Proceedings of Machine Learning Research vol 201:1–42, 2023 34th International Conference on Algorithmic Learning Theory

Adversarial Online Multi-Task Reinforcement Learning

Quan Nguyen MANHQUAN233@GMAIL.COM
Department of Computer Science, University of Victoria

Nishant A. Mehta NMEHTA@UVIC.CA
Department of Computer Science, University of Victoria

Editors: Shipra Agrawal and Francesco Orabona

Abstract

We consider the adversarial online multi-task reinforcement learning setting, where in each of K episodes the learner is given an unknown task taken from a finite set of M unknown finite-horizon MDP models. The learner's objective is to minimize its regret with respect to the optimal policy for each task. We assume the MDPs in $\mathcal{M}$ are well-separated under a notion of λ-separability, and show that this notion generalizes many task-separability notions from previous works. We prove a minimax lower bound of $\Omega(K\sqrt{DSAH})$ on the regret of any learning algorithm and an instance-specific lower bound of $\Omega(\frac{K}{\lambda^2})$ in sample complexity for a class of *uniformly good* cluster-then-learn algorithms. We use a novel construction called *2-JAO MDP* for proving the instance-specific lower bound. The lower bounds are complemented with a polynomial time algorithm that obtains $\tilde{O}(\frac{K}{\lambda^2})$ sample complexity guarantee for the clustering phase and $\tilde{O}(\sqrt{MK})$ regret guarantee for the learning phase, indicating that the dependency on K and $\frac{1}{\lambda^2}$ is tight.

Keywords: reinforcement learning, adversarial online multi-task learning, clustering

1. Introduction

The majority of theoretical works in online reinforcement learning (RL) have focused on single-task settings in which the learner is given the same task in every episode. In practice, an autonomous agent might face a sequence of different tasks. For example, an automatic medical diagnosis system could be given an arbitrarily ordered sequence of patients who are suffering from an unknown set of variants of a virus. In this example, the system needs to classify and learn the appropriate treatment for each variant of the virus. This example is an instance of the adversarial online multi-task episodic RL setting, an important learning setting for which the theoretical understanding is rather limited.

The framework commonly used in existing theoretical works is an episodic setting of K episodes; in each episode an unknown Markov decision process (MDP) from a finite set $\mathcal{M}$ of size M is given to the learner. When $M = 1$, the setting reduces to single-task episodic RL. Most existing algorithms for single-task episodic RL are based on aggregating samples in all episodes to obtain sub-linear bounds on various notions of regret (Azar et al., 2017; Jin et al., 2018; Simchowitz and Jamieson, 2019) or finite (ϵ, δ)-PAC bounds on the sample complexity of exploration (Dann and Brunskill, 2015). When $M > 1$, without any assumptions on the common structure of the tasks, aggregating samples from different tasks could produce negative transfer (Brunskill and Li, 2013). To avoid negative transfer, existing works (Brunskill and Li, 2013; Hallak et al., 2015; Kwon et al., 2021) assumed that there exists some notion of task-separability that defines how different the tasks in $\mathcal{M}$ are. Based on this notion of separability, most existing algorithms followed a two-phase cluster-then-learn paradigm that first attempts to figure out which MDP is being given and then uses the samples from the previous episodes of the same MDP for learning. However, most existing works employ strong assumptions such that the tasks are given stochastically following a fixed distribution (Azar et al., 2013; Brunskill and Li, 2013; Steimle et al., 2021; Kwon et al., 2021) or the task-separability notion allows the MDPs to be distinguished in a small number of exploration steps (Hallak et al., 2015; Kwon et al., 2021). These strong assumptions become the main theoretical challenges towards understanding this setting.

Our goal in this work is to study the adversarial setting with a more general task-separability notion, in which the aformentioned strong assumptions do not hold. Specifically, the learner makes no statistical assumptions on the sequence of tasks; the task in each episode can be either the same or different from the tasks in any other episodes. Moreover, the difference between the tasks in two consecutive episodes can be large (linear in the length of the episodes) so that algorithms based on a fixed budget for total variation such as RestartQ-UCB (Mao et al., 2021) cannot be applied. The performance of the learner is measured by its regret with respect to an omniscient agent that knows which tasks are coming in every episode and the optimal policies for these tasks. We consider the same cluster-then-learn paradigm of the previous works and focus on the following two questions:

- *Is there a task-separability notion that generalizes the notions from previous works while still enabling tasks to be distinguished by a cluster-then-learn algorithm with polynomial time and sample complexity? If so, what is the optimal sample complexity of clustering under this notion?*
- *Is there a polynomial time cluster-then-learn algorithm that simultaneously obtains near-optimal sample complexity in the clustering phase and near-optimal regret guarantee for the learning phase in the adversarial setting?*

We answer both questions positively. For the first question, we introduce the notion of λ-separability, a task-separability notion that generalizes the task-separability definitions in previous works in the same setting (Brunskill and Li, 2013; Hallak et al., 2015; Kwon et al., 2021). Definition 1 formally defines λ-separability. A more informal version of λ-separability has appeared in the discounted setting of Concurrent PAC RL (Guo and Brunskill, 2015) where multiple MDPs are learned concurrently; however the implications on the episodic sequential setting and the tightness of their results were lacking. In essence, λ-separability assumes that between every pair of MDPs in $\mathcal{M}$, there exists some state-action pair whose transition functions are well-separated in ℓ_1-norm. This setting is more challenging than the one considered by Hallak et al. (2015) where *all* state-action pairs are well-separated. In Appendix B, we show that λ-separability is more general than the entropy-based separability defined in Kwon et al. (2021) and thus requires novel approaches to exploring and clustering samples from different episodes. Under this notion of λ-separability, we show an instance-specific lower bound[1] $\Omega(\frac{K}{\lambda^2})$ on both the sample complexity and regret of the clustering phase for a class of cluster-then-learn algorithms that includes most of the existing works.

To answer the second question, we propose a new cluster-then-learn algorithm, AOMultiRL, which obtains a regret upper bound of $\tilde{O}\left(\frac{K}{\lambda^2} + \sqrt{MK}\right)$ (the $\tilde{O}$ hides logarithmic terms). This upper bound indicates that the linear dependency on K and λ^2 in the lower bounds are tight. The $\tilde{O}(\sqrt{MK})$ upper bound in the learning phase is near-optimal because if the identity of the model is revealed to a learner at the beginning of every episode (so that no clustering is necessary), there

1. Here and throughout the introduction, we suppress factors related to the MDPs such that the number of states and actions and the horizon length in all the bounds.

exists a straightforward $\Omega(\sqrt{MK})$ lower bound obtained by combining the lower bound for the single-task episodic setting of Domingues et al. (2021) and Cauchy-Schwarz inequality. In the stochastic setting, the L-UCRL algorithm (Kwon et al., 2021) obtains $O(\sqrt{MK})$ regret with respect to the optimal policy of a partially observable MDP (POMDP) setting that does not know the identity of the MDPs in each episode; thus their notion of regret is weaker than the one in our work.

Overview of Techniques

- In Section 3, we present two lower bounds. The first is a minimax lower bound $\Omega(K\sqrt{SAH})$ on the total regret of any algorithm. This result uses the construction of JAO MDPs in Jaksch et al. (2010). The second is a $\Omega\left(\frac{K}{\lambda^2}\right)$ instance-specific lower bound on the sample complexity and regret of the clustering phase for a class of *uniformly good* cluster-then-learn algorithms when both λ and M are sufficiently large. The instance-specific lower bound relies on the novel construction of *2-JAO MDP*, a hard instance combining two JAO MDPs in which one is the minimax lower bound instance and the other satisfies λ-separability. We show that learning 2-JAO MDPs is fundamentally a two-dimensional extension of the problem of finding a biased coin among a collection of fair coins (e.g. Tulsiani, 2014), for which information theoretic techniques of the one-dimensional problem can be adapted.
- In Section 4, we show that AOMultiRL obtains a regret upper bound of $\tilde{O}\left(\frac{K}{\lambda^2} + \sqrt{MK}\right)$. The main idea of AOMultiRL is based on the observation that a fixed horizon of order $\Theta(\frac{1}{\lambda^2})$ with a small constant factor is sufficient to obtain a λ-dependent coarse estimate of the transition functions of all state-action pairs. In turn, this coarse estimate is sufficient to have high-probability guarantees for the correctness of the clustering phase. This allows AOMultiRL to have a fixed horizon for the learning phase and be able to apply single-task RL algorithms with theoretical guarantees such as UCBVI-CH (Azar et al., 2017) in the learning phase.

Our paper is structured as follows: Section 2 formally sets up the problem. Section 3 presents the lower bounds. AOMultiRL and its regret upper bound are shown in Section 4. The appendix contains formal proofs of all results. We defer detailed discussion on related works to Appendix A.

2. Problem Setup

Our learning setting consists of K episodes. In episode $k = 1, 2, \ldots, K$, an adversary chooses an unknown Markov decision process (MDP) m^k from a set of finite-horizon tabular stationary MDP models $\mathcal{M} = \{(\mathcal{S}, \mathcal{A}, H, P_i, r) : i = 1, 2, \ldots, M\}$ where $r : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ is the shared reward function, $\mathcal{S}$ is the set of states with size S, $\mathcal{A}$ is the set of actions with size A, H is the length of each episode, and $P_i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ is the transition function where $P_i(s'|s, a)$ specifies the probability of being in state s' after taking action a at state s. The state space $\mathcal{S}$ and action space $\mathcal{A}$ are known and shared between all models; however, the transition functions are distinct and unknown. Following a common practice in single-task RL literature (Azar et al., 2017; Jin et al., 2018), we assume that the reward function is known and deterministic, however our techniques and results extend to the setting of unknown stochastic r. Furthermore, the MDPs are assumed to be communicating with a finite diameter D (Jaksch et al., 2010). A justification for this assumption on the diameter is provided in Section 2.1.

The adversary also chooses the initial state s_1^k. The policy π_k of the learner in episode k is a collection of H functions $\pi^k = \{\pi_{k,h} : \mathcal{S} \mapsto \mathcal{A}\}$, which can be non-stationary and history-dependent. The value function of π_k starting in state s at step h is the expected rewards obtained by following π_k for $H - h + 1$ steps $V_h^{\pi_k}(s) = \mathbb{E}[\sum_{h'=h}^{H} r(s_{h'}^k, \pi_h^k(s_{h'}^k)) \mid s_h^k = s]$, where the expectation is taken with respect to the stochasticity in m^k and π^k. Let $V_1^{k,*}$ denote the value function of the optimal policy in episode k.

The performance of the learner is measured by its regret with respect to the optimal policies in every episode:

$$\text{Regret}(K) = \sum_{k=1}^{K} [V_1^{k,*} - V_1^{\pi_k}](s_1^k). \tag{1}$$

Let $[M] = \{1, 2, \dots, M\}$. We assume that the MDPs in $\mathcal{M}$ are λ-*separable*:

Definition 1 (λ-separability) *Let $\lambda > 0$ and consider set of MDP models $\mathcal{M} = \{m_1, \dots, m_M\}$ with M models. For all $(i,j) \in [M] \times [M]$ and $i \neq j$, the λ-distinguishing set for two models m_i and m_j is defined as the set of state-action pairs such that the ℓ_1 distance between $P_i(s,a)$ and $P_j(s,a)$ is larger than λ: $\Gamma_{i,j}^{\lambda} = \{(s,a) \in \mathcal{S} \times \mathcal{A} : \|P_i(s,a) - P_j(s,a)\| \geq \lambda\}$, where $\|\cdot\|$ denotes the ℓ_1-norm and $P_i(s,a) = P_i(\cdot \mid s,a)$.*

The set $\mathcal{M}$ is λ-separable if for every two models m_i, m_j in $\mathcal{M}$, the set $\Gamma_{i,j}^{\lambda}$ is non-empty:

$$\forall i,j \in [M], i \neq j : \Gamma_{i,j}^{\lambda} \neq \emptyset.$$

In addition, λ is called a separation level of $\mathcal{M}$, and we say a state-action pair (s,a) is λ-distinguishing for two models m_i and m_j if $\|P_i(s,a) - P_j(s,a)\| > \lambda$.

We use the following notion of a λ-distinguishing set for a collection of MDP models $\mathcal{M}$:

Definition 2 (λ-distinguishing set) *Given a λ-separable set of MDPs $\mathcal{M}$, a λ-distinguishing set of $\mathcal{M}$ is a set of state-action pairs $\Gamma^{\lambda} \subseteq \mathcal{S} \times \mathcal{A}$ such that for all $i,j \in [M], \Gamma_{i,j}^{\lambda} \cap \Gamma^{\lambda} \neq \emptyset$. In particular, the set $\Gamma = \cup_{i,j} \Gamma_{i,j}^{\lambda}$ is a λ-distinguishing set of $\mathcal{M}$.*

By definition, a state-action pair can be λ-distinguishing for some pairs of models and not λ-distinguishing for other pairs of models.

2.1. Assumption on the finite diameter of the MDPs

In this work, all MDPs are assumed to be communicating. We employ the following formal definition and assumption commonly used in literature (Jaksch et al., 2010; Brunskill and Li, 2013; Sun and Huang, 2020; Tarbouriech et al., 2021):

Definition 3 *((Jaksch et al., 2010)) Given an ergodic Markov chain $\mathcal{F}$, let $T_{s,s'}^{\mathcal{F}} = \inf\{t > 0 \mid s_t = s', s_0 = s\}$ be the first passage time for two states s, s' on $\mathcal{F}$. Then the hitting time of a unichain MDP G is $T_G = \max_{s,s' \in S} \max_{\pi} \mathbb{E}[T_{s,s'}^{\mathcal{F}_\pi}]$, where $\mathcal{F}_\pi$ is the Markov chain induced by π on G. In addition, $T'_G = \max_{s,s' \in S} \min_{\pi} \mathbb{E}[T_{s,s'}^{\mathcal{F}_\pi}]$ is the diameter of G.*

Assumption 4 *The diameter of all MDPs in $\mathcal{M}$ are bounded by a constant D.*

While this finite diameter assumption is common in undiscounted and discounted single-task setting (Jaksch et al., 2010; Guo and Brunskill, 2015), it is not necessary in the episodic single-task setting (Jin et al., 2018; Mao et al., 2021). Therefore, it is important to justify this assumption in the episodic multi-task setting. In the episodic single-task setting, for any initial state s_1, the average time between any pair of states reachable from s_1 is bounded $2H$; hence, H plays the same role as D (Domingues et al., 2021). This allows the learner to visit and gather state-transition samples in each state multiple times and construct accurate estimates of the model.

However, in the multi-task setting, the same initial state s_1 in one episode might belong to a different MDP than the state s_1 in the previous episodes. Therefore, the set of reachable states and their state-transition distributions could change drastically. Hence, it is important that the λ-distinguishing state-action pairs be reachable from any initial state s_1 for the learner to recognize which MDP it is in and use the samples appropriately. Otherwise, combining samples from different MDPs could lead to negative transfer. Conversely, if the MDPs are allowed to be non-communicating, the component that makes them λ-separable might be unreachable from other components. In this case, the adversary can pick the initial states in these components and block the learner from accessing the λ-distinguishing state-actions. A construction that formalizes this argument is shown at the end of Section 3.

3. Minimax and Instance-Dependent Lower Bounds

We first show that if λ is sufficiently small and $M = \Theta(SA)$, then the setting is uninteresting in the sense that one cannot do much better than learning every episode individually without any transfer, leading to an expected regret that grows linearly in the number of episodes K.

Lemma 5 (Minimax Lower Bound) *Suppose $S, A \geq 10, D \geq 20 \log_A(S)$ and $H \geq DSA$ are given. Let $\lambda = \Theta(\sqrt{\frac{SA}{HD}})$. There exists a set of λ-separable MDPs $\mathcal{M}$ of size $M = \frac{SA}{4}$, each with S states, A actions, diameter at most D and horizon H such that if the tasks are chosen uniformly at random from $\mathcal{M}$, the expected regret of any sequence of policies $(\pi_k)_{k=1,\ldots,K}$ over K episodes is*

$$\mathbb{E}[\text{Regret}(K)] \geq \Omega\left(K\sqrt{DSAH}\right).$$

Proof (Sketch) We construct $\mathcal{M}$ so that each MDP in $\mathcal{M}$ is a JAO MDP (Jaksch et al., 2010) of two states $\{0, 1\}$, $\frac{SA}{4}$ actions and diameter $\frac{D}{4}$. Figure 1 (left) illustrates the structure of a JAO MDP. State 0 has no reward, while state 1 has reward $+1$. Each model has a unique best action a^* that starts from 0 and goes to 1. The pair $(0, a^*)$ is a λ-distinguishing state-action pair.

