $o_{h+1}, a_{h+1}, \cdots, o_{h+\bar{W}}$ when executing policy π_{sep} starting at step $h + 1$ in POMDP θ, conditional on the history τ_h (see also (29)). Therefore, for any policy π, it holds that

$$\mathbb{P}_\theta^{\pi \circ_{h+1} \pi_{\mathrm{sep}}}(\overline{\tau}_{h+\bar{W}}) = \mathbf{e}_{(o_{h+1}, a_{h+1}, \cdots, o_{h+\bar{W}})}^\top \mathbf{B}_\theta(o_h, a_h) \cdots \mathbf{B}_\theta(o_1, a_1) \mathbf{q}_{\theta,0} \times \pi(\tau_h). \tag{62}$$

7. Strictly speaking, θ also specifies $\mathcal{Z}_\theta$, its own latent state space. For notational simplicity, we always omit the subscript θ of the state space $\mathcal{Z}$ in the following analysis.

In particular, we can now express TV distance between model as difference between operators:

$$\begin{aligned} &D_{\mathrm{TV}}\left(\mathbb{P}_{\theta}^{\pi\circ_{h+1}\pi_{\mathrm{sep}}},\mathbb{P}_{\bar\theta}^{\pi\circ_{h+1}\pi_{\mathrm{sep}}}\right) \\ &=\frac{1}{2}\sum_{\tau_h}\pi(\tau_h)\times\left\|\mathbf{B}_\theta(o_h,a_h)\cdots\mathbf{B}_\theta(o_1,a_1)\mathbf{q}_{\theta,0}-\mathbf{B}_{\bar\theta}(o_h,a_h)\cdots\mathbf{B}_{\bar\theta}(o_1,a_1)\mathbf{q}_{\bar\theta,0}\right\|_1. \end{aligned} \tag{63}$$

Also, we denote $\mathbf{q}_\theta(\tau_h)=\left[\mathbb{P}_\theta((o_{h+1},a_{h+1},\cdots,o_{h+\bar W})=\cdot|\tau_h,\pi_{\mathrm{sep}})\right]\in\Delta(\mathcal{T})$, then we also have

$$\mathbf{B}_\theta(o_h,a_h)\cdots\mathbf{B}_\theta(o_1,a_1)\mathbf{q}_{\theta,0}=\mathbf{q}_\theta(\tau_h)\times\mathbb{P}_\theta(\tau_h), \tag{64}$$

where we recall the notation $\mathbb{P}_\theta(\tau_h)=\mathbb{P}_\theta(o_{1:h}|\mathrm{do}(a_{1:h}))$.

Another important fact is that, for any 1-step policy $\pi:\mathcal{O}\to\Delta(\mathcal{A})$ and $\mathbf{q}\in\mathbb{R}^{\mathcal{T}}$,

$$\sum_{o,a}\pi(a|o)\times\|\mathbf{B}_\theta(o,a)\mathbf{q}\|_1\leqslant\left\|\mathbb{K}_\theta^+\mathbf{q}\right\|_1, \tag{65}$$

$$\sum_{o,a}\pi(a|o)\times\left\|\mathbb{K}_\theta^+\mathbf{B}_\theta(o,a)\mathbf{q}\right\|_1\leqslant\left\|\mathbb{K}_\theta^+\mathbf{q}\right\|_1. \tag{66}$$

This is because $\|\mathbb{K}_\theta\|_1\leqslant 1$, $\|\mathbb{T}_{\theta,a}\|_1\leqslant 1$, and $\sum_{o,a}\pi(a|o)\mathbb{O}_\theta(o|z)=1$ for any $z\in\mathcal{Z}$. Hence, we can apply (66) recursively to show that, for any h-step policy π,

$$\sum_{\tau_h}\pi(\tau_h)\times\|\mathbf{B}_\theta(o_h,a_h)\cdots\mathbf{B}_\theta(o_1,a_1)\mathbf{q}\|_1\leqslant\left\|\mathbb{K}_\theta^+\mathbf{q}\right\|_1. \tag{67}$$

Proposition E.11 *For each pair of models $\theta,\bar\theta\in\Theta$, we define $\bar{\mathcal{E}}^{\theta;\bar\theta}:\mathbb{R}^{\mathcal{T}}\to\mathbb{R}$ as follows:*

$$\bar{\mathcal{E}}^{\theta;\bar\theta}(\mathbf{q}):=\frac{1}{2}\max_{\pi':\mathcal{O}\to\Delta(\mathcal{A})}\sum_{o,a}\pi'(a|o)\times\left\|\mathbb{K}_\theta^+(\mathbf{B}_\theta(o,a)-\mathbf{B}_{\bar\theta}(o,a))\mathbf{q}\right\|_1 \tag{68}$$

For each step h, define[8]

$$\mathcal{E}^{\theta;\bar\theta}(\tau_h):=\bar{\mathcal{E}}^{\theta;\bar\theta}(\mathbf{q}_{\bar\theta}(\tau_h)),\qquad \mathcal{E}_0^{\theta;\bar\theta}:=\frac{1}{2}\left\|\mathbb{K}_\theta^+(\mathbf{q}_{\theta,0}-\mathbf{q}_{\bar\theta,0})\right\|_1.$$

Then it holds that

$$D_{\mathrm{TV}}\left(\mathbb{P}_\theta^{\pi\circ_W\pi_{\mathrm{sep}}},\mathbb{P}_{\bar\theta}^{\pi\circ_W\pi_{\mathrm{sep}}}\right)\leqslant\mathcal{E}_0^{\theta;\bar\theta}+\sum_{h=1}^{W-1}\mathbb{E}_{\bar\theta}^{\pi}\mathcal{E}^{\theta;\bar\theta}(\tau_{h-1}). \tag{69}$$

Conversely, it holds

$$(\mathcal{E}_0^{\theta;\bar\theta})^2+\sum_{h=1}^{W-1}\mathbb{E}_{\bar\theta}^{\pi}\mathcal{E}^{\theta;\bar\theta}(\tau_{h-1})^2\leqslant 8AW\Lambda_{\exp}^2D_{\mathrm{H}}^2\left(\mathbb{P}_\theta^{\varphi(\pi)},\mathbb{P}_{\bar\theta}^{\varphi(\pi)}\right). \tag{70}$$

8. The error functional might seem strange at first glance, but it can be regarded as a counterpart of the decomposition (31) for MDP. Indeed, when $\widetilde\Theta$ is a class of MDP models (i.e. $\mathcal{Z}=\mathcal{O}=\mathcal{S}$ and $\mathbb{K}=\mathbb{O}=I_{\mathcal{S}}$), then

$$\mathcal{E}^{\theta;\bar\theta}(\tau_{h-1})=\mathbb{E}_{s_h|\tau_{h-1},\bar\theta}\max_a D_{\mathrm{TV}}\left(\mathbb{T}_\theta(\cdot|s_h,a),\mathbb{T}_{\bar\theta}(\cdot|s_h,a)\right).$$

Proof. Before presenting the proof, we first introduce some notations. We abbreviate $\mathbf{B}_\theta(o_1, a_1, \cdots, o_l, a_l) = \mathbf{B}_\theta(o_l, a_l) \cdots \mathbf{B}_\theta(o_1, a_1)$. For a trajectory $\tau_H = (o_1, a_1, \cdots, o_H, a_H)$, we write $\tau_{h':h} = (o_{h'}, a_{h'}, \cdots, o_h, a_h)$ and $\overline{\tau}_{h':h} = (o_{h'}, a_{h'}, \cdots, o_h)$.

Using (63), we have

$$
\begin{aligned}
&2D_{\mathrm{TV}}\left(\mathbb{P}_\theta^{\pi \circ_W \pi_{\mathrm{sep}}}, \mathbb{P}_{\bar\theta}^{\pi \circ_W \pi_{\mathrm{sep}}}\right) \\
&\overset{(63)}{=} \sum_{\tau_{W-1}} \pi(\tau_{W-1}) \times \left\| \mathbf{B}_\theta(o_{W-1}, a_{W-1}) \cdots \mathbf{B}_\theta(o_1, a_1) \mathbf{q}_{\theta,0} - \mathbf{B}_{\bar\theta}(o_{W-1}, a_{W-1}) \cdots \mathbf{B}_{\bar\theta}(o_1, a_1) \mathbf{q}_{\bar\theta,0} \right\|_1 \\
&\leqslant \sum_{\tau_{W-1}} \pi(\tau_{W-1}) \left\| \mathbf{B}_\theta(\tau_{1:W-1}) (\mathbf{q}_{\theta,0} - \mathbf{q}_{\bar\theta,0}) \right\|_1 \\
&\quad + \sum_{\tau_{W-1}} \pi(\tau_{W-1}) \times \sum_{h=1}^{W-1} \left\| \mathbf{B}_\theta(\tau_{h+1:W-1}) (\mathbf{B}_\theta(o_h, a_h) - \mathbf{B}_{\bar\theta}(o_h, a_h)) \mathbf{B}_{\bar\theta}(\tau_{1:h-1}) \mathbf{q}_{\bar\theta,0} \right\|_1 \\
&\overset{(67)}{\leqslant} \frac{1}{2} \left\| \mathbb{K}_\theta^+ (\mathbf{q}_{\theta,0} - \mathbf{q}_{\bar\theta,0}) \right\|_1 + \frac{1}{2} \sum_{h=1}^{W-1} \sum_{\tau_h} \pi(\tau_h) \times \left\| \mathbb{K}_\theta^+ (\mathbf{B}_\theta(o_h, a_h) - \mathbf{B}_{\bar\theta}(o_h, a_h)) \mathbf{B}_{\bar\theta}(\tau_{1:h-1}) \mathbf{q}_{\bar\theta,0} \right\|_1 \\
&\overset{(64)}{=} \frac{1}{2} \left\| \mathbb{K}_\theta^+ (\mathbf{q}_{\theta,0} - \mathbf{q}_{\bar\theta,0}) \right\|_1 + \frac{1}{2} \sum_{h=1}^{W-1} \sum_{\tau_h} \pi(\tau_h) \times \left\| \mathbb{K}_\theta^+ (\mathbf{B}_\theta(o_h, a_h) - \mathbf{B}_{\bar\theta}(o_h, a_h)) \mathbf{q}_{\bar\theta}(\tau_{h-1}) \right\|_1 \times \mathbb{P}_\theta(\tau_{h-1}) \\
&= \mathcal{E}_0^{\theta;\bar\theta} + \frac{1}{2} \sum_{h=1}^{W-1} \sum_{\tau_{h-1}} \sum_{o_h, a_h} \mathbb{P}_\theta^\pi(\tau_{h-1}) \times \pi(a_h | \tau_{h-1}, o_h) \times \left\| \mathbb{K}_\theta^+ (\mathbf{B}_\theta(o_h, a_h) - \mathbf{B}_{\bar\theta}(o_h, a_h)) \mathbf{q}_{\bar\theta}(\tau_{h-1}) \right\|_1 \\
&\leqslant \mathcal{E}_0^{\theta;\bar\theta} + \sum_{h=1}^{W-1} \sum_{\tau_{h-1}} \mathbb{P}_\theta^\pi(\tau_{h-1}) \times \mathcal{E}^{\theta;\bar\theta}(\mathbf{q}_{\bar\theta}(\tau_{h-1})),
\end{aligned}
$$

