A. The ZERORMAX algorithm

RMAX is a well-known PAC exploration algorithm (Brafman & Tennenholtz, 2002). Here, we show that a modified version of RMAX, which we call ZERORMAX[4], addresses the reward-free exploration setting. The difference between ZERORMAX and RMAX is that we set the reward in "known" states to 0 instead of the true reward, which explains the name. We briefly describe the algorithm and derive the PAC bound relying heavily on prior arguments. Details about RMAX and its analysis can be found in prior work (Brafman & Tennenholtz, 2002; Kakade, 2003).

Following the reward-free exploration framework proposed in Section 2, the ZERORMAX algorithm first collects samples without knowledge about reward (exploration) and then computes a policy for each configuration of reward function (planning). We define set of known states $\mathcal{K}$ to be

$$\mathcal{K} := \{(s,h) : \forall a \in \mathcal{A}, N_h(s,a) \geq m\}$$

where $N_h(s,a)$ counts how many times s has been visited and a was taken in the h-th step and m is a parameter to be specified later. The set $\mathcal{K}$ contains states that we have visited enough times to estimate the corresponding transition kernel, and is typically referred to as the "known set" in the literature. For (s,h) not in $\mathcal{K}$, we call them "unknown."

Now ZERORMAX explores as follows. In each episode $i \in [N]$, the agent has a known set $\mathcal{K}_i$ and

1. builds an empirical MDP $\hat{\mathcal{M}}_{i,\mathcal{K}_i}$ with parameters

$$\mathbb{P}_h(\cdot|s,a) = \begin{cases} \hat{\mathbb{P}}_{h,i}(\cdot|s,a) & \text{if } (s,h) \in \mathcal{K}_i \\ \mathbb{1}\{s' = s\} & \text{otherwise} \end{cases} \qquad r_h(s,a) = \begin{cases} 0 \text{ if } (s,h) \in \mathcal{K}_i \\ 1 \text{ otherwise} \end{cases} \tag{7}$$

   where $\mathbb{P}_{h,i}$ is the empirical estimation of $\mathbb{P}_h$ in the i-th episode.

2. computes $\pi_i = \pi^\star_{\hat{\mathcal{M}}_{i,\mathcal{K}_i}}$ on $\hat{\mathcal{M}}_{i,\mathcal{K}_i}$ by value iteration.

3. samples a trajectory from the environment following π_i.

4. constructs $\mathcal{K}_{i+1}$ for the next episode

For the planning phase, we first sample an index $i \in [N]$ uniformly and construct the MDP $\hat{\mathcal{M}}_{i,\mathcal{K}_i}$. Then given reward function, we can just perform value iteration on $\hat{\mathcal{M}}_{i,\mathcal{K}_i}$, which gives us a near optimal policy.

A.1. Analysis

A central concept for analyzing the sample complexity of ZERORMAX is the escape probability, which is the probability of visiting the unknown states. Formally,

$$p^\pi_{\mathcal{K}} = \mathbb{P}_{\mathcal{M},\pi}\{\exists (s_h,h) \;\; s.t. (s_h,h) \notin \mathcal{K}\}$$

The above definition also depends on the corresponding MDP $\mathcal{M}$. Since we only care about the escape probability w.r.t the true MDP $\mathcal{M}$, we will omit this dependence. The key observation is that there cannot be too many episodes where the escape probability is large. The inuition is that, if the escape probability is big, then the agent will soon visit an unknown states. However, the agent can visit unknown states at most mSA times in total.

Lemma A.1 (Lemma 8.5.2 in (Kakade, 2003))**.** *Let π_i be the policy followed in the i^{th}episode and $\mathcal{K}_i$ be corresponding set of known states. Then with probability $1-p$, there can be at most $\mathcal{O}\left(\frac{mSA}{\varepsilon}\log\frac{SANH}{p}\right)$ episodes where $p^{\pi_i}_{\mathcal{K}_i} > \varepsilon$.*

As a result, we have the following corollary.

Corollary A.2. *If we sample i uniformly from 1 to K, then with probability $1-p-\mathcal{O}\left(\frac{mSA}{\varepsilon N}\log\frac{SANH}{p}\right)$, we have $p^{\pi_i}_{\mathcal{K}_i} \leq \varepsilon$.*

In what follows, we focus on a single "good" episode i where $p^{\pi_i}_{\mathcal{K}_i} \leq \varepsilon$. Since we focus on a single episode, let us denote $\mathcal{K}_i$ by $\mathcal{K}$ and π_i by $\pi^\star_{\hat{\mathcal{M}}_{\mathcal{K}}}$. There are three MDPs of interest, with important details presented in Table 1.

[4] ZERORMAX is basically the exploration part of E^3 algorithm (Kearns & Singh, 2002)

	$\mathcal{M}$	$\mathcal{M}_{\mathcal{K}}$	$\hat{\mathcal{M}}_{\mathcal{K}}$
Known ($\mathcal{K}$)	$=\mathcal{M}$	$=\mathcal{M}$	$\approx\mathcal{M}$
Unknown	$=\mathcal{M}$	self loop	self loop

Table 1: A comparison between the three MDPs involved taken from (Jiang, 2019).

$\mathcal{M}$ is the true MDP of interest, that we will use to measure the performance of the policy we find in the planning phase. $\hat{\mathcal{M}}_{\mathcal{K}}$ is the MDP we use for computing policies in both exploration and planning phases. The final MDP, $\mathcal{M}_{\mathcal{K}}$ is an intermediate MDP which agrees with $\mathcal{M}$ on the known set but follows self-loops in the unknown states. Our plan is to prove with high probability, the value of any policyπ on $\mathcal{M}$ and $\hat{\mathcal{M}}_{\mathcal{K}}$ are close, which implies the desired sample complexity result using the same argument as in Theorem 3.5.

The first step is to prove that for any policy π, the values on $\mathcal{M}_{\mathcal{K}}$ and $\hat{\mathcal{M}}_{\mathcal{K}}$ are similar.

Lemma A.3. *With probability $1-p$, for any policy π and reward function r,*

$$\left|\mathbb{E}_{s_1\sim\mathbb{P}_1}[V^{\pi}_{1,\hat{\mathcal{M}}_{\mathcal{K}}}(s_1;r)-V^{\pi}_{1,\mathcal{M}_{\mathcal{K}}}(s_1;r)]\right|\leq\mathcal{O}\left(H^2\sqrt{\frac{S}{m}\log\frac{SANH}{p}}\right).$$

Proof. We apply Lemma C.1 to $\mathcal{M}_{\mathcal{K}}$ and $\hat{\mathcal{M}}_{\mathcal{K}}$, since the reward function is the same and the transition kernel is the same for unknown states,

$$\begin{aligned}\left|\mathbb{E}_{s_1\sim\mathbb{P}_1}[V^{\pi}_{1,\hat{\mathcal{M}}_{\mathcal{K}}}(s_1;r)-V^{\pi}_{1,\mathcal{M}_{\mathcal{K}}}(s_1;r)]\right|&\leq\mathbb{E}_{M_{\mathcal{K}},\pi}\left\{\sum_{h=1}^{H}\mathbb{1}\left\{(s_h,h)\in\mathcal{K}\right\}\left|\left(\mathbb{P}_h-\hat{\mathbb{P}}_h\right)V^{\pi}_{h+1,\hat{\mathcal{M}}_{\mathcal{K}}}(s_h,a_h)\right|\right\}\\&\leq\mathcal{O}\left(H^2\sqrt{\frac{S}{m}\log\frac{SANH}{p}}\right).\end{aligned}$$

□

The second step is to prove that for any policy π, the values on $\mathcal{M}_{\mathcal{K}}$ and $\mathcal{M}$ are similar, which is less straightforward.

Lemma A.4. *With probability $1-p$ and i is a "good" episode, for any policy π,*

$$\left|\mathbb{E}_{s_1\sim\mathbb{P}_1}[V^{\pi}_{1,\hat{\mathcal{M}}_{\mathcal{K}}}(s_1;r)-V^{\pi}_{1,\mathcal{M}_{\mathcal{K}}}(s_1;r)]\right|\leq H^3\varepsilon+\mathcal{O}\left(H^4\sqrt{\frac{S}{m}\log\frac{SANH}{p}}\right).$$

Proof. Notice that for any policy π, if we can upper bound the escape probability, then $\mathcal{M}_{\mathcal{K}}$ and $\mathcal{M}$ must be similar for this policy. Fortunately, this is actually the case, due to our setting of the reward function in the exploration phase, following (7). Then by definition for any s,

$$\mathbb{E}_{s_1\sim\mathbb{P}_1}V^{\pi}_{\mathcal{M}_{\mathcal{K}}}(s_1)\geq p^{\pi}_{\mathcal{K}},\quad\text{and}\quad Hp^{\pi}_{\mathcal{K}}\geq\mathbb{E}_{s_1\sim\mathbb{P}_1}V^{\pi}_{\mathcal{M}_{\mathcal{K}}}(s_1).$$

and using Lemma A.3,

$$\mathbb{E}_{s_1\sim\mathbb{P}_1}V^{\pi}_{\hat{\mathcal{M}}_{\mathcal{K}}}(s_1)\geq p^{\pi}_{\mathcal{K}}-\mathcal{O}\left(H^2\sqrt{\frac{S}{m}\log\frac{SANH}{p}}\right)$$

However, since we are considering a good episode, we know that for the optimal policy on $\hat{\mathcal{M}}_{\mathcal{K}}$, $\pi^*_{\hat{\mathcal{M}}_{\mathcal{K}}}$, we have $p_{\mathcal{K}}^{\pi^*_{\hat{\mathcal{M}}_{\mathcal{K}}}}\leq\varepsilon$. Therefore,

$$\begin{aligned}&H\varepsilon+\mathcal{O}\left(H^2\sqrt{\frac{S}{m}\log\frac{SANH}{p}}\right)\geq Hp_{\mathcal{K}}^{\pi^*_{\hat{\mathcal{M}}_{\mathcal{K}}}}+\mathcal{O}\left(H^2\sqrt{\frac{S}{m}\log\frac{SANH}{p}}\right)\\\geq&\mathbb{E}_{s_1\sim\mathbb{P}_1}V_{\mathcal{M}_{\mathcal{K}}}^{\pi^*_{\hat{\mathcal{M}}_{\mathcal{K}}}}(s_1)+\mathcal{O}\left(H^2\sqrt{\frac{S}{m}\log\frac{SANH}{p}}\right)\geq\mathbb{E}_{s_1\sim\mathbb{P}_1}V_{\hat{\mathcal{M}}_{\mathcal{K}}}^{\pi^*_{\hat{M}_{\mathcal{K}}}}(s_1)\geq\mathbb{E}_{s_1\sim\mathbb{P}_1}V^{\pi}_{\hat{\mathcal{M}}_{\mathcal{K}}}(s_1)\end{aligned}$$

$$\geq p_{\mathcal{K}}^{\pi} - \mathcal{O}\left(H^2\sqrt{\frac{S}{m}\log\frac{SANH}{p}}\right)$$

and as a result

$$p_{\mathcal{K}}^{\pi} \leq H\varepsilon + \mathcal{O}\left(H^2\sqrt{\frac{S}{m}\log\frac{SANH}{p}}\right).$$

Now notice $\mathcal{M}_{\mathcal{K}}$ and $\mathcal{M}$ are only different on unknown states, which will not influence the agent unless the agent escapes from $\mathcal{K}$. Using Lemma C.1 on $\mathcal{M}_{\mathcal{K}}$ and $\mathcal{M}$ we have

$$\left|\mathbb{E}_{s_1\sim\mathbb{P}_1}[V_{1,\hat{\mathcal{M}}_{\mathcal{K}}}^{\pi}(s_1;r) - V_{1,\mathcal{M}_{\mathcal{K}}}^{\pi}(s_1;r)]\right| \leq H^3\varepsilon + \mathcal{O}\left(H^4\sqrt{\frac{S}{m}\log\frac{SANH}{p}}\right). \qquad \square$$

Finally we can put everything together. Again following the argument in Theorem 3.5, we have

Theorem A.5. *With probability* $1 - 2p - \mathcal{O}\left(\frac{mSA}{\varepsilon K}\log\frac{SANH}{p}\right)$*, given any reward function, the* ZERORMAX *algorithm can output a policy* π *such that*

$$\mathbb{E}_{s_1\sim\mathbb{P}_1}[V_{1,\mathcal{M}}^{*}(s_1) - V_{1,\mathcal{M}}^{\pi}(s_1)] \leq H^3\varepsilon + \mathcal{O}\left(H^4\sqrt{\frac{S}{m}\log\frac{SANH}{p}}\right).$$

Now we can set the parameters m and ε. To make $\mathbb{E}_{s_1\sim\mathbb{P}_1}[V_{1,\mathcal{M}}^{*}(s_1) - V_{1,\mathcal{M}}^{\pi}(s_1)] \leq \epsilon$, we need $m \geq \Omega\left(\frac{SH^8}{\epsilon^2}\log\frac{SAKH}{p}\right)$ and $\varepsilon \leq \mathcal{O}\left(\epsilon/H^3\right)$. This means we must set

$$N \geq \Omega\left(\frac{H^{11}S^2A}{\epsilon^3 p}\left(\log\frac{SANH}{p}\right)^2\right)$$

or equivalently,

$$N \geq \Omega\left(\frac{H^{11}S^2A}{\epsilon^3 p}\left(\log\frac{SAH}{p\epsilon}\right)^2\right)$$

This sample complexity is quite poor because it scales with ϵ^{-3} and polynomially, rather than logarithmically, with $1/p$.