A JAO MDP can be converted to an MDP with S states, A actions and diameter D, and this type of MDP gives the minimax lower bound proof in the undiscounted setting (Jaksch et al., 2010). The adversary selects a model from $\mathcal{M}$ uniformly at random, and so previous episodes provide no useful information for the current episode; hence, the regret of any learner is equal to the sum of its K one-episode learning regrets. The one-episode learning regret for JAO MDPs is known to be $\Omega(\sqrt{DSAH})$ when comparing against the optimal infinite-horizon average reward. For JAO MDPs, the optimal infinite horizon policy is also optimal for finite horizon; so, we can use a geometric convergence result from Markov chain theory (Levin et al., 2008) to convert this lower bound to a lower bound of the standard finite-horizon regret of the same order, giving the result. ∎

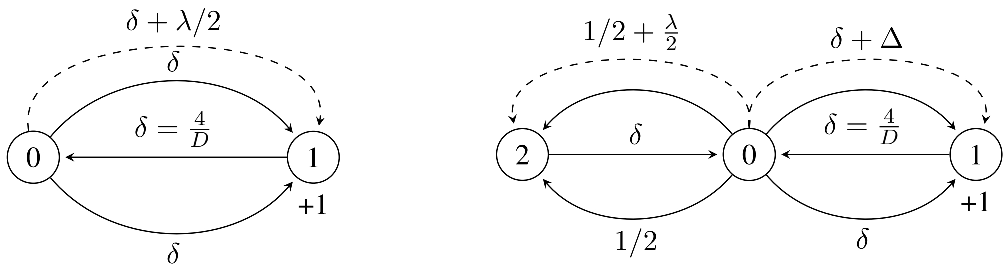

Figure 1: A JAO MDP (left) and a 2-JAO MDP (right). Only state 1 has reward $+1$. The dashed arrows indicate the best actions.

Using the same technique in the proof of Lemma 5, we can show that applying UCRL2 (Jaksch et al., 2010) in every episode individually leads to a regret upper bound of $O\left(KDS\sqrt{AH\ln H}\right)$. This implies that learning every episode individually already gives a near-optimal regret guarantee.

Remark 6 *Our proof for Lemma 5 contains a simple yet rigorous proof for the mixing-time argument used in Mao et al. (2021); Jin et al. (2018). This argument claims that for JAO MDPs, when the diameter is sufficiently small compared to the horizon, the optimal H-step value function V_1^* in the regret of the episodic setting can be replaced by the optimal average reward $\rho^* H$ in the undiscounted setting without changing the order of the lower bound. To the best of our knowledge, our proof is the first rigorous proof for this argument that applies for any number of episodes including $K = 1$. Domingues et al. (2021) provide an alternative proof; however the results therein hold in a different setting where K is sufficiently large and the horizon H can be much smaller than D.*

We emphasize that the lower bound in Lemma 5 holds for *any* learning algorithms. This result motivates the more interesting setting in which λ is a fixed and large constant independent of H. In this case, we are interested in an instance-specific lower bound. For multi-armed bandits, instance-specific lower bounds are constructed with respect to a class of *uniformly good* learning algorithms (Lai and Robbins, 1985). In our setting, we focus on defining a class of uniformly good algorithms that include the cluster-then-learn algorithms in the previous works for multi-task PAC RL settings such as Finite-Model-RL (Brunskill and Li, 2013) and PAC-EXPLORE (Guo and Brunskill, 2015). We consider a class of MDPs and a cluster-then-learn algorithm uniformly good if they satisfy an intuitive property: for any MDP in that class, the algorithm should be able to correctly classify whether a cluster of samples is from that MDP or not with an arbitrarily low (but not zero) failure probability, provided that the horizon H is sufficiently long for the algorithm to collect enough samples. The following definition formalizes this idea.

Definition 7 (PAC identifiability of MDPs) *A set of models $\mathcal{M}$ of size M is PAC identifiable if there exists a function $f : (0, 1) \mapsto \mathbb{N}$, a sample collection policy π and a classification algorithm $\mathcal{C}$ with the following property: for every $p \in (0, 1)$, for each model $1 \le m \le M$ in $\mathcal{M}$, if π is run for $f(p)$ steps and the state-transition samples are given to $\mathcal{C}$, then the algorithm $\mathcal{C}$ returns the correct identity of m with probability at least $1-p$, where the probability is taken over all possible sequence of $f(p)$ samples collected by running π on m for $f(p)$ steps. The smallest choice of function $f(p)$ among all possible choices is called the sample complexity of model identification of $\mathcal{M}$.*

The clustering algorithm in a cluster-then-learn framework solves a problem different from classification: they only need to tell whether a cluster of samples belong to the same or different distribution than another cluster of samples, not the identity of the distribution. We can reduce one problem to the other by the following construction: consider the adversary that gives all M models in the first M episodes. After the first M episodes, there are M clusters of samples, each corresponding to one model in $\mathcal{M}$. Once the learner has constructed M different clusters, from the episode $M+1$, the clustering problem is as hard as classification since identifying the right cluster immediately implies the identity of the MDP where the samples come from, and vice versa. Hence, we can apply the sample complexity of classification to that of clustering.

Next, we show the lower bound on the sample complexity of model identification for the class of λ-separable communicating MDPs.

Lemma 8 *For any $S, A \geq 20, D \geq 16$ and $\lambda \in (0, \frac{1}{2}]$, there exists a PAC identifiable λ-separable set of MDPs $\mathcal{M}$ of size $\frac{SA}{12}$, each with at most S states, A actions and diameter D such that for any classification algorithm $\mathcal{C}$, if the number of state-transition samples given to $\mathcal{C}$ is less than $\frac{SA}{180\lambda^2}$ then for at least one MDP in $\mathcal{M}$, algorithm $\mathcal{C}$ fails to identify that MDP with probability at least $\frac{1}{2}$.*

Proof (Sketch) The set $\mathcal{M}$ is a set of 2-JAO MDPs, shown in Figure 1 (right). Each 2-JAO MDP combines two JAO MDPs with the same number of actions and with diameter in the range $[\frac{D}{2}, D]$; one is λ-separable and one is the hard instance for the minimax lower bound of Jaksch et al. (2010). Rewards exist only in the part containing the hard instance. If a learner completely ignores the λ-separable part, by Lemma 5 the learner cannot do much better than just learning every episode individually. On the other hand, with enough samples from the λ-separable part, the learner can identify the MDP and use the samples collected in the previous episodes of the same MDP to accelerate learning the hard instance part. However, the λ-separable part is also a JAO MDP, for which no useful information from previous episodes can help identify the MDP in the current episode.

Only the actions at state 0 are λ-distinguishing and can be used to identify the MDPs. Taking an action in state 0 can be seen as flipping a coin: heads for transitioning to another state and tails for staying in state 0. Identifying a 2-JAO MDP reduces to the problem of using at most H coin flips to identify, in a $Q \times 2$ matrix of coins, a row j that has coins that are slightly different from the others. The first column has fair coins except in row j, where the success probability is $\frac{1}{2} + \lambda$. The second column coins with success probability of $\delta \leq \frac{1}{4}$ except in row j, where the coin is upwardly biased by $\Delta \leq \lambda$. Lemma 23 and Corollary 24 in the appendix show a $\Omega\left(\frac{Q}{\lambda^2}\right)$ lower bound on the number of coin flips on the first column (the left part of the 2-JAO MDP), implying the desired result. ■

Lemma 8 imply that for 2-JAO MDPs, any uniformly good model identification algorithm needs to collect at least $\Omega\left(\frac{SA}{\lambda^2}\right)$ samples from state 0 on the left part. Whenever an action towards state 2 is taken from state 0, the learner may end up in state 2. Once in state 2, the learner needs to get back to state 0 to obtain the next useful sample. The expected number of actions needed to get back to state 0 from state 2 is $\frac{1}{\delta} = \frac{D}{4}$. This implies the following two lower bounds on the horizon of the clustering phase and the total regret of any cluster-then-learn algorithms.

Corollary 9 *For any $S, A \geq 20, D \geq 16$ and $\lambda \in (0, 1]$, there exists a PAC identifiable λ-separable set of MDPs $\mathcal{M}$ of size $M = \frac{SA}{12}$, each with S states, A actions and diameter D such that for any uniformly good cluster-then-learn algorithm, to find the correct cluster with probability* *of at least $\frac{1}{2}$, the expected number of exploration steps needed in the clustering phase is $\Omega(\frac{DSA}{\lambda^2})$. Furthermore, the expected regret over K episodes of the same algorithm is*

$$\mathbb{E}[\text{Regret}(K)] \geq \Omega\left(\frac{KDSA}{\lambda^2}\right).$$

Proof (Sketch) In the lower bound construction, the learner is assumed to know everything about the set of models, including their optimal policies. Hence, after having identified the model in the clustering phase, the learner can follow the optimal policy in the learning phase and incur a small regret of at most $\frac{D}{2}$ in this phase. Therefore, the regret is dominated by the regret in the clustering phase, which is of order $\frac{DSA}{\lambda^2}$. ■

Remark 10 *The lower bound in Corollary 9 holds for a particular class of uniformly good cluster-then-learn algorithms under an adaptive adversary. It remains an open question whether this lower bound holds for any algorithms, not just cluster-then-learn.*

Remark 11 *Corollary 9 implies that, without further assumptions, it is not possible to improve the $\frac{1}{\lambda^2}$ dependency on λ. At the first glance this seems to contradict the existing results in bandits and online learning literature, where the regret bound depends on $\frac{1}{\text{gap}}$ where* gap *is the the difference in expected reward between the best arm and the sub-optimal arms. However, λ does not play the same role as the gaps in bandits. Observe that on the 2-JAO MDPs, the set of arms with positive reward is only in the right JAO MDP. The lower-bound learner knows this, but chooses to pull the arms on the left JAO MDP (with zero-reward) to collect side information that helps learn the right part faster. In this analogy, λ does not play the same role as the gaps in bandits, since the learner already knows the arms on the left JAO MDP are suboptimal. The role of λ is in model identification, for which similar $\frac{1}{\lambda^2}$ lower bounds are known (e.g. Tulsiani, 2014).*

Finally, we construct a non-communicating variant of the 2-JAO MDP to show that the finite diameter assumption is necessary. Figure 3 in Appendix C illustrates this construction. On this variant, all the transitions from state 0 to state 2 are reversed. In addition, no actions take state 0 to state 2, making this MDP non-communicating. A set of these non-communicating MDPs is still λ-separable due to the state-action pairs that start at state 2. However, by setting the initial state to 0, the adversary can force the learner to operate only on the right part, regardless of how large λ is.

4. Non-Asymptotic Upper Bounds

We propose and analyze AOMultiRL, a polynomial time cluster-then-learn algorithm that obtains a high-probability regret bound of $\tilde{O}(\frac{KDSA}{\lambda^2} + H^{3/2}\sqrt{MSAK})$. In each episode, the learner starts with the clustering phase to identify the cluster of samples generated in previous episodes that has the same task. Once the right cluster is identified, the learner can use the samples from previous episodes in the learning phase.

A fundamental difference between the undiscounted infinite horizon setting considered in previous works (Guo and Brunskill, 2015; Brunskill and Li, 2013) and the episodic finite horizon in our work is the horizon of the two phases. In previous works, different episodes might have different horizons for the clustering phase depending on whether the learner decides to start exploration at all (Brunskill and Li, 2015) or which state-action pairs are to be explored (Brunskill and Li, 2013). This poses a challenge for the episodic finite-horizon setting, because a varying horizon for the clustering phase leads to a varying horizon for the learning phase. Thus, standard single-task algorithms that rely on a fixed horizon such as UCBVI (Azar et al., 2017) and StrongEuler (Simchowitz and Jamieson, 2019) cannot be applied directly. From an algorithmic standpoint, for a fixed horizon H, a non-asymptotic bound on the horizon of the clustering phase is necessary so that the learner knows exactly whether H is large enough and when to stop collecting samples.

AOMultiRL alleviates this issue by setting a fixed horizon for the clustering phase, which reduces the learning phase to standard single-task episodic RL. First, we state an assumption on the ergodicity of the MDPs.

Assumption 12 *The hitting times of all MDPs in $\mathcal{M}$ are bounded by a known constant $\tilde{D}$.*

The main purpose of Assumption 12 is simplifying the computation of a non-asymptotic upper bound for the clustering phase in order to focus the exposition on the main ideas. We discuss a method for removing this assumption in Appendix G.

Algorithm 1 outlines the main steps of our approach. Given a set Γ^α of α-distinguishing state-action pairs, in the clustering phase the learner employs a history-dependent policy specified by Algorithm 2, `ExploreID`, to collect at least N samples for each state-action pair in Γ^α, where N will be determined later. Once all (s, a) in Γ^α have been visited at least N times, Algorithm 3, `IdentifyCluster`, computes the empirical means of the transition function of these (s, a) and then compares them with those in each cluster to determine which cluster contains the samples from the same task (or none do, in which case a new cluster is created). For the rest of the episode, the learner uses the UCBVI-CH algorithm (Azar et al., 2017) to learn the optimal policy.

The algorithms and results up to Theorem 16 are presented for a general set Γ^α. Since Γ^α is generally unknown, Corollary 17 shows the result for $\alpha = \lambda$ and $\Gamma^\alpha = \mathcal{S} \times \mathcal{A}$.

4.1. The Exploration Algorithm

Given a collection $\mathcal{B}$ of tuples (s, a, s'), the empirical transition functions estimated by $\mathcal{B}$ are

$$\hat{P}_{\mathcal{B}}(s' \mid s, a) = \begin{cases} \frac{N_{\mathcal{B}}(s,a,s')}{N_{\mathcal{B}}(s,a)} & \text{if } N_{\mathcal{B}}(s,a) > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $$N_{\mathcal{B}}(s,a,s') = \sum_{(x,y,z)\in\mathcal{B}} \mathbb{I}\{x = s, y = a, z = s'\}, \qquad N_{\mathcal{B}}(s,a) = \sum_{s'\in\mathcal{S}} N_{\mathcal{B}}(s,a,s')$$

are the number of instances of (s, a, s') and (s, a) in $\mathcal{B}$, respectively.

For each episode k, let P^k denote the transition function of the task m^k and $\mathcal{B}_k$ denote the collection of samples (s_h, a_h, s_{h+1}) collected during the learning phase. The empirical means $\hat{P}^k$ estimated using samples in $\mathcal{B}_k$ are $\hat{P}^k = \hat{P}_{\mathcal{B}_k}$. The value of N can be chosen so that for all $(s, a) \in \Gamma^\alpha$, with high probability $\hat{P}^k(s, a)$ is close to $P^k(s, a)$. Specifically, we find that if N is large enough so that $\hat{P}^k(s, a)$ is within $\lambda/8$ in ℓ_1 norm of the true function $P^k(s, a)$, then the right cluster can be identified in every episode. The exact value of N is given in the following lemma.

Algorithm 1: Adversarial online multi-task RL

Input: Number of models M, number of episodes K, MDPs parameters $\mathcal{S}, \mathcal{A}, H, \tilde{D}, \lambda$, probability p, separation level α and an α-distinguishing set Γ^α.