where the last two lines follow from the definition (68). This completes the proof of (69).

Next, we proceed to prove (70). By definition,

$$
\begin{aligned}
2\mathcal{E}^{\theta;\bar\theta}(\tau_h) = &\max_{\pi'} \sum_{o,a} \pi'(a|o) \times \left\| \mathbb{K}_\theta^+ (\mathbf{B}_\theta(o, a) - \mathbf{B}_{\bar\theta}(o, a)) \mathbf{q}_{\bar\theta}(\tau_{h-1}) \right\|_1 \\
\leqslant &\max_{\pi'} \sum_{o,a} \pi'(a|o) \times \left\| \mathbb{K}_\theta^+ (\mathbf{B}_\theta(o, a) \mathbf{q}_\theta(\tau_{h-1}) - \mathbf{B}_{\bar\theta}(o, a) \mathbf{q}_{\bar\theta}(\tau_{h-1})) \right\|_1 \\
&+ \max_{\pi'} \sum_{o,a} \pi'(a|o) \times \left\| \mathbb{K}_\theta^+ \mathbf{B}_\theta(o, a) (\mathbf{q}_\theta(\tau_{h-1}) - \mathbf{q}_{\bar\theta}(\tau_{h-1})) \right\|_1.
\end{aligned}
$$

For the first term, notice that for any $o \in \mathcal{O}$, $a \in \mathcal{A}$,

$$
\mathbf{B}_\theta(o, a) \mathbf{q}_\theta(\tau_{h-1}) = \left[\mathbb{P}_\theta(o_h = o, \overline{\tau}_{h+1:h+\bar W} = \cdot | \tau_{h-1}, a_h = a, a_{h+1:h+\bar W} \sim \pi_{\mathrm{sep}}) \right] \in \mathbb{R}^{\mathcal{T}}.
$$

Therefore, for any step $1 \leqslant h \leqslant W - 1$ and any 1-step policy $\pi' : \mathcal{O} \to \Delta(\mathcal{A})$, we have

$$
\begin{aligned}
&\sum_{o,a} \pi'(a|o) \times \left\| \mathbb{K}_\theta^+ (\mathbf{B}_\theta(o, a) \mathbf{q}_\theta(\tau_{h-1}) - \mathbf{B}_{\bar\theta}(o, a) \mathbf{q}_{\bar\theta}(\tau_{h-1})) \right\|_1 \\
&\leqslant \Lambda_{\exp} \sum_{o,a} \pi'(a|o) \times \left\| \mathbf{B}_\theta(o, a) \mathbf{q}_\theta(\tau_{h-1}) - \mathbf{B}_{\bar\theta}(o, a) \mathbf{q}_{\bar\theta}(\tau_{h-1}) \right\|_1
\end{aligned}
$$

$$= 2\Lambda_{\exp} D_{\mathrm{TV}} \left(\mathbb{P}_\theta(\overline{\tau}_{h:h+\bar{W}} = \cdot | \tau_{h-1}, \pi' \circ \pi_{\mathrm{sep}}), \mathbb{P}_{\bar{\theta}}(\overline{\tau}_{h:h+\bar{W}} = \cdot | \tau_{h-1}, \pi' \circ \pi_{\mathrm{sep}}) \right),$$

where the inequality uses the fact that $\|\mathbb{K}_\theta^+\|_1 \leqslant \Lambda_{\exp}$ for all $\theta \in \Theta$. Furthermore,

$$\begin{aligned}
&\frac{1}{2} D_{\mathrm{TV}}^2 \left(\mathbb{P}_\theta(\overline{\tau}_{h:h+\bar{W}} = \cdot | \tau_{h-1}, \pi' \circ \pi_{\mathrm{sep}}), \mathbb{P}_{\bar{\theta}}(\overline{\tau}_{h:h+\bar{W}} = \cdot | \tau_{h-1}, \pi' \circ \pi_{\mathrm{sep}}) \right) \\
\leqslant\ & D_{\mathrm{H}}^2 \left(\mathbb{P}_\theta(\overline{\tau}_{h:h+\bar{W}} = \cdot | \tau_{h-1}, \pi' \circ \pi_{\mathrm{sep}}), \mathbb{P}_{\bar{\theta}}(\overline{\tau}_{h:h+\bar{W}} = \cdot | \tau_{h-1}, \pi' \circ \pi_{\mathrm{sep}}) \right) \\
\leqslant\ & \sum_{a \in \mathcal{A}} D_{\mathrm{H}}^2 \left(\mathbb{P}_\theta(\overline{\tau}_{h:h+\bar{W}} = \cdot | \tau_{h-1}, a \circ \pi_{\mathrm{sep}}), \mathbb{P}_{\bar{\theta}}(\overline{\tau}_{h:h+\bar{W}} = \cdot | \tau_{h-1}, a \circ \pi_{\mathrm{sep}}) \right) \\
=\ & A D_{\mathrm{H}}^2 \left(\mathbb{P}_\theta(\overline{\tau}_{h:h+\bar{W}} = \cdot | \tau_{h-1}, \mathrm{Unif}(\mathcal{A}) \circ \pi_{\mathrm{sep}}), \mathbb{P}_{\bar{\theta}}(\overline{\tau}_{h:h+\bar{W}} = \cdot | \tau_{h-1}, \mathrm{Unif}(\mathcal{A}) \circ \pi_{\mathrm{sep}}) \right),
\end{aligned}$$

where the second inequality uses the fact that squared Hellinger distance is an f-divergence. For the second term, by the definition of $\mathbf{B}_\theta$, we have

$$\begin{aligned}
&\sum_{o,a} \pi'(a|o) \times \left\| \mathbb{K}_\theta^+ \mathbf{B}_\theta(o,a)(\mathbf{q}_\theta(\tau_{h-1}) - \mathbf{q}_{\bar{\theta}}(\tau_{h-1})) \right\|_1 \overset{(66)}{\leqslant} \left\| \mathbb{K}_\theta^+ (\mathbf{q}_\theta(\tau_{h-1}) - \mathbf{q}_{\bar{\theta}}(\tau_{h-1})) \right\|_1 \\
\leqslant\ & \Lambda_{\exp} \|\mathbf{q}_\theta(\tau_{h-1}) - \mathbf{q}_{\bar{\theta}}(\tau_{h-1})\|_1 \\
=\ & \Lambda_{\exp} \cdot 2 D_{\mathrm{TV}} \left(\mathbb{P}_\theta(\overline{\tau}_{h:h+\bar{W}-1} = \cdot | \tau_{h-1}, \pi_{\mathrm{sep}}), \mathbb{P}_{\bar{\theta}}(\overline{\tau}_{h:h+\bar{W}-1} = \cdot | \tau_{h-1}, \pi_{\mathrm{sep}}) \right)
\end{aligned}$$

Combining the inequalities above and applying Lemma A.5, we obtain

$$\begin{aligned}
\mathbb{E}_{\bar{\theta}}^{\pi} \mathcal{E}^{\theta;\bar{\theta}}(\tau_{h-1})^2 \leqslant 4 A \Lambda_{\exp}^2 D_{\mathrm{H}}^2 \left(\mathbb{P}_\theta^{\pi \circ_h \mathrm{Unif}(\mathcal{A}) \circ_{h+1} \pi_{\mathrm{sep}}}, \mathbb{P}_{\bar{\theta}}^{\pi \circ_h \mathrm{Unif}(\mathcal{A}) \circ_{h+1} \pi_{\mathrm{sep}}} \right) \\
+ 4 \Lambda_{\exp}^2 D_{\mathrm{H}}^2 \left(\mathbb{P}_\theta^{\pi \circ_h \pi_{\mathrm{sep}}}, \mathbb{P}_{\bar{\theta}}^{\pi \circ_h \pi_{\mathrm{sep}}} \right).
\end{aligned} \tag{71}$$

