Optimism in Face of a Context: Regret Guarantees for Stochastic Contextual MDP

Orin Levy orinlevy@mail.tau.ac.il
Tel Aviv University
Israel

Yishay Mansour mansour.yishay@gmail.com
Tel Aviv University and Google Research
Israel

Abstract

We present regret minimization algorithms for stochastic contextual MDPs under minimum reachability assumption, using an access to an offline least square regression oracle. We analyze three different settings: where the dynamics is known, where the dynamics is unknown but independent of the context and the most challenging setting where the dynamics is unknown and context-dependent. For the latter, our algorithm obtains $\tilde{O}\left(\max\{H, \frac{1}{p_{min}}\}H|S|^{3/2}\sqrt{|A|T\log\frac{\max\{|\mathcal{G}|,|\mathcal{P}|\}}{\delta}}\right)$ regret bound, with probability $1-\delta$, where $\mathcal{P}$ and $\mathcal{G}$ are finite and realizable function classes used to approximate the dynamics and rewards respectively, p_{min} is the minimum reachability parameter, S is the set of states, A the set of actions, H the horizon, and T the number of episodes. To our knowledge, our approach is the first optimistic approach applied to contextual MDPs with general function approximation (i.e., without additional knowledge regarding the function class, such as it being linear and etc.). In addition, we present a lower bound of $\Omega(\sqrt{TH|S||A|\ln(|\mathcal{G}|)/\ln(|A|)})$, on the expected regret which holds even in the case of known dynamics.

Keywords: Reinforcement Learning, Regret Minimization, Contextual MDP, Function Approximation

1. Introduction

Markov decision processes (MDPs) have been extensively studied, and are commonly used to described dynamic environments. MDPs characterize a variety of real-life tasks and applications including: advertising, healthcare, games, robotics and more, where at each episode an agent interacts with the environment with the goal of maximizing her return. (See, e.g., Sutton and Barto (2018).)

In many applications, in each episode, there are additional external factors that affect the environment, which we refer to as the *context*. One way to handle this is to extend the state space to include the context. This approach has the disadvantage of greatly increasing the state space, and hence the complexity of learning and even the representation of a policy. An alternative approach, is to keep a small state space, and regard the context as an additional side-information. Contextual Markov Decision Process (CMDP) describes such a model, where for each context there is a potentially different optimal policy. The CMDP model was first presented by Hallak et al. (2015).

CMDPs are useful to model many user-driven applications, where the context is a user-related information which influences the optimal decision making. One natural application is in recommendation systems, where two different users might behave completely different from one another, hence, a single MDP can not describe them both. In recommendation systems, users behavior can be characterized using a side information about them, such as age, gender, interest fields and hobbies. We refer to that information as *context*. CMDP defines a mapping from context to a related MDP, and the optimal policy given a context is the optimal policy in the related MDP.

Our contributions. We present regret minimization algorithms for CMDP under three different settings: (1) known dynamics, (2) unknown context-independent dynamics and (3) unknown context-dependent dynamics, which is the most challenging. In all settings we assume an access a least square regression oracle, and finite function classes $\mathcal{G}$ and $\mathcal{P}$ used to approximate the rewards and dynamics, respectively. In addition, we assume minimum reachability, where any policy for any context has a probability of at least p_{min} to reach any state. For the known dynamics setting we

show $\tilde{O}\left(\max\{1/p_{min}, H\}|S|\sqrt{T|A|\log(|\mathcal{G}|/\delta)}\right)$ regret. For the unknown context-independent dynamics we also show an $\tilde{O}\left(\max\{H, 1/p_{min}\}|S|\sqrt{T|A|\log(|\mathcal{G}|/\delta)}\right)$ regret. For the unknown context-dependent dynamics we show an $\tilde{O}\left(\max\{H, 1/p_{min}\}H|S|^{3/2}\sqrt{|A|T\log(\max\{|\mathcal{G}|, |\mathcal{P}|\}/\delta)}\right)$ regret. All the bounds hold with high probability. Lastly, we present a lower bound of $\Omega(\sqrt{TH|S||A|\ln(|\mathcal{G}|)/\ln(|A|)})$ on the expected regret.

Our approach applies the "optimism in face of uncertainty" principle to CMDPs and achieves a sub-linear regret. Our algorithms and analysis were inspired by the optimistic approach of Xu and Zeevi (2020) for learning contextual multi armed bandits using least square regression oracle. We extended their approach to handle CMDPs and even a context-dependent dynamics.

2. Related Work

Contextual Reinforcement Leaning. CMDP was first introduce by Hallak et al. (2015). Modi et al. (2018) gives a general framework for deriving generalization bounds for smooth CMDPs and finite linear combination of MDPs. Modi and Tewari (2020) gives a regret bound for Generalized Linear Models (GLMs). Our function approximation framework is more general than GLM.

Foster et al. (2021) present a new statistical complexity measure, for interactive decision making, and show an application of it to obtain regret upper bound for Contextual RL. They assume an access to an online estimation oracle with regret guarantees, that maximizes over models and policies together. It is unclear when is this oracle implementable in polynomial time. In contrast, we make a significantly weaker and standard assumption regarding a regression oracle. Another difference is that we use an optimistic approach while they use the inverse gap minimization technique. (More details later.)

Jiang et al. (2017) present OLIVE which is sample efficient for Contextual Decision Processes (CDP) with low Bellman rank. We do not make any assumptions on the Bellman rank.

Levy and Mansour (2022) consider the sample complexity of learning CMDPs using function approximation. They provide the first general and efficient reduction from CMDP to offline supervised learning. Their sample complexity varies from $O(1/\epsilon^2)$ to $O(1/\epsilon^8)$, depending on the setting. We, in contrast, consider regret minimization and obtain $\tilde{O}(\sqrt{T})$ regret under the minimum reachability assumption.

Contextual Bandits. Contextual bandits (CMAB) are a natural extension of the Multi-Arm Bandit (MAB), augmented by a context which influences the rewards Slivkins (2019); Lattimore and Szepesvári (2020). Agarwal et al. (2014) use efficiently an optimization oracle to derive an optimal regret bound. Regression based approaches appear in Agarwal et al. (2012); Foster et al. (2018); Foster and Rakhlin (2020); Simchi-Levi and Xu (2021). We differ from CMAB, since our main challenge is the dynamics, and the need to optimize future rewards, which is the case in most RL settings.

Xu and Zeevi (2020) present the first optimistic algorithm for CMAB. They assume an access to a least-square regression oracle and achieve $\tilde{O}(\sqrt{T|A|\log|\mathcal{F}|})$ regret, where $\mathcal{F}$ is a finite and realizable function class uses to approximate the rewards. Our algorithms and analysis are inspired by their optimistic approach and we extend it to CMDP.

Inverse Gap Minimization (IGM) technique. Foster and Rakhlin (2020); Simchi-Levi and Xu (2021) apply the IGM technique to CMAB and obtain $\tilde{O}(\sqrt{T|A|})$ regret, assuming an access to a least square regression oracle. However, we do not see any straight-forward extension of their approach to CMDP which is both computationally efficient and has an optimal regret, under the same least-square oracle assumption (even when the dynamics is known to the learner). Foster et al. (2021) apply IGM to CMDP and obtain optimal regret. However they use the much strong online estimation oracle as discussed above. (For extended related work overview, see Appendix A.)

3. Preliminaries and Notations

Markov Decision Process (MDP) is a tuple (S, A, P, r, s_0, H), where (1) S is a finite state space, (2) A is a finite action space, (3) $s_0 \in S$ is the unique start state, (4) $P(\cdot|s, a)$ defines the transition probability function, i.e., $P(s'|s, a)$ is the probability that we reach state s' given that we are in state s and perform action a, (5) $R(s, a) \in [0, 1]$ is a random

variable for the reward of performing action a in state s, and $r(s,a)$ is its expectation, i.e., $r(s,a) = \mathbb{E}[R(s,a)|s,a]$, and (6) H is the finite horizon.

The state space is decomposed into $H+1$ disjoint subsets (layers) $S_0, S_1, \ldots, S_{H-1}, S_H$ such that transitions are only possible between consecutive layers (i.e., loop-free). In addition, $S_H = \{s_H\}$, meaning there is a unique final state with reward 0.

Policy. A *stochastic policy* π is a mapping from states to distribution over actions, i.e., $\pi : S \to \Delta(A)$. A *deterministic policy* π is a mapping from states to actions, i.e., $\pi : S \to A$.

Occupancy measure. Let $q_h(s,a|\pi,P)$ denote the probability of reaching state $s \in S$ and performing action $a \in A$ at time $h \in [H]$ of an episode generated using policy π and dynamics P. Let $q_h(s|\pi,P) = \sum_{a \in A} q_h(s,a|\pi,P)$, be the probability to visit state $s \in S$ at time h.

Episode and trajectory. At the start of each episode we select a policy π. The episode starts at the unique initial state s_0. In state $s_h \in S_h$, we play action $a_h \sim \pi(\cdot|s_h)$, observe a reward $r_h \sim R(s_h,a_h)$ and move to $s_{h+1} \sim P(\cdot|s_h,a_h)$. We generate a trajectory $\sigma_{H+1} = (s_0, a_0, r_0, s_1, \ldots, s_{H-1}, a_{H-1}, r_{H-1}, s_H)$ of length $H+1$.

Value functions. Given a policy π and a MDP $M = (S, A, P, r, s_0, H)$, the $h \in [H-1]$ stage value function of a state $s \in S_h$ is defined as $V^{\pi}_{M,h}(s) = \mathbb{E}_{\pi,M}[\sum_{k=h}^{H-1} r(s_k, \pi(s_k))|s_h = s]$ and for an action $a \in A$ we have $Q^{\pi}_{M,h}(s,a) = \mathbb{E}_{\pi,M}[\sum_{k=h}^{H-1} r(s_k,\pi(s_k))|s_h = s, a_h = a]$. For brevity, when $h = 0$ we denote $V^{\pi}_{M,0}(s_0) := V^{\pi}_M(s_0)$.

Optimal policy $\pi^{\star}_M$ for MDP M satisfies, for every stage $h \in [H-1]$ and a state $s \in S_h$, $\pi^{\star}_{M,h}(s) \in \arg\max_{\pi}\{V^{\pi}_{M,h}(s)\}$, and w.l.o.g it is a deterministic policy.

Planning. Given an MDP $M = (S, A, P, r, s_0, H)$ the algorithm $\texttt{Planning}(M)$ returns an optimal policy $\pi^{\star}_M$ and its value $V^{\star}_M(s_0)$ and runs in time $O(|S|^2\,|A|\,H)$.

Contextual Markov Decision Process (CMDP) is a tuple $(\mathcal{C}, S, A, \mathcal{M})$ where $\mathcal{C} \subseteq \mathbb{R}^{d'}$ is the context space, S the state space and A the action space. The mapping $\mathcal{M}$ maps a context $c \in \mathcal{C}$ to a MDP $\mathcal{M}(c) = (S, A, P^c_{\star}, r^c_{\star}, s_0, H)$, where $r^c_{\star}(s,a) = \mathbb{E}[R^c_{\star}(s,a)|c,s,a]$, $R^c_{\star}(s,a) \sim \mathcal{D}_{c,s,a}$.

There is an unknown distribution $\mathcal{D}$ over the context space $\mathcal{C}$, and for each episode a context is sampled i.i.d. from $\mathcal{D}$. For mathematical convenience, we assume the context space is finite (but potentially huge). Our results naturally extend to infinite contexts space.

Context-Independent and Context-Dependent dynamics. A CMDP has *context-independent* dynamics when the context effects only the rewards function, while the dynamics are identical for all contexts, i.e., $P^c_{\star} = P$ for any context c. A *context-dependent* dynamics has a potentially different dynamics $P^c_{\star}$ for each context c. We consider both context-independent and context-dependent dynamics.

Context-dependent policies. A stochastic context-dependent policy $\pi = (\pi(c;\cdot) : S \to \Delta(A))_{c \in \mathcal{C}}$ maps a context $c \in \mathcal{C}$ to a stochastic policy $\pi(c;\cdot) : S \to \Delta(A)$. A deterministic context-dependent policy $\pi = (\pi(c;\cdot) : S \to A))_{c \in \mathcal{C}}$ maps a context $c \in \mathcal{C}$ to a policy $\pi(c;\cdot) : S \to A$. Let $\Pi_{\mathcal{C}}$ denote the class of all deterministic context-dependent policies. A (deterministic) context-dependent policy $\pi^{\star} \in \Pi_{\mathcal{C}}$ is *optimal* if for all $c \in \mathcal{C}$ it holds that $\pi^{\star}(c;\cdot) \in \arg\max_{\pi : S \to A} V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)$.

Minimum reachability. We assume that there exists $p_{min} \in (0,1]$ such that for every context $c \in \mathcal{C}$, layer $h \in [H-1]$ and state $s_h \in S^c_h$, any context-dependent policy $\pi \in \Pi_{\mathcal{C}}$ satisfies $q_h(s_h|\pi(c;\cdot), P^c_{\star}) \geq p_{min}$. Let $q(s|\pi(c;\cdot), P^c_{\star})$ denote the probability of visiting state s when playing π on the dynamics $P^c_{\star}$. When the dynamics is layered and loop-free, then $q(s|\pi(c;\cdot), P^c_{\star}) = q_h(s|\pi(c;\cdot), P^c_{\star}) \geq p_{min}$ if and only if $s \in S^c_h$. We remark that our minimum reachability assumption is much more refined than that usually used in RL literature, that $P^c_{\star}(s'|s,a) \geq p_{min}$ for every context c and (s,a,s'). Clearly, this requirement implies our minimum reachability, but the other direction does not necessarily hold.

Interaction protocol. In each episode $t = 1, 2, ..., T$ the agent: (1) Observes context $c_t \in \mathcal{C}$. (2) Chooses a policy π_t (based on c_t and the observed history). (3) Observes a trajectory of π_t in $\mathcal{M}(c_t)$.

Trajectories and History. Each episode is of length H. A trajectory of length $h \in [H]$, which we denote $\sigma_h = (c; s_0, a_0, r_0, \ldots, s_{h-1}, a_{h-1}, r_{h-1})$, is generated using the dynamics $P^c_{\star}$ and the played policy $\pi(c;\cdot)$. We

denote the history up to time $t-1$ by $\mathbb{H}_{t-1} = (\sigma^1_{H+1}, \ldots, \sigma^{t-1}_{H+1})$ where σ^i_{H+1} is the trajectory of length $H+1$ observed in time $i \in [t-1]$, i.e., $\sigma^i_{H+1} = (c_i, s^i_0, a^i_0, r^i_0, \ldots, s^i_{H-1}, a^i_{H-1}, r^i_{H-1}, s_H)$.

Offline least square regression (LSR) oracle solves the optimization problem $\hat{f} \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (f(x_i) - y_i)^2$, given a data set $D = \{(x_i, y_i)\}_{i=1}^n$ and a function class $\mathcal{F}$.

Reward function approximation. We consider a finite function class $\mathcal{G} \subseteq (\mathcal{C} \times A \to [0,1])$ to approximate the context-dependent rewards function of each state $s \in S$. Many times it would be more convenient to consider a finite function class $\mathcal{F} = \mathcal{G}^S$ where $f \in \mathcal{F}$ are functions of the form $f(c,s,a) = g_s(c,a)$ where $g_s \in \mathcal{G}$. Note that, $\log(|\mathcal{F}|) = |S| \log(|\mathcal{G}|)$. Our algorithms get as input the finite function class $\mathcal{F} \subseteq (\mathcal{C} \times S \times A \to [0,1])$. Each function $f \in \mathcal{F}$ maps context $c \in \mathcal{C}$, state $s \in S$ and action $a \in A$ to a (approximate) reward $r \in [0,1]$. We use $\mathcal{F}$ to approximate the context-dependent rewards function using the LSR oracle under the following realizability assumption.

Assumption 1 (rewards realizability) *We assume that $\mathcal{F}$ is realizable, meaning, there exists a function $f_\star \in \mathcal{F}$ such that $f_\star(c,s,a) = r^c_\star(s,a) = \mathbb{E}[R^c_\star(s,a)|c,s,a]$.*

For mathematical convenience, we state our algorithms and regret upper bounds in terms of the cardinality of $|\mathcal{F}|$, and use the cardinality of $|\mathcal{G}| = S^{-1} \log|\mathcal{F}|$ for our lower bound. We present a comparison between the bounds in Section 8.

Dynamics function approximation. For the unknown context-independent dynamics case we simply use a tabular approximation (see Section 5). For the unknown context-dependent case, our algorithm gets as input a finite function class $\mathcal{P} \subseteq S \times (S \times A \times \mathcal{C}) \to [0,1]$, where every function $P \in \mathcal{P}$ satisfies $\sum_{s' \in S} P(s'|s,a,c) = 1$ for all $c \in \mathcal{C}$ and $(s,a) \in S \times A$. We use $\mathcal{P}$ to approximate the context-dependent dynamics using LSR oracle under the following realizability assumption.

Assumption 2 (Dynamics Realizability) *We assume that $\mathcal{P}$ is realizable, meaning there exist a function $P_\star \in \mathcal{P}$ which is the true context-dependent dynamics.*

Learning goal. Our goal is to minimize the regret, relative to the value of the optimal context-dependent policy $\pi^\star$, which defined as $\text{Regret}_T := \sum_{t=1}^{T} V^{\pi^\star(c_t;\cdot)}_{\mathcal{M}(c_t)}(s_0) - V^{\pi_t(c_t;\cdot)}_{\mathcal{M}(c_t)}(s_0)$, where c_t and $\pi_t \in \Pi_{\mathcal{C}}$ are the context and the selected policy at round t. Denote the expected regret $\text{E.Regret}_T := \mathbb{E}[\text{Regret}_T]$ where the expectation is over the contexts, the randomization of the algorithm and the history.

Mathematical notations. We denote expectation by $\mathbb{E}[\cdot]$, variance by $\mathbb{V}[\cdot]$ and probabilities by $\mathbb{P}[\cdot]$. The indicator function is $\mathbb{I}[G]$ returns 1 if event G holds and 0 otherwise.

4. Known Dynamics

In this section, we present regret minimization algorithm (see Algorithm 1) for contextual MDPs under the minimum reachability assumption, where the context-dependent dynamics $P^c_\star$ is known to the learner, for every context $c \in \mathcal{C}$. We remark that the minimum reachability parameter p_{min} is unknown to the learner. This section sets the main building blocks of our approach, which we will later be extend to handle the unknown dynamics cases.

Algorithm outline. For the first $|A|$ rounds, in round $i \in [|A|]$ the agent plays the policy $\pi_i \in \Pi_{\mathcal{C}}$ that always selects action a_i, regardless of the context and the state. At every round $t > |A|$ we approximate the context-dependent rewards function using a least-square minimizer. Using it, we build an "optimistic in expectation" rewards function, and compute an optimal policy for the optimistic model. We run it to generate a trajectory and update the oracle. Here, we take advantage of the ability to compute the optimal policy $\pi_k(c;\cdot)$ for every context $c \in \mathcal{C}$ separately, for all $k = |A|+1, \ldots, t$, to obtain computationally efficient algorithm (we discuss this challenge in Subsection 4.1).

Remark 1 *Since the CMDP is layered, for every context $c \in \mathcal{C}$, layer $h \in [H]$ and state $s_h \in S^c_h$ we have $q_h(s_h|\pi_i(c;\cdot), P^c_\star) = q(s|\pi_i(c;\cdot), P^c_\star)$. For convenience, in Algorithm 1 we use q to compute the approximated rewards function, but in the regret analysis we use q_h.*

Algorithm 1 Regret Minimization for CMDP with Known Dynamics (RM-KD)

inputs: MDP parameters: S , A, $P_\star$, s_0, H. Confidence $\delta > 0$ and tuning parameters $\{\beta_t\}_{t=1}^T$.
Initialization: In round $i \leq |A|$, run policy $\pi_i(c; s) := a_i$
for round $t = |A| + 1, \ldots, T$ **do**
  compute $\hat{f}_t \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{t-1} \sum_{h=0}^{H-1} (f(c_i, s_h^i, a_h^i) - r_h^i)^2$ using the LSR oracle
  observe context $c_t \in \mathcal{C}$
  for $k = |A| + 1, |A| + 2, \ldots, t$ **do**
    compute $\forall (s, a) \in S \times A: \quad \hat{r}_k^{c_t}(s, a) = \hat{f}_k(c_t, s, a) + \frac{\beta_k}{\sum_{i=1}^{k-1} \mathbb{I}[a=\pi_i(c_t;s)] q(s|\pi_i(c_t;\cdot), P_\star^{c_t})}$
    define $\widehat{\mathcal{M}}_k(c_t) = (S, A, P_\star^{c_t}, \hat{r}_k^{c_t}, s_0, H)$
    compute $\pi_k(c_t; \cdot) \in \arg\max_{\pi \in S \to A} V^\pi_{\widehat{\mathcal{M}}_k(c_t)}(s_0)$ using a planning algorithm
  play $\pi_t(c_t; \cdot)$ and update oracle using $\sigma^t = (c_t, s_0^t, a_0^t, r_0^t, s_1^t, \ldots, s_{H-1}^t, a_{H-1}^t, r_{H-1}^t, s_H^t)$

Analysis Outline. Our analysis consists of four main steps.

Step 1: Establish uniform convergence bound over any $t \geq 2$ and a fixed sequence of functions $f_2, f_3, \ldots \in \mathcal{F}$ (see Lemma 16). The bound implies for the least square minimizers sequence $\{\hat{f}_t\}_{t=|A|+1}^T$ and any $\delta \in (0, 1)$, that with probability at least $1 - \delta/2$ for all $t \geq 2$ it holds that

$$\sum_{i=1}^{t-1} \sum_{h=0}^{H-1} \mathop{\mathbb{E}}_{c_i, s_h^i, a_h^i} \left[(\hat{f}_t(c_i, s_h^i, a_h^i) - f_\star(c_i, s_h^i, a_h^i))^2 | \mathbb{H}_{i-1} \right] \leq 68H \log(4|\mathcal{F}|t^3/\delta).$$

Step 2: Constructing a confidence bound over the value of any given policy w.r.t the true rewards function $f_\star$ and the least square minimizer at round t, $\hat{f}_t$. The confidence bound holds with high probability, in expectation over the contexts (i.e., "optimism in expectation"). In Lemma 18 we show that w.p. at least $1 - \delta/2$ for all $t > |A|$ and any policy $\pi \in \Pi_\mathcal{C}$ it holds that

$$\left| \mathbb{E}_c \left[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0) \right] - \mathbb{E}_c \left[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{f}_t, P_\star)}(c)}(s_0) \right] \right| \leq \sqrt{\phi_t(\pi)} \cdot \sqrt{68H \log(4|\mathcal{F}|t^3/\delta)},$$

where $\phi_t(\pi) := \mathbb{E}_c \left[\sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \frac{q_h(s_h|\pi(c;\cdot), P_\star^c)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)] q_h(s_h|\pi_i(c;\cdot), P_\star^c)} \right]$ is the contextual potential of π at round t, $\mathcal{M}^{(f,P_\star)}(c) = (S, A, P_\star^c, f(c, \cdot, \cdot), s_0, H)$ for any $f \in \mathcal{F}$. π_i is the selected policy at round i.

Step 3: Relax the confidence bound of step 2 to be additive. In Lemma 19 we show that with probability at least $1 - \delta/2$, for all $t > |A|$ and any policy $\pi \in \Pi_\mathcal{C}$ for $\beta_t = \sqrt{\frac{17t \log(4|\mathcal{F}|t^3/\delta)}{|S||A|}}$ it holds that

$$\left| \mathbb{E}_c \left[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0) \right] - \mathbb{E}_c \left[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{f}_t, P_\star)}(c)}(s_0) \right] \right| \leq \beta_t \cdot \phi_t(\pi) + \beta_t \cdot \frac{H\,|S|\,|A|}{t}.$$

Step 4: Bound the cumulative contextual potential ϕ_t over every round $t \geq |A| + 1$.
In Lemma 21 we show that for any sequence of played policies $\{\pi_t \in \Pi_\mathcal{C}\}_{t=1}^T$ it holds that $\sum_{t=|A|+1}^T \phi_t(\pi_t) \leq \frac{|S||A|}{p_{min}}(1 + \log(T/|A|))$. In addition, if for all t, $\pi_t = \pi$, for any $\pi \in \Pi_c$, we have an improved bound where the $|S|$ factor is replaced with H factor.

By combining all the steps and applying Azuma's inequality, we obtain the following regret bound. (Also, see Theorem 25 and 24, and Corollary 26).

Theorem 2 (regret bound) *For any $T > |A|$, finite function class $\mathcal{F}$ and $\delta \in (0, 1)$ let $\beta_t = \sqrt{\frac{17 \log(4|\mathcal{F}|t^3/\delta)t}{|S||A|}}$ for all $t \in [T]$. Then, with probability at least $1 - \delta$ the following holds.*

$$\textit{Regret}_T(\textit{RM-KD}) \leq \tilde{O}\left(\max\left\{ \frac{1}{p_{min}}, H \right\} \sqrt{T|S||A| \log(|\mathcal{F}|/\delta)} \right).$$

Corollary 3 (regret bound in terms of $\mathcal{G}$) *Under the same conditions of Theorem 2, with probability at least $1-\delta$ it holds that Regret$_T$(RM-KD) $\leq \tilde{O}\left(\max\left\{\frac{1}{p_{min}}, H\right\}|S|\sqrt{T|A|\log(|\mathcal{G}|/\delta)}\right)$.*

We remark that in all of our algorithms, for $T \leq |A|$ the regret is bounded trivially by $|A|H$.

4.1 Main Technical Challenges and Our Technique

Following steps 2 and 3, a natural "optimistic in expectation" strategy is to select at round t

$$\pi_t \in \arg\max_{\pi\in\Pi_{\mathcal{C}}}\left\{\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{f}_t,P_\star)}(c)}(s_0)] + \beta_t\cdot\phi_t(\pi)\right\} = \arg\max_{\pi\in\Pi_{\mathcal{C}}}\left\{\mathbb{E}_c[V^{\pi(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)]\right\}.$$

This approach has an obvious three major drawbacks:
(1) The distribution over the contexts, $\mathcal{D}$, is unknown. Hence, we can not compute $\mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)\right]$, for any policy π.
(2) Even when $\mathcal{D}$ is known, computing $\pi_t \in \Pi_{\mathcal{C}}$ is intractable when the context space $\mathcal{C}$ is large.
(3) The representation of a context-dependent policy π_t scales with the size of the context space $|\mathcal{C}|$, which can be huge.

We overcome those hurdles using two observations. The first observation is that

$$\max_{\pi\in\Pi_{\mathcal{C}}}\left\{\mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)\right]\right\} = \mathbb{E}_c\left[\max_{\pi(c;\cdot)\in S\to A}\{V^{\pi(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)\}\right].$$

We conclude that to compute a context-dependent policy $\pi_t \in \Pi_{\mathcal{C}}$ which maximizing LHS, we can compute for each context $c \in \mathcal{C}$ separately, a policy $\pi_t(c;\cdot) : S \to A$ that is optimal for $\widehat{\mathcal{M}}_t(c)$. For each context $c \in \mathcal{C}$ separately, solving the maximization problem in RHS can be done efficiently using a standard planning algorithm.

The second observation is that in every round t, we do not have to know the full representation of π_k, for all $k \leq t$, but only the mappings $\{\pi_k(c_i;\cdot)\}_{k=1}^t$ for the observed contexts $\{c_i\}_{i=1}^t$.

By taking an advantage of those two observations, at every round $t \leq T$, we compute $\pi_k(c_t;\cdot)$ for all $k \leq t$, which can be done in $poly(|S|,|A|,H,t)$ using a planning algorithm. This gives us an efficient algorithm which does not depend on $|\mathcal{C}|$.

5. Unknown and Context-Independent Dynamics

In this section, we assume the dynamics is unknown to the learner, but is independent of the context. Meaning, for all $c \in \mathcal{C}$, $P^c_\star = P_\star$. In addition, we assume the learner knows the (context-independent) partition of the states space to layers,$S = \{S_0,\ldots,S_H\}$, and the minimum reachability parameter.

Algorithm overview. Similarly to Algorithm RM-KD (Algorithms 1 or 4), we define an optimistic-in-expectation rewards function, but, since the dynamics is unknown, we replace $q(s|\pi_i(c_t;\cdot),P^{c_t}_\star)$ with its lower bound p_{min}, and clipping the rewards function to $[0,2]$. Denote by $N_t(s,a)$ and $N_t(s,a,s')$ the number of visits to (s,a) and (s,a,s'), respectively, up to round t. To approximate the dynamics, we use a tabular approximation and maintain the following confidence bounds over it $\xi_t(s,a) = \sqrt{\frac{|S|+1\log(4|S||A|T^2/\delta)}{\max\{1,N_t(s,a)\}}}$ for all $(s,a) \in S\times A$.

At round t, we compute an optimistic model w.r.t the rewards function $\hat{r}^{c_t}_t$ and a deterministic policy $\pi_t(c_t;\cdot)$ using Algorithm FOA (see Algorithm 7). Algorithm FOA gets a rewards function as an input, and compute the dynamics and deterministic policy which maximize the value of the resulting MDP, under the constraints that the dynamics is within the confidence interval. We remark that the resulting optimistic approximated dynamics is context-dependent, since it was computed w.r.t the context-dependent approximated rewards function. For more details, see Appendix C, Subsection C.2.

Analysis overview. We construct confidence intervals for both the dynamics and rewards. For the analysis, we define an intermediate CMDP where for all $t > |A|$ and context $c \in \mathcal{C}$, $\mathcal{M}^{(\hat{r}_t,P_\star)}(c) = (S,A,P_\star,\hat{r}^c_t,s_0,H)$. Let $\pi^\star \in \Pi_{\mathcal{C}}$ be an optimal policy of the true CMDP.

Algorithm 2 (sketch) Regret Minimization for Unknown Context Independent Dynamics (RM-UCID)

for round $t > |A|$ **do**
  compute $\hat{f}_t \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{t-1} \sum_{h=0}^{H-1} (f(c_i, s_h^i, a_h^i) - r_h^i)^2$ using the LSR oracle
  compute the empirical model for all $(s, a, s') \in S \times A \times S$: $\bar{P}_k(s'|s,a) = \frac{N_k(s,a,s')}{\max\{1, N_k(s,a)\}}$
  observe context $c_t \in \mathcal{C}$
  for $k = |A|+1, \ldots, t$ **do**
    compute $\forall (s,a) \in S \times A$: $\hat{r}_k^{c_t}(s,a) = \hat{f}_k(c_t, s, a) + \min\left\{ \frac{\beta_k}{p_{min} \sum_{i=1}^{k-1} \mathbb{I}[a = \pi_i(c_t; s)]}, 1 \right\}$
    find optimistic model $\widehat{\mathcal{M}}_k(c_t) = (S, A, \hat{P}_k^{c_t}, \hat{r}_k^{c_t}, s_0, H)$ and policy $\pi_k(c_t; \cdot)$ using Algorithm FOA
  play $\pi_t(c_t; \cdot)$ and observe trajectory σ^t. Update counters and LSR oracle using σ^t

Analysing the error caused by the rewards approximation. Similar to the analysis for the known dynamics (Section 4), we show in Lemma 38 that with high probability, for all $t > |A|$, and any policy $\pi \in \Pi_{\mathcal{C}}$ the followings hold. (1) $\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(\hat{r}_t, P_\star)(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] \leq 2\beta_t \cdot \psi_t(\pi) + \beta_t \cdot \frac{H|S||A|}{t}$.
(2) $\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(\hat{r}_t, P_\star)(c)}(s_0)] \leq \beta_t \cdot \psi_t(\pi) + \beta_t \cdot \frac{H|S||A|}{t}$,
where $\psi_t(\pi) := \mathbb{E}_c\left[\sum_{h=0}^{H-1} \sum_{s_h \in S_h} \frac{q_h(s_h, \pi(c;s_h)|\pi(c;\cdot), P_\star)}{p_{min} \sum_{i=1}^{t-1} \mathbb{I}[\pi(c;s_h) = \pi_i(c;s_h)]}\right]$ is the contextual potential at round t.

Analysing the error caused by the dynamics approximation. We show that with high probability the following good event hold. For all $t > |A|$ and $(s,a) \in S \times A$, we have $\|\bar{P}_t(\cdot|s,a) - P_\star(\cdot|s,a)\|_1 \leq \xi_t(s,a)$. In Lemma 32 we show that under the good event, our optimistic approximated model $\widehat{\mathcal{M}}_t(c) = (S, A, \widehat{P}_t^c, \hat{r}_t^c, s_0, H)$ and the selected policy $\pi_t(c;\cdot)$ satisfy for all $c \in \mathcal{C}$ and $t > |A|$ that $V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0) \geq V^{\pi^\star(c;\cdot)}_{\mathcal{M}(\hat{r}_t^c, P_\star)}(s_0)$. When combining that with the confidence intervals over the rewards, we obtain (see Lemma 43), for all $t > |A|$, that

$$\mathbb{E}_c\left[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0)\right] \leq \mathbb{E}_c\left[V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)\right] + \beta_t \cdot \psi_t(\pi^\star) + \beta_t \cdot \frac{H|S||A|}{t}.$$

Moreover, Lemma 35 shows that under that good event, with high probability we have

$$\sum_{t=|A|+1}^{T} \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(\hat{r}_t, P_\star)(c)}] \leq 4H\sqrt{|S| + 2\log(4|S||A|T^2/\delta)}\left(\sqrt{2TH\log(8/\delta)} + 2\sqrt{|S||A|T}\right).$$

Lastly, we bound $\sum_{t=|A|+1}^{T} \psi_t(\pi^\star)$ and $\sum_{t=|A|+1}^{T} \psi_t(\pi_t)$ in Appendix C. By combining all the above, and applying Azuma's inequality, we obtain the following regret bound (also see Theorems 45 and 44, and Corollary 46).

Theorem 4 (regret bound) *For any $T > |A|$, finite function class $\mathcal{F}$ and $\delta \in (0,1)$ for the choice of $\beta_t = \sqrt{\frac{17\log(8|\mathcal{F}|t^3/\delta)t}{|S||A|}}$ for all $t \in [T]$, with probability at least $1-\delta$ the following holds.*

$$\textit{Regret}_T(\textit{RM-UCID}) \leq \tilde{O}\left(H|S|\sqrt{T|A|}\log(1/\delta) \;+\; \max\{H, 1/p_{min}\}\sqrt{\log(|\mathcal{F}|/\delta)T|S||A|}\right).$$

Corollary 5 (regret bound in terms of $\mathcal{G}$) *Under the same conditions of Theorem 4, with probability at least $1-\delta$ it holds that $\textit{Regret}_T(\textit{RM-UCID}) \leq \tilde{O}\left(\max\left\{\frac{1}{p_{min}}, H\right\}|S|\sqrt{T|A|\log(|\mathcal{G}|/\delta)}\right)$.*

6. Unknown and Context-Dependent Dynamics

In this section, we consider the most challenging case, where the dynamics is unknown and context-dependent. We assume an access to a finite function class $\mathcal{P} \subseteq (S \times (S \times A \times \mathcal{C}) \to [0,1])$, for which every function $P \in \mathcal{P}$ satisfies

$\sum_{s' \in S} P(s'|s,a,c) = 1, \quad \forall c \in \mathcal{C},\ \forall (s,a) \in S \times A$. We use $\mathcal{P}$ to approximate the context-dependent dynamics under the dynamics realizability assumption (see Assumption 2).

Algorithm outline. In Algorithm RM-UCDD (see Algorithms 3 or 8), we approximate both the rewards and the dynamics using LSR oracle. The first $|A|$ rounds are initialization rounds, as before. At round $t > |A|$, we compute the approximated rewards function for the context c_t as before, but for the dynamics, we use the least square minimizer $\widehat{P}_t$. We define the approximate model for c_t, compute an optimal policy for it $\pi_t(c_t;\cdot)$ and run it to generate a trajectory and update the oracles. We feed the LSR oracle for the dynamics with samples of the form $((c_t, s_h^t, a_h^t, s'), \mathbb{I}[s' = s_{h+1}^t])$ for all $t \in [T]$, $h \in [H-1]$ for every $s' \in S$.

Algorithm 3 Regret Minimization for CMDP with Unknown Context-Dependent Dynamics

inputs: MDP parameters: S, A, H, s_0. Confidence $\delta > 0$ and tuning parameters $\{\beta_t\}_{t=1}^T$. Minimum reachability parameter $p_{min} > 0$.
Initialization: in round $i \leq |A|$, run policy $\pi_i(c;s) := a_i$
for round $t = |A|+1, \ldots, T$ **do**
  compute $\hat{f}_t \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{t-1} \sum_{h=0}^{H-1} (f(c_i, s_h^i, a_h^i) - r_h^i)^2$ using the LSR oracle, and
  $\widehat{P}_t \in \arg\min_{\tilde{P} \in \mathcal{P}} \sum_{i=1}^{t-1} \sum_{h=0}^{H-1} \sum_{s' \in S} (\tilde{P}^{c_i}(s'|s_h^i, a_h^i) - \mathbb{I}[s' = s_{h+1}^i])^2$ using LSR oracle
  observe context $c_t \in \mathcal{C}$.
  for $k = |A|+1, |A|+2, \ldots, t$ **do**
    compute $\forall (s,a) \in S \times A\ \ \hat{r}_k^{c_t}(s,a) = \hat{f}_k(c_t, s, a) + \min\left\{ \frac{1}{p_{min}} \frac{\beta_k}{\sum_{i=1}^{k-1} \mathbb{I}[a = \pi_i(c_t;s)]}, 1 \right\}$
    define $\widehat{\mathcal{M}}_k(c_t) = (S, A, \widehat{P}_k^{c_t}, \hat{r}_k^{c_t}, s_0, H)$
    compute $\pi_k(c_t, \cdot) \in \arg\max_{\pi \in S \to A} V^{\pi}_{\widehat{\mathcal{M}}_k(c_t)}(s_0)$ using planing algorithm
  play $\pi_t(c_t, \cdot)$, observe trajectory σ^t and update oracles

Analysis outline. In the analysis, we define the following intermediate MDPs for any context $c \in \mathcal{C}$: (1) $\mathcal{M}^{(\hat{r}_t, P)}(c) = (S, A, P^c, \hat{r}_t^c, s_0, H)$ for context-dependent dynamics $P \in \mathcal{P}$, where $\hat{r}_t^c$ is the approximated rewards function in round t, which defined in Algorithm 3. By definition, $\widehat{\mathcal{M}}_t(c) = \mathcal{M}^{(\hat{r}_t, \widehat{P}_t)}(c)$. (2) Let $\mathcal{M}^{(f, P_\star)}(c) = (S, A, P_\star^c, f(c,\cdot,\cdot), s_0, H)$ for any function $f \in \mathcal{F}$, where $P_\star^c$ is the true dynamics associated with the context c. By definition, $\mathcal{M}(c) = \mathcal{M}^{(f_\star, P_\star)}(c)$. In addition, we use the notations of ϕ_t and ψ_t defined in previous sections.

Analysing the error caused by the rewards approximation. Similar the known dynamics setting (see Section 4), we show (see Lemma 51) that with probability at least $1 - \delta/4$, for all $t > |A|$ and $\pi \in \Pi_{\mathcal{C}}$ the following holds.
(1) $\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t, P_\star)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] \leq \beta_t \left(2\psi_t(\pi) + \frac{H|S||A|}{t}\right)$,
(2) $\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t, P_\star)}(c)}(s_0)] \leq \beta_t \left(\psi_t(\pi) + \frac{H|S||A|}{t}\right)$.

Analysing the error caused by the dynamics approximation.

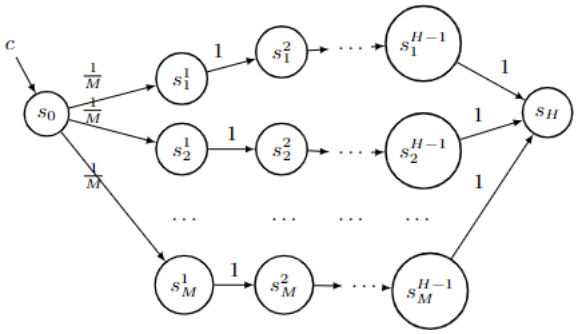
Key observation. Let $\mathcal{B}$ be a random variable which generates the next state s_{h+1} given the true dynamics associated with c, $P_\star^c$, the state s_h and the action a_h. $\mathcal{B}$ is defined as $\mathcal{B}(P_\star^c, s_h, a_h) \sim P_\star^c(\cdot|s_h, a_h)$. Our observation is that since the CMDP is layered, given the context c_t state s_h^t and action a_h^t, we have that the random variables $\mathcal{B}(P_\star^{c_t}, s_h^t, a_h^t)$ and $(s_0^t, a_0^t, s_1^t, \ldots, s_{h-1}^t, a_{h-1}^t)$ are independent random variables. Using that observation, we are able to extend our uniform convergence bound to the dynamics approximation. Hence, we can apply the four steps strategy above for the dynamics approximation as well.