Compute

$p_1 = p/3, N = \frac{256}{\lambda^2}\max\{S, \ln(\frac{K|\Gamma^\alpha|}{p_1})\}, \delta = \alpha - \lambda/4, H_0 = 12D|\Gamma^\alpha|N$

Initialize $\mathcal{C} \leftarrow \emptyset$

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

  Initialize $\mathcal{B}_k \leftarrow \emptyset$

  The environment chooses a task m^k

  Observe the initial state s_1

  for $h = 1, \ldots, H_0$ **do**

    $a_h = \texttt{ExploreID}(s_h, \Gamma^\alpha)$

    Observe s_{h+1} and r_{h+1}

    Add (s_h, a_h, s_{h+1}) to $\mathcal{B}_k$

  $id \leftarrow \texttt{IdentifyCluster}(\mathcal{B}_k, \Gamma^\alpha, \mathcal{C}, \delta)$

  if $id \geq 1$ **then** $\mathcal{C}_{id}^{model} = \mathcal{C}_{id}^{model} \cup \mathcal{B}_k$

  else

    $id \leftarrow |\mathcal{C}| + 1$

    $\mathcal{C}_{id}^{model} = \mathcal{B}_k, \mathcal{C}_{id}^{regret} = \emptyset$

    $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{C}_{id}$

  $\pi_k = \texttt{UCBVI-CH}(\mathcal{C}_{id}^{regret})$

  for $h = H_0 + 1, \ldots, H$ **do**

    $a_h = \pi_k(h, s_h)$

    Observe s_{h+1} and r_{h+1}

    $\mathcal{C}_{id}^{regret} = \mathcal{C}_{id}^{regret} \cup (s_h, a_h, s_{h+1})$

Algorithm 2: ExploreID

Input: Episode k, state s, set Γ^α and number N

Set $\mathcal{G}(s) = \left\{ \begin{array}{l} a \in \mathcal{A} : \\ (s,a) \in \Gamma^\alpha, N_{\mathcal{B}_k}(s,a) < N \end{array} \right\}$

if $\mathcal{G}(s) \neq \emptyset$ **then**

  return $\arg\max_{a \in \mathcal{G}(s)} N_{\mathcal{B}_k}(s,a)$

else

  return $\arg\max_{a \in \mathcal{A}} \sum_{s'=1}^{S} \hat{P}^k(s' \mid s,a)\mathbb{I}\{\mathcal{G}(s') \neq \emptyset\}$

Algorithm 3: Identify Cluster

Input: Episode k, set Γ^α, clusters $\mathcal{C}$, and threshold δ

for $c = 1, \ldots, \|\mathcal{C}\|$ **do**

  Initialize $\texttt{id} \leftarrow c$

  for $(s,a) \in \Gamma$ **do**

    if $\left\|[\hat{P}_c - \hat{P}^k](s,a)\right\| > \delta$ **then**

      $\texttt{id} \leftarrow 0$

      break;

  if $\texttt{id} == c$ **then**

    return $\texttt{id}$;

return 0;

Lemma 13 *Suppose the learner is given a constant $p_1 \in (0,1)$ and a α-distinguishing set $\Gamma^\alpha \subseteq \mathcal{S} \times \mathcal{A}$. If each state-action pair in Γ^α is visited at least $N = \frac{256}{\lambda^2}\max\{S, \ln(\frac{K|\Gamma^\alpha|}{p_1})\}$ times during the clustering phase of each episode $k = 1, 2, \ldots, K$, then with probability at least $1 - p_1$, the event*

$$\mathcal{E}_k^{\Gamma^\alpha} = \left\{\forall (s,a) \in \Gamma^\alpha, \left\|P^k(s,a) - \hat{P}^k(s,a)\right\| \leq \frac{\lambda}{8}\right\} \text{ holds for all } k \in [K].$$

The exploration in AOMultiRL is modelled as an instance of the active model estimation problem (Tarbouriech et al., 2020). Given the current state s, if there exists an action a such that $(s,a) \in \Gamma^\alpha$ and (s,a) has not been visited at least N times, this action will be chosen (with ties broken by selecting the most chosen action). Otherwise, the algorithm chooses an action that has the highest estimated probability of leading to an under-sampled state-action pair in Γ^α. The following lemma computes the number of steps H_0 in the clustering phase.

Lemma 14 *Consider p_1 and N defined in Lemma 13. By setting*

$$H_0 = 12\tilde{D}|\Gamma^\alpha|N = \frac{3072\tilde{D}|\Gamma^\alpha|}{\lambda^2}\max\{S, \ln(\frac{K|\Gamma^\alpha|}{p_1})\},$$

with probability at least $1 - p_1$, Algorithm 2 visits each state-action pair in Γ^α at least N times during the clustering phase in each of the K episodes.

4.2. The Clustering Algorithm

Denote by $\mathcal{C}$ the set of clusters, $C = |\mathcal{C}|$ the number of clusters and $\mathcal{C}_i$ the i^{th} cluster. Each $\mathcal{C}_i$ is a collection of two multisets $\mathcal{C}_i^{model}, \mathcal{C}_i^{regret} \subset \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ which contain the (s, a, s') samples collected during the clustering and learning phases, respectively. Formally, up to episode k we have

$$\mathcal{C}_i^{model} = \cup_{k'=1}^{k-1}\{(s_h^{k'}, a_h^{k'}, s_{h+1}^{k'}) : h \le H_0, \text{id}^{k'} = i\},$$
$$\mathcal{C}_i^{regret} = \cup_{k'=1}^{k-1}\{(s_h^{k'}, a_h^{k'}, s_{h+1}^{k'}) : h > H_0, \text{id}^{k'} = i\},$$

where s_h^k and a_h^k are the state and action at time step h of episode k, respectively and $id^{k'}$ is the cluster index returned by Algorithm 3 in episode k'.

Let $\hat{P}_i = \hat{P}_{\mathcal{C}_i^{model}}$ denote the empirical means estimated using samples in $\mathcal{C}_i^{model}$. For each episode k, from Lemma 14 with high probability after the first H_0 steps each state-action pair $(s, a) \in \Gamma^\alpha$ has been visited at least N times. Algorithm 3 determines the right cluster for a task by computing the ℓ_1 distance between $\hat{P}^k$ and the empirical transition function $\hat{P}_i$ for each cluster $i = 1, 2, \dots, C$. If there exists an $(s, a) \in \Gamma^\alpha$ such that the distance is larger than a certain threshold δ, i.e., $\left\|[\hat{P}_i - \hat{P}^k](s, a)\right\| > \delta$, then the algorithm concludes that the task belongs to another cluster. Otherwise, the task is considered to belong to cluster i. We set $\delta = \alpha - \lambda/4$. The following lemma shows that with this choice of δ, the right cluster is identified by Algorithm 3 in all episodes.

Lemma 15 *Consider a λ-separable set of MDPs $\mathcal{M}$ and an α-distinguishing set Γ^α where $\alpha \ge \lambda/2$. If the events $\mathcal{E}_k^{\Gamma^\alpha}$ defined in Lemma 13 hold for all $k \in [K]$, then with the distance threshold $\delta = \alpha - \lambda/4$ Algorithm 3 always produces a correct output in each episode: the trajectories of the same model in two different episodes are clustered together and no two trajectories of two different models are in the same cluster.*

Once the clustering phase finishes, the learner enters the learning phase and uses the UCBVI-CH algorithm (Azar et al., 2017) to learn the optimal policy for this phase. In principle, almost all standard single-task RL algorithms with a near-optimal regret guarantee can be used for this phase. We chose UCBVI-CH to simplify the analysis and make the exposition clear.

To simulate the standard single-task episodic learning setting, the learner only uses the samples in $\mathcal{C}_i^{regret}$ for regret minimization. Theorem 16 states a regret bound for Algorithm 1.

Theorem 16 *For any failure probability $p \in (0, 1)$, with probability at least $1 - p$ the regret of Algorithm 1 is bounded as*

$$\textit{Regret}(K) \le 2KH_0 + 67H_1^{3/2}L\sqrt{MSAK} + 15MS^2AH_1^2L^2,$$

where $H_0 = 12\tilde{D}|\Gamma^\alpha|N$, $N = \frac{256}{\lambda^2}\max\{S, \ln(\frac{3K|\Gamma^\alpha|}{p})\}$, $H_1 = H - H_0$, and $L = \ln(15SAKHM/p)$.

For $K > MS^3AH$, the first two terms are the most significant. The $2KH_0$ term accounts for the clustering phase and the fact that the exploration policy might lead the learner to an undesirable state after H_0 steps. The $\tilde{O}(\sqrt{K})$ term comes from the fact that the learning phase is equivalent to episodic single-task learning with horizon H_1. When $H \gg H_0$, the sub-linear bound on the learning phase is a major improvement compared to the $O(K\sqrt{HSA})$ bound of the strategy that learns each episode individually.

By setting $\Gamma^\alpha = \mathcal{S} \times \mathcal{A}$ and $\alpha = \lambda$, we obtain

Corollary 17 *For any failure probability $p \in (0, 1)$, with probability at least $1 - p$, by setting $\Gamma^\alpha = \mathcal{S} \times \mathcal{A}$ with $\alpha = \lambda$, the regret of Algorithm 1 is*

$$\textit{Regret}(K) \leq O\left(\frac{K\tilde{D}SA}{\lambda^2} \ln\left(\frac{KSA}{p}\right) + H^{3/2}L\sqrt{MSAK}\right). \tag{2}$$

where $L = \ln(15SAKH_1M/p)$.

Time Complexity The clustering algorithm runs once in each episode, which leads to time complexity of $O(MSA + H)$. When $H \gg H_0$, the overall time complexity is dominated by the learning phase, which is $O(HSA)$ for UCBVI-CH.

Remark 18 *Instead of clustering, a different paradigm involves actively merging samples from different MDPs to learn a model that is an averaged estimate of the MDPs in $\mathcal{M}$. The best regret guarantee in this paradigm, to the best of our knowledge, is $\tilde{O}(S^{1/3}A^{1/3}B^{1/3}H^{5/3}K^{2/3})$, where B is a variation budget, achieved by RestartQ-UCB (Mao et al., 2021, Theorem 3). In our setting, if the adversary frequently alternates between tasks then $B = \Omega(KH\lambda)$ and therefore this bound becomes $\tilde{O}(\lambda^{1/3}S^{1/3}A^{1/3}H^2K)$, which is larger than the trivial bound KH and worse than the bound in Corollary 17. If the adversary selects tasks so that B is small i.e. $B = o(K)$ then the bound offered by RestartQ-UCB is better since it is sub-linear in K. Note that this does not contradict the lower bound result in Section 3, since the lower bound is constructed with an adversary that selects tasks uniformly at random, and hence B is linear in K.*

4.3. Learning a distinguishing set when M is small

As pointed out by Brunskill and Li (2013), for all $\alpha > 0$, the size of the smallest α-distinguishing set of $\mathcal{M}$ is at most $\binom{M}{2}$. If $M^2 \ll SA$ and such a set is known to the learner, then the clustering phase only need collect samples from this set instead of the full $\mathcal{S} \times \mathcal{A}$ set of state-action pairs. However, in general this set is not known. We show that if the adversary is weaker so that all models are guaranteed to appear at least once early on, the learner will be able to discover a $\frac{\lambda}{2}$-distinguishing set $\hat{\Gamma}$ of size at most $\binom{M}{2}$. Specifically, we employ the following assumption:

Assumption 19 *There exists an unknown constant $K_1 \geq M$ satisfying $K_1SA < K$ such that after at most K_1 episodes, each model in $\mathcal{M}$ has been given to the learner at least once.*

In order to discover $\hat{\Gamma}$, the learner uses Algorithm 4, which consists of two stages:

- Stage 1: the learner starts by running Algorithm 1 with the λ-distinguishing set candidate $\mathcal{S} \times \mathcal{A}$ until the number of clusters is M. With high probability, each cluster corresponds to a model. At the end of stage 1, the learner uses the empirical estimates in all clusters $\hat{P}_i$ for $i \in [M]$ to construct a $\lambda/2$-distinguishing set $\hat{\Gamma}$ for $\mathcal{M}$.

Algorithm 4: AOMultiRL with all models being given at least once

Input: Number of models M, number of episodes K, MDPs parameters $\mathcal{S}, \mathcal{A}, H, \tilde{D}, \lambda$, probability p

Stage 1: Run Algorithm 1 with the distinguishing set $\Gamma^{\alpha} = \mathcal{S} \times \mathcal{A}$ and $\alpha = \lambda$ until the number of clusters is M;

for $i, j \in [M] \times [M], i \neq j$ **do**
    $\hat{\Gamma}_{i,j} = \emptyset$;
    for $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**
        if $\left\| \hat{P}_i(s, a) - \hat{P}_j(s, a) \right\| > 3\lambda/4$ **then**
            $\hat{\Gamma}_{i,j} = \hat{\Gamma}_{i,j} \cup (s, a)$;
            break

$\hat{\Gamma} = \cup_{i,j} \hat{\Gamma}_{i,j}$;

Stage 2: Run Algorithm 1 with distinguishing set $\hat{\Gamma}$ and $\alpha = \lambda/2$ for $K_2 = K - K_1$ episodes.

- Stage 2: the learner runs Algorithm 1 with the distinguishing set $\hat{\Gamma}$ as an input.

Extracting $\lambda/2$-distinguishing pairs: After K_1 episodes, with high probability there are M clusters corresponding to M models. For two clusters i and j, the set $\hat{\Gamma}_{i,j}$ contains the first state-action pair (s, a) that satisfies $\left\| \hat{P}_i(s, a) - \hat{P}_j(s, a) \right\| > 3\lambda/4$. With high probability, every $(s, a) \in \Gamma_{i,j}$ satisfies this condition, hence $\hat{\Gamma}_{i,j} \neq \emptyset$.

Let $i^\star \in [M]$ denote the index of the MDP model corresponding to cluster i. For all $(s, a) \in \hat{\Gamma}_{i,j}$, by the triangle inequality, we have

$$\|P_{i^\star} - P_{j^\star}\| \geq \left\| \hat{P}_i - \hat{P}_j \right\| - \left\| \hat{P}_i - P_{i^\star} + P_{j^\star} - \hat{P}_j \right\| > 3\lambda/4 - (\lambda/8 + \lambda/8) = \lambda/2,$$

where (s, a) is omitted for brevity. It follows that the set $\hat{\Gamma} = \cup_{i,j} \hat{\Gamma}_{i,j}$ is $\lambda/2$-distinguishing and $|\hat{\Gamma}| \leq \binom{M}{2}$. Although $\lambda/2$ is smaller than the λ-separation level of Γ, it is sufficient for the conditions in Lemma 15 to hold. Thus, with high probability the clustering algorithm in stage 2 works correctly. The next theorem shows the regret guarantee of Algorithm 4.

Theorem 20 *Under Assumption 19, With probability at least $1 - p$, the regret of Algorithm 4 is*

$$\textit{Regret}(K) = O\left(\frac{K\tilde{D}M^2}{\lambda^2} \ln \frac{KM^2}{p} + H^{3/2} L \sqrt{MKSA} \right),$$

where $H_{0,M} = \frac{3072\tilde{D}M^2}{\lambda^2} \max\{S, \ln(\frac{3KM^2}{p})\}$ *and* $L = \ln(15SAKH_1M/p)$.

Compared to Corollary 17, Theorem 20 improves the clustering phase's dependency from SA to M^2. This implies that if the number of models is small and all models appear relatively early, we can discover a $\lambda/2$-distinguishing set quickly without increasing the order of the total regret bound.

5. Conclusion

In this paper, we studied the adversarial online multi-task RL setting with the tasks belonging to a finite set of well-separated models. We used a general notion of task-separability, which we call λ-separability. Under this notion, we proved a minimax regret lower bound that applies to all algorithms and an instance-specific regret lower bound that applies to a class of uniformly good cluster-then-learn algorithms. We further proposed AOMultiRL, a polynomial time cluster-then-learn algorithm that obtains a nearly-optimal instance-specific regret upper bound. These results addressed two fundamental aspects of online multi-task RL, namely learning an adversarial task sequence and learning under a general task-separability notion. Adversarial online multi-task learning remains challenging when the diameter and the number of models are unknown; this is left for future work.

Acknowledgments

This work was supported by the NSERC Discovery Grant RGPIN-2018-03942.

References

David Abel, Yuu Jinnai, Sophie Yue Guo, George Konidaris, and Michael Littman. Policy and value transfer in lifelong reinforcement learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 20–29. PMLR, 10–15 Jul 2018.

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2007.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002. doi: 10.1137/S0097539701398375. URL https://doi.org/10.1137/S0097539701398375.

Mohammad Gheshlaghi Azar, Alessandro Lazaric, and Emma Brunskill. Sequential transfer in multi-armed bandit with finite set of models. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272. PMLR, 06–11 Aug 2017.

Emma Brunskill and Lihong Li. Sample complexity of multi-task reinforcement learning. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, page 122–131, Arlington, Virginia, USA, 2013. AUAI Press.

Emma Brunskill and Lihong Li. The online discovery problem and its application to lifelong reinforcement learning. *CoRR*, abs/1506.03379, 2015.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 578–598. PMLR, 16–19 Mar 2021. URL https://proceedings.mlr.press/v132/domingues21a.html.

Zhaohan Guo and Emma Brunskill. Concurrent PAC RL. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2624–2630. AAAI Press, 2015. ISBN 0262511290.

Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual Markov Decision Processes, 2015.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, and Shie Mannor. RL for latent MDPs: Regret guarantees and a lower bound. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

Erwan Lecarpentier, David Abel, Kavosh Asadi, Yuu Jinnai, Emmanuel Rachelson, and Michael L. Littman. Lipschitz lifelong reinforcement learning. In *AAAI*, 2021.

David Asher Levin, Yuval Peres, and Elizabeth Lee Wilmer. *Markov Chains and Mixing Times*. American Mathematical Soc., 2008. ISBN 9780821886274. URL http://pages.uoregon.edu/dlevin/MARKOV/.