B. MaxEnt Exploration

Another approach for reward-free exploration was studied in (Hazan et al., 2019). They consider the infinite horizon discounted setting with discount factor γ, and they show that with $\tilde{O}(\frac{S^2A}{\varepsilon^3(1-\gamma)^2})$ trajectories of length $\tilde{O}(\frac{\log S}{\varepsilon^{-1}\log(1/\gamma)})$, they can find a policy $\hat{\pi}$ such that

$$\frac{1}{S}\sum_{s}\log(d_{\hat{\pi}}(s)) \geq \max_{\pi}\frac{1}{S}\sum_{s}\log(d_{\pi}(s)) - \varepsilon$$

where $d_{\pi}(s) = (1-\gamma)\sum_{t=1}^{\infty}\gamma^t d_{t,\pi}(s)$ and $d_{t,\pi}(s) = \mathbb{P}[s_t = s \mid \pi]$. This claim is their Corollary 4.6, which uses a smoothing argument to address the fact that the objective function as stated is not defined everywhere.

For reward free exploration, we want to use this guarantee to establish a condition similar to the conclusion of Theorem 3.3. For the sake of contradiction, suppose there exists some policy $\tilde{\pi}$ and some state $\tilde{s}$ such that

$$\frac{d_{\tilde{\pi}}(\tilde{s})}{d_{\hat{\pi}}(\tilde{s})} > 4S.$$

We want to show that the non-Markovian mixture policy $(1-\alpha)\hat{\pi} + \alpha\tilde{\pi}$ for some $\alpha > 0$ demonstrates that $\hat{\pi}$ violates its near-optimality guarantee for the optimization problem. To do this, we lower bound the difference in objective values between the mixture policy and $\hat{\pi}$:

$$
\begin{aligned}
&\frac{1}{S}\sum_s \log((1-\alpha)d_{\hat{\pi}}(x) + \alpha d_{\tilde{\pi}}(s)) - \log(d_{\hat{\pi}}(s)) = \frac{1}{S}\sum_s \log\left(1 - \alpha\frac{d_{\hat{\pi}(s)} - d_{\tilde{\pi}}(s)}{d_{\hat{\pi}}(s)}\right)\\
&\geq \frac{S-1}{S}\log(1-\alpha) + \frac{1}{S}\log(1+\alpha(4S-1))\\
&\geq \frac{S-1}{S}\frac{-\alpha}{1-\alpha} + \frac{1}{S}\frac{\alpha(4S-1)}{1+\alpha(4S-1)}\\
&= \frac{\alpha}{S}\left(\frac{4S}{1+\alpha(4S-1)} - \frac{1}{1+\alpha(4S-1)} - \frac{(S-1)}{1-\alpha}\right).
\end{aligned}
$$

Here we are using that $\log(1-x_1+x_2)$ is monotonically increasing in x_2 so we use the lower bound of $4S$ on $\tilde{s}$ and the trivial lower bound of 0 on all of the other states. We also use that $\log(1+x) \geq \frac{x}{1+x}$, which holds for any $x > -1$. The expression inside the parenthesis can be simplified to

$$
\frac{3S + S\alpha - 4S^2\alpha}{(1-\alpha)(1+\alpha(4S-1))}.
$$

At this point we can see that if $\alpha \geq 1/S$ then this expression is negative, so the mixture policy with large α does not yield any improvement in objective. On the other hand, for any $\alpha < 1/S$ then this inner expression is $\Theta(S)$. So if we set $\alpha = \Theta(1/S)$ the overall improvement in objective is $\Omega(1/S)$. This means that if we want establish the guarantee in Theorem 3.3, we must set $\varepsilon = 1/S$, at which point the overall sample complexity scales with S^5, which is quite poor.

Note that this calculation shows that $O(S^5)$ samples is sufficient for the maximum entropy approach to find a suitable exploratory policy, but we do not claim that it is necessary for this method. A sharper analysis may be possible, but we are not aware of any such results.

C. Proof for Main Results

In this section, we present proofs for results in Section 3.

C.1. Exploration Phase

We begin with the proof of Lemma 3.4, which is a simple modification of the Theorem 1 in (Zanette & Brunskill, 2019).

Proof of Lemma 3.4. WLOG, we can assume s_1 is fixed. This is because for s_1 stochastic from $\mathbb{P}_1$, we can simply add an artificial step before the first step of MDP, which always starts from the same state s_0, has only one action, and the transition to s_1 satisfies $\mathbb{P}_1$. This creates a new MDP with fixed initial state with length $H+1$, which is equivalent to the original MDP.

We use an alternative upper-bound for equation (156) in (Zanette & Brunskill, 2019), which gives:

$$
\begin{aligned}
&\frac{1}{N_0 H}\sum_{k=1}^{N_0}\mathbb{E}_{\pi_k}\left[(\sum_{h=1}^{H} r(s_h,a_h) - V_1^{\pi_k}(s_1))^2 \middle| s_1\right]\\
\leq &\frac{2}{N_0 H}\sum_{k=1}^{N_0}\mathbb{E}_{\pi_k}\left[(\sum_{h=1}^{H} r(s_h,a_h))^2 + (V_1^{\pi_k}(s_1))^2 \middle| s_1\right]\\
\stackrel{(i)}{\leq} &\frac{2}{N_0 H}\sum_{k=1}^{N_0}\mathbb{E}_{\pi_k}\left[\sum_{h=1}^{H} r(s_h,a_h) + V_1^{\pi_k}(s_1) \middle| s_1\right]\\
\leq &\frac{4}{N_0 H}\sum_{k=1}^{N_0} V_1^{\pi_k}(s_1) \leq \frac{4}{H}V_1^{\star}(s_1)
\end{aligned}
$$

where π_k is the policy used in EULER in the k-th episode. Step (i) is because using the reward function designed in Line 4 in Algorithm 2, we have all reward equal to zero except one state. Therefore, we have $\sum_{h=1}^{H} r(s_h, a_h) \leq 1$ and $V_1^\pi(s_1) \leq 1$. Therefore, we have replace the upper bound $\mathcal{G}^2$ in (156) of (Zanette & Brunskill, 2019) by $4V_1^\star(s_1)$.

This allows us also replace the $\mathcal{G}^2$ in Theorem 1 of (Zanette & Brunskill, 2019) by $4V_1^\star(s_1)$, which gives the regret of algorithm (note (Zanette & Brunskill, 2019) is for stationary MDP, while our paper is for non-stationary MDP, thus S in (Zanette & Brunskill, 2019) need to be replaced by SH in our paper due to state augmentation, which creates new states as (s,h)):

$$\sum_{k=1}^{N_0}[V_1^\star(s_1) - V^{\pi_k}(s_1)] \leq \tilde{\mathcal{O}}(\sqrt{V_1^\star(s_1)SAT} + S^2AH^4)$$

Finally, plug in $T = N_0 H$, we finish the proof. □

Now we can prove the main result in this section.

Proof of Theorem 3.3. In the following we can fix a state (s,h) and consider the corresponding policy given by EULER. Remember in our setting (Line 4 in Algorithm 2),

$$\mathbb{E}_{s_1\sim\mathbb{P}_1} V_1^\star(s_1) = \max_\pi P_h^\pi(s)$$

Therefore the regret guarantee Lemma 3.4 implies

$$\max_\pi P_h^\pi(s) - \frac{1}{N_0}\sum_{\pi\in\Phi^{(s,h)}} P_h^\pi(s) \leq c_0\sqrt{\frac{SAH\iota_0 \cdot \max_\pi P_h^\pi(s)}{N_0}} + \frac{S^2AH^4\iota_0^3}{N_0}$$

for some absolute constant c_0. Therefore, in order to make the following true

$$\max_\pi P_h^\pi(s) - \frac{1}{N_0}\sum_{\pi\in\Phi^{(s,h)}} P_h^\pi(s) \leq \frac{1}{2}\max_\pi P_h^\pi(s)$$

We simply need to choose N_0 large enough so that:

$$\sqrt{\frac{SAH\iota_0 \cdot \max_\pi P_h^\pi(s)}{N_0}} \leq c_1 \cdot \max_\pi P_h^\pi(s)$$
$$\frac{S^2AH^4\iota_0^3}{N_0} \leq c_1 \cdot \max_\pi P_h^\pi(s)$$

for a sufficient small absolute constant c_1. Combining with the fact that for δ-significant (s,h), $\max_\pi P_h^\pi(s) \geq \delta$, we know choosing $N_0 = \mathcal{O}(S^2AH^4\iota_0^3/\delta)$ is sufficient. As a result, we have

$$\max_\pi \frac{P_h^\pi(s)}{\frac{1}{N_0}\sum_{\pi\in\Phi^{(s,h)}} P_h^\pi(s)} \leq 2$$

Since Algorithm 2 sets all policy in $\Phi^{(s,h)}$ to choose action uniformly randomly at (s,h), this implies

$$\max_{\pi,a} \frac{P_h^\pi(s,a)}{\frac{1}{N_0}\sum_{\pi\in\Phi^{(s,h)}} P_h^\pi(s,a)} \leq 2A$$

Finally, we can apply the same argument for all δ-significant (s,h), and let $\Psi = \cup\{\Phi^{(s,h)}\}_{(s,h)}$ which gives:

$$\forall\ \delta\text{-significant } (s,h), \quad \max_{\pi,a} \frac{P_h^\pi(s,a)}{\frac{1}{N_0 SH}\sum_{\pi\in\Psi} P_h^\pi(s,a)} \leq 2SAH.$$

This finishes the proof. □

C.2. Planning Phase

The following lemma (E.15 in (Dann et al., 2017)) will be useful to characterize the difference between $V_h^\pi(s;r)$ and $\hat{V}_h^\pi(s;r)$.

Lemma C.1 (Lemma E.15 in (Dann et al., 2017))**.** *For any two MDPs $\mathcal{M}'$ and $\mathcal{M}''$ with rewards r' and r'' and transition probabilities $\mathbb{P}'$ and $\mathbb{P}''$, the difference in values V', V'' with respect to the same policy π can be written as*

$$V_h'(s) - V_h''(s) = \mathbb{E}_{\mathcal{M}'',\pi}\left[\sum_{i=h}^{H}[r_i'(s_i,a_i) - r_i''(s_i,a_i) + (\mathbb{P}_i' - \mathbb{P}_i'')V_{i+1}'(s_i,a_i)]\middle| s_h = s\right]$$

With this decomposition in mind, we can prove Lemma 3.6.