Step 1: Establish uniform convergence bound over any $t \geq 2$ and a fixed sequence of functions $P_2, P_3, \ldots \in \mathcal{P}$ (see Lemma 57). The bound implies that for the least square minimizers sequence $\{\widehat{P}_t\}_{t=|A|+1}^T$ with high probability, for all $t > |A|$, the following holds,

$$\sum_{i=1}^{t-1} \sum_{h=0}^{H-1} \mathop{\mathbb{E}}_{c_i, s_h^i, a_h^i} \left[\sum_{s' \in S} (\widehat{P}_t^{c_i}(s'|s_h^i, a_h^i) - P_\star^{c_i}(s'|s_h^i, a_h^i))^2 | \mathbb{H}_{i-1} \right] \leq 72H|S| \log(8|\mathcal{P}|t^3/\delta).$$

Step 2: Constructing a confidence bound over the value of any given policy w.r.t the approximated and true dynamics, where the rewards function is the approximated rewards at round t. The confidence bound holds with high probability,

Figure 1: Lower bound illustration



in expectation over the contexts (i.e., "optimism in expectation"). In Lemma 58 we show that with probability at least $1-\delta/4$ for all $t > |A|$ and any policy $\pi \in \Pi_{\mathcal{C}}$ it holds that

$$\left|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t,P_\star)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t,\hat{P}_t)}(c)}(s_0)]\right| \leq 2\sqrt{H|S|\phi_t(\pi)} \cdot \sqrt{72H^2|S|\log(8|\mathcal{P}|t^3/\delta)}.$$

Step 3: Relax the confidence bound in step 2 to be additive. In Lemma 59 we show that with high probability, for all $t > |A|$ and any policy $\pi \in \Pi_{\mathcal{C}}$ for $\gamma_t = \sqrt{\frac{72t\log(8|\mathcal{P}|t^3/\delta)}{|S||A|}}$ it holds that

$$\left|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t,P_\star)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t,\hat{P}_t)}(c)}(s_0)]\right| \leq \gamma_t H|S|\phi_t(\pi) + \gamma_t\frac{H^2|S|^2|A|}{t}.$$

Step 4: Bounding the sum of contextual potentials similarly to shown for the rewards.

Using all the above, we obtain the optimism lemma (Lemma 61) which states that under the good events of step 2 for both the dynamics and rewards approximation, for all $t > |A|$ it hold that

$$\mathbb{E}_c[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)] \leq H|S|\gamma_t \cdot \phi_t(\pi^\star) + \gamma_t\frac{H^2|S|^2|A|}{t} + \beta_t \cdot \psi_t(\pi^\star) + \beta_t\frac{H|S||A|}{t},$$

yielding the following regret bound. (Also, see Theorems 63 and 62, and Corollary 64)

Theorem 6 (regret bound) *For any $\delta \in (0,1)$, $T > |A|$ and finite function classes $\mathcal{F}$ and $\mathcal{P}$, for the choice of $\beta_t = \sqrt{\frac{17t\log(8|\mathcal{F}|t^3/\delta)}{|S||A|}}$ and $\gamma_t = \sqrt{\frac{72t\log(8|\mathcal{P}|t^3/\delta)}{|S||A|}}$ for all $t \in [T]$, with probability at least $1-\delta$ the following holds.*

$$\textit{Regret}_T(\textit{RM-UCDD}) \leq \tilde{O}\left(\max\{H, 1/p_{min}\}H|S|^{3/2}\sqrt{|A|T\log(\max\{|\mathcal{F}|,|\mathcal{P}|\}/\delta)}\right).$$

Corollary 7 (regret bound in terms of $\mathcal{G}$) *Under the same conditions of Theorem 6, with probability at least $1-\delta$ it holds that $\textit{Regret}_T(\textit{RM-UCDD}) \leq \tilde{O}\left(\max\{H, 1/p_{min}\}H|S|^{3/2}\sqrt{|A|T\log(\max\{|\mathcal{G}|,|\mathcal{P}|\}/\delta)}\right)$.*

7. Lower bound

We present a lower bound for layered CMDP, where the dynamics is known and context-independent, which based on the lower bound for CMAB presented by Agarwal et al. (2012), in which $K = |A|$, $\mathcal{G} \subseteq (\mathcal{C} \times A \to [0,1])$ and $N \in \mathbb{N}$.

Theorem 8 (Theorem 5.1, Agarwal et al. (2012)) *For every N and K such that $\ln N/\ln K \leq T$, and every algorithm $\mathfrak{A}$, there exist a functions class $\mathcal{G}$ of cardinality at most N and a distribution $D(c,r)$ for which the realizability assumption holds, but the expected regret of $\mathfrak{A}$ is $\Omega(\sqrt{KT\ln N/\ln K})$.*

In the following we present a lower bound for horizon $H \geq 2$.

Theorem 9 (Lower bound for CMDP) *Let $\delta \in (0,1)$, horizon $H \geq 2$ and $M, N \in \mathbb{N}$.*
Let $T \geq 8M \log \frac{|S|}{\delta} + 2M \ln N / \ln |A|$ and consider a CMDP $(\mathcal{C}, S, A, \mathcal{M})$ for which $|S| = M \cdot (H-1) + 2$.
Then, for any algorithm $\mathfrak{A}$, there exist a base function class $\mathcal{G} \subseteq (\mathcal{C} \times A \to [0,1])$ of cardinality at most N and a distribution $D(c, s, a, r)$ for which the realizability assumption holds for $\mathcal{F} = \mathcal{G}^S$, and, with probability at least $1-\delta$, the expected regret of $\mathfrak{A}$ is $\Omega\left(\sqrt{TH|S||A|\ln(N)/\ln(|A|)}\right)$.

proof idea. Solving the CMDP illustrated in Figure 1 is equivalent to solving $M(H-1)+1$ CMAB problems. Hence, the theorem follows by Theorem 8. (See Appendix E.)

8. Discussion

To the best of our knowledge, this is the first work that obtains sub-liner regret bounds using general function approximation (i.e., without additional structural assumption regarding the CMDP or the functions that use for approximation) and to present an expected regret lower bound. Our results can be extended to infinite function classes using covering numbers analysis (see Shalev-Shwartz and Ben-David (2014)). We leave that for future work. Our algorithms has $poly(|S|, |A|, H, T)$ running time and space complexity, assuming an efficient least-square regression oracle.

The main advantages of our technique.
(1) We present a novel confidence interval for general function approximation, which requires no additional knowledge regarding the function class.
(2) Our algorithms do no fully represent the selected context-dependent policy at each time step, as the representation of it scales linearly in the context space size $|\mathcal{C}|$, which can be huge, but rather compute it only for the sequence of observed contexts.
(3) To the best of our knowledge, our approach is the first application of the optimism in face of uncertainty principle to CMDPs using general function approximation. We believe the our optimistic approach can be extend to other related setting such as MDPs with rich observations or large states space.

Tightness of our bounds. For comparison, consider our regret upper bounds in terms of the base class $\mathcal{G}$ cardinality. *Known context-dependent* and *Unknown context-independent* dynamics: $\widetilde{O}\left(\max\{H, 1/p_{min}\}\,|S|\sqrt{T|A|\log(|\mathcal{G}|/\delta)}\right)$. *Unknown context-dependent* dynamics: $\widetilde{O}\left(\max\{H, 1/p_{min}\}H|S|^{3/2}\sqrt{|A|T\log(\max\{|\mathcal{G}|, |\mathcal{P}|\}/\delta)}\right)$. Recall our lower bound is $\Omega\left(\sqrt{TH|S||A|\ln(|\mathcal{G}|)/\ln(|A|)}\right)$. While our dependency in T, $|A|$ and $|\mathcal{G}|$ is near-optimal, bridging the gap, in p_{min} especially, is an important open question.

Following the above, the main limitation of our approach is the minimum reachability assumption. It allows us to limit the exploration-exploitation trade-off only to the actions selection, since any state is reached with probability at least $p_{min} > 0$. Hence, we avoid the need to explicitly explore the dynamics. In particular, in Algorithm RM-UCDD we avoid the need to add bonus that encourages dynamics exploration. Nevertheless, the exploration task is far from trivial, due to the context-dependent environment and the lack any additional knowledge regarding the function class. An interesting direction for future research is to obtain $\widetilde{O}(\sqrt{T})$ regret bound, without our reachability assumption.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), by the Israel Science Foundation(grant number 993/17), Tel Aviv University Center for AI and Data Science (TAD), and the Yandex Initiative for Machine Learning at Tel Aviv University.

References

Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert Schapire. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, pages 19–26. PMLR, 2012.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014.

Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar. Reinforcement learning of contextual mdps using spectral methods. *arXiv preprint arXiv:1611.03907*, 2016.

Peter Bartlett, Varsha Dani, Thomas Hayes, Sham Kakade, Alexander Rakhlin, and Ambuj Tewari. High-probability regret bounds for bandit online linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory-COLT 2008*, pages 335–342. Omnipress, 2008.

Yonathan Efroni, Lior Shani, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. *arXiv preprint arXiv:2002.08243*, 2020.

Hamid Eghbal-zadeh, Florian Henkel, and Gerhard Widmer. Learning to infer unseen contexts in causal contextual reinforcement learning. In *Self-Supervision for Reinforcement Learning Workshop - ICLR 2021*, 2021a.

Hamid Eghbal-zadeh, Florian Henkel, and Gerhard Widmer. Context-adaptive reinforcement learning using unsupervised learning of context variables. In *NeurIPS 2020 Workshop on Pre-registration in Machine Learning*, pages 236–254. PMLR, 2021b.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

Dylan Foster, Alekh Agarwal, Miroslav Dudik, Haipeng Luo, and Robert Schapire. Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 1539–1548. PMLR, 2018.

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

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

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

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Contextual-mdps for pac-reinforcement learning with rich observations. *CoRR*, abs/1602.02722, 2016.

Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, and Shie Mannor. RL for latent mdps: Regret guarantees and a lower bound. In *NeurIPS*, 2021.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Orin Levy and Yishay Mansour. Learning efficiently function approximation for contextual mdp. *arXiv preprint arXiv:2203.00995*, 2022.

Aditya Modi and Ambuj Tewari. No-regret exploration in contextual reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, pages 829–838. PMLR, 2020.

Aditya Modi, Nan Jiang, Satinder Singh, and Ambuj Tewari. Markov decision processes with continuous side information. In *Algorithmic Learning Theory*, pages 597–618. PMLR, 2018.

Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, pages 5478–5486. PMLR, 2019.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 2021.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Found. Trends Mach. Learn.*, 12(1-2):1–286, 2019.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

Yunbei Xu and Assaf Zeevi. Upper counterfactual confidence bounds: a new optimism principle for contextual bandits. *arXiv preprint arXiv:2007.07876*, 2020.

Yan Zhang and Michael M Zavlanos. Transfer reinforcement learning under unobserved contextual information. In *2020 ACM/IEEE 11th International Conference on Cyber-Physical Systems (ICCPS)*, pages 75–86. IEEE, 2020.

Appendix A. Extended Related Work

Contextual Reinforcement Leaning. CMDP was first introduce by Hallak et al. (2015). Modi et al. (2018) gives a general framework for deriving generalization bounds as a function of the covering number for smooth CMDPs and contextual linear combination of d MDPs, where d is a finite number of MDPs, and the context-space is the $d-1$ probability simplex. Modi and Tewari (2020) give a regret analysis for Generalized Linear Models (GLMs). Our function approximation framework is much more general than GLM and smooth CMDP.

Foster et al. (2021) present new statistical complexity measure, the decision-estimation coefficient, for interactive decision making. They show an application of it to obtain regret upper bound for Contextual RL. They assume an access to an online estimation oracle with regret guarantees, denote it **Est**. Their reference model class is the class of all contextual MDPs $\mathcal{M}$ and randomized policies Π. Given the observations and played policies up to the current time step, the online estimation oracle returns an estimated CMDP. Given the current context, they use it to compute a distribution over policies, and sample the played policy from it. They obtain $\tilde{O}(\sqrt{T \cdot \textbf{Est}})$ regret.

The main disadvantages of their approach are that their online estimation oracle is very strong and might be computationally inefficient. It is unclear whether their algorithmic results can be extended to support offline oracles for estimation. In addition, the sample complexity of such an oracle is unclear. Moreover, the relation between their new complexity measure and known complexity measures for function approximation (i.e, VC/Pseudo/Fat-shattering/Natrajan dimension) that are commonly-used in offline supervised learning is unclear. Their results are very general and capture many RL settings. In contrast, we use a standard and efficient offline least square oracle to build an approximated optimistic CMDP from scratch. Hence, we have much refined assumptions which allows us to use standard offline supervised learning tolls.

Jiang et al. (2017) consider Contextual Decision Processes (CDP) with low Bellman rank, and present OLIVE, which is sample efficient for CDPs with a small Bellman rank. We do not make any assumptions on the Bellman rank.

Levy and Mansour (2022) consider learning CMDPs using function approximation. They assume an access to an ERM oracle and derive sample complexity bounds to compute ϵ-optimal policy under four different settings: where the dynamics known and unknown, and context dependent or context-free. They assume no additional structural assumption regarding the CMDP. Their method provide the first general and efficient reduction from CMDP to offline supervised learning. However, their sample complexity bounds are not optimal, as our lower bound shows. We, in contrast, consider online learning problem where the goal is regret minimization. We obtain $\sqrt{T}$ regret under the minimum reachability assumption while Levy and Mansour (2022) has no such an assumption but obtain higher dependence on T.

Contextual Bandits. Contextual bandits (CMAB) are a natural extension of the Multi-Arm Bandit (MAB), augmented by a context which influences the rewards Slivkins (2019); Lattimore and Szepesvári (2020). Agarwal et al. (2014) use efficiently an optimization oracle to derive an optimal regret bound. Regression based approaches appear in Agarwal et al. (2012); Foster et al. (2018); Foster and Rakhlin (2020); Simchi-Levi and Xu (2021). We differ from CMAB, since our main challenge is the dynamics, and the need to optimize future rewards, which is the case in most RL settings.

Xu and Zeevi (2020) present the first optimistic algorithm for CMAB. They assume an access to a least-square regression oracle and achieve $\tilde{O}(\sqrt{T|A|\log|\mathcal{F}|})$ regret, where $\mathcal{F}$ is a finite and realizable function class uses to approximate the rewards. They also show a result for infinite function class using covering numbers analysis. We adapt their approach and extend it to CMDP.

Inverse Gap Minimization (IGM) technique. Foster and Rakhlin (2020); Simchi-Levi and Xu (2021) apply the IGM technique to CMAB and obtain $\tilde{O}(\sqrt{T|A|})$ regret, assuming an access to an online/offline least square regression oracle, respectively. However, we do not see any straight-forward extension of their approach to CMDP which is both computationally efficient and has an optimal regret, under the same least-square oracle assumption (even when the dynamics is known to the learner). In more detail, consider the following application of IGM to CMDP. At time t: (1) Compute an approximation for the context-dependent rewards function $\hat{f}_t$. (2) Observe a context c_t and choose policy π_t according to a distribution over deterministic policies. The probability of a policy π is proportional to $1/(V_t^{\pi_t^\star}(s_0) - V_t^{\pi}(s_0))$, where $\pi_t^\star$ is the optimal policy for the rewards $\hat{f}_t$. (3) Experience trajectory and update the function approximation. Using the analysis of Foster and Rakhlin (2020); Simchi-Levi and Xu (2021) one can obtain $\tilde{O}(\sqrt{T \cdot 2^{|S||A|}})$ regret and it is computationally inefficient. A similar approach using the Q function will also similarly fail. Consider at time t selecting a stochastic policy such that $\pi_t(a|s)$ is proportional to $1/(Q_t^{\pi_t}(s, \pi_t^\star(s)) - Q_t^{\pi_t}(s,a))$.

This attempt will fail due to the lack of optimism and the changing-per-context optimal policy. Foster et al. (2021) apply IGM to CMDP and obtain optimal regret. However they use the strong online estimation oracle discussed above.

Additional works. There are works that considered the case were the contexts are unobservable Kwon et al. (2021); Eghbal-zadeh et al. (2021a,b), latent states Krishnamurthy et al. (2016), spectral methods Azizzadenesheli et al. (2016), transfer learning Zhang and Zavlanos (2020), and more. All those issue are somewhat unrelated to our main focus.

Appendix B. Known Dynamics

B.1 Assumptions

Known context-dependent dynamics. We assume that there is a mapping $P_\star$ from context c to a dynamics $P^c_\star$. In the known dynamics case the learner knows this mapping.

Recall we assume that for each context c, the related MDP $\mathcal{M}(c)$ is layered and loop free. We remark that the partition to layers is context-dependent. Since the dynamics is known, the partition of the state space into layers can be easily computed for every context $c \in \mathcal{C}$. Hence, we assume that for each context, the learner also knows the partition to layers, which naturally reveals to her when the dynamics is known.

For every context c, we denote by $S^c_0, S^c_1, \ldots, S^c_{H-1}, S^c_H$ the disjoint layers, and $\forall c \in \mathcal{C}.\ \ S = \bigcup_{h\in[H]} S^c_h$. Recall that for every context c we have that $S^c_0 = \{s_0\}$ and $S^c_H = \{s_H\}$, meaning there are unique start and final states, which are the same for all of the contexts.

Clearly, the above assumptions do not limit the generality of our results.

Recall the definition of a context-dependent policy.

Definition 10 (Context-dependent policy) *A stochastic context-dependent policy $\pi = (\pi(c;\cdot) : S \to \Delta(A))_{c\in\mathcal{C}}$ maps a context $c \in \mathcal{C}$ to a policy $\pi(c;\cdot) : S \to \Delta(A)$.*

A deterministic context-dependent policy $\pi = (\pi(c;\cdot) : S \to A))_{c\in\mathcal{C}}$ maps a context $c \in \mathcal{C}$ to a policy $\pi(c;\cdot) : S \to A$. Let $\Pi_\mathcal{C}$ denote the class of all deterministic context-dependent policies.

Minimum reachability. We assume that there exists $p_{min} \in (0, 1]$ such that for every context $c \in \mathcal{C}$, layer $h \in [H-1]$ and a state $s_h \in S^c_h$ for any context-dependent policy $\pi \in \Pi_\mathcal{C}$ it holds that $q_h(s_h|\pi(c;\cdot), P^c_\star) \geq p_{min}$. We remark that p_{min} is not (explicitly) given to the learner in this section.

Recall $q(s|\pi(c;\cdot), P^c_\star)$ denotes the probability of visiting state $s \in S$ when playing $\pi(c;\cdot)$, which is policy π defines for the context c, on the dynamics $P^c_\star$. When the dynamics is layered and loop-free, then $q(s|\pi(c;\cdot), P^c_\star) = q_h(s|\pi(c;\cdot), P^c_\star) \geq p_{min}$ if and only if $s \in S_h$.

Reward approximation using least-square regression. Our algorithms get as input a finite function class $\mathcal{F} \subseteq \mathcal{C} \times S \times A \to [0, 1]$. Each function $f \in \mathcal{F}$ maps a context $c \in \mathcal{C}$, state $s \in S$ and an action $a \in A$ to a reward $r \in [0, 1]$. We use $\mathcal{F}$ to approximate the rewards function using LSR oracle.

Assumption 3 (rewards realizability) *We assume that $\mathcal{F}$ is realizable, meaning, there exist a function $f_\star \in \mathcal{F}$ such that for every context $c \in \mathcal{C}$, state $s \in S$ and action $a \in A$,*

$$f_\star(c, s, a) = r^c_\star(s, a) = \mathbb{E}[R^c_\star(s, a)|c, s, a].$$

B.2 Main Technical Challenges of the Optimism in Expectation Approach

In Lemma 19 we will later present, we show that with high probability, for all $t > |A|$ it holds that

$$\begin{aligned}
&\left|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{f}_t,P_\star)(c)}(s_0)]\right| \\
\leq &\beta_t \cdot \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right] + \beta_t\frac{H\cdot|S|\cdot|A|}{t},
\end{aligned}$$

where $\widehat{f}_t$ is the approximated rewards function at time t, $\mathcal{M}^{(\widehat{f}_t,P_\star)}(c) = (S, A, P_\star^c, \widehat{f}_t(c,\cdot,\cdot), s_0, H)$, π_i is the context-dependent policy selected in round $i \in [t-1]$ and β_t is a parameter related to the function class $\mathcal{F}$.

Consider the approximated MDP $\widehat{\mathcal{M}}_t(c)$ defined in Algorithm RM-KD (Algorithm 4). For every context-dependent policy $\pi \in \Pi_{\mathcal{C}}$, consider the following explicit representation of the value function for any given context $c \in \mathcal{C}$.

$$\begin{aligned}
&V^{\pi(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0) = \sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|P_\star^c,\pi(c;\cdot))\cdot\widehat{r}_t^c(s_h,a_h) \\
&= \sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|P_\star^c,\pi(c;\cdot))\cdot\left(\widehat{f}_t(c,s_h,a_h) + \frac{\beta_t}{\sum_{i=1}^{t-1}\mathbb{I}[a_h=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right) \\
&= \sum_{h=0}^{H-1}\sum_{s_h\in S_h^c} q_h(s_h|P_\star^c,\pi(c;\cdot))\cdot\left(\widehat{f}_t(c,s_h,\pi(c;s_h)) + \frac{\beta_t}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right) \\
&= V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{f}_t,P_\star)(c)}(s_0) + \sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi(c;\cdot),P_\star^c)\cdot\beta_t}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}.
\end{aligned} \tag{1}$$

Hence, following the above, a natural "optimistic in expectation" approach is to choose at time t the policy

$$\begin{aligned}
\pi_t &\in \arg\max_{\pi\in\Pi_{\mathcal{C}}}\left\{\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{f}_t,P_\star)(c)}(s_0)] + \beta_t\cdot\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]\right\} \\
&= \arg\max_{\pi\in\Pi_{\mathcal{C}}}\{\mathbb{E}_c[V^{\pi(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)]\},
\end{aligned}$$

where the equality follows from the value function derivation in equation (1).

This approach has three major drawbacks.

1. The distribution over the context $\mathcal{D}$ is unknown. Hence, we can not compute $\mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)\right]$, for any policy π.
2. Even when $\mathcal{D}$ is known, computing $\pi_t \in \Pi_{\mathcal{C}}$ is intractable when the context space $\mathcal{C}$ is large.
3. The representation of a context-dependent policy π scales with the size of the context space $|\mathcal{C}|$, which can be huge.

We overcome those hurdles using two observations.

Observation 1: it holds that

$$\max_{\pi\in\Pi_{\mathcal{C}}}\left\{\mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)\right]\right\} = \mathbb{E}_c\left[\max_{\pi(c;\cdot)\in S\to A}\{V^{\pi(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)\}\right].$$

We conclude that to compute a context-dependent policy $\pi_t \in \Pi_{\mathcal{C}}$ which maximizing LHS, we can compute for each context $c \in \mathcal{C}$ separately, a policy $\pi_t(c;\cdot) : S \to A$ that is optimal for $\widehat{\mathcal{M}}_t(c)$. For each context $c \in \mathcal{C}$ separately, solving the maximization problem in RHS can be done efficiently using a standard planning algorithm.

Observation 2: in each round $t \leq T$, we do not have to know the full representation of π_k, for all $k \leq t$. Notice that at every round $t \leq T$, we only use the mappings $\{\pi_k\}_{k=1}^t$ for the contexts $\{c_k\}_{k=1}^t$.

By taking an advantage of those two observation, at every round $t \leq T$, we compute $\pi_k(c_t;\cdot)$ for all $k \leq t$, which can be done in $poly(|S|,|A|,H,t)$ using a planing algorithm. By that, we overcome the hardness of fully computing π_k, and avoid the memory-overload of saving $\{\pi_k\}_{k=1}^T$ full representation.

B.3 Regret Minimization Algorithm

For the known dynamics case, our algorithm works as follows.

Rounds $i = 1, 2, \ldots, |A|$ are initialization rounds, where for each action $a_i \in A$ in turn, the agent simply runs the policy π_i that always selects action a_i, for any observed context $c_i \in \mathcal{C}$ (and any visited state $s \in S$).

In rounds $t = |A|+1, \ldots, T$ the agent observe context $c_t \in \mathcal{C}$. Then the agent does the following computation. For every (previous) round $k = |A|+1, \ldots, t-1$ the agent computes the policy $\pi_k(c_t;\cdot)$ which is the optimal policy of $\widehat{\mathcal{M}}_k(c_t)$. I.e., this is the optimal policy for c_t with respect to the approximated model that uses the function $\widehat{f}_k$, which was computed in round $k \leq t$. The agent computes $\pi_k(c_t;\cdot)$ in the following manner.

(1) The agent computes the optimistic reward function of round k for the context c_t, using $\widehat{f}_k$ and the policies $\{\pi_i(c_t;\cdot)\}_{i=1}^{k-1}$. The rewards function is

$$\forall (s,a) \in S \times A: \quad \widehat{r}_k^{c_t}(s,a) = \widehat{f}_k(c_t,s,a) + \frac{\beta_k}{\sum_{i=1}^{k-1} \mathbb{I}[a=\pi_i(c_t;s)] q(s|\pi_i(c_t;\cdot),P_\star^{c_t})}.$$

We remark that since $P_\star^{c_t}$ is known, $q(s|\pi_i(c_t;\cdot),P_\star^{c_t})$ can also be computed in polynomial time in $|S|,|A|,H$, for all $s \in S$.

(2) The agent defines the approximated MDP for c_t at round k, $\widehat{\mathcal{M}}_k(c_t) = (S, A, P_\star^{c_t}, \widehat{r}_k^{c_t}, s_0, H)$.
(3) The agent computes $\pi_k(c_t,\cdot) \in \arg\max_{\pi \in S \to A} V^\pi_{\widehat{\mathcal{M}}_k(c_t)}(s_0)$ using a planing algorithm.

The agent computes $\pi_t(c_t;\cdot)$ in the same way, and then run it to generate trajectory σ^t. Lastly, the agent updates the least square regression (LSR) oracle using σ^t. (For more details, see Algorithm 4.)

Algorithm 4 Regret Minimization for CMDP with Known Dynamics (RM-KD)

inputs:
- MDP parameters: S , A, $P_\star$, s_0, H.
- Confidence parameter δ and tuning parameters $\{\beta_t\}_{t=1}^T$.

initialization: for the first $|A|$ rounds, for each action a_i in turn, run once the policy $\pi_i(c,s) = a_i$ that at any state s plays action a_i, regardless of the context c.
for round $t = |A|+1, \ldots, T$ **do**
  compute $\widehat{f}_t \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{t-1} \sum_{h=0}^{H-1} (f(c_i, s_h^i, a_h^i) - r_h^i)^2$ using the LSR oracle.
  observe context $c_t \in \mathcal{C}$.
  for $k = |A|+1, |A|+2, \ldots, t$ **do**
    compute $\forall (s,a) \in S \times A: \quad \widehat{r}_k^{c_t}(s,a) = \widehat{f}_k(c_t,s,a) + \frac{\beta_k}{\sum_{i=1}^{k-1} \mathbb{I}[a=\pi_i(c_t,s)] q(s|\pi_i(c_t,\cdot),P_\star^{c_t})}$.
    define $\widehat{\mathcal{M}}_k(c_t) = (S, A, P_\star^{c_t}, \widehat{r}_k^{c_t}, s_0, H)$.
    compute $\pi_k(c_t,\cdot) \in \arg\max_{\pi \in S \to A} V^\pi_{\widehat{\mathcal{M}}_k(c_t)}(s_0)$ using a planning algorithm.
  play $\pi_t(c_t,\cdot)$ and update oracle using $\sigma^t = (c_t, s_0^t, a_0^t, r_0^t, s_1^t, \ldots, s_{H-1}^t, a_{H-1}^t, r_{H-1}^t, s_H^t)$.

Observation 1 *For any $|A|+1 \leq t \leq T$, given the policy $\pi_t(c_t;\cdot)$, the trajectory σ^t is independent from the previous trajectories $\sigma^1, \ldots \sigma^{t-1}$.*

Observation 2 *Since the CMDP is layered, for every $c \in \mathcal{C}$, $h \in [H]$ and $s_h \in S_h^c$ we have $q_h(s_h|\pi_i(c;\cdot),P_\star^c) = q(s|\pi_i(c;\cdot),P_\star^c)$.*

Corollary 11 *Following Observation 2, the following two definitions of the rewards function are equivalent.*

$$\begin{aligned}&\forall h \in [H-1],\ s_h \in S_h^c,\ a_h \in A: \\ &\widehat{r}_k^c(s_h,a_h) = \widehat{f}_k(c,s_h,a_h) + \frac{\beta_k}{\sum_{i=1}^{k-1}\mathbb{I}[a_h=\pi_i(c,s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)},\end{aligned} \tag{2}$$

and

$$\forall (s,a)\in S\times A:\ \ \widehat{r}_k^c(s,a) = \widehat{f}_k(c,s,a) + \frac{\beta_k}{\sum_{i=1}^{k-1}\mathbb{I}[a=\pi_i(c,s)]q(s|\pi_i(c;\cdot),P_\star^c)}. \tag{3}$$

For convenience, in Algorithm 4 we use Definition 3 the approximated rewards function, but in the regret analysis we use Definition 2.

B.4 Regret Analysis

B.4.1 Analysis Outline

In the following regent analysis, for any $t > |A|$ we define the following MDPs for every context $c \in \mathcal{C}$.

1. $\mathcal{M}^{(f,P_\star)}(c) = (S, A, P_\star^c, f(c,\cdot,\cdot), s_0, H)$, for any $f \in \mathcal{F}$.
2. $\mathcal{M}^{(f_\star,P_\star)}(c)$ is the true model, where $f_\star(c,\cdot,\cdot) = r_\star^c$ is the true context dependent rewards and $P_\star^c$ is the true context-dependent dynamics, which we also denote by $\mathcal{M}(c)$.
3. $\widehat{\mathcal{M}}_t(c) = \mathcal{M}^{(\widehat{r}_t,P_\star)}(c)$ is the approximated MDP defined in Algorithm 4.

Inspired by the analysis of Xu and Zeevi (2020) for CMAB, our analysis is based on four main steps.

Step 1: Establish uniform convergence bound for any $t \geq 2$ and a fixed sequence of functions $f_2, f_3, \ldots \in \mathcal{F}$ which states the following (see Lemma 16).

For any $\delta \in (0,1)$, with probability at least $1-\delta/2$ it holds that

$$\begin{aligned}&\sum_{i=1}^{t-1}\mathbb{E}\left[\sum_{h=0}^{H-1}(f_t(c_i,s_h^i,a_h^i) - f_\star(c_i,s_h^i,a_h^i))^2|\mathbb{H}_{i-1}\right] \\ &\leq 68H\log(4|\mathcal{F}|t^3/\delta) + 2\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}(f_t(c_i,s_h^i,a_h^i)-r_h^i)^2 - (f_\star(c_i,s_h^i,a_h^i)-r_h^i)^2).\end{aligned}$$

To derive the above, we use Lemmas 12 and 15.

Step 2: Constructing a confidence bound over the value of any given context-dependent policy with respect to the true rewards function $f_\star$ and the least square minimizer at time t, $\widehat{f}_t$. The confidence bound holds with high probability.

In Lemma 18 we show that w.p. at least $1-\delta/2$ for all $t > |A|$ and every context-dependent policy $\pi \in \Pi_{\mathcal{C}}$ it holds that

$$\left|\mathbb{E}_c\left[V_{\mathcal{M}(c)}^{\pi(c;\cdot)}(s_0)\right] - \mathbb{E}_c\left[V_{\mathcal{M}^{(\widehat{f}_t,P_\star)}(c)}^{\pi(c;\cdot)}(s_0)\right]\right| \leq \sqrt{\phi_t(\pi)}\cdot\sqrt{68H\log(4|\mathcal{F}|t^3/\delta)},$$

where $\phi_t(\pi)$ is the contextual potential of policy $\pi \in \Pi_{\mathcal{C}}$ at time $|A| < t \leq T$ which is defined as follows.

$$\phi_t(\pi) := \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right].$$

Step 3: Relax the confidence bound in step 2 to be additive. In Lemma 19 we show that with probability at least $1-\delta/2$, for all $t > |A|$ and any policy $\pi \in \Pi_{\mathcal{C}}$ for $\beta_t = \sqrt{\frac{17t\log(4|\mathcal{F}|t^3/\delta)}{|S||A|}}$ it holds that

$$\left|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(\hat{f}_t,P_\star)(c)}(s_0)]\right| \leq \beta_t \cdot \phi_t(\pi) + \beta_t \cdot \frac{H\,|S|\,|A|}{t}.$$

Step 4: Bounding the cumulative contextual potential ϕ_t over every round $t > |A|$.
In Lemma 21 we show that for any sequence of selected context-dependent policies $\{\pi_t \in \Pi_{\mathcal{C}}\}_{t=1}^T$ it holds that

$$\sum_{t=|A|+1}^{T} \phi_t(\pi_t) \leq \frac{|S||A|}{p_{min}}(1+\log(T/|A|)).$$

In addition, if for all t, $\pi_t = \pi$, for any context dependent policy $\pi \in \Pi_c$, we have an improved bound where the $|S|$ factor is replaced with H factor:

$$\sum_{t=|A|+1}^{T} \phi_t(\pi) \leq \frac{H|A|}{p_{min}}(1+\log(T/|A|)).$$

Deriving regret bound. Using the results of steps 2 and 3, we derive the optimism lemma (Lemma 23). The lemma states that under the good event established in step 2, the following holds for any $t > |A|$.

$$\mathbb{E}_c\left[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0)\right] \leq \mathbb{E}_c\left[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}\right] + 2\beta_t \cdot \phi_t(\pi_t) + 2\beta_t\frac{H|S||A|}{t},$$

where $\pi^\star \in \Pi_{\mathcal{C}}$ is an optimal context-dependent policy, and $\pi_t \in \Pi_{\mathcal{C}}$ is the selected context-dependent policy at round t.

By summing the above inequality for all $t > |A|$ and applying the resoults of step 4, in Theorem 25 (and 24) we obtain a regret bound (and expected regret bound) of

$$\tilde{O}\left(\max\left\{\frac{\sqrt{T|S||A|\log\frac{|\mathcal{F}|}{\delta}}}{p_{min}}, H\sqrt{T|S||A|\log\frac{|\mathcal{F}|}{\delta}}, |A|H\right\}\right),$$

which holds with probability at least $1-\delta$ for all $T \geq 1$.

In the following analysis, we use the explicit expression of the contextual potential ϕ_t.

B.4.2 STEP 1: ESTABLISHING UNIFORM CONVERGENCE BOUND OVER $\mathcal{F}$

The following lemma is an adaption of Lemma 4.2 from Agarwal et al. (2012).

Lemma 12 *Fix a function $f \in \mathcal{F}$. Suppose we sample context c and rewards function r from the data distribution $\mathcal{D}$ and state-action pair (s,a) from arbitrary distribution such that r and (s,a) are conditionally independent given c. Define the random variable*

$$Y_{c,s,a,r} = (f(c,s,a) - r(s,a))^2 - (f_\star(c,s,a) - r(s,a))^2.$$

Then, the followings hold.

1. $\mathbb{E}_{c,s,a,r}[Y_{c,s,a,r}] = \mathbb{E}_{c,s,a}\left[(f(c,s,a) - f_\star(c,s,a))^2\right]$.
2. $\mathbb{V}_{c,s,a,r}[Y_{c,s,a,r}] \leq 4\,\mathbb{E}_{c,s,a,r}[Y_{c,s,a,r}]$
3. $\mathbb{V}\left[\sum_{h=0}^{H-1} Y_{c,s_h,a_h,r_h}\right] \leq 4H\mathbb{E}\left[\sum_{h=0}^{H-1} Y_{c,s_h,a_h,r_h}\right]$.

Proof Let us rearrange the definition of $Y_{c,s,a,r}$ as

$$Y_{c,s,a,r} = (f(c,s,a) - f_\star(c,s,a))(f(c,s,a) + f_\star(c,s,a) - 2r(s,a)). \tag{4}$$

Hence, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{c,s,a,r}[Y_{c,s,a,r}] &= \mathop{\mathbb{E}}_{c,s,a,r}[(f(c,s,a) - f_\star(c,s,a))(f(c,s,a) + f_\star(c,s,a) - 2r(s,a))] \\
&= \mathop{\mathbb{E}}_{c,s,a} \mathop{\mathbb{E}}_{r|c}[(f(c,s,a) - f_\star(c,s,a))(f(c,s,a) + f_\star(c,s,a) - 2r(s,a))] \\
&= \mathop{\mathbb{E}}_{c,s,a}[(f(c,s,a) - f_\star(c,s,a))(f(c,s,a) + f_\star(c,s,a) - 2 \mathop{\mathbb{E}}_{r|c}[r(s,a)])] \\
&= \mathop{\mathbb{E}}_{c,s,a}\left[(f(c,s,a) - f_\star(c,s,a))^2\right],
\end{aligned}$$

where the third identity uses the realizability assumption that $f_\star(c,s,a) = \mathbb{E}_{r|c}[r(s,a)]$, proving the first part of the lemma.

For the second part, note that f, $f_\star$ and r are all between 0 and 1. Hence using equation (4) we obtain

$$\begin{aligned}
Y^2_{c,s,a,r} =& (f(c,s,a) - f_\star(c,s,a))^2 (f(c,s,a) + f_\star(c,s,a) - 2r(s,a))^2 \\
\le& 4(f(c,s,a) - f_\star(c,s,a))^2,
\end{aligned} \tag{5}$$

yielding the second part of the lemma as

$$\mathop{\mathbb{V}}_{c,s,a,r}[Y_{c,s,a,r}] \le \mathop{\mathbb{E}}_{c,s,a,r}[Y^2_{c,s,a,r}] \le 4 \mathop{\mathbb{E}}_{c,s,a}[(f(c,s,a) - f_\star(c,s,a))^2] = 4 \mathop{\mathbb{E}}_{c,s,a,r}[Y_{c,s,a,r}].$$

For the third part, by norms inequality and inequality (5) we obtain,

$$\left(\sum_{h=0}^{H-1} Y_{c,s_h,a_h,r_h}\right)^2 \le H \sum_{h=0}^{H-1} Y^2_{c,s_h,a_h,r_h} \le 4H \sum_{h=0}^{H-1} (f(c,s_h,a_h) - f_\star(c,s_h,a_h))^2,$$

yielding the third part of the lemma as

$$\mathbb{V}\left[\sum_{h=0}^{H-1} Y_{c,s_h,a_h,r_h}\right] \le \mathbb{E}\left[\left(\sum_{h=0}^{H-1} Y_{c,s_h,a_h,r_h}\right)^2\right] \le 4H\,\mathbb{E}\left[\sum_{h=0}^{H-1} (f(c,s_h,a_h) - f_\star(c,s,a))^2\right] = 4H\,\mathbb{E}\left[\sum_{h=0}^{H-1} Y_{c,s_h,a_h,r_h}\right].$$

■

Lemma 13 (Freedman's inequality Bartlett et al. (2008)) *suppose $Z_1, Z_2, \ldots, Z_t$ is a martingale difference sequence with $|Z_i| \le b$ for all $i = 1, 2, \ldots, t$. Then for any $\delta < 1/e^2$ with probability at least $1 - \log_2(t)\delta$,*

$$\sum_{i=1}^{t} Z_i \le 4\sqrt{\sum_{i=1}^{t} \mathbb{V}[Z_i | Z_1, \ldots, Z_{i-1}] \log(1/\delta)} + 2b \log(1/\delta).$$

Definition 14 *For every round $t \ge 1$, layer $h \in [H-1]$ and a function $f \in \mathcal{F}$ we define the random variable*

$$Y_{f,c_t,s_h^t,a_h^t,r_h^t} = (f(c_t, s_h^t, a_h^t) - r_h^t)^2 - (f_\star(c_t, s_h^t, a_h^t) - r_h^t)^2$$

where $(c_t, s_h^t, a_h^t) \sim \mathcal{D}(c_t) \cdot q_h(s_h^t, a_h^t | \pi_t(c_t; \cdot), P_\star^{c_t})$ and $r_h^t \sim R_\star^{c_t}(s_h^t, a_h^t)$.

In the proof of the following lemma, we use the following obvious observations.

Observation 3 *For all $i \in [t-1]$ and $h \in \{0, \dots H-1\}$ we have that*

$$(c_i, s_h^i, a_h^i) \sim \mathcal{D}(c_i) \cdot q_h(s_h^i, a_h^i | \pi_i(c_i; \cdot), P_\star^{c_i}).$$

In addition, π_i is determined completely by the history $\mathbb{H}_{i-1}$. Hence, by linearity of expectation it holds for any $f \in \mathcal{F}$ that,

$$\mathbb{E}\left[\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i} \middle| \mathbb{H}_{i-1}\right] = \sum_{h=0}^{H-1} \mathop{\mathbb{E}}_{c_i,s_h^i,a_h^i,r_h^i}\left[Y_{f,c_i,s_h^i,a_h^i,r_h^i} \middle| \mathbb{H}_{i-1}\right].$$

Similarly,

$$\mathbb{E}\left[\sum_{h=0}^{H-1} (f(c_i, s_h^i, a_h^i) - f_\star(c_i, s_h^i, a_h^i))^2 \middle| \mathbb{H}_{i-1}\right] = \sum_{h=0}^{H-1} \mathop{\mathbb{E}}_{c_i,s_h^i,a_h^i}\left[(f(c_i, s_h^i, a_h^i) - f_\star(c_i, s_h^i, a_h^i))^2 \middle| \mathbb{H}_{i-1}\right].$$

Observation 4 *For any $f \in \mathcal{F}$, we have that $\{Z_i\}_{i=1}^{t-1}$ is a martingale difference sequence of length $t-1$ where*

$$Z_i := \mathbb{E}\left[\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i} \middle| \mathbb{H}_{i-1}\right] - \sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i}$$

where the filtration is $\{\mathbb{H}_i\}_{i=1}^{t-1}$, and $\mathbb{H}_0$ is the empty history. Recall that for all $i \geq 1$ we defined that $\mathbb{H}_i = (\sigma^1, \dots, \sigma^i)$. (Clearly, Z_i is determined given $\mathbb{H}_0, \mathbb{H}_1, \dots, \mathbb{H}_i$ and $\mathbb{E}[Z_i | \mathbb{H}_1, \dots, \mathbb{H}_{i-1}] = 0$ for all $i \in \{1, 2, \dots, t-1\}$.)