Weichao Mao, Kaiqing Zhang, Ruihao Zhu, David Simchi-Levi, and Tamer Basar. Near-optimal model-free reinforcement learning in non-stationary episodic MDPs. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7447–7458. PMLR, 18–24 Jul 2021. URL http://proceedings.mlr.press/v139/mao21b.html.

Igal Sason. On reverse Pinsker inequalities. *arXiv preprint arXiv:1503.07118*, 2015.

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular MDPs. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Lauren N. Steimle, David L. Kaufman, and Brian T. Denton. Multi-model Markov decision processes. *IISE Transactions*, 53(10):1124–1139, 2021. doi: 10.1080/24725854.2021.1895454.

Yanchao Sun and Furong Huang. Can agents learn by analogy? An inferable model for PAC reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, page 1332–1340, Richland, SC, 2020. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450375184.

Jean Tarbouriech, Shubhanshu Shekhar, Matteo Pirotta, Mohammad Ghavamzadeh, and Alessandro Lazaric. Active model estimation in Markov Decision Processes. In *Uncertainty in Artificial Intelligence*, 2020. URL http://proceedings.mlr.press/v124/tarbouriech20a/tarbouriech20a-supp.pdf.

Jean Tarbouriech, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. A provably efficient sample collection strategy for reinforcement learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=AvVDR8R-kQX.

Madhur Tulsiani. Lecture 6, lecture notes in information and coding theory, October 2014. URL https://home.ttic.edu/~madhurt/courses/infotheory2014/l6.pdf.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Appendix A. Related Work

Stochastic Online Multi-task RL. The Finite-Model-RL algorithm (Brunskill and Li, 2013) considered the stochastic setting with infinite-horizon MDPs and focused on deriving a sample complexity of exploration in a (ϵ, δ)-PAC setting. As shown by Dann et al. (2017), even an optimal (ϵ, δ)-PAC bound can only guarantee a necessarily sub-optimal $O(K_m^{2/3})$ regret bound for each task $m \in [M]$ that appears in K_m episodes, leading to an overall $O(M^{1/3}K^{2/3})$ regret bound for the learning phase in the multi-task setting. The Contextual MDPs algorithm by Hallak et al. (2015) is capable of obtaining a $O(\sqrt{K})$ regret bound in the learning phase after the right cluster has been identified; however their clustering phase has exponential time complexity in K. The recent L-UCRL algorithm (Kwon et al., 2021) considered the stochastic finite-horizon setting and reduced the problem to learning the optimal policy of a POMDP. Under a set of assumptions that allow the clusters to be discovered in $O(\text{polylog}(MSA))$, L-UCRL is able to obtain an overall $O(\sqrt{MK})$ regret with respect to a POMDP planning oracle which aims to learn a policy that maximizes the expected single-task return when a task is randomly drawn from a known distribution of tasks. In contrast, our work adopts a stronger notion of regret that encourages the learner to maximize its expected return for a sequence of tasks chosen by an adversary. When the models are bandits instead of MDPs, Azar et al. (2013) use spectral learning to estimate the mean reward of the arms in all models and obtains an upper bound linear in K.

Lifelong RL. Learning a sequence of related tasks is more well-studied in the lifelong learning literature. Recent works in lifelong RL (Abel et al., 2018; Lecarpentier et al., 2021) often focus on the setting where tasks are drawn from an unknown distribution of MDPs and there exists some similarity measure between MDPs that support transfer learning. Our work instead focuses on learning the dissimilarity between tasks for the clustering phase and avoiding negative transfer.

Active model estimation The exploration in AOMultiRL is modelled after the active model estimation problem (Tarbouriech et al., 2020), which is often presented in PAC-RL setting. Several recent works on active model estimation are PAC-Explore (Guo and Brunskill, 2015), FW-MODEST (Tarbouriech et al., 2020), $\beta-$curious walking (Sun and Huang, 2020), and GOSPRL (Tarbouriech et al., 2021). The $\Theta(\tilde{D}|\Gamma^\alpha|N)$ bound on the horizon of clustering in Lemma 14 has the same $O(S^2A)$ dependency on the number of states and actions as the state-of-the-art bound by GOSPRL (Tarbouriech et al., 2021) for the active model estimation problem. The main drawback is that H_0 depends linearly on the hitting time $\tilde{D}$ and not the diameter D of the MDPs. As the hitting time is often strictly larger than the diameter (Jaksch et al., 2010; Tarbouriech et al., 2021), this dependency on $\tilde{D}$ is sub-optimal. On the other hand, AOMultiRL is substantially less computationally expensive than GOSPRL since there is no shortest-path policy computation involved.

Appendix B. The generality of λ-separability notion

In this section, we show that the general separation notion in Definition 1 defines a broader class of online multi-task RL problems that extends the entropy-based separation assumption in the latent MDPs setting (Kwon et al., 2021). We start by restating the entropy-based separation condition of Kwon et al. (2021):

Definition 21 *Let Π denote the class of all history-dependent and possibly non-Markovian policies, and let $\tau \sim (m, \pi)$ be a trajectory of length H sampled from MDP m by a policy $\pi \in \Pi$. The set $\mathcal{M}$ is well-separated if the following condition holds:*

$$\forall m, m' \in \mathcal{M}, m' \neq m, \pi \in \Pi, \Pr_{\tau\sim(m,\pi)}\left(\frac{\Pr_{m',\pi}(\tau)}{\Pr_{m,\pi}(\tau)} > (\epsilon_p/M)^{c_1}\right) < (\epsilon_p/M)^{c_2}, \tag{3}$$

where $\epsilon_p \in (0,1)$ is a target failure probability, $c_1 \geq 4, c_2 \geq 4$ are universal constants and $\Pr_{m,\pi}(\tau)$ is the probability that τ is realized when running policy π on model m.

The following lemma constructs a set $\mathcal{M}$ of just two models that satisfy the λ-separability condition but not the entropy-based separation condition.

Lemma 22 *Given any $\lambda \in (0,1), \epsilon_p \in (0,1), H > 0$ and any constants $c_1, c_2 \geq 4$, there exists a set of MDPs $\mathcal{M} = \{m_1, m_2\}$ with horizon H that is λ-separable but is not well-separated in the sense of Definition 21.*

Proof Consider the set $\mathcal{M}$ with $M = 2, \mathcal{S} = \{s^1, s^2, s^3\}, \mathcal{A} = \{a^1, a^2\}$ in Figure 2. Both m_1 and m_2 have the same transition functions in all state-action pairs except for (s^1, a^1):

$$\begin{aligned}
&\mathbb{P}_1(s^2 \mid s^1, a^1) = \lambda \\
&\mathbb{P}_1(s^3 \mid s^1, a^1) = 1 - \lambda \\
&\mathbb{P}_2(s^2 \mid s^1, a^1) = \lambda/2 \\
&\mathbb{P}_2(s^3 \mid s^1, a^1) = 1 - \lambda/2.
\end{aligned}$$

It follows that the ℓ_1 distance between $P_1(s^1, a^1)$ and $P_2(s^1, a^1)$ is

$$\begin{aligned}
\left\|P_1(s^1, a^1) - P_2(s^1, a^1)\right\| &= \left\|P_1(s^2 \mid s^1, a^1) - P_2(s^2 \mid s^1, a^1)\right\| + \left\|P_1(s^3 \mid s^1, a^1) - P_2(s^3 \mid s^1, a^1)\right\| \\
&= 2(\lambda - \lambda/2) \\
&= \lambda.
\end{aligned}$$

As a result, this set $\mathcal{M}$ is λ-separable. However, any deterministic policy that takes action a_2 in s_1 and an arbitrary action in s_2 and s_3 will induce the same Markov chain on two MDP models. Thus, the entropy-based separation definition does not apply. An example of such a policy is shown below.

Consider running the following deterministic policy on model m_1:

$$\begin{aligned}
\pi(s^1) &= a^2 \\
\pi(s^2) &= a^1 \\
\pi(s^3) &= a^1.
\end{aligned}$$

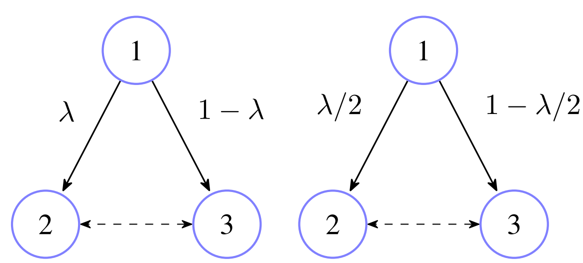

Figure 2: An instance of λ-separable LMDPs where Definition 21 does not apply

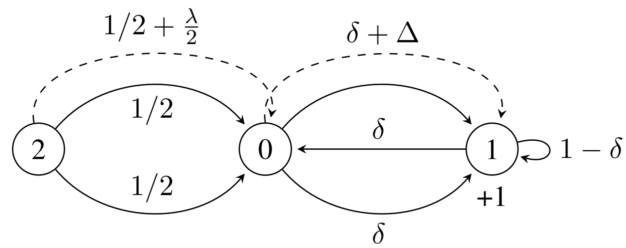

Figure 3: A non-communicating 2-JAO MDP. There are no rewards at states 0 and 2, while state 1 has reward $+1$. We set $\Delta = \Theta(\sqrt{\frac{SA}{HD}})$. The dashed arrows indicate the unique actions with highest transition probabilities on the left and right parts of the MDP. No actions take state 0 to state 2, making this MDP non-communicating.

Consider an arbitrary trajectory τ. The probability that this trajectory is realized with respect to both models is

$$\Pr_{m_1,\pi}(\tau) = \prod_{t=1}^{H} P_1(s_{t+1} \mid s_t, a_t) \tag{4}$$

$$= \prod_{t=1}^{H} P_1(s_{t+1} \mid s_t, \pi(s_t)) \tag{5}$$

$$= \prod_{t=1}^{H} P_2(s_{t+1} \mid s_t, a_t) \qquad \text{since } (s_t, \pi(s_t)) \neq (s^1, a^1) \tag{6}$$

$$= \Pr_{m_2,\pi}(\tau). \tag{7}$$

As a result, for all τ,

$$\frac{\Pr_{m_2,\pi}(\tau)}{\Pr_{m_1,\pi}(\tau)} = 1, \tag{8}$$

which implies that

$$\Pr_{\tau\sim m_1,\pi}\left(\frac{\Pr_{m_2,\pi}(\tau)}{\Pr_{m_1,\pi}(\tau)} > (\epsilon_p/M)^{c_1}\right) = \Pr_{\tau\sim m_1,\pi}(1 > (\epsilon_p/M)^{c_1}) = 1, \tag{9}$$

which is larger than $(\epsilon_p/M)^{c_2}$. ■

Appendix C. Proofs of the lower bounds

Lemma 5 (Minimax Lower Bound) *Suppose $S, A \geq 10, D \geq 20\log_A(S)$ and $H \geq DSA$ are given. Let $\lambda = \Theta(\sqrt{\frac{SA}{HD}})$. There exists a set of λ-separable MDPs $\mathcal{M}$ of size $M = \frac{SA}{4}$, each with S states, A actions, diameter at most D and horizon H such that if the tasks are chosen uniformly at random from $\mathcal{M}$, the expected regret of any sequence of policies $(\pi_k)_{k=1,\ldots,K}$ over K episodes is*

$$\mathbb{E}[\text{Regret}(K)] \geq \Omega\left(K\sqrt{DSAH}\right).$$

Proof We construct $\mathcal{M}$ in the following way: each MDP in $\mathcal{M}$ is a JAO MDP (Jaksch et al., 2010) of two states and SA actions and diameter $D' = D/4$. The translation from this JAO MDP to an MDP with S states, A actions and diameter D is straightforward (Jaksch et al., 2010). State 1 has reward $+1$ while state 0 has no reward. In state 0, for all actions the probability of transitioning to state 1 is δ except for one best action where this probability is $\delta + \lambda/2$. Every MDP in $\mathcal{M}$ has a unique best action: for $i = 1, \dots, SA$, the i^{th} action is the best action in the MDP m_i. The starting state is always $s_1 = 0$.

We consider a learner who knows all the parameters of models in $\mathcal{M}$, except the identity of the task m^k given in episode k. We employ the following information-theoretic argument from Mao et al. (2021): when the task m^k in episode k is chosen uniformly at random from $\mathcal{M}$, no useful information from the previous episodes can help the learner identify the best action in m^k. This is true since all the information in the previous episodes is samples from the MDPs in $\mathcal{M}$, which provide no further information than the parameters of the models in $\mathcal{M}$. Since $\mathcal{M} = SA$, all actions (from state 0) are equally probable to be the best action in m^k. Therefore, the learner is forced to learn m^k from scratch. It follows that the total regret of the learner is the sum of the one-episode-learning regrets in every episode:

$$\text{Regret}(K) = \sum_{k=1}^{K} R^k,$$

where $R^k = V_1^*(s_1) - V_1^{\pi_k}(s_1)$ is the one-episode-learning regret in episode k. The one-episode-learning is equivalent to the learning in the undiscounted setting with horizon H. Applying the lower bound result for the undiscounted setting in Jaksch et al. (2010, Theorem 5) obtains that for all π_k,

$$\rho^* H - \mathbb{E}_{m^k \sim \mathcal{M}} V_1^{\pi_k}(s_1) \geq \Omega(\sqrt{DSAH}),$$

where $\rho^* = \frac{\delta + \lambda/2}{2\delta + \lambda/2}$ is the average reward of the optimal policy (Jaksch et al., 2010). Note since only state 1 has reward $+1$, ρ^* is also the stationary probability that the optimal learner is at state 1.

Next, we show that for all $H \geq 2$ and $m^k \in \mathcal{M}$, it holds that $|V_1^* - \rho^* H| \leq \frac{D}{2}$. The optimal policy on all m^k induces a Markov chain between two states with transition matrix

$$\begin{bmatrix} 1 - \delta - \lambda/2 & \delta + \lambda/2 \\ \delta & 1 - \delta \end{bmatrix}.$$

Let $\mathbb{P}_{m^k}(s_t = 1 \mid s_1 = 0)$ be the probability that the Markov chain is in state 1 after t time steps with the initial state $s_1 = 0$. Let $\Delta_t = \mathbb{P}_{m^k}(s_t = 1 \mid s_1 = 0) - \rho^*$. Obviously, $\Delta_1 = -\rho^*$. By Levin et al. (2008, Equation 1.8), we have $\Delta_t = (1 - 2\delta - \lambda/2)^{t-1}\Delta_1$. It follows that, for the optimal policy,

$$
\begin{aligned}
V_1^*(s_1) &= \sum_{t=1}^{H} \mathbb{P}_{m^k}(s_t = 1 \mid s_1 = 0) && (10)\\
&= \sum_{t=1}^{H} (\Delta_t + \rho^*) && (11)\\
&= \rho^* H + \sum_{t=1}^{H} \Delta_t && (12)\\
&= \rho^* H + \sum_{t=1}^{H} (1 - 2\delta - \lambda/2)^{t-1} \Delta_1 && (13)\\
&= \rho^* H + \Delta_1 \frac{1 - (1 - 2\delta - \lambda/2)^H}{2\delta + \lambda/2}. && (14)
\end{aligned}
$$

Hence,

$$
\begin{aligned}
|V_1^*(s_1) - \rho^* H| &= \left| \Delta_1 \frac{1 - (1 - 2\delta - \lambda/2)^H}{2\delta + \lambda/2} \right| && (15)\\
&\le \left| \frac{\Delta_1}{2\delta + \lambda/2} \right| && (16)\\
&= \frac{\rho^*}{2\delta + \lambda/2} && (17)\\
&\le \frac{1}{2\delta + \lambda/2} && (18)\\
&\le \frac{1}{2\delta} && (19)\\
&= \frac{D}{2}, && (20)
\end{aligned}
$$

where the last equality follows from $\delta = \frac{D}{4}$.

For any $H \ge DSA$ and $S, A \ge 2$, we have $\sqrt{HDSA} \ge DSA \ge 4D$, and hence $\sqrt{HDSA} - \frac{D}{2} \ge \frac{\sqrt{HDSA}}{2}$. We conclude that

$$
\begin{aligned}
\mathbb{E}[\text{Regret}(K)] &= \sum_{k=1}^{K} \mathbb{E}[R^k]\\
&= \sum_{k=1}^{K} \mathbb{E}[V_1^* - V_1^{\pi_k}](s_1)\\
&\ge \sum_{k=1}^{K} (\rho^* H - \frac{D}{2} - V_1^{\pi_k}(s_1))\\
&= \Omega(K\sqrt{DHSA}).
\end{aligned}
$$

■

The upper bound of UCRL2 can be proved similarly: Theorem 2 in Jaksch et al. (2010) states that for any $p \in (0, 1)$, by running UCRL2 with failure parameter p, we obtain that for any initial state s_1 and any $H > 1$, with probability at least $1 - p$,

$$\rho^* H - \sum_{h=1}^{H} r_h \leq O\left(DS\sqrt{AH \ln \frac{H}{p}}\right). \tag{21}$$

Setting $p = \frac{1}{H}$ and trivially bound the regret in the failure cases by H to obtain

$$\rho^* H - E[\sum_{h=1}^{H} r_h] \leq O\left(DS\sqrt{AH \ln H^2}\right) + \frac{1}{H} \times H \tag{22}$$

$$= O\left(DS\sqrt{AH \ln H}\right). \tag{23}$$

This bound holds across all episodes, hence the total regret bound with respect to $\rho^* H$ is $O\left(KDS\sqrt{AH \ln H}\right)$. Combining this with the fact that $V_1^*(s_1) \leq \rho^* H + \frac{D}{2}$, we obtain the upper bound.