Proof of Lemma 3.6. In this section, we always use $\mathbb{E}$ to denote the expectation under the true MDP $\mathcal{M}$. Using Lemma C.1 on $\mathcal{M}$ (the true MDP) and $\hat{\mathcal{M}}$ (the empirical version), we have

$$|\mathbb{E}_{s_1\sim\mathbb{P}_1}\{\hat{V}_1^\pi(s_1;r) - V_1^\pi(s_1;r)\}| \le |\mathbb{E}_\pi\sum_{h=1}^{H}(\hat{\mathbb{P}}_h - \mathbb{P}_h)\hat{V}_{h+1}^\pi(s_h,a_h)| \le \mathbb{E}_\pi\sum_{h=1}^{H}|(\hat{\mathbb{P}}_h - \mathbb{P}_h)\hat{V}_{h+1}^\pi(s_h,a_h)|$$

Let $\mathcal{S}_h^\delta := \{s : \max_\pi P_h^\pi(s) \ge \delta\}$ be the set of δ-significant states in the h-th step. We further have:

$$\mathbb{E}_\pi|(\hat{\mathbb{P}}_h - \mathbb{P}_h)\hat{V}_{h+1}^\pi(s_h,a_h)| \le \underbrace{\sum_{a,s\in\mathcal{S}_h^\delta}|(\hat{\mathbb{P}}_h - \mathbb{P}_h)\hat{V}_{h+1}^\pi(s,a)|P_h^\pi(s,a)}_{\xi_h} + \underbrace{\sum_{a,s\notin\mathcal{S}_h^\delta}|(\hat{\mathbb{P}}_h - \mathbb{P}_h)\hat{V}_{h+1}^\pi(s,a)|P_h^\pi(s,a)}_{\zeta_h}$$

By definition of insignificant state, we have:

$$\zeta_h \le H\sum_{a,s\notin\mathcal{S}_h^\delta}P_h^\pi(s,a) = H\sum_{s\notin\mathcal{S}_h^\delta}P_h^\pi(s) \le H\sum_{s\notin\mathcal{S}_h^\delta}\delta \le HS\delta. \tag{8}$$

On the other hand, by Cauchy-Shwartz inequality, we have:

$$\xi_h \le \left[\sum_{a,s\in\mathcal{S}_h^\delta}|(\hat{\mathbb{P}}_h - \mathbb{P}_h)\hat{V}_{h+1}^\pi(s,a)|^2P_h^\pi(s,a)\right]^{\frac{1}{2}} = \left[\sum_{a,s\in\mathcal{S}_h^\delta}|(\hat{\mathbb{P}}_h - \mathbb{P}_h)\hat{V}_{h+1}^\pi(s,a)|^2P_h^\pi(s)\pi_h(a|s)\right]^{\frac{1}{2}}$$

We note since $\hat{V}_{h+1}^\pi$ only depends on π at $h+1,\cdots,H$ steps, it does not depends on π_h. Therefore, we have:

$$\begin{aligned}\sum_{a,s\in\mathcal{S}_h^\delta}|(\hat{\mathbb{P}}_h - \mathbb{P}_h)\hat{V}_{h+1}^\pi(s,a)|^2P_h^\pi(s)\pi_h(a|s) &\le \max_{\pi_h'}\sum_{a,s\in\mathcal{S}_h^\delta}|(\hat{\mathbb{P}}_h - \mathbb{P}_h)\hat{V}_{h+1}^\pi(s,a)|^2P_h^\pi(s)\pi_h'(a|s)\\ &= \max_{\nu:\mathcal{S}\to\mathcal{A}}\sum_{a,s\in\mathcal{S}_h^\delta}|(\hat{\mathbb{P}}_h - \mathbb{P}_h)\hat{V}_{h+1}^\pi(s,a)|^2P_h^\pi(s)\mathbb{1}\{a=\nu(s)\}\end{aligned}$$

where the last step is because the maximization over π_h' achieves at deterministic polices.

Recall that by preconditions, we have 4 holds for $\delta = \epsilon/(2SH^2)$. That is, for any $s\in\mathcal{S}_h^\delta$ we always have

$$\max_{\tilde{\pi}}\frac{P_h^{\tilde{\pi}}(s,a)}{\mu_h(s,a)} \le 2SAH$$

Therefore, for any (s,a) pair, we can design a policy π' so that $\pi_{h'}' = \pi_{h'}$ for all $h' < h$, and $\pi_h'(s) = a$. This will give that

$$P_h^\pi(s) = P_h^{\pi'}(s) = P_h^{\pi'}(s,a) \le 2SAH\mu_h(s,a)$$

which gives:

$$
\begin{aligned}
&\sum_{a,s\in\mathcal{S}_h^\delta} |(\hat{\mathbb{P}}_h-\mathbb{P}_h)\hat{V}_{h+1}^\pi(s,a)|^2 P_h^\pi(s)\mathbb{1}\{a=\nu(s)\}\\
\le&2SAH\sum_{a,s\in\mathcal{S}_h^\delta} |(\hat{\mathbb{P}}_h-\mathbb{P}_h)\hat{V}_{h+1}^\pi(s,a)|^2 \mu_h(s)\mathbb{1}\{a=\nu(s)\}\\
\le&2SAH\sum_{s,a} |(\hat{\mathbb{P}}_h-\mathbb{P}_h)\hat{V}_{h+1}^\pi(s,a)|^2 \mu_h(s)\mathbb{1}\{a=\nu(s)\}\\
=&2SAH\mathbb{E}_{\mu_h}|(\hat{\mathbb{P}}_h-\mathbb{P}_h)\hat{V}_{h+1}^\pi(s,a)|^2 \mathbb{1}\{a=\nu(s)\}
\end{aligned}
$$

By Lemma C.2, we have:

$$
\mathbb{E}_{\mu_h}|(\hat{\mathbb{P}}_h-\mathbb{P}_h)\hat{V}_{h+1}^\pi(s,a)|^2 \mathbb{1}\{a=\nu(s)\}\le\mathcal{O}\left(\frac{H^2S}{N}\log(\frac{AHN}{p})\right)
$$

Therefore, combine all equations above, we have

$$
|\mathbb{E}_{s_1\sim\mathbb{P}_1}\{\hat{V}_1^\pi(s_1;r)-V_1^\pi(s_1;r)\}|\le\mathcal{O}(\sqrt{\frac{H^5S^2A}{N}\log(\frac{AHN}{p})})+H^2S\delta
$$

Recall our choice $\delta=\epsilon/(2SH^2)$ and $N\ge c\frac{H^5S^2A}{\epsilon^2}\log(\frac{SAH}{p\epsilon})$ for sufficiently large absolute constant c, which finishes the proof. □

Lemma C.2. *Suppose $\hat{\mathbb{P}}$ is the empirical transition matrix formed by sampling according to μ distribution for N samples, then with probability at least $1-p$, we have for any $h\in[H]$:*

$$
\max_{G:\mathcal{S}\to[0,H]}\max_{\nu:\mathcal{S}\to\mathcal{A}}\mathbb{E}_{\mu_h}|(\hat{\mathbb{P}}_h-\mathbb{P}_h)G(s,a)|^2\mathbb{1}\{a=\nu(s)\}\le\mathcal{O}\left(\frac{H^2S}{N}\log(\frac{AHN}{p})\right)
$$

Proof. Define random variable

$$
X_i=(\hat{\mathbb{P}}_hG(s_i,a_i)-G(s_i'))^2-(\mathbb{P}_hG(s_i,a_i)-G(s_i'))^2
$$

where $(s_i,a_i,s_i')\sim\mu_h\times\mathbb{P}_h(\cdot|s_i,a_i)$ is the i-th sample in level h we collect.

Also we define

$$
Y_i=X_i\mathbb{1}\{a_i=\nu(s_i)\}.
$$

To simplify the notation, when some property of Y_i holds for any i, we just use the notation Y to describe a generic Y_i.

We first state some properties of the random variables Y_i, which are justified at the end of the proof.

- (Expection) $\mathbb{E}Y=\mathbb{E}_{\mu_h}|(\hat{\mathbb{P}}_h-\mathbb{P}_h)G(s,a)|^2\mathbb{1}\{a=\nu(s)\}$
- (Empirical risk minimization) $\sum_{i=1}^N Y_i\le 0$
- (Self-bounded) Var$\{Y\}\le 4H^2\mathbb{E}Y$

Given these three properties, now we are ready to apply Berstein's inequality to $(\sum_{i=1}^N Y_i)/N$. Since we are taking maximum over ν and $G(s)$ and $\hat{\mathbb{P}}$ is random, we need to cover all the possible values of $\hat{\mathbb{P}}G(s,a)\mathbb{1}\{a=\nu(s)\}$ and $\mathbb{P}G(s,a)\mathbb{1}\{a=\nu(s)\}$ to ε accuracy to make Bernstein's inequality hold. For ν, there are A^S deterministic policies in total. Given a fixed ν, $\hat{\mathbb{P}}G(s,a)\mathbb{1}\{a=\nu(s)\}$ and $\mathbb{P}G(s,a)\mathbb{1}\{a=\nu(s)\}$ can be covered by $(H/\varepsilon)^{2S}$ values by boundedness condition because for $a\neq\nu(s)$ they are always 0. The overall approximation error will be at most $12H\varepsilon$ by boundedness condition.

As a result, with probability at least $1-p/H$, for any ν, $G(s)$ and $\hat{\mathbb{P}}$,

$$\begin{aligned}
&\mathbb{E}_{\mu_h}|(\hat{\mathbb{P}}_h-\mathbb{P}_h)G(s,a)|^2\mathbb{1}\{a=\nu(s)\}=\mathbb{E}Y\le\mathbb{E}Y-\frac{1}{N}\sum_{i=1}^N Y_i\\
\le&\sqrt{\frac{2\text{Var}\{Y\}\log((\frac{H}{\varepsilon})^{2S}\cdot A^S\cdot\frac{H}{p})}{N}}+\frac{H^2\log((\frac{H}{\varepsilon})^{2S}\cdot A^S\cdot\frac{H}{p})}{3N}+12H\varepsilon\\
\le&\sqrt{\frac{2\text{Var}\{Y\}[2S\log(\frac{HA}{\varepsilon})+\log\frac{H}{p}]}{N}}+\frac{H^2[2S\log(\frac{HA}{\varepsilon})+\log\frac{H}{p}]}{3N}+12H\varepsilon
\end{aligned}$$

We can simply choose $\varepsilon=HS/36N$ and thus

$$\begin{aligned}
&\mathbb{E}_{\mu_h}|(\hat{\mathbb{P}}_h-\mathbb{P}_h)G(s,a)|^2\mathbb{1}\{a=\nu(s)\}\\
\le&\sqrt{8H^2\mathbb{E}_{\mu_h}|(\hat{\mathbb{P}}_h-\mathbb{P}_h)G(s,a)|^2\mathbb{1}\{a=\nu(s)\}\frac{2S\log(\frac{36AN}{S})+\log\frac{H}{p}}{N}}+\frac{H^2[2S\log(\frac{36AN}{S})+\log\frac{H}{p}+S]}{3N}
\end{aligned}$$

Solving this quadratic formula we get

$$\mathbb{E}_{\mu_h}|(\hat{\mathbb{P}}_h-\mathbb{P}_h)G(s,a)|^2\mathbb{1}\{a=\nu(s)\}\le\mathcal{O}(\frac{H^2S}{N}\log(\frac{ANH}{p}))$$

Since the above upper bound holds for arbitrary ν, $G(s)$ and $\mathbb{P}_h$,

$$\max_{G:\mathcal{S}\to[0,H]}\max_{\nu:\mathcal{S}\to\mathcal{A}}\mathbb{E}_{\mu_h}|(\hat{\mathbb{P}}_h-\mathbb{P}_h)G(s,a)|^2\mathbb{1}\{a=\nu(s)\}\le\mathcal{O}\left(\frac{H^2S}{N}\log(\frac{AHN}{p})\right)$$

Taking union bound w.r.t. h, the claim holds for any h with probability $1-p$.