In addition, since $\mathbb{E}\left[\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i} \middle| \mathbb{H}_{i-1}\right]$ is deretmined given $\mathbb{H}_{i-1}$, it holds that

$$\mathbb{V}\left[Z_i \mid \mathbb{H}_{i-1}\right] = \mathbb{V}\left[\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i} \middle| \mathbb{H}_{i-1}\right].$$

Lemma 15 (uniform convergence over $\mathcal{F}$) *For a fixed $t \geq 2$ and a fixed $\delta_t \in (0, 1/e^2)$ with probability at least $1 - \log_2((t-1)H)\delta_t$, it holds that*

$$\begin{aligned}
&\sum_{i=1}^{t-1} \sum_{h=0}^{H-1} \mathop{\mathbb{E}}_{c_i,s_h^i,a_h^i}\left[(f(c_i, s_h^i, a_h^i) - f_\star(c_i, s_h^i, a_h^i))^2 | \mathbb{H}_{i-1}\right] \\
&= \sum_{i=1}^{t-1} \mathbb{E}\left[\sum_{h=0}^{H-1} (f(c_i, s_h^i, a_h^i) - f_\star(c_i, s_h^i, a_h^i))^2 \middle| \mathbb{H}_{i-1}\right] \\
&\leq 68H \log(|\mathcal{F}|/\delta_t) + 2\sum_{i=1}^{t-1} \sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i}.
\end{aligned}$$

uniformly for all $f \in \mathcal{F}$.

Proof Fix a function $f \in \mathcal{F}$, and consider the random variable defined in 14,

$$Y_{f,c_t,s_h^t,a_h^t,r_h^t} = (f_t(c_i, s_h^i, a_h^i) - r_h^i)^2 - (f_\star(c_i, s_h^i, a_h^i) - r_h^i)^2.$$

Notice that $|Y_{f,c_t,s_h^t,a_h^t,r_h^t}| \leq 1$ for any function $f \in \mathcal{F}$, round $t \geq 2$, layer $h \in [H-1]$, state $s_h^t \in S_h^{c_t}$, action $a_h^t \in A$ and observed reward r_h^t. Hence,

$$\left|\sum_{h=0}^{H-1} Y_{f,c_t,s_h^t,a_h^t,r_h^t}\right| \leq \sum_{h=0}^{H-1} |Y_{f,c_t,s_h^t,a_h^t,r_h^t}| \leq H.$$

By Freedman's inequality (Lemma 13), for $\delta_t < 1/e^2$, with probability at least $1 - \log_2(t-1)\delta_t/|\mathcal{F}|$ it holds that

$$\sum_{i=1}^{t-1} \mathbb{E}\left[\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i} \middle| \mathbb{H}_{i-1}\right] - \sum_{i=1}^{t-1}\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i}$$
$$\leq 4\sqrt{\sum_{i=1}^{t-1} \mathbb{V}\left[\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i} \middle| \mathbb{H}_{i-1}\right] \log(|\mathcal{F}|/\delta_t)} + 2H\log(|\mathcal{F}|/\delta_t).$$

By Lemma 12 for all $i \in \{1, 2, \ldots, t-1\}$ it holds that

$$\mathbb{V}\left[\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i} \middle| \mathbb{H}_{i-1}\right] \leq 4H\mathbb{E}\left[\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i} \middle| \mathbb{H}_{i-1}\right].$$

Therefore,

$$\sum_{i=1}^{t-1} \mathbb{E}\left[\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i} \middle| \mathbb{H}_{i-1}\right]$$
$$\leq 4\sqrt{\sum_{i=1}^{t-1} \mathbb{V}\left[\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i} \middle| \mathbb{H}_{i-1}\right] \log(|\mathcal{F}|/\delta_t)} + 2H\log(|\mathcal{F}|/\delta_t) + \sum_{i=1}^{t-1}\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i}$$
$$\leq 8\sqrt{H\sum_{i=1}^{t-1} \mathbb{E}\left[\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i} \middle| \mathbb{H}_{i-1}\right] \log(|\mathcal{F}|/\delta_t)} + 2H\log(|\mathcal{F}|/\delta_t) + \sum_{i=1}^{t-1}\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i}.$$

We have

$$\sum_{i=1}^{t-1} \mathbb{E}\left[\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i} \middle| \mathbb{H}_{i-1}\right] - 8\sqrt{H\sum_{i=1}^{t-1} \mathbb{E}\left[\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i} \middle| \mathbb{H}_{i-1}\right] \log(|\mathcal{F}|/\delta_t)}$$
$$\leq 2H\log(|\mathcal{F}|/\delta_t) + \sum_{i=1}^{t-1}\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i}.$$

We add to both sides $16H\log(|\mathcal{F}|/\delta_t)$, and this implies for any $f \in \mathcal{F}$ that

$$\left(\sqrt{\sum_{i=1}^{t-1} \mathbb{E}\left[\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i} \middle| \mathbb{H}_{i-1}\right]} - 4\sqrt{H\log(|\mathcal{F}|/\delta_t)}\right)^2 \leq 18H\log(|\mathcal{F}|/\delta_t) + \sum_{i=1}^{t-1}\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i}. \quad (6)$$

Yielding,

$$\sqrt{\sum_{i=1}^{t-1} \mathbb{E}\left[\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i} \middle| \mathbb{H}_{i-1}\right]} \leq 4\sqrt{H\log(|\mathcal{F}|/\delta_t)} + \sqrt{18H\log(|\mathcal{F}|/\delta_t) + \sum_{i=1}^{t-1}\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i}}. \quad (7)$$

And then,

$$\sum_{i=1}^{t-1} \mathbb{E}\left[\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i} \middle| \mathbb{H}_{i-1}\right] \leq \left(4\sqrt{H\log(|\mathcal{F}|/\delta_t)} + \sqrt{18H\log(|\mathcal{F}|/\delta_t) + \sum_{i=1}^{t-1}\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i}}\right)^2. \quad (8)$$

This inequality further implies (using that $(a+b)^2 \leq 2(a^2+b^2)$ for all $a, b \in \mathbb{R}$) that for any $f \in \mathcal{F}$ it holds that

$$\sum_{i=1}^{t-1} \mathbb{E}\left[\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i} \middle| \mathbb{H}_{i-1}\right] \leq 68H\log(|\mathcal{F}|/\delta_t) + 2\sum_{i=1}^{t-1}\sum_{h=0}^{H-1} Y_{f,c_i,s_h^i,a_h^i,r_h^i}. \quad (9)$$

Lastly, by combining inequality (9) with Lemma 12 we obtain the lemma since,

$$\begin{aligned}
&\sum_{i=1}^{t-1}\mathbb{E}\left[\sum_{h=0}^{H-1}(f(c_i,s_h^i,a_h^i)-f_\star(c_i,s_h^i,a_h^i))^2|\mathbb{H}_{i-1}\right]\\
&=\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\mathop{\mathbb{E}}_{c_i,s_h^i,a_h^i}\left[(f(c_i,s_h^i,a_h^i)-f_\star(c_i,s_h^i,a_h^i))^2|\mathbb{H}_{i-1}\right] && \text{(By Observation 3)}\\
&=\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\mathop{\mathbb{E}}_{c_i,s_h^i,a_h^i,r_h^i}\left[Y_{f,c_i,s_h^i,a_h^i,r_h^i}\;\middle|\;\mathbb{H}_{i-1}\right] && \text{(Lemma 12)}\\
&=\sum_{i=1}^{t-1}\mathbb{E}\left[\sum_{h=0}^{H-1}Y_{f,c_i,s_h^i,s_h^i,r_h^i}\;\middle|\;\mathbb{H}_{i-1}\right] && \text{(By Observation 3)}\\
&\le 68H\log(|\mathcal{F}|/\delta_t)+2\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}Y_{f,c_i,s_h^i,a_h^i,r_h^i}. && \text{(By inequality 9)}
\end{aligned}$$

■

Lemma 16 (uniform convergence over all sequences of estimators) *For any $\delta\in(0,1)$, with probability at least $1-\delta/2$ it holds that*

$$\begin{aligned}
&\sum_{i=1}^{t-1}\mathbb{E}\left[\sum_{h=0}^{H-1}(f_t(c_i,s_h^i,a_h^i)-f_\star(c_i,s_h^i,a_h^i))^2|\mathbb{H}_{i-1}\right]\\
&=\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\mathop{\mathbb{E}}_{c_i,s_h^i,a_h^i}\left[(f_t(c_i,s_h^i,a_h^i)-f_\star(c_i,s_h^i,a_h^i))^2|\mathbb{H}_{i-1}\right]\\
&\le 68H\log(4|\mathcal{F}|t^3/\delta)+2\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}(f_t(c_i,s_h^i,a_h^i)-r_h^i)^2-(f_\star(c_i,s_h^i,a_h^i)-r_h^i)^2.
\end{aligned}$$

simultaneously for all $t\ge 2$ and any fixed sequence of functions $f_2,f_3,\ldots\in\mathcal{F}$.

Proof For a fixed $\delta\in(0,1)$, take $\delta_t=\delta/4t^3$ and apply union bound to Lemma 15 with all $t\ge 2$. We have,

$$\sum_{t=1}^{\infty}\delta_t\log(t-1)=\sum_{t=1}^{\infty}\delta/4t^3\log(t-1)\le\sum_{t=1}^{\infty}\frac{\delta}{4t^2}\le\frac{\delta}{2}.$$

Hence, by Lemma 15 with probability at least $1-\delta/2$ it holds that

$$\begin{aligned}
&\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\mathop{\mathbb{E}}_{c_i,s_h^i,a_h^i}\left[(f_t(c_i,s_h^i,a_h^i)-f_\star(c_i,s_h^i,a_h^i))^2|\mathbb{H}_{i-1}\right]\\
&=\sum_{i=1}^{t-1}\mathbb{E}\left[\sum_{h=0}^{H-1}(f_t(c_i,s_h^i,a_h^i)-f_\star(c_i,s_h^i,a_h^i))^2|\mathbb{H}_{i-1}\right]\\
&\le 68H\log(4|\mathcal{F}|t^3/\delta)+2\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}Y_{f_t,c_t,s_h^i,a_h^i,r_h^i}\\
&=68H\log(4|\mathcal{F}|t^3/\delta)+2\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}(f_t(c_i,s_h^i,a_h^i)-r_h^i)^2-(f_\star(c_i,s_h^i,a_h^i)-r_h^i)^2.
\end{aligned}$$

simultaneously for every $t\ge 2$ and any fixed sequence of functions $f_2,f_3,\ldots\in\mathcal{F}$. ■

B.4.3 STEP 2: CONSTRUCTING CONFIDENCE BOUND OVER POLICIES WITH RESPECT TO REWARDS APPROXIMATION

Remark 17 *Recall $f_\star(c,s,a) = r^c_\star(s,a)$. Hence, for any policy $\pi \in \Pi_\mathcal{C}$ and a context $c \in \mathcal{C}$ we have the following,*

$$\begin{aligned} V^{\pi(c;\cdot)}_{\mathcal{M}(c)} &= \sum_{h=0}^{H-1} \sum_{s_h \in S^c_h} \sum_{a_h \in A} q_h(s_h,a_h|\pi(c;\cdot),P^c_\star)\cdot r^c_\star(s_h,a_h) \\ &= \sum_{h=0}^{H-1} \sum_{s_h \in S^c_h} \sum_{a_h \in A} q_h(s_h,a_h|\pi(c;\cdot),P^c_\star)\cdot f_\star(c;s_h,a_h). \end{aligned} \tag{10}$$

Lemma 18 (confidence bound over policies w.r.t rewards approximation) *Consider Algorithm 4 that at each initialization round $t \le |A|$, plays the policy that always choose action a_t, and at each round $t \ge |A|+1$ selects π_t based on the history $\mathbb{H}_{t-1}$.*

Then, for any $\delta \in (0,1)$, with probability at least $1-\delta/2$ for all $t > |A|$ and any context-dependent policy $\pi \in \Pi_\mathcal{C}$ the following holds.

$$\begin{aligned} &\left|\mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)\right] - \mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{f}_t,P_\star)(c)}(s_0)\right]\right| \\ &\le \sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S^c_h} \frac{q_h(s_h|\pi(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}\right]}\cdot\sqrt{68H\log(4|\mathcal{F}|t^3/\delta)}, \end{aligned}$$

where $\mathcal{M}^{(f,P_\star)}(c) = (S,A,P^c_\star,f(c,\cdot,\cdot),s_0,H)$ for any $f \in \mathcal{F}$, and $\mathcal{M}(c) = \mathcal{M}^{(f_\star,P_\star)}(c)$.

Proof For any function $f \in \mathcal{F}$ we consider the random variable defined in 14

$$Y_{f,c_t,s^t_h,a^t_h,r^t_h} = (f(c_t,s^t_h,a^t_h) - r^t_h)^2 - (f_\star(c_t,s^t_h,a^t_h) - r^t_h)^2$$

for any $t \ge 1$, context $c_t \in \mathcal{C}$, layer $h \in [H-1]$, state $s^t_h \in S^{c_t}_h$ and action $a^t_h \in A$. We remark that $(c_t,s^t_h,a^t_h) \sim \mathcal{D}(c_t)\cdot q_h(s^t_h,a^t_h|\pi_t(c_t;\cdot),P^{c_t}_\star)$ and $r^t_h \sim R^{c_t}_\star(s^t_h,a^t_h)$.

We now prove the following auxiliary claim,

Claim 1 *For all $t \ge 2$ it holds that*

$$\begin{aligned} &\sum_{i=1}^{t-1}\sum_{h=1}^{H-1} \mathop{\mathbb{E}}_{c_i,s^i_h,a^i_h}\left[(\widehat{f}_t(c_i,s^i_h,a^i_h) - f_\star(c_i,s^i_h,a^i_h))^2|\mathbb{H}_{i-1}\right] = \\ &\mathbb{E}_c\left[\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\sum_{s_h\in S^c_h} q_h(s_h|\pi_i(c;\cdot),P^c_\star)(\widehat{f}_t(c,s_h,\pi_i(c;s_h)) - f_\star(c,s_h,\pi_i(c;s_h)))^2\right]. \end{aligned}$$

Proof Recall that for every round $i \in \{1,2,\ldots,|A|\}$, π_i is a deterministic policy that always plays action a_i. For all $i > |A|$ we have that π_i is determined completely by the history $\mathbb{H}_{i-1}$, and is a deterministic context-dependent policy. Hence, for any function $f \in \mathcal{F}$, round $i \in \{1,\ldots,t-1\}$ and layer $h \in [H-1]$ it holds that

$$\begin{aligned} &\mathop{\mathbb{E}}_{c_i,s^i_h,a^i_h}\left[(f(c_i,s^i_h,a^i_h) - f_\star(c_i,s^i_h,a^i_h))^2|\mathbb{H}_{i-1}\right] \\ &= \mathop{\mathbb{E}}_{c_i,s^i_h}\left[(f(c_i,s^i_h,\pi_i(c_i;s^i_h)) - f_\star(c_i,s^i_h,\pi_i(c_i;s^i_h)))^2|\mathbb{H}_{i-1}\right] \\ &= \mathbb{E}_c\left[\mathbb{E}_{s_h}\left[(f(c,s_h,\pi_i(c;s_h)) - f_\star(c,s_h,\pi_i(c;s_h)))^2|\mathbb{H}_{i-1},c\right]\right] \\ &= \mathbb{E}_c\left[\sum_{s_h\in S^c_h} q_h(s_h|\pi_i(c;\cdot),P^c_\star)(f(c,s_h,\pi_i(c;s_h)) - f_\star(c,s_h,\pi_i(c;s_h)))^2\right], \end{aligned} \tag{11}$$

where the first equality is because $a_h^i = \pi_i(c_i; s_h^i)$ and π_i is determined deterministically given $\mathbb{H}_{i-1}$. The second equality is since c_i is independent of $\mathbb{H}_{i-1}$. The third equality is an explicit representation of the expectation over s_h given the context c and the history $\mathbb{H}_{i-1}$, since π_i is determined completely by $\mathbb{H}_{i-1}$.

By summing over $i = 1, 2, \ldots, t-1$ and $h \in [H-1]$ we obtain the claim since,

$$\begin{aligned}
&\sum_{i=1}^{t-1}\sum_{h=1}^{H-1} \mathop{\mathbb{E}}_{c_i, s_h^i, a_h^i}\left[(\widehat{f}_t(c_i, s_h^i, a_h^i) - f_\star(c_i, s_h^i, a_h^i))^2 | \mathbb{H}_{i-1}\right] = \\
=&\sum_{i=1}^{t-1}\sum_{h=1}^{H-1} \mathbb{E}_c\left[\sum_{s_h \in S_h^c} q_h(s_h|\pi_i(c;\cdot), P_\star^c)(\widehat{f}_t(c, s_h, \pi_i(c; s_h)) - f_\star(c, s_h, \pi_i(c; s_h)))^2\right] && \text{(By equation (11))} \\
=&\mathbb{E}_c\left[\sum_{i=1}^{t-1}\sum_{h=1}^{H-1}\sum_{s_h \in S_h^c} q_h(s_h|\pi_i(c;\cdot), P_\star^c)(\widehat{f}_t(c, s_h, \pi_i(c; s_h)) - f_\star(c, s_h, \pi_i(c; s_h)))^2\right], \\
& && \text{(By linearity of expectation)}
\end{aligned}$$

as stated.

■

Returning to the proof of the lemma, by Lemma 16, for any $\delta \in (0,1)$ we have with probability at least $1 - \delta/2$ that

$$\begin{aligned}
&\sum_{i=1}^{t-1}\sum_{h=0}^{H-1} \mathop{\mathbb{E}}_{c_i, s_h^i, a_h^i}\left[(f_t(c_i, s_h^i, a_h^i) - f_\star(c_i, s_h^i, a_h^i))^2 | \mathbb{H}_{i-1}\right] \\
&\leq 68H\log(4|\mathcal{F}|t^3/\delta) + 2\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}(f_t(c_i, s_h^i, a_h^i) - r_h^i)^2 - (f_\star(c_i, s_h^i, a_h^i) - r_h^i)^2,
\end{aligned}$$

simultaneously for all $t \geq |A| + 1$ and any fixed sequence of functions $f_{|A|+1}, f_{|A|+2}, \ldots \in \mathcal{F}$.

Since $\{\widehat{f}_t\}_{t=|A|+1}^T$ are the least square minimizers at time t, it holds that

$$\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}(\widehat{f}_t(c_i, s_h^i, a_h^i) - r_h^i)^2 - (f_\star(c_i, s_h^i, a_h^i) - r_h^i)^2 \leq 0.$$

Recall that $\mathcal{M}(c) = \mathcal{M}^{(f_\star, P_\star)}(c)$ for all $c \in \mathcal{C}$.

By all the above, with probability at least $1 - \delta/2$ for all $t \geq |A| + 1$ and any context-dependent policy $\pi \in \Pi_{\mathcal{C}}$ it holds that

$$\begin{aligned}
&\left|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{f}_t, P_\star)}(c)}(s_0) - V^{\pi(c;\cdot)}_{\mathcal{M}^{(f_\star, P_\star)}(c)}(s_0)]\right| \\
=&\left|\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h \in S_h^c}\sum_{a_h \in A} q_h(s_h, a_h|\pi(c;\cdot), P_\star^c)(\widehat{f}_t(c, s_h, a_h) - f_\star(c, s_h, a_h))\right]\right| && \text{(By definition)} \\
=&\left|\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h \in S_h^c} q_h(s_h|\pi(c;\cdot), P_\star^c)(\widehat{f}_t(c, s_h, \pi(c; s_h)) - f_\star(c, s_h, \pi(c; s_h)))\right]\right| \\
& && (\pi(c;\cdot) \text{ is a deterministic policy for all } c \in \mathcal{C}) \\
\leq&\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h \in S_h^c} q_h(s_h|\pi(c;\cdot), P_\star^c)\,|f_t(c, s_h, \pi(c; s_h)) - f_\star(c, s_h, \pi(c; s_h))|\right] && \text{(By triangle inequality)}
\end{aligned}$$

$$=\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}(\sqrt{q_h(s_h|\pi(c;\cdot),P_\star^c)})^2\frac{\sqrt{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}}{\sqrt{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}}\right.$$

$$\left.\cdot\left|\widehat{f}_t(c,s_h,\pi(c;s_h))-f_\star(c,s_h,\pi(c;s_h))\right|\right] \qquad \text{(Multiplication in } \tfrac{\sqrt{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}}{\sqrt{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}}\text{)}$$

$$=\sum_{c\in\mathcal{C}}\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sqrt{\mathcal{D}(c)}(\sqrt{q_h(s_h|\pi(c;\cdot),P_\star^c)})^2\frac{1}{\sqrt{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}} \quad \text{(Re-arranging)}$$

$$\cdot\sqrt{\mathcal{D}(c)}\sqrt{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)|\widehat{f}_t(c,s_h,\pi(c;s_h))-f_\star(c,s_h,\pi(c;s_h))|}$$

$$\leq\sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h^2(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]} \qquad \text{(By Cauchy-Schwartz inequality)}$$

$$\cdot\sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)(\widehat{f}_t(c,s_h,\pi(c;s_h))-f_\star(c,s_h,\pi(c;s_h)))^2\right]}$$

$$\leq\sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]} \quad \text{(By } q_h^2(s_h|\pi(c;\cdot),P_\star^c)\leq q_h(s_h|\pi(c;\cdot),P_\star^c)\text{)}$$

$$\cdot\sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)(\widehat{f}_t(c,s_h,\pi(c;s_h))-f_\star(c,s_h,\pi(c;s_h)))^2\right]}$$

$$=\sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi,P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]}$$

(The non-zero terms are where $\pi_i(c;s_h)=\pi(c;s_h)$)

$$\cdot\sqrt{\mathbb{E}_c\left[\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)(\widehat{f}_t(c,s_h,\pi_i(c;s_h))-f_\star(c,s_h,\pi_i(c;s_h)))^2\right]}$$

$$\leq\sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]}$$

(Removing the indicators only increase the sum)

$$\cdot\sqrt{\mathbb{E}_c\left[\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}q_h(s_h|\pi_i(c;\cdot),P_\star^c)(\widehat{f}_t(c,s_h,\pi_i(c;s_h))-f_\star(c,s_h,\pi_i(c;s_h)))^2\right]}$$

$$=\sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]} \qquad \text{(By Claim 1)}$$

$$\cdot \sqrt{\sum_{i=1}^{t-1}\sum_{h=1}^{H-1} \mathop{\mathbb{E}}_{c_i,s_h^i,a_h^i}\left[(\widehat{f}_t(c_i,s_h^i,a_h^i) - f_\star(c_i,s_h^i,a_h^i))^2 | \mathbb{H}_{i-1}\right]}$$

$$\leq \sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c} \frac{q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]} \cdot \sqrt{68H\log(4|\mathcal{F}|t^3/\delta)}.$$

(By Lemma 16 combined with the fact that $\widehat{f}_t$ is the least-square minimizer)

Lastly, we remark that by choice of π_i for $i \in \{1,2,\ldots,|A|\}$, and the minimum reachability assumption for any deterministic policy $\pi \in \Pi_{\mathcal{C}}$, layer $h \in [H-1]$, state $s_h \in S_h^c$ and $t > |A|$ it holds that

$$\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c) \geq p_{min}\cdot\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)] \geq p_{min} > 0,$$

hence the above is well defined. ■

B.4.4 STEP 3: RELAX THE CONFIDENCE BOUND TO BE ADDITIVE

Lemma 19 (the "square trick" relaxation) *Under the good event of Lemma 18, for all $t > |A|$ and any context-dependent policy $\pi \in \Pi_{\mathcal{C}}$ it holds that*

$$\left|\mathbb{E}_c[V_{\mathcal{M}(c)}^{\pi(c;\cdot)}(s_0)] - \mathbb{E}_c[V_{\mathcal{M}(\widehat{f}_t,P_\star)(c)}^{\pi(c;\cdot)}(s_0)]\right|$$
$$\leq \beta_t\cdot\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right] + \beta_t\cdot\frac{H|S||A|}{t},$$

where $\beta_t = \sqrt{\frac{17t\log(4|\mathcal{F}|t^3/\delta)}{|S||A|}}$.

Proof Consider the following derivation, for $\beta_t = \sqrt{\frac{17t\log(4|\mathcal{F}|t^3/\delta)}{|S||A|}}$.

$$\left|\mathbb{E}_c\left[V_{\mathcal{M}(c)}^{\pi(c;\cdot)}\right] - \mathbb{E}_c\left[V_{\mathcal{M}(\widehat{f}_t,P_\star)(c)}^{\pi(c;\cdot)}\right]\right|$$

$$\leq \sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]}\cdot\sqrt{68H\log(4|\mathcal{F}|t^3/\delta)} \qquad \text{(By Lemma 18)}$$

$$= \sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{\beta_t\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]}\cdot\sqrt{\frac{1}{\beta_t}68H\log(4|\mathcal{F}|t^3/\delta)}$$

$$= \sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{\beta_t\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]}\cdot\sqrt{H\sqrt{\frac{|S||A|}{t}}\frac{68\log(4|\mathcal{F}|t^3/\delta)}{\sqrt{17\log(4|\mathcal{F}|t^3/\delta)}}}$$

(By β_t choice)

$$= \sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{\beta_t\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]}\cdot\sqrt{4H\cdot\sqrt{\frac{|S||A|}{t}}\frac{(\sqrt{17\log(4|\mathcal{F}|t^3/\delta)})^2}{\sqrt{17\log(4|\mathcal{F}|t^3/\delta)}}}$$

$$
\begin{aligned}
&= \sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{\beta_t\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]}\cdot\sqrt{4H\cdot\sqrt{\frac{|S||A|}{t}}\sqrt{17\log(4|\mathcal{F}|t^3/\delta)}} \\
&= \sqrt{2\cdot\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{\beta_t\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]} \\
&\quad\cdot\sqrt{2H\cdot\sqrt{\frac{|S|\cdot|A|}{t}}\sqrt{\frac{t^2}{|S|^2\cdot|A|^2}}\sqrt{\frac{|S|^2\cdot|A|^2}{t^2}}\sqrt{17\log(4|\mathcal{F}|t^3/\delta)}} \\
&= \sqrt{2\cdot\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{\beta_t\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]}\cdot\sqrt{2H\cdot\beta_t\sqrt{\frac{|S|^2\cdot|A|^2}{t^2}}} \\
&= 2\cdot\sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{\beta_t\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]}\cdot\sqrt{\beta_t\cdot\frac{H|S||A|}{t}} \\
&\le \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{\beta_t\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]+\beta_t\cdot\frac{H|S||A|}{t}. \qquad \text{(Since } 2ab\le a^2+b^2\text{)}
\end{aligned}
$$

■

B.4.5 STEP 4: BOUNDING THE SUM OF CONTEXTUAL POTENTIAL FUNCTIONS

Let us define the contextual potential function in round t, for $T \ge t > |A|$.

Definition 20 *We denote by $\phi_t(\pi)$ the contextual potential of a context-dependent policy $\pi \in \Pi_{\mathcal{C}}$ at round $|A| < t \le T$ which is defined as follows.*

$$
\phi_t(\pi) := \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right],
$$

where $\{\pi_t \in \Pi_{\mathcal{C}}\}_{t=1}^T$ is the sequence of policies selected by Algorithm 4.

In the following lemma, we bound the sum of contextual potential functions, over the rounds $t = |A|+1, \dots, T$.

Lemma 21 (contextual potential) *Let $\{\pi_t \in \Pi_{\mathcal{C}}\}_{t=1}^T$ be the sequence of policies selected by Algorithm 4. Then, for all $T > |A|$ the followings hold.*

1. *For any policy $\pi \in \Pi_{\mathcal{C}}$ it holds that*

$$
\begin{aligned}
\sum_{t=|A|+1}^{T}\phi_t(\pi) &= \sum_{t=|A|+1}^{T}\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right] \\
&\le \frac{H|A|}{p_{min}}(1+\log(T/|A|)).
\end{aligned}
$$

2. *For the sequence* $\{\pi_t \in \Pi_{\mathcal{C}}\}_{t=1}^{T}$ *it holds that*

$$\sum_{t=|A|+1}^{T} \phi_t(\pi_t) = \sum_{t=|A|+1}^{T} \mathbb{E}_c \left[\sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \frac{q_h(s_h|\pi_t(c;\cdot), P_\star^c)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi_t(c;s_h) = \pi_i(c;s_h)] q_h(s_h|\pi_i(c;\cdot), P_\star^c)} \right]$$
$$\leq \frac{|S||A|}{p_{min}}(1 + \log(T/|A|)).$$

Proof For a fixed context $c \in \mathcal{C}$ and any context-dependent policy $\pi \in \Pi_{\mathcal{C}}$ it holds that

$$\begin{aligned}
&\sum_{t=|A|+1}^{T} \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \frac{q_h(s_h|\pi(c;\cdot), P_\star^c)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c;s_h) = \pi_i(c;s_h)] q_h(s_h|\pi_i(c;\cdot), P_\star^c)} \\
\leq &\sum_{t=|A|+1}^{T} \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \frac{q_h(s_h|\pi(c;\cdot), P_\star^c)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c;s_h) = \pi_i(c;s_h)] \cdot p_{min}} \\
= &\frac{1}{p_{min}} \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} q_h(s_h|\pi(c;\cdot), P_\star^c) \sum_{t=|A|+1}^{T} \frac{1}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c;s_h) = \pi_i(c;s_h)]} \\
\leq &\frac{1}{p_{min}} \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} q_h(s_h|\pi(c;\cdot), P_\star^c) \sum_{a_h \in A} \sum_{i=1}^{\sum_{t=1}^{T} \mathbb{I}[\pi_t(c;s_h)=a_h]} \frac{1}{i} \\
\leq &\frac{1}{p_{min}} \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} q_h(s_h|\pi(c;\cdot), P_\star^c) \sum_{a_h \in A} \left(1 + \log\left(\sum_{t=1}^{T} \mathbb{I}[\pi_t(c;s_h) = a_h]\right)\right) \\
\leq &\frac{1}{p_{min}}(H \cdot |A| + H \cdot |A| \log(T/|A|)) \\
= &\frac{H|A|}{p_{min}}(1 + \log(T/|A|)),
\end{aligned}$$

where the last inequality is due to Jensen's inequality. By taking expectation over c in both sides of the above inequality, we obtain the first part of the lemma.

The proof of the second part of the lemma is similar.

For a fixed context $c \in \mathcal{C}$ it holds that

$$\begin{aligned}
&\sum_{t=|A|+1}^{T} \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \frac{q_h(s_h|\pi_t(c;\cdot), P_\star^c)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi_t(c;s_h) = \pi_i(c;s_h)] q_h(s_h|\pi_i(c;\cdot), P_\star^c)} \\
\leq &\sum_{t=|A|+1}^{T} \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \frac{q_h(s_h|\pi_t(c;\cdot), P_\star^c)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi_t(c;s_h) = \pi_i(c;s_h)] \cdot p_{min}} \\
\leq &\frac{1}{p_{min}} \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \sum_{t=|A|+1}^{T} \frac{1}{\sum_{i=1}^{t-1} \mathbb{I}[\pi_t(c;s_h) = \pi_i(c;s_h)]} \\
\leq &\frac{1}{p_{min}} \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \sum_{a_h \in A} \sum_{i=1}^{\sum_{t=1}^{T} \mathbb{I}[\pi_t(c;s_h)=a_h]} \frac{1}{i} \\
\leq &\frac{1}{p_{min}} \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \sum_{a_h \in A} \left(1 + \log\left(\sum_{t=1}^{T} \mathbb{I}[\pi_t(c;s_h) = a_h]\right)\right) \\
\leq &\frac{1}{p_{min}}(|S| \cdot |A| + |S| \cdot |A| \log(T/|A|)) \qquad \text{(Since } \textstyle\bigcup_{h=0}^{H} S_h^c = S\text{)}
\end{aligned}$$

$$= \frac{|S||A|}{p_{min}}(1 + \log(T/|A|)),$$

where the last inequality is due to Jensen's inequality. By taking expectation over c in both sides of the above inequality, we obtain the second part of the lemma. ■

B.4.6 DERIVING REGRET BOUND

Lemma 22 (equivalence of policies) *For any mapping $\widetilde{\mathcal{M}}$ from a context $c \in \mathcal{C}$ to a MDP, it holds that*

$$\max_{\pi \in \Pi_{\mathcal{C}}} \mathbb{E}_c \left[V^{\pi(c;\cdot)}_{\widetilde{\mathcal{M}}(c)}(s_0) \right] = \mathbb{E}_c \left[\max_{\pi \in S \to A} V^{\pi}_{\widetilde{\mathcal{M}}(c)}(s_0) \right].$$

Proof Holds trivially. ■

Lemma 23 (policy difference) *Under the good event of Lemma 18 , for any $t > |A|$ it holds that*

$$\mathbb{E}_c \left[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0) \right] \leq \mathbb{E}_c \left[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)} \right] + 2\beta_t \cdot \mathbb{E}_c \left[\sum_{h=0}^{H-1} \sum_{s_h \in S^c_h} \frac{q_h(s_h|\pi_t(c;\cdot), P^c_\star)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi_t(c;s_h) = \pi_i(c;s_h)] q_h(s_h|\pi_i(c;\cdot), P^c_\star)} \right] + 2\beta_t \frac{H|S||A|}{t}.$$

where $\pi^\star \in \Pi_{\mathcal{C}}$ is an optimal context-dependent policy, and π_t is the context-dependent policy selected in round t.

Proof Under the good event of Lemma 18, using Lemma 19, for every $t > |A|$, consider the following derivation.

$$\begin{aligned}
&\mathbb{E}_c[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0)] \\
&\leq \mathbb{E}_c[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(\hat{f}_t, P_\star)(c)}(s_0)] + \mathbb{E}_c \left[\sum_{h=0}^{H-1} \sum_{s_h \in S^c_h} \frac{\beta_t \cdot q_h(s_h|\pi^\star(c;\cdot), P^c_\star)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi^\star(c;s_h) = \pi_i(c;s_h)] q_h(s_h|\pi_i(c;\cdot), P^c_\star)} \right] + \beta_t \frac{H|S||A|}{t} \\
&\qquad \text{(By Lemma 19 applied for } \pi^\star\text{)} \\
&\leq \max_{\pi \in \Pi_{\mathcal{C}}} \left\{ \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(\hat{f}_t, P_\star)(c)}(s_0)] + \mathbb{E}_c \left[\sum_{h=0}^{H-1} \sum_{s_h \in S^c_h} \frac{\beta_t \cdot q_h(s_h|\pi(c;\cdot), P^c_\star)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c;s_h) = \pi_i(c;s_h)] q_h(s_h|\pi_i(c;\cdot), P^c_\star)} \right] \right\} + \beta_t \frac{H|S||A|}{t} \\
&= \max_{\pi \in \Pi_{\mathcal{C}}} \left\{ \mathbb{E}_c \left[V^{\pi(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0) \right] \right\} + \beta_t \frac{H|S||A|}{t} \qquad \text{(By equation (1))} \\
&= \mathbb{E}_c \left[V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0) \right] + \beta_t \frac{H|S||A|}{t} \qquad \text{(By Lemma 22 applied on the mapping } \widehat{\mathcal{M}}_t \text{ and } \pi_t \text{ choice)} \\
&= \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(\hat{f}_t, P_\star)(c)}] + \mathbb{E}_c \left[\sum_{h=0}^{H-1} \sum_{s_h \in S^c_h} \frac{\beta_t \cdot q_h(s_h|\pi_t(c;\cdot), P^c_\star)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi_t(c;s_h) = \pi_i(c;s_h)] q_h(s_h|\pi_i(c;\cdot), P^c_\star)} \right] + \beta_t \frac{H|S||A|}{t} \\
&\qquad \text{(By equation (1))} \\
&\leq \mathbb{E}_c \left[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)} \right] + 2\beta_t \cdot \mathbb{E}_c \left[\sum_{h=0}^{H-1} \sum_{s_h \in S^c_h} \frac{q_h(s_h|\pi_t(c;\cdot), P^c_\star)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi_t(c;s_h) = \pi_i(c;s_h)] q_h(s_h|\pi_i(c;\cdot), P^c_\star)} \right] + 2\beta_t \frac{H|S||A|}{t}, \\
&\qquad \text{(By Lemma 19 applied for } \pi_t\text{)}
\end{aligned}$$

as stated. ■

Theorem 24 (expected regret bound) *For any $T \geq 1$, finite functions class $\mathcal{F}$ and $\delta \in (0,1)$ let $\beta_t = \sqrt{\frac{17t\log(4|\mathcal{F}|t^3/\delta)}{|S||A|}}$ for all $t \in [T]$. Then, with probability at least $1-\delta$ it holds that*

$$\textit{E.Regret}_T(RM-KD) \leq \tilde{O}\left(\frac{\sqrt{T|S||A|\log\frac{|\mathcal{F}|}{\delta}}}{p_{min}} + H\sqrt{T|S||A|\log\frac{|\mathcal{F}|}{\delta}} + |A|H\right).$$

Proof We prove the theorem under the good event stated in Lemma 18, which holds with probability at least $1-\delta/2$. By the optimism lemma (Lemma 23) , under the good event, for all $t > |A|$ it holds that

$$\begin{aligned}\mathbb{E}_c\left[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0) - V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)\right] \leq & 2\beta_t \cdot \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi_t(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi_t(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}\right] \\ & + 2\beta_t\frac{H|S||A|}{t}.\end{aligned} \tag{12}$$

In expectation we have

$$\begin{aligned}\text{E.Regret}_T(RM-KD) =& \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_c\left[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0) - V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)\right]\right] \\ =& \sum_{t=1}^{T}\mathbb{E}_{\mathbb{H}_{t-1}}\left[\mathbb{E}_c\left[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0) - V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)\middle|\mathbb{H}_{t-1}\right]\right].\end{aligned}$$

(By linearity of expectation, since the algorithm is deterministic.)

Thus, since π_t is determined completely by the history $\mathbb{H}_{t-1}$, by Azuma's inequality, with probability at least $1-\delta/2$ over the policies π_t, the following holds.

$$\text{E.Regret}_T(RM-KD) \leq \sum_{t=1}^{T}\mathbb{E}_c\left[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0) - V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)\right] + 2H\sqrt{2T\log(4/\delta)} \quad \text{(By Azuma's inequality)}$$

$$\leq \sum_{t=|A|+1}^{T}\mathbb{E}_c[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0) - V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)] + 2H\sqrt{2T\log(4/\delta)} + |A|H$$

$$\leq \sum_{t=|A|+1}^{T}\left(2\beta_t\cdot\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi_t(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi_t(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}\right] + 2\beta_t\frac{H|S||A|}{t}\right)$$
$$+ 2H\sqrt{2T\log(4/\delta)} + |A|H \quad \text{(By inequality (12))}$$

$$\leq 2\beta_T\cdot\sum_{t=|A|+1}^{T}\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi_t(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi_t(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}\right]$$
$$+ \sum_{t=|A|+1}^{T}2\beta_t\frac{H|S||A|}{t} + 2H\sqrt{2T\log(4/\delta)} + |A|H \quad \text{(Since } \beta_T \geq \beta_t \text{ for all } t \leq T.\text{)}$$

$$\leq 2\beta_T\cdot\frac{|S||A|}{p_{min}}(1+\log(T/|A|)) + \sum_{t=|A|+1}^{T}2\beta_t\frac{H|S||A|}{t}$$
$$+ 2H\sqrt{2T\log(4/\delta)} + |A|H \quad \text{(By Lemma 21 part 2)}$$

$$\leq 2\frac{|S||A|}{p_{min}}\sqrt{\frac{17\log(4|\mathcal{F}|T^3/\delta)T}{|S||A|}}(1+\log(T/|A|)) \quad \text{(By } \beta_t \text{ choice)}$$

$$
\begin{aligned}
&+ 2\sqrt{17\log(4|\mathcal{F}|T^3/\delta)}\sum_{t=|A|+1}^{T}\sqrt{\frac{t}{|S||A|}}\frac{H|S||A|}{t} \\
&+ 2H\sqrt{2T\log(4/\delta)} + |A|H \\
\leq& 2\frac{(1+\log(T/|A|))}{p_{min}}\cdot\sqrt{17\log(4|\mathcal{F}|T^3/\delta)\cdot T|S||A|} + 2H\sqrt{17\log(4|\mathcal{F}|T^3/\delta)T|S||A|} \\
&+ 2H\sqrt{2T\log(4/\delta)} + |A|H \\
=& \tilde{O}\left(\frac{\sqrt{T|S||A|\log\frac{|\mathcal{F}|}{\delta}}}{p_{min}} + H\sqrt{T|S||A|\log\frac{|\mathcal{F}|}{\delta}} + |A|H\right).
\end{aligned}
$$

By union bound, the expected regret bound above holds with probability at least $1-\delta$. ■

Consider the regret, which defined as

$$\text{Regret}_T(\text{ALG}) := \sum_{t=1}^{T} V_{\mathcal{M}(c_t)}^{\pi^\star(c_t;\cdot)} - V_{\mathcal{M}(c_t)}^{\pi_t(c_t;\cdot)}.$$

The following theorem establish regret bound, which holds with high probability.