Notice that for step $h \geqslant 2$, we have

$$D_{\mathrm{H}}^2 \left(\mathbb{P}_\theta^{\pi \circ_h \pi_{\mathrm{sep}}}, \mathbb{P}_{\bar{\theta}}^{\pi \circ_h \pi_{\mathrm{sep}}} \right) \leqslant A D_{\mathrm{H}}^2 \left(\mathbb{P}_\theta^{\pi \circ_{h-1} \mathrm{Unif}(\mathcal{A}) \circ_h \pi_{\mathrm{sep}}}, \mathbb{P}_{\bar{\theta}}^{\pi \circ_{h-1} \mathrm{Unif}(\mathcal{A}) \circ_h \pi_{\mathrm{sep}}} \right),$$

and we also have

$$\begin{aligned}
\mathcal{E}_0^{\theta;\bar{\theta}} = \frac{1}{2} \left\| \mathbb{K}_\theta^+ (\mathbf{q}_{\theta,0} - \mathbf{q}_{\bar{\theta},0}) \right\|_1 \leqslant \Lambda_{\exp} D_{\mathrm{TV}} \left(\mathbb{P}_\theta^{\pi_{\mathrm{sep}}}(\overline{\tau}_{1:\bar{W}} = \cdot), \mathbb{P}_{\bar{\theta}}^{\pi_{\mathrm{sep}}}(\overline{\tau}_{1:\bar{W}} = \cdot) \right) \\
\leqslant \sqrt{2} \Lambda_{\exp} D_{\mathrm{H}} \left(\mathbb{P}_\theta^{\pi_{\mathrm{sep}}}, \mathbb{P}_{\bar{\theta}}^{\pi_{\mathrm{sep}}} \right).
\end{aligned} \tag{72}$$

Combining the inequalities above completes the proof of (70). □

Proposition E.12 *Suppose that $D = \mathrm{rank}(\mathbb{T}_{\theta^\star})$, $\beta \geqslant 1$, and $(\theta^1, \pi^1), \cdots, (\theta^K, \pi^K)$ is a sequence of (POMDP, policy) pairs such that for all $k \in [K]$,*

$$\sum_{t<k} D_{\mathrm{H}}^2 \left(\mathbb{P}_{\theta^k}^{\varphi(\pi^t)}, \mathbb{P}_{\theta^\star}^{\varphi(\pi^t)} \right) \leqslant M.$$

Then it holds that

$$\sum_{k=1}^{K} D_{\mathrm{TV}} \left(\mathbb{P}_{\theta^k}^{\pi^k \circ_W \pi_{\mathrm{sep}}}, \mathbb{P}_{\theta^\star}^{\pi^k \circ_W \pi_{\mathrm{sep}}} \right) \lesssim \sqrt{\Lambda_{\exp}^2 A D W^2 \tilde{\iota} \cdot K M},$$

where $\tilde{\iota} = \log \left(1 + \frac{2 \Lambda_{\exp}^2 K D}{A M} \right)$.

Proof. Using Proposition E.11, we have

$$\sum_{k=1}^{K} D_{\mathrm{TV}}\left(\mathbb{P}_{\theta^k}^{\pi^k \circ_W \pi_{\mathrm{sep}}}, \mathbb{P}_{\theta^\star}^{\pi^k \circ_W \pi_{\mathrm{sep}}}\right) \leqslant \sum_{k=1}^{K} 1 \wedge \mathcal{E}_0^{\theta^k;\theta^\star} + \sum_{h=1}^{W-1}\sum_{k=1}^{K} 1 \wedge \mathbb{E}_{\theta^\star}^{\pi^k} \mathcal{E}^{\theta^k;\theta^\star}(\tau_{h-1}), \tag{73}$$

and for any pair of (t,k),

$$(\mathcal{E}_0^{\theta^k;\theta^\star})^2 + \sum_{h=1}^{W-1} \mathbb{E}_{\theta^\star}^{\pi^t} \mathcal{E}^{\theta^k;\theta^\star}(\tau_{h-1})^2 \leqslant 8AW\Lambda_{\exp}^2 D_{\mathrm{H}}^2\left(\mathbb{P}_{\theta^k}^{\varphi(\pi^t)}, \mathbb{P}_{\theta^\star}^{\varphi(\pi^t)}\right).$$

In particular, for any $k \in [K]$,

$$\sum_{t<k} (\mathcal{E}_0^{\theta^k;\theta^\star})^2 + \sum_{h=1}^{W-1}\sum_{t<k} \mathbb{E}_{\theta^\star}^{\pi^t} \mathcal{E}^{\theta^k;\theta^\star}(\tau_{h-1})^2 \leqslant 8AW\Lambda_{\exp}^2 M. \tag{74}$$

It remains to apply Proposition A.8 to bridge between (73) and (74).

For each $k \in [K]$, define $f_k = \bar{\mathcal{E}}^{\theta^k;\theta^\star} : \mathbb{R}^{\mathcal{T}} \to \mathbb{R}$. By definition, f_k takes the form

$$f_k(x) = \max_{\pi} \sum_{o,a,s} \left|\left\langle x, y_{k,(o,a),\pi} \right\rangle\right|$$

where $y_{k,(o,a),\pi}^\top = \pi(a|o) \times \mathbf{e}_s^\top \mathbb{K}_{\theta^k}^+ (\mathbf{B}_{\theta^k}(o,a) - \mathbf{B}_{\theta^\star}(o,a))$. It is also easy to verify that $f_k(x) \leqslant 2\Lambda_{\exp}^2 \|x\|_1$ using $\left\|\mathbb{K}_\theta^+\right\|_1 \leqslant \Lambda_{\exp}$ and $\left\|\mathbb{K}_{\theta^\star}^+\right\|_1 \leqslant \Lambda_{\exp}$. Furthermore, for each step $1 \leqslant h \leqslant W-1$, the set

$$\mathcal{X}_h := \left\{\mathbf{q}_{\theta^\star}(\tau_{h-1}) : \tau_{h-1} \in (\mathcal{O} \times \mathcal{A})^{h-1}\right\}$$

spans a subspace of dimension at most D.

Therefore, applying Proposition A.8 yields that for each $1 \leqslant h \leqslant W-1$

$$\sum_{k=1}^{K} 1 \wedge \mathbb{E}_{\theta^\star}^{\pi^k} \mathcal{E}^{\theta^k;\theta^\star}(\tau_{h-1}) \lesssim \sqrt{D\tilde{\iota}\left[K \cdot AM + \sum_{k=1}^{K}\sum_{t<k} \mathbb{E}_{\theta^\star}^{\pi^t} \mathcal{E}^{\theta^k;\theta^\star}(\tau_{h-1})^2\right]}, \tag{75}$$

where $\tilde{\iota} = \log(1 + 2\Lambda_{\exp}^2 DK/AM)$. Similarly, treating $\mathcal{E}_0^{\theta^k;\theta^\star}$ as a function over the singleton set, we also have

$$\sum_{k=1}^{K} 1 \wedge \mathcal{E}_0^{\theta^k;\theta^\star} \lesssim \sqrt{\tilde{\iota}\left[KAM + \sum_{k=1}^{K}\sum_{t<k} (\mathcal{E}_0^{\theta^k;\theta^\star})^2\right]}$$

Combining the two inequalities above with (73) and (74), we obtain

$$\sum_{k=1}^{K} D_{\mathrm{TV}}\left(\mathbb{P}_{\theta^k}^{\pi^k \circ_W \pi_{\mathrm{sep}}}, \mathbb{P}_{\theta^\star}^{\pi^k \circ_W \pi_{\mathrm{sep}}}\right) \leqslant \sum_{k=1}^{K} 1 \wedge \mathcal{E}_0^{\theta^k;\theta^\star} + \sum_{h=1}^{W-1}\sum_{k=1}^{K} 1 \wedge \mathbb{E}_{\theta^\star}^{\pi^k} \mathcal{E}^{\theta^k;\theta^\star}(\tau_{h-1})$$

$$\lesssim \sqrt{DW\tilde{\iota}\left[KAM + \sum_{k=1}^{K}\sum_{t<k}\left((\mathcal{E}_0^{\theta^k;\theta^\star})^2 + \sum_{h=1}^{W-1}\mathbb{E}_{\theta^\star}^{\pi^t}\mathcal{E}^{\theta^k;\theta^\star}(\tau_{h-1})^2\right)\right]}$$
$$\lesssim \sqrt{DW\iota \cdot K \cdot \Lambda_{\rm exp}^2 AWM},$$

where the first inequality is (73), the second inequality follows from Cauchy-Schwarz, and the last inequality follows from (74) and the given condition. $\square$

Proof of Theorem E.5 Recall that Θ is a class of LMDP with common state space $\mathcal{S}$. For each LMDP $\theta \in \Theta$, we construct a POMDP $\mathsf{pomdp}(\theta)$ with latent state space $\mathcal{Z} = \mathcal{S} \times \mathrm{supp}(\rho_\theta)$ and observation space $\mathcal{O} = \mathcal{S}$ as follows:

- The initial state is $\tilde{s}_1 = (s_1, m)$, where $m \sim \rho_\theta$, $s_1 \sim \mu_{\theta,m}$.
- The state $\tilde{s} = (s, m)$ always emits $o = s$ as the observation. After an action a is taken, the next state is generated as $\tilde{s}' = (s', m)$ where $s' \sim \mathbb{T}_{\theta,m}(\cdot|s, a)$.