Lemma 8 *For any $S, A \geq 20, D \geq 16$ and $\lambda \in (0, \frac{1}{2}]$, there exists a PAC identifiable λ-separable set of MDPs $\mathcal{M}$ of size $\frac{SA}{12}$, each with at most S states, A actions and diameter D such that for any classification algorithm $\mathcal{C}$, if the number of state-transition samples given to $\mathcal{C}$ is less than $\frac{SA}{180\lambda^2}$ then for at least one MDP in $\mathcal{M}$, algorithm $\mathcal{C}$ fails to identify that MDP with probability at least $\frac{1}{2}$.*

Before showing the proof of Lemma 8, we consider the following auxiliary problem: Suppose we are given three constants $\delta, \lambda, \epsilon \in (0, \frac{1}{4}]$ and a set of $2Q$ coins. The coins are arranged into a $Q \times 2$ table of Q rows and 2 columns so that each cell contains exactly one coin. The rows are indexed from 1 to Q and the columns are indexed from 1 to 2. In the first column, all coins are fair except for one coin at row θ which is biased with probability of heads equal to $\frac{1}{2} + \lambda$. In the second column, all coins have probability of heads equal to δ except for the coin at row θ which has probability of heads $\delta + \epsilon$. In this setting, row θ is a special row that contains the most biased coins in the two columns. The objective is to find this special row θ after at most H coin flips, where $H > 0$ is a constant representing a fixed budget. Note that if we ignore the second column, then this problem is reduced to the well-known problem of identifying one biased coin in a collection of Q-coins (Tulsiani, 2014).

Let N_1, N_2 be the number of flips an algorithm performs on the first and second column, respectively. For a fixed global budget H, after $\tau = N_1 + N_2 \leq H$ coin flips, the algorithm recommends $\hat{\theta}$ as its prediction for θ. Note that τ is a random stopping time which can depend on any information the algorithm observes up to time τ. Let X_t be the random variable for the outcome of t^{th} flip, and $X_1^\tau = (X_1, X_2, \ldots, X_\tau)$ be the sequence of outcomes after τ flips. For $j \in [Q]$, let $\mathbb{P}_j$ denote the probability measure induced by Alg corresponding to the case when $\theta = j$. We first show that if the algorithm fails to flip the coins sufficiently many times in both columns, then for some θ the probability of failure is at least $\frac{1}{2}$.

Lemma 23 *Let $Q \geq 12, C_1 = 40$ and $C_2 = 64$. For any algorithm* Alg*, if*

$$N_1 \leq T_1 := \frac{Q}{4C_1\lambda^2} \quad \textit{and} \quad N_2 \leq T_2 := \frac{Q(\delta+\epsilon)}{4C_2\epsilon^2},$$

then there exists a set $J \subseteq [Q]$ with $|J| \geq \frac{Q}{6}$ such that

$$\forall j \in J,\ \mathbb{P}_j[\hat{\theta} = j] \leq \frac{1}{2}.$$

The proof uses a reasonably well-known reverse Pinsker inequality (Sason, 2015, Equation 10):

Let P and Q be probability measures over a common discrete set. Then

$$KL(P \,\|\, Q) \leq \frac{4\log_2 e}{\min_x Q(x)} \cdot D_{TV}(P \,\|\, Q)^2. \tag{24}$$

where D_{TV} is the total variation distance. In the particular case where P and Q are Bernoulli distributions with success probabilities p and $q \leq \frac{1}{2}$ respectively, we get

$$KL(P \,\|\, Q) \leq \frac{4\log_2 e}{q} \cdot (p-q)^2. \tag{25}$$

Proof (of Lemma 23) As reasoned in the proof for the lower bound of multi-armed bandits (Auer et al., 2002), we can assume that Alg is deterministic[2]. Our proof closely follows the main steps in the proof of Tulsiani (2014) for the setting where there is only one column. We will lower bound the probability of mistake of Alg based on its behavior on a hypothetical instance where $\lambda = \epsilon = 0$.

To account for algorithms which do not exhaust both budgets T_1 and T_2, we introduce two "dummy coins" by adding a zero'th row with two identical coins, solely for the analysis. These two coins have the same mean of 1 under all Q models and hence flipping either of them provides no information. An algorithm which wishes to stop in a round $\tau < H$ will simply flip any dummy coin in the remaining rounds $\tau+1, \tau+2, \dots, H$. This way, we have the convenient option of always working with a sequence of outcomes X_1^H in the analysis.

Let $\mathbb{P}_0$ and $\mathbb{E}_0$ denote the probability and expectation over X_1^H taken on the hypothetical instance with $\lambda = \epsilon = 0$, respectively. Let $a_t = (a_{t,0}, a_{t,1}) \in \{0, 1, \dots, Q\} \times \{1, 2\}$ be the coin that the algorithm flips in step t. Let $x_t \in \{0, 1\}$ denote the outcome of a_t where 0 is tails and 1 is heads.

The number of flips the coin in row i, column k is

$$N_{i,k} = \sum_{t=1}^{T} \mathbb{I}\{a_t = (i,k)\}.$$

By the earlier definition of N_k for $k \in \{1, 2\}$, we have

$$N_1 = \sum_{i=1}^{Q} N_{i,1},$$
$$N_2 = \sum_{i=1}^{Q} N_{i,2}.$$

2. Deterministic conditional on the random history

We define

$$J_1 := \left\{ i \in [Q] \colon \left(\mathbb{E}_0[N_{i,1}] \le \frac{4T_1}{Q} \right) \wedge \left(\mathbb{E}_0[N_{i,2}] \le \frac{4T_2}{Q} \right) \right\}.$$

Clearly, at most $\frac{Q}{4}$ rows i satisfy $\mathbb{E}_0[N_{i,1}] > \frac{4T_1}{Q}$ and, similarly, at most $\frac{Q}{4}$ rows i satisfy $\mathbb{E}_0[N_{i,2}] > \frac{4T_2}{Q}$. Therefore, $|J_1| \ge Q - 2 \cdot \frac{Q}{4} = \frac{Q}{2}$.

We also define

$$J_2 := \left\{ i \in [Q] \colon \mathbb{P}_0(\hat{\theta} = i) \le \frac{3}{Q} \right\}.$$

As at most $\frac{Q}{3}$ arms i can satisfy $\mathbb{P}_0(\hat{\theta} = i) > \frac{3}{Q}$, it holds that $|J_2| \ge \frac{2Q}{3}$.

Consequently, defining $J := J_1 \cap J_2$, we have $|J| \ge \frac{Q}{6}$.

For any $j \in J$, we have

$$|\mathbb{P}_j[c^* = j] - \mathbb{P}_0[c^* = j]| = |\mathbb{E}_j[\mathbb{I}\{c^* = j\}] - \mathbb{E}_0[\mathbb{I}\{c^* = j\}]| \tag{26}$$

$$\le \frac{1}{2} \left\| \mathbb{P}_0(X_1^H) - \mathbb{P}_j(X_1^H) \right\|_1 \tag{27}$$

$$\le \frac{1}{2} \sqrt{2 \ln 2 KL(P_0(X_1^H) \parallel \mathbb{P}_j(X_1^H))}, \tag{28}$$

where the first inequality follows from Auer et al. (2002, Equation 28) since the final output c^* is a function of the outcomes X_1^H, and the last inequality is Pinsker inequality.

Since Alg is deterministic, the flip a_t at step t is fully determined given the previous outcomes x_1^{t-1}. Applying the chain rule for KL-divergences (Cover and Thomas, 2006, Theorem 2.5.3) we obtain

$$KL(P_0(X_1^H) \parallel \mathbb{P}_j(X_1^H)) = \sum_{t=1}^{H} \sum_{x_1^{t-1}} \mathbb{P}_0[x_{1:t-1}] KL(\mathbb{P}_0[x_t] \parallel \mathbb{P}_j[x_t] \mid x_1^{t-1}).$$

Note that x_t is the result of a single coin flip. When $a_{t,0} \neq j$, the KL-divergence is zero since the two instances have the identical coins on both columns. When $a_{t,0} = j$, the KL-divergence is either $B_1 = KL(\frac{1}{2} \parallel \frac{1}{2} + \lambda)$ or $B_2 = KL(\delta \parallel \delta + \epsilon)$, depending on whether $a_{t,1} = 1$ or $a_{t,1} = 2$, respectively. It follows that

$$\begin{aligned} KL(P_0(X_1^H) \parallel \mathbb{P}_j(X_1^H)) &= \sum_{t=1}^{H} \sum_{x_{1:t-1}} \mathbb{P}_0[x_{1:t-1}] \left(\mathbb{I}\{a_t = (j,1)\} B_1 + \mathbb{I}\{a_t = (j,2)\} B_2 \right) \\ &= \mathbb{E}_0[N_{j,1}] B_1 + \mathbb{E}_0[N_{j,2}] B_2 \\ &\le \frac{4T_1}{Q} B_1 + \frac{4T_2}{Q} B_2 \\ &\le \frac{B_1}{C_1 \lambda^2} + \frac{(\delta + \epsilon) B_2}{C_2 \epsilon^2} \end{aligned}$$

Since $\lambda \le \frac{1}{4}$ and $\delta + \epsilon \le \frac{1}{2}$, we can bound $B_1 \le \frac{5\lambda^2}{2 \ln 2}$ (Tulsiani, 2014) and $B_2 \le \frac{4 \log_2(e) \epsilon^2}{\delta + \epsilon}$. Consequently,

$$KL(P_0(X_1^H) \parallel \mathbb{P}_j(X_1^H)) \le \frac{5}{(2 \ln 2) C_1} + \frac{4 \log_2(e)}{C_2}$$

Plugging this into Equation 28 and applying $Q \geq 12$, we obtain

$$\begin{aligned}
\mathbb{P}_j[\hat{\theta} = j] &\leq \mathbb{P}_0[\hat{\theta} = j] + \frac{1}{2}\sqrt{2\ln 2\left(\frac{5}{(2\ln 2)C_1} + \frac{4\log_2(e)}{C_2}\right)} \\
&= \frac{3}{Q} + \frac{1}{2}\sqrt{\frac{5}{C_1} + \frac{8}{C_2}} \\
&\leq \frac{3}{12} + \frac{1}{2}\sqrt{\frac{5}{40} + \frac{8}{64}} \\
&= \frac{1}{2}.
\end{aligned}$$

■

The next result shows that if ϵ is sufficiently small, then any algorithm has to flip the coins in the first column sufficiently many times; otherwise the probability of failure is at least $\frac{1}{2}$.

Corollary 24 *Let Q, C_1 and C_2 be the constants defined in Lemma 23. Let $H > 0$ be the budget for the number of flips on both columns. If $\epsilon = \frac{1}{20}\sqrt{\frac{Q\delta}{H}}$, then for any algorithm* Alg*, if*

$$N_1 \leq \frac{Q}{4C_1\lambda^2},$$

then there exists a set $J \subseteq [Q]$ with $|J| \geq \frac{Q}{6}$ such that

$$\forall j \in J, \mathbb{P}_j[\hat{\theta} = j] \leq \frac{1}{2}.$$

Proof We will show that when $\epsilon = \frac{1}{20}\sqrt{\frac{Q\delta}{H}}$, the inequality $N_2 \leq T_2 = \frac{Q(\delta+\epsilon)}{4C_2\epsilon^2}$ holds trivially for any $N_2 \leq H$ (recall that H is the fixed budget for the total number of coin flips). The result then follows directly from Lemma 23. We have

$$\begin{aligned}
T_2 = \frac{Q(\delta+\epsilon)}{4C_2\epsilon^2} &\geq \frac{Q\delta}{4C_2\epsilon^2} \\
&= \frac{Q\delta}{256\epsilon^2} \qquad \text{since } C_2 = 64 \\
&= \frac{400}{256}H \\
&> H \\
&\geq N_2,
\end{aligned}$$

which implies that $N_2 \leq T_2$ always holds for any $N_2 \leq H$.

■

We are now ready to prove Lemma 8.

Proof (of Lemma 8) We construct $\mathcal{M}$ as the set of $\frac{SA}{12}$ 2-JAO MDPs in Figure 1 (right). Each MDP has a left part and a right part, where each part is a JAO MDP. The left part of the MDP m_i consists of two states $\{0, 2\}$ and $\frac{SA}{12}$ actions numbered from 1 to $\frac{SA}{12}$, where all actions from state 0 transition to state 2 with probability of $\frac{1}{2}$ or stay at state 0 with probability $\frac{1}{2}$, except for the i^{th} action that transitions to state 2 with probability $\frac{1}{2} + \frac{\lambda}{2}$ and stays at state 0 with probability $\frac{1}{2} - \frac{\lambda}{2}$. The right part of the i^{th} MDP consists of two states $\{0, 1\}$ and also $\frac{SA}{12}$ actions numbered from 1 to $\frac{SA}{12}$, where all actions from state 0 transition to state 1 with probability of $\delta = \frac{4}{D} \leq \frac{1}{4}$ or stays at state 0 with probability $1 - \delta$, except for the i^{th} action that transitions to state 2 with probability $\delta + \Delta$ and stays at state 0 with probability $1 - \delta - \Delta$. We set $\Delta = \frac{1}{20}\left(\sqrt{\frac{SA}{3HD}}\right)$. We will show the conversion from these 2-JAO MDPs to MDPs with S states and A actions later.

Since each model in $\mathcal{M}$ has a distinct index for the actions on both parts that transitions from 0 to 1 and 2 with probability higher than any other actions, identifying a model in $\mathcal{M}$ is equivalent to identifying this distinct action. Each action on both parts can be seen as a (possibly biased) coin, where the probability of getting tails is equal to the probability of ending up in state 0 when the action is taken. Thus, the problem of identifying this distinct action index reduces to the above auxiliary problem of identifying the row of the most biased coins, where taking an action from state 0 is equivalent to flipping a coin, $Q = \frac{SA}{12} \geq 12, \epsilon = \Delta$ and λ is replaced by $\lambda/2$. Corollary 24 states that for every algorithm, if the number of coin flips on the first column is less than $\frac{SA}{480\lambda^2}$, then there exists a set of size at least $\frac{SA}{72}$ positions of the row with the most biased coins such that the algorithm fails to find the biased coin with probability at least $\frac{1}{2}$. Correspondingly, for any model classification algorithm, if the number of state-transition samples from state 0 towards state 2 (i.e. the first column) is less than $\frac{SA}{480\lambda^2}$ then the algorithm fails to identify the model for at least $\frac{SA}{72}$ MDPs in $\mathcal{M}$.

Finally, we show the conversion from the 2-JAO MDP to an MDP with S states and A actions. The conversion is almost identical to that of Jaksch et al. (2010), which starts with an *atomic* 2-JAO MDP of three states and $A' = \frac{A}{2}$ actions and builds an A'-ary tree from there. Assuming A' is an even positive number, each part of the atomic 2-JAO MDP has $\frac{A'}{2}$ actions. We make $\frac{S}{3}$ copies of these atomic 2-JAO MDPs, where only one of them has the best action on the right part. Arranging $\frac{S}{3}$ copies of these atomic 2-JAO MDPs and connecting their states 0 by $A - A'$ connections, we obtain an A'-ary tree which represents a composite MDP with at most S states, A actions and diameter D. The transitions of the $A - A'$ actions on the tree are defined identically to that of Jaksch et al. (2010): self-loops for states 1 and 2, deterministic connections to the state 0 of other nodes on the tree for state 0. By having $\delta = \frac{4}{D}$ in each atomic 2-JAO MDP, the diameter of this composite MDP is at most $\frac{2}{\delta} + \log_{A'} \frac{S}{3} \leq D$. This composite MDP is harder to explore and learn than the 2-JAO MDP with three states and $\frac{SA}{6}$ actions, and hence all the lower bound results apply.