Theorem 25 (regret bound) *For every $T \geq 1$, finite functions class $\mathcal{F}$ and $\delta \in (0,1)$ let $\beta_t = \sqrt{\frac{17t\log(4|\mathcal{F}|t^3/\delta)}{|S||A|}}$ for all $t \in [T]$. Then, with probability at least $1-\delta$ it holds that*

$$\textit{Regret}_T(RM-KD) \leq \tilde{O}\left(\frac{\sqrt{T|S||A|\log\frac{|\mathcal{F}|}{\delta}}}{p_{min}} + H\sqrt{T|S||A|\log\frac{|\mathcal{F}|}{\delta}} + |A|H\right).$$

Proof We prove the theorem under the good event stated in Lemma 18, which holds with probability at least $1-\delta/2$. By the optimism lemma (Lemma 23), under the good event, for all $t > |A|$ it holds that

$$
\begin{aligned}
\mathbb{E}_c\left[V_{\mathcal{M}(c)}^{\pi^\star(c;\cdot)}(s_0) - V_{\mathcal{M}(c)}^{\pi_t(c;\cdot)}(s_0)\right] \leq& 2\beta_t\cdot\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi_t(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi_t(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right] \\
&+ 2\beta_t\frac{H|S||A|}{t}.
\end{aligned} \tag{13}
$$

Recall that $\mathbb{H}_t = (\sigma^1,\ldots,\sigma^t)$. Consider the martingale difference sequence $\{Y_t\}_{t=1}^T$ and the filtration $\{\mathbb{H}_t\}_{t=1}^T$ where

$$Y_t := V_{\mathcal{M}(c_t)}^{\pi^\star(c_t,\cdot)}(s_0) - V_{\mathcal{M}(c_t)}^{\pi_t(c_t,\cdot)}(s_0) - \mathbb{E}_{c_t}\left[V_{\mathcal{M}(c_t)}^{\pi^\star(c_t,\cdot)}(s_0) - V_{\mathcal{M}(c_t)}^{\pi_t(c_t,\cdot)}(s_0)\middle|\mathbb{H}_{t-1}\right].$$

Clearly, for all t, $|Y_t| \leq 2H$, Y_t is determined completely by $\mathbb{H}_t$ and $\mathbb{E}\left[Y_t|\mathbb{H}_{t-1}\right] = 0$ since

$$
\begin{aligned}
\mathbb{E}\left[Y_t|\mathbb{H}_{t-1}\right] =& \mathbb{E}_{c_t}\left[V_{\mathcal{M}(c_t)}^{\pi^\star(c_t,\cdot)}(s_0) - V_{\mathcal{M}(c_t)}^{\pi_t(c_t,\cdot)}(s_0)\middle|\mathbb{H}_{t-1}\right] - \mathbb{E}_{c_t}\left[V_{\mathcal{M}(c_t)}^{\pi^\star(c_t,\cdot)}(s_0) - V_{\mathcal{M}(c_t)}^{\pi_t(c_t,\cdot)}(s_0)\middle|\mathbb{H}_{t-1}\right] \\
=& 0. \qquad \text{(Since } \pi_t \text{ is determined by } \mathbb{H}_{t-1}\text{, and } c_t \text{ and } \pi^\star \text{ are independent of the history } \mathbb{H}_{t-1}\text{)}
\end{aligned}
$$

Hence, by Azuma's inequality, with probability at least $1-\delta/2$ it holds that

$$\text{Regret}_T(RM-KD) \leq \sum_{t=1}^{T}\mathbb{E}_{c_t}\left[V_{\mathcal{M}(c_t)}^{\pi^\star(c_t;\cdot)}(s_0) - V_{\mathcal{M}(c_t)}^{\pi_t(c_t;\cdot)}(s_0)\middle|\mathbb{H}_{t-1}\right] + 2H\sqrt{2T\log(4/\delta)}.$$

Since π_t is determined completely (and deterministically) by $\mathbb{H}_{t-1}$, and c_t and $\pi^\star$ are independent of the history $\mathbb{H}_{t-1}$, we can omit the conditioning on $\mathbb{H}_{t-1}$. Thus, we obtain,

$$\begin{aligned}
&\text{Regret}_T(RM-KD) \leq \sum_{t=1}^{T} \mathbb{E}_c\left[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0) - V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)\right] + 2H\sqrt{2T\log(4/\delta)} && \text{(By Azuma's inequality)}\\
&\leq \sum_{t=|A|+1}^{T} \mathbb{E}_c[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0) - V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)] + 2H\sqrt{2T\log(4/\delta)} + |A|H\\
&\leq \sum_{t=|A|+1}^{T}\left(2\beta_t\cdot\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S^c_h}\frac{q_h(s_h|\pi_t(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi_t(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}\right] + 2\beta_t\frac{H|S||A|}{t}\right)\\
&\quad + 2H\sqrt{2T\log(4/\delta)} + |A|H && \text{(By inequality (13))}\\
&\leq 2\beta_T\cdot\sum_{t=|A|+1}^{T}\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S^c_h}\frac{q_h(s_h|\pi_t(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi_t(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}\right]\\
&\quad + \sum_{t=|A|+1}^{T} 2\beta_t\frac{H|S||A|}{t} + 2H\sqrt{2T\log(4/\delta)} + |A|H && \text{(Since } \beta_T\geq\beta_t \text{ for all } t\leq T.)\\
&\leq 2\beta_T\cdot\frac{|S||A|}{p_{min}}(1+\log(T/|A|)) + \sum_{t=|A|+1}^{T} 2\beta_t\frac{H|S||A|}{t}\\
&\quad + 2H\sqrt{2T\log(4/\delta)} + |A|H && \text{(By Lemma 21 part 2)}\\
&\leq 2\frac{|S||A|}{p_{min}}\sqrt{\frac{17\log(4|\mathcal{F}|T^3/\delta)T}{|S||A|}}(1+\log(T/|A|)) && \text{(By } \beta_t \text{ choice)}\\
&\quad + 2\sqrt{17\log(4|\mathcal{F}|T^3/\delta)}\sum_{t=|A|+1}^{T}\sqrt{\frac{t}{|S||A|}}\frac{H|S||A|}{t}\\
&\quad + 2H\sqrt{2T\log(4/\delta)} + |A|H\\
&\leq 2\frac{(1+\log(T/|A|)}{p_{min}}\cdot\sqrt{17\log(4|\mathcal{F}|T^3/\delta)\cdot T|S||A|} + 2H\sqrt{17\log(4|\mathcal{F}|T^3/\delta)T|S||A|}\\
&\quad + 2H\sqrt{2T\log(4/\delta)} + |A|H\\
&= \tilde{O}\left(\frac{\sqrt{T|S||A|\log\frac{|\mathcal{F}|}{\delta}}}{p_{min}} + H\sqrt{T|S||A|\log\frac{|\mathcal{F}|}{\delta}} + |A|H\right).
\end{aligned}$$

By union bound, the regret bound above holds with probability at least $1-\delta$. ■

Corollary 26 (regret bound in terms of $\mathcal{G}$) *For every $T\geq 1$, finite function class $\mathcal{G}$ ($\mathcal{F}=\mathcal{G}^S$) and $\delta\in(0,1)$, the following holds with probability at least $1-\delta$, for the same choice of parameters $\{\beta_t\}_{t\in[T]}$.*

$$\textit{Regret}_T(RM-KD) \leq \tilde{O}\left(\frac{|S|\sqrt{T|A|\log\frac{|\mathcal{G}|}{\delta}}}{p_{min}} + H|S|\sqrt{T|A|\log\frac{|\mathcal{G}|}{\delta}} + |A|H\right).$$

Proof Plug $\log(|\mathcal{F}|)=|S|\log(|\mathcal{G}|)$ in the bound of Theorem 25. ■

Appendix C. Unknown and Context-Independent Dynamics

C.1 Assumptions and Notations

Unknown and context-independent dynamics. Meaning, for every context $c \in \mathcal{C}$ we have $P_\star^c = P_\star$, i.e., all the contexts has the same dynamics. In addition, $P_\star$ is unknown to the learner. Recall we assume the CMDP is layered. Since the dynamics is context-independent so is the partition of the states space into layers. Hence, we denote by $S_0, S_1, \ldots S_H$ the disjoint layers of the CMDP. As before, $S_0 = \{s_0\}$, $S_H = \{s_H\}$, where s_0, s_H are unique start and final states, respectively, and $S = \bigcup_{h\in[H]} S_h$. We assume that the (context-independent) partition to layers is known to the learner.

Known minimum reachability parameter. We assume that there exists $p_{min} \in (0, 1]$ such that for each layer $h \in [H]$, state $s_h \in S_h$ and context $c \in \mathcal{C}$ for every context-dependent policy $\pi \in \Pi_{\mathcal{C}}$ it holds that $q_h(s_h|\pi(c;\cdot), P_\star) \geq p_{min}$. We remark in this section, we assume that p_{min} is **known** to the learner.

Q-function. Given a policy π and a MDP $M = (S, A, P, r, s_0, H)$, the $h \in [H-1]$ stage Q-function of a state $s \in S_h$ and an action $a \in A$ is defined as

$$Q^\pi_{M,h}(s,a) = \mathbb{E}_{\pi,M}\Big[\sum_{k=h}^{H-1} r(s_k, \pi(s_k))|s_h = s, a_h = a\Big].$$

For completeness we define $Q^\pi_{M,H}(s,a) = 0$ for all $(s,a) \in S \times A$. For brevity, when $h = 0$ we denote $Q^\pi_{M,0}(s_0,a) := Q^\pi_M(s_0,a)$.

Recall Bellman's equations for the Q-function. For every layer $h \in [H-1]$, state $s \in S_h$ and an action $a \in A$ it holds that

$$Q^\pi_{M,h}(s,a) = r(s,a) + \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\left[V^\pi_{M,h+1}(s')\right].$$

C.2 Algorithm

In Algorithm RM-UCID (Algorithm 5), the first $|A|$ rounds are initialization rounds, where in round $i \in \{1, 2, \ldots, |A|\}$ the agent plays the policy π_i which always choose action a_i, i.e., $\pi_i(c; s) = a_i$ for every context $c \in \mathcal{C}$ and state $s \in S$. She updates the LSR oracle and the counters $N_i(s,a)$, $N_i(s,a,s')$ for all $(s,a,s') \in S \times A \times S$ using the observed trajectory σ^i.

At every round $t = |A| + 1, \ldots, T$ the agent observes a context c_t and computes the optimistic approximated model $\widehat{\mathcal{M}}_t(c)$ and its optimal deterministic policy $\pi_t(c_t;\cdot)$.

Similarly to Algorithm RM-KD (i.e., Algorithm 4), to compute the optimistic rewards function for the context c_t, $\widehat{r}^{c_t}_t$, the agent needs to compute $\pi_k(c_t;\cdot)$ for all $k = |A|+1, \ldots, t-1$. To compute $\pi_k(c_t;\cdot)$ she needs to compute both $\widehat{r}^{c_t}_k$ and the optimistic dynamics $\widehat{P}^{c_t}_k$, for all $k = |A|+1, \ldots, t-1$. Hence, the agent performs the following steps.

For all $k = |A|+1, \ldots, t-1$:

(1) The agent computes the empirical dynamics at round k, which defined as, $\bar{P}_k(s'|s,a) = \frac{N_k(s,a,s')}{\max\{1, N_k(s,a)\}}$ for all $(s,a,s') \in S \times A \times S$.

(2) The agent computes an approximated rewards function $\widehat{r}^{c_t}_k$ for the context c_t using the approximation computed in round k, $\widehat{f}_k$, the policies $\{\pi_i(c_t;\cdot)\}_{i=1}^{k-1}$ and the minimum reachability parameter p_{min}.

(3) The agent computes the optimistic dynamics $\widehat{P}^{c_t}_k$ and a deterministic context-dependent policy $\pi_k(c_t;\cdot)$ by invoking Algorithm FOA (see Algorithm 7). We now briefly describe Algorithm FOA. Algorithm FOA finds an optimistic dynamics and policy, with respect to the rewards function $\widehat{r}^{c_t}_k$, as follows.

(a) The first step is to solve the linear program (14), and by that compute dynamics $\widehat{P}^{c_t}_k$ and a stochastic policy $\pi^{\widehat{q}^{c_t}_k}$ which achieve maximal value for the rewards function $\widehat{r}^{c_t}_k$, under the constraints that for all $(s,a) \in S \times A$, the optimistic dynamics satisfies that $\left\|\widehat{P}^{c_t}_k - \bar{P}_k\right\|_1 \leq \xi_k(s,a)$, where ξ_k is a confidence bound defined in the algorithm.

(b) The second step is to invoke Algorithm FDP (see Algorithm 6) which return a deterministic policy $\pi_k(c_t;\cdot)$ which obtains value that equals or higher to that of $\pi^{\widehat{q}_k^{c_t}}$ on the MDP $\widehat{\mathcal{M}}_t(c_t)$ which defined by the rewards function $\widehat{r}_k^{c_t}$ and the dynamics $\widehat{P}_k^{c_t}$.

Lastly, the agent computes $\pi_t(c_t;\cdot)$ in the same manner described bellow. Then she runs $\pi_t(c_t;\cdot)$ to generate a trajectory σ^t and update the LSR oracle and the counters using σ^t.

For more details, see Algorithm 5.

Algorithm 5 Regret Minimization for Unknown Context Independent Dynamics (RM-UCID)

inputs:

- MDP parameters: $S=\{S_0,S_1,\ldots,S_H\}$, A, H, s_0
- Confidence parameter δ and tuning parameters $\{\beta_t\}_{t=1}^T$.
- Minimum reachability parameter p_{min}.

initialize counters $N_t(s,a)=0$, $N_t(s,a,s')=0$ for all $(s,a,s')\in S\times A\times S, t\in\{1,2,\ldots,T\}$.
for round $i=1,\ldots,|A|$ **do**
  run policy π_i which always selects action a_i
  update counters according to the observed tuples (s_h^i,a_h^i,s_{h+1}^i) for all $i\in[|A|],h\in[H-1]$
for round $t=|A|+1,\ldots,T$ **do**
  compute $\widehat{f}_t\in\arg\min_{f\in\mathcal{F}}\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}(f(c_i,s_h^i,a_h^i)-r_h^i)^2$ using the LSR oracle
  observe context $c_t\in\mathcal{C}$
  for $k=|A|+1,\ldots,t$ **do**
    compute for all $(s,a,s')\in S\times A\times S:\ \ \bar{P}_k(s'|s,a)=\frac{N_k(s,a,s')}{\max\{1,N_k(s,a)\}}$
    compute $\forall(s,a)\in S\times A:\ \widehat{r}_k^{c_t}(s,a)=\widehat{f}_k(c_t,s,a)+\min\left\{\frac{1}{p_{min}}\cdot\frac{\beta_k}{\sum_{i=1}^{k-1}\mathbb{I}[a=\pi_i(c_t,s)]},1\right\}$
    find optimistic model $\widehat{\mathcal{M}}_k(c_t)=(S,A,\widehat{P}_k^{c_t},\widehat{r}_k^{c_t},s_0,H)$ and a policy for it $\pi_k(c_t,\cdot)$ using algorithm FOA (Algorithm 7)
  play $\pi_t(c_t,\cdot)$ and observe trajectory $\sigma^t=(c_t,s_0^t,a_0^t,r_0^t,s_1^t,\ldots,s_{H-1}^t,a_{H-1}^t,r_{H-1}^t,s_H^t)$.
  update counters:

$$N_{t+1}(s,a)=N_t(s,a)+\mathbb{I}[(s,a)\in\sigma^t],$$
$$N_{t+1}(s,a,s')=N_t(s,a,s')+\mathbb{I}[(s,a,s')\in\sigma^t]\ \ \forall(s,a,s')\in S\times A\times S$$

Algorithm 6 Find Deterministic Policy (FDP)

inputs:

- Layered MDP $M=(S,A,P,r,s_0,H)$, where $S=\{S_0,S_1,\ldots,S_H\}$.
- Stochastic (time-invariant) Markovian policy π.

for $h=H,H-1,\ldots,0$ **do**
  compute $Q^{\pi}_{M,h}(s,a)$ for all $(s,a)\in S_h\times A$ ▷ can be computed in $poly(|S|,|A|,H)$ time.
  for $s\in S_h$ **do**
    choose $\pi'(s)\in\arg\max_{a\in A}Q^{\pi}_{M,h}(s,a)$
return: π'

Remark 27 *The optimistic approximated dynamics at round t, $\widehat{P}_t$, is context-dependent dynamics although the true dynamics is not. This is since the rewards function is context-dependent and for every context the optimistic dynamics is computed with respect to the rewards function of that context.*

Algorithm 7 Find Optimistic Approximation (FOA)

1: **inputs:**

- MDP parameters: $S = \{S_0, S_1, \ldots, S_H\}$, A, s_0, H.
- Confidence parameter δ.
- Round k, counters $N_k(s,a)$ for all $(s,a) \in S \times A$, and the empirical dynamics $\bar{P}_k$.
- Context c and the approximated rewards function $\widehat{r}_k^c$.

2: compute confidence intervals over the dynamics $\xi_k(s,a) = \sqrt{\frac{|S|+\lambda}{\max\{1, N_k(s,a)\}}}$, where $\lambda = 2\log(4|S||A|T^2/\delta)$.

3: solve the following linear program

$$
\begin{aligned}
&\max_{q} \sum_{h=0}^{H-1} \sum_{s \in S_h} \sum_{a \in A} \widehat{r}_k^c(s,a) \cdot \sum_{s' \in S} q(s,a,s') \\
&\text{subject to:} \\
&q(s,a,s') \in [0,1], \quad \forall h \in [H-1], (s,a,s') \in S \times A \times S \\
&\sum_{s \in S_h} \sum_{a \in A} \sum_{s' \in S_{h+1}} q(s,a,s') = 1, \quad \forall h \in [H-1] \\
&\sum_{s' \in S_{h+1}} \sum_{a \in A} q(s,a,s') = \sum_{s' \in S_{h-1}} \sum_{a \in A} q(s',a,s), \quad \forall h \in \{1, \ldots, H-1\}, s \in S_h \\
&\sum_{s' \in S_{h+1}} \left| q(s,a,s') - \bar{P}_k(s'|s,a) \cdot \sum_{s'' \in S_{h+1}} q(s,a,s'') \right| \leq \xi_k(s,a) \cdot \sum_{s'' \in S_{h+1}} q(s,a,s'') \\
&\forall h \in [H-1], (s,a) \in S_h \times A
\end{aligned}
\tag{14}
$$

4: denote by $\widehat{q}_k^c$ the solution of LP 14 and compute the induced policy and dynamics

$$
\pi^{\widehat{q}_k^c}(a|s) = \frac{\sum_{s' \in S_{h+1}} \widehat{q}_k^c(s,a,s')}{\sum_{a' \in A} \sum_{s' \in S_{h+1}} \widehat{q}_k^c(s,a',s')}, \quad \forall h \in [H-1], s \in S_h, a \in A
$$

and

$$
\widehat{P}_k^c(s'|s,a) = \frac{\widehat{q}_k^c(s,a,s')}{\sum_{s'' \in S_{h+1}} \widehat{q}_k^c(s,a,s'')}, \quad \forall h \in [H-1], s \in S_h, a \in A
$$

let $\widehat{\mathcal{M}}_k(c) = (S, A, \widehat{P}_k^c, \widehat{r}_k^c, s_0, H)$

5: compute a deterministic policy $\pi_k(c,\cdot) \leftarrow \texttt{FDP}(\widehat{\mathcal{M}}_k(c), \pi^{\widehat{q}_k^c})$

6: **return:** $\pi_k(c,\cdot)$ and $\widehat{\mathcal{M}}_k(c) = (S, A, \widehat{P}_k^c, \widehat{r}_k^c, s_0, H)$ that solve 18

C.3 Regret Analysis

For every $t > |A|$ we define the following MDPs for every context $c \in \mathcal{C}$,

1. $\mathcal{M}^{(f,P_\star)}(c) = (S, A, P_\star, f(c,\cdot,\cdot), s_0, H)$, for any $f \in \mathcal{F}$ and the true dynamics $P_\star$.
2. $\mathcal{M}^{(f_\star,P_\star)}(c)$ is the true model, where $f_\star$ is the true context dependent rewards and $P_\star$ is the true dynamics, which we also denote by $\mathcal{M}(c)$.
3. $\mathcal{M}^{(\widehat{r}_t,P_\star)}(c) = (S, A, P_\star, \widehat{r}_t^c, s_0, H)$ is the model where $\widehat{r}_t^c$ is the approximated rewards in round t defined in Algorithm 5 and $P_\star$ is the true dynamics.
4. $\widehat{\mathcal{M}}_t(c) = \mathcal{M}^{(\widehat{r}_t,\widehat{P}_t)}(c)$ is the approximated MDP in time $t > |A|$ defined in Algorithm 5.

C.3.1 ANALYSIS OUTLINE

We analyse the regret of Algorithm RM-UCID under the following two good events.

Event G_1: confidence bound over policies w.r.t rewards estimation. Intuitively, the good event G_1 states that the following confidence bound over the expected value difference holds for every $t > |A|$ and context-dependent policy $\pi \in \Pi_{\mathcal{C}}$. We show that G_1 holds with high probability.

Formally, in Lemma 37 we prove that with probability at least $1 - \delta/4$, for all $t > |A|$ and every context-dependent policy $\pi \in \Pi_{\mathcal{C}}$ it holds that

$$\begin{aligned}&\left|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{f}_t,P_\star)}(c)}(s_0)]\right| \\ \leq &\sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\frac{q_h(s_h|\pi,P)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star)}\right]}\cdot\sqrt{68H\log(8|\mathcal{F}|t^3/\delta)}.\end{aligned}$$

We derive that good event similarly to shown for the known dynamics setting. It follows using the results of Lemma 36. For more details, see Subsection C.3.3.

Event G_2: Confidence interval over the empirical dynamics. Intuitively, it states that the following confidence bound over the empirical dynamics $\bar{P}_t$ holds, for every round $t \in \{1, 2, \ldots, T\}$ and state-action pair $(s, a) \in S \times A$. We show that G_2 holds with high probability.

Formally,

Lemma 28 *With probability at least $1 - \delta/4$, for every $t \in \{1, 2, \ldots, T\}$, $k \in \{1, 2, \ldots, t\}$ and state-action pair $(s, a) \in S \times A$ it holds that*

$$\|P_\star(\cdot|s,a) - \bar{P}_k(\cdot|s,a)\|_1 \leq \xi_k(s,a),$$

where $\xi_k(s,a) = \sqrt{\frac{|S|+2\log(4|S||A|T^2/\delta)}{\max\{1,N_k(s,a)\}}}$.

Proof The lemma holds by Bretagnolle Huber-Carol inequality (see Lemma 74) and union bound for all $t \in \{1, 2, \ldots, T\}$, $k \in \{1, 2, \ldots, t\}$ and $(s, a) \in S \times A$. ■

In the following, we first analyse the error caused by the dynamic approximation (see Sub-subsection C.3.2). The analysis consist of a two main steps.

Step 1: In Lemma 32 we show that under the good event G_2, for all $t > |A|$ and every context $c \in \mathcal{C}$, it holds that

$$V^{\pi_t(c,\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0) \geq V^{\pi^\star(c,\cdot)}_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}, \tag{15}$$

where $\pi^\star$ is the optimal context-dependent policy and π_t is the selected context-dependent policy at round t. To prove the above inequality, we use the results of Lemmas 29 and 30.

Step 2: In Lemma 35 we show that under the good event G_2, with probability at least $1-\delta/4$ it holds that

$$\sum_{t=|A|+1}^{T} \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,\widehat{P}_t)(c)}] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,P_\star)(c)}] \leq 4H\sqrt{|S|+2\log(4|S||A|T^2/\delta)}\left(\sqrt{2TH\log(8/\delta)}+2\sqrt{|S||A|T}\right). \tag{16}$$

To prove the lemma, we first show the approximate rewards function is bounded for every context $c \in \mathcal{C}$ (see Lemma 33), and then use value difference lemma (Lemma 73), and the good event assumption to obtain the above bound.

To analyse the error caused by the rewards approximation (see Subsubsections C.3.4)), we repeat the four steps analysis presented for the known dynamics setting (see Appendix B). We use steps 1 and 2 (Lemmas 36 and 37) to derive the good even G_1. In step 3, Lemma 38, we relax the confidence bound to be additive. In step 4 (Lemma 41) we bound the sum of contextual potential functions.

Using the results of steps 3 and 4, in Lemma 42 we obtain for all $T > |A|$ that

$$\begin{aligned}&\sum_{t=|A|+1}^{T} \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,P_\star)(c)}] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}] \\ \leq& H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} + 2(1+\log(T/|A|))\frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|S||A|}}{p_{min}}.\end{aligned} \tag{17}$$

To derive the regret bound, using inequality (15) we prove the optimism lemma (Lemma 43) which states that under the good events G_1 and G_2 for all $t > |A|$ it holds that

$$\mathbb{E}_c\left[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0)\right] - \mathbb{E}_c\left[V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)\right] \leq \beta_t\left(\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\frac{q_h(s_h,\pi^\star(c;s_h)|\pi^\star(c;\cdot),P_\star)}{p_{min}\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]}\right]+\frac{H|S||A|}{t}\right).$$

Lastly, in Theorems 45 and 44 we combine the result of the optimism lemma (Lemma 43), with inequalities (16) and (17) and the cumulative contextual potential bounds (Lemma 41) to obtain a regret bound (and an expected regret bound) of

$$\tilde{O}\left(H|S|\sqrt{T|A|}\log\frac{1}{\delta}+\max\{H,1/p_{min}\}\cdot\sqrt{|S||A|T\log\frac{|\mathcal{F}|}{\delta}}+|A|H\right).$$

which holds with high probability.

We remark that in the following analysis, we use the explicit expression of the contextual potential ψ_t.

C.3.2 ANALYSING THE ERROR CAUSED BY THE DYNAMICS APPROXIMATION

Analysis of Algorithm FDP. In the analysis, we use the following lemma which states that the deterministic policy computed by Algorithm FDP has equal or higher value then the stochastic policy it gets as an input.

Lemma 29 *For any Layered MDP M and stochastic policy π, Algorithm FDP (i.e., Algorithm 6) returns a deterministic policy π' for which $V^{\pi'}_M(s_0) \geq V^{\pi}_M(s_0)$.*

Proof Fix a Layered MDP M with H layers and a stochastic policy π. Let π' denote the deterministic policy returned by algorithm FDP for M and π.

We prove by backwards induction on the horizon h that for all $h \in [H]$ and $s \in S_h$ it holds that

$$V^{\pi'}_{M,h}(s) \geq V^{\pi}_{M,h}(s).$$

The above yields the lemma for $h = 0$ and s_0.

Base case, $h = H$. Holds trivially since $V^{\pi'}_{M,H}(s) = V^{\pi}_{M,H}(s) = 0$, for all $s \in S_H$.

Induction step. Assume the induction hypothesis holds for all $k \in \{h+1, \ldots, H\}$. We prove it holds for h.

For any $s \in S_h$ consider the following derivation,

$$\begin{aligned}
V^{\pi}_{M,h}(s) &= \sum_{a \in A} \pi(a|s) \cdot Q^{\pi}_{M,h}(s,a) \\
&\leq \max_{a \in A} Q^{\pi}_{M,h}(s,a) \\
&= Q^{\pi}_{M,h}(s, \pi'(s)) && \text{(By } \pi' \text{ definition)} \\
&= r(s,\pi'(s)) + \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,\pi'(s))} \left[V^{\pi}_{M,h+1}(s') \right] && \text{(By Bellman equations for the } Q \text{ function)} \\
&\leq r(s,\pi'(s)) + \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,\pi'(s))} \left[V^{\pi'}_{M,h+1}(s') \right] && \text{(By the induction hypothesis)} \\
&= V^{\pi'}_{M,h}(s). && \text{(By Bellman equations for the value function, since } \pi' \text{ is a deterministic policy)}
\end{aligned}$$

Hence, the lemma follows. ■

Analysis of Algorithm FOA. The following lemmas shows that algorithm FOA returns an optimistic dynamics and stochastic policy, with respect to the approximated rewards function.

Lemma 30 *Assume the good event G_2 holds. Then, for all $k > |A|$ and a context $c \in \mathcal{C}$ the linear program (14) is equivalent to the following constraint maximization problem,*

$$\begin{aligned}
&\max_{(\pi,P)} V^{\pi}_{\mathcal{M}^{(\widehat{r}_k,P)}(c)}(s_0) \\
&\quad \textit{subject to:} \\
&\quad \pi \in S \to \Delta(A) \\
&\quad \forall (s,a) \in S \times A: \quad P(\cdot|s,a) \in \Delta(S) \\
&\quad \forall (s,a) \in S \times A: \quad \|P(\cdot|s,a) - \bar{P}_k(\cdot|s,a)\|_1 \leq \xi_k(s,a)
\end{aligned} \tag{18}$$

where $\bar{P}_k$ is the the empirical dynamics, $\widehat{r}_k$ is the approximated reward function at round k and we define $\mathcal{M}^{(\widehat{r}_k,P)}(c) := (S, A, P, \widehat{r}^c_k, s_0, H)$.

Proof Fix round $k > |A|$ and a context $c \in \mathcal{C}$.

Consider the following extended definition of the occupancy measure presented by Rosenberg and Mansour (2019) for any dynamics P and stochastic Markovian policy π,

$$q^{P,\pi}(s,a,s') := \mathbb{P}_{P,\pi}[s_h = s, a_h = a, s_{h+1} = s'], \;\; \forall h \in [H-1], (s,a,s') \in S_h \times A \times S_{h+1}.$$

Let $\Delta(M)$ denote the set of all extended occupancy measures.
We have that $q \in \Delta(M)$ if and only if q satisfies the following requirements.

1. $\sum_{s \in S_h} \sum_{a \in A} \sum_{s' \in S_{h+1}} q(s,a,s') = 1, \;\; \forall h \in [H-1]$.
2. $\sum_{s' \in S_{h+1}} \sum_{a \in A} q(s,a,s') = \sum_{s' \in S_{h-1}} \sum_{a \in A} q(s',a,s), \;\; \forall h \in \{1, \ldots, H-1\}, s \in S_h$.

Rosenberg and Mansour (2019) showed that $q \in \Delta(M)$ if and only if there exist a pair of stochastic Markovian policy and dynamics (π^q, P^q) for which

$$\pi^q(a|s) = \frac{\sum_{s' \in S_{h+1}} q(s,a,s')}{\sum_{a' \in A} \sum_{s' \in S_{h+1}} q(s,a',s')}, \;\; \forall h \in [H-1], s \in S_h, a \in A,$$

and

$$P^q(s'|s,a) = \frac{q(s,a,s')}{\sum_{s'' \in S_{h+1}} q(s,a,s'')}, \quad \forall h \in [H-1], s \in S_h, a \in A.$$

Using the extended occupancy measure definition, for $\mathcal{M}^{(\widehat{r}_k, P^q)}(c) = (S, A, P^q, \widehat{r}_k^c, s_0, H)$ and π^q it holds that

$$V^{\pi^q}_{\mathcal{M}^{(\widehat{r}_k, P^q)}(c)}(s_0) = \sum_{h=0}^{H-1} \sum_{s \in S_h} \sum_{a \in A} \widehat{r}_k^c(s,a) \sum_{s' \in S_{h+1}} q(s,a,s').$$

Hence, the objective functions of both maximization problems are equivalent.

In the following we show that the constraints of both maximization problems are equivalent as well.

Let $\bar{P}_k$ be the empirical dynamics at round k. In the maximization problem (18) we have the constraint $\|P(\cdot|s,a) - \bar{P}_k(\cdot|s,a)\|_1 \leq \xi_k(s,a)$, for all $(s,a) \in S \times A$. Rosenberg and Mansour (2019) showed we can equivalently apply the following constraint on the extended occupancy measure q

$$\sum_{s' \in S_{h+1}} \left| q(s,a,s') - \bar{P}_k(s'|s,a) \cdot \sum_{s'' \in S_{h+1}} q(s,a,s'') \right| \leq \xi_k(s,a) \cdot \sum_{s'' \in S_{h+1}} q(s,a,s''), \quad \forall h \in [H-1], (s,a) \in S_h \times A$$

and obtain that $\|P^q(\cdot|s,a) - \bar{P}_k(\cdot|s,a)\|_1 \leq \xi_k(s,a)$, for all $(s,a) \in S \times A$.

Hence, we showed the constrains on (π, P) in (18) are equivalent to the constrains on q in (14), and so are objective functions we maximize. Hence, the two optimization problems are equivalent.

■

Corollary 31 *Assume the good event G_2 holds. For all $k > |A|$ and any context $c \in \mathcal{C}$, let $\widehat{q}_k^c$ be an optimal solution for LP (14). Let us define the induced policy and dynamics,*

$$\pi^{\widehat{q}_k^c}(a|s) = \frac{\sum_{s' \in S_{h+1}} \widehat{q}_k^c(s,a,s')}{\sum_{a' \in A} \sum_{s' \in S_{h+1}} \widehat{q}_k^c(s,a',s')}, \quad \forall h \in [H-1], s \in S_h, a \in A,$$

and

$$\widehat{P}_k^c(s'|s,a) = \frac{\widehat{q}_k^c(s,a,s')}{\sum_{s'' \in S_{h+1}} \widehat{q}_k^c(s,a,s'')}, \quad \forall h \in [H-1], s \in S_h, a \in A.$$

Then, $\left(\pi^{\widehat{q}_k^c}, \widehat{P}_k^c\right)$ is an optimal solution of the maximization problem (18).

Moreover, $\pi^{\widehat{q}_k^c}$ is an optimal stochastic policy for the MDP $\mathcal{M}^{(\widehat{r}_k, \widehat{P}_k^c)}(c) := (S, A, \widehat{P}_k^c, \widehat{r}_k^c, s_0, H)$.

Proof The first part of the corollary is immediately implied by Lemma 30.

For the second part, assume that there exist a round k and a context c such that $\pi^{\widehat{q}_k^c}$ is not an optimal policy of the MDP $\mathcal{M}^{(\widehat{r}_k, \widehat{P}_k^c)}(c)$. Meaning, there exists a policy $\pi' : S \to \Delta(A)$ for which

$$V^{\pi'}_{\mathcal{M}^{(\widehat{r}_k, \widehat{P}_k^c)}(c)}(s_0) > V^{\pi^{\widehat{q}_k^c}}_{\mathcal{M}^{(\widehat{r}_k, \widehat{P}_k^c)}(c)}(s_0).$$

Under the good event G_2, $\left(\pi', \widehat{P}_k^c\right)$ is a feasible solution to the maximization problem (18), for the context c at round k, while $\left(\pi^{\widehat{q}_k^c}, \widehat{P}_k^c\right)$ is an optimal solution. Hence,

$$V^{\pi'}_{\mathcal{M}^{(\widehat{r}_k, \widehat{P}_k^c)}(c)}(s_0) \leq V^{\pi^{\widehat{q}_k^c}}_{\mathcal{M}^{(\widehat{r}_k, \widehat{P}_k^c)}(c)}(s_0),$$

a contradiction. ■

Lemma 32 (optimality of $(\pi_t, \widehat{P}_t)$ w.r.t $\widehat{r}_t$) *Assume the good event G_2 holds. Then, for all $t > |A|$ and every context $c \in \mathcal{C}$, it holds that*

$$V^{\pi_t(c,\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0) \geq V^{\pi^\star(c,\cdot)}_{\mathcal{M}(\widehat{r}_t, P_\star)(c)}.$$

Proof Fix a round $t > |A|$ and a context $c \in \mathcal{C}$. Under the good event G_2 the true dynamics $P_\star$ satisfies for all $k \in \{1, 2, \ldots, t\}$ that

$$\|P_\star - \bar{P}_k\|_1 \leq \xi_k(s,a), \;\; \forall (s,a) \in S \times A.$$

Let $\widehat{q}^c_t$ be the solution of LP (14) for round t and the context c. Let $\pi^{\widehat{q}^c_t}$ and $\widehat{P}^c_t$ be the induced policy and dynamics. By Corollary 31 we have that $(\pi^{\widehat{q}^c_t}, \widehat{P}^c_t)$ is an optimal solution for the maximization problem (18). Since $(\pi^\star(c;\cdot), P_\star)$ is a feasible solution for (18), by the optimality of $(\pi^{\widehat{q}^c_t}, \widehat{P}^c_t)$ and $\widehat{\mathcal{M}}_t(c)$ definition we have that

$$V^{\pi^{\widehat{q}^c_t}}_{\widehat{\mathcal{M}}_t(c)}(s_0) = V^{\pi^{\widehat{q}^c_t}}_{\mathcal{M}(\widehat{r}_t, \widehat{P}^c_t)(c)}(s_0) \geq V^{\pi^\star(c,\cdot)}_{\mathcal{M}(\widehat{r}_t, P_\star)(c)}.$$

Lastly, since $\pi_t(c;\cdot)$ is the deterministic policy returned by Algorithm FDP (Algorithm 6) for the inputs $\widehat{\mathcal{M}}_t(c)$ and $\pi^{\widehat{q}^c_t}$, by Lemma 29 it holds that

$$V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0) \geq V^{\pi^{\widehat{q}^c_t}}_{\widehat{\mathcal{M}}_t(c)}(s_0),$$

yielding the lemma since

$$V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0) \geq V^{\pi^{\widehat{q}^c_t}}_{\widehat{\mathcal{M}}_t(c)}(s_0) \geq V^{\pi^\star(c,\cdot)}_{\mathcal{M}(\widehat{r}_t, P_\star)(c)}.$$

■

Bound on the Cumulative Error of the Approximated Dynamics. We now bound the cumulative value difference caused by the dynamics approximation.