The transition matrix of $\mathsf{pomdp}(\theta)$ specified above can also be written as

$$\mathbb{T}_{\mathsf{pomdp}(\theta)} = \mathrm{diag}\left(\mathbb{T}_{\theta,m}\right)_{m\in\mathrm{supp}(\rho_\theta)},$$

up to reorganization of coordinates. Therefore, we have $\mathrm{rank}(\mathbb{T}_{\mathsf{pomdp}(\theta^\star)}) \leqslant Ld$.

Because $\mathcal{O} = \mathcal{S}$, any policy for the LMDP θ is a policy for the POMDP $\mathsf{pomdp}(\theta)$, and vice versa. Furthermore, it is easy to verify that for any policy π, the trajectory distribution $\mathbb{P}^\pi_{\mathsf{pomdp}(\theta)}(\tau_H = \cdot)$ agrees with the distribution $\mathbb{P}^\pi_\theta(\tau_H = \cdot)$. Hence, for each $\theta \in \Theta$,

$$\mathbb{K}_{\mathsf{pomdp}(\theta)} = \mathrm{diag}\left(\mathbb{M}^\theta_{*,\bar{W}}(\pi_{\rm sep}, s)\right)_{s\in\mathcal{S}},$$

where we denote

$$\mathbb{M}^\theta_{*,\bar{W}}(\pi_{\rm sep}, s) := [\mathbb{M}^\theta_{m,\bar{W}}(\pi_{\rm sep}, s)]_{m\in\mathrm{supp}(\rho_\theta)} \in \mathbb{R}^{(\mathcal{A}\times\mathcal{S})^{\bar{W}-1}\times\mathrm{supp}(\rho_\theta)}.$$

By Lemma A.7, as long as $\varpi(\bar{W}) \geqslant \log(2L)$, for each $(s, m) \in \mathcal{Z}$, there exists a left inverse of $\mathbb{M}^\theta_{*,\bar{W}}(\pi_{\rm sep}, s)$ with ℓ_1 norm bounded by 2. In particular, we apply Lemma A.7 to conclude the existence of a left inverse with the desired norm bound for each block of the block diagonal matrix $\mathbb{K}_{\mathsf{pomdp}(\theta)}$. Therefore, there exists a left inverse of $\mathbb{K}_{\mathsf{pomdp}(\theta)}$ with ℓ_1 norm bounded by 2, and hence $\Lambda_{\rm exp} \leqslant 2$.

Therefore, we can now apply Proposition E.12 to complete the proof of Theorem E.5. $\square$

E.7. A sufficient condition for Assumption 4.7

The following proposition indicates that Assumption 4.7 is not that strong as it may seem: it holds for a broad class of LMDPs under relatively mild assumptions on the support of each MDP instance.

Proposition E.13 *Suppose that there is a policy π_0 and parameter $W_0 \geqslant \varpi^{-1}(3\log(L/\alpha_0))$, such that for each $\theta \in \Theta$, the LMDP M_θ is ϖ-separated under π_0, and there exists $\mu_\theta : \mathcal{S} \to \Delta(\mathcal{S})$ so that*

$$\mathbb{T}^{\pi_0}_{\theta,m}(s_{W_0} = s'|s_1 = s) \geqslant \alpha_0\mu_\theta(s'|s), \qquad \forall m \in \mathrm{supp}(\rho_\theta), s, s' \in \mathcal{S}.$$

Let $\pi_{\rm sep} = \pi_0 \circ_{W_0} \pi_0$. Then Assumption 4.7 holds with $W_{\rm exp} = 2W_0$ and $\alpha = \frac{\alpha_0}{32}$.

For the sake of notational simplicity, we first prove a more abstract version of Proposition E.13.

Proposition E.14 *For measurable spaces $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z} := \mathcal{Y} \times \mathcal{X}$, consider the class of transition kernels from $\mathcal{X}$ to $\mathcal{Z}$:*

$$\mathcal{Q} = \{\mathbb{Q} : \mathcal{X} \to \Delta(\mathcal{Z})\}.$$

For any $\mathbb{Q} \in \mathcal{Q}$, we define $\mathbb{Q}^{\otimes 2} : \mathcal{X} \to \Delta(\mathcal{Z} \times \mathcal{Z})$ as follows: for any $x_0 \in \mathcal{X}$, $\mathbb{Q}^{\otimes 2}(\cdot|x_0)$ is the probability distribution of (z, z'), where $z = (Y, x) \sim \mathbb{Q}(\cdot|x)$, $z' = (Y', x') \sim \mathbb{Q}(\cdot|x)$.

Suppose that $\mathbb{Q}_m \in \mathcal{Q}$ are transition kernels such that for all $m \neq l$,

$$D_{\mathrm{B}}(\mathbb{Q}_m(\cdot|x), \mathbb{Q}_l(\cdot|x)) \geqslant 3\log(L/\alpha), \qquad \forall x \in \mathcal{X}.$$

Further assume that there exists $\mu : \mathcal{X} \to \Delta(\mathcal{X})$ such that

$$\mathbb{Q}_m(x|x_0) \geqslant \alpha\mu(x|x_0), \qquad \forall m \in [L]. \tag{76}$$

Then for any $\mathbb{Q} \in \mathcal{Q}$, $x_0 \in \mathcal{X}$, and $p \in \Delta([L])$, we have

$$D_{\mathrm{TV}}\left(\mathbb{E}_{m\sim p}\mathbb{Q}_m^{\otimes 2}(\cdot|x_0), \mathbb{Q}^{\otimes 2}(\cdot|x_0)\right) \geqslant \frac{\alpha}{32}(1 - \max_m p_m).$$

Proof. In the following, we fix any given $\mathbb{Q} \in \mathcal{Q}$, $x_0 \in \mathcal{X}$, and $p \in \Delta([L])$. Let $\widetilde{\mathbb{P}}$ be the probability distribution of (m, z, z'), where $m \sim p$, $z = (Y, x) \sim \mathbb{Q}_m(\cdot|x_0)$, and $z' = (Y', x') \sim \mathbb{Q}_m(\cdot|x_0)$ (i.e. $(z, z') \sim \mathbb{Q}_m^{\otimes 2}(\cdot|x_0)$). Also, let $\mathbb{P} = \mathbb{E}_{m\sim p}\mathbb{Q}_m^{\otimes 2}(\cdot|x_0)$ be the marginal distribution of $(z, z') \sim \widetilde{\mathbb{P}}$. We also omit x_0 from the conditional probabilities when it is clear from the context.

By Lemma A.4, it holds that

$$\mathbb{E}_{(Y,x)\sim\mathbb{P}}\big[D_{\mathrm{TV}}\left(\mathbb{P}(z' = \cdot|Y, x), \mathbb{Q}^{\otimes 2}(z' = \cdot|Y, x)\right)\big] \leqslant 2D_{\mathrm{TV}}\left(\mathbb{P}, \mathbb{Q}^{\otimes 2}\right).$$

We also have

$$\mathbb{E}_{x'\sim\mathbb{P}}\big[D_{\mathrm{TV}}\left(\mathbb{P}(z' = \cdot|x), \mathbb{Q}^{\otimes 2}(z' = \cdot|x)\right)\big] \leqslant 2D_{\mathrm{TV}}\left(\mathbb{P}, \mathbb{Q}^{\otimes 2}\right).$$

Notice that the conditional distribution $\mathbb{Q}^{\otimes 2}(z' = \cdot|Y, x) = \mathbb{Q}(z' = \cdot|x)$ only depends on x, and hence by triangle inequality,

$$\mathbb{E}_{(Y,x)\sim\mathbb{P}}\big[D_{\mathrm{TV}}\left(\mathbb{P}(z' = \cdot|Y, x), \mathbb{P}(z' = \cdot|x)\right)\big] \leqslant 4D_{\mathrm{TV}}\left(\mathbb{P}, \mathbb{Q}^{\otimes 2}\right)$$

Further notice that

$$\mathbb{P}(z' = \cdot|Y, x) = \mathbb{E}_{m|Y,x}\big[\mathbb{Q}_m(z' = \cdot|x)\big], \qquad \mathbb{P}(z' = \cdot|x) = \mathbb{E}_{m|x}\big[\mathbb{Q}_m(z' = \cdot|x)\big].$$

Hence, by Lemma A.6, we have

$$D_{\mathrm{TV}}\left(\mathbb{P}(z' = \cdot|Y, x), \mathbb{P}(z' = \cdot|x)\right) \geqslant \frac{1}{2}D_{\mathrm{TV}}\left(\widetilde{\mathbb{P}}(m = \cdot|Y, x), \widetilde{\mathbb{P}}(m = \cdot|x)\right).$$

Next, using the definition of TV distance (which is a f-divergence, see e.g. Polyanskiy and Wu (2014)), we can show that

$$\mathbb{E}_{(Y,x)\sim\widetilde{\mathbb{P}}}\Big[D_{\mathrm{TV}}\left(\widetilde{\mathbb{P}}(m=\cdot|Y,x),\widetilde{\mathbb{P}}(m=\cdot|x)\right)\Big]=\mathbb{E}_{(m,x)\sim\widetilde{\mathbb{P}}}\Big[D_{\mathrm{TV}}\left(\widetilde{\mathbb{P}}(Y=\cdot|m,x),\widetilde{\mathbb{P}}(Y=\cdot|x)\right)\Big].$$

We know

$$\mathbb{P}(Y=\cdot|m,x)=\mathbb{Q}_m(Y=\cdot|x_0,x),\qquad \mathbb{P}(Y=\cdot|x)=\mathbb{E}_{m|x}[\mathbb{Q}_m(zy=\cdot|x_0,x)],$$

and hence combining the inequalities above gives

$$4D_{\mathrm{TV}}\left(\mathbb{P},\mathbb{Q}^{\otimes 2}\right)\geqslant\mathbb{E}_{(m,x)\sim\mathbb{P}}\big[D_{\mathrm{TV}}\left(\mathbb{Q}_m(Y=\cdot|x_0,x),\mathbb{E}_{m'|x}[\mathbb{Q}_{m'}(Y=\cdot|x_0,x)]\right)\big].\tag{77}$$