■

Corollary 9 *For any $S, A \geq 20, D \geq 16$ and $\lambda \in (0, 1]$, there exists a PAC identifiable λ-separable set of MDPs $\mathcal{M}$ of size $M = \frac{SA}{12}$, each with S states, A actions and diameter D such that for any uniformly good cluster-then-learn algorithm, to find the correct cluster with probability of at least $\frac{1}{2}$, the expected number of exploration steps needed in the clustering phase is $\Omega(\frac{DSA}{\lambda^2})$. Furthermore,* *the expected regret over K episodes of the same algorithm is*

$$\mathbb{E}[\text{Regret}(K)] \geq \Omega\left(\frac{KDSA}{\lambda^2}\right).$$

Proof As argued in Section 3, we can apply the sample complexity of the classification algorithm onto that of the clustering algorithm. Using the same set $\mathcal{M}$ of 2-JAO MDPs constructed in the proof of Lemma 8, for any given MDP $\mathcal{M}$, any PAC classification learner has to be in state 0 and takes at least $Z = \Omega(\frac{SA}{\lambda^2})$ actions from state 0 to state 2. If the learner stays at state 0, then it can take the next action from 0 to 2 in the next time step. However, if the learner transitions to state 2, then it has to wait until it gets back to state 0 to take the next action. Let Z_2 denote the number of times the learner ends up in state 2 after taking Z actions on the left part from state 0. Since every action from 0 to 2 has probability at least $\frac{1}{2}$ of ending up in state 2, we have

$$\mathbb{E}[Z_2 \mid Z] \geq \frac{Z}{2}. \tag{29}$$

Since every action from state 2 transitions to state 0 with the same probability of $\delta = \Theta(\frac{1}{D})$, every time the learner is in state 2, the expected number of time steps it needs to get back to state 0 is $\Theta(\frac{1}{\delta}) = \Theta(D)$. Hence, the expected number of time steps the learner needs to get back to state 0 after Z_2 times being in state 2 is $\Theta(Z_2 D)$. We conclude that for any PAC learner, the expected number of exploration steps needed to identify the model with probability of correct at least $\frac{1}{2}$ is at least

$$\mathbb{E}[Z + Z_2 D] \geq \Omega(ZD) = \Omega\left(\frac{DSA}{\lambda^2}\right). \tag{30}$$

Next, we lower bound the expected regret of the same algorithm. Let H_0 be the number of time steps the algorithm spends on the left part and H_1 on the right part of each model in $\mathcal{M}$. Note that H_0 and H_1 are random variables. Recall that the right part of each MDP in $\mathcal{M}$ resembles the JAO MDP in the minimax lower bound proof in Lemma 5, hence we can apply the regret formula of the JAO MDP for 2-JAO MDP and obtain that the regret in each episode is of the same order as

$$\begin{aligned}
\text{Regret} &= \rho^* H - \mathbb{E}[\sum_{h=1}^{H} r(s_h, a_h)] && (31)\\
&= \rho^* E[H_0 + H_1] - \mathbb{E}[\mathbb{E}[\sum_{h=1}^{H_0} r(s_h, a_h)] + \mathbb{E}[\sum_{h=H_0+1}^{H} r(s_h, a_h)] \mid H_0, H_1] && (32)\\
&= \rho^* E[H_0 + H_1] - \mathbb{E}[\mathbb{E}[\sum_{h=H_0+1}^{H} r(s_h, a_h) \mid H_0, H_1]] && (33)\\
&= \rho^* E[H_0] + E\left[\left(\rho^* H_1 - \mathbb{E}[\sum_{h=H_0+1}^{H} r(s_h, a_h)]\right) \mid H_1\right] && (34)\\
&\geq \Omega\left(\rho^* E[H_0]\right) - \frac{D}{2} && (35)\\
&= \Omega\left(\frac{DSA}{\lambda^2}\right), && (36)
\end{aligned}$$

where

- the second equality follows from $H = H_0 + H_1$,
- the third equality follows from the fact that the H_0 time steps spent on the left part of the MDP returns no rewards,
- the fourth equality follows from the linearity of expectation,
- the inequality follows from $H_1 = H - H_0$ and equation 20,
- the last equality follows from $\rho^* = \frac{\delta+\Delta}{2\delta+\Delta} \geq \frac{1}{2}$ for all $\delta, \Delta > 0$ and $E[H_0] \geq \Omega\left(\frac{DSA}{\lambda^2}\right)$.

We conclude that the expected regret over K episodes is at least

$$\Omega(\mathbb{E}[KH_0]) = \Omega\left(\frac{KDSA}{\lambda^2}\right).$$

■

Appendix D. Proofs of the upper bounds

First, we state the following concentration inequality for vector-valued random variables by Weissman et al. (2003).

Lemma 25 (Weissman et al. (2003)) *Let P be a probability distribution on the set $\mathcal{S} = \{1, \ldots, S\}$. Let $\mathcal{X}^N$ be a set of N i.i.d samples drawn from P. Then, for all $\epsilon > 0$:*

$$\Pr(\left\|P - \hat{P}_{\mathcal{X}^N}\right\| \geq \epsilon) \leq (2^S - 2)e^{-N\epsilon^2/2}.$$

Using Lemma 25, we can show that $N = O(\frac{S}{\lambda^2})$ samples are sufficient for each $(s, a) \in \Gamma$ so that with high probability, the empirical means of the transition function $\hat{P}_{\mathcal{B}}(\cdot \mid s, a)$ are within $\lambda/8$ of their true values, measured in ℓ_1 distance.

Corollary 26 *Denote $p_1 \in (0, 1)$. If a state-action pair (s, a) is visited at least*

$$N = \frac{256}{\lambda^2} \max\{S, \ln(1/p_1)\} \tag{37}$$

times, then with probability at least $1 - p_1$,

$$\left\|P(s, a) - \hat{P}_{\mathcal{X}^N}(s, a)\right\| \leq \lambda/8.$$

Proof We simplify the bound in Lemma 25 as follows:

$$\Pr(\left\|P - \hat{P}_{\mathcal{X}^N}\right\| \geq \epsilon) \leq (2^S - 2)e^{-N\epsilon^2/2} \leq e^{S - N\epsilon^2/2}$$

Next, we substitute $\epsilon = \lambda/8$ into the right hand side and solve the following inequality for N:

$$e^{S-N\lambda^2/128} \le p_1$$

to obtain $N \ge \frac{128}{\lambda^2}(S + \ln(1/p_1))$. Thus $N = \frac{256}{\lambda^2}\max\{S, \ln(1/p_1)\}$ satisfies this condition. ■

Taking a union bound of the result in Corollary 26 over all state-action pairs in the set Γ of all episodes from 1 to K, we obtain Lemma 13.

Next, we show the proof of Lemma 14. The proof strategy is similar to that of Auer and Ortner (2007); Sun and Huang (2020).

Lemma 14 *Consider p_1 and N defined in Lemma 13. By setting*

$$H_0 = 12\tilde{D}|\Gamma^\alpha|N = \frac{3072\tilde{D}|\Gamma^\alpha|}{\lambda^2}\max\{S, \ln(\frac{K|\Gamma^\alpha|}{p_1})\},$$

with probability at least $1 - p_1$, Algorithm 2 visits each state-action pair in Γ^α at least N times during the clustering phase in each of the K episodes.

Proof The history-dependent exploration policy in Algorithm 2 visits an under-sampled state-action pair in Γ^α whenever possible; otherwise it starts a sequence of steps that would lead to such a state-action pair. In the latter case, denote the current state of the learner by s and the number of steps needed to travel from s' to an under-sampled state s by $T(s', s)$. By Assumption 4 and using Markov inequality, we have

$$\Pr(T(s', s) > 2\tilde{D}) \le \frac{E[T(s', s)]}{2\tilde{D}} \le \frac{\tilde{D}}{2\tilde{D}} = \frac{1}{2}.$$

It follows that $\Pr(T(s', s) > 2\tilde{D}) \le 1/2$. In other words, in every interval of $2\tilde{D}$ time steps, the probability of visiting an under-sampled state-action pair in Γ^α is at least $1/2$. Over such n intervals, the expected number of such visits is lower bounded by $n/2$. Fix a $(s, a) \in \Gamma^\alpha$. Let V_n denote number of visits to $(s, a) \in \Gamma^\alpha$ after n intervals. Using a Chernoff bound for Poisson trials, we have

$$\Pr(V_n \ge (1 - \epsilon)n/2) \ge 1 - e^{-\epsilon^2 n/4}$$

for any $\epsilon \in (0, 1)$. Setting $\epsilon = 1 - 2N/m$ and solving

$$e^{-(1-2N/n)^2 n/4} \le p_1$$

for n, we obtain

$$n \ge 2(N + \ln(1/p_1)) + 2\sqrt{2N\ln(1/p_1) + (\ln(1/p_1))^2}. \tag{38}$$

By definition of N,

$$\begin{aligned} 2N\ln(1/p_1) + (\ln(1/p_1))^2 &\le (1 + \frac{512}{\lambda^2})\max\{S, \ln(1/p_1)\}^2 \\ &\le \left(\frac{256}{\lambda}\max\{S, \ln(1/p_1)\}\right)^2 \\ &\le N^2. \end{aligned}$$

We also have $N \geq \ln(1/p)$. Overall, $n = 6N$ satisfies the condition in Equation 38. Taking a union bound over all $(s,a) \in \Gamma^\alpha$ and noting that each interval has length $2\tilde{D}$ steps, the total number of identifying steps needed is $H_0 = 2\tilde{D}n|\Gamma^\alpha| = 12\tilde{D}|\Gamma^\alpha|N$. ■

To prove Lemma 15, we state the following auxiliary proposition and its corollary.

Proposition 27 *Suppose we are given a probability distribution P over $\mathcal{S} = 1, \dots, S$, a constant $\epsilon > 0$ and two set of samples $\mathcal{X} = (X_1, \dots, X_{N_\mathcal{X}})$ and $\mathcal{Y} = (Y_1, \dots, Y_{N_\mathcal{Y}})$ drawn from P such that $\left\| P - \hat{P}_\mathcal{X} \right\| \leq \epsilon$ and $\left\| P - \hat{P}_\mathcal{Y} \right\| \leq \epsilon$. Then,*

$$\left\| P - \hat{P}_{\mathcal{X}\cup\mathcal{Y}} \right\| \leq \epsilon.$$

Proof Let $N_\mathcal{X}(s)$ and $N_\mathcal{Y}(s)$ denote the number of samples of $s \in [S]$ in $\mathcal{X}$ and $\mathcal{Y}$, respectively. We have:

$$\begin{aligned}
\left\| P - \hat{P}_{\mathcal{X}\cup\mathcal{Y}} \right\| &= \sum_{s=1}^{S} |P(s) - \frac{N_\mathcal{X}(s) + N_\mathcal{Y}(s)}{N_\mathcal{X} + N_\mathcal{Y}}| && (39)\\
&= \frac{1}{N_\mathcal{X} + N_\mathcal{Y}} \sum_{s=1}^{S} |N_\mathcal{X} P(s) - N_\mathcal{X}(s) + N_\mathcal{Y} P(s) - N_\mathcal{Y}(s)| && (40)\\
&\leq \frac{1}{N_\mathcal{X} + N_\mathcal{Y}} \sum_{s=1}^{S} (|N_\mathcal{X} P(s) - N_\mathcal{X}(s)| + |N_\mathcal{Y} P(s) - N_\mathcal{Y}(s)|) \quad \text{(triangle inequality)} && (41)\\
&= \frac{1}{N_\mathcal{X} + N_\mathcal{Y}} \left(N_\mathcal{X} \sum_{s=1}^{S} |P(s) - \frac{N_\mathcal{X}(s)}{N_\mathcal{X}}| \right) + \frac{1}{N_\mathcal{X} + N_\mathcal{Y}} \left(N_\mathcal{Y} \sum_{s=1}^{S} |P(s) - \frac{N_\mathcal{Y}(s)}{N_\mathcal{Y}}| \right) && (42)\\
&= \frac{1}{N_\mathcal{X} + N_\mathcal{Y}} (N_\mathcal{X} \left\| P - \hat{P}_\mathcal{X} \right\|_1 + N_\mathcal{Y} \left\| P - \hat{P}_\mathcal{Y} \right\|) && (43)\\
&\leq \epsilon && (44)
\end{aligned}$$

■

Corollary 28 *Suppose we are given a probability distribution P over $\mathcal{S} = 1, \dots, S$, a constant $\epsilon > 0$ and a finite number of set of samples $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_t$ such that $\left\| P - \hat{P}_{\mathcal{X}_i} \right\| \leq \epsilon$ for all $i = 1, 2, \dots, t$. Then,*

$$\left\| P - \hat{P}_{\cup_{i=1,\dots,t} \mathcal{X}_i} \right\| \leq \epsilon. \tag{45}$$

Proof (Of Lemma 15) The proof is by induction. The claim is trivially true for the first episode ($k = 1$). For an episode $k > 1$, assume that the outputs of the Algorithm 3 are correct until the beginning of this episode. We consider two cases:

- When the task m_k has never been given to the learner before episode k.

  Consider an arbitrary existing cluster c. Denote by $i \in [M]$ the identity of the model to which the samples in c belong, $j \in [M]$ the identity of the task m_k, and (s,a) in $\Gamma^\alpha_{i,j}$ a state-action pair that distinguishes these two models. Under the definition of $\Gamma^\alpha_{i,j}$, the result in Lemma 13 and the result in Corollary 28, the following three inequalities hold true:

$$\begin{aligned}\|[P_i - P_j](s,a)\| &> \alpha \\ \left\|[P_j - \hat{P}_{\mathcal{B}_k}](s,a)\right\| &\leq \lambda/8 \\ \left\|[P_i - \hat{P}_c](s,a)\right\| &\leq \lambda/8.\end{aligned}$$

  From here, we omit the (s,a) and write P for $P(s,a)$ when no confusion is possible. Applying the triangle inequality twice, we obtain:

$$\begin{aligned}\left\|\hat{P}_c - \hat{P}_{\mathcal{B}_k}\right\| &\geq \|P_i - P_j\| - (\left\|P_i - \hat{P}_c\right\| + \left\|P_j - \hat{P}_{\mathcal{B}_k}\right\|) \\ &> \alpha - (\lambda/8 + \lambda/8) \\ &= \delta.\end{aligned}$$

  It follows that the `break` condition in Algorithm 3 is satisfied, and the correct value of 0 is returned. A new cluster is created containing only the samples generated by the new task m_k.

- When the task m_k has been given to the learner before episode k.

  In this case, there exists a cluster c' containing the samples generated from model j. Using a similar argument in the previous part, we have that whenever the iteration in Algorithm 3 reaches a cluster c whose identity $i \neq j$, the `break` condition is true for at least one $(s,a) \in \Gamma^\alpha$, and the algorithm moves to the next cluster. When the iteration reaches cluster c', for all $(s,a) \in \tilde{\Gamma}^\alpha$, we have:

$$\begin{aligned}\left\|\hat{P}_{\mathcal{B}_k} - \hat{P}_{c'}\right\| &\leq \left\|\hat{P}_{\mathcal{B}_k} - P_j\right\| + \left\|P_j - \hat{P}_{c'}\right\| \\ &\leq \lambda/8 + \lambda/8 = \lambda/4 \\ &\leq \delta.\end{aligned}$$

  Hence, the `break` condition is false for all $(s,a) \in \Gamma$, and thus the algorithm returns `id` $= c'$ as expected.

By induction, under event $\mathcal{E}_\Gamma$, Algorithm 3 always produces correct outputs throughout the K episodes. ∎

We can now state the regret bound of Algorithm 1 where the regret minimization algorithm in every episode is UCBVI-CH (Azar et al., 2017). For each state-action pair (s,a) in episode k, UCBVI-CH needs a bonus term defined as

$$b_k(s,a) = 7H_1L_k\sqrt{\frac{1}{N_k^{regret}(s,a)}},$$

where $L_k = \ln(5SAK_{m_k}H_1/p_1)$, $N_k^{regret}(s,a)$ is the total number of visits to (s,a) in the learning phase before episode k, and K_{m^k} is the total number of episodes in which the model m^k is given to the learner. However, K_{m^k} is unknown to the learner. We instead upper bound K_{m^k} by K and modify the bonus term as

$$b'_k(s,a) = 7H_1L\sqrt{\frac{1}{N_k^{regret}(s,a)}} \tag{46}$$

where $L = \ln(5SAKHM/p_1)$. Since $b'_k \geq b_k$, this algorithm still retain the optimism principle needed for UCBVI-CH. The total regret of each model in $\mathcal{M}$ is bounded by the following result, whose proof is in Appendix E.