Lemma 33 (bound on the approximated rewards function) *For every round $t > |A|$ and a context $c \in \mathcal{C}$ it holds that $\widehat{r}^c_t : S \times A \to [0, 2]$, where $\widehat{r}^c_t$ is the rewards function defined in Algorithm 5.*

Proof Fix a round $t > |A|$ and a context $c \in \mathcal{C}$. By definition we have

$$\widehat{r}^c_t(s,a) = \widehat{f}_t(c,s,a) + \min\left\{ \frac{1}{p_{min}} \frac{\beta_t}{\sum_{i=1}^{t-1} \mathbb{I}[a = \pi_i(c;s)]}, 1 \right\}.$$

By $\mathcal{F}$ definition, it holds that $\widehat{f}_t : \mathcal{C} \times S \times A \to [0,1]$. In addition, by π_i choice for every $i \in \{1, \ldots, |A|\}$ it holds that

$$\sum_{i=1}^{t-1} \mathbb{I}[a = \pi_i(c;s)] \geq 1 \;\; \forall (s,a) \in S \times A.$$

Hence, for all $(s,a) \in S \times A$,

$$\frac{1}{p_{min}} \frac{\beta_t}{\sum_{i=1}^{t-1} \mathbb{I}[a = \pi_i(c;s)]} \geq 0.$$

Overall, for every state $s \in S$ and action $a \in A$ it holds that

$$0 \leq \widehat{r}^c_t(s,a) = \widehat{f}_t(c,s,a) + \min\left\{ \frac{1}{p_{min}} \frac{\beta_t}{\sum_{i=1}^{t-1} \mathbb{I}[a = \pi_i(c;s)]}, \; 1 \right\} \leq \widehat{f}_t(c,s,a) + 1 \leq 2.$$

Hence, the lemma follows. ■

Corollary 34 (value bound) *For every dynamics P, round $t > |A|$ and a context $c \in \mathcal{C}$ it holds that*

$$|V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,P)(c),h}(s)| \leq 2H \quad \forall \pi \in \Pi_{\mathcal{C}},\ s \in S,\ h \in [H].$$

Lemma 35 (cumulative value error due to dynamics approximation) *Assume the good event G_2 holds. Let $\{\pi_t \in \Pi_{\mathcal{C}}\}_{t=1}^T$ be the sequence of selected (deterministic) context-dependent policies. Then, with probability at least $1-\delta/4$ it holds that*

$$\sum_{t=|A|+1}^{T} \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,\widehat{P}_t)(c)}] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,P_\star)(c)}] \leq 4H\sqrt{|S| + 2\log(4|S||A|T^2/\delta)}\left(\sqrt{2TH\log(8/\delta)} + 2\sqrt{|S||A|T}\right).$$

Proof Assume the good event G_2 holds. Using value-difference Lemma 73 and Corollary 34 we obtain

$$\begin{aligned}
&\sum_{t=|A|+1}^{T} \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,\widehat{P}_t)(c)}] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,P_\star)(c)}] \\
&= \sum_{t=|A|+1}^{T} \mathbb{E}_c\left[\mathbb{E}_{\pi_t(c;\cdot),P_\star}\left[\sum_{h=0}^{H-1}\sum_{s'\in S}(\widehat{P}^c_t(s'|s_h,a_h) - P_\star(s'|s_h,a_h))V^{\pi_t(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,\widehat{P}_t)(c),h+1}(s')\right]\right] \\
&\qquad \text{(Value difference Lemma 73)} \\
&\leq 2H \sum_{t=|A|+1}^{T} \mathbb{E}_c\left[\mathbb{E}_{\pi_t(c;\cdot),P_\star}\left[\sum_{h=0}^{H-1}\|\widehat{P}^c_t(\cdot|s_h,a_h) - P_\star(\cdot|s_h,a_h)\|_1\right]\right] \qquad \text{(By Corollary 34)} \\
&\leq 2H \sum_{t=|A|+1}^{T} \mathbb{E}_c\left[\mathbb{E}_{\pi_t(c;\cdot),P_\star}\left[\sum_{h=0}^{H-1}\|\widehat{P}^c_t(\cdot|s_h,a_h) - \bar{P}_t(\cdot|s_h,a_h)\|_1 + \|\bar{P}_t(\cdot|s_h,a_h) - P_\star(\cdot|s_h,a_h)\|_1\right]\right] \\
&\qquad \text{(By triangle inequality)} \\
&\leq 2H \sum_{t=|A|+1}^{T} \mathbb{E}_c\left[\mathbb{E}_{\pi_t(c;\cdot),P_\star}\left[\sum_{h=0}^{H-1} 2\xi_t(s_h,a_h)\right]\right] \qquad \text{(Since } G_2 \text{ holds)} \\
&= 2H\sqrt{|S| + 2\log(4|S||A|T^2/\delta)} \sum_{t=|A|+1}^{T} \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A}\frac{q_h(s_h,a_h|\pi_t(c;\cdot),P_\star)}{\sqrt{\max\{1,N_t(s_h,a_h)\}}}\right] \\
&\qquad \text{(Since the MDP is layered)} \\
&\leq 2H\sqrt{|S| + 2\log(4|S||A|T^2/\delta)} \sum_{t=1}^{T} \mathbb{E}_{c_t}\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A}\frac{q_h(s_h,a_h|\pi_t(c_t;\cdot),P_\star)}{\sqrt{\max\{1,N_t(s_h,a_h)\}}}\right] \\
&\qquad \text{(By adding non-negative terms and "renaming" the context)} \\
&= 4H\sqrt{|S| + 2\log(4|S||A|T^2/\delta)}\sum_{t=1}^{T}\sum_{h=0}^{H-1}\left(\mathbb{E}_{c_t}\left[\sum_{s_h\in S_h}\sum_{a_h\in A}\frac{q_h(s_h,a_h|\pi_t(c_t;\cdot),P_\star)}{\sqrt{\max\{1,N_t(s_h,a_h)\}}}\right]\right. \\
&\quad \left. - \sum_{s_h\in S_h}\sum_{a_h\in A}\frac{\mathbb{I}[c_t,t,h,s_h,a_h]}{\sqrt{\max\{1,N_t(s_h,a_h)\}}} + \sum_{s_h\in S_h}\sum_{a_h\in A}\frac{\mathbb{I}[c_t,t,h,s_h,a_h]}{\sqrt{\max\{1,N_t(s_h,a_h)\}}}\right) \\
&\qquad \text{(By linearity of } \mathbb{E} \text{ and adding and subtracting indicators, see explanation bellow)} \\
&\underbrace{\leq}_{(i)} 4H\sqrt{|S| + 2\log(4|S||A|T^2/\delta)}\left(\sqrt{2TH\log(8/\delta)} + 2\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A}\frac{\mathbb{I}[c_t,t,h,s_h,a_h]}{\sqrt{\max\{1,N_t(s_h,a_h)\}}}\right) \\
&\qquad \text{(By Azuma's inequality, holds with probability at least } 1-\delta/4) \\
&\leq 4H\sqrt{|S| + 2\log(4|S||A|T^2/\delta)}\left(\sqrt{2TH\log(8/\delta)} + 2\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A}\sum_{i=1}^{\max\{1,N_t(s_h,a_h)\}}\frac{1}{\sqrt{i}}\right)
\end{aligned}$$

$$=4H\sqrt{|S|+2\log(4|S||A|T^2/\delta)}\left(\sqrt{2TH\log(8/\delta)}+\underbrace{2\sum_{t=1}^{T}\sum_{s\in S}\sum_{a\in A}\sum_{i=1}^{\max\{1,N_t(s,a)\}}\frac{1}{\sqrt{i}}}_{\leq\sqrt{|S||A|T}}\right)$$

(Since $S=\bigcup_{h=0}^{H-1}S_h$)

$$\leq 4H\sqrt{|S|+2\log(4|S||A|T^2/\delta)}\left(\sqrt{2TH\log(8/\delta)}+2\sqrt{|S||A|T}\right).$$

We denote by $\mathbb{I}[c_t,t,h,s_h,a_h]$ an indicator which indicates whether at time-step h of round t, when playing $\pi_t(c_t;\cdot)$, the agent visited s_h and played a_h where the context is c_t. Inequality (i) holds with probability at least $1-\delta/4$ by Azuma's inequality.

■

C.3.3 DERIVING THE GOOD EVENT FOR THE REWARDS APPROXIMATION

Step 1: Establishing uniform convergence bound over $\mathcal{F}$

Lemma 36 (uniform convergence over all sequences of estimators) *For any $\delta\in(0,1)$, with probability at least $1-\delta/4$ it holds that*

$$\begin{aligned}
&\sum_{i=1}^{t-1}\mathbb{E}\left[\sum_{h=0}^{H-1}(f_t(c_i,s_h^i,a_h^i)-f_\star(c_i,s_h^i,a_h^i))^2|\mathbb{H}_{i-1}\right]\\
&=\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\mathop{\mathbb{E}}_{c_i,s_h^i,a_h^i}\left[(f_t(c_i,s_h^i,a_h^i)-f_\star(c_i,s_h^i,a_h^i))^2|\mathbb{H}_{i-1}\right]\\
&\leq 68H\log(8|\mathcal{F}|t^3/\delta)+2\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}(f_t(c_i,s_h^i,a_h^i)-r_h^i)^2-(f_\star(c_i,s_h^i,a_h^i)-r_h^i)^2.
\end{aligned}$$

The above holds simultaneously for any $t\geq 2$ and a fixed sequence of functions $f_2,f_3,\ldots\in\mathcal{F}$.

Proof For a fixed $\delta\in(0,1)$, take $\delta_t=\delta/8t^3$ and apply union bound to Lemma 15 with all $t\geq 2$. The proof is identical to Lemma 16 ■

Step 2: Constructing confidence bound over policies with respect to the rewards approximation

Lemma 37 (confidence of policies w.r.t rewards approximation) *Consider Algorithm 5 that at each initialization round $t\leq|A|$, plays the policy that always choose action a_t, and at each round $t\geq|A|+1$ selects π_t based on the history $\mathbb{H}_{t-1}$.*

Then, for any $\delta\in(0,1)$ with probability at least $1-\delta/4$ for all $t\geq|A|+1$ and every policy $\pi\in\Pi_{\mathcal{C}}$, it holds that

$$\begin{aligned}
&|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(f_\star,P_\star)(c)}(s_0)]-\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(\hat{f}_t,P_\star)(c)}(s_0)]|\\
&\leq\sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\frac{q_h(s_h|\pi(c;\cdot),P_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star)}\right]}\cdot\sqrt{68H\log(8|\mathcal{F}|t^3/\delta)}.
\end{aligned}$$

Proof The Lemma is obtained using the same derivation presented in Lemma 18, where for every context $c\in\mathcal{C}$ we have $P_\star^c=P_\star$, using Lemma 36 (instead of Lemma 16). ■

C.3.4 ANALYSING THE ERROR DUE TO THE REWARDS APPROXIMATION

Step 3: Relax the confidence bound to be additive

Lemma 38 (the "square trick" relaxation for unknown and context-free dynamics) *Under the good event G_1 (which stated in Lemma 37), we have for any $t > |A|$ and policy $\pi \in \Pi_{\mathcal{C}}$*

$$\begin{aligned}&\left|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{f}_t,P_\star)}(c)}(s_0)]\right| \\ \leq &\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\frac{\beta_t\cdot q_h(s_h|\pi(c;\cdot),P_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star)}\right] + \beta_t\frac{H|S||A|}{t},\end{aligned}$$

where $\beta_t = \sqrt{\frac{17t\log(8|\mathcal{F}|t^3/\delta)}{|S||A|}}$.

The above implies that

$$\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}(s_0)] \leq \beta_t\cdot\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\frac{q_h(s_h,\pi(c;s_h)|\pi(c;\cdot),P_\star)}{p_{min}\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]}\right] + \beta_t\frac{H|S||A|}{t}, \quad (19)$$

and

$$\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] \leq 2\beta_t\cdot\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\frac{q_h(s_h,\pi(c;s_h)|\pi(c;\cdot),P_\star)}{p_{min}\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]}\right] + \beta_t\frac{H|S||A|}{t}, \quad (20)$$

where $\widehat{r}_t$ is the context-dependent rewards function defined in Algorithm 5.

Proof Using the same derivation presented in Lemma 19, where $\beta_t = \sqrt{\frac{17t\log(8|\mathcal{F}|t^3/\delta)}{|S||A|}}$ we obtain the first part of the lemma. For the second part, by minimum reachability it holds that

$$\begin{aligned}&\left|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{f}_t,P_\star)}(c)}(s_0)]\right| \\ &\leq \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\frac{\beta_t\cdot q_h(s_h|\pi(c;\cdot),P_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star)}\right] + \beta_t\frac{H|S||A|}{t} \\ &\leq \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\frac{1}{p_{min}}\frac{\beta_t\cdot q_h(s_h|\pi(c;\cdot),P_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]}\right] + \beta_t\frac{H|S||A|}{t}.\end{aligned} \quad (21)$$

For every context $c\in\mathcal{C}$ (we will later take the expectation over c), by the value function, definition for every $t\geq|A|+1$ and $\pi\in\Pi_{\mathcal{C}}$ it holds that

$$\begin{aligned}&V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}(s_0) \\ =&\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A}q_h(s_h,a_h|\pi(c;\cdot),P_\star)\cdot\widehat{r}^c_t(s_h,a_h) \\ =&\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A}q_h(s_h,a_h|\pi(c;\cdot),P_\star)\cdot\left(\widehat{f}_t(c,s_h,a_h)+\min\left\{\frac{1}{p_{min}}\cdot\frac{\beta_k}{\sum_{i=1}^{k-1}\mathbb{I}[a=\pi_i(c_t,s)]},1\right\}\right) \\ =&\sum_{h=0}^{H-1}\sum_{s_h\in S_h}q_h(s_h,\pi(c;s_h)|\pi(c;\cdot),P_\star)\cdot\left(\widehat{f}_t(c,s_h,\pi(c;s_h))+\min\left\{\frac{1}{p_{min}}\cdot\frac{\beta_k}{\sum_{i=1}^{k-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c_t,s)]},1\right\}\right) \\ =&V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{f}_t,P_\star)}(c)}(s_0)+\sum_{h=0}^{H-1}\sum_{s_h\in S_h}q_h(s_h,\pi(c;s_h)|\pi(c;\cdot),P_\star)\cdot\min\left\{\frac{1}{p_{min}}\cdot\frac{\beta_k}{\sum_{i=1}^{k-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c_t,s)]},1\right\}.\end{aligned} \quad (22)$$

The third and forth equalities above are since $\pi(c;\cdot)$ is deterministic, thus

$$\sum_{a_h \in A} q_h(s_h, a_h|\pi(c;\cdot), P_\star) = \sum_{a_h \in A} q_h(s_h|\pi(c;\cdot), P_\star) \cdot \pi(a_h|c; s_h) = q_h(s_h, \pi(c; s_h)|\pi(c;\cdot), P_\star) = q_h(s_h|\pi(c;\cdot), P_\star) \cdot 1.$$

We remark that $\frac{1}{p_{min}} \frac{\beta_t}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c;s)=\pi_i(c;s)}} > 0$ for every c, s and π by our initialization.

Thus, equation (22) further implies that

$$V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t, P_\star)}(c)}(s_0) \geq V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{f}_t, P_\star)}(c)}(s_0), \tag{23}$$

and, using that $\sum_i a_i \cdot \min\{b_i, c_i\} \leq \sum_i a_i \cdot b_i$ where $a_i, b_i, c_i \geq 0$ for all i, we also have

$$V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t, P_\star)}(c)}(s_0) \leq V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{f}_t, P_\star)}(c)}(s_0) + \sum_{h=0}^{H-1} \sum_{s_h \in S_h} q_h(s_h|\pi(c;\cdot), P_\star) \cdot \frac{1}{p_{min}} \cdot \frac{\beta_t}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c; s_h) = \pi_i(c; s_h)]}. \tag{24}$$

Clearly, inequalities (23) and (24) also hold when taking the expectation over the context c on both sides.

To obtain inequality (19) we combine inequalities (23) and (21):

$$\begin{aligned}
\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t, P_\star)}(c)}(s_0)] &\leq \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{f}_t, P_\star)}(c)}(s_0)] && \text{(By inequality (23))} \\
&\leq \mathbb{E}_c\left[\sum_{h=0}^{H-1} \sum_{s_h \in S_h} \frac{1}{p_{min}} \frac{\beta_t \cdot q_h(s_h|\pi(c;\cdot), P_\star)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c; s_h) = \pi_i(c; s_h)]}\right] + \beta_t \frac{H|S||A|}{t}. \\
& && \text{(By inequality (21))}
\end{aligned}$$

To obtain inequality (20) we combine inequalities (24) and (21):

$$\begin{aligned}
&\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t, P_\star)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] \\
\leq &\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{f}_t, P_\star)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] + \mathbb{E}_c\left[\sum_{h=0}^{H-1} \sum_{s_h \in S_h} \frac{1}{p_{min}} \frac{\beta_t \cdot q_h(s_h|\pi(c;\cdot), P_\star)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c; s_h) = \pi_i(c; s_h)]}\right] \\
& && \text{(By inequality (24))} \\
\leq &\left|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{f}_t, P_\star)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)]\right| + \mathbb{E}_c\left[\sum_{h=0}^{H-1} \sum_{s_h \in S_h} \frac{1}{p_{min}} \frac{\beta_t \cdot q_h(s_h|\pi(c;\cdot), P_\star)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c; s_h) = \pi_i(c; s_h)]}\right] \\
\leq &\mathbb{E}_c\left[\sum_{h=0}^{H-1} \sum_{s_h \in S_h} \frac{1}{p_{min}} \frac{\beta_t \cdot q_h(s_h|\pi(c;\cdot), P_\star)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c; s_h) = \pi_i(c; s_h)]}\right] + \beta_t \frac{H|S||A|}{t} && \text{(By inequality (21))} \\
&+ \mathbb{E}_c\left[\sum_{h=0}^{H-1} \sum_{s_h \in S_h} \frac{1}{p_{min}} \frac{\beta_t \cdot q_h(s_h|\pi(c;\cdot), P_\star)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c; s_h) = \pi_i(c; s_h)]}\right] \\
= &2\mathbb{E}_c\left[\sum_{h=0}^{H-1} \sum_{s_h \in S_h} \frac{1}{p_{min}} \frac{\beta_t \cdot q_h(s_h, |\pi(c;\cdot), P_\star)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c; s_h) = \pi_i(c; s_h)]}\right] + \beta_t \frac{H|S||A|}{t},
\end{aligned}$$

which completes the proof of the lemma. ■

Step 4: Bounding the sum of contextual potential functions. We consider the following two contextual potential functions in round t, for $T \geq t > |A|$ and a context-depended policy $\pi \in \Pi_{\mathcal{C}}$.

We define $\psi_t(\pi)$,

Definition 39 *We denote by $\psi_t(\pi)$ the contextual potential of a context-dependent policy $\pi \in \Pi_{\mathcal{C}}$ at round $|A| < t \leq T$ using p_{min} as follows.*

$$\psi_t(\pi) := \mathbb{E}_c \left[\sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \frac{q_h(s_h|\pi(c;\cdot), P_\star^c)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c;s_h) = \pi_i(c;s_h)] p_{min}} \right],$$

where $\{\pi_t \in \Pi_{\mathcal{C}}\}_{t=1}^T$ is the sequence of policies selected by Algorithm 5.

In addition, recall the definition of $\phi_t(\pi)$ from Section B.

Definition 40 *We denote by $\phi_t(\pi)$ the contextual potential of a context-dependent policy $\pi \in \Pi_{\mathcal{C}}$ at round $|A| < t \leq T$ which is defined as follows.*

$$\phi_t(\pi) := \mathbb{E}_c \left[\sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \frac{q_h(s_h|\pi(c;\cdot), P_\star^c)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c;s_h) = \pi_i(c;s_h)] q_h(s_h|\pi_i(c;\cdot), P_\star^c)} \right],$$

where $\{\pi_t \in \Pi_{\mathcal{C}}\}_{t=1}^T$ is the sequence of policies selected by Algorithm 5.

In addition, recall that in our setting, $P_\star^c = P_\star$ for every context $c \in \mathcal{C}$.

In the following lemma, we bound the sum of contextual potential functions, over the rounds $t = |A| + 1, \ldots, T$.

Lemma 41 (contextual potential lemma for unknown context-independent dynamics) *Let $\{\pi_t \in \Pi_{\mathcal{C}}\}_{t=1}^T$ be the sequence of policies selected by Algorithm 5. Then, for all $T > |A|$ the followings hold.*

1. *For any policy $\pi \in \Pi_{\mathcal{C}}$ it holds that*

$$\begin{aligned}\sum_{t=|A|+1}^{T} \phi_t(\pi) = \sum_{t=|A|+1}^{T} \mathbb{E}_c \left[\sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \frac{q_h(s_h|\pi(c;\cdot), P_\star)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c;s_h) = \pi_i(c;s_h)] q_h(s_h|\pi_i(c;\cdot), P_\star^c)} \right] \\ \leq \frac{H|A|}{p_{min}} (1 + \log(T/|A|)).\end{aligned}$$

2. *For the sequence $\{\pi_t \in \Pi_{\mathcal{C}}\}_{t=1}^T$ it holds that*

$$\begin{aligned}\sum_{t=|A|+1}^{T} \phi_t(\pi_t) = \sum_{t=|A|+1}^{T} \mathbb{E}_c \left[\sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \frac{q_h(s_h|\pi_t(c;\cdot), P_\star)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi_t(c;s_h) = \pi_i(c;s_h)] q_h(s_h|\pi_i(c;\cdot), P_\star^c)} \right] \\ \leq \frac{|S||A|}{p_{min}} (1 + \log(T/|A|)).\end{aligned}$$

3. *For any policy $\pi \in \Pi_{\mathcal{C}}$ it holds that*

$$\begin{aligned}\sum_{t=|A|+1}^{T} \psi_t(\pi) = \sum_{t=|A|+1}^{T} \mathbb{E}_c \left[\sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \frac{q_h(s_h|\pi(c;\cdot), P_\star)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c;s_h) = \pi_i(c;s_h)] p_{min}} \right] \\ \leq \frac{H|A|}{p_{min}} (1 + \log(T/|A|)).\end{aligned}$$

4. *For the sequence $\{\pi_t \in \Pi_{\mathcal{C}}\}_{t=1}^T$ it holds that*

$$\begin{aligned}\sum_{t=|A|+1}^{T} \psi_t(\pi_t) = \sum_{t=|A|+1}^{T} \mathbb{E}_c \left[\sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \frac{q_h(s_h|\pi_t(c;\cdot), P_\star)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi_t(c;s_h) = \pi_i(c;s_h)] p_{min}} \right] \\ \leq \frac{|S||A|}{p_{min}} (1 + \log(T/|A|))\end{aligned}$$

Proof The proofs of parts 1 and 2 are identical to the proof of Lemma 21 where for all $c \in \mathcal{C}$ we have $P^c\star = P_\star$ and for all $h \in [H-1]$ we have $S_h^c = S_h$.

The proof of 3 and 4 are also similar. For 3, consider the following derivation.

For a fixed context $c \in \mathcal{C}$, given any policy $\pi \in \Pi_{\mathcal{C}}$ it holds that,

$$\begin{aligned}
&\sum_{t=|A|+1}^{T}\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\frac{1}{p_{min}}\frac{q_h(s_h|\pi(c;\cdot),P_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]}\\
=&\frac{1}{p_{min}}\sum_{h=0}^{H-1}\sum_{s_h\in S_h}q_h(s_h|\pi(c;\cdot),P_\star)\sum_{t=|A|+1}^{T}\frac{1}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]}\\
\leq&\frac{1}{p_{min}}\sum_{h=0}^{H-1}\sum_{s_h\in S_h}q_h(s_h|\pi(c;\cdot),P_\star)\sum_{a_h\in A}\sum_{i=1}^{\sum_{t=1}^{T}\mathbb{I}[\pi_t(c;s_h)=a_h]}\frac{1}{i}\\
\leq&\frac{1}{p_{min}}\sum_{h=0}^{H-1}\sum_{s_h\in S_h}q_h(s_h|\pi(c;\cdot),P_\star)\sum_{a_h\in A}\left(1+\log\left(\sum_{t=1}^{T}\mathbb{I}[\pi_t(c;s_h)=a_h]\right)\right)\\
\leq&\frac{H|A|}{p_{min}}(1+\log(T/|A|)),
\end{aligned}$$

where the last inequality is due to Jensen's inequality. By taking expectation over c on both sides of the above inequality, we obtain part 3 of the lemma.

For part 4 we similarly have

$$\begin{aligned}
&\sum_{t=|A|+1}^{T}\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\frac{1}{p_{min}}\frac{q_h(s_h|\pi_t(c;\cdot),P_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]}\\
\leq&\frac{1}{p_{min}}\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{t=|A|+1}^{T}\frac{1}{\sum_{i=1}^{t-1}\mathbb{I}[\pi_t(c;s_h)=\pi_i(c;s_h)]}\\
\leq&\frac{1}{p_{min}}\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A}\sum_{i=1}^{\sum_{t=1}^{T}\mathbb{I}[\pi_t(c;s_h)=a_h]}\frac{1}{i}\\
\leq&\frac{1}{p_{min}}\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A}\left(1+\log\left(\sum_{t=1}^{T}\mathbb{I}[\pi_t(c;s_h)=a_h]\right)\right)\\
\leq&\frac{|S||A|}{p_{min}}(1+\log(T/|A|)),
\end{aligned}$$

where the last inequality is due to Jensen's inequality. By taking expectation over c on both sides of the above inequality, we obtain part 4 of the lemma. ■

Lemma 42 (cumulative value error due to rewards approximation) *Under the good event G_1, for all $T > |A|$ it holds that*

$$\begin{aligned}
&\sum_{t=|A|+1}^{T}\mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,P_\star)}(c)]-\mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}]\\
\leq&H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T}+2(1+\log(T/|A|))\frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|S||A|}}{p_{min}}.
\end{aligned}$$

Proof Assume the good event G_1 holds and consider the following derivation.

$$
\begin{aligned}
&\sum_{t=|A|+1}^{T} \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}] \\
\leq &\sum_{t=|A|+1}^{T} \frac{\beta_t H|S||A|}{t} + 2\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\frac{1}{p_{min}}\frac{\beta_t\cdot q_h(s_h,\pi_t(c;s_h)|\pi_t(c;\cdot),P_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi_t(c;s_h)=\pi_i(c;s_h)]}\right] && \text{(By Lemma 38, equation (20))} \\
= &H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|}\sum_{t=|A|+1}^{T}\frac{1}{\sqrt{t}} && \text{(By } \beta_t \text{ choice)} \\
&+ 2\cdot\beta_T\sum_{t=|A|+1}^{T}\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\frac{1}{p_{min}}\frac{q_h(s_h,\pi_t(c;s_h)|P,\pi_t(c;\cdot))}{\sum_{i=1}^{t-1}\mathbb{I}[\pi_t(c;s_h)=\pi_i(c;s_h)]}\right] && \text{(Since } \beta_T\geq\beta_t \text{ for all } t\leq T.) \\
\leq &H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} && \text{(Since } \textstyle\sum_{t=1}^{T}1/\sqrt{t}\leq\sqrt{T}.) \\
&+ 2\beta_T\frac{|S||A|}{p_{min}}(1+\log(T/|A|)) && \text{(By Lemma 41, part 4.)} \\
= &H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} + 2\sqrt{\frac{17\log(8|\mathcal{F}|T^3/\delta)T}{|S||A|}}\frac{|S||A|}{p_{min}}(1+\log(T/|A|)) \\
\leq &H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} + 2(1+\log(T/|A|))\frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|S||A|}}{p_{min}},
\end{aligned}
$$

as stated. ■

C.3.5 REGRET BOUND

Recall that for every $t > |A|$ and $\pi \in \Pi_{\mathcal{C}}$ it holds that

$$
\begin{aligned}
&V^{\pi(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0) = \sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A} q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}^c_t)\cdot\widehat{r}^c_t(s_h,a_h) \\
&= \sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A} q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}^c_t)\cdot\left(\widehat{f}_t(c,s_h,a_h)+\min\left\{\frac{1}{p_{min}}\cdot\frac{\beta_k}{\sum_{i=1}^{k-1}\mathbb{I}[a_h=\pi_i(c_t,s)]},1\right\}\right) \\
&= \sum_{h=0}^{H-1}\sum_{s_h\in S_h} q_h(s_h,\pi(c;s_h)|\pi(c;\cdot),\widehat{P}^c_t)\cdot\left(\widehat{f}_t(c,s_h,\pi(c;s_h))+\min\left\{\frac{1}{p_{min}}\cdot\frac{\beta_k}{\sum_{i=1}^{k-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c_t,s)]},1\right\}\right) \\
&= V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{f}_t,\widehat{P}_t)}(c)}(s_0) + \sum_{h=0}^{H-1}\sum_{s_h\in S_h} q_h(s_h,\pi(c;s_h)|\pi(c;\cdot),\widehat{P}^c_t)\cdot\min\left\{\frac{1}{p_{min}}\cdot\frac{\beta_k}{\sum_{i=1}^{k-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c_t,s)]},1\right\}.
\end{aligned}
\tag{25}
$$

Lemma 43 (optimism) *Let $\pi^\star \in \Pi_{\mathcal{C}}$ be an optimal policy of the true CMDP. Under the good events G_1 and G_2, for all $t > |A|$ it holds that*

$$
\mathbb{E}_c\left[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0)\right] - \mathbb{E}_c\left[V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)\right] \leq \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\frac{1}{p_{min}}\frac{\beta_t\cdot q_h(s_h,\pi^\star(c;s_h)|\pi^\star(c;\cdot),P_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]}\right] + \beta_t\frac{H|S||A|}{t}.
$$

Proof Assume the good events hold and consider the following derivation.

$$
\mathbb{E}_c\left[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0)\right]
$$

$$\leq \mathbb{E}_c\left[V^{\pi^\star(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t,P_\star)}(c)}(s_0)\right] + \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\frac{1}{p_{min}}\frac{\beta_t\cdot q_h(s_h,\pi^\star(c;s_h)|\pi^\star(c;\cdot),P_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]}\right] + \beta_t\frac{H|S||A|}{t}$$

(Lemma 38, equation 19.)

$$\leq \mathbb{E}_c\left[V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(c)(s_0)\right] + \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\frac{1}{p_{min}}\frac{\beta_t\cdot q_h(s_h,\pi^\star(c;s_h)|\pi^\star(c;\cdot),P_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]}\right] + \beta_t\frac{H|S||A|}{t} \quad \text{(By Lemma 32)}$$

■

Theorem 44 (expected regret bound) *For every $T \geq 1$, finite functions class $\mathcal{F}$, and $\delta \in (0,1)$ with probability at least $1-\delta$ it holds that*

$$\begin{aligned}&\textit{E.Regret}_T(RM-UCID)\\ &\leq \tilde{O}\left(H|S|\sqrt{T|A|}\log\frac{1}{\delta} + \max\{H, 1/p_{min}\}\cdot\sqrt{|S||A|T\log\frac{|\mathcal{F}|}{\delta}} + |A|H\right),\end{aligned}$$

for $\beta_t = \sqrt{\frac{17t\log(8|\mathcal{F}|t^3/\delta)}{|S||A|}}$.

Proof We bound the expected regret under the good events G_1 and G_2. Since $G_1 \cap G_2$ holds with probability at least $1-\delta/2$, the theorem will follow by union bound.

By Azuma's inequality, with probability at least $1-\delta/4$ over the policies π_t the following holds.

$$\text{E.Regret}_T(RM-UCID) = \sum_{t=1}^{T}\mathbb{E}_{\mathbb{H}_{t-1}}\left[\mathbb{E}_c[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)]\right]$$

$$\leq \sum_{t=1}^{T}\left(\mathbb{E}_c[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)]\right) + 2H\sqrt{2T\log(8/\delta)} \quad \text{(By Azuma's inequality)}$$

$$\leq \sum_{t=|A|+1}^{T}\left(\mathbb{E}_c[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)]\right) + 2H\sqrt{2T\log(8/\delta)} + |A|H$$

$$\leq \sum_{t=|A|+1}^{T}\left(\mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)]\right) + \sum_{t=|A|+1}^{T}\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\frac{1}{p_{min}}\frac{\beta_t\cdot q_h(s_h,\pi^\star(c;s_h)|\pi^\star(c;\cdot),P_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]}\right]$$
$$+ \sum_{t=|A|+1}^{T}\beta_t\frac{H|S||A|}{t} + 2H\sqrt{2T\log(8/\delta)} + |A|\cdot H \quad \text{(By the optimism lemma, Lemma 43)}$$

$$= \sum_{t=|A|+1}^{T}\left(\mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t,\hat{P}_t)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t,P_\star)}(c)}(s_0)] + \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t,P_\star)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)]\right)$$

(By adding and subtracting $\mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t,P_\star)}(c)}(s_0)]$)

$$+ \sum_{t=|A|+1}^{T}\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\frac{1}{p_{min}}\frac{\beta_t\cdot q_h(s_h,\pi^\star(c;s_h)|\pi^\star(c;\cdot),P_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]}\right]$$
$$+ \sum_{t=|A|+1}^{T}\beta_t\frac{H|S||A|}{t} + 2H\sqrt{2T\log(8/\delta)} + |A|\cdot H$$

$$\leq \sum_{t=|A|+1}^{T}\left(\mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t,\hat{P}_t)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t,P_\star)}(c)}(s_0)] + \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t,P_\star)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)]\right)$$

$$
\begin{aligned}
&+ \beta_T \cdot \mathbb{E}_c \left[\sum_{t=|A|+1}^{T} \sum_{h=0}^{H-1} \sum_{s_h \in S_h} \frac{1}{p_{min}} \frac{q_h(s_h, \pi^\star(c;s_h)|\pi^\star(c;\cdot), P_\star)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi^\star(c;s_h) = \pi_i(c;s_h)]} \right] && (\beta_T \geq \beta_t, \ \ \forall t \in \{1,2,\ldots,T\}.) \\
&+ \sum_{t=|A|+1}^{T} \beta_t \frac{H|S||A|}{t} + 2H\sqrt{2T\log(8/\delta)} + |A| \cdot H \\
=& \underbrace{\sum_{t=|A|+1}^{T} \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(\widehat{r}_t, \widehat{P}_t)(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(\widehat{r}_t, P_\star)(c)}(s_0)]}_{(\star)} \\
&+ \underbrace{\sum_{t=|A|+1}^{T} \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(\widehat{r}_t, P_\star)(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)]}_{(\star\star)} \\
&+ \beta_T \cdot \sum_{t=|A|+1}^{T} \mathbb{E}_c \left[\sum_{h=0}^{H-1} \sum_{s_h \in S_h} \frac{1}{p_{min}} \frac{q_h(s_h, \pi^\star(c;s_h)|\pi^\star(c;\cdot), P_\star)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi^\star(c;s_h) = \pi_i(c;s_h)]} \right] \\
&+ \sum_{t=|A|+1}^{T} \beta_t \frac{H|S||A|}{t} + 2H\sqrt{2T\log(8/\delta)} + |A| \cdot H \\
\leq& (\star) + (\star\star) \\
&+ H(1+\log(T/|A|)) \frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|A|}}{p_{min} \cdot \sqrt{|S|}} && \text{(By Lemma 41 part 3 and } \beta_T \text{ choice)} \\
&+ H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} && \text{(By } \beta_t \text{ choice and } \textstyle\sum_{t=1}^{T} 1/\sqrt{t} \leq \sqrt{T}.) \\
&+ 2H\sqrt{2T\log(8/\delta)} + |A| \cdot H.
\end{aligned}
$$

By Lemma 35 under the good event G_2 we have with probability at least $1 - \delta/4$ that

$$
(\star) \leq 4H\sqrt{|S| + 2\log(4|S||A|T^2/\delta)} \left(\sqrt{2TH\log(8/\delta)} + 2\sqrt{|S||A|T} \right).
$$

By Lemma 42, under the good event G_1 it holds that

$$
\begin{aligned}
(\star\star) =& \sum_{t=|A|+1}^{T} \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(\widehat{r}_t, P_\star)(c)}] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}] \\
\leq& H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} + 2(1+\log(T/|A|)) \frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|S||A|}}{p_{min}}.
\end{aligned}
$$

By plugging the upper bounds on $(\star)$ and $(\star\star)$ we obtain

$$
\begin{aligned}
&\text{E.Regret}_T(RM - UCID) \leq (\star) + (\star\star) \\
&+ H(1+\log(T/|A|)) \frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|A|}}{p_{min} \cdot \sqrt{|S|}} + H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} \\
&+ 2H\sqrt{2T\log(8/\delta)} + |A| \cdot H \\
=& 4H\sqrt{|S| + 2\log(4|S||A|T^2/\delta)} \left(\sqrt{2TH\log(8/\delta)} + 2\sqrt{|S||A|T} \right) \\
&+ H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} + 2(1+\log(T/|A|)) \frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|S||A|}}{p_{min}}
\end{aligned}
$$

$$
\begin{aligned}
&+ H(1+\log(T/|A|))\frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|A|}}{p_{min}\cdot\sqrt{|S|}} + H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} + 2H\sqrt{2T\log(8/\delta)} + |A|\cdot H \\
=&4H\left(\sqrt{2TH\log(8/\delta)}\sqrt{|S|+2\log(4|S||A|T^2/\delta)} + 2\sqrt{|S||A|T}\sqrt{|S|+2\log(4|S||A|T^2/\delta)}\right) \\
&+ 2H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} + 3(1+\log(T/|A|))\frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|S||A|}}{p_{min}} \\
&+ 2H\sqrt{2T\log(8/\delta)} + |A|H \\
\le&4H\left(\sqrt{2T|S|H\log(8/\delta) + 4TH\log(8/\delta)\log(4|S||A|T^2/\delta)} + 4|S|\sqrt{|A|T\log(4|S||A|T^2/\delta)}\right) \\
&+ 2H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} + 3(1+\log(T/|A|))\frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|S||A|}}{p_{min}} \\
&+ 2H\sqrt{2T\log(8/\delta)} + |A|H \\
\le&4H\left(\sqrt{16T|S|H\log^2(4|S||A|T^2/\delta)} + 4|S|\sqrt{|A|T\log(4|S||A|T^2/\delta)}\right) \\
&+ 2H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} + 3(1+\log(T/|A|))\frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|S||A|}}{p_{min}} \\
&+ 2H\sqrt{2T\log(8/\delta)} + |A|H \\
\le&4H\left(\sqrt{16T|S|^2|A|\log^2(4|S||A|T^2/\delta)} + 4|S|\sqrt{|A|T\log(4|S||A|T^2/\delta)}\right) && (H\le|S|) \\
&+ 2H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} + 3(1+\log(T/|A|))\frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|S||A|}}{p_{min}} \\
&+ 2H\sqrt{2T\log(8/\delta)} + |A|H \\
\le&4H\left(8|S|\sqrt{|A|T}\log(4|S||A|T^2/\delta)\right) \\
&+ 2H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} + 3(1+\log(T/|A|))\frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|S||A|}}{p_{min}} \\
&+ 2H\sqrt{2T\log(8/\delta)} + |A|H \\
\le&32H|S|\sqrt{|A|T}\log(4|S||A|T^2/\delta) \\
&+ 2H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} + 3(1+\log(T/|A|))\frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|S||A|}}{p_{min}} \\
&+ 2H\sqrt{2T\log(8/\delta)} + |A|H \\
=&\tilde{O}\left(H|S|\sqrt{T|A|}\log\frac{1}{\delta} + \max\{H, 1/p_{min}\}\cdot\sqrt{|S||A|T\log\frac{|\mathcal{F}|}{\delta}} + |A|H\right).
\end{aligned}
$$

Overall, by union bounds the above holds with probability at least $1-\delta$. ■

Recall the regret, which defined as $\text{Regret}_T(\text{ALG}) := \sum_{t=1}^{T} V_{\mathcal{M}(c_t)}^{\pi^\star(c_t;\cdot)} - V_{\mathcal{M}(c_t)}^{\pi_t(c_t;\cdot)}$.

Theorem 45 (regret bound) *For every $T \ge 1$, finite functions class $\mathcal{F}$ and $\delta \in (0,1)$ with probability at least $1-\delta$ it holds that*

$$
\textit{Regret}_T(RM-UCID) \le \tilde{O}\left(H|S|\sqrt{T|A|}\log\frac{1}{\delta} + \max\{H, 1/p_{min}\}\cdot\sqrt{|S||A|T\log\frac{|\mathcal{F}|}{\delta}} + |A|H\right).
$$

for the choice in $\beta_t = \sqrt{\frac{17t\log(8|\mathcal{F}|t^3/\delta)}{|S||A|}}$ for all $t \in [T]$.

Proof We bound the regret under the good events G_1 and G_2. Since $G_1 \cap G_2$ holds with probability at least $1-\delta/2$, the theorem will follow by union bound.