Consider the set

$$\mathcal{X}_+=\{x\in\mathcal{X}:D_{\mathrm{B}}(\mathbb{Q}_m(Y=\cdot|x_0,x),\mathbb{Q}_l(Y=\cdot|x_0,x))\geqslant\log L,\ \forall m\neq l\}.$$

For any $x\in\mathcal{X}_+$, by Lemma A.6, we have

$$D_{\mathrm{TV}}\left(\mathbb{Q}_m(Y=\cdot|x_0,x),\mathbb{E}_{m'|x}[\mathbb{Q}_{m'}(Y=\cdot|x_0,x)]\right)\geqslant\frac{1}{2}\Big(1-\widetilde{\mathbb{P}}(m|x)\Big).$$

Therefore, combining the above inequality with (77) gives

$$\begin{aligned}4D_{\mathrm{TV}}\left(\mathbb{P},\mathbb{Q}^{\otimes 2}\right)&\geqslant\mathbb{E}_{(m,x)\sim\widetilde{\mathbb{P}}}\big[D_{\mathrm{TV}}\left(\mathbb{Q}_m(Y=\cdot|x_0,x),\mathbb{E}_{m'|x}[\mathbb{Q}_{m'}(Y=\cdot|x_0,x)]\right)\big]\\&\geqslant\frac{1}{2}\mathbb{E}_{(m,x)\sim\widetilde{\mathbb{P}}}\Big[\mathbf{1}\{x\in\mathcal{X}_+\}\left(1-\widetilde{\mathbb{P}}(m|x)\right)\Big]\\&\geqslant\frac{1}{2}\mathbb{E}_x\Big[\mathbf{1}\{x\in\mathcal{X}_+\}\min_m\left(1-\widetilde{\mathbb{P}}(m|x)\right)\Big]\end{aligned}$$

By definition,

$$1-\widetilde{\mathbb{P}}(m|x)=\sum_{l\neq m}\widetilde{\mathbb{P}}(l|x)=\frac{\sum_{l\neq m}p_l\mathbb{Q}_l(x|x_0)}{\mathbb{P}(x)}.$$

Therefore,

$$\begin{aligned}\mathbb{E}_x\Big[\mathbf{1}\{x\in\mathcal{X}_+\}\min_m\left(1-\widetilde{\mathbb{P}}(m|x)\right)\Big]&=\sum_{x\in\mathcal{X}_+}\min_m\sum_{l\neq m}p_l\mathbb{Q}_l(x|x_0)\\&\overset{(76)}{\geqslant}\sum_{x\in\mathcal{X}_+}\min_m\sum_{l\neq m}p_l\cdot\alpha\mu(x)\\&=\alpha\mu(\mathcal{X}_+)(1-\max_m p_m).\end{aligned}$$

It remains to prove that $\mu(\mathcal{X}_+)\geqslant\frac{1}{2}$. For each pair of $m\neq l$, consider the set

$$\mathcal{X}_{m,l}:=\{x\in\mathcal{X}:D_{\mathrm{B}}(\mathbb{Q}_m(Y=\cdot|x_0,x),\mathbb{Q}_l(Y=\cdot|x_0,x))<\log L\}.$$

By definition,

$$\exp\left(-D_{\mathrm{B}}(\mathbb{Q}_m(z=\cdot|x_0),\mathbb{Q}_l(z=\cdot|x_0))\right)$$

$$
\begin{aligned}
&= \sum_{x\in\mathcal{X}} \sqrt{\mathbb{Q}_m(x|x_0)\mathbb{Q}_l(x|x_0)} \exp\left(-D_{\mathrm{B}}(\mathbb{Q}_m(Y=\cdot|x_0,x),\mathbb{Q}_l(Y=\cdot|x_0,x))\right) \\
&> \sum_{x\in\mathcal{X}_{m,l}} \sqrt{\mathbb{Q}_m(x|x_0)\mathbb{Q}_l(x|x_0)} \cdot \frac{1}{L} \\
&\geqslant \alpha\mu(\mathcal{X}_{m,l}) \cdot \frac{1}{L}.
\end{aligned}
$$

Therefore, by the fact that $D_{\mathrm{B}}(\mathbb{Q}_m(z=\cdot|x_0),\mathbb{Q}_l(z=\cdot|x_0)) \geqslant 3\log(L/\alpha)$, we know that $\mu(\mathcal{X}_{m,l}) \leqslant \frac{1}{L}$ for all $m \neq l$, and hence

$$
1-\mu(\mathcal{X}_+) \leqslant \sum_{m<l} \mu(\mathcal{X}_{m,l}) \leqslant \frac{1}{2}.
$$

The proof is completed by combining the inequalities above. □

Proof of Proposition E.13. We only need to demonstrate how to apply Proposition E.14. We abbreviate $W = W_0$ in the following proof. Take $\mathcal{X} = \mathcal{S}$, $\mathcal{Y} = \mathcal{A} \times (\mathcal{S}\times\mathcal{A})^{W-2}$, with variable $x_0 = s_1, Y = (a_1, s_2, \cdots, a_{W-1}), x = s_W$. Let

$$
\mathbb{Q}_m = \mathbb{T}^{\pi_{\mathrm{sep}}}_{\theta,m}((a_1,s_2,\cdots,s_W)=\cdot|s_1=\cdot) \in \mathcal{Q}, \qquad m\in[L].
$$

Then, we can identify $\mathbb{Q}_m^{\otimes 2}$ as

$$
\mathbb{Q}_m^{\otimes 2} = \mathbb{T}^{\pi_{\mathrm{sep}}}_{\theta,m}((a_1,s_2,\cdots,s_{2W-1})=\cdot|s_1=\cdot).
$$

We also have $\mathbb{Q}_m(x|x_0) = \mathbb{T}^{\pi_{\mathrm{sep}}}_{\theta,m}(s_W = s'|s_0 = s)$. Therefore, we can indeed apply Proposition E.14 and the proof is hence completed. □

Appendix F. Proofs for Section 5

F.1. Proof of Theorem 5.2

We first prove the following lemma.

Lemma F.1 *Suppose that the policy $\widehat{\pi}$ is returned by Algorithm 2. Then for any policy π, it holds that*

$$
V(\widehat{\pi}) \geqslant V(\pi) - \mathbb{P}^{\pi}(m^\star \neq m(\overline{\tau}_W)).
$$

Proof. For any policy π and trajectory $\overline{\tau}_h$, we consider the value π given the trajectory $\overline{\tau}_h$:

$$
V^{\pi}(\overline{\tau}_h) := \mathbb{E}^{\pi}\left[\sum_{h'=h}^{H} R_h(s_h,a_h)\middle|\overline{\tau}_h\right].
$$

In particular, for trajectory $\overline{\tau}_W = (s_1, a_1, \cdots, s_W)$, we have

$$
V^{\pi}(\overline{\tau}_W) = \mathbb{E}^{\pi}\left[\sum_{h=W}^{H} R_h(s_h,a_h)\middle|\overline{\tau}_W\right]
$$

$$= \sum_{m\in[L]} \widetilde{\mathbb{P}}(m|\overline{\tau}_W) \cdot \mathbb{E}_m^{\pi(\cdot|\overline{\tau}_W)} \left[\sum_{h=W}^{H} R_h(s_h, a_h) \middle| s_W \right],$$

where the expectation $\mathbb{E}_m^{\pi(\cdot|\overline{\tau}_W)}$ is taken over the probability distribution of $(s_{W+1:H}, a_{W:H})$ induced by executing the policy $\pi(\cdot|\overline{\tau}_W)$ in MDP M_m with starting state s_W. Therefore, because $\widehat{V}_{m,W}$ is exactly the optimal value function in MDP M_m (at step W), we know that

$$\mathbb{E}_m^{\pi(\cdot|\overline{\tau}_W)} \left[\sum_{h=W}^{H} R_h(s_h, a_h) \middle| s_W \right] \leqslant \widehat{V}_{m,W}(s_W).$$