Finally we give the proofs for the claimed three properties of Y_i. We begin with the expectation property:

$$\begin{aligned}
\mathbb{E}Y=&\mathbb{E}_{s,a\sim\mu_h}\mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)}\{\mathbb{1}\{a=\nu(s)\}[(\hat{\mathbb{P}}_hG(s,a)-G(s'))^2-(\mathbb{P}_hG(s,a)-G(s'))^2]\}\\
\overset{(i)}{=}&2\mathbb{E}_{s,a\sim\mu_h}\mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)}\{\mathbb{1}\{a=\nu(s)\}(\hat{\mathbb{P}}_h-\mathbb{P}_h)G(s,a)(\mathbb{P}_hG(s,a)-G(s'))\}\\
&+\mathbb{E}_{\mu_h}|(\hat{\mathbb{P}}_h-\mathbb{P}_h)G(s,a)|^2\mathbb{1}\{a=\nu(s)\}\\
\overset{(ii)}{=}&\mathbb{E}_{\mu_h}|(\hat{\mathbb{P}}_h-\mathbb{P}_h)G(s,a)|^2\mathbb{1}\{a=\nu(s)\}
\end{aligned}$$

where (i) is by $b^2-d^2=(b-d+d)^2-d^2=(b-d)^2+2b(d-b)$ with $b=\hat{\mathbb{P}}_hG(s,a)-G(s')$ and $d=\mathbb{P}_hG(s,a)-G(s')$ and (ii) is because $\mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)}\{G(s')\}=\mathbb{P}_hG(s,a)$.

The emipirical risk minimization property is true because the evaluation rule is essentially minimizing the empirical Bellman error for each (s,a) pair separately. Mathematically,

$$\hat{\mathbb{P}}_hG(s,a)=\arg\max_g\sum_{i=1}^N\mathbb{1}\{s_i=s,a_i=a\}(g-G(s'))^2$$

The self-bounded property is because

$$\begin{aligned}
&\text{Var}\{Y\}\le\mathbb{E}(Y)^2\\
\overset{(i)}{=}&\mathbb{E}\{\mathbb{1}\{a=\nu(s)\}[(\hat{\mathbb{P}}_h-\mathbb{P}_h)G(s,a)]^2[(\hat{\mathbb{P}}_h+\mathbb{P}_h)G(s,a)-2G(s')]^2\}\\
\le&4H^2\mathbb{E}_{\mu_h}|(\hat{\mathbb{P}}_h-\mathbb{P}_h)G(s,a)|^2\mathbb{1}\{a=\nu(s)\}\\
=&4H^2\mathbb{E}Y
\end{aligned}$$

where (i) by $b^2-d^2=(b+d)(b-d)$ with $b=\hat{\mathbb{P}}_hG(s,a)-G(s')$ and $d=\mathbb{P}_hG(s,a)-G(s')$. $\square$

C.3. Proof of Theorem 3.1

Putting everything together we can prove the main theorem.

Proof of Theorem 3.1. We only need to choose the parameter δ and N_0. From the proof of Lemma 3.6 we can see, we need $\delta = \epsilon/(2SH^2)$ and thus $N_0 \geq cS^3AH^6\iota^3/\epsilon$. Since we need N_0 episodes for each (s,h), the total number episodes required for finding Ψ is $\mathcal{O}(cS^4AH^7\iota^3/\epsilon)$, which gives the second term in (3). The proof is completed by combining Theorem 3.5, which gives the first term in (3). $\square$

C.4. Approximate MDP Solvers

The convergence of NPG is well studied in (Agarwal et al., 2019) (tabluar & infinite horizon) and (Cai et al., 2019) (linear approximation). For completeness we give a full proof of convergence rate of NPG algorithm in episodic setting.

Since we only need to prove the guarantee on the true MDP, we will not distinguish true MDP $\mathcal{M}$ and estimated MDP $\hat{\mathcal{M}}$ here. Remember the NPG is defined by

$$\pi_h^{(0)}(a|s) = 1/A$$

and

$$\pi_h^{(t+1)}(a|s) = \pi_h^{(t)}(a|s)\exp\{\eta(Q_h^{(t)}(s,a) - V_h^{(t)}(s))\}/Z_h^{(t)}(s)$$

where $Q_h^{(t)}(s,a) := Q_h^{\pi^{(t)}}(s,a)$ is computed following the value iteration procedure. Similarly we define $V_h^{(t)}(s) := V_h^{\pi^{(t)}}(s)$. The normalization constant can be written explicitly as

$$Z_h^{(t)}(s) := \sum_{a\in\mathcal{A}} \pi_h^{(t)}(a|s)\exp\{\eta[Q_h^{(t)}(s,a) - V_h^{(t)}(s)]\}$$

Notice the definition of the normalization constant is not unique. Here we choose the form that makes the following proof simpler but different choice will essentially gives exactly the same algorithm.

We begin with a lemma showing that the value function monotonically increases.

Lemma C.3 (Lemma 5.8 in (Agarwal et al., 2019))**.** *Following the NPG iterations,*

$$V_h^{(t+1)}(s;r) - V_h^{(t)}(s;r) \geq \frac{1}{\eta}\sum_{h'=h}^{H} \mathbb{E}_{s_{h'}\sim M,\pi^{(t+1)}}\{\log Z_{h'}^{(t)}(s_{h'})|s_h = s\} \geq 0$$

In particular,

$$\log Z_h^{(t)}(s_h) \leq \eta[V_h^{(t+1)}(s_h;r) - V_h^{(t)}(s_h;r)]$$

Proof. By performance difference lemma (Kakade & Langford, 2002),

$$\begin{aligned}
&V_h^{(t+1)}(s;r) - V_h^{(t)}(s;r)\\
=&\sum_{h'=h}^{H} \mathbb{E}_{\pi^{(t+1)}}\{\sum_{a\in\mathcal{A}} \pi_{h'}^{(t+1)}(a|s_{h'})[Q_{h'}^{(t)}(s_{h'},a) - V_{h'}^{(t)}(s_{h'})]|s_h = s\}\\
=&\frac{1}{\eta}\sum_{h'=h}^{H} \mathbb{E}_{\pi^{(t+1)}}\{\sum_{a\in\mathcal{A}} \pi_{h'}^{(t+1)}(a|s_{h'})\log\frac{\pi_{h'}^{(t+1)}(a|s_{h'})Z_{h'}^{(t)}(s_{h'})}{\pi_{h'}^{(t)}(a|s_{h'})}|s_h = s\}\\
=&\frac{1}{\eta}\sum_{h'=h}^{H} \mathbb{E}_{\pi^{(t+1)}}\{\mathrm{KL}(\pi_{h'}^{(t+1)}(s_{h'})||\pi_{h'}^{(t)}(s_{h'})) + \log Z_{h'}^{(t)}(s_{h'})|s_h = s\}\\
\geq&\frac{1}{\eta}\sum_{h'=h}^{H} \mathbb{E}_{\pi^{(t+1)}}\{\log Z_{h'}^{(t)}(s_{h'})|s_h = s\}\\
\overset{(i)}{\geq}&0
\end{aligned}$$

where (i) is by for any h and s,

$$\begin{aligned}\log Z_h^{(t)}(s) =& \log\{\sum_{a\in\mathcal{A}} \pi_h^{(t)}(a|s)\exp\{\eta[Q_h^{(t)}(s,a) - V_h^{(t)}(s)]\}\}\\ \geq& \eta \sum_{a\in\mathcal{A}} \pi_h^{(t)}(a|s)[Q_h^{(t)}(s,a) - V_h^{(t)}(s)]\\ =&0\end{aligned}$$

because $V_h^{(t)}(s) = \sum_{a\in\mathcal{A}} \pi_h^{(t)}(a|s)Q_h^{(t)}(s,a)$ by definition. □

Equipped with the monotone property, we can simply prove an upper bound for the cumulative regret, which immediately implies the convergence rate for the last iteration.

Proof of Proposition 3.7. Again by performance difference lemma,

$$\begin{aligned}&\mathbb{E}_{s_1\sim\mathbb{P}_1}\{V_1^\star(s_1;r) - V_1^{(t)}(s_1;r)\}\\ =&\sum_{h=1}^H \mathbb{E}_{\pi^\star}\{\sum_{a\in\mathcal{A}} \pi_h^\star(a|s)[Q_h^{(t)}(s,a) - V_h^{(t)}(s)]\}\\ =&\frac{1}{\eta}\sum_{h=1}^H \mathbb{E}_{\pi^\star}\{\sum_{a\in\mathcal{A}} \pi_h^\star(a|s_h)\log\frac{\pi_h^{(t+1)}(a|s_h)Z_h^{(t)}(s_h)}{\pi_h^{(t)}(a|s_h)}\}\\ =&\frac{1}{\eta}\sum_{h=1}^H \mathbb{E}_{\pi^\star}\{\mathrm{KL}(\pi_h^\star(s_h)||\pi_h^{(t)}(s_h)) - \mathrm{KL}(\pi_h^\star(s_h)||\pi_h^{(t+1)}(s_h)) + \log Z_h^{(t)}(s_h)\}\end{aligned}$$

Now we can upper bound the regret of $\pi^{(T-1)}$ by upper bound the cumulative regret using Lemma C.3

$$\begin{aligned}&\mathbb{E}_{s_1\sim\mathbb{P}_1}\{V_1^\star(s_1;r) - V_1^{(T-1)}(s_1;r)\}\\ \leq&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{s_1\sim\mathbb{P}_1}\{V_1^\star(s_1;r) - V_1^{(t)}(s_1;r)\}\\ \leq&\frac{1}{\eta T}\sum_{t=0}^{T-1}\sum_{h=1}^H \mathbb{E}_{\pi^\star}\{\mathrm{KL}(\pi_h^\star(s_h)||\pi_h^{(t)}(s_h)) - \mathrm{KL}(\pi_h^\star(s_h)||\pi_h^{(t+1)}(s_h)) + \log Z_h^{(t)}(s_h)\}\\ \leq&\frac{1}{\eta T}\sum_{h=1}^H \mathbb{E}_{\pi^\star}\{\mathrm{KL}(\pi_h^\star(s_h)||\pi_h^{(0)}(s_h))\} + \frac{1}{\eta T}\sum_{t=0}^{T-1}\sum_{h=1}^H \mathbb{E}_{\pi^\star}\{\log Z_h^{(t)}(s_h)\}\\ \stackrel{(i)}{\leq}&\frac{H\log A}{\eta T} + \frac{1}{T}\sum_{h=1}^H\sum_{t=0}^{T-1}[V_h^{(t+1)}(s_h;r) - V_h^{(t)}(s_h;r)]\\ \leq&\frac{H\log A}{\eta T} + \frac{1}{T}\sum_{h=1}^H V_h^{(T)}(s_h;r)\\ \leq&\frac{H\log A}{\eta T} + \frac{H^2}{T}\end{aligned}$$

where (i) is by using Lemma C.3. □

D. Proof of Lower Bound

In this section, we prove our lower bound, Theorem 4.1. First, we develop further notation in Section D.1 which will aid in distinguishing between multiple possible instances. Next, Section D.2 states Lemma D.2, the formal analogue of Lemma 4.2, which describes a lower bound for learning transitions at a single state. Then, Section D.3 embeds the construction to obtain an instance where the learner to learn transitions at n states, yielding the lower bound Theorem 4.1. Finally, Section D.4 details the proof of the 1-state lower bound, Lemma 4.2.

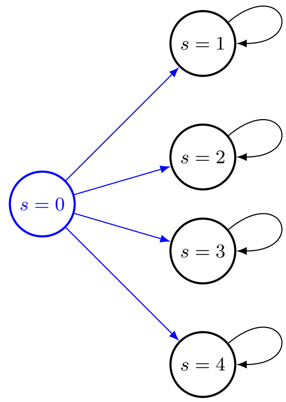

Figure 3: The agent begins in stage $s = 0$, and moves to states $s \in [2n]$, $n = 2$. Different actions correspond to different probability distributions over next states $s \in [2n]$. States $s \in [2n]$ are absorbing, and rewards are action-independent. Lemma 4.2 shows that this construction requires the learner to learn $\Omega(n)$ bits about the transition probabilities $p(\cdot \mid 0, a)$.

D.1. Preliminaries

Environments, Transition Classes, Reward Classes To formalize our embedding a one-state instance into a larger MDP, the following formalities are helpful: we define an environment $\mathscr{E} = (\mathcal{X}, A, H)$ as a triple specifying a finite state space $\mathcal{X}$, number of actions A, and horizon H. For a fixed environment, a transition class $\mathscr{P}$ is a class of transition and initital state distributions, denoted by $\mathbb{P}$; a reward class $\mathscr{R}$ is a family of reward functions $r : (\mathcal{X}, A) \to [0, 1]$. Given a reward vector r and transition vector $\mathbb{P}$, we let $\mathsf{mdp}(\mathbb{P}, r)$ denote the with-reward MDP induced by $\mathbb{P}$ and r. We denote value of a policy π on $\mathsf{mdp}(\mathbb{P}, r)$ by $V^{\pi}(\mathbb{P}, r)$.