Consider the martingale difference sequence $\{Y_t\}_{t=1}^T$ where the filtration is $\{\mathbb{H}_t\}_{t=1}^T$ for

$$Y_t := V_{\mathcal{M}(c_t)}^{\pi^\star(c_t,\cdot)}(s_0) - V_{\mathcal{M}(c_t)}^{\pi_t(c_t,\cdot)}(s_0) - \mathbb{E}_{c_t}\left[V_{\mathcal{M}(c_t)}^{\pi^\star(c_t,\cdot)}(s_0) - V_{\mathcal{M}(c_t)}^{\pi_t(c_t,\cdot)}(s_0)\middle|\mathbb{H}_{t-1}\right].$$

Clearly, for all t, $|Y_t| \le 2H$, Y_t is determined completely by the history $\mathbb{H}_t$ and $\mathbb{E}_{c_t}[Y_t|\mathbb{H}_{t-1}] = 0$. Thus, by Azuma's inequality, with probability at least $1-\delta/4$ it holds that

$$\begin{aligned}
&\text{Regret}_T(RM-UCID) = \sum_{t=1}^T V_{\mathcal{M}(c_t)}^{\pi^\star(c_t;\cdot)}(s_0) - V_{\mathcal{M}(c_t)}^{\pi_t(c_t;\cdot)}(s_0)\\
&\le \sum_{t=1}^T \left(\mathbb{E}_{c_t}[V_{\mathcal{M}(c_t)}^{\pi^\star(c_t;\cdot)}(s_0)|\mathbb{H}_{t-1}] - \mathbb{E}_{c_t}[V_{\mathcal{M}(c_t)}^{\pi_t(c_t;\cdot)}(s_0)|\mathbb{H}_{t-1}]\right) + 2H\sqrt{2T\log(8/\delta)} && \text{(By Azuma's inequality)}\\
&= \sum_{t=1}^T \left(\mathbb{E}_c[V_{\mathcal{M}(c)}^{\pi^\star(c;\cdot)}(s_0)] - \mathbb{E}_c[V_{\mathcal{M}(c)}^{\pi_t(c;\cdot)}(s_0)]\right) + 2H\sqrt{2T\log(8/\delta)}\\
&\quad \text{(Since } \pi_t \text{ is determined by } \mathbb{H}_{t-1}\text{, and } c_t,\pi^\star \text{ is independent of the history, we can omit the conditioning in } \mathbb{H}_{t-1}\text{)}\\
&\le \sum_{t=|A|+1}^T \left(\mathbb{E}_c[V_{\mathcal{M}(c)}^{\pi^\star(c;\cdot)}(s_0)] - \mathbb{E}_c[V_{\mathcal{M}(c)}^{\pi_t(c;\cdot)}(s_0)]\right) + 2H\sqrt{2T\log(8/\delta)} + |A|H\\
&\le \sum_{t=|A|+1}^T \left(\mathbb{E}_c[V_{\widehat{\mathcal{M}}_t(c)}^{\pi_t(c;\cdot)}(s_0)] - \mathbb{E}_c[V_{\mathcal{M}(c)}^{\pi_t(c;\cdot)}(s_0)]\right) + \sum_{t=|A|+1}^T \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\frac{1}{p_{min}}\frac{\beta_t\cdot q_h(s_h,\pi^\star(c;s_h)|\pi^\star(c;\cdot),P_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]}\right]\\
&\quad + \sum_{t=|A|+1}^T \beta_t\frac{H|S||A|}{t} + 2H\sqrt{2T\log(8/\delta)} + |A|H && \text{(By the optimism lemma, Lemma 43)}\\
&= \sum_{t=|A|+1}^T \left(\mathbb{E}_c[V_{\mathcal{M}^{(\widehat{r}_t,\widehat{P}_t)}(c)}^{\pi_t(c;\cdot)}(s_0)] - \mathbb{E}_c[V_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}^{\pi_t(c;\cdot)}(s_0)] + \mathbb{E}_c[V_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}^{\pi_t(c;\cdot)}(s_0)] - \mathbb{E}_c[V_{\mathcal{M}(c)}^{\pi_t(c;\cdot)}(s_0)]\right)\\
&\qquad \text{(By adding and subtracting } \mathbb{E}_c[V_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}^{\pi_t(c;\cdot)}(s_0)]\text{)}\\
&\quad + \sum_{t=|A|+1}^T \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\frac{1}{p_{min}}\frac{\beta_t\cdot q_h(s_h,\pi^\star(c;s_h)|\pi^\star(c;\cdot),P_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]}\right]\\
&\quad + \sum_{t=|A|+1}^T \beta_t\frac{H|S||A|}{t} + 2H\sqrt{2T\log(8/\delta)} + |A|H\\
&\le \sum_{t=|A|+1}^T \left(\mathbb{E}_c[V_{\mathcal{M}^{(\widehat{r}_t,\widehat{P}_t)}(c)}^{\pi_t(c;\cdot)}(s_0)] - \mathbb{E}_c[V_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}^{\pi_t(c;\cdot)}(s_0)] + \mathbb{E}_c[V_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}^{\pi_t(c;\cdot)}(s_0)] - \mathbb{E}_c[V_{\mathcal{M}(c)}^{\pi_t(c;\cdot)}(s_0)]\right)\\
&\quad + \beta_T\cdot\mathbb{E}_c\left[\sum_{t=|A|+1}^T\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\frac{1}{p_{min}}\frac{q_h(s_h,\pi^\star(c;s_h)|\pi^\star(c;\cdot),P_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]}\right] && (\beta_T\ge\beta_t,\ \ \forall t\in\{1,2,\ldots,T\}.)\\
&\quad + \sum_{t=|A|+1}^T \beta_t\frac{H|S||A|}{t} + 2H\sqrt{2T\log(8/\delta)} + |A|H\\
&= \underbrace{\sum_{t=|A|+1}^T \mathbb{E}_c[V_{\mathcal{M}^{(\widehat{r}_t,\widehat{P}_t)}(c)}^{\pi_t(c;\cdot)}(s_0)] - \mathbb{E}_c[V_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}^{\pi_t(c;\cdot)}(s_0)]}_{(\star)}
\end{aligned}$$

$$
\begin{aligned}
&+ \underbrace{\sum_{t=|A|+1}^{T} \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,P_\star)(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)]}_{(\star\star)} \\
&+ \beta_T \cdot \sum_{t=|A|+1}^{T} \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h} \frac{1}{p_{min}} \frac{q_h(s_h,\pi^\star(c;s_h)|\pi^\star(c;\cdot),P_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]}\right] \\
&+ \sum_{t=|A|+1}^{T} \beta_t \frac{H|S||A|}{t} + 2H\sqrt{2T\log(8/\delta)} + |A|H \\
\leq &(\star) + (\star\star) \\
&+ H(1+\log(T/|A|))\frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|A|}}{p_{min}\cdot\sqrt{|S|}} && \text{(By Lemma 41 part 3 and } \beta_T \text{ choice)} \\
&+ H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} && \text{(By } \beta_t \text{ choice and } \textstyle\sum_{t=1}^{T} 1/\sqrt{t} \leq \sqrt{T}.) \\
&+ 2H\sqrt{2T\log(8/\delta)} + |A|H.
\end{aligned}
$$

By Lemma 35 under the good event G_2 with probability at least $1-\delta/4$ it holds that

$$
(\star) \leq 4H\sqrt{|S|+2\log(4|S||A|T^2/\delta)}\left(\sqrt{2TH\log(8/\delta)} + 2\sqrt{|S||A|T}\right).
$$

By Lemma 42, under the good event G_1 it holds that

$$
\begin{aligned}
(\star\star) = &\sum_{t=|A|+1}^{T} \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,P_\star)(c)}] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}] \\
\leq &H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} + 2(1+\log(T/|A|))\frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|S||A|}}{p_{min}}.
\end{aligned}
$$

By plugging the upper bounds on $(\star)$ and $(\star\star)$ we obtain

$$
\begin{aligned}
&\text{Regret}_T(RM-UCID) \leq (\star) + (\star\star) \\
&+ H(1+\log(T/|A|))\frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|A|}}{p_{min}\cdot\sqrt{|S|}} + H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} + 2H\sqrt{2T\log(8/\delta)} + |A|\cdot H \\
=&4H\sqrt{|S|+2\log(4|S||A|T^2/\delta)}\left(\sqrt{2TH\log(8/\delta)} + 2\sqrt{|S||A|T}\right) \\
&+ H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} + 2(1+\log(T/|A|))\frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|S||A|}}{p_{min}} \\
&+ H(1+\log(T/|A|))\frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|A|}}{p_{min}\cdot\sqrt{|S|}} + H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} + 2H\sqrt{2T\log(8/\delta)} + |A|\cdot H \\
=&4H\left(\sqrt{2TH\log(8/\delta)}\sqrt{|S|+2\log(4|S||A|T^2/\delta)} + 2\sqrt{|S||A|T}\sqrt{|S|+2\log(4|S||A|T^2/\delta)}\right) \\
&+ 2H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} + 3(1+\log(T/|A|))\frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|S||A|}}{p_{min}} \\
&+ 2H\sqrt{2T\log(8/\delta)} + |A|H \\
\leq &4H\left(\sqrt{2T|S|H\log(8/\delta) + 4TH\log(8/\delta)\log(4|S||A|T^2/\delta)} + 4|S|\sqrt{|A|T\log(4|S||A|T^2/\delta)}\right) \\
&+ 2H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} + 3(1+\log(T/|A|))\frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|S||A|}}{p_{min}}
\end{aligned}
$$

$$
\begin{aligned}
&+ 2H\sqrt{2T\log(8/\delta)} + |A|H \\
\leq& 4H\left(\sqrt{16T|S|H\log^2(4|S||A|T^2/\delta)} + 4|S|\sqrt{|A|T\log(4|S||A|T^2/\delta)}\right) \\
&+ 2H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} + 3(1+\log(T/|A|))\frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|S||A|}}{p_{min}} \\
&+ 2H\sqrt{2T\log(8/\delta)} + |A|H \\
\leq& 4H\left(\sqrt{16T|S|^2|A|\log^2(4|S||A|T^2/\delta)} + 4|S|\sqrt{|A|T\log(4|S||A|T^2/\delta)}\right) && (H \leq |S|) \\
&+ 2H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} + 3(1+\log(T/|A|))\frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|S||A|}}{p_{min}} \\
&+ 2H\sqrt{2T\log(8/\delta)} + |A|H \\
\leq& 4H\left(8|S|\sqrt{|A|T}\log(4|S||A|T^2/\delta)\right) \\
&+ 2H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} + 3(1+\log(T/|A|))\frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|S||A|}}{p_{min}} \\
&+ 2H\sqrt{2T\log(8/\delta)} + |A|H \\
\leq& 32H|S|\sqrt{|A|T}\log(4|S||A|T^2/\delta) \\
&+ 2H\sqrt{17\log(8|\mathcal{F}|T^3/\delta)|S||A|T} + 3(1+\log(T/|A|))\frac{\sqrt{17\log(8|\mathcal{F}|T^3/\delta)T|S||A|}}{p_{min}} \\
&+ 2H\sqrt{2T\log(8/\delta)} + |A|H \\
=& \tilde{O}\left(H|S|\sqrt{T|A|}\log\frac{1}{\delta} + \max\{H, 1/p_{min}\}\cdot\sqrt{|S||A|T\log\frac{|\mathcal{F}|}{\delta}} + |A|H\right).
\end{aligned}
$$

Overall, by union bounds the above holds with probability at least $1-\delta$. ■

Corollary 46 (regret bound in terms of $\mathcal{G}$) *For every $T \geq 1$, finite functions class $\mathcal{G}$ ($\mathcal{F} = \mathcal{G}^S$) and $\delta \in (0,1)$ the following holds with probability at least $1-\delta$ for the same choice of parameters $\{\beta_t\}_{t\in[T]}$.*

$$
\textit{Regret}_T(RM-UCID) \leq \tilde{O}\left(H|S|\sqrt{T|A|}\log\frac{1}{\delta} + \max\{H, 1/p_{min}\}\cdot|S|\sqrt{|A|T\log\frac{|\mathcal{G}|}{\delta}} + |A|H\right).
$$

Proof Plug $\log(|\mathcal{F}|) = |S|\log(|\mathcal{G}|)$ in the bound of Theorem 45. ■

C.3.6 COMPUTATIONAL EFFICIENCY

The following establish the computational efficiency of Algorithm RM-UCID.

Lemma 47 *Under the good event G_2, for all $t > |A|$ and any context $c \in \mathcal{C}$ there exist a feasible solution to LP (14) and Algorithm FOM 7 run in $poly(|S|,|A|,H)$ time, assuming an efficient least square regression oracle.*

Proof Assume the good event G_2 holds. Then, for all $t > |A|$ and a context $c \in \mathcal{C}$, the true dynamics $P_\star$ and the optimal policy $\pi^\star(c;\cdot)$ induces an extended occupancy measure $q_\star^c$ which is a feasible solution for LP (14).

For the running time, LP (14) is a linear program with $poly(|S|, |A|, H)$ constrains and variables, hence it can be solved in $poly(|S|, |A|, H)$ time (using interior point methods, for example). Algorithm FDP (Algorithm 6) also run in $poly(|S|, |A|, H)$ time. Hence, overall, Algorithm 7 run in $poly(|S|, |A|, H)$ time. ■

Corollary 48 *The overall running time of Algorithm RM-UCID (Algorithm 5) is in* $poly(|S|, |A|, H, T)$.

Appendix D. Unknown and Context-Dependent Dynamics

D.1 Notations and Assumptions

Unknown and context-dependent dynamics. Meaning, for every context $c \in \mathcal{C}$ we have different dynamics $P_\star^c$ and is unknown to the learner. Recall we assume layered CMDP. For any context c, we denote by $S_0^c, S_1^c, \ldots, S_H^c$ the disjoint layers, and for every context $c \in \mathcal{C}$ it holds that $S = \bigcup_{h \in [H]} S_h^c$.

Known minimum reachability parameter. We assume that there exists $p_{min} \in (0, 1]$ such that for every context $c \in \mathcal{C}$, layer $h \in [H]$, state $s_h \in S_h^c$ an context-dependent policy $\pi \in \Pi_{\mathcal{C}}$ it holds that

$$q_h(s_h | \pi(c; \cdot), P_\star^c) \geq p_{min}.$$

We remark that p_{min} is **known** to the learner in this section.

Dynamics approximation using least-square regression. Let $\mathcal{P} \subseteq S \times (S \times A \times \mathcal{C}) \to [0, 1]$ be finite context-dependent dynamics class, i.e., every function $\widetilde{P} \in \mathcal{P}$ satisfies

$$\sum_{s' \in S} \widetilde{P}(s' | s, a, c) = 1, \quad \forall c \in \mathcal{C},\ \forall (s, a) \in S \times A.$$

Given a context-dependent dynamics P and a context c we denote the dynamics P induced by c by P^c, i.e., $P^c(s'|s, a) = P(s'|s, a, c)$.

The random variable $\mathcal{B}$. For every context $c \in \mathcal{C}$, and state-action pair $(s, a) \in S \times A$ we define a random variable $\mathcal{B}(P_\star^c, s, a) \in S$ which returns the next state s' that observed after action a was played in state s and the dynamics is $P_\star^c$. Observe that $\mathcal{B}$ satisfies the following,

$$\mathcal{B}(P_\star^c, s, a) \sim P_\star^c(\cdot | s, a),$$

and by definition it holds that

$$\mathbb{E}\left[\mathbb{I}[s' = \mathcal{B}(P_\star^c, s, a)]\right] = P_\star^c(s'|s, a) \ \ \forall s' \in S.$$

Assumption 4 (dynamics realizability) *We assume that $\mathcal{P}$ is realizable, meaning there exist a function $P_\star \in \mathcal{P}$ which is the true dynamics.*

As in previous sections, we assume an access to a least square regression (LSR) oracle which solves

$$\widehat{P} \in \arg\min_{\widetilde{P} \in \mathcal{P}} \sum_{((c,s,a,s'),b) \in D} (\widetilde{P}^c(s'|s, a) - b)^2$$

given a data set $D \in ((\mathcal{C} \times S \times A \times S) \times \{0, 1\})^n$ which contains samples, each of a context c, state s, action a, next state s' and a bit $b \in \{0, 1\}$ which indicate whether s' was observed after action a was played in state s, for every state $s' \in S$.

D.2 Algorithm

In Algorithm RM-UCDD (Algorithm 8), we approximate both the rewards and the dynamics using LSR oracle. The first $|A|$ rounds are initialization rounds, where in round $i \in \{1, 2, \ldots, |A|\}$ the agent plays the policy π_i which always chooses action a_i, i.e., $\pi_i(c; s) = a_i$ for every context $c \in \mathcal{C}$ and state $s \in S$.

At round $t \geq |A| + 1$, the agent computes the approximated rewards function for the context c_t using the previously-selected policies $\{\pi_k(c_t; \cdot)\}_{k=1}^{t-1}$ in the same iterative computation we present in the previous sections.

To approximate the dynamics, the agent uses the least square minimizer $\widehat{P}_t$. She defines the approximate MDP for the context c_t, $\widehat{\mathcal{M}}_t(c_t) = (S, A, \widehat{P}_t^{c_t}, \widehat{r}_t^{c_t}, s_0, H)$, computes an optimal policy for it $\pi_t(c_t; \cdot)$ and run it to generate a trajectory and update the oracles.

We remark that we feed the LSR oracle for the dynamics with samples of the form $((c_t, s_h^t, a_h^t, s'), \mathbb{I}[s' = s_{h+1}^t])$ for all $t \in \{1, 2, \ldots, T\}$, $h \in [H-1]$ and $s' \in S$, to approximate the distribution $P_\star^c(\cdot|s, a)$ over S for each state-action pair $(s, a) \in S \times A$.

Algorithm 8 Regret Minimization for CMDP with Unknown Context-Dependent Dynamics (RM-UCDD)

inputs:
- MDP parameters: S , A, H, s_0.
- Confidence parameter δ and tuning parameters $\{\beta_t\}_{t=1}^T$.
- Minimum reachability parameter p_{min}.

initialization: for the first $|A|$ rounds, for each action a_i in turn i, run the policy π_i that at any state s plays action a_i, regardless of the context.

for episode $t = |A| + 1, \ldots, T$ **do**

  compute $\widehat{f}_t \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{t-1} \sum_{h=0}^{H-1} (f(c_i, s_h^i, a_h^i) - r_h^i)^2$ using the LSR oracle

  compute $\widehat{P}_t \in \arg\min_{\widetilde{P} \in \mathcal{P}} \sum_{i=1}^{t-1} \sum_{h=0}^{H-1} \sum_{s' \in S} (\widetilde{P}^{c_i}(s'|s_h^i, a_h^i) - \mathbb{I}[s' = s_{h+1}^i])^2$ using the LSR oracle

  observe context $c_t \in \mathcal{C}$

  for $k = |A| + 1, |A| + 2, \ldots, t$ **do**

    compute for all $(s, a) \in S \times A$:

$$\widehat{r}_k^{c_t}(s, a) = \widehat{f}_k(c_t, s, a) + \min\left\{ \frac{1}{p_{min}} \frac{\beta_k}{\sum_{i=1}^{k-1} \mathbb{I}[a = \pi_i(c_t, s)]}, 1 \right\}$$

    define the approximated MDP $\widehat{\mathcal{M}}_k(c_t) = (S, A, \widehat{P}_k^{c_t}, \widehat{r}_k^{c_t}, s_0, H)$

    compute $\pi_k(c_t, \cdot) \in \arg\max_{\pi \in S \to A} V^\pi_{\widehat{\mathcal{M}}_k(c_t)}(s_0)$ using planing algorithm

  play $\pi_t(c_t, \cdot)$ and observe trajectory $\sigma^t = (c_t, s_0^t, a_0^t, r_0^t, s_1^t, \ldots, s_{H-1}^t, a_{H-1}^t, r_{H-1}^t, s_H^t)$

  update the LSR oracles using σ^t

D.3 Regret Analysis

D.3.1 Analysis Outline

For the analysis, for every $t > |A|$ we define the following intermediate MDPs for every context $c \in \mathcal{C}$.

1. $\mathcal{M}^{(f,P)}(c) = (S, A, P^c, f(c, \cdot, \cdot), s_0, H)$, for any $f \in \mathcal{F}$ and $P \in \mathcal{P}$.
2. $\mathcal{M}^{(f_\star, P_\star)}(c)$ is the true model, where $f_\star$ is the true context dependent rewards and $P_\star$ is the true dynamics, which we also denote by $\mathcal{M}(c)$.
3. $\widehat{\mathcal{M}}_t(c) = \mathcal{M}^{(\widehat{r}_t, \widehat{P}_t)}(c)$ is the approximated MDP defined in Algorithm 8.

4. $\mathcal{M}^{(\widehat{r}_t,P_\star)}(c) = (S, A, P^c_\star, \widehat{r}^c_t, s_0, H)$, where $\widehat{r}^c_t$ is the context-dependent approximated rewards function defined in Algorithm 8 and $P^c_\star$ is the true context-dependent dynamics.

For brevity, in the analysis outline we use the following notations for the contextual potential functions at round t.

$$\phi_t(\pi) := \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S^c_h}\frac{q_h(s_h|\pi(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}\right],$$

and

$$\psi_t(\pi) := \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S^c_h}\frac{q_h(s_h|\pi(c;\cdot),P^c_\star)}{p_{min}\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]}\right].$$

We remark that in the detailed analysis, we use the explicit expressions of the context potential functions ϕ_t and ψ_t.

In the following, we analyse the error caused by the dynamics approximation and the rewards approximation separately.

Analysing the error caused by the rewards approximation. (see Sub-subsection D.3.2). Using a similar analysis to that showed for the known dynamics setting (Appendix B), we show (in Lemma 51) that with probability at least $1-\delta/4$ for all $t \geq |A|+1$ and context-dependent policy $\pi \in \Pi_{\mathcal{C}}$ the following holds.

$$\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] \leq 2\beta_t\cdot\psi_t(\pi) + \beta_t\cdot\frac{H|S||A|}{t},$$

and

$$\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}(s_0)] \leq \beta_t\cdot\psi_t(\pi) + \beta_t\cdot\frac{H|S||A|}{t}.$$

Analysing the error caused by the dynamics approximation. (See Sub-subsection D.3.3) We extend the four steps strategy applied to the rewards approximation, to analyse the dynamics approximation. Such an extension is possible, due to the following key observation.

Key observation. Recall $\mathcal{B}$ is a random variable which generate the next state s_{h+1} for the context c given the true dynamics associated with c, $P^c_\star$, the current state s_h and the played action a_h. Recall that by definition, $\mathcal{B}(P^c_\star, s_h, a_h) \sim P^c_\star(\cdot|s_h,a_h)$.

Observation 5 *Since the CMDP is layered and loop free, given the context c_t state s^t_h and action a^t_h, we have that the random variables $\mathcal{B}(P^{c_t}_\star, s^t_h, a^t_h)$ and $(s^t_0, a^t_0, s^t_1, \ldots, s^t_{h-1}, a^t_{h-1})$ are independent random variables.*

Using Observation 5, we are able to adapt Lemma 12, for dynamics approximation using least-square regression. For more details, see Lemma 54.

Using Lemma 54 we extend our uniform convergence bound to the dynamics approximation, in Lemma 56 and apply the four steps strategy present for the rewards to the dynamics approximation.

Step 1: establish uniform convergence bound over any $t \geq 2$ and a fixed sequence of functions $P_2, P_3, \ldots \in \mathcal{P}$ which states the following (see Lemma 57 for more details).

For every $\delta \in (0,1)$, with probability at least $1-\delta/4$ it holds that

$$\sum_{i=1}^{t-1}\mathbb{E}\left[\sum_{h=0}^{H-1}\sum_{s'\in S}(P^{c_i}_t(s'|s^i_h,a^i_h) - P^{c_i}_\star(s'|s^i_h,a^i_h))^2|\mathbb{H}_{i-1}\right] \leq 72H|S|\log(8|\mathcal{P}|t^3/\delta)$$
$$+ 2\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\sum_{s'\in S}\left(P^{c_i}_t(s'|s^i_h,a^i_h) - \mathbb{I}\left[s'=s^{(i,h)}_{\mathcal{B}}\right]\right)^2 - \left(P^{c_i}_\star(s'|s^i_h,a^i_h) - \mathbb{I}\left[s'=s^{(i,h)}_{\mathcal{B}}\right]\right)^2,$$

where $s^{(i,h)}_{\mathcal{B}} \sim \mathcal{B}(P^{c_i}_\star, s^i_h, a^i_h)$.

We prove the above in Lemma 57, using the results of Lemmas 54 and 56.

Step 2: constructing a confidence bound over the expected value of any context-dependent policy with respect to the approximated and true dynamics, which holds with high probability.

Formally, in Lemma 58 we show that with probability at least $1-\delta/4$ for all $t \geq |A|+1$ and every context-dependent policy $\pi \in \Pi_{\mathcal{C}}$ it holds that

$$\left|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t,P_\star)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t,\hat{P}_t)}(c)}(s_0)]\right| \leq 2\sqrt{H|S|\phi_t(\pi)} \cdot \sqrt{72H^2|S|\log(8|\mathcal{P}|t^3/\delta)}.$$

Step 3: Relax the confidence bound in step 2 to be additive. In Lemma 59 we show that under the good event of step 2, for all $t \geq |A|+1$ and every context-dependent policy $\pi \in \Pi_{\mathcal{C}}$, for $\gamma_t = \sqrt{\frac{72t\log(8|\mathcal{P}|t^3/\delta)}{|S||A|}}$, it holds that

$$\left|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t,P_\star)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t,\hat{P}_t)}(c)}(s_0)]\right| \leq \gamma_t H|S|\phi_t(\pi) + \gamma_t\frac{H^2|S|^2|A|}{t}.$$

Step 4: Bounding the sum of contextual potentials similarly to shown for the rewards. For more details, see Lemma 60.

Using all the above, we obtain the optimism lemma (see Lemma 61) which states that under the good events of step 2 for both the dynamics and rewards approximations, for all $t \geq |A|+1$ it hold that

$$\mathbb{E}_c[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)] \leq H|S|\gamma_t \cdot \phi_t(\pi^\star) + \gamma_t\frac{H^2|S|^2|A|}{t} + \beta_t \cdot \psi_t(\pi^\star) + \beta_t\frac{H|S||A|}{t},$$

yielding the following regret bound (see Theorem 63) (and an expected regret bound, see Theorem 62).

Theorem 49 (regret bound) *For every $\delta \in (0,1)$, $T > |A|$ and finite function classes $\mathcal{F}$ and $\mathcal{P}$, let $\beta_t = \sqrt{\frac{17t\log(8|\mathcal{F}|t^3/\delta)}{|S||A|}}$ and $\gamma_t = \sqrt{\frac{72t\log(8|\mathcal{P}|t^3/\delta)}{|S||A|}}$.*

Then, with probability at least $1-\delta$ it holds that

$$\textit{Regret}_T(RM-UCDD) \leq \tilde{O}\left(\max\{H, 1/p_{min}\} \cdot \left(H|S|^{3/2}\sqrt{|A|T\log\frac{|\mathcal{P}|}{\delta}} + \sqrt{T|S||A|\log\frac{|\mathcal{F}|}{\delta}}\right)\right).$$

D.3.2 CONSTRUCTING CONFIDENCE BOUND FOR REWARDS APPROXIMATION

Step 1: Establishing uniform convergence bound over $\mathcal{F}$. We use Lemma 36, presented in Appendix C.

Step 2: Constructing confidence bound over policies with respect to the rewards approximation.

Lemma 50 (confidence of policies w.r.t rewards approximation) *Consider Algorithm 8 that at each initialization round $t \leq |A|$, plays the policy that always choose action a_t, and at each round $t \geq |A|+1$ selects π_t based on the history $\mathbb{H}_{t-1}$.*

Then, for every $\delta \in (0,1)$ with probability at least $1-\delta/4$ for all $t \geq |A|+1$ and every context-dependent policy $\pi \in \Pi_{\mathcal{C}}$, the following holds.

$$\begin{aligned}&\left|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(f_\star,P_\star)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\hat{f}_t,P_\star)}(c)}(s_0)]\right| \\ &\leq \sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]} \cdot \sqrt{68H\log(8|\mathcal{F}|t^3/\delta)}.\end{aligned}$$

Proof The Lemma is obtained using the same derivation presented in Lemma 18, using Lemma 36 (instead of Lemma 16) ■

Step 3: Relax the confidence bound to be additive.

Lemma 51 (the "square trick" relaxation for rewards and context-dependent dynamics) *Under the good event stated in Lemma 50 for every $t \geq |A|+1$ and a context-dependent policy $\pi \in \Pi_{\mathcal{C}}$ it holds that*

$$\begin{aligned}&\left|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(\hat{f}_t,P_\star)(c)}(s_0)]\right| \\ \leq &\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S^c_h}\frac{\beta_t\cdot q_h(s_h|\pi(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}\right] + \beta_t\frac{H|S||A|}{t},\end{aligned}$$

where $\beta_t = \sqrt{\frac{17t\log(8|\mathcal{F}|t^3/\delta)}{|S||A|}}$.

The above implies in particular that

$$\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(\hat{r}_t,P_\star)(c)}(s_0)] \leq \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S^c_h}\frac{1}{p_{min}}\frac{\beta_t\cdot q_h(s_h|\pi(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]}\right] + \beta_t\frac{H|S||A|}{t}, \quad (26)$$

and

$$\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(\hat{r}_t,P_\star)(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] \leq 2\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S^c_h}\frac{1}{p_{min}}\frac{\beta_t\cdot q_h(s_h|\pi(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]}\right] + \beta_t\frac{H|S||A|}{t}, \quad (27)$$

where $\widehat{r}^c_t$ is the rewards function defined in Algorithm 8.

Proof Using the same derivation presented in Lemma 19, where $\beta_t = \sqrt{\frac{17t\log(8|\mathcal{F}|t^3/\delta)}{|S||A|}}$ we obtain the first part of the lemma.

For the second part, by the minimum reachability assumption it holds that

$$\begin{aligned}&\left|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(\hat{f}_t,P_\star)(c)}(s_0)]\right| \\ \leq &\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S^c_h}\frac{\beta_t\cdot q_h(s_h|\pi(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}\right] + \beta_t\frac{H|S||A|}{t} \\ \leq &\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S^c_h}\frac{1}{p_{min}}\frac{\beta_t\cdot q_h(s_h|\pi(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]}\right] + \beta_t\frac{H|S||A|}{t}.\end{aligned} \quad (28)$$

For every context $c \in \mathcal{C}$ (later we take the expectation over c), by the value function definition, for any $t \geq |A| + 1$ and a context-dependent $\pi \in \Pi_{\mathcal{C}}$ the following holds.

$$
\begin{aligned}
&V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t, P_\star)}(c)}(s_0) = \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \sum_{a_h \in A} q_h(s_h, a_h|\pi(c;\cdot), P^c_\star) \cdot \widehat{r}^c_t(s_h, a_h) \\
&= \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \sum_{a_h \in A} q_h(s_h, a_h|\pi(c;\cdot), P^c_\star) \cdot \left(\widehat{f}_t(c, s_h, a_h) + \min\left\{ \frac{1}{p_{min}} \frac{\beta_t}{\sum_{i=1}^{t-1} \mathbb{I}[a_h = \pi_i(c; s_h)]}, 1 \right\} \right) \\
&= \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} q_h(s_h, \pi(c; s_h)|\pi(c;\cdot), P^c_\star) \cdot \left(\widehat{f}_t(c, s_h, \pi(c; s_h)) + \min\left\{ \frac{1}{p_{min}} \frac{\beta_t}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c; s_h) = \pi_i(c; s_h)]}, 1 \right\} \right) \\
&= \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} q_h(s_h|\pi(c;\cdot), P^c_\star) \cdot \left(\widehat{f}_t(c, s_h, \pi(c; s_h)) + \min\left\{ \frac{1}{p_{min}} \frac{\beta_t}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c; s_h) = \pi_i(c; s_h)]}, 1 \right\} \right) \\
&= V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{f}_t, P_\star)}(c)}(s_0) + \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} q_h(s_h|\pi(c;\cdot), P^c_\star) \cdot \min\left\{ \frac{1}{p_{min}} \frac{\beta_t}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c; s_h) = \pi_i(c; s_h)]}, 1 \right\},
\end{aligned}
\tag{29}
$$

where the third and forth equations are since $\pi(c;\cdot)$ is deterministic, hence

$$
\sum_{a_h \in A} q_h(s_h, a_h|\pi(c;\cdot), P^c_\star) = \sum_{a_h \in A} q_h(s_h|\pi(c;\cdot), P^c_\star) \cdot \pi(a_h|c; s_h) = q_h(s_h, \pi(c; s_h)|\pi(c;\cdot), P^c_\star) = q_h(s_h|\pi(c;\cdot), P^c_\star) \cdot 1.
$$

We remark that $\frac{1}{p_{min}} \frac{\beta_t}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c;s) = \pi_i(c;s)]} > 0$ for every c, s and π.

Thus, equation (29) further implies that

$$
V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t, P_\star)}(c)}(s_0) \geq V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{f}_t, P_\star)}(c)}(s_0), \tag{30}
$$

and, using that $\sum_i a_i \cdot \min\{b_i, c_i\} \leq \sum_i a_i \cdot b_i$ where $a_i, b_i, c_i \geq 0$ for all i, we also have

$$
V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t, P_\star)}(c)}(s_0) \leq V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{f}_t, P_\star)}(c)}(s_0) + \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} q_h(s_h|\pi(c;\cdot), P^c_\star) \cdot \frac{1}{p_{min}} \frac{\beta_t}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c; s_h) = \pi_i(c; s_h)]}. \tag{31}
$$

Clearly, inequalities (30) and (31) also hold when taking the expectation over the context c, on both sides.

To obtain inequality (26) we combine inequalities (30) and (28):

$$
\begin{aligned}
\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t, P_\star)}(c)}(s_0)] &\leq \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{f}_t, P_\star)}(c)}(s_0)] && \text{(By inequality (30))} \\
&\leq \mathbb{E}_c\left[\sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \frac{1}{p_{min}} \frac{\beta_t \cdot q_h(s_h|\pi(c;\cdot), P^c_\star)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c; s_h) = \pi_i(c; s_h)]} \right] + \beta_t \frac{H|S||A|}{t}. && \text{(By inequality (28))}
\end{aligned}
$$

To obtain inequality (27) we combine inequalities (31) and (28):

$$
\begin{aligned}
&\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t, P_\star)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] \\
&\leq \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{f}_t, P_\star)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)] + \mathbb{E}_c\left[\sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \frac{1}{p_{min}} \frac{\beta_t \cdot q_h(s_h|\pi(c;\cdot), P^c_\star)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c; s_h) = \pi_i(c; s_h)]} \right] \\
&\hspace{30em} \text{(By inequality (31))}
\end{aligned}
$$

$$\leq \left|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{f}_t,P_\star)(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(c)}(s_0)]\right| + \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{1}{p_{min}}\frac{\beta_t\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]}\right]$$

$$\leq \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{1}{p_{min}}\frac{\beta_t\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]}\right] + \beta_t\frac{H|S||A|}{t} \qquad \text{(By inequality (28))}$$

$$+\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{1}{p_{min}}\frac{\beta_t\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]}\right]$$

$$= 2\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{1}{p_{min}}\frac{\beta_t\cdot q_h(s_h,|\pi(c;\cdot),P^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]}\right] + \beta_t\frac{H|S||A|}{t},$$

as stated. ■

Step 4: Bounding the contextual potential for both dynamics and rewards. We show in Lemma 60 bounds on the cumulative contextual potential functions that use for both the dynamics an rewards approximation analysis.

D.3.3 CONSTRUCTING CONFIDENCE BOUND FOR DYNAMICS APPROXIMATION

Step 0: Bounding the approximated rewards function. The following lemma shows that the approximated rewards function is bounded in $[0,2]$ for every $t > |A|$ any any context $c \in \mathcal{C}$.

Lemma 52 (approximated rewards are bounded) *For every $t \geq |A|+1$ and a context c it holds that $\widehat{r}_t^c : S\times A \to [0,2]$, where $\widehat{r}_t^c$ is the context-dependent rewards function defined in round t of Algorithm 8.*

Proof Fix $t \geq |A|+1$ and a context $c \in \mathcal{C}$. By definition,

$$\widehat{r}_t^c(s,a) = \widehat{f}_t(c,s,a) + \min\left\{\frac{1}{p_{min}}\frac{\beta_t}{\sum_{i=1}^{t-1}\mathbb{I}[a=\pi_i(c;s)]}, 1\right\}.$$

By $\mathcal{F}$ definition, $\widehat{f}_t : \mathcal{C}\times S\times A \to [0,1]$. In addition, by π_i choice in every initialization round $i \in \{1,2,\ldots,|A|\}$ it holds that

$$\sum_{i=1}^{t-1}\mathbb{I}[a=\pi_i(c;s)] \geq 1 \;\; \forall (s,a,c) \in S\times A\times\mathcal{C}.$$

Hence,

$$\frac{1}{p_{min}}\frac{\beta_t}{\sum_{i=1}^{t-1}\mathbb{I}[a=\pi_i(c;s)]} \geq 0 \;\; \forall (s,a,c) \in S\times A\times\mathcal{C}.$$

Overall, for all $s\in S$, $a\in A$, $c\in\mathcal{C}$ it holds that

$$0 \leq \widehat{r}_t^c(s,a) = \widehat{f}_t(c,s,a) + \min\left\{\frac{1}{p_{min}}\frac{\beta_t}{\sum_{i=1}^{t-1}\mathbb{I}[a=\pi_i(c;s)]}, 1\right\} \leq \widehat{f}_t(c,s,a) + 1 \leq 2,$$

which implies the lemma. ■

Corollary 53 (value bound) *For any dynamics P, round $t \geq |A|+1$ and a context $c\in\mathcal{C}$ it holds that*

$$|V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,P)(c),h}(s)| \leq 2H \quad \forall\pi\in\Pi_{\mathcal{C}},\; s\in S,\; h\in[H].$$

Step 1: Establishing uniform convergence bound over $\mathcal{P}$. Let $\mathcal{B}$ be a random variable which generate the next state s_{h+1} given the true dynamics associated with c, $P^c_\star$, the state s_h and the action a_h. $\mathcal{B}$ is defined as follows,

$$\mathcal{B}(P^c_\star, s_h, a_h) \sim P^c_\star(\cdot|s_h, a_h).$$

By definition, $\mathcal{B}$ satisfies the following for all $s' \in S$.

$$\mathbb{E}\left[\mathbb{I}[s' = \mathcal{B}(P^c_\star, s_h, a_h)]\middle| c, s_h, a_h\right] = P^c_\star(s'|s_h, a_h).$$

Observation 6 *Given the context c_t state s^t_h and action a^t_h, we have that the random variables $\mathcal{B}(P^{c_t}_\star, s^t_h, a^t_h)$ and $(s^t_0, a^t_0, s^t_1, \ldots, s^t_{h-1}, a^t_{h-1})$ are independent random variables.*

Conclusion 1 *Our samples for the dynamics approximation satisfies the requirements of Lemma 54 (see bellow).*

The following is an adaption of Lemma 4.2 from Agarwal et al. (2012) for the dynamics.

Lemma 54 *Fix a function $\widetilde{P} \in \mathcal{P}$. Suppose we sample context c using the distribution $\mathcal{D}$ and than sample state s_h and action a_h using some policy π. Let $s' \in S$ denote a possible next state and define the random variable*

$$Y_{c,s_h,a_h,\mathcal{B}} = \sum_{s'} (\widetilde{P}^c(s'|s_h, a_h) - \mathbb{I}[s' = s_{\mathcal{B}}])^2 - (P^c_\star(s'|s_h, a_h) - \mathbb{I}[s' = s_{\mathcal{B}}])^2,$$

where $s_{\mathcal{B}} \sim \mathcal{B}(P^c_\star, s_h, a_h)$.

Then, the followings hold.

1. $\mathbb{E}_{c,s_h,a_h,\mathcal{B}}[Y_{c,s_h,a_h,\mathcal{B}}] = \mathbb{E}_{c,s_h,a_h}\left[\sum_{s' \in S}\left(\widetilde{P}^c(s'|s_h, a_h) - P^c_\star(s'|s_h, a_h)\right)^2\right]$.
2. $\mathbb{V}_{c,s_h,a_h,\mathcal{B}}[Y_{c,s_h,a_h,\mathcal{B}}] \leq 4|S| \cdot \mathbb{E}_{c,s_h,a_h,\mathcal{B}}[Y_{c,s_h,a_h,\mathcal{B}}]$.
3. $\mathbb{V}\left[\sum_{h=0}^{H-1} Y_{c,s_h,a_h,\mathcal{B}}\right] \leq 4H|S| \cdot \mathbb{E}\left[\sum_{h=0}^{H-1} Y_{c,s_h,a_h,\mathcal{B}}\right]$.

Proof Let us rearrange $Y_{c,s_h,a_h,\mathcal{B}}$ as

$$Y_{c,s_h,a_h,\mathcal{B}} = \sum_{s' \in S} (\widetilde{P}^c(s'|s_h, a_h) - P^c_\star(s'|s_h, a_h))(\widetilde{P}^c(s'|s_h, a_h) + P^c_\star(s'|s_h, a_h) - 2\mathbb{I}[s' = s_{\mathcal{B}}]). \tag{32}$$

Consider the following derivation.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{c,s_h,a_h,\mathcal{B}}[Y_{c,s_h,a_h,\mathcal{B}}] \\
&= \mathop{\mathbb{E}}_{c,s_h,a_h,\mathcal{B}}\left[\sum_{s'}\left(\widetilde{P}^c(s'|s_h, a_h) - P^c_\star(s'|s_h, a_h)\right)\left(\widetilde{P}^c(s'|s_h, a_h) + P^c_\star(s'|s_h, a_h) - 2\mathbb{I}[s' = s_{\mathcal{B}}]\right)\right] \\
&= \mathop{\mathbb{E}}_{c,s_h,a_h}\left[\mathop{\mathbb{E}}_{\mathcal{B}}\left[\sum_{s'}(\widetilde{P}^c(s'|s_h, a_h) - P^c_\star(s'|s_h, a_h))(\widetilde{P}^c(s'|s_h, a_h) + P^c_\star(s'|s_h, a_h) - 2\mathbb{I}[s' = s_{\mathcal{B}}])\middle| c, s_h, a_h\right]\right] \\
&= \mathop{\mathbb{E}}_{c,s_h,a_h}\left[\sum_{s'}\left(\widetilde{P}^c(s'|s_h, a_h) - P^c_\star(s'|s_h, a_h)\right)\left(\widetilde{P}^c(s'|s_h, a_h) + P^c_\star(s'|s_h, a_h) - 2\mathbb{E}_{\mathcal{B}}\left[\mathbb{I}[s' = s_{\mathcal{B}}]|c, s_h, a_h\right]\right)\right] \\
&= \mathop{\mathbb{E}}_{c,s_h,a_h}\left[\sum_{s'}\left(\widetilde{P}^c(s'|s_h, a_h) - P^c_\star(s'|s_h, a_h)\right)^2\right],
\end{aligned}$$

where the second identity is since $s_{\mathcal{B}} \sim \mathcal{B}(P^c_\star, s_h, a_h)$ and the randomness of $\mathcal{B}$ is determined completely by c, s_h and a_h.

The third inequality is since the only ransom variable that depends on $\mathcal{B}$ is $s_{\mathcal{B}}$.

The forth identity uses $\mathcal{B}$ definition which implies that for all $s' \in S$ it holds that $P^c_\star(s'|s,a) = \mathbb{E}_{\mathcal{B}}[\mathbb{I}[s' = \mathcal{B}(P^c_\star, s, a)]|c, s, a]$, proving the first part of the lemma.