Hence, we have

$$\begin{aligned}
V^{\pi}(\overline{\tau}_W) &\leqslant \sum_{m\in[L]} \widetilde{\mathbb{P}}(m|\overline{\tau}_W)\widehat{V}_{m,W}(s_W) \\
&\leqslant \widetilde{\mathbb{P}}(m(\overline{\tau}_W)|\overline{\tau}_W) \cdot \widehat{V}_{m(\overline{\tau}_W),W}(s_W) + \sum_{m\neq m(\overline{\tau}_W)} \widetilde{\mathbb{P}}(m|\overline{\tau}_W) \\
&= \widehat{V}(\overline{\tau}_W) + \widetilde{\mathbb{P}}(m^\star \neq m(\overline{\tau}_W)|\overline{\tau}_W),
\end{aligned}$$

where the last line follows from the definition of $\widehat{V}$ in Algorithm 2. On the other hand, we also have

$$\begin{aligned}
V^{\widehat{\pi}}(\overline{\tau}_W) &= \sum_{m\in[L]} \widetilde{\mathbb{P}}(m|\overline{\tau}_W) \cdot \mathbb{E}_m^{\widehat{\pi}(\cdot|\overline{\tau}_W)} \left[\sum_{h=W}^{H} R_h(s_h, a_h) \middle| s_W \right] \\
&\geqslant \widetilde{\mathbb{P}}(m(\overline{\tau}_W)|\overline{\tau}_W) \cdot \mathbb{E}_{m(\overline{\tau}_W)}^{\widehat{\pi}(\cdot|\overline{\tau}_W)} \left[\sum_{h=W}^{H} R_h(s_h, a_h) \middle| s_W \right] \\
&= \widetilde{\mathbb{P}}(m(\overline{\tau}_W)|\overline{\tau}_W) \cdot \mathbb{E}_{m(\overline{\tau}_W)} \left[\sum_{h=W}^{H} R_h(s_h, a_h) \middle| s_W, \text{for each } h \geqslant W, a_h = \pi_h^{(m(\overline{\tau}_W))}(s_h) \right] \\
&= \widetilde{\mathbb{P}}(m(\overline{\tau}_W)|\overline{\tau}_W) \cdot \widehat{V}_{m(\overline{\tau}_W),W}(s_W) = \widehat{V}(\overline{\tau}_W),
\end{aligned}$$

where the last line is because $\widehat{V}_{m,W}(s_W)$ is exactly the expected cumulative reward if the agent starts at step W and state s_W, and executes π_m afterwards. Combining the inequalities above, we obtain

$$V^{\pi}(\overline{\tau}_W) - \widetilde{\mathbb{P}}(m^\star \neq m(\overline{\tau}_W)|\overline{\tau}_W) \leqslant V^{\widehat{\pi}}(\overline{\tau}_W).$$

By recursively using the definition of $\widehat{\pi}$, we can show that for each step $h = W, W-1, \cdots, 1$,

$$V^{\pi}(\overline{\tau}_h) - \widetilde{\mathbb{P}}(m^\star \neq m(\overline{\tau}_W)|\overline{\tau}_h) \leqslant V^{\widehat{\pi}}(\overline{\tau}_h).$$

The desired result follows as

$$V(\pi) - \widetilde{\mathbb{P}}^{\pi}(m^\star \neq m(\overline{\tau}_W)) = \mathbb{E}\Big[V^{\pi}(\overline{\tau}_1) - \widetilde{\mathbb{P}}(m^\star \neq m(\overline{\tau}_W)|\overline{\tau}_1)\Big] \leqslant \mathbb{E}\Big[V^{\widehat{\pi}}(\overline{\tau}_1)\Big] = V(\widehat{\pi}).$$

□

Proof of Theorem 5.2 Let $\pi_\star$ be an optimal policy such that M is ϖ-separated under $\pi_\star$. By Proposition 4.1, we know that $\mathbb{P}^{\pi_\star}(m^\star \neq m(\overline{\tau}_W)) \leqslant L\exp(-\varpi(W)) \leqslant \varepsilon$. Therefore, Lemma F.1 implies $V(\widehat{\pi}) \geqslant V(\pi_\star) - \varepsilon = V^\star - \varepsilon$. The time complexity follows immediately from the definition of Algorithm 2. $\square$

F.2. Embedding 3SAT problem to LMDP

Suppose that Φ is a 3SAT formula with n variables $x_1, \cdots, x_n$ and N clauses $C_1, \cdots, C_N$, and $\mathcal{A} = \{0,1\}^w$. Consider the corresponding LMDP M_Φ constructed as follows.

- The horizon length is $H = \lceil n/w \rceil + 1$.
- The state space is $\mathcal{S} = \left\{ s_\ominus^1, s_\ominus^2, \cdots, s_\ominus^{H-1}, s_\oplus \right\}$, and the action space is $\mathcal{A}$.
- $L = N$, and the mixing weight is $\rho = \mathrm{Unif}([N])$.
- For each $m \in [N]$, the MDP M_m is given as follows.
  - The initial state is $s_\ominus^1$.
  - At state $s_\ominus^h$, taking action $a \in \mathcal{A}_{m,h}$ leads to $s_\oplus$, where
  $$\begin{aligned}\mathcal{A}_{m,h} := & \left\{ a \in \{0,1\}^w : \exists j \in [w] \text{ such that } a[j] = 1 \text{ and the clause } C_m \text{ contains } x_{w(h-1)+j} \right\} \\ & \bigcup \left\{ a \in \{0,1\}^w : \exists j \in [w] \text{ such that } a[j] = 0 \text{ and the clause } C_m \text{ contains } \neg x_{w(h-1)+j} \right\}.\end{aligned}$$
  For action $a \notin \mathcal{A}_{m,h}$, taking action a leads to $s_\ominus^{\min\{h+1, H-1\}}$.
- The reward function is given by $R_h(s,a) = \mathbf{1}\{s = s_\oplus, h = H\}$.

The basic property of M_Φ is that, the optimal value of the LMDP M_Φ encodes the satisfiability of the formula Φ. More concretely, if taking an action sequence $a_{1:H-1}$ leads to $s_\oplus$ for all $m \in [N]$, then the first n bits of the sequence $(a_1, \cdots, a_{H-1})$ gives a satisfying assignment of Φ. Conversely, any satisfying assignment of Φ gives a corresponding action sequence such that taking it leads to $s_\oplus$ always. On the other hand, if Φ is not satisfiable, then for any action sequence $a_{1:H-1}$, there must be a latent index $m \in [N]$ such that taking $a_{1:H-1}$ leads to $s_\ominus^{H-1}$ in MDP M_m. To summarize, we have the following fact.

Claim. The optimal value $V^\star$ of M_Φ equals 1 if and only if Φ is satisfiable. Furthermore, when Φ is not satisfiable, $V^\star \leqslant 1 - \frac{1}{m}$.

Based on the LMDP M_Φ described above, we construct a "perturbed" version $\widetilde{M}_\Phi$ that is δ-strongly separated.

- Pick $d = \lceil 11\log(2N) \rceil$ and invoke Lemma F.5 to generates a sequence $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N \in \{-1,+1\}^d$, such that for all $i \neq j$, $i, j \in [N]$,
  $$\|\mathbf{x}_i - \mathbf{x}_j\|_1 \geqslant \frac{d}{2}, \qquad \|\mathbf{x}_i + \mathbf{x}_j\|_1 \geqslant \frac{d}{2}.$$
  We also set $\bar{\delta} = 4\delta$, and for each $m \in [N]$, we define
  $$\mu_m^+ = \left[\frac{1+\bar{\delta}\mathbf{x}_m[1]}{2d}; \frac{1-\bar{\delta}\mathbf{x}_m[1]}{2d}; \cdots; \frac{1+\bar{\delta}\mathbf{x}_m[d]}{2d}; \frac{1-\bar{\delta}\mathbf{x}_m[d]}{2d} \right] \in \Delta([2d]),$$

$$\mu_m^- = \left[\frac{1-\bar{\delta}\mathbf{x}_m[1]}{2d}; \frac{1+\bar{\delta}\mathbf{x}_m[1]}{2d}; \cdots; \frac{1-\bar{\delta}\mathbf{x}_m[d]}{2d}; \frac{1+\bar{\delta}\mathbf{x}_m[d]}{2d}\right] \in \Delta([2d]).$$

- The state space is $\tilde{\mathcal{S}} = \mathcal{S} \times [2d]$, the action space is $\mathcal{A}$, and the horizon length is H.
- $L' = 2N$, and the mixing weight is $\rho' = \mathrm{Unif}([2N])$
- For each $m \in [N]$, we set $\widetilde{M}_{2m-1} = M_m \otimes \mu_m^+$ and $\widetilde{M}_{2m} = M_m \otimes \mu_m^-$ (recall our definition in Definition D.6).
- The reward function is given by $R_h((s,o),a) = \mathbf{1}\{s = s_\oplus, h = H\}$.

Proposition F.2 *In the LMDP $\widetilde{M}_\Phi$ described above, for any policy class Π that contains $\mathcal{A}^H$, we have*

$$\max_{\pi \in \Pi} V(\pi) = \begin{cases} 1, & \Phi \textit{ is satisifiable}, \\ \leqslant 1 - \frac{(1-\bar{\delta}^2)^{(H-1)/2}}{N}, & \textit{otherwise}. \end{cases}$$

Proof. By our construction, regardless of the actions taken, we always have $\tilde{s}_H[1] = s_\oplus$ or $\tilde{s}_H[1] = s_\ominus^{H-1}$. Therefore, for any policy π,

$$V(\pi) = \mathbb{P}^\pi(\tilde{s}_H[1] = s_\oplus) = 1 - \mathbb{P}^\pi(\tilde{s}_H[1] = s_\ominus^{H-1}).$$

By construction, any reachable trajectory that ends with $\tilde{s}_H[1] = s_\ominus^{H-1}$ must take the form

$$(s_\ominus^1, o_1), a_1, \cdots, (s_\ominus^{H-1}, o_{H-1}), a_{H-1}, (s_\ominus^{H-1}, o_H).$$

Further, for each $m \in [N]$, in the MDP $\widetilde{M}_{2m-1}$ and $\widetilde{M}_{2m}$, $\tilde{s}_H[1] = s_\ominus^{H-1}$ if and only if $a_{1:H-1} \notin \mathbb{A}_{\mathsf{sat},m}$, where we define

$$\mathbb{A}_{\mathsf{sat},m} = \left\{a_{1:H-1} \in \mathcal{A}^{H-1} : \text{for some } h \in [H-1], a_h \in \mathcal{A}_{m,h}\right\} \subset \mathcal{A}^{H-1}.$$