Reward-Free MDP Algorithm A reward-free MDP algorithm Alg is algorithm which collects a random number K trajectories from a given reward-free MDP, and then, when given a sequence of reward vectors $r^{(1)}, r^{(2)}, \dots, r^{(N)}$, returns a sequence of policies $\pi^{(1)}, \pi^{(2)}, \dots, \pi^{(N)}$. We let $\mathbf{E}_{\mathbb{P},\mathsf{Alg}}[\cdot]$ denote the expectation under the joint law prescribed by the explortion phase of algorithm Alg and transition operator $\mathbb{P}$.

Correctness Given $\epsilon, p \in (0, 1)$, say that a reward-free MDP algorithm $(\epsilon, p,)$-learns a a problem class $\mathscr{M} := (\mathscr{E}, \mathscr{R}, \mathscr{P})$ if, for any transition operator $\mathbb{P} \in \mathscr{P}$, for any finite sequence of reward vectors $r^{(1)}, \dots, r^{(N)} \in \mathscr{R}$, Alg returns a sequence policies $\pi^{(1)}, \dots, \pi^{(N)}$, such that, with probability $1 - p$, the following holds

$$V^{\pi^{(i)}}(\mathbb{P}, r^{(i)}) \geq \max_{\pi} V^{\pi}(\mathbb{P}, r^{(i)}) - \epsilon, \quad \forall i \in [N].$$

For the lower bound, we allow the policies π prescribed by Alg to be arbitrary randomized mappings form observed histories, that is, Alg selects a random seed ξ from some distribution; that is the policy at stage h is a map

$$\pi_h : (s_1, \dots, s_h, a_1, \dots, a_{h-1}, \xi) \to [A].$$

D.2. Learning A Single Instance

In this section, we define a triple $(\mathscr{E}, \mathscr{R}, \mathscr{P})$ on $\mathcal{O}(n)$-states which forces the learner to spend $\Omega(nA/\epsilon^2)$ trajectories to learn the transition probabilities at a given state.

As described in Figure 3, the hard instances consist of reward-free MDPs that begin in a fixed initial state, and transition to one of $2n$ terminal states according to an unknown transition distribution. The transitions are all taken to be $\epsilon/2n$-close to uniform in the ℓ_∞ norm, which helps with the embedding later on. For simplicitiy, the rewards are taken to depend only on states but not on actions. We formalize these instances in the following definition:

Definition D.1 (Hard Transitions and Rewards at Single State)**.** For parameters $n, A \geq 1$ and A, we define the problem class $\mathscr{M}_{\text{single}}(\epsilon; n, A) : (\mathscr{E}_{\text{single}}(n), \mathscr{P}_{\text{single}}(\epsilon; n, A), \mathscr{R}_{\text{single}}(n, A))$ as the triple with the following consitutents:

1. The environment $\mathscr{E}_{\mathrm{single}}(n)$ is

$$\mathscr{E}_{\mathrm{single}}(n, A) = (\mathcal{X}_{\mathrm{single}}(n), A, 2), \quad \text{where } \mathcal{X}_{\mathrm{single}}(n) := \{0, 1, \dots, 2n\}$$

2. For a given $\epsilon \in (0, 1)$, we define the transition class $\mathscr{P}_{\mathrm{single}}(\epsilon; n, A)$ as the set of transition operator on $\mathscr{E}_{\mathrm{single}}(n, A)$, parameterized by vectors q, which begin at state $x_1 = 0$, and always transition to a state $x_2 \in \{1, \dots, 2n\}$ with near-uniform probability, and remain at that state for the remainder of the episode. Formally,

$$\mathscr{P}_{\mathrm{single}}(\epsilon; n, A) := \Big\{ \mathbb{P}[x_1 = 0] = 1, |\mathbb{P}[x' = s \mid x = 0, a] - \tfrac{1}{2n}| \le \frac{1}{2n}\epsilon \\ \mathbb{P}[x' = s \mid x = s, a] = 1 \ \forall a \in [A], s \in [2n], \Big\}.$$

3. We define the hard reward class $\mathscr{R}_{\mathrm{single}}(n, A)$ as the set of rewards which as the set of rewards which assign 0 reward to state 0, and an action-independent reward to each state $s \in [2n]$. Formally, we define $\mathscr{R}_{\mathrm{single}}(n, A) := \left\{ r_\nu : r_\nu(0, \cdot) = 0,\, r_\nu(x, \cdot) = \nu[x], \quad \nu \in [0, 1]^{2n} \right\}$.

Lemma D.2 (Formal Statement of Lemma 4.2)**.** *Fix $\epsilon \le 1$, $p \le 1/2$, $A \ge 2$, and suppose that $n \ge c_0 \log_2 A$ for universal constants c_0. Then, there exists a distribution $\mathcal{D}$ over transition vectors $\mathbb{P} \in \mathscr{P}_{\mathrm{single}}(\epsilon; n, A)$ such that any algorithm which $(\epsilon/12, p)$-learns the class $\mathscr{M}_{\mathrm{single}}(\epsilon; n, A)$ satisfies*

$$\mathbf{E}_{\mathbb{P} \sim \mathcal{D}}\, \mathbf{E}_{\mathbb{P}, \mathsf{Alg}}[K] \gtrsim \frac{nA}{\epsilon^2}\,.$$

Due to its level of technical, the proof of Lemma D.2 is given in Section D.4.

D.3. Learning Transitions at n states: Proof of Theorem 4.1

Let $n \ge 2$ be a power of two, which we ultimately will choose to be $\Omega(S)$. This means that $\ell_0 := \log_2 n \in \mathbb{N}$ is integral, and define the layered state space:

$$\mathcal{X} := \left\{ (x, \ell) : x \in [2^\ell],\, \ell \in \{0, 1, \dots, \ell_0 + 1\} \right\}$$

The cardinality of the state space is bounded as $|\mathcal{X}| \le 1 + 2 + \dots + n/2 + n + 2n \le 4n$. Hence, we shall chose n to be the largest power of two such that $4n \le S$. Note then that $n = \Omega(S)$ as long as $S \ge C$ for a universal constant C. We will establish our lower bound for the environment $\mathscr{E}_{\mathrm{embed}} = (\mathcal{X}, A, H)$, that is, with state space $\mathcal{X}$; the lower bound extends to an MDP wiht desired state space of size S by augmenting the MDP with isolated, univistable states.

Description of Transition Class Let us define the class $\mathscr{P}_{\mathrm{embd}}$. First, we require that the states (x, ℓ) for $\ell \in [\ell_0]$ form a dyadic tree, whose transitions are all known to the learner. That is, for $\mathbb{P} \in \mathscr{P}_{\mathrm{embd}}$,

$$\begin{aligned} &\mathbb{P}[s_1 = (0, 1)] = 1 \\ &\mathbb{P}[s' = (x, \ell + 1) \mid s = (x, \ell), a = 1] = 1, \quad \ell \in \{0, 1 \dots, \ell_0 - 1\} \\ &\mathbb{P}[x' = (2^\ell + x, \ell - 1) \mid s = (x, \ell), a] = 1, \quad \ell \in \{0, 1, \dots, \ell_0 - 1\},\, a > 1. \end{aligned}$$

In words, $\mathbb{P}$ starts at $(1, 1)$, moves leftward with action $a = 1$, and rightward with actions $a > 1$. At each state $s = (x, \ell_0)$, the learn learner faces transitions described by some $\mathbb{P}^{(x)}_{\mathrm{single}} \in \mathscr{P}_{\mathrm{single}}(\epsilon_0)$ for $\epsilon_0 = 1/8H$: specifically, we stipulate that states (x, ℓ_0) always transition to states $(x', \ell_0 + 1)$, which are absorbing:

$$\begin{aligned} &\forall P \in \mathscr{P}_{\mathrm{embd}}, x \in [n], \text{ there exists a } \mathbb{P}^{(x)}_{\mathrm{single}} \in \mathscr{P}_{\mathrm{single}}(\epsilon_0) \text{ such that } : \\ &\mathbb{P}[s' = (x', \ell_0 + 1) \mid s = (x, \ell_0), a] = \mathbb{P}^{(x)}_{\mathrm{single}}[s' = x' \mid s = 0, a],\ \forall a \in [A],\, x' \in [2n]. \\ &\mathbb{P}[s' = (x', \ell_0 + 1) \mid s = (x', \ell_0 + 1), a] = 1,\ \forall a \in [A] \end{aligned}$$

Thus, there is a bijection between instances $\mathbb{P} \in \mathscr{P}_{\mathrm{embd}}$ and tuples $(\mathbb{P}^{(1)}_{\mathrm{single}}, \dots, \mathbb{P}^{(n)}_{\mathrm{single}}) \in \mathscr{P}^n_{\mathrm{single}}$.

Description of Reward Class Define the reward class $\mathscr{R}_{\text{embed}} = \{r_{x,\nu}\}$ considering for action-independent rewards

$$r_{x,\nu}(s,a) = \begin{cases} 0 & s = (x', \ell),\ \ell < \ell_0, \\ 0 & s = (x', \ell_0) \text{ and } x' \neq x \\ 1 & s = (x, \ell_0) \\ r_\nu[x'] & s = (x', \ell_0 + 1). \end{cases}$$

In other words, the learner recieves reward 1 at state (x, ℓ_0), rewards r_ν at terminal states (x', ℓ_0+1), and 0 elsewhere. We now establish that any policy which is ϵ-optimal under reward $r_{x,\nu}$ must visit $(y, \ell_{\max})$ with sufficiently high probability:

Lemma D.3. *Suppose that a (possibly randomized, non-Markovian) policy π satisfies, for $\epsilon \le 1/4$ and $\epsilon_0 \le 1/8H$,*

$$V^\pi(\mathbb{P}, r_{x,\nu}) \ge \max_{\pi'} V^{\pi'}(\mathbb{P}, r_{x,\nu}) - \epsilon, \quad \forall i \in [N].$$

Then, $\mathbb{P}^\pi[s_{\ell_0+1} = (x, \ell_{\max})] \ge \frac{1}{2}$.

Proof. Due to the structure of the transitions and rewards, the value of any policy π is

$$V^\pi(\mathbb{P}, r_{x,\nu}) = \mathbb{P}^\pi[s_{\ell_0+1} = (x, \ell_0)] + (H - \ell_0 - 1)\sum_{x'=1}^{2n} \nu(x')\mathbb{P}^\pi[s_{\ell_0+2} = (x, \ell_0)]$$

Since the transitions from (x', ℓ_0) to (x'', ℓ_0+1) is $\epsilon_0/2n$-away from uniform in ℓ_∞, we can also see that $\mathbb{P}^\pi[s_{\ell_0+2} = (x, \ell_0)] \in (\frac{1}{2n} - \epsilon, \frac{1}{2n} + \epsilon)$. Thus, letting $\overline{\nu} := \frac{1}{2n}\sum_{x'=1}^{2n} \nu[x']$, we have

$$\left|(H - \ell_0 - 1)\sum_{x'=1}^{2n} \nu(x')\mathbb{P}^\pi[s_{\ell_0+2} = (x, \ell_0)] - (H - \ell_0 - 1)\overline{\nu}\right| \le (H - \ell_0 - 1)\epsilon_0 \le \frac{1}{8}.$$

This entails that

$$|V^\pi(\mathbb{P}, r_{x,\nu}) - (H - \ell_0 - 1)\overline{\nu} - \mathbb{P}^\pi[s_{\ell_0+1} = (x, \ell_0)]| \le \frac{1}{8}.$$

Consequently, by considering a policy π' which always visits state $s_{\ell_0+1} = (x, \ell_0)$ (this can be achieved due to the deterministic behavior of the actions),

$$\max_{\pi'} V^{\pi'}(\mathbb{P}, r_{x,\nu}) - V^\pi(\mathbb{P}, r_{x,\nu}) \ge 1 - \mathbb{P}^\pi[s_{\ell_0+1} = (x, \ell_0)] - 2\cdot\frac{1}{8} = \frac{3}{4} - \mathbb{P}^\pi[s_{\ell_0+1} = (x, \ell_0)].$$