For the second part, note that $\widetilde{P}^c(s'|s_h, a_h)$, $P^c_\star(s'|s_h, a_h)$ and $\mathbb{I}[s' = s_{\mathcal{B}}]$ are all between 0 and 1. Hence from equation (32) we obtain

$$\begin{aligned}
Y^2_{c,s_h,a_h,\mathcal{B}} =& \left(\sum_{s'}(\widetilde{P}^c(s'|s_h,a_h) - P^c_\star(s'|s_h,a_h))(\widetilde{P}^c(s'|s_h,a_h) + P^c_\star(s'|s_h,a_h) - 2\mathbb{I}[s' = s_{\mathcal{B}}])\right)^2 \\
\leq& |S| \sum_{s'}(\widetilde{P}^c(s'|s_h,a_h) - P^c_\star(s'|s_h,a_h))^2(\widetilde{P}^c(s'|s_h,a_h) + P^c_\star(s'|s_h,a_h) - 2\mathbb{I}[s' = s_{\mathcal{B}}])^2 \\
\leq& 4|S| \sum_{s'}(\widetilde{P}^c(s'|s_h,a_h) - P^c_\star(s'|s_h,a_h))^2,
\end{aligned}$$

yielding the second part of the lemma since

$$\begin{aligned}
\mathop{\mathbb{V}}_{c,s_h,a_h,\mathcal{B}}[Y_{c,s_h,a_h,\mathcal{B}}] \leq& \mathop{\mathbb{E}}_{c,s_h,a_h,\mathcal{B}}[Y^2_{c,s_h,a_h,\mathcal{B}}] \\
\leq& 4|S| \mathop{\mathbb{E}}_{c,s_h,a_h}\left[\sum_{s'}(\widetilde{P}^c(s'|s_h,a_h) - P^c_\star(s'|s_h,a_h))^2\right] \\
=& 4|S| \mathop{\mathbb{E}}_{c,s_h,a_h,\mathcal{B}}[Y_{c,s_h,a_h,\mathcal{B}}].
\end{aligned}$$

For the third part, by norms inequality and part 2 of the lemma we have

$$\left(\sum_{h=0}^{H-1} Y_{c,s_h,a_h,\mathcal{B}}\right)^2 \leq H \sum_{h=0}^{H-1} Y^2_{c,s_h,a_h,\mathcal{B}} \leq 4H|S| \sum_{s'}(\widetilde{P}^c(s'|s_h,a_h) - P^c_\star(s'|s_h,a_h))^2,$$

yielding the third part of the lemma as

$$\begin{aligned}
\mathbb{V}\left[\sum_{h=0}^{H-1} Y_{c,s_h,a_h,\mathcal{B}}\right] \leq& \mathbb{E}\left[\left(\sum_{h=0}^{H-1} Y_{c,s_h,a_h,\mathcal{B}}\right)^2\right] \\
\leq& 4H|S|\,\mathbb{E}\left[\sum_{h=0}^{H-1}\sum_{s'}(\widetilde{P}^c(s'|s_h,a_h) - P^c_\star(s'|s_h,a_h))^2\right] \\
=& 4H|S|\,\mathbb{E}\left[\sum_{h=0}^{H-1} Y_{c,s_h,a_h,\mathcal{B}}\right],
\end{aligned}$$

as stated. ■

Definition 55 *For every round $t \geq 1$, layer $h \in [H-1]$ and a function $\widetilde{P} \in \mathcal{P}$ we define the random variable*

$$Y_{\widetilde{P},c_t,s^t_h,a^t_h,s^t_{h+1}} = \sum_{s' \in S}\left(\widetilde{P}^{c_t}(s'|s^t_h,a^t_h) - \mathbb{I}\left[s' = s^t_{s+1}\right]\right)^2 - \left(P^{c_t}_\star(s'|s^t_h,a^t_h) - \mathbb{I}\left[s' = s^t_{h+1}\right]\right)^2$$

where $(c_t, s^t_h, a^t_h) \sim \mathcal{D}(c_t) \cdot q_h(s^t_h, a^t_h|\pi_t(c_t;\cdot), P^{c_t}_\star)$ and $s^t_{h+1} \sim \mathcal{B}(P^{c_t}_\star, s^t_h, a^t_h)$.

Observation 7 *For all $i \in [t-1]$ and $h \in [H-1]$ we have that*

$$(c_i, s_h^i, a_h^i) \sim \mathcal{D}(c_i) \cdot q_h(s_h^i, a_h^i | \pi_i(c_i; \cdot), P_\star^{c_i}).$$

Recall that for all $i \leq |A|$ π_i is selected deterministically and for all $i > |A|$ π_i is determined completely by the history $\mathbb{H}_{i-1}$. Hence, by linearity of expectation, for all $\widetilde{P} \in \mathcal{P}$ it holds that

$$\mathbb{E}\left[\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i} \middle| \mathbb{H}_{i-1}\right] = \sum_{h=0}^{H-1} \mathop{\mathbb{E}}_{c_i,s_h^i,a_h^i,s_{h+1}^i} \left[Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i} \middle| \mathbb{H}_{i-1}\right].$$

Similarly,

$$\mathbb{E}\left[\sum_{h=0}^{H-1} \sum_{s' \in S} (\widetilde{P}^{c_i}(s'|s_h^i, a_h^i) - P_\star^{c_i}(s'|s_h^i, a_h^i))^2 \middle| \mathbb{H}_{i-1}\right] = \sum_{h=0}^{H-1} \mathop{\mathbb{E}}_{c_i,s_h^i,a_h^i} \left[\sum_{s' \in S} (\widetilde{P}^{c_i}(s'|s_h^i, a_h^i) - P_\star^{c_i}(s'|s_h^i, a_h^i))^2 \middle| \mathbb{H}_{i-1}\right].$$

Observation 8 *Since for all $i \leq |A|$ π_i is selected deterministically and for all $i > |A|$ π_i is determined completely by the history $\mathbb{H}_{i-1}$, for any function $\widetilde{P} \in \mathcal{P}$, it holds that $\{Z_i(\widetilde{P})\}_{i=1}^{t-1}$ is a martingale difference sequence of length $t-1$, where the filtration is $\{\mathbb{H}_i\}_{i=1}^{t-1}$ and*

$$Z_i(\widetilde{P}) := \mathbb{E}\left[\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i} \middle| \mathbb{H}_{i-1}\right] - \sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i}.$$

(Recall that for all $i \geq 1$ we defined that $\mathbb{H}_i = (\sigma^1, \ldots, \sigma^i)$ and $\mathbb{H}_0$ is the empty history.)

In addition, since $\mathbb{E}\left[\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i} \middle| \mathbb{H}_{i-1}\right]$ is determined given $\mathbb{H}_{i-1}$, it holds that

$$\mathbb{V}\left[Z_i(\widetilde{P}) \,|\mathbb{H}_{i-1}\right] = \mathbb{V}\left[\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i} \middle| \mathbb{H}_{i-1}\right].$$

Lemma 56 (uniform convergence over $\mathcal{P}$) *For a fixed $t \geq 2$ and a fixed $\delta_t \in (0, 1/e^2)$ with probability at least $1 - \log_2(t-1)\delta_t$ it holds that*

$$\begin{aligned}
&\sum_{i=1}^{t-1} \sum_{h=0}^{H-1} \mathop{\mathbb{E}}_{c_i,s_h^i,a_h^i} \left[\sum_{s' \in S} (\widetilde{P}^{c_i}(s'|s_h^i, a_h^i) - P_\star^{c_i}(s'|s_h^i, a_h^i))^2 \middle| \mathbb{H}_{i-1}\right] \\
&= \sum_{i=1}^{t-1} \mathbb{E}\left[\sum_{h=0}^{H-1} \sum_{s' \in S} (\widetilde{P}^{c_i}(s'|s_h^i, a_h^i) - P_\star^{c_i}(s'|s_h^i, a_h^i))^2 \middle| \mathbb{H}_{i-1}, \right] \\
&\leq 72H|S| \log(|\mathcal{P}|/\delta_t) + 2 \sum_{i=1}^{t-1} \sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i}.
\end{aligned}$$

uniformly over all $\widetilde{P} \in \mathcal{P}$.

Proof Fix a function $\widetilde{P} \in \mathcal{P}$. Consider the following random variables defined in 55

$$Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i} = \sum_{s' \in S} \left(\widetilde{P}^{c_i}(s'|s_h^i, a_h^i) - \mathbb{I}\left[s' = s_{h+1}^i\right]\right)^2 - \left(P_\star^{c_i}(s'|s_h^i, a_h^i) - \mathbb{I}\left[s' = s_{h+1}^i\right]\right)^2$$

where $s_{h+1}^i \sim \mathcal{B}(P_\star^{c_i}, s_h^i, a_h^i)$, for all $i \in \{1, 2, \ldots, t-1\}$ and $h \in [H-1]$.

Notice that $|Y_{\widetilde{P},c_t,s_h^i,a_h^i,s_{h+1}^i}| \leq 2$ for any $\widetilde{P}, c_i, s_h^i, a_h^i, s_{h+1}^i$.

Hence, $\left|\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i}\right| \leq 2H$ for all $i \in \{1, 2, \ldots, t-1\}$.

By Freedman's inequality (Lemma 13) and Observation 8, for $\delta_t < 1/e^2$, with probability at least $1 - \log_2(t-1)\delta_t/|\mathcal{P}|$ it holds that

$$\sum_{i=1}^{t-1} \mathbb{E}\left[\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i} \middle| \mathbb{H}_{i-1}\right] - \sum_{i=1}^{t-1}\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i}$$
$$\leq 4\sqrt{\sum_{i=1}^{t-1} \mathbb{V}\left[\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i} \middle| \mathbb{H}_{i-1}\right] \log(|\mathcal{P}|/\delta_t)} + 2 \cdot 2H \log(|\mathcal{P}|/\delta_t).$$

By Lemma 54 part 4, for all $i \in \{1, 2, \ldots, t-1\}$ it holds that

$$\mathbb{V}\left[\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i} \middle| \mathbb{H}_{i-1}\right] \leq 4H|S|\mathbb{E}\left[\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i} \middle| \mathbb{H}_{i-1}\right].$$

Therefore, by combine the two inequalities we obtain

$$\sum_{i=1}^{t-1} \mathbb{E}\left[\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i} \middle| \mathbb{H}_{i-1}\right]$$
$$\leq 4\sqrt{\sum_{i=1}^{t-1} \mathbb{V}\left[\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i} \middle| \mathbb{H}_{i-1}\right] \log(|\mathcal{P}|/\delta_t)} + 2 \cdot 2H \log(|\mathcal{P}|/\delta_t) + \sum_{i=1}^{t-1}\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i}$$
$$\leq 8\sqrt{H|S|\sum_{i=1}^{t-1} \mathbb{E}\left[\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i} \middle| \mathbb{H}_{i-1}\right] \log(|\mathcal{P}|/\delta_t)} + 4H|S|\log(|\mathcal{P}|/\delta_t) + \sum_{i=1}^{t-1}\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i}.$$

This implies that for all $\widetilde{P} \in \mathcal{P}$,

$$\sum_{i=1}^{t-1} \mathbb{E}\left[\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i} \middle| \mathbb{H}_{i-1}\right] - 8\sqrt{H|S|\sum_{i=1}^{t-1} \mathbb{E}\left[\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i} \middle| \mathbb{H}_{i-1}\right] \log(|\mathcal{P}|/\delta_t)}$$
$$\leq 4H|S|\log(|\mathcal{P}|/\delta_t) + \sum_{i=1}^{t-1}\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i}.$$

By adding $16H|S|\log(|\mathcal{P}|/\delta_t)$ to both sides of the inequality we obtain

$$\left(\sqrt{\sum_{i=1}^{t-1} \mathbb{E}\left[\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i} \middle| \mathbb{H}_{i-1}\right]} - 4\sqrt{H|S|\log(|\mathcal{P}|/\delta_t)}\right)^2 \leq 20H|S|\log(|\mathcal{P}|/\delta_t) + \sum_{i=1}^{t-1}\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i}.$$

Hence,

$$\sqrt{\sum_{i=1}^{t-1} \mathbb{E}\left[\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i} \middle| \mathbb{H}_{i-1}\right]} \leq 4\sqrt{H|S|\log(|\mathcal{P}|/\delta_t)} + \sqrt{20H|S|\log(|\mathcal{P}|/\delta_t) + \sum_{i=1}^{t-1}\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i}},$$

yielding,

$$\sum_{i=1}^{t-1} \mathbb{E}\left[\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i} \middle| \mathbb{H}_{i-1}\right] \leq \left(4\sqrt{H|S|\log(|\mathcal{P}|/\delta_t)} + \sqrt{20H|S|\log(|\mathcal{P}|/\delta_t) + \sum_{i=1}^{t-1}\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i}}\right)^2.$$

The latter inequality further implies (using $(a+b)^2 \leq 2a^2 + 2b^2$) that for all $\widetilde{P} \in \mathcal{P}$ it holds that

$$\sum_{i=1}^{t-1} \mathbb{E}\left[\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i} \middle| \mathbb{H}_{i-1}\right] \leq 72H|S|\log(|\mathcal{P}|/\delta_t) + 2\sum_{i=1}^{t-1}\sum_{h=0}^{H-1} Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i}. \tag{33}$$

Lastly, by combining inequality (33) with Lemma 54 we obtain the lemma since

$$\begin{aligned}
&\sum_{i=1}^{t-1}\mathbb{E}\left[\sum_{h=0}^{H-1}\sum_{s'\in S}(\widetilde{P}^{c_i}(s'|s_h^i,a_h^i)-P_\star^{c_i}(s'|s_h^i,a_h^i))^2 \;\middle|\; \mathbb{H}_{i-1}\right]\\
&=\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\mathop{\mathbb{E}}_{c_i,s_h^i,a_h^i}\left[\sum_{s'\in S}(\widetilde{P}^{c_i}(s'|s_h^i,a_h^i)-P_\star^{c_i}(s'|s_h^i,a_h^i))^2 \;\middle|\; \mathbb{H}_{i-1}\right] && \text{(By Observation 7)}\\
&=\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\mathop{\mathbb{E}}_{c_i,s_h^i,a_h^i,s_{h+1}^i}\left[Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i} \;\middle|\; \mathbb{H}_{i-1}\right] && \text{(By Lemma 54 part 1)}\\
&=\sum_{i=1}^{t-1}\mathbb{E}\left[\sum_{h=0}^{H-1}Y_{\widetilde{P},c_i,s_h^i,s_h^i,s_{h+1}^i} \;\middle|\; \mathbb{H}_{i-1}\right] && \text{(By Observation 7)}\\
&\le 72H|S|\log(|\mathcal{P}|/\delta_t)+2\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}Y_{\widetilde{P},c_i,s_h^i,a_h^i,s_{h+1}^i}. && \text{(By inequality 33)}
\end{aligned}$$

■

Lemma 57 (uniform convergence over any sequence of estimators for the dynamics) *For any $\delta\in(0,1)$, with probability at least $1-\delta/4$ it holds that*

$$\begin{aligned}
&\sum_{i=1}^{t-1}\mathbb{E}\left[\sum_{h=0}^{H-1}\sum_{s'\in S}(P_t^{c_i}(s'|s_h^i,a_h^i)-P_\star^{c_i}(s'|s_h^i,a_h^i))^2|\mathbb{H}_{i-1}\right]\\
=&\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\mathop{\mathbb{E}}_{c_i,s_h^i,a_h^i}\left[\sum_{s'\in S}(P_t^{c_i}(s'|s_h^i,a_h^i)-P_\star^{c_i}(s'|s_h^i,a_h^i))^2|\mathbb{H}_{i-1}\right]\\
\le&72H|S|\log(8|\mathcal{P}|t^3/\delta)\\
&+2\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\sum_{s'\in S}\left(P_t^{c_i}(s'|s_h^i,a_h^i)-\mathbb{I}\left[s'=s_{h+1}^i\right]\right)^2-\left(P_\star^{c_i}(s'|s_h^i,a_h^i)-\mathbb{I}\left[s'=s_{h+1}^i\right]\right)^2,
\end{aligned}$$

where $s_{h+1}^i\sim\mathcal{B}(P_\star^{c_i},s_h^i,a_h^i)$.

The above holds for any $t\ge 2$ and a fixed sequence of probability functions $P_2,P_3,\ldots\in\mathcal{P}$.

Proof For a fixed $\delta\in(0,1)$, take $\delta_t=\delta/8t^3$ and apply union bound to Lemma 56 with all $t\ge 2$. We have,

$$\sum_{t=1}^{\infty}\delta_t\log(t-1)=\sum_{t=1}^{\infty}\delta/8t^3\log(t-1)\le\sum_{t=1}^{\infty}\frac{\delta}{8t^2}\le\frac{\delta}{4},$$

Hence, by Lemma 56 we have with probability at least $1-\frac{\delta}{4}$ that

$$\begin{aligned}
&\sum_{i=1}^{t-1}\mathbb{E}\left[\sum_{h=0}^{H-1}\sum_{s'\in S}(P_t^{c_i}(s'|s_h^i,a_h^i)-P_\star^{c_i}(s'|s_h^i,a_h^i))^2\middle|\mathbb{H}_{i-1}\right]\\
&=\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\mathop{\mathbb{E}}_{c_i,s_h^i,a_h^i}\left[\sum_{s'\in S}(P_t^{c_i}(s'|s_h^i,a_h^i)-P_\star^{c_i}(s'|s_h^i,a_h^i))^2\middle|\mathbb{H}_{i-1}\right]\\
&\le 72H|S|\log(8|\mathcal{P}|t^3/\delta)+2\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}Y_{P_t,c_t,s_h^i,a_h^i,s_{h+1}^i}\\
&=72H|S|\log(8|\mathcal{P}|t^3/\delta)+2\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\sum_{s'\in S}\left(P_t^{c_i}(s'|s_h^i,a_h^i)-\mathbb{I}\left[s'=s_{h+1}^i\right]\right)^2-\left(P_\star^{c_i}(s'|s_h^i,a_h^i)-\mathbb{I}\left[s'=s_{h+1}^i\right]\right)^2.
\end{aligned}$$

where $s^i_{h+1} \sim \mathcal{B}(P^{c_i}_\star, s^i_h, a^i_h)$. The above holds for any $t \geq 2$ and a fixed sequence $P_2, P_3, \ldots \in \mathcal{P}$. ■

Step 2: Constructing confidence bound over policies with respect to the dynamics approximation.

Lemma 58 (confidence of policies over dynamics approximation) *Consider Algorithm 8 that at each initialization round $t \leq |A|$, plays the policy that always choose action a_t, and at each round $t \geq |A|+1$ selects π_t based on the history $\mathbb{H}_{t-1}$.*

Then, for any $\delta \in (0,1)$ with probability at least $1-\delta/4$ for all $t \geq |A|+1$ and every policy $\pi \in \Pi_{\mathcal{C}}$ the following holds.

(1) $\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t,\widehat{P}_t)}(c)}(s_0)]$

$$\leq \sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h \in S^c_h} \frac{2H \cdot |S| \cdot q_h(s_h|\pi(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}\right]} \cdot \sqrt{2 \cdot 72 \cdot H^2 \cdot |S| \cdot \log(8|\mathcal{P}|t^3/\delta)}.$$

(2) $\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t,\widehat{P}_t)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}(s_0)]$

$$\leq \sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h \in S^c_h} \frac{2H \cdot |S| \cdot q_h(s_h|\pi(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}\right]} \cdot \sqrt{2 \cdot 72 \cdot H^2 \cdot |S| \cdot \log(8|\mathcal{P}|t^3/\delta)}.$$

The proof is similar to shown for the rewards in Lemma 18.

Proof For any $\widetilde{P} \in \mathcal{P}$ consider the random variables defined in 55

$$Y_{\widetilde{P},c_i,s^i_h,a^i_h,s^i_{h+1}} = \sum_{s' \in S}(\widetilde{P}^{c_i}(s'|s^i_h,a^i_h) - \mathbb{I}[s'=s^i_{h+1}])^2 - (P^{c_i}_\star(s'|s^i_h,a^i_h) - \mathbb{I}[s'=s^i_{h+1}])^2,$$

where $s^i_{h+1} \sim \mathcal{B}(P^{c_i}_\star, s^i_h, a^i_h)$, and $(c_t, s^t_h, a^t_h) \sim \mathcal{D}(c_t) \cdot q_h(s^t_h, a^t_h|\pi_t(c_t;\cdot), P^{c_t})$, for all $t \in \{1,2,\ldots,T\}$ and $h \in [H-1]$.

We first show the following auxiliary claim.

Claim 2 *For all $t \geq 2$ it holds that*

$$\sum_{i=1}^{t-1}\sum_{h=0}^{H-1} \mathop{\mathbb{E}}_{c_i,s^i_h,a^i_h}\left[\sum_{s' \in S}(\widehat{P}^{c_i}_t(s'|s^i_h,a^i_h) - P^{c_i}_\star(s'|s^i_h,a^i_h))^2 | \mathbb{H}_{i-1}\right] =$$

$$\mathbb{E}_c\left[\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\sum_{s_h \in S^c_h} q_h(s_h|\pi_i(c;\cdot),P^c_\star)\sum_{s' \in S}(\widehat{P}^{c_i}_t(s'|s^i_h,\pi_i(c_i,s^i_h)) - P^{c_i}_\star(s'|s^i_h,\pi_i(c_i,s^i_h)))^2\right]$$

Proof Recall that for all $i \in \{1,2,\ldots,|A|\}$, π_i is a deterministically selected policy uses for initialization. For all $i > |A|$, π_i is determined completely by the history $\mathbb{H}_{i-1}$, recalling that for all i, π_i is a deterministic policy.

Hence, for any function $\widetilde{P} \in \mathcal{P}$, round $i \in \{1, \ldots, t-1\}$ and layer $h \in [H-1]$ the following holds.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{c_i, s_h^i, a_h^i}\left[\sum_{s' \in S}(\widetilde{P}^{c_i}(s'|s_h^i, a_h^i) - P_\star^{c_i}(s'|s_h^i, a_h^i))^2 \Big| \mathbb{H}_{i-1}\right] \\
=&\mathop{\mathbb{E}}_{c_i, s_h^i}\left[\sum_{s' \in S}(\widetilde{P}^{c_i}(s'|s_h^i, \pi_i(c_i; s_h^i)) - P_\star^{c_i}(s'|s_h^i, \pi_i(c_i; s_h^i)))^2 \Big| \mathbb{H}_{i-1}\right] \\
=&\mathbb{E}_c\left[\mathop{\mathbb{E}}_{s_h}\left[\sum_{s' \in S}(\widetilde{P}^{c}(s'|s_h, \pi_i(c; s_h)) - P_\star^{c}(s'|s_h, \pi_i(c; s_h)))^2 \Big| \mathbb{H}_{i-1}, c\right]\right] \\
=&\mathbb{E}_c\left[\sum_{s_h \in S_h^c} q_h(s_h|\pi_i(c;\cdot), P_\star^c) \sum_{s' \in S}(\widetilde{P}^{c}(s'|s_h, \pi_i(c; s_h)) - P_\star^{c}(s'|s_h, \pi_i(c; s_h)))^2\right],
\end{aligned} \tag{34}$$

where the first equality is because $a_h^i = \pi_i(c_i; s_h^i)$ and π_i is determined completely given $\mathbb{H}_{h-1}$. The second equality is because c_i is independent of $\mathbb{H}_{i-1}$, but s_h^i is dependent on both c_i and $\mathbb{H}_{i-1}$. (The dependency of s_h^i in $\mathbb{H}_{i-1}$ is through the policy π_i). The third equality is an explicit representation of the expectation over s_h given the context c and the history $\mathbb{H}_{i-1}$ since π_i is determined by $\mathbb{H}_{i-1}$.

By summing over $i = 1, 2, \ldots, t-1$ and $h \in [H-1]$ we obtain the claim since,

$$\begin{aligned}
&\sum_{i=1}^{t-1}\sum_{h=1}^{H-1} \mathop{\mathbb{E}}_{c_i, s_h^i, a_h^i}\left[\sum_{s' \in S}(\widetilde{P}^{c_i}(s'|s_h^i, a_h^i) - P_\star^{c_i}(s'|s_h^i, a_h^i))^2 | \mathbb{H}_{i-1}\right] = \\
=&\sum_{i=1}^{t-1}\sum_{h=1}^{H-1} \mathbb{E}_c\left[\sum_{s_h \in S_h^c} q_h(s_h|\pi_i(c;\cdot), P_\star^c) \sum_{s' \in S}(\widetilde{P}^{c}(s'|s_h, \pi_i(c; s_h)) - P_\star^{c}(s'|s_h, \pi_i(c; s_h)))^2\right] \qquad \text{(By equation 34)} \\
=&\mathbb{E}_c\left[\sum_{i=1}^{t-1}\sum_{h=1}^{H-1}\sum_{s_h \in S_h^c} q_h(s_h|\pi_i(c;\cdot), P_\star^c) \sum_{s' \in S}(\widetilde{P}^{c}(s'|s_h, \pi_i(c; s_h)) - P_\star^{c}(s'|s_h, \pi_i(c; s_h)))^2\right], \\
&\qquad \text{(By linearity of expectation)}
\end{aligned}$$

as stated. ■

We now return to the proof of the lemma. By Lemma 57, for any $\delta \in (0,1)$ with probability at least $1 - \delta/4$ it holds that

$$\begin{aligned}
&\sum_{i=1}^{t-1}\mathbb{E}\left[\sum_{h=0}^{H-1}\sum_{s' \in S}(P_t^{c_i}(s'|s_h^i, a_h^i) - P_\star^{c_i}(s'|s_h^i, a_h^i))^2 | \mathbb{H}_{i-1}\right] \\
=&\sum_{i=1}^{t-1}\sum_{h=0}^{H-1} \mathop{\mathbb{E}}_{c_i, s_h^i, a_h^i}\left[\sum_{s' \in S}(P_t^{c_i}(s'|s_h^i, a_h^i) - P_\star^{c_i}(s'|s_h^i, a_h^i))^2 | \mathbb{H}_{i-1}\right] \\
\leq& 72H|S|\log(8|\mathcal{P}|t^3/\delta) + 2\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\sum_{s' \in S}\left(P_t^{c_i}(s'|s_h^i, a_h^i) - \mathbb{I}\Big[s' = s_{h+1}^i\Big]\right)^2 - \left(P_\star^{c_i}(s'|s_h^i, a_h^i) - \mathbb{I}\Big[s' = s_{h+1}^i\Big]\right)^2.
\end{aligned}$$

The above hold simultaneously for all $t \geq |A| + 1$.

Since $\widehat{P}_t$ is the solution for the least square regression, it holds that

$$\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\sum_{s' \in S}(\widehat{P}_t^{c_i}(s'|s_h^i, a_h^i) - \mathbb{I}[s' = s_{h+1}^i])^2 - (P_\star^{c_i}(s'|s_h^i, a_h^i) - \mathbb{I}[s' = s_{h+1}^i])^2 \leq 0.$$

For item (1), using the above and a simplification of the value difference lemma, (see Lemma 71), for every context-dependent policy $\pi \in \Pi_{\mathcal{C}}$ and round $t > |A|$ the following holds.

$$
\begin{aligned}
&\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,P_\star)(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,\widehat{P}_t)(c)}(s_0)] \\
=&\mathbb{E}_c\left[\mathbb{E}_{\pi(c;\cdot),P^c_\star}\left[\sum_{h=0}^{H-1}\sum_{s'\in S}(P^c_\star(s'|s_h,a_h)-\widehat{P}^c_t(s'|s_h,a_h))V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,\widehat{P}_t),h+1}(s')\right]\middle| s_0\right] && \text{(Lemma 71)} \\
\leq&\mathbb{E}_c\left[\mathbb{E}_{\pi(c;\cdot),P^c_\star}\left[\sum_{h=0}^{H-1}\sum_{s'\in S}|P^c_\star(s'|s_h,a_h)-\widehat{P}^c_t(s'|s_h,a_h)|V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,\widehat{P}_t),h+1}(s')\right]\middle| s_0\right] && \text{(By triangle inequality)} \\
\leq&\mathbb{E}_c\left[\mathbb{E}_{\pi(c;\cdot),P^c_\star}\left[\sum_{h=0}^{H-1}\sum_{s'\in S}|P^c_\star(s'|s_h,a_h)-\widehat{P}^c_t(s'|s_h,a_h)|\cdot 2\cdot H\right]\middle| s_0\right] && \text{(By Corollary 53)} \\
\leq&\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S^c_h}\sum_{a_h\in A}q_h(s_h|\pi(c;\cdot),P^c_\star)\pi(a_h|c;s_h)\sum_{s'\in S}|P^c_\star(s'|s_h,a_h)-\widehat{P}^c_t(s'|s_h,a_h)|2H\right] \\
& && \text{(Explicit representation of the expectation using occupancy measure)} \\
=&\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S^c_h}q_h(s_h|\pi(c;\cdot),P^c_\star)\sum_{s'\in S}|P^c_\star(s'|s_h,\pi(c;s_h))-\widehat{P}^c_t(s'|s_h,\pi(c;s_h))|2H\right] \\
& && \text{(Since } \pi \text{ is a deterministic context-dependent policy)} \\
=&\mathbb{E}_c\Big[\sum_{h=0}^{H-1}\sum_{s_h\in S^c_h}(\sqrt{2H}\sqrt{q_h(s_h|\pi(c;\cdot),P^c_\star)})^2 && \text{(By multiplication in } \tfrac{\sqrt{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}}{\sqrt{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}}\text{)} \\
&\cdot\frac{\sqrt{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_j)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}}{\sqrt{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}}\cdot\sum_{s'\in S}|P^c_\star(s'|s_h,\pi(c;s_h))-\widehat{P}^c_t(s'|s_h,\pi(c;s_h))|\Big] \\
=&\sum_{c\in\mathcal{C}}\sum_{h=0}^{H-1}\sum_{s_h\in S^c_h}\sum_{s'\in S}\sqrt{\mathcal{D}(c)}\sqrt{2H}(\sqrt{q_h(s_h|\pi(c;\cdot),P^c_\star)})^2\frac{1}{\sqrt{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}} \\
& && \text{(Re-arranging)} \\
&\cdot\sqrt{\mathcal{D}(c)}\sqrt{2H}\sqrt{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)|P^c_\star(s'|s,\pi(c;s))-\widehat{P}^c_t(s'|s,\pi(c;s))|} \\
\leq&\sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S^c_h}\sum_{s'\in S}\frac{2H\cdot q^2_h(s_h|\pi,P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}\right]} && \text{(By Cauchy-Schwartz)} \\
&\cdot\sqrt{2H\cdot\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S^c_h}\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)\sum_{s'\in S}(P^c_\star(s'|s,\pi(c;s))-\widehat{P}^c_t(s'|s,\pi(c;s)))^2\right]} \\
\leq&\sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S^c_h}\frac{2H\cdot|S|\cdot q_h(s_h|\pi(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}\right]}\sqrt{2H} \\
& && \text{(By } q^2_h(s_h|\pi(c;\cdot),P^c_\star)\leq q_h(s_h|\pi(c;\cdot),P^c_\star)\text{)}
\end{aligned}
$$

$$\cdot \sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{i=1}^{t-1} q_h(s_h|\pi_i(c;\cdot),P_\star^c)\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]\sum_{s'\in S}(\widehat{P}_t^c(s'|s_h,\pi(c;s_h))-P_\star^c(s'|s_h,\pi(c;s_h)))^2\right]}$$

$$\leq \sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{2H\cdot|S|\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]}$$

(The non-zero terms are where $\pi_i(c;s_h)=\pi(c;s_h)$)

$$\cdot \sqrt{2H\cdot\mathbb{E}_c\left[\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c} q_h(s_h|\pi_i(c;\cdot),P_\star^c)\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]\sum_{s'\in S}(\widehat{P}_t^c(s'|s_h,\pi_i(s_h))-P_\star^c(s'|s_h,\pi_i(s_h)))^2\right]}$$

$$\leq \sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{2H\cdot|S|\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]}$$

(Removing the indicators can only increase the sum)

$$\cdot \sqrt{2H\cdot\mathbb{E}_c\left[\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c} q_h(s_h|\pi_i(c;\cdot),P_\star^c)\sum_{s'\in S}(\widehat{P}_t^c(s'|s_h,\pi_i(s_h))-P_\star^c(s'|s_h,\pi_i(s_h)))^2\right]}$$

$$= \sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{2H\cdot|S|\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]}. \qquad \text{(By Claim 2)}$$

$$\cdot \sqrt{2H\sum_{i=1}^{t-1}\sum_{h=1}^{H-1}\mathop{\mathbb{E}}_{c_i,s_h^i,a_h^i}\left[\sum_{s'\in S}(\widehat{P}_t^c(s'|s_h^i,a_h^i)-P_\star^c(s'|s_h^i,a_h^i))^2|\mathbb{H}_{i-1}\right]}$$

$$\leq \sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{2H\cdot|S|\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]}\cdot\sqrt{2\cdot H\cdot 72H|S|\log(8|\mathcal{P}|t^3/\delta)}.$$

(By Lemma 57 combined with the fact that $\widehat{P}_t$ is the least-square minimizer)

The above proves (1).

For (2), by Lemma 73 it holds that,

$$\begin{aligned}
&\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,\widehat{P}_t)(c)}(s_0)]-\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,P_\star)(c)}(s_0)]\\
=&\mathbb{E}_c\left[\mathbb{E}_{\pi(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1}\sum_{s'\in S}(\widehat{P}_t^c(s'|s_h,a_h)-P_\star^c(s'|s_h,a_h))V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,\widehat{P}_t),h+1}(s')\right]\middle|s_0\right] \qquad \text{(Lemma 73)}\\
\leq&\mathbb{E}_c\left[\mathbb{E}_{\pi(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1}\sum_{s'\in S}\left|P_\star^c(s'|s_h,a_h)-\widehat{P}_t^c(s'|s_h,a_h)\right|V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,\widehat{P}_t),h+1}(s')\right]\middle|s_0\right].
\end{aligned}$$

Now, using an identical derivation to that showed above (from the third inequality on), we obtain (2).

Lastly, we remark that by the choice of π_i for all $i\in\{1,2,\ldots,|A|\}$, and the minimum reachability assumption for any deterministic context-dependent policy $\pi\in\Pi_{\mathcal{C}}$, context $c\in\mathcal{C}$, layer $h\in[H-1]$ and state $s_h\in S_h^c$ it holds that

$$\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)\geq p_{min}\cdot\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]\geq p_{min}>0,$$

hence the above is well defined. ■

Step 3: Relax the confidence bound to be additive.

Lemma 59 (the "square trick" relaxation for dynamics approximation) *Under the good event of Lemma 58 for all $t > |A|$ and ant context-dependent policy $\pi \in \Pi_{\mathcal{C}}$ the followings hold for $\gamma_t = \sqrt{\frac{72t\log(8|\mathcal{P}|t^3/\delta)}{|S||A|}}$.*

$$(1)\quad \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,P_\star)(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,\widehat{P}_t)(c)}(s_0)]$$
$$\leq \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{\gamma_t\cdot H\cdot|S|\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right] + \gamma_t\frac{H^2\cdot|S|^2\cdot|A|}{t}.$$

$$(2)\quad \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,\widehat{P}_t)(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,P_\star)(c)}(s_0)]$$
$$\leq \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{\gamma_t\cdot H\cdot|S|\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right] + \gamma_t\frac{H^2\cdot|S|^2\cdot|A|}{t}.$$

Proof For item (1), consider the following derivation, where $\gamma_t = \sqrt{\frac{72t\log(8|\mathcal{P}|t^3/\delta)}{|S||A|}}$.

$$\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,P_\star)(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{r}_t,\widehat{P}_t)(c)}(s_0)]$$
$$\leq \sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{2\cdot H\cdot|S|\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]}\cdot\sqrt{2\cdot 72\cdot H^2\cdot|S|\cdot\log(8|\mathcal{P}|t^3/\delta)}$$
(Lemma 58 part (1))
$$= \sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{\gamma_t\cdot 2\cdot H\cdot|S|\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]}\cdot\sqrt{\frac{1}{\gamma_t}2\cdot H^2\cdot|S|\cdot 72\log(8|\mathcal{P}|t^3/\delta)}$$
$$= \sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{\gamma_t\cdot 2\cdot H\cdot|S|\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]}\cdot\sqrt{\sqrt{\frac{|S||A|}{t}}\frac{2\cdot H^2\cdot|S|\cdot 72\log(8|\mathcal{P}|t^3/\delta)}{\sqrt{72\cdot\log(8|\mathcal{P}|t^3/\delta)}}}$$
$$= \sqrt{\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{\gamma_t\cdot 2\cdot H\cdot|S|\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]}\cdot\sqrt{2\cdot H^2\cdot|S|\cdot\sqrt{\frac{|S||A|}{t}}\sqrt{72\log(8|\mathcal{P}|t^3/\delta)}}$$
$$= \sqrt{2\cdot\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{\gamma_t\cdot H\cdot|S|\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]}$$
$$\cdot\sqrt{2\cdot H^2\cdot|S|\cdot\sqrt{\frac{|S||A|}{t}}\sqrt{\frac{t^2}{|S|^2|A|^2}}\sqrt{\frac{|S|^2|A|^2}{t^2}}\sqrt{72\log(8|\mathcal{P}|t^3/\delta)}}$$
$$= \sqrt{2\cdot\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{\gamma_t\cdot H\cdot|S|\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]}\cdot\sqrt{2\cdot H^2\cdot|S|\cdot\gamma_t\sqrt{\frac{|S|^2|A|^2}{t^2}}}$$

$$=2\sqrt{\cdot\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{\gamma_t\cdot H\cdot|S|\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]\cdot\sqrt{\gamma_t\frac{H^2\cdot|S|^2\cdot|A|}{t}}}$$

$$\leq\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{\gamma_t\cdot H\cdot|S|\cdot q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]+\gamma_t\frac{H^2\cdot|S|^2\cdot|A|}{t}.$$

(Since $2ab\leq a^2+b^2$ for all $a,b\geq 0$)

The above proves (1). We obtain (2) using an identical derivation, where in the first inequality we use item (2) of Lemma 58 (instead of item (1)). ■

Step 4: Bounding the contextual potential for both dynamics and rewards. As in previous sections, we consider $\phi_t(\pi)$ and $\psi_t(\pi)$, the two contextual potential functions in round t, for $T\geq t>|A|$ and a context-depended policy $\pi\in\Pi_{\mathcal{C}}$.

Recall their definitions.

$$\phi_t(\pi):=\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right],$$

and

$$\psi_t(\pi):=\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]p_{min}}\right],$$

where $\{\pi_t\in\Pi_{\mathcal{C}}\}_{t=1}^T$ is the sequence of policies selected by Algorithm 8.

In the following lemma, we bound the sum of contextual potential functions, over the rounds $t=|A|+1,\ldots,T$.

Lemma 60 (contextual potential lemma for dynamics approximation) *Let $\{\pi_t\in\Pi_{\mathcal{C}}\}_{t=1}^T$ be the sequence of policies selected by Algorithm 8. Then, for all $T>|A|$ the followings hold.*

1. *For any policy $\pi\in\Pi_{\mathcal{C}}$ it holds that*

$$\sum_{t=|A|+1}^{T}\phi_t(\pi)=\sum_{t=|A|+1}^{T}\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]$$
$$\leq\frac{H|A|}{p_{min}}(1+\log(T/|A|)).$$

2. *For the sequence $\{\pi_t\in\Pi_{\mathcal{C}}\}_{t=1}^T$ it holds that*

$$\sum_{t=|A|+1}^{T}\phi_t(\pi_t)=\sum_{t=|A|+1}^{T}\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi_t(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi_t(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]$$
$$\leq\frac{|S||A|}{p_{min}}(1+\log(T/|A|)).$$

3. *For any policy $\pi\in\Pi_{\mathcal{C}}$ it holds that*

$$\sum_{t=|A|+1}^{T}\psi_t(\pi)=\sum_{t=|A|+1}^{T}\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi(c;s_h)=\pi_i(c;s_h)]p_{min}}\right]$$
$$\leq\frac{H|A|}{p_{min}}(1+\log(T/|A|)).$$

4. *For the sequence* $\{\pi_t \in \Pi_{\mathcal{C}}\}_{t=1}^{T}$ *it holds that*

$$\sum_{t=|A|+1}^{T} \psi_t(\pi_t) = \sum_{t=|A|+1}^{T} \mathbb{E}_c \left[\sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \frac{q_h(s_h|\pi_t(c;\cdot), P_\star^c)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi_t(c;s_h) = \pi_i(c;s_h)] p_{min}} \right] \leq \frac{|S||A|}{p_{min}}(1 + \log(T/|A|))$$

Proof The proofs of parts 1 and 2 are presented in Lemma 21.

The proof of 3 and 4 are also similar. For 3, consider the following.

For a fixed context $c \in \mathcal{C}$, given any policy $\pi \in \Pi_{\mathcal{C}}$ it holds that,

$$\begin{aligned}
&\sum_{t=|A|+1}^{T} \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \frac{1}{p_{min}} \frac{q_h(s_h|\pi(c;\cdot), P_\star^c)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c;s_h) = \pi_i(c;s_h)]} \\
=&\frac{1}{p_{min}} \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} q_h(s_h|\pi(c;\cdot), P_\star^c) \sum_{t=|A|+1}^{T} \frac{1}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c;s_h) = \pi_i(c;s_h)]} \\
\leq&\frac{1}{p_{min}} \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} q_h(s_h|\pi(c;\cdot), P_\star^c) \sum_{a_h \in A} \sum_{i=1}^{\sum_{t=1}^{T} \mathbb{I}[\pi_t(c;s_h)=a_h]} \frac{1}{i} \\
\leq&\frac{1}{p_{min}} \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} q_h(s_h|\pi(c;\cdot), P_\star^c) \sum_{a_h \in A} \left(1 + \log\left(\sum_{t=1}^{T} \mathbb{I}[\pi_t(c;s_h) = a_h]\right)\right) \\
\leq&\frac{H|A|}{p_{min}}(1 + \log(T/|A|)),
\end{aligned}$$

where the last inequality is due to Jensen's inequality. By taking expectation over c on both sides of the above inequality, we obtain part 3 of the lemma.