Lemma 29 *With probability at least $1-p_1$, applying UCBVI-CH with the bonus term b'_k defined in Equation 46, each task m in $\mathcal{M}$ has a total regret of*

$$\textit{Regret}(m, K_m) \leq K_m(H_0 + D) + 67H_1^{3/2}L\sqrt{SAK_m} + 15S^2A^2H_1^2L^2$$

Theorem 16 *For any failure probability $p \in (0,1)$, with probability at least $1-p$ the regret of Algorithm 1 is bounded as*

$$\textit{Regret}(K) \leq 2KH_0 + 67H_1^{3/2}L\sqrt{MSAK} + 15MS^2AH_1^2L^2,$$

where $H_0 = 12\tilde{D}|\Gamma^\alpha|N$, $N = \frac{256}{\lambda^2}\max\{S, \ln(\frac{3K|\Gamma^\alpha|}{p})\}$, $H_1 = H - H_0$, and $L = \ln(15SAKHM/p)$.

Proof Summing up the regret for all $m \in \mathcal{M}$ and applying the Cauchy-Schwarz inequality, Lemma 29 together with Lemma 15 and Lemma 14 imply that with probability $1-p$, the total regret is bounded by

$$\text{Regret}(K) \leq K(H_0 + D) + 67H_1L\sqrt{MSAKH_1} + 15MS^2AH_1^2L^2. \tag{47}$$

Note that the bound in Equation 47 is tighter than the bound in Theorem 16. To obtain the bound in Theorem 16, notice that $D \leq \tilde{D} \leq H_0$ and thus $K(H_0 + D) \leq K(H_0 + H_0) = 2KH_0$. ■

Theorem 20 *Under Assumption 19, With probability at least $1-p$, the regret of Algorithm 4 is*

$$\textit{Regret}(K) = O\left(\frac{K\tilde{D}M^2}{\lambda^2}\ln\frac{KM^2}{p} + H^{3/2}L\sqrt{MKSA}\right),$$

where $H_{0,M} = \frac{3072\tilde{D}M^2}{\lambda^2}\max\{S, \ln(\frac{3KM^2}{p})\}$ and $L = \ln(15SAKH_1M/p)$.

Proof In stage 1, as the distinguishing set has size $|\tilde{\Gamma}| = SA$, the number of time steps needed in the clustering phase is

$$H_{0,1} = 12\tilde{D}|\tilde{\Gamma}|N_1 = 12DSAN_1,$$

where $N_1 = \frac{256}{\lambda^2}\max\{S, \ln(\frac{3KSA}{p})\}$.

In stage 2, the length of the clustering phase is

$$H_{0,2} = 12\tilde{D}|\hat{\Gamma}|N_2,$$

where $N_2 = \frac{256}{\lambda^2}\max\{S, \ln(\frac{3K|\hat{\Gamma}|}{p})\}$.

Substituting $H_{0,1}$ and $H_{0,2}$ into Theorem 16, we obtain the regret bound of stage 1 and stage 2:

$$\text{Regret}_{Stage1} \leq 2K_1H_{0,1} + 67(H_{1,1})^{3/2}L_1\sqrt{MSAK_1} + 15MS^2A(H_{1,1})^2L_1^2,$$

where $L_1 = \ln(\frac{15MSAKH_{1,1}}{p})$ and $H_{1,1} = H - H_{0,1}$.

$$\text{Regret}_{Stage2} \leq 2K_2H_{0,2} + 67H_{1,2}^{3/2}L_2\sqrt{MSAK_2} + 15MS^2AH_{1,2}^2L_2^2,$$

where $L_2 = \ln(\frac{15MSAKH_{1,2}}{p})$ and $H_{1,2} = H - H_{0,2}$.

Since $H_{0,1} \geq H_{0,2}$, we have $H_{1,1} \leq H_{1,2}$. Using the assumption that $K_1SA < K_2$ and the Cauchy-Schwarz inequality for the sum $\sqrt{K_1} + \sqrt{K_2}$, we obtain

$$\text{Regret}(K) = \text{Regret}_{Stage1} + \text{Regret}_{Stage2} \tag{48}$$
$$\leq 4KH_{0,2} + 67H_{1,2}^{3/2}L_2\sqrt{2MSAK} + 30MS^2AH_{1,2}^2L_2^2. \tag{49}$$

By having $|\hat{\Gamma}| \leq \binom{M}{2} \leq M^2$, $H_{1,2} \leq H$ and $\max\{L_1, L_2\} \leq L$, we obtain

$$\text{Regret}(K) \leq 4KH_{0,M} + 67H^{3/2}L\sqrt{2MSAK} + 30MS^2AH^2L^2. \tag{50}$$

where $H_{0,M} = \frac{3072\tilde{D}M^2}{\lambda^2}\max\{S, \ln(\frac{3KM^2}{p})\}$.

■

Appendix E. Per-model Regret analysis

First, we prove the following lemma which upper bound the per-episode regret as a function of H_0 and the regret of the clustering phase.

Lemma 30 *The regret of Algorithm 1 in episode k is*

$$\Delta_k = [V_1^{k,*} - V_1^{\pi_k}](s_1^k) \leq H_0 + D + \max_{s\in\mathcal{S}}[V_{H_0+1}^{k,*} - V_{H_0+1}^{\pi_k}](s).$$

Proof Denote by $\Pr(s_h^k = s \mid s_1, \pi)$ the probability of visiting state s at time h when the learner follows a (possibly non-stationary) policy π in model m^k starting from state s_1. The regret of task

m in a single episode $k \in \mathcal{K}_m$ can be written as

$$\begin{aligned}
\Delta_k &= [V_1^{k,*} - V_1^{\pi_k}](s_1^k) \\
&= E[\sum_{h=1}^{H} r(s_h, a_h) \mid s_1 = s_1^k, a_h = \pi_k^*(s_h)] - E[\sum_{h=1}^{H} r(s_h, a_h) \mid s_1 = s_1^k, a_h = \pi_k(s_h)] \\
&= \left(E[\sum_{h=1}^{H_0} r(s_h, a_h) \mid s_1 = s_1^k, a_h = \pi_k^*(s_h)] + \sum_{s \in \mathcal{S}} \mathrm{Pr}_m(s_{H_0+1}^k = s \mid s_1^k, \pi_k^*) V_{H_0+1}^{k,*}(s) \right) \\
&\quad - \left(E[\sum_{h=1}^{H_0} r(s_h, a_h) \mid s_1 = s_1^k, a_h = \pi_k(s_h)] + \sum_{s \in \mathcal{S}} \mathrm{Pr}_m(s_{H_0+1}^k = s \mid s_1^k, \pi_k) V_{H_0+1}^{\pi_k}(s) \right) \\
&\leq H_0 + \sum_{s \in \mathcal{S}} \mathrm{Pr}_m(s_{H_0+1}^k = s \mid s_1^k, \pi_k^*) V_{H_0+1}^{k,*}(s) - \sum_{s \in \mathcal{S}} \mathrm{Pr}_m(s_{H_0+1}^k = s \mid s_1^k, \pi_k) V_{H_0+1}^{\pi_k}(s) \\
&= H_0 + \left(\sum_{s \in \mathcal{S}} \mathrm{Pr}_m(s_{H_0+1}^k = s \mid s_1^k, \pi_k^*) V_{H_0+1}^{k,*}(s) - \sum_{s \in \mathcal{S}} \mathrm{Pr}_m(s_{H_0+1}^k = s \mid s_1^k, \pi_k) V_{H_0+1}^{k,*}(s) \right) \\
&\quad + \sum_{s \in \mathcal{S}} \mathrm{Pr}_m(s_{H_0+1}^k = s \mid s_1^k, \pi_k) [V_{H_0+1}^{k,*} - V_{H_0+1}^{\pi_k}](s) \\
&\leq H_0 + \underbrace{\left(\max_{s \in \mathcal{S}} V_{H_0+1}^{k,*}(s) - \min_{s \in \mathcal{S}} V_{H_0+1}^{k,*}(s) \right)}_{(\clubsuit)} + \max_{s \in \mathcal{S}} [V_{H_0+1}^{k,*} - V_{H_0+1}^{\pi_k}](s).
\end{aligned}$$

The first inequality follows from the assumption that $r(s,a) \in [0,1]$ for all (s,a). The second inequality follows the fact that

$$\begin{aligned}
\sum_{s \in \mathcal{S}} \mathrm{Pr}_m(s_{H_0+1}^k = s \mid s_1^k, \pi_k^*) V_{H_0+1}^{k,*}(s) &\leq \sum_{s \in \mathcal{S}} \mathrm{Pr}_m(s_{H_0+1}^k = s \mid s_1^k, \pi_k^*) \max_{x \in \mathcal{S}} V_{H_0+1}^{k,*}(x) \\
&= \left(\max_{x \in \mathcal{S}} V_{H_0+1}^{k,*}(x) \right) \sum_{s \in \mathcal{S}} \mathrm{Pr}_m(s_{H_0+1}^k = s \mid s_1^k, \pi_k^*) \\
&= \max_{x \in \mathcal{S}} V_{H_0+1}^{k,*}(x),
\end{aligned}$$

and

$$\begin{aligned}
\sum_{s \in \mathcal{S}} \mathrm{Pr}_m(s_{H_0+1}^k = s \mid s_1^k, \pi_k) V_{H_0+1}^{k,*}(s) &\geq \sum_{s \in \mathcal{S}} \mathrm{Pr}_m(s_{H_0+1}^k = s \mid s_1^k, \pi_k) \min_{x \in \mathcal{S}} V_{H_0+1}^{k,*}(x) \\
&= \left(\min_{x \in \mathcal{S}} V_{H_0+1}^{k,*}(x) \right) \sum_{s \in \mathcal{S}} \mathrm{Pr}_m(s_{H_0+1}^k = s \mid s_1^k, \pi_k) \\
&= \min_{x \in \mathcal{S}} V_{H_0+1}^{k,*}(x).
\end{aligned}$$

Furthermore, since $V^{k,*}_{H_0+1}(s) \geq V^{\pi_k}_{H_0+1}(s)$ for all $s \in \mathcal{S}$, we have

$$\begin{aligned}\sum_{s\in\mathcal{S}} \Pr_m(s^k_{H_0+1} = s \mid s^k_1, \pi_k)[V^{k,*}_{H_0+1} - V^{\pi_k}_{H_0+1}](s) &\leq \sum_{s\in\mathcal{S}} \Pr_m(s^k_{H_0+1} = s \mid s^k_1, \pi_k) \max_{x\in\mathcal{S}}[V^{k,*}_{H_0+1} - V^{\pi_k}_{H_0+1}](x) \\ &= \max_{x\in\mathcal{S}}[V^{k,*}_{H_0+1} - V^{\pi_k}_{H_0+1}](x) \sum_{s\in\mathcal{S}} \Pr_m(s^k_{H_0+1} = s \mid s^k_1, \pi_k) \\ &= \max_{x\in\mathcal{S}}[V^{k,*}_{H_0+1} - V^{\pi_k}_{H_0+1}](x).\end{aligned}$$

For each state s, the value of $V^{k,*}_h(s)$ is the expected total $(H-h)$-step reward of an optimal non-stationary $(H-h)$ step policy starting in state s on the MDP m. Thus, the term (♣) represents the *bounded span* of the finite-step value function in MDP m. Applying equation 11 of Jaksch et al. (2010), the span of the value function is bounded by the diameter of the MDP. We obtain for all h

$$\max_{s\in\mathcal{S}} V^{k,*}_h(s) - \min_{s\in\mathcal{S}} V^{k,*}_h(s) \leq D.$$

It follows that

$$\Delta_k \leq H_0 + D + \max_{s\in\mathcal{S}}[V^{k,*}_{H_0+1} - V^{\pi_k}_{H_0+1}](s).$$

■

Denote $\mathcal{K}_m$ the set of episodes where the model m is given to the learner. The total regret of the learner in episodes $\mathcal{K}_m$ is

$$\begin{aligned}\text{Regret}(m, K_m) &= \sum_{k\in\mathcal{K}_m} \Delta_k \\ &\leq K_m(H_0 + D) + \underbrace{\sum_{k\in\mathcal{K}_m} \max_{s\in S}[V^{k,*}_{H_0+1} - V^{\pi_k}_{H_0+1}](s)}_{(\heartsuit)}.\end{aligned}$$

The policy π_k from time step $H_0 + 1$ to H is the UCBVI-CH algorithm (Azar et al., 2017). Therefore, the term $(\heartsuit)$ corresponds to the total regret of UCBVI-CH in an adversarial setting in which the starting state s^k_1 in each episode is chosen by an adversary that maximizes the regret in each episode. In Appendix F, we given a simplified analysis for UCBVI-CH and show that with probability at least $1 - p_1/M$,

$$(\heartsuit) = \sum_{k\in\mathcal{K}_m} \max_{s\in S}[V^{k,*}_{H_0+1} - V^{\pi_k}_{H_0+1}](s) \leq 67H_1^{3/2}L\sqrt{SAK_m} + 15S^2A^2H_1^2L^2. \tag{51}$$

The proof of Lemma 29 is completed by plugging the bound of (2) in Equation 51 to obtain

$$\begin{aligned}\text{Regret}(m, K_m) &= \sum_{k\in\mathcal{K}_m} \Delta_k \\ &\leq K_m(H_0 + D) + 67H_1^{3/2}L\sqrt{SAK_m} + 15S^2A^2H_1^2L^2.\end{aligned}$$

Algorithm 5: UCBVI

Input: Failure probability p

Initialize an empty collection $\mathcal{B}$;

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

  $Q_{k,h} =$ `UCB-Q-Values` $(\mathcal{B}, p)$;

  for $h = 1, \ldots, H$: **do**

    Take action $a_{k,h} = \arg\max_a Q_{k,h}(s_h^k, a)$;

    Add $(s_h^k, a_h^k, s_{h+1}^k)$ to $\mathcal{B}$;

Algorithm 6: UCB-Q-Values with Hoeffding bonus

Input: Collection $\mathcal{B}$, probability p

Compute, for all $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$

  $N_k(s, a, s') = \sum_{(x,a',y)\in\mathcal{B}} \mathbb{I}(x = s, a' = a, y = s')$

  $N_k(s, a) = \sum_{s' \in S} N_k(s, a, s')$;

For all $(s, a) \in \{(s, a) : N_k(s, a) > 0\}$, compute

  $\hat{P}_k(s' \mid s, a) = \frac{N_k(s,a,s')}{N_k(s,a)}$

  $b_{k,h}(s, a) = 7HL\sqrt{\frac{1}{N_k(s,a)}}$ where $L = \ln(5SAKH/p)$;

Initialize $V_{k,H+1}(s) = 0$ for all $x \in \mathcal{S}$;

for $h = H, H - 1, \ldots, 1$: **do**

  for $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**

    if $N_k(s, a) > 0$ **then**

      $Q_{k,h}(s, a) = \min\{H, r(s, a) + \left(\sum_{s' \in \mathcal{S}} \hat{P}_k(s' \mid s, a)V_{k,h+1}(s')\right) + b_{k,h}(s, a)\}$

    else

      $Q_{k,h} = H$

    $V_{k,h}(s) = \max_a Q_{k,h}(s, a)$

Appendix F. A simplified analysis for UCBVI-CH

In section, we construct a simplified analysis for the UCBVI-CH algorithm in Azar et al. (2017). The proof largely follows the existing constructions in Azar et al. (2017), with two differences: the definition of "typical" episodes and the analysis are tailored specifically for the Chernoff-type bonus of UCBVI-CH, without being complicated by handling of the variances for the Bernstein-type bonus of UCBVI-BF in Azar et al. (2017). For completeness, the full UCBVI-CH algorithm from Azar et al. (2017) is shown in Algorithms 5 and 6.

Notation. In this section, we consider the standard single-task episodic RL setting in Azar et al. (2017) where the learner is given the same MDP $(\mathcal{S}, \mathcal{A}, H, P, r)$ in K episodes. We assume the reward function $r : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ is deterministic and known. The state and action spaces $\mathcal{S}$ and $\mathcal{A}$ are discrete spaces with size S and A, respectively. Denote by p the failure probability and let $L = \ln(5SAKH/p)$. We assume the product $SAKH$ is sufficiently large that $L > 1$.

Let V_1^* denote the optimal value function and $V_1^{\pi_k}$ the value function of the policy π_k of the UCBVI-CH agent in episode k. The regret is defined as follows.

$$\text{Regret}(K) = \sum_{k=1}^{K} \delta_{k,1}, \tag{52}$$

where $\delta_{k,h} = [V_h^* - V_h^{\pi_k}](s_h^k)$.

Denote by $N_k(s,a)$ the number of visits to the state-action pair (s,a) up to the beginning of episode k.