In order for the above to be at most $1/4$, we must have that $\mathbb{P}^\pi[s_{\ell_0+1} = (x, \ell_0)] \ge 1/2$. □

Concluding the Proof of Theorem 4.1 To prove Theorem 4.1, we use the following lemma:

Lemma D.4 (Embedding Correspondence)**.** *Suppose that $H \ge (2\ell_0 + 2)$. Then there exists a correspondence Ψ, which does not dependent on $\mathbb{P} \in \mathscr{P}_{\text{embd}}$ or $r_{y,\nu} \in \mathscr{R}_{\text{embed}}$ (but possibly on ϵ, n, A, H) which operates as follows: Given a policy π for $\mathscr{E}_{\text{embed}}$, $\Psi[\pi] = (\pi^{(1)}, \dots, \pi^{(n)})$ returns an n-tuple of policies for $\mathscr{E}_{\text{single}}(n, A)$ with the following property: For any $\mathbb{P} \equiv (\mathbb{P}^{(1)}_{\text{single}}, \dots, \mathbb{P}^{(n)}_{\text{single}}) \in \mathscr{P}_{\text{embd}}$ and $r_{x,\nu} \in \mathscr{R}_{\text{embed}}$,*

$$\text{If } V^\pi(\mathbb{P}, r_{x,\nu}) \ge \max_{\pi'} V^{\pi'}(\mathbb{P}, r_{x,\nu}) - \epsilon, \quad \forall x \in [n], \quad V^{\pi^{(x)}}(\mathbb{P}^{(x)}_{\text{single}}, r_\nu) \ge \max_{\pi'} V^{\pi'}(\mathbb{P}^{(x)}_{\text{single}}, r_\nu).$$

Proof of Lemma D.4. We directly construct the map Ψ. Observe that policies $\pi^{(x)}$ on the single state environment can be discred by a distribution over which actions $a \in [A]$ they select at the initial state x. Thus identifying policies as elements of $\Delta(A)$, we set

$$\pi^{(x)}[a] := \begin{cases} \mathbb{P}^\pi[a_{\ell_0+1} = a \mid s_{\ell_0+1} = (x, \ell_0)] & \mathbb{P}^\pi[s_{\ell_0+1} = (x, \ell_0)] > 0 \\ \text{arbitrary} & \text{otherwise} \end{cases}$$

as the marginal distribution of actions selected when $s_{\ell_0+1} = (x, \ell_0 + 1)$. Observe that the above conditional probabilites *do not* depend on $\mathbb{P} \in \mathscr{P}_{\mathrm{embd}}$ since the dynamics up to $h = \ell_0 + 1$ are identical for all instances. By considing a policy which coincides with π until $s_{\ell_0+1} = (x, \ell_0)$ and swtiches to playing optimally, we can lower bound the subopitmality of π by

$$\max_{\pi'} V^{\pi'}(\mathbb{P}, r_{x,\nu}) - V^{\pi}(\mathbb{P}, r_{x,\nu}) \geq$$

$$\mathbb{P}^{\pi}[s_{\ell_0+1} = (x, \ell_0)] \cdot (H - \ell_0 - 1) \left(\max_{\pi'} V^{\pi}(\mathbb{P}^{(x)}_{\text{single}}, r_\nu) - V^{\pi^{(x)}}(\mathbb{P}^{(x)}_{\text{single}}, r_\nu) \right)$$

In particular, if π is $\epsilon \leq 1/4$-suboptimal, then Lemma D.3 ensures $\mathbb{P}^{\pi}[s_{\ell_0+1} = (x, \ell_0)] \geq 1/2$. Since $H \geq 2(\ell_0 + 1)$ by assumption, we have

$$\epsilon \geq \max_{\pi'} V^{\mathcal{M},\pi'} - V^{\mathcal{M},\pi} \geq \frac{H}{4} \left(\max_{\pi'} V^{\pi}(\mathbb{P}^{(x)}_{\text{single}}, r_\nu) - V^{\pi^{(x)}}(\mathbb{P}^{(x)}_{\text{single}}, r_\nu) \right),$$

Therefore, $\max_{\pi'} V^{\pi}(\mathbb{P}^{(x)}_{\text{single}}, r_\nu) - V^{\pi^{(x)}}(\mathbb{P}^{(x)}_{\text{single}}, r_\nu) \leq \frac{4\epsilon}{H}$, as needed. □

We now conclude with the proof of our main theorem:

Proof of Theorem 4.1. Let Alg be (ϵ, p)-correct on the class $(\mathscr{E}_{\mathrm{embed}}, \mathscr{P}_{\mathrm{embd}}, \mathscr{R}_{\mathrm{embed}})$. Then, for any $x \in [2n]$, we simulate obtain a $(4\epsilon/H, p)$-correct algorithm for $\mathscr{M}_{\text{single}}(4\epsilon/H; n, A)$ as follows:

1. Exploration: Let $\mathcal{D}$ be the distribution over $\mathbb{P}_{\text{single}} \in \mathscr{P}_{\text{single}}$ from Lemma D.2. Draw a tuple $\mathbb{P}^{\neq x} = (\mathbb{P}^{(x')}_{\text{single}})_{x' \neq x}$ of $n - 1$ distributions i.i.d from $\mathcal{D}$, and let $\mathsf{Alg}^{(x, \mathbb{P}^{\neq x})}_{\text{single}}$ denote the algorithm induced by embeding the instance in $\mathscr{M}_{\text{single}}(4\epsilon/H; n, A)$ at stage x of the embedding construction, running Alg on this embedded instance
2. Planning: When queried given a reward vector $r_\nu \in \mathscr{R}_{\text{single}}$, use Alg to compute a policy π for reward vector $r_{x,\nu} \in \mathscr{R}_{\mathrm{embed}}$, and return the policy $\pi^{(x)}$ dicated by the corresponding ψ.

Since Alg is (ϵ, p)-correct and $\epsilon \leq 1/4$, the correspondence Ψ ensures that for any draw of $\mathbb{P}^{\neq x}$, $\mathsf{Alg}^{(x, \mathbb{P}^{\neq x})}_{\text{single}}$ is $(4\epsilon/H, p)$-correct. Let $K^{(x, \mathbb{P}^{\neq x})}$ denote the random number of episodes collected by $\mathsf{Alg}^{(x, \mathbb{P}^{\neq x})}_{\text{single}}$ in the exploration phase, Thus, if $\epsilon \leq \min\{\frac{1}{4}, \frac{H}{48}\}$, and $n \geq c_0 \log_2 A$ for the appropriate c_0 specified in Lemma D.2, the Lemma D.2 entails

$$\mathbf{E}_{\mathbb{P}_{\text{single}} \sim \mathcal{D}} \mathbf{E}_{\mathbb{P}_{\text{single}}, \mathsf{Alg}^{(x, \mathbb{P}^{\neq x})}_{\text{single}}} [K^{(x, \mathbb{P}^{\neq x})}] \gtrsim \frac{nAH^2}{\epsilon^2}.$$

By taking an expectation over $\mathbb{P}^{\neq x}$, we have

$$\mathbf{E}_{\mathbb{P}^{\neq x} \sim \mathcal{D}^{n-1}, \mathbb{P}_{\text{single}} \sim \mathcal{D}} \mathbf{E}_{\mathbb{P}_{\text{single}}, \mathsf{Alg}^{(x, \mathbb{P}^{\neq x})}_{\text{single}}} [K^{(x, \mathbb{P}^{\neq x})}] \gtrsim \frac{nAH^2}{\epsilon^2}.$$

Note then that, if $N_K(x)$ denotes the number of times that the original Alg visits state (x, ℓ_0), then, by Fubini's theorem and the contruction of $\mathsf{Alg}^{(x, \mathbb{P}^{\neq x})}_{\text{single}}$, the expectation of $N_K(x)$ under probabilities drawn uniform from $\mathcal{D}^n$ is euqal to the expectation of $K^{(x, \mathbb{P}^{\neq x})}$ where $\mathbb{P}^{\neq x}$ is drawm uniformly from $\mathcal{D}^{n-1}$, and then the transition $\mathbb{P}_{\text{single}}$ is selected. Formally,

$$\mathbf{E}_{\mathbb{P}^{\neq x} \sim \mathcal{D}^{n-1}, \mathbb{P}_{\text{single}} \sim \mathcal{D}} \mathbf{E}_{\mathbb{P}_{\text{single}}, \mathsf{Alg}^{(x, \mathbb{P}^{\neq x})}_{\text{single}}} [K^{(x, \mathbb{P}^{\neq x})}] = \mathbf{E}_{\mathbb{P} \equiv (\mathbb{P}^{(1)}_{\text{single}}, \dots, \mathbb{P}^{(n)}_{\text{single}}) \sim \mathcal{D}^n} \mathbf{E}_{\mathbb{P}, \mathsf{Alg}} [K_x]$$

This implies that

$$\mathbf{E}_{\mathbb{P} = (\mathbb{P}^{(1)}_{\text{single}}, \dots, \mathbb{P}^{(n)}_{\text{single}}) \sim \mathcal{D}^n} \mathbf{E}_{\mathbb{P}, \mathsf{Alg}} [K_x] \gtrsim \frac{nAH^2}{\epsilon^2}.$$

Since the number of episodes K encounted by Alg is equal to $\sum_{x=1}^n K_x$ (the agent visits exactly one state of the form (x, ℓ_0) per episode), we have

$$\mathbf{E}_{\mathbb{P}=(\mathbb{P}^{(1)}_{\mathrm{single}},\dots,\mathbb{P}^{(n)}_{\mathrm{single}})\sim\mathcal{D}^n}\mathbf{E}_{\mathbb{P},\mathsf{Alg}}[K] \gtrsim \sum_{x=1}^n \frac{nAH^2}{\epsilon^2} = \frac{n^2AH^2}{\epsilon^2}.$$

Since $S/8 \le n \le S$, for the above conditions to hold, it suffices that, for a sufficiently large constant C, $S \ge C\log_2 A$, $\epsilon \le \min\{\frac{1}{4}, \frac{H}{48}\}$, and $H \ge C\log_2 S$. Moreover, $\frac{n^2AH^2}{\epsilon^2} = \Omega(\frac{S^2AH^2}{\epsilon^2})$, as needed. □

D.4. Proof of Lemma D.2

A packing of reward-free MDPs The first step is to construct a family of transition probabilities $\mathbb{P}_J \in \mathscr{P}(\epsilon; n, A)$ which witness the lower bound. Let $\mathbf{1}$ denote the all ones vector on $[2n]$. To construct the packing, we define the set of binary vectors

$$\mathcal{K} := \left\{v \in \{-1,1\}^{2n} \ : \mathbf{1}^\top v = 0\right\}.$$

For a cardinality parameter M to be chosen shortly, we consider a packing of vectors

$$\mathcal{V}_{A,M} := \{v_{a,j} \in \mathcal{K} : a \in [A], j \in [M]\}$$

Throughout, we shall consider packings $\mathcal{V}_{A,M}$ which are *uncorrelated* in the following sense:

Definition D.5 (Uncorrelated)**.** For $\gamma \in (0,1)$, we say that $\mathcal{V}_{A,M}$ is γ-uncorrelated if, for any pair $(a,j),(a',j')$ with *either* $a \ne a'$ or $j \ne j'$, it holds that $|\langle v_{a,j}, v_{a',j'}\rangle| < 2n\gamma$..