Therefore, for any reachable trajectory τ_{H-1} that leads to $\tilde{s}_H[1] = s_\ominus^{H-1}$, we have

$$\begin{aligned}
\tau_{H-1} &= ((s_\ominus^1, o_1), a_1, \cdots, (s_\ominus^{H-1}, o_{H-1}), a_{H-1}), \\
\mathbb{P}^\pi(\tau_{H-1}) &= \frac{1}{2N} \sum_{l=1}^{2N} \mathbb{P}^\pi_{\widetilde{M}_l}(\tau_{H-1}) \\
&= \frac{1}{N} \sum_{m=1}^{N} \mathbf{1}\{a_{1:H-1} \in \mathbb{A}_{\mathsf{sat},m}\} \cdot \pi(\tau_{H-1}) \cdot \left(\prod_{h=1}^{H-1} \mu_m^+(o_h) + \prod_{h=1}^{H-1} \mu_m^-(o_h)\right)
\end{aligned}$$

where by convention we write

$$\pi(\tau_{H-1}) = \prod_{h=1}^{H-1} \pi(a_h | (s_\ominus^1, o_1), a_1, \cdots, (s_\ominus^h, o_h)),$$

and we abbreviate this quantity as $p_\pi(a_{1:H}|o_{1:H})$. Then, we have

$$1 - V(\pi) = \mathbb{P}^\pi(\tilde{s}_H[1] = s_\ominus^{H-1})$$

$$= \sum_{\substack{\text{reachable } \tau_{H-1} \text{ that} \\ \text{leads to } \tilde{s}_H[1]=s_{\ominus}^{H-1}}} \mathbb{P}^{\pi}(\tau_{H-1})$$

$$= \sum_{(o_{1:H-1},a_{1:H-1})} \frac{1}{2m}\sum_{i=1}^{m} \mathbf{1}\{a_{1:H-1}\in\mathbb{A}_{\mathsf{sat},m}\}\cdot p_\pi(a_{1:H}|o_{1:H})\cdot\left(\prod_{h=1}^{H-1}\mu_m^+(o_h)+\prod_{h=1}^{H-1}\mu_m^-(o_h)\right).$$

By Lemma F.3, it holds that

$$\prod_{h=1}^{H-1}\mu_m^+(o_h)+\prod_{h=1}^{H-1}\mu_m^-(o_h) \geqslant \frac{2(1-\bar{\delta}^2)^{\lfloor (H-1)/2\rfloor}}{(2d)^{H-1}}.$$

Hence, we have

$$\begin{aligned}
1-V(\pi) &\geqslant \frac{1}{m}\sum_{i=1}^{m}\sum_{(o_{1:H-1},a_{1:H-1})}\mathbf{1}\{a_{1:H-1}\in\mathbb{A}_{\mathsf{sat},m}\}\cdot p_\pi(a_{1:H}|o_{1:H})\cdot\frac{2(1-\bar{\delta}^2)^{\lfloor (H-1)/2\rfloor}}{(2d)^{H-1}} \\
&= (1-\bar{\delta}^2)^{\lfloor (H-1)/2\rfloor}\sum_{a_{1:H-1}}\frac{\#\{m\in[N]: a_{1:H-1}\notin\mathbb{A}_{\mathsf{sat},m}\}}{N}\times\frac{1}{(2d)^H}\sum_{o_{1:H-1}}p_\pi(a_{1:H}|o_{1:H}) \\
&\geqslant (1-\bar{\delta}^2)^{\lfloor (H-1)/2\rfloor}\cdot\min_{a_{1:H-1}}\frac{\#\{m\in[N]: a_{1:H-1}\notin\mathbb{A}_{\mathsf{sat},m}\}}{N}\cdot\sum_{a_{1:H-1}}\frac{1}{(2d)^H}\sum_{o_{1:H-1}}p_\pi(a_{1:H}|o_{1:H}) \\
&= (1-\bar{\delta}^2)^{\lfloor (H-1)/2\rfloor}\cdot\min_{a_{1:H-1}}\frac{\#\{m\in[N]: a_{1:H-1}\notin\mathbb{A}_{\mathsf{sat},m}\}}{N},
\end{aligned}$$

where the last line is because

$$\sum_{a_{1:H-1}}\sum_{o_{1:H-1}}p_\pi(a_{1:H}|o_{1:H}) = (2d)^H.$$

Therefore, if Φ is not satisfiable, then for any action sequence $a_{1:H}$, there must exist $m\in[N]$ such that $a_{1:H}\notin\mathbb{A}_{\mathsf{sat},m}$. This is because if $a_{1:H}\in\mathbb{A}_{\mathsf{sat},m}$ for all $m\in[N]$, then the first n bits of the sequence $(a_1,\cdots,a_{H-1})$ gives a satisfying assignment of Φ. Thus, in this case, for any policy π,

$$1-V(\pi)\geqslant\frac{(1-\bar{\delta}^2)^{\lfloor (H-1)/2\rfloor}}{m}.$$

On the other hand, if Φ is satisfiable, then there is an action sequence $a_{1:H-1}\in\mathbb{A}_{\mathsf{sat},m}$ for all $m\in[N]$, and hence $V(a_{1:H-1})=1$. Combining these complete the proof. □

Lemma F.3 *For any reals $\lambda_1,\cdots,\lambda_k\in[-1,1]$ and $\delta\in[0,1)$, it holds that*

$$\prod_{i=1}^{k}(1+\delta\lambda_i)+\prod_{i=1}^{k}(1-\delta\lambda_i)\geqslant 2(1-\delta^2)^{\lfloor k/2\rfloor}.$$

Proof. Notice that the LHS is a linear function of λ_i for each i (fixing other λ_j's). Therefore, we only need to consider the case $\lambda_i\in\{-1,1\}$. Suppose that $\lambda_1,\cdots,\lambda_k$ has r many 1's and s many -1's ($r+s=k$), and w.l.o.g $r\geqslant s$. Then for $t=r-s\geqslant 0$,

$$\prod_{i=1}^{k}(1+\delta\lambda_i)+\prod_{i=1}^{k}(1-\delta\lambda_i)=(1+\delta)^r(1-\delta)^s+(1+\delta)^s(1-\delta)^r$$

$$
\begin{aligned}
&= (1-\delta^2)^s\left[(1+\delta)^t + (1-\delta)^t\right] \\
&\geqslant 2(1-\delta^2)^s \geqslant 2(1-\delta^2)^{\lfloor k/2 \rfloor}.
\end{aligned}
$$

□

F.3. Proof of Proposition 5.1

Suppose that a 3SAT formula Φ with n variables and N clauses are given. Then, we can pick $w = 1$, $\delta = \frac{1}{\sqrt{n}}$, $\varepsilon = \frac{c}{N}$ for some small constant c, and the LMDP $\widetilde{M}_\Phi$ constructed above has $H = n+1$, $L = 2N$, $S = Hd$, $A = 2$, and it is δ-strongly separated. Further, we have $\max\left\{L, S, A, H, \varepsilon^{-1}, \delta^{-1}\right\} \leqslant \mathcal{O}(n+N)$, and $\widetilde{M}_\Phi$ can be computed in $\mathrm{poly}(n, N)$ time. Therefore, if we can solve any given δ-strong separated LMDP in polynomial time, we can determine the satisfiability of any given 3SAT formula Φ in polynomial time by solving $\widetilde{M}_\Phi$, which implies that NP=P. □

F.4. Proof of Theorem 5.4

Suppose that there is an algorithm $\mathfrak{A}$ that contradicts the statement of Theorem 5.4.

Fix a given 3-SAT formula Φ with n variables and N clauses is given (we assume $N \leqslant n^3$ without loss of generality), we proceed to determine the satisfiability of Φ in $2^{o(n)}$-time using $\mathfrak{A}$.

Pick $t = t_n \in \mathbb{N}$ to be the minimal integer such that

$$
200n \leqslant \frac{\log(1/\varepsilon_t) \cdot \lfloor \log A_t \rfloor}{\delta_t^2}. \tag{78}
$$

We then consider $\varepsilon = \varepsilon_t$, $w = \lfloor \log A_t \rfloor$, $A = 2^w$, $\delta = \frac{1}{\delta_t}$, and $\mathcal{A} = \{0,1\}^w$.

Now, consider the LMDP $\widetilde{M}_\Phi$ constructed in Appendix F.2 based on $(\Phi, \mathcal{A}, \delta)$. We know that $\widetilde{M}_\Phi$ is δ-strongly separated, and we also have

$$
L = 2N \leqslant 2n^3, \qquad S = nd \leqslant \mathcal{O}(n \log n), \qquad H = \lceil \frac{n}{w} \rceil + 1 \leqslant n+1.
$$

In the following, we show that (9) and (78) (with suitably chosen C) ensures that

$$
\varepsilon < \varepsilon' := \frac{(1-\bar{\delta}^2)^{(H-1)/2}}{3N}.
$$

By definition,

$$
\log(1/\varepsilon') = \frac{(H-1)\log\frac{1}{1-\bar{\delta}^2}}{2} + \log(3N) \leqslant \frac{2\bar{\delta}^2}{1-\bar{\delta}^2}\lceil \frac{n}{w} \rceil + \log(3N) \leqslant \frac{128\delta^2}{3}\frac{n}{w} + 3\log(n) + 4.
$$

Therefore, by (78), we have $\log(1/\varepsilon') < \log(1/\varepsilon)$ if we have $\frac{3}{4}\log(1/\varepsilon) > 3\log n + 4$, or equivalently $e^6 n^4 \leqslant \varepsilon^{-1}$. This is indeed insured by (9).