We call an episode k "typical" if all state-action pairs visited in episode k have been visited at least H times at the beginning of episode k. The set of typical episodes is defined as follows.

$$[K]_{typ} = \{i \in [K] : \forall h \in [H], N_i(s_h^i, a_h^i) \geq H\}. \tag{53}$$

Equation 52 can be written as

$$\begin{aligned}
\text{Regret}(K) &= \sum_{k \notin [K]_{typ}} \delta_{k,1} + \sum_{k \in [K]_{typ}} \delta_{k,1} \\
&\leq \sum_{k \notin [K]_{typ}} H + \sum_{k \in [K]_{typ}} \delta_{k,1} \\
&\leq SAH^2 + \sum_{k \in [K]_{typ}} \delta_{k,1}.
\end{aligned} \tag{54}$$

The first inequality follows from the trivial upper bound of the regret in an episode $\delta_{k,1} \leq H$. The second inequality comes from the fact that each state-action pair can cause at most H episodes to be non-typical; therefore there are at most SAH non-typical episodes.

Next, we have:

$$\sum_{k \in [K]_{typ}} \delta_{k,1} = \sum_{k}^{K} \delta_{k,1} \mathbb{I}\{k \in [K]_{typ}\}. \tag{55}$$

From here we write $\mathbb{I}_k = \mathbb{I}\{k \in [K]_{typ}\}$ for brevity.

Lemma 3 in Azar et al. (2017) implies that, for all $k \in [K]$,

$$\delta_{k,1} \leq e \sum_{h=1}^{H} \left[\varepsilon_{k,h} + 2\sqrt{L}\bar{\varepsilon}_{k,h} + c_{1,k,h} + b_{k,h} + c_{4,k,h} \right]. \tag{56}$$

where $c_{4,k,h} = \frac{4SH^2L}{N_k(s_h^k, a_h^k)}$, $\varepsilon_{k,h}$ and $\bar{\varepsilon}_{k,h}$ are martingale difference sequences which, by Lemma 5 in Azar et al. (2017), satisfy

$$\begin{aligned}
\sum_{k=1}^{K} \sum_{h=1}^{H} \varepsilon_{k,h} &\leq H\sqrt{KHL} \\
\sum_{k=1}^{K} \sum_{h=1}^{H} \bar{\varepsilon}_{k,h} &\leq \sqrt{KH},
\end{aligned} \tag{57}$$

and $c_{1,k,h}$ is a confidence interval to be defined later.

Plugging Equation 56 into Equation 55 and combining with Equation 57, we obtain:

$$\begin{aligned}
\sum_{k\in[K]_{typ}} \delta_{k,1} &\leq e \sum_{k=1}^{K} \left(\sum_{h=1}^{H} \left[\varepsilon_{k,h} + 2\sqrt{L}\bar{\varepsilon}_{k,h} + c_{1,k,h} + b_{k,h} + c_{4,k,h} \right] \right) \mathbb{I}_k \\
&= e \left[\left(\sum_{k=1}^{K} \mathbb{I}_k \sum_{h=1}^{H} (\varepsilon_{k,h} + 2\sqrt{L}\bar{\varepsilon}_{k,h}) \right) + \left(\sum_{k=1}^{K} \mathbb{I}_k \sum_{h=1}^{H} (b_{k,h} + c_{1,k,h} + c_{4,k,h}) \right) \right] \\
&\leq e \left[\left(\sum_{k=1}^{K} \sum_{h=1}^{H} (\varepsilon_{k,h} + 2\sqrt{L}\bar{\varepsilon}_{k,h}) \right) + \left(\sum_{k=1}^{K} \sum_{h=1}^{H} (b_{k,h}\mathbb{I}_k + c_{1,k,h}\mathbb{I}_k + c_{4,k,h}\mathbb{I}_k) \right) \right] \\
&\leq e \left[\left(H\sqrt{KHL} + 2\sqrt{L}\sqrt{KH} \right) + \left(\sum_{k=1}^{K} \sum_{h=1}^{H} (b_{k,h}\mathbb{I}_k + c_{1,k,h}\mathbb{I}_k + c_{4,k,h}\mathbb{I}_k) \right) \right] \\
&= e \left[\left((H+2)\sqrt{KHL} \right) + \left(\sum_{k=1}^{K} \sum_{h=1}^{H} (b_{k,h}\mathbb{I}_k + c_{1,k,h}\mathbb{I}_k + c_{4,k,h}\mathbb{I}_k) \right) \right]
\end{aligned}$$

Note that the second inequality follows from the fact that $\mathbb{I}_k \leq 1$, and the last inequality follows directly from Equation 57.

Let $\mathbb{I}_{k,h} = \mathbb{I}\{N_k(s_h^k, a_h^k) \geq H\}$. By the definition of a "typical" episode, $\mathbb{I}_k = 1$ implies that $\mathbb{I}_{k,h} = 1$ for all h. It follows that $\mathbb{I}_k \leq \mathbb{I}_{k,h}$. Thus,

$$\sum_{k\in[K]_{typ}} \delta_{k,1} \leq e \left((H+2)\sqrt{KHL} + \sum_{i=1}^{K} \sum_{j=1}^{H} (b'_{k,h} + c'_{1,k,h} + c'_{4,k,h}) \right), \tag{58}$$

where $b'_{k,h} = b_{k,h}\mathbb{I}_{k,h}$, $c'_{1,k,h} = c_{1,k,h}\mathbb{I}_{k,h}$ and $c'_{4,k,h} = c_{4,k,h}\mathbb{I}_{k,h}$.

Next, we compute $c_{1,k,h}$. In Equation (32) in Azar et al. (2017), $c_{1,k,h}$ corresponds to the confidence interval of

$$(\hat{P}_h^\pi - P_h^\pi)V_{h+1}^*(s_h^k) = \sum_{s'\in\mathcal{S}} \left[\hat{P}(s' \mid s_h^k, a_h^k) - P_h(s' \mid s_h^k, a_h^k) \right] V_{h+1}^*(s').$$

Equation (9) in Azar et al. (2017) computes a confidence interval for this term using the Bernstein inequality. Instead, we use the Hoeffding inequality and obtain

$$[(\hat{P}_h^\pi - P_h^\pi)V_{h+1}^*] \leq H\sqrt{\frac{L}{2N_k(s_h^k, a_h^k)}} = c_{1,k,h}. \tag{59}$$

Combining Equations 59, 58 and 54, the total regret is bounded as

$$\text{Regret} \leq SAH^2 + e \left((H+2)\sqrt{KHL} + \underbrace{\sum_{k=1}^{K} \sum_{h=1}^{H} (b'_{k,h} + c'_{1,k,h} + c'_{4,k,h})}_{(a)} \right) \tag{60}$$

where $b'_{k,h} = \frac{7HL\mathbb{I}_{k,h}}{\sqrt{N_k(s_h^k, a_h^k)}}$, $c'_{1,k,h} = \frac{H\sqrt{L}\mathbb{I}_{k,h}}{\sqrt{2N_k(s_h^k, a_h^k)}}$ and $c'_{4,k,h} = \frac{4SH^2L\mathbb{I}_{k,h}}{N_k(s_h^k, a_h^k)}$.

We focus on the third and dominant term (a). As $b_{k,h} \geq c_{1,k,h}$, this term can be upper bounded by

$$\begin{aligned}(a) &\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\left[\frac{8HL\mathbb{I}_{k,h}}{\sqrt{N_k(s_h^k, a_h^k)}} + \frac{4SH^2L\mathbb{I}_{k,h}}{N_k(s_h^k, a_h^k)}\right] \quad \text{(since } L > 1\text{)} \\ &= 8HL\underbrace{\sum_{i=1}^{K}\sum_{j=1}^{H}\frac{\mathbb{I}_{k,h}}{\sqrt{N_k(s_h^k, a_h^k)}}}_{(b)} + 4SH^2L\underbrace{\sum_{i=1}^{K}\sum_{j=1}^{H}\frac{\mathbb{I}_{k,h}}{N_k(s_h^k, a_h^k)}}_{(c)}.\end{aligned} \tag{61}$$

We bound (b) and (c) separately.

First, we bound (b). We introduce the following lemma, which is an analogy to Lemma 19 in Jaksch et al. (2010) in the finite-horizon setting.

Lemma 31 *Let $H \geq 1$. For any sequence of numbers $z_1, \ldots, z_n$ with $0 \leq z_k \leq H$, consider the sequence $Z_0, Z_1, \ldots Z_n$ defined as*

$$\begin{aligned}&Z_0 \geq H \\ &Z_k = Z_{k-1} + z_k \qquad \text{for } k \geq 1.\end{aligned}$$

Then, for all $n \geq 1$,

$$\sum_{k=1}^{n}\frac{z_k}{\sqrt{Z_{k-1}}} \leq (\sqrt{2}+1)\sqrt{Z_n}.$$

Using Lemma 31, we can bound (b) by Lemma 32.

Lemma 32 *Denote $v_i(s,a) = \sum_{j=1}^{H}\mathbb{I}(a_{i,j} = a, s_{i,j} = s)$ the number of times the state-action pair (s,a) is visited during episode i, and let $\tau(s,a) = \arg\min_{k\in[K]}\{N_k(s,a) \geq H\}$ be the first episode where the state-action pair (s,a) is visited at least H times. Then,*

$$(b) \leq (\sqrt{2}+1)\sqrt{SAKH}. \tag{62}$$

Proof By definition, $N_i(s,a) = \sum_{k=1}^{i-1} v_k(s,a)$. Regrouping the sum in (b) by (s,a), we have

$$\begin{aligned}
(b) &= \sum_{s,a} \sum_{i=1}^{K} \frac{v_i(s,a)}{\sqrt{N_i(s,a)}} \mathbb{I}\{N_i(s,a) \geq H\} \\
&= \sum_{s,a} \left(\sum_{i=1}^{\tau(s,a)-1} \frac{v_i(s,a)}{\sqrt{N_i(s,a)}} \mathbb{I}\{N_i(s,a) \geq H\} + \sum_{i=\tau(s,a)}^{K} \frac{v_i(s,a)}{\sqrt{N_i(s,a)}} \right) \\
&= \sum_{s,a} \sum_{i=\tau(s,a)}^{K} \frac{v_i(s,a)}{\sqrt{N_i(s,a)}} \\
&\leq \sum_{s,a} (\sqrt{2}+1) \sqrt{N_K(s,a) + v_K(s,a)} \\
&\leq (\sqrt{2}+1)\sqrt{SAKH}.
\end{aligned}$$

where the last two inequalities follow from Lemma 31, the Cauchy-Schwarz inequality and the fact that $\sum_{s,a} N_K(s,a) \leq KH$. ■

In order to bound the term (c) in Equation 61, we use the following lemma, which is a variant of Lemma 31 and was stated in Azar et al. (2017) without proof.

Lemma 33 *Let $H \geq 1$. For any sequence of numbers $z_1, \ldots, z_n$ with $0 \leq z_k \leq H$, consider the sequence $Z_0, Z_1, \ldots Z_n$ defined as*

$$\begin{aligned}
&Z_0 \geq H \\
&Z_k = Z_{k-1} + z_k \qquad \textit{for } k \geq 1.
\end{aligned}$$

Then, for all $n \geq 1$,

$$\sum_{k=1}^{n} \frac{z_k}{Z_{k-1}} \leq \sum_{j=1}^{Z_n - Z_0} \frac{1}{j} \leq \ln(Z_n - Z_0) + 1.$$

Proof The second half follows immediately from existing results for the partial sum of the harmonic series. We prove the first half of the inequality by induction. By definition of the two sequences, $Z_k \geq H \geq 1$ and $z_k \leq H \leq Z_{k-1}$ for all k. At $n = 1$, if $z_1 = 0$ then the inequality trivially holds. If $z_1 > 0$, then $Z_1 - Z_0 = z_1$ and

$$\frac{z_1}{Z_0} \leq \frac{z_1}{H} = \left(\underbrace{\frac{1}{H} + \cdots + \frac{1}{H}}_{z_1 \text{ terms}} \right) \leq 1 + \frac{1}{2} + \cdots + \frac{1}{z_1}$$

since $z_1 \leq H$.

For $n > 1$, by the induction hypothesis, we have

$$\begin{aligned}
\sum_{k=1}^{n} \frac{z_k}{Z_{k-1}} &= \sum_{k=1}^{n-1} \frac{z_k}{Z_{k-1}} + \frac{z_n}{Z_{n-1}} \\
&\leq \left(\sum_{j=1}^{Z_{n-1}-Z_0} \frac{1}{j} \right) + \frac{z_n}{Z_{n-1}} \\
&= \left(\sum_{j=1}^{Z_{n-1}-Z_0} \frac{1}{j} \right) + \left(\underbrace{\frac{1}{Z_{n-1}} + \cdots + \frac{1}{Z_{n-1}}}_{z_n \text{terms}} \right) \\
&\leq \left(\sum_{j=1}^{Z_{n-1}-Z_0} \frac{1}{j} \right) + \left(\frac{1}{Z_{n-1} - Z_0 + 1} + \cdots + \frac{1}{Z_{n-1} - Z_0 + z_n} \right) \\
&= \sum_{j=1}^{Z_n - Z_0} \frac{1}{j},
\end{aligned}$$

where the last inequality follows from $z_n \leq Z_0$. Therefore, the induction hypothesis holds for all $n \geq 1$. ■

Using Lemma 33, the term (c) can be bounded similarly to term (b) as follows:

Lemma 34 *With $v_i(s, a)$ and $\tau(s, a)$ defined in Lemma 32, we have*

$$(c) \leq SAL + SA.$$

Proof We write (c) as

$$\begin{aligned}
(c) &= \sum_{i=1}^{K} \sum_{j=1}^{H} \frac{\mathbb{I}\{N_i(s, a) \geq H\}}{N_i(s_{i,j}, a_{i,j})} \\
&= \sum_{s,a} \sum_{i=1}^{K} \frac{v_i(s, a)}{N_i(s, a)} \mathbb{I}\{N_i(s, a) \geq H\} \\
&\leq \sum_{s,a} \left(\sum_{i=1}^{\tau(s,a)-1} \frac{v_i(s, a)}{N_i(s, a)} \mathbb{I}\{N_i(s, a) \geq H\} + \sum_{i=\tau(s,a)}^{K} \frac{v_i(s, a)}{N_i(s, a)} \right) \\
&= \sum_{s,a} \sum_{i=\tau(s,a)}^{K} \frac{v_i(s, a)}{N_i(s, a)} \\
&\leq \sum_{s,a} \left(\ln \left(N_K(s, a) + v_K(s, a) - N_{\tau(s,a)}(s, a) \right) + 1 \right)
\end{aligned}$$

where the last inequality follows from Lemma 33. Trivially bounding the logarithm term by $\ln(KH)$, we obtain

$$(c) \leq SA\ln(KH) + SA \leq SAL + SA.$$

■

Combining Lemma 32 and Lemma 34, we obtain

$$\begin{aligned}(a) &\leq 8HL((\sqrt{2}+1)\sqrt{SAKH}) + 4SH^2L(SAL+SA) \\ &\leq 20HL\sqrt{SAKH} + 5S^2AH^2L^2.\end{aligned}$$

Substituting this into Equation 60, we obtain

$$\begin{aligned}\text{Regret} &\leq SAH^2 + e(H+2)\sqrt{KHL} + e20HL\sqrt{SAKH} + e5S^2AH^2L^2 \\ &\leq 67HL\sqrt{SAKH} + 15S^2AH^2L^2.\end{aligned}$$

Appendix G. Removing the assumption on the hitting time

GOSPRL (Tarbouriech et al., 2021, Lemma 3) guaranteed that in the undiscounted infinite horizon setting, with $H_0 = O(\frac{DS^2A}{\lambda^2})$, Lemma 14 holds with high probability. Thus, in the episodic finite horizon setting, by setting $H_0 = c\frac{DS^2A}{\lambda^2}$ for some appropriately large constant $c > 0$ and applying GOSPRL in each episode we obtain a tight bound in the dependency of K and λ for communicating MDPs. One difficulty in this approach is both c and D are unknown. One possible way to overcome this is to apply the doubling-trick as following: at the beginning of episode k, we set $H_0 = c_k\frac{S^2A}{\lambda^2}$, where $c_1 = 1$. If the learner successfully visits every state-action pair at least N times after H_0 steps, we set $c_{k+1} = c_k$. Otherwise, $c_{k+1} = 2c_k$. There are at most $\log_2(cD)$ episodes with failed exploration until c_k is large enough so that with high probability, all the subsequent episodes will have successful explorations. Moreover, the horizons of the clustering and learning phases change at most $\log_2(cD)$ times. The full analysis of this approach is not in the scope of this paper and is left to future work.