The following lemma shows that the exist γ-uncorrelated packings of size $e^{\Omega(n\gamma^2)}$:

Lemma D.6. *Fix* $\gamma \in (0,1)$*, and suppose that* $2\log(M) \le n\gamma^2 - \log(4n) - 2\log(A)$*. Then, there exists a* γ*-uncorrelated packing* $\mathcal{V}_{A,M}$*.*

Proof Sketch. We use the probabilistic method. Specifically, we draw $v_{a,j} \overset{\mathrm{unif}}{\sim} \mathcal{K}$, and can bound $\langle v_{a,j}, v_{a',j'}\rangle$ with high-probability Chernoff bounds. Taking a union bound shows that an uncorrelated packings arise from this construction with non-zero probability. A full proof is given in in Section D.4.1. □

Given a γ-uncorrelated packing $\mathcal{V}_{A,M}$, define transition vectors

$$q_{a,j} := q_0 + \frac{\epsilon}{2n} v_{a,j}, \text{ where } q_0 = \frac{1}{2n}\mathbf{1}.$$

Since $\epsilon \le 1$ and $\mathbf{1}^\top v_{a,j_a} = 0$, $q_{j,a} \in \Delta(2n)$. Wet indices J denote tuples $J = (J_1,\dots,J_A) \in [M]^A$, let $q_J(\cdot, a) = q_{a,J_a}$, and define $\mathbb{P}_J$ as the instance $\mathbb{P}_{q_J}$, where $\mathbb{P}_q$ is as in Definition []. Formally,

$$\mathbb{P}_J: \quad \mathbb{P}^{\mathbb{P}_J}[s_1 = 0] = 1,\ \mathbb{P}^{\mathbb{P}_J}[s_2 = 0] = 0,\ \forall s \in [2n],\ \mathbb{P}^{\mathbb{P}_J}[s_2 = s \mid s_1 = 0, a] = q_J(s,a) = q_{a,J_a}(s)$$

Lower Bound for Estimating the Packing Instance: Let us suppose we have an exploration algorithm $\mathsf{Alg}_{\mathrm{est}}$ which, for any $\mathbb{P}_J$, collects (a possibly random number) K trajectories, and returns estimates $\widehat{J}_1,\dots,\widehat{J}_A$ of $J_1,\dots,J_A$. Our first step is to establish a lower bound on K assuming that $\mathsf{Alg}_{\mathrm{est}}$ satisfies a uniform correctness guarantee:

Lemma D.7. *For any* $\mathsf{Alg}_{\mathrm{est}}$ *satisfying the guarantee*

$$\forall J \in [A]^M,\ \mathbb{P}_{\mathbb{P}_J,\mathsf{Alg}_{\mathrm{est}}}\left[\widehat{J}_a = J_a\, \forall a \in [A]\right] \ge 1 - a. \tag{9}$$

Then, we must have

$$\mathbf{E}_{J \overset{\mathrm{unif}}{\sim}[A]^M}\mathbf{E}_{\mathbb{P}_J,\mathsf{Alg}_{\mathrm{est}}}[K] \ge A \cdot \frac{(1-p)\log M - \log 2}{\epsilon^2}$$

The above bound essentially follows from an application of Fano's inequality, and is proven in Section D.4.2. In particular, if we take say $p = 1/2$, and require $M = e^{\Omega(S)}$, then we have $\mathbf{E}_{J \overset{\mathrm{unif}}{\sim}[A]^M}\mathbf{E}^{\mathbb{P}_J,\mathsf{Alg}_{\mathrm{est}}}[K] \gtrsim \frac{SA}{\epsilon^2}$, as desired.

Estimation Reduces to Exploration Of course, the above bound applies only to an estimation algorithm $\mathsf{Alg}_{\mathrm{est}}$, but our intent is to establish lower bounds for exploration algorithms. In the following lemma, we state that if the packing is suffciently uncorrelated, then we can convert an $(\epsilon/24, p)$-correct exploration algorithm into an Algorithm $\mathsf{Alg}_{\mathrm{est}}$ satisfying Eq. (9).

Lemma D.8. *Suppose* Alg *is* $(\epsilon/24, p)$*-correct on the class* $\mathscr{M}_{\mathrm{single}}(\epsilon, n, A)$*, and that the packing* $\mathcal{V}_{M,A}$ *is* $\gamma = 1/10$*-uncorrelated. Then, there is an algorithm* $\mathsf{Alg}_{\mathrm{est}}$ *which collects* K *trajectories according to* Alg*, and satisfies Eq. 9.*

Proof Sketch. Consider reward vectors r_ν induced by $\nu_{a,j,a_2,j_2} \propto 2q_{a,j} - q_{a_2,j_2}$. These reward vectors can be used to "pick out" q_{a,J_a} as follows. For a given a, we show that on the good exploration event, Alg returns policies with $\mathbb{P}[\widehat{\pi}_1^\nu(0) = a] > 1/2$ for all $\nu = \nu_{a,J_a,a_2,j_2}$ ranging across a_2, j_2. However, for $j \neq J_a$, we show that on this good event there exists some a_2, j_2 for which Alg returns policies with $\mathbb{P}[\widehat{\pi}_1^\nu(0) = a] < 1/2$. Hence, we can estimate q_{a,J_a} by finding the (say, the first) index j for which $\mathbb{P}[\widehat{\pi}_1^\nu(0) = a] > 1/2$ for all $\nu = \nu_{a,j,a_2,j_2}$, ranging across a_2, j_2. A full proof is given in Section D.4.3. □

As a consequence, we find that if $\gamma \le 1/10$ and Alg is $(\epsilon/24, p)$-correct,

$$\mathbf{E}_{J \overset{\mathrm{unif}}{\sim} [A]^M} \mathbf{E}_{\mathbb{P}_J, \mathsf{Alg}}[K] \ge A \cdot \frac{(1-p)\log M - \log 2}{\epsilon^2}$$

In particular, if $\log M \ge 4 \log 2$ and $p \le 1/2$, then,

$$\mathbf{E}_{J \overset{\mathrm{unif}}{\sim} [A]^M} \mathbf{E}_{\mathbb{P}_J, \mathsf{Alg}}[K] \ge A \cdot \frac{\log M}{4\epsilon^2} \tag{10}$$

Concluding the proof Take $\gamma = 1/10$. For constants c_0, c_1 sufficiently large, we can ensure that if $n \ge c_0 \log_2 A$, then $M = e^{-n/c_1}$ statisfies $2\log(M) \le n\gamma^2 - \log(4n) - 2\log(A)$ and $\log M \ge 4 \log 2$. Thus, we can construct a γ-uncorrelated packing of cardinality $\log M \ge n/c_1$,

$$\mathbf{E}_{J \overset{\mathrm{unif}}{\sim} [A]^M} \mathbf{E}_{\mathbb{P}_J, \mathsf{Alg}}[K] \ge A \cdot \frac{n}{4c_1\epsilon^2},$$

as needed. □

D.4.1. Proof of Lemma D.6

We begin with the following concentration inequality:

Lemma D.9. *For any fixed* (a, j) *and* (a', j')*, we have*

$$\mathbb{P}[|\langle v_{a,j}, v_{a',j'}\rangle| \ge 2n\gamma] \le e^{\log(4n) - n\gamma^2}.$$

Proof. By permuting coordinates, we may assume that

$$v_{a',j'}[s] = \begin{cases} 1 & s \in [n] \\ -1 & s \in \{n+1, \dots, 2n\} \end{cases}.$$

Then,

$$\begin{aligned} \langle v_{a,j}, v_{a',j'}\rangle &= 2|\{s \in [n] : v_{a,j}[s] = 1\}| - 2(n - |\{s \in [n] : v_{a,j}[s] = 1\}|) \\ &= 2n - 4|\{s \in [n] : v_{a,j}[s] = 1\}| := 2n - 4Z, \end{aligned}$$

where we set $Z = |\{s \in [n] : v_{a,j}[s] = 1\}|$. Hence, if $|\langle v_{a,j}, v_{a',j'}\rangle| \ge 2\gamma n$, we need

$$\left|\frac{Z}{n} - \frac{1}{2}\right| \ge \frac{\gamma}{2}.$$

Now, we have that for $i \in [n]$,

$$\mathbb{P}[Z = i] < \frac{\binom{n}{i} \cdot \binom{n}{n-i}}{\sum_{i=0}^{n} \binom{n}{i} \cdot \binom{n}{n-i}} = \frac{\binom{n}{i}^2}{\sum_{i=0}^{n} \binom{n}{i}^2} < n \frac{\binom{n}{i}^2}{\left(\sum_{i=0}^{n} \binom{n}{i}\right)^2} = n\mathbb{P}_{W \sim \mathrm{Binom}(n,1/2)}[W = i]^2.$$

Hence,

$$\begin{aligned}
\mathbb{P}\left[\left|\frac{Z}{n} - \frac{1}{2}\right| \geq \frac{\gamma}{2}\right] &\leq n \sum_{i: |\frac{i}{n} - \frac{1}{2}| \geq \frac{\gamma}{2}} \mathbb{P}_{W \sim \mathrm{Binom}(n,1/2)}[W = i]^2 \\
&\leq n \left(\sum_{i: |\frac{i}{n} - \frac{1}{2}| \geq \frac{\gamma}{2}} \mathbb{P}_{W \sim \mathrm{Binom}(n,1/2)}[W = i]\right)^2 \\
&= n \left(\mathbb{P}_{W \sim \mathrm{Binom}(n,1/2)}\left[\left|\frac{W}{n} - \frac{1}{2}\right| \geq \frac{\gamma}{2}\right]\right)^2 \qquad \leq n(2e^{-2(\gamma/2)^2 n})^2 = e^{\log(4n) - n\gamma^2}
\end{aligned}$$

□

We now finish the proof of our intended lemma:

Proof of Lemma D.6. By a union bound over at most $A^2 M^2 - 1$ pairs $(a, j), (a', j')$, there exists a γ-uncorrelated packing for any M satisfying

$$A^2 M^2 e^{\log(4n) - n\gamma^2} \leq 1$$

Taking logarithms, we require $2\log(M) \leq n\gamma^2 - \log(4n) - 2\log(A)$.

□

D.4.2. Proof of Lemma D.7

To begin, let us state a variant of Fano's inequality, which replaces mutual-information with an arbitrary comparison measure:

Lemma D.10 (Fano's Inequality). *Consider M probability measures $\mathbb{P}_1, \dots, \mathbb{P}_M$ on a space Ω. Then for any estimator $\widehat{j}$ on Ω and any comparison law $\mathbb{P}_0$ on Ω,*

$$\frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_j\left[\widehat{j} \neq j\right] \geq 1 - \frac{\log 2 + \frac{1}{M}\sum_{j=1}^{M} \mathrm{KL}(\mathbb{P}_j, \mathbb{P}_0)}{\log M}$$

Proof. This follows from the standard statement of Fano's inequality, where we use that

$$\inf_{\mathbb{P}_0} \frac{1}{M} \sum_{j=1}^{M} \mathrm{KL}(\mathbb{P}_j, \mathbb{P}_0) = \frac{1}{M} \sum_{j=1}^{M} \mathrm{KL}\left(\mathbb{P}_j, \frac{1}{M} \sum_{j'=1}^{M} \mathbb{P}_{j'}\right)$$

For reference, see e.g. Equation (11) in (Chen et al., 2016). □

We will apply Fano's inequality of each $a \in [A]$. To begin, for a fixed $J \in [M]^A$ and $a \in [A]$, let us define the laws "$\mathbb{P}_j$". We let $\mathbb{P}_{J,a,j}$ denote the reward-free MDP with starting at $x = 0$ deterministically, and with transitions

$$\mathbb{P}^{\mathbb{P}_{J,a,j}}[s \mid x_1 = 0, a_1 = a'] = \begin{cases} q_{a,j}[s] & a' = a \\ q_{a',J_{a'}}[s] & a' \neq a. \end{cases}$$

For fixed J, a, we let $\mathbb{P}_{j;J,a}$ denote the joint law induced by $\mathsf{Alg}_{\mathrm{est}}$ and $\mathbb{P}_{J,a,j}$. For the comparison measure, let $\mathbb{P}_{J,a,0}$ denote the analogous MDP to $\mathbb{P}_{J,a,j}$, but where $\mathbb{P}^{\mathbb{P}_{J,a,j}}[s \mid x_1 = 0, a_1 = a] = q_0$ for the fixed action a. We let $\mathbb{P}_{0;J,a}$ denote the law induced by $\mathsf{Alg}_{\mathrm{est}}$ and $\mathbb{P}_{J,a,j}$. Then, Fano's iqequality implies that

$$\forall J, a, \quad (1-p)\log M - \log 2 \le \frac{1}{M}\sum_{j=1}^{M} \mathrm{KL}(\mathbb{P}_{J,a,j}, \mathbb{P}_{0;J,a}). \tag{11}$$

Now, observe that the laws $\mathbb{P}_{J,a,j}$ and $\mathbb{P}_{0;J,a}$ only differ due to transitions selecting action $a_1 = a$. Under the first law, these have distribution $\mathrm{Multinomial}(q_{a,j})$, and under the second, $\mathrm{Multinomial}(q_0)$. Let $N_K(a = a_1)$ denote the expected number of times algorithm $\mathsf{Alg}_{\mathrm{est}}$ selects action $a_1 = a$ at time step 1. From a Wald's identity argument (see e.g. (Kaufmann et al., 2016)), we have