Next, consider running $\mathfrak{A}$ on $(\widetilde{M}_\Phi, \varepsilon)$, and let $\widehat{V}$ be the value returned by $\mathfrak{A}$. By Proposition F.2, we have the follow facts: (a) If $\widehat{V} \geqslant 1-\varepsilon$, then Φ is satisfiable. (b) If $\widehat{V} < 1-\varepsilon$, then Φ is not satisfiable. Therefore, we can use $\mathfrak{A}$ to determine the satisfiability of Φ in time $A^{o(\delta^{-2}\log(1/\varepsilon))} + \mathrm{poly}(n)$. Notice that our choice of t ensures that $\log(1/\varepsilon_t) w \delta_t^{-2} \leqslant 3200n$, and hence we actually determine the satisfiability of Φ in $2^{o(n)}$-time, which contradicts Conjecture 5.3. □

F.5. Proof of Theorem 5.5

Suppose that there is an algorithm $\mathfrak{A}$ that contradicts the statement of Theorem 5.5.

Fix a given 3-SAT formula Φ with n variables and N clauses is given (we assume $N \leqslant n^3$ without loss of generality), we proceed to determine the satisfiability of Φ in $2^{o(n)}$-time using $\mathfrak{A}$.

Pick $t = t_n \in \mathbb{N}$ to be the minimal integer such that

$$Cn\lceil \log_2 N \rceil \leqslant \frac{\log L_t \cdot \lfloor \log A_t \rfloor}{\delta_t^2}, \tag{79}$$

where C is a large absolute constant. We then consider $L = 2^{\log L_t}$, $w = \lfloor \log A_t \rfloor$, $A = 2^w$, $\delta = \frac{1}{\delta_t}$, and $\mathcal{A} = \{0,1\}^w$.

Let M_Φ be the LMDP with action set $\mathcal{A}$, horizon $H = \lceil n/w \rceil + 1$ constructed in Appendix F.2.

Further, we choose $r = \lceil \log_2 N \rceil$, $d = \lfloor \frac{\log L_t}{r} \rfloor$. By our choice (79), we can ensure the presumption $d \geqslant C_0 H \delta^2$ of Lemma F.6 holds, which implies that we can construct a $(N, H, \delta, r2^{-c_0 d}, 2^{dr})$-family over $[2d]^r$ in time $\mathrm{poly}(2^{dr}) \leqslant \mathrm{poly}(L)$. Denote $\mathcal{Q}$ be such a family, and we consider $M_\Phi \otimes \mathcal{Q}$, which is a δ-strongly separated LMDPs family with $S = (2d)^r H$ and hence $\log S \leqslant \mathcal{O}(\log\log L_t)$ by (10) (because $n \leqslant \mathrm{poly}\log L_t$ using (79)).

Consider running $\mathfrak{A}$ on $M_\Phi \otimes \mathcal{Q}$ with $\varepsilon = \frac{1}{3N}$, and let $\widehat{V}$ be the value returned by $\mathfrak{A}$. Let V_Φ be the optimal value of M_Φ, $V_{M,\Phi}$ be the optimal value of $M_\Phi \otimes \Phi$. Then by Proposition D.11, it holds that

$$V_\Phi \leqslant V_{M,\Phi} \leqslant r2^{-c_0 d} + V_\Phi.$$

Hence, as long as $r2^{-c_0 d} < \frac{1}{3N}$ (which is ensured by condition (10)), we have the follow facts: (a) If $V_\Phi = 1$, then $\widehat{V} \geqslant 1 - \frac{1}{3N}$. (b) If $V_\Phi \leqslant 1 - \frac{1}{N}$, then $\widehat{V} < 1 - \frac{1}{3N}$. Notice that a special case of Proposition F.2 is that, when Φ is satisfiable, then $V_\Phi = 1$, and otherwise $V_\Phi \leqslant 1 - \frac{1}{N}$. Therefore, we can use $\mathfrak{A}$ to determine the satisfiability of Φ in time $A^{o\left(\delta^{-2}\frac{\log L}{\log\log L}\right)} + \mathrm{poly}(L)$. Notice that our choice of t ensures that $(\log L_t)(\lfloor \log A_t \rfloor)\delta_t^{-2} \leqslant 16Cn\lceil \log_2 N \rceil$, and hence $\log L = o(n)$, and

$$\frac{\log A \log L}{\delta^2 \log\log L} = \mathcal{O}(n).$$

Therefore, given $\mathfrak{A}$, we can construct a $2^{o(n)}$-time algorithm for 3SAT, a contradiction. □

F.6. Technical lemmas

Lemma F.4 *There is a procedure such that, for any input integer $N \geqslant 2$ and $d \geqslant \lceil 11 \log N \rceil$, compute a sequence $\mathbf{x}_1, \cdots, \mathbf{x}_N \in \{-1,+1\}^d$ such that $\|\mathbf{x}_i - \mathbf{x}_j\|_1 \geqslant \frac{d}{2} \forall i \neq j$, with running time* $\mathrm{poly}(2^d)$.

Proof. Consider the following procedure: We maintain two set $\mathcal{U}, \mathcal{V}$, and we initialize $\mathcal{U} = \{\}, \mathcal{V} = \{-1,1\}^d$. At each step, we pick a $\mathbf{x} \in \mathcal{V}$, add $\mathbf{x}$ to $\mathcal{U}$, and remove all $\mathbf{y} \in \mathcal{V}$ such that $\|\mathbf{y} - \mathbf{x}\|_1 < \frac{d}{2}$. The procedure ends when $\mathcal{V}$ is empty or $|\mathcal{U}| = N$.

We show that this procedure must end with $|\mathcal{U}| = N$. Notice that for any $\mathbf{x}, \mathbf{y} \in \{-1, 1\}^d$, we have $\|\mathbf{x} - \mathbf{y}\|_1 < \frac{d}{2}$ only when $\mathbf{x}, \mathbf{y}$ differs by at most $i < \frac{d}{4}$ coordinates. Therefore, at each step, we remove at most

$$M = \sum_{i=0}^{\lceil d/4 \rceil - 1} \binom{d}{i}$$

elements in $\mathcal{V}$. Hence, it remains to show that $\frac{2^d}{M} \geqslant N$.

Denote $k = \lceil d/4 \rceil - 1$. Then we have

$$M = \sum_{i=0}^{d} \binom{d}{i} \leqslant \left(\frac{ed}{k}\right)^k \leqslant \left(\frac{ed}{d/4}\right)^{d/4} = \exp\left(\frac{1 + 2\log 2}{4} d\right),$$

and hence $\frac{2^d}{M} > \exp(d/11) \geqslant N$ as claimed. □

Repeating the argument above, we can also prove the following result.

Lemma F.5 *There is a procedure such that, for any input integer $N \geqslant 2$ and $d \geqslant \lceil 11 \log(2N) \rceil$, compute a sequence $\mathbf{x}_1, \cdots, \mathbf{x}_N \in \{-1, +1\}^d$ such that for any $i \neq j$,*

$$\|\mathbf{x}_i - \mathbf{x}_j\|_1 \geqslant \frac{d}{2}, \qquad \|\mathbf{x}_i + \mathbf{x}_j\|_1 \geqslant \frac{d}{2}$$

with running time $\mathrm{poly}(2^d)$.

Lemma F.6 *There is a procedure such that, for any input $r, d, H \geqslant 2$ and $\delta \in (0, \frac{1}{4}]$ satisfying $d \geqslant C_0 H \delta^2$, compute a $(2^r, H, \delta, \gamma, 2^{dr})$-family over $[2d]^r$, with $\gamma \leqslant r 2^{-c_0 d}$, with running time* $\mathrm{poly}(2^{dr})$.

Proof. We first invoke the procedure of Lemma F.4 to compute $\mathbf{x}_1, \cdots, \mathbf{x}_N \in \{-1, 1\}^d$ such that $\|\mathbf{x}_i - \mathbf{x}_j\|_1 \geqslant \frac{d}{2}$ and $N > \exp(d/11)$. Consider the distribution $\mu_i = \mathbb{Q}_{\bar{\delta}\mathbf{x}_i} \in \Delta([2d])$ for each $i \in [N]$, where we set $\bar{\delta} = 4\delta$. Clearly, we have $D_{\mathrm{TV}}(\mu_i, \mu_j) \geqslant \delta$ for $i \neq j$.

Notice that for $K = \lceil d/60 \rceil$, we have $N > \binom{K+d-1}{d} + 1$, and hence by Corollary D.16, there exists $\xi_0, \xi_1 \in \Delta([N])$ such that $\mathrm{supp}(\xi_0) \cup \mathrm{supp}(\xi_1) = \varnothing$ and

$$D_{\mathrm{TV}}^2\left(\mathbb{E}_{i\sim\xi_0}[\mu_i^{\otimes n}], \mathbb{E}_{i\sim\xi_1}[\mu_i^{\otimes n}]\right) \leqslant \sum_{k=K}^{H} \left(\frac{eH\bar{\delta}^2}{K}\right)^k.$$

Therefore, as long as $d \geqslant 120eH\bar{\delta}^2$, $\mathcal{Q} = \{(\xi_0, \xi_1), (\mu_1, \cdots, \mu_N)\}$ is a $(2, H, \delta, 2^{-\frac{K-1}{2}}, N)$-family over $[2d]$. Further, invoking Lemma D.9 yields $\mathcal{Q}'$, a $(2^r, H, \delta, r2^{-\frac{K-1}{2}}, N^r)$-family over $[2d]^r$.

By the proof of Corollary D.16, ξ_0, ξ_1 can be computed in $\mathrm{poly}(N)$ time, and $\mathcal{Q}'$ can also be computed from $\mathcal{Q}$ in time $\mathrm{poly}(2^{dr})$ by going through the proof of Lemma D.9. Combining the results above completes the proof. □