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{J,a,j}, \mathbb{P}_{0;J,a}) &= \mathbf{E}_{\mathbb{P}_{J,a,j},\mathsf{Alg}_{\mathrm{est}}}[N_K(a_1 = a)]\, \mathrm{KL}(\mathrm{Multinomial}(q_{a,j}), \mathrm{Multinomial}(q_{a,0}))\\
&= \mathbf{E}_{\mathbb{P}_{J,a,j},\mathsf{Alg}_{\mathrm{est}}}[N_K(a_1 = a)] \sum_{s=1}^{2n} \frac{1+\epsilon v_{j,a}[s]}{2n}\log(1+\epsilon v_{j,a}[s])\\
&\overset{(i)}{\le} \mathbf{E}_{\mathbb{P}_{J,a,j},\mathsf{Alg}_{\mathrm{est}}}[N_K(a_1 = a)] \sum_{s=1}^{2n} \frac{\epsilon v_{j,a} + \epsilon^2 v_{j,a}[s]^2}{2n}\\
&\overset{(ii)}{\le} \epsilon^2 \cdot \mathbf{E}_{\mathbb{P}J,a,j,\mathsf{Alg}_{\mathrm{est}}}[N_K(a_1 = a)]
\end{aligned}$$

where (i) uses $1+\epsilon v_{j,a}[s] \ge 0$ and the identity $\log(1+x) \le x$, and (ii) uses the fact that $v_{j,a}[s]^2 = 1$ and $\sum_{s=1}^{2n} v_{j,a}[s] = 0$ for $v_{j,a} \in \mathcal{K}$. Thus, by Eq 11,

$$\forall J, a, \quad \frac{(1-p)\log M - \log 2}{\epsilon^2} \le \frac{1}{M}\sum_{j=1}^{M} \mathbf{E}_{\mathbb{P}_{J,a,j},\mathsf{Alg}_{\mathrm{est}}}[N_K(a_1 = a)].$$

By taking an expectation over index tuples J drawn uniformly from $[A]^M$, we have

$$\begin{aligned}
\forall a, \quad \frac{(1-p)\log M - \log 2}{\epsilon^2} &\le \frac{1}{M}\sum_{j=1}^{M} \mathbf{E}_{J \overset{\mathrm{unif}}{\sim} [A]^M} \mathbf{E}_{\mathbb{P}_{J,a,j},\mathsf{Alg}_{\mathrm{est}}}\left[N_K(a_1 = a)\right]\\
&= \mathbf{E}_{J \overset{\mathrm{unif}}{\sim} [A]^M} \mathbf{E}_{\mathbb{P}_{J},\mathsf{Alg}_{\mathrm{est}}}\left[N_K(a_1 = a)\right],
\end{aligned}$$

where the last line follows that $\mathbb{P}_{J,a,j} = \mathbb{P}_{J'}$ for some J' and that, by symmetry, each index J' has equal weight when averaged over both $J \in [A]^M$ and $j \in [M]$. Summing over $a \in [A]$, we have

$$A \cdot \frac{(1-p)\log M - \log 2}{\epsilon^2} \le \mathbf{E}_{J \overset{\mathrm{unif}}{\sim} [A]^M} \mathbf{E}_{\mathbb{P}_{J},\mathsf{Alg}_{\mathrm{est}}}\left[\sum_{a=1}^{A} N_K(a_1 = a)\right] = \mathbf{E}_{J \overset{\mathrm{unif}}{\sim} [A]^M} \mathbf{E}_{\mathbb{P}_{J},\mathsf{Alg}_{\mathrm{est}}}[K].$$

D.4.3. Proof of Lemma D.8

Let us now show that $(\epsilon/12, p)$-learning implies the existence of an algorithm $\mathsf{Alg}_{\mathrm{est}}$ satisfying Eq. 9, provided the packing is sufficiently uncorrelated. Introduce the vectors

$$\nu_{a_1,a_2,j_1,j_2} := \frac{1}{3}v_{a_1,j_1} + \frac{1}{6}v_{a_2,j_2} + \frac{1}{2}\mathbf{1},$$

which can be checked to lie $[0,1]^{2n}$. We shall establish the following lemma, which says that for sufficciently uncorrelated packings, the vectors $\nu_{(\dots)}$ witness separations between q_{a_1,j_1} and q_{a_2,j_2} for different actions a_1, a_2:

Lemma D.11. *Fix $a_1 \in [A]$ and $j_1 \in [M]$, and suppose the packing is $\gamma = 1/10$-uncorrelated: Then, for any $a_2 \ne a_1$ and $j_2 \in [M]$, the following holds*

$$\begin{aligned}
\min_{a_2', j_2'} \langle q_{a_1,j_1} - q_{a_2,j_2}, \nu_{a_1,a_2',j_1,j_2'}\rangle &> \frac{\epsilon}{12}\\
\forall j_1' \ne j_1,\ \min_{a_2', j_2'} \langle q_{a_1,j_1} - q_{a_2,j_2}, \nu_{a_1,a_2',j_1',j_2'}\rangle &< -\frac{\epsilon}{12}
\end{aligned}$$

Proof of Lemma D.11.

$$\begin{aligned}\langle q_{a_1,j_1} - q_{a_2,j_2}, \nu_{a_1',a_2',j_1',j_2'}\rangle &= \frac{\epsilon}{2n}\langle v_{a_1,j_1} - v_{a_2,j_2}, \nu_{a_1',a_2',j_1',j_2'}\rangle \\ &= \frac{\epsilon}{12n}\langle v_{a_1,j_1} - v_{a_2,j_2}, 2v_{a_1',j_1'} - v_{a_2',j_2'}\rangle,\end{aligned}$$

where we use the fact that $v_{a,j}^\top \mathbf{1} = 1$ for all a, j. If $a_1' = a_1$ and $j_1' = j_1$, and the packing is $\gamma \le 1/6$-uncorrelated

$$\begin{aligned}\langle q_{a_1,j_1} - q_{a_2,j_2}, \nu_{a_1,a_2',j_1,j_2'}\rangle &= \frac{\epsilon}{12n}\langle v_{a_1,j_1} - v_{a_2,j_2}, 2v_{a_1,j_1} - v_{a_2',j_2'}\rangle \\ &= \frac{\epsilon}{12n}\left(2\langle v_{a_1,j_1}, v_{a_1,j_1}\rangle - 2\langle v_{a_2,j_2}, v_{a_1,j_1}\rangle + \langle v_{a_1,j_1}, v_{a_2',j_2'}\rangle - \langle v_{a_2,j_2}, v_{a_2',j_2'}\rangle\right) \\ &> \frac{\epsilon}{12n}\left(4n - 4\gamma n - 2n - 2n\gamma\right) \\ &\ge \frac{\epsilon}{12n}\left(2n - 6n\gamma\right) = \frac{\epsilon}{12}.\end{aligned}$$

On the other hand, if $j_1 \neq j_1'$, but $(a_2, j_2) = (a_2', j_2')$ then a similar computation reveals that for $\gamma \le 1/10$,

$$\langle q_{a_1,j_1} - q_{a_2,j_2}, \nu_{a_1,a_2,j_1',j_2}\rangle < \frac{\epsilon}{12n}\left(10\gamma n - 2n\rangle\right) < \frac{-\epsilon}{12}.$$

□

We can now conclude the proof of our reduction:

Proof of Lemma D.8. Suppose that Alg is run on $\mathbb{P}_J$ for $J \in [M]^A$. Further, recall the rewards r_ν which assign reward of $r_\nu(s,a) = \mathbf{I}(s \in [2n])\nu(s)$. By $(\epsilon/24, p)$-correctness of Alg, then with probability $1-p$, Alg computes policies $\widehat{\pi}_\nu$ which satisfies the following bound simultaneously for all $\nu \in \{\nu_{a_1,a_2,j_1,j_2}\}$:

$$\max_\pi V^\pi(\mathbb{P}_J, r_\nu) - V^{\widehat{\pi}_\nu}(\mathbb{P}_J, r_\nu) \le \epsilon/24. \tag{12}$$

For a possibly randomized policy, we use the shorthand $\pi[a]$ to denote the probability of selecting a at the initial state 0; that is $\mathbb{P}^\pi[a_1 = a]$. Now, Consider the following procedure: for each $a \in [A]$, estimate J_a by returning the first $j \in [M]$ for which

$$\forall a_2', j_2', \quad \widehat{\pi}_{\nu_{a,a_2',j,j_2'}}[a] > 1/2. \tag{13}$$

We conclude our proof by showing that, on the good event Eq. (12), the condition in Eq. (13) holds if and only if $j = J_a$. To this end, define the short hand

$$q_\pi := \sum_{a'} \pi[a'] q_{a',J_{a'}}$$

Then, we have that

$$\max_\pi V^\pi(\mathbb{P}_J, r_\nu) - V^{\widehat{\pi}_\nu}(\mathbb{P}_J, r_\nu) = \max_\pi \langle q_\pi - q_{\widehat{\pi}_\nu}, \nu\rangle,$$

so that on the good event of Eq. 12, we have

$$\max_\pi \langle q_\pi - q_{\widehat{\pi}_\nu}, \nu\rangle \le \frac{\epsilon}{24}.$$

True Positive for $j = J_a$: First let's show that Equation 13 holds for $j = J_a$. Indeed, if it does not, then there exists some a_2', j_2' for which $\mathbb{P}[\widehat{\pi}_{\nu_{a,j,a_2',j_2'}}[a]] \le 1/2$, and (setting $\nu = \nu_{a,j,a_2',j_2'}$ for shorthand in $\widehat{\pi}^\nu$)

$$\epsilon/24 \ge \max_\pi \langle q_\pi - q_{\widehat{\pi}_\nu}, \nu\rangle,$$

$$
\begin{aligned}
&\geq \left\langle q_{a,J_a} - q_{\widehat{\pi}^\nu},\, \nu_{a,j,a_2',j_2'} \right\rangle && \text{(choose } \pi[a] = 1\text{)}\\
&= \sum_{a' \neq a} \widehat{\pi}_\nu[a'] \left\langle q_{a,J_a} - q_{a',J_{a'}},\, \nu_{a,j,a_2',j_2'} \right\rangle\\
&\geq \underbrace{(1 - \widehat{\pi}_\nu[a])}_{\geq 1/2} \cdot \underbrace{\min_{a' \neq a} \left\langle q_{a,J_a} - q_{a',J_{a'}},\, \nu_{a,j,a_2',j_2'} \right\rangle}_{> \epsilon/12 \text{ by Lemma D.11}} > \frac{\epsilon}{24},
\end{aligned}
$$

yielding a contradiction.

True Negative for $j \neq J_a$: On the other hand, for $j \neq J_a$ suppose that for all all $a_2' \neq a$ and all $j_2' \in [M]$, $\mathbb{P}[\widehat{\pi}_1^{\nu_{a,j,a_2',j_2'}}(0) = a] > 1/2$. Then, considering $a_2' = a_2$ and $j_2' = J_{a_2}$, we have (setting $\nu = \nu_{a,j,a_2,J_{a_2}}$ for shorthand in $\widehat{\pi}^\nu$)

$$
\begin{aligned}
\epsilon/24 &\geq \max_{a'} \left\langle q_{a',J_{a'}} - q_{\widehat{\pi}^\nu},\, \nu_{a,j,a_2,J_2} \right\rangle\\
&\geq \left\langle q_{a_2,J_{a_2}} - q_{\widehat{\pi}^\nu},\, \nu_{a,j,a_2,J_2} \right\rangle\\
&\geq \underbrace{\widehat{\pi}_\nu[a_2]}_{\geq \widehat{\pi}_\nu[a] > 1/2} \cdot \underbrace{\min_{a' \neq a_2} \left\langle q_{a_2,J_{a_2}} - q_{a',J_{a'}},\, \nu_{a,j,a_2',j_2'} \right\rangle}_{> \epsilon/12 \text{ by Lemma D.11}} > \frac{\epsilon}{24},
\end{aligned}
$$

again drawing a contradiction. $\square$