For part 4 we similarly have

$$\begin{aligned}
&\sum_{t=|A|+1}^{T} \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \frac{1}{p_{min}} \frac{q_h(s_h|\pi_t(c;\cdot), P_\star)}{\sum_{i=1}^{t-1} \mathbb{I}[\pi(c;s_h) = \pi_i(c;s_h)]} \\
\leq&\frac{1}{p_{min}} \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \sum_{t=|A|+1}^{T} \frac{1}{\sum_{i=1}^{t-1} \mathbb{I}[\pi_t(c;s_h) = \pi_i(c;s_h)]} \\
\leq&\frac{1}{p_{min}} \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \sum_{a_h \in A} \sum_{i=1}^{\sum_{t=1}^{T} \mathbb{I}[\pi_t(c;s_h)=a_h]} \frac{1}{i} \\
\leq&\frac{1}{p_{min}} \sum_{h=0}^{H-1} \sum_{s_h \in S_h^c} \sum_{a_h \in A} \left(1 + \log\left(\sum_{t=1}^{T} \mathbb{I}[\pi_t(c;s_h) = a_h]\right)\right) \\
\leq&\frac{|S||A|}{p_{min}}(1 + \log(T/|A|)),
\end{aligned}$$

where the last inequality is due to Jensen's inequality. By taking expectation over c on both sides of the above inequality, we obtain part 4 of the lemma. ■

D.3.4 REGRET BOUND

Lemma 61 (optimism) *Under the good events of Lemmas 50 and 58 for any $t \geq |A|+1$ it holds that*

$$\begin{aligned}
&\mathbb{E}_c[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)]\\
\leq &\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h \in S^c_h} \frac{\gamma_t \cdot H \cdot |S| \cdot q_h(s_h|\pi^\star(c;\cdot), P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}\right] + \gamma_t\frac{H^2|S|^2|A|}{t}\\
&+\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h \in S^c_h}\frac{1}{p_{min}}\frac{\beta_t \cdot q_h(s_h|\pi^\star(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]}\right] + \beta_t\frac{H|S||A|}{t},
\end{aligned}$$

where $\beta_t = \sqrt{\frac{17t\log(8|\mathcal{F}|t^3/\delta)}{|S||A|}}$ *and* $\gamma_t = \sqrt{\frac{72t\log(8|\mathcal{P}|t^3/\delta)}{|S||A|}}$.

Proof Assume the good event holds and consider the following derivation.

$$\begin{aligned}
&\mathbb{E}_c[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0)]\\
\leq &\mathbb{E}_c[V^{\pi^\star(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t,P_\star)}(c)}(s_0)] + \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h \in S^c_h}\frac{1}{p_{min}}\frac{\beta_t \cdot q_h(s_h|\pi^\star(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]}\right] + \beta_t\frac{H|S||A|}{t}\\
&\text{(By Lemma 51 equation (26))}\\
\leq &\mathbb{E}_c[V^{\pi^\star(c;\cdot)}_{\mathcal{M}^{(\hat{r}_t,\hat{P}_t)}(c)}(s_0)] + \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h \in S^c_h} \frac{\gamma_t \cdot H \cdot |S| \cdot q_h(s_h|\pi^\star(c;\cdot), P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}\right] + \gamma_t\frac{H^2|S|^2|A|}{t}\\
&+\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h \in S^c_h}\frac{1}{p_{min}}\frac{\beta_t \cdot q_h(s_h|\pi^\star(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]}\right] + \beta_t\frac{H|S||A|}{t} \qquad \text{(By Lemma 59 part (1))}\\
= &\mathbb{E}_c[V^{\pi^\star(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)] + \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h \in S^c_h} \frac{\gamma_t \cdot H \cdot |S| \cdot q_h(s_h|\pi^\star(c;\cdot), P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}\right]\\
&+\gamma_t\frac{H^2|S|^2|A|}{t} + \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h \in S^c_h}\frac{1}{p_{min}}\frac{\beta_t \cdot q_h(s_h|\pi^\star(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]}\right] + \beta_t\frac{H|S||A|}{t} \qquad \text{(By } \widehat{\mathcal{M}}_t(c) \text{ definition)}\\
\leq &\mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)] + \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h \in S^c_h} \frac{\gamma_t \cdot H \cdot |S| \cdot q_h(s_h|\pi^\star(c;\cdot), P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}\right]\\
&+\gamma_t\frac{H^2|S|^2|A|}{t} + \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h \in S^c_h}\frac{1}{p_{min}}\frac{\beta_t \cdot q_h(s_h|\pi^\star(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]}\right] + \beta_t\frac{H|S||A|}{t},\\
&(\pi_t(c;\cdot) \text{ is the optimal policy of } \widehat{\mathcal{M}}_t(c).)
\end{aligned}$$

■

Theorem 62 (expected regret bound) *For any $T \geq 1$, finite functions classes $\mathcal{F}$ and $\mathcal{P}$ and $\delta \in (0,1)$, with probability at least $1-\delta$ it holds that*

$$\textit{E.Regret}_T(RM-UCDD) \leq \tilde{O}\left(\max\{H, 1/p_{min}\}\cdot\left(H|S|^{3/2}\sqrt{|A|T\log\frac{|\mathcal{P}|}{\delta}} + \sqrt{T|S||A|\log\frac{|\mathcal{F}|}{\delta}}\right) + |A|H\right),$$

for $\beta_t = \sqrt{\frac{17t\log(8|\mathcal{F}|t^3/\delta)}{|S||A|}}$ *and* $\gamma_t = \sqrt{\frac{72t\log(8|\mathcal{P}|t^3/\delta)}{|S||A|}}$, *for all* $t \in \{1,2,\ldots,T\}$.

Proof We derive an expected regret bound under the good events of Lemmas 50 and 58 that hold with probability at least $1-\delta/2$.

Assume the good events hold. By Azuma's inequality, with probability at least $1-\delta/2$ over the policies π_t, the following holds.

$$\begin{aligned}
&\text{E.Regret}_T(RM-UCDD) = \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_c[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)]\right] \\
&= \sum_{t=1}^{T}\mathbb{E}_{\mathbb{H}_{t-1}}\left[\mathbb{E}_c[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0) - V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)|\mathbb{H}_{t-1}]\right] \\
&\le \sum_{t=1}^{T}\mathbb{E}_c[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)] + 2H\sqrt{2T\log(4/\delta)} && \text{(By Azuma's inequality)} \\
&\le \sum_{t=|A|+1}^{T}\mathbb{E}_c[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)] + 2H\sqrt{2T\log(4/\delta)} + |A|H \\
&\le \sum_{t=|A|+1}^{T}\mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)] \\
&\quad + \sum_{t=|A|+1}^{T}\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{\gamma_t\cdot H\cdot|S|\cdot q_h(s_h|\pi^\star(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}\right] + \sum_{t=|A|+1}^{T}\gamma_t\frac{H^2|S|^2|A|}{t} \\
&\quad + \sum_{t=|A|+1}^{T}\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{1}{p_{min}}\frac{\beta_t\cdot q_h(s_h|\pi^\star(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]}\right] + \sum_{t=|A|+1}^{T}\beta_t\frac{H|S||A|}{t} \\
&\quad + 2H\sqrt{2T\log(4/\delta)} + |A|H && \text{(By the optimism lemma (Lemma 61))} \\
&= \sum_{t=|A|+1}^{T}\mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}(s_0)] + \sum_{t=|A|+1}^{T}\mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)] \\
&\qquad\qquad \text{(By adding and subtracting } \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}(s_0)]) \\
&\quad + H\cdot|S|\cdot\gamma_T\sum_{t=|A|+1}^{T}\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi^\star(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P^c_\star)}\right] + \sum_{t=|A|+1}^{T}\gamma_t\frac{H^2|S|^2|A|}{t} \\
&\qquad\qquad \text{(Since for all } t\in\{1,2,\ldots,T\},\ \gamma_T\ge\gamma_t) \\
&\quad + \sum_{t=|A|+1}^{T}\beta_T\cdot\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{1}{p_{min}}\frac{q_h(s_h|\pi^\star(c;\cdot),P^c_\star)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]}\right] + \sum_{t=|A|+1}^{T}\beta_t\frac{H|S||A|}{t} \\
&\qquad\qquad \text{(Since for all } t\in\{1,2,\ldots,T\},\ \beta_T\ge\beta_t) \\
&\quad + 2H\sqrt{2T\log(4/\delta)} + |A|H \\
&= \sum_{t=|A|+1}^{T}\mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t,\widehat{P}_t)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}(s_0)] && \text{(By } \widehat{\mathcal{M}}_t(c) \text{ definition)} \\
&\quad + \sum_{t=|A|+1}^{T}\mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)]
\end{aligned}$$

$$
\begin{aligned}
&+ H \cdot |S| \cdot \gamma_T \cdot \frac{H|A|}{p_{min}}(1+\log(T/|A|)) && \text{(By Lemma 60 part 1)}\\
&+ \sum_{t=|A|+1}^{T} \gamma_t \frac{H^2|S|^2|A|}{t}\\
&+ \beta_T \cdot \frac{H|A|}{p_{min}}(1+\log(T/|A|)) && \text{(By Lemma 60 part 1)}\\
&+ \sum_{t=|A|+1}^{T} \beta_t \frac{H|S||A|}{t} + 2H\sqrt{2T\log(4/\delta)} + |A|H\\
\leq & \sum_{t=|A|+1}^{T} \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c} \frac{\gamma_t \cdot H\cdot|S|\cdot q_h(s_h|\pi_t(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi_t(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right] + \sum_{t=|A|+1}^{T}\gamma_t\frac{H^2|S|^2|A|}{t}\\
& && \text{(By Lemma 59, part (2))}\\
&+ 2\sum_{t=|A|+1}^{T} \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c} \frac{1}{p_{min}}\frac{\beta_t \cdot q_h(s_h|\pi_t(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi_t(c;s_h)=\pi_i(c;s_h)]}\right] + \sum_{t=|A|+1}^{T}\beta_t\frac{H|S||A|}{t}\\
& && \text{(By Lemma 51, equation (27))}\\
&+ H \cdot |S| \cdot \gamma_T \cdot \frac{H|A|}{p_{min}}(1+\log(T/|A|))\\
&+ \sum_{t=|A|+1}^{T} \gamma_t \frac{H^2|S|^2|A|}{t}\\
&+ \beta_T \cdot \frac{H|A|}{p_{min}}(1+\log(T/|A|))\\
&+ \sum_{t=|A|+1}^{T} \beta_t \frac{H|S||A|}{t} + 2H\sqrt{2T\log(4/\delta)} + |A|H\\
\leq & \gamma_T \cdot H\cdot|S|\cdot\sum_{t=|A|+1}^{T} \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c} \frac{q_h(s_h|\pi_t(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi_t(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]\\
& && \text{(Since } \gamma_T \geq \gamma_t\text{, for all } t\in\{1,2,\ldots,T\}\text{)}\\
&+ 2\beta_T\sum_{t=|A|+1}^{T} \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c} \frac{1}{p_{min}}\frac{q_h(s_h|\pi_t(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi_t(c;s_h)=\pi_i(c;s_h)]}\right] && \text{(Since } \beta_T \geq \beta_t\text{, for all } t\in\{1,2,\ldots,T\}\text{)}\\
&+ H \cdot |S| \cdot \gamma_T \cdot \frac{H|A|}{p_{min}}(1+\log(T/|A|))\\
&+ 2\sum_{t=|A|+1}^{T} \gamma_t \frac{H^2|S|^2|A|}{t}\\
&+ \beta_T \cdot \frac{H|A|}{p_{min}}(1+\log(T/|A|))\\
&+ 2\sum_{t=|A|+1}^{T} \beta_t \frac{H|S||A|}{t} + 2H\sqrt{2T\log(4/\delta)} + |A|H\\
\leq & \gamma_T \cdot H\cdot|S|\cdot\frac{|S||A|}{p_{min}}(1+\log(T/|A|)) && \text{(By Lemma 60, part 2)}
\end{aligned}
$$

$$
\begin{aligned}
&+ 2\beta_T \frac{|S||A|}{p_{min}}(1+\log(T/|A|)) && \text{(By Lemma 60, part 4)}\\
&+ H\cdot|S|\cdot\gamma_T\cdot\frac{H|A|}{p_{min}}(1+\log(T/|A|))\\
&+ 2\sum_{t=|A|+1}^{T}\gamma_t\frac{H^2|S|^2|A|}{t}\\
&+ \beta_T\cdot\frac{H|A|}{p_{min}}(1+\log(T/|A|))\\
&+ 2\sum_{t=|A|+1}^{T}\beta_t\frac{H|S||A|}{t} + 2H\sqrt{2T\log(4/\delta)} + |A|H\\
\leq&\gamma_T\cdot H\cdot|S|\cdot\frac{|S||A|}{p_{min}}(1+\log(T/|A|))\\
&+ 2\beta_T\frac{|S||A|}{p_{min}}(1+\log(T/|A|))\\
&+ H\cdot|S|\cdot\gamma_T\cdot\frac{H|A|}{p_{min}}(1+\log(T/|A|))\\
&+ 2H^2\cdot|S|^{3/2}\cdot\sqrt{|A|}\sqrt{72\log(8|\mathcal{P}|T^3/\delta)}\sum_{t=|A|+1}^{T}\frac{1}{\sqrt{t}} && \text{(By } \gamma_t \text{ choice)}\\
&+ \beta_T\cdot\frac{H|A|}{p_{min}}(1+\log(T/|A|))\\
&+ 2H\cdot\sqrt{|S|\cdot|A|}\sqrt{17\log(8|\mathcal{F}|T^3/\delta)}\sum_{t=|A|+1}^{T}\frac{1}{\sqrt{t}} && \text{(By } \beta_t \text{ choice)}\\
&+ 2H\sqrt{2T\log(4/\delta)} + |A|H\\
\leq&\gamma_T\cdot H\cdot|S|\cdot\frac{|S||A|}{p_{min}}(1+\log(T/|A|))\\
&+ 2\beta_T\frac{|S||A|}{p_{min}}(1+\log(T/|A|))\\
&+ H\cdot|S|\cdot\gamma_T\cdot\frac{H|A|}{p_{min}}(1+\log(T/|A|))\\
&+ 2H^2\cdot|S|^{3/2}\cdot\sqrt{|A|T\cdot 72\log(8|\mathcal{P}|T^3/\delta)}\\
&+ \beta_T\cdot\frac{H|A|}{p_{min}}(1+\log(T/|A|))\\
&+ 2H\cdot\sqrt{|S||A|T}\cdot\sqrt{17\log(8|\mathcal{F}|T^3/\delta)}\\
&+ 2H\sqrt{2T\log(4/\delta)} + |A|H\\
\leq&\frac{H|S|^{3/2}\sqrt{T\cdot|A|72\log(8|\mathcal{P}|T^3/\delta)}(1+\log(T/|A|))}{p_{min}} && \text{(By } \gamma_T \text{ choice)}\\
&+ \frac{2\sqrt{T|S||A|17\log(8|\mathcal{F}|T^3/\delta)}(1+\log(T/|A|))}{p_{min}} && \text{(By } \beta_T \text{ choice)}\\
&+ \frac{H^2\cdot\sqrt{T|A||S|72\log(8|\mathcal{P}|T^3/\delta)}\cdot(1+\log(T/|A|))}{p_{min}} && \text{(By } \gamma_T \text{ choice)}\\
&+ 2H^2\cdot|S|^{3/2}\cdot\sqrt{T\cdot|A|}\sqrt{72\log(8|\mathcal{P}|T^3/\delta)}
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{H\sqrt{T|A|17\log(8|\mathcal{F}|T^3/\delta)}\cdot(1+\log(T/|A|))}{p_{min}\cdot\sqrt{|S|}} && \text{(By } \beta_T \text{ choice)}\\
&+2H\cdot\sqrt{T\cdot|S|\cdot|A|}\sqrt{17\log(8|\mathcal{F}|T^3/\delta)}\\
&+2H\sqrt{2T\log(4/\delta)}+|A|H\\
\leq& 2\frac{H|S|^{3/2}\sqrt{T\cdot|A|72\log(8|\mathcal{P}|T^3/\delta)}(1+\log(T/|A|))}{p_{min}} && \text{(Since } H\leq|S|\text{)}\\
&+3\frac{\sqrt{T|S||A|17\log(8|\mathcal{F}|T^3/\delta)}(1+\log(T/|A|))}{p_{min}}\\
&+2H^2\cdot|S|^{3/2}\cdot\sqrt{T\cdot|A|}\sqrt{72\log(8|\mathcal{P}|T^3/\delta)}\\
&+2H\cdot\sqrt{T\cdot|S|\cdot|A|}\sqrt{17\log(8|\mathcal{F}|T^3/\delta)}\\
&+2H\sqrt{2T\log(4/\delta)}+|A|H\\
=&\tilde{O}\left(\max\{H,1/p_{min}\}\cdot\left(H|S|^{3/2}\sqrt{T|A|\log\frac{|\mathcal{P}|}{\delta}}+\sqrt{T|S||A|\log\frac{|\mathcal{F}|}{\delta}}\right)+|A|H\right\}\right).
\end{aligned}
$$

Since both good events hold with probability at least $1-\delta/2$, by union bound combined with Azuma's inequality we obtain the theorem. ■

Recall the regret, which defined as $\text{Regret}_T(\text{ALG}) := \sum_{t=1}^{T} V_{\mathcal{M}(c_t)}^{\pi^\star(c_t;\cdot)} - V_{\mathcal{M}(c_t)}^{\pi_t(c_t;\cdot)}$.

Theorem 63 (regret bound) *For any $T\geq 1$, finite functions classes $\mathcal{F}$ and $\mathcal{P}$ and $\delta\in(0,1)$, with probability at least $1-\delta$ it holds that*

$$
\textit{Regret}_T(RM-UCDD)\leq\tilde{O}\left(\max\{H,1/p_{min}\}\cdot\left(H|S|^{3/2}\sqrt{|A|T\log\frac{|\mathcal{P}|}{\delta}}+\sqrt{T|S||A|\log\frac{|\mathcal{F}|}{\delta}}\right)+|A|H\right),
$$

where $\beta_t=\sqrt{\frac{17\log(8|\mathcal{F}|t^3/\delta)t}{|S||A|}}$ and $\gamma_t=\sqrt{\frac{72t\log(8|\mathcal{P}|t^3/\delta)}{|S||A|}}$, for all $t\in[T]$.

Proof We derive a regret bound under the good events of Lemmas 50 and 58. Both events hold with probability at least $1-\delta/2$.

Consider the martingale difference sequence $\{Y_t\}_{t=1}^T$ and the filtration $\{\mathbb{H}_t\}_{t=1}^T$ where

$$
Y_t := V_{\mathcal{M}(c_t)}^{\pi^\star(c_t,\cdot)}(s_0)-V_{\mathcal{M}(c_t)}^{\pi_t(c_t,\cdot)}(s_0)-\mathbb{E}_{c_t}\left[V_{\mathcal{M}(c_t)}^{\pi^\star(c_t,\cdot)}(s_0)-V_{\mathcal{M}(c_t)}^{\pi_t(c_t,\cdot)}(s_0)\middle|\mathbb{H}_{t-1}\right].
$$

Clearly, for all t, $|Y_t|\leq 2H$, Y_t is determined completely by the histories $\mathbb{H}_1,\ldots,\mathbb{H}_t$ and $\mathbb{E}_{c_t}[Y_t|\mathbb{H}_{t-1}]=0$. Hence, by Azuma's inequality, with probability at least $1-\delta/2$ it holds that

$$
\begin{aligned}
&\text{Regret}_T(RM-UCDD)=\sum_{t=1}^{T}V_{\mathcal{M}(c_t)}^{\pi^\star(c_t;\cdot)}(s_0)-V_{\mathcal{M}(c_t)}^{\pi_t(c_t;\cdot)}(s_0)\\
\leq&\sum_{t=1}^{T}\mathbb{E}_{c_t}[V_{\mathcal{M}(c_t)}^{\pi^\star(c_t;\cdot)}(s_0)|\mathbb{H}_{t-1}]-\mathbb{E}_{c_t}[V_{\mathcal{M}(c_t)}^{\pi_t(c_t;\cdot)}(s_0)|\mathbb{H}_{t-1}]+2H\sqrt{2T\log(4/\delta)} && \text{(By Azuma's inequality)}\\
=&\sum_{t=1}^{T}\mathbb{E}_{c}[V_{\mathcal{M}(c)}^{\pi^\star(c;\cdot)}(s_0)]-\mathbb{E}_{c}[V_{\mathcal{M}(c)}^{\pi_t(c;\cdot)}(s_0)]+2H\sqrt{2T\log(4/\delta)}
\end{aligned}
$$

(Since π_t is determined by $\mathbb{H}_{t-1}$, and c_t,$\pi^\star$ are independent of the history, we can omit the conditioning on $\mathbb{H}_{t-1}$)

$$\leq \sum_{t=|A|+1}^{T} \mathbb{E}_c[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)] + 2H\sqrt{2T\log(4/\delta)} + |A|H$$

$$\leq \sum_{t=|A|+1}^{T} \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)]$$

$$+ \sum_{t=|A|+1}^{T} \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c} \frac{\gamma_t \cdot H\cdot|S|\cdot q_h(s_h|\pi^\star(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right] + \sum_{t=|A|+1}^{T}\gamma_t\frac{H^2|S|^2|A|}{t}$$

$$+ \sum_{t=|A|+1}^{T} \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c} \frac{1}{p_{min}}\frac{\beta_t \cdot q_h(s_h|\pi^\star(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]}\right] + \sum_{t=|A|+1}^{T}\beta_t\frac{H|S||A|}{t}$$

$$+ 2H\sqrt{2T\log(4/\delta)} + |A|H \qquad \text{(By the optimism lemma (Lemma 61))}$$

$$= \sum_{t=|A|+1}^{T} \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}(s_0)] + \sum_{t=|A|+1}^{T} \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)]$$

(By adding and subtracting $\mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}(s_0)]$)

$$+ H\cdot|S|\cdot\gamma_T \sum_{t=|A|+1}^{T} \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c} \frac{q_h(s_h|\pi^\star(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right] + \sum_{t=|A|+1}^{T}\gamma_t\frac{H^2|S|^2|A|}{t}$$

(Since for all $t\in\{1,2,\ldots,T\}$, $\gamma_T\geq\gamma_t$)

$$+ \sum_{t=|A|+1}^{T} \beta_T\cdot\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c} \frac{1}{p_{min}}\frac{q_h(s_h|\pi^\star(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi^\star(c;s_h)=\pi_i(c;s_h)]}\right] + \sum_{t=|A|+1}^{T}\beta_t\frac{H|S||A|}{t}$$

(Since for all $t\in\{1,2,\ldots,T\}$, $\beta_T\geq\beta_t$)

$$+ 2H\sqrt{2T\log(4/\delta)} + |A|H$$

$$= \sum_{t=|A|+1}^{T} \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t,\widehat{P}_t)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}(s_0)] \qquad \text{(By } \widehat{\mathcal{M}}_t(c) \text{ definition)}$$

$$+ \sum_{t=|A|+1}^{T} \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}^{(\widehat{r}_t,P_\star)}(c)}(s_0)] - \mathbb{E}_c[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)]$$

$$+ H\cdot|S|\cdot\gamma_T\cdot\frac{H|A|}{p_{min}}(1+\log(T/|A|)) \qquad \text{(By Lemma 60 part 1)}$$

$$+ \sum_{t=|A|+1}^{T}\gamma_t\frac{H^2|S|^2|A|}{t}$$

$$+ \beta_T\cdot\frac{H|A|}{p_{min}}(1+\log(T/|A|)) \qquad \text{(By Lemma 60 part 3)}$$

$$+ \sum_{t=|A|+1}^{T}\beta_t\frac{H|S||A|}{t} + 2H\sqrt{2T\log(4/\delta)} + |A|H$$

$$\leq \sum_{t=|A|+1}^{T} \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c} \frac{\gamma_t \cdot H\cdot|S|\cdot q_h(s_h|\pi_t(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi_t(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right] + \sum_{t=|A|+1}^{T}\gamma_t\frac{H^2|S|^2|A|}{t}$$

(By Lemma 59, part (2))

$$+ 2\sum_{t=|A|+1}^{T}\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{1}{p_{min}}\frac{\beta_t\cdot q_h(s_h|\pi_t(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi_t(c;s_h)=\pi_i(c;s_h)]}\right]+\sum_{t=|A|+1}^{T}\beta_t\frac{H|S||A|}{t}$$

(By Lemma 51, equation (27))

$$+ H\cdot|S|\cdot\gamma_T\cdot\frac{H|A|}{p_{min}}(1+\log(T/|A|))$$
$$+\sum_{t=|A|+1}^{T}\gamma_t\frac{H^2|S|^2|A|}{t}$$
$$+\beta_T\cdot\frac{H|A|}{p_{min}}(1+\log(T/|A|))$$
$$+\sum_{t=|A|+1}^{T}\beta_t\frac{H|S||A|}{t}+2H\sqrt{2T\log(4/\delta)}+|A|H$$

$$\leq\gamma_T\cdot H\cdot|S|\cdot\sum_{t=|A|+1}^{T}\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{q_h(s_h|\pi_t(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi_t(c;s_h)=\pi_i(c;s_h)]q_h(s_h|\pi_i(c;\cdot),P_\star^c)}\right]$$

(Since $\gamma_T\geq\gamma_t$, for all $t\in\{1,2,\ldots,T\}$)

$$+ 2\beta_T\sum_{t=|A|+1}^{T}\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\frac{1}{p_{min}}\frac{q_h(s_h|\pi_t(c;\cdot),P_\star^c)}{\sum_{i=1}^{t-1}\mathbb{I}[\pi_t(c;s_h)=\pi_i(c;s_h)]}\right]$$ (Since $\beta_T\geq\beta_t$, for all $t\in\{1,2,\ldots,T\}$)

$$+ H\cdot|S|\cdot\gamma_T\cdot\frac{H|A|}{p_{min}}(1+\log(T/|A|))$$
$$+ 2\sum_{t=|A|+1}^{T}\gamma_t\frac{H^2|S|^2|A|}{t}$$
$$+\beta_T\cdot\frac{H|A|}{p_{min}}(1+\log(T/|A|))$$
$$+ 2\sum_{t=|A|+1}^{T}\beta_t\frac{H|S||A|}{t}+2H\sqrt{2T\log(4/\delta)}+|A|H$$

$$\leq\gamma_T\cdot H\cdot|S|\cdot\frac{|S||A|}{p_{min}}(1+\log(T/|A|))$$ (By Lemma 60, part 2)

$$+ 2\beta_T\frac{|S||A|}{p_{min}}(1+\log(T/|A|))$$ (By Lemma 60, part 4)

$$+ H\cdot|S|\cdot\gamma_T\cdot\frac{H|A|}{p_{min}}(1+\log(T/|A|))$$
$$+ 2\sum_{t=|A|+1}^{T}\gamma_t\frac{H^2|S|^2|A|}{t}$$
$$+\beta_T\cdot\frac{H|A|}{p_{min}}(1+\log(T/|A|))$$
$$+ 2\sum_{t=|A|+1}^{T}\beta_t\frac{H|S||A|}{t}+2H\sqrt{2T\log(4/\delta)}+|A|H$$

$$\leq\gamma_T\cdot H\cdot|S|\cdot\frac{|S||A|}{p_{min}}(1+\log(T/|A|))$$

$$
\begin{aligned}
&+ 2\beta_T \frac{|S||A|}{p_{min}}(1+\log(T/|A|)) \\
&+ H\cdot|S|\cdot\gamma_T\cdot\frac{H|A|}{p_{min}}(1+\log(T/|A|)) \\
&+ 2H^2\cdot|S|^{3/2}\cdot\sqrt{|A|}\sqrt{72\log(8|\mathcal{P}|T^3/\delta)}\sum_{t=|A|+1}^{T}\frac{1}{\sqrt{t}} && \text{(By } \gamma_t \text{ choice)} \\
&+ \beta_T\cdot\frac{H|A|}{p_{min}}(1+\log(T/|A|)) \\
&+ 2H\cdot\sqrt{|S|\cdot|A|}\sqrt{17\log(8|\mathcal{F}|T^3/\delta)}\sum_{t=|A|+1}^{T}\frac{1}{\sqrt{t}} && \text{(By } \beta_t \text{ choice)} \\
&+ 2H\sqrt{2T\log(4/\delta)}+|A|H \\
\leq&\gamma_T\cdot H\cdot|S|\cdot\frac{|S||A|}{p_{min}}(1+\log(T/|A|)) \\
&+ 2\beta_T\frac{|S||A|}{p_{min}}(1+\log(T/|A|)) \\
&+ H\cdot|S|\cdot\gamma_T\cdot\frac{H|A|}{p_{min}}(1+\log(T/|A|)) \\
&+ 2H^2\cdot|S|^{3/2}\cdot\sqrt{|A|T\cdot 72\log(8|\mathcal{P}|T^3/\delta)} \\
&+ \beta_T\cdot\frac{H|A|}{p_{min}}(1+\log(T/|A|)) \\
&+ 2H\cdot\sqrt{|S||A|T}\cdot\sqrt{17\log(8|\mathcal{F}|T^3/\delta)} \\
&+ 2H\sqrt{2T\log(4/\delta)}+|A|H \\
\leq&\frac{H|S|^{3/2}\sqrt{T\cdot|A|72\log(8|\mathcal{P}|T^3/\delta)}(1+\log(T/|A|))}{p_{min}} && \text{(By } \gamma_T \text{ choice)} \\
&+ \frac{2\sqrt{T|S||A|17\log(8|\mathcal{F}|T^3/\delta)}(1+\log(T/|A|))}{p_{min}} && \text{(By } \beta_T \text{ choice)} \\
&+ \frac{H^2\cdot\sqrt{T|A||S|72\log(8|\mathcal{P}|T^3/\delta)}\cdot(1+\log(T/|A|))}{p_{min}} && \text{(By } \gamma_T \text{ choice)} \\
&+ 2H^2\cdot|S|^{3/2}\cdot\sqrt{T\cdot|A|}\sqrt{72\log(8|\mathcal{P}|T^3/\delta)} \\
&+ \frac{H\sqrt{T|A|17\log(8|\mathcal{F}|T^3/\delta)}\cdot(1+\log(T/|A|))}{p_{min}\cdot\sqrt{|S|}} && \text{(By } \beta_T \text{ choice)} \\
&+ 2H\cdot\sqrt{T\cdot|S|\cdot|A|}\sqrt{17\log(8|\mathcal{F}|T^3/\delta)} \\
&+ 2H\sqrt{2T\log(4/\delta)}+|A|H \\
\leq&2\frac{H|S|^{3/2}\sqrt{T\cdot|A|72\log(8|\mathcal{P}|T^3/\delta)}(1+\log(T/|A|))}{p_{min}} && \text{(Since } H\leq|S|\text{)} \\
&+ 3\frac{\sqrt{T|S||A|17\log(8|\mathcal{F}|T^3/\delta)}(1+\log(T/|A|))}{p_{min}} \\
&+ 2H^2\cdot|S|^{3/2}\cdot\sqrt{T\cdot|A|}\sqrt{72\log(8|\mathcal{P}|T^3/\delta)} \\
&+ 2H\cdot\sqrt{T\cdot|S|\cdot|A|}\sqrt{17\log(8|\mathcal{F}|T^3/\delta)}
\end{aligned}
$$

$$
\begin{aligned}
&+ 2H\sqrt{2T\log(4/\delta)} + |A|H \\
=&\tilde{O}\left(\max\{H, 1/p_{min}\} \cdot \left(H|S|^{3/2}\sqrt{T|A|\log\frac{|\mathcal{P}|}{\delta}} + \sqrt{T|S||A|\log\frac{|\mathcal{F}|}{\delta}} \right) + |A|H \right\} \right).
\end{aligned}
$$

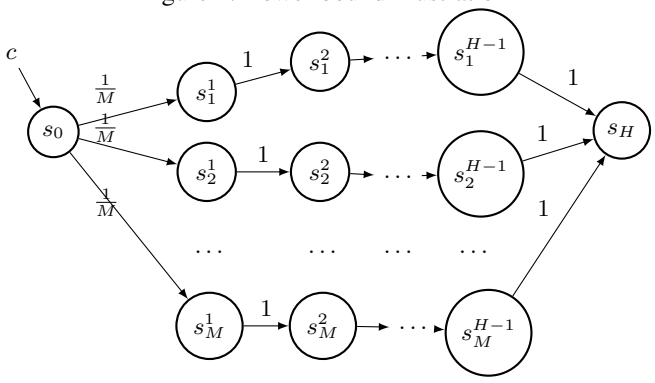
Since the good events hold w.p. at least $1-\delta/2$, by union bound combined with Azuma's inequality we obtain the theorem. ■

Corollary 64 (regret bound in terms of $\mathcal{G}$) *For every $T \geq 1$, finite functions class $\mathcal{G}$ ($\mathcal{F} = \mathcal{G}^S$) and $\mathcal{P}$ and $\delta \in (0,1)$ the following holds with probability at least $1-\delta$ for the same choice of parameters $\{\beta_t, \gamma_t\}_{t\in[T]}$.*

$$
\textit{Regret}_T(RM - UCDD) \leq \tilde{O}\left(\max\{H, 1/p_{min}\} \cdot \left(H|S|^{3/2}\sqrt{T|A|\log\frac{|\mathcal{P}|}{\delta}} + |S|\sqrt{T|A|\log\frac{|\mathcal{G}|}{\delta}} \right) \right).
$$

Proof Plug $\log(|\mathcal{F}|) = |S|\log(|\mathcal{G}|)$ in the bound of Theorem 63. ■

Figure 2: Lower bound illustration



Appendix E. Lower Bound

We present a lower bound for layered CMDP using the lower bound for CMAB presented by Agarwal et al. (2012), in which $K = |A|$, $\mathcal{G} \subseteq \mathcal{C} \times A \to [0,1]$ and $N \in \mathbb{N}$.

Theorem 65 (Theorem 5.1, Agarwal et al. (2012)) *For every N and K such that $\ln N / \ln K \leq T$, and every algorithm $\mathfrak{A}$, there exist a functions class $\mathcal{G}$ of cardinality at most N and a distribution $D(c,r)$ for which the realizability assumption holds, but the expected regret of $\mathfrak{A}$ is $\Omega(\sqrt{KT \ln N / \ln K})$.*

We present a lower bound for **layered CMDP**, where the dynamics is known and context-independent. The rewards are context-dependent. Clearly, it implies a lower bound for the unknown dynamics setting. We remark that similarly to the non-contextual MDP, in the non-layered case, an additional H factor is expected.

In addition we remark that if the horizon length is $H = 1$, since we assume that there is a unique start state and the CMDP is layered, the lower bound for CMAB stated in Theorem 65 holds. In the following theorem we show a lower bound for horizon $H \geq 2$.

Theorem 66 (lower bound for CMDP) *Let $\delta \in (0,1)$, horizon $H \geq 2$ and $M, N \in \mathbb{N}$.*
Let $T \geq 8M \log \frac{|S|}{\delta} + 2M \ln N / \ln |A|$ and consider a CMDP $(\mathcal{C}, S, A, \mathcal{M})$ for which $|S| = M \cdot (H-1) + 2$.

Then, for any algorithm $\mathfrak{A}$, there exist a base function class $\mathcal{G} \subseteq (\mathcal{C} \times A \to [0,1])$ of cardinality at most N, and a distribution $D(c,s,a,r)$ for which the realizability assumption holds for $\mathcal{F} = \mathcal{G}^S$, and, with probability at least $1-\delta$, the expected regret of $\mathfrak{A}$ is $\Omega\left(\sqrt{TH|S||A| \ln(N) / \ln(|A|)}\right)$.

Proof Let $\delta \in (0,1)$, horizon $H \geq 2$, $N, M \in \mathbb{N}$, and T that satisfy the requirements stated in the theorem.

Consider the following layered CMDP, $(\mathcal{C}, S, A, \mathcal{M})$, where $\mathcal{C} \subseteq \mathbb{R}^d$, $\mathcal{M}$ maps a context $c \in \mathcal{C}$ to the MDP $\mathcal{M}(c) = (S, A, P_\star^c, r_\star^c, s_0, H)$ that is defined as follows.
$S = \{S_0, S_1, \ldots, S_H\}$ is a layered states space contains $M \cdot (H-1) + 2$ states for which $S_0 = \{s_0\}$ and $S_H = \{s_H\}$, meaning, s_0 and s_H are unique start and final states, respectively. In addition, for all $i \in \{1, 2, \ldots, H-1\}$ we have $S_i = \{s_1^i, s_2^i, \ldots, s_M^i\}$, meaning, each layer $i \in \{1, \ldots, H-1\}$ contains $M \geq 1$ states and the layers are disjoint.

Let $A = \{a_1, \ldots, a_K\}$ be a set of K actions. We define a context-independent dynamics $P_\star^c = P$ for every context $c \in \mathcal{C}$, as follows.

$$P(s|s_0, a) = \frac{1}{M} \cdot \mathbb{I}[s \in S_1], \quad \forall a \in A.$$
$$P(s|s_j^i, a) = \mathbb{I}[s = s_j^{i+1}], \quad \forall a \in A,\ i \in \{1, \ldots, H-2\},\ j \in \{1, \ldots, M\}.$$

$$P(s|s_j^{H-1}, a) = \mathbb{I}[s = s_H],\ \forall a \in A,\ j \in \{1, \ldots, M\}.$$

For illustration, see Figure 2.

We assume that the dynamics P is known to the learner. $r_\star^c$ is an unknown context-dependent (expected) rewards function.

Remark 67 *In layers $1, 2, \ldots H-1$ the dynamics is deterministic and for any choice of an action $a \in A$ the agent will move to the exactly same state (regardless of the context). In addition, the start state s_0 is unique and identical for all of the contexts, and the final state s_H has zero rewards. Hence, since the dynamics is known to the learner, minimizing regret in this CMDP is equivalent to minimizing regret in $|S| - 1$ CMAB problems.*

By all the above and since the horizon is H, as explain in Remark 67, solving the above CMDP problem is equivalent to solving $|S| - 1 = M \cdot (H-1) + 1$ (unrelated) CMAB problems, one for each state $s \in S \setminus \{s_H\}$.

For $|A| = K$ and N that satisfy the requirements stated in the theorem, by Theorem 65, for any algorithm (for CMAB) there exists a realizable functions class $\mathcal{G} \subseteq (\mathcal{C} \times A \to [0, 1])$ for which $|\mathcal{G}| \leq N$ and the expected regret of the algorithm on the appropriate CMAB problem is $\Omega(\sqrt{T|A| \ln N / \ln |A|})$.

Remark 68 *While the function class $\mathcal{G}$ is the same for all states, the true rewards function of an individual state s ,might be different. Meaning, for two states $s \neq s'$, the true reward functions are $g_s^\star, g_{s'}^\star \in \mathcal{G}$ (respectively) are not necessarily identical. Hence, the CMAB problems defined by s and s' are different, and unrelated.*

Hence, for any algorithm, consider the function class $\mathcal{F} = \mathcal{G}^S \subseteq (\mathcal{C} \times S \times A \to [0, 1])$, where $\mathcal{G}$ is the "hard" function class for the CMAB problem.

Consider a T rounds run of any algorithm $\mathfrak{A}$. For every state $s \in S$ and round $t \in \{1, 2, \ldots, T\}$, let X_s^t be a Bernoulli random variable which indicates whether state s was visited in round t. By our construction, for all $s \in S \setminus \{s_0, s_H\}$ and $t \in \{1, 2, \ldots, T\}$ we have $\mathbb{E}[X_s^t] = \frac{1}{M}$. For all $s \in S$ let $X_s = \sum_{t=1}^T X_s^t$. It holds that $\mathbb{E}[X_s] = \frac{T}{M}$ for all $s \in S \setminus \{s_0, s_H\}$.

For the start and final states s_0 and s_H we have that $X_{s_0} = X_{s_H} = T$ with probability 1.

Hence, by multiplicative Chernoff bound, for all $s \in S \setminus \{s_0, s_H\}$ it holds that

$$\Pr\left[X_s \leq \frac{1}{2} \cdot \frac{T}{M}\right] \leq \exp\left(\frac{-T/M \cdot 0.25}{2}\right) = \exp\left(\frac{-T}{8M}\right).$$

Thus, by union bound

$$\Pr\left[\exists s \in S :\ X_s \leq \frac{T}{2M}\right] \leq \sum_{s \in S} \Pr\left[X_s \leq \frac{T}{2M}\right] \leq |S| \exp\left(\frac{-T}{8M}\right),$$

which implies that

$$\Pr\left[\forall s \in S :\ X_s \geq \frac{T}{2M}\right] \geq 1 - |S| \exp\left(\frac{-T}{8M}\right).$$

Hence, for $T \geq 8M \cdot \log \frac{|S|}{\delta}$ it holds that

$$\Pr\left[\forall s \in S :\ X_s \geq \frac{T}{2M}\right] \geq 1 - \delta. \tag{35}$$

Let E.Regret(X, s) denote the expected regret of state s given that it was visited X times.

By all the above, we have for any algorithm $\mathfrak{A}$ with probability at least $1 - \delta$ that

$$\text{E.Regret}_T(\mathfrak{A}) = \sum_{s \in S \setminus \{s_H\}} \mathbb{E}_{X_s}\left[\ \text{E.Regret}(X_s, s)\ \right]$$

$$
\begin{aligned}
&\geq \sum_{s\in S\setminus\{s_H\}} \text{E.Regret}\left(\frac{T}{2M}, s\right) && \text{(Holds w.p. at least } 1-\delta\text{, by inequality (35))}\\
&\geq \Omega\left(\sum_{s\in S\setminus\{s_H\}} \sqrt{\frac{T}{M}\cdot|A|\ln(N)/\ln(|A|)}\right) && \text{(By Theorem 65, since } \tfrac{T}{2M}\geq \ln N/\ln|A|\text{)}\\
&= \Omega\left(|S|\sqrt{\frac{T}{M}\cdot|A|\ln(N)/\ln(|A|)}\right)\\
&= \Omega\left(|S|\sqrt{\frac{TH}{|S|}\cdot|A|\ln(N)/\ln(|A|)}\right) && \text{(Since } M=\tfrac{|S|-2}{H-1}\text{)}\\
&= \Omega\left(\sqrt{T\cdot H\cdot|S|\cdot|A|\ln(N)/\ln(|A|)}\right).
\end{aligned}
$$

The above holds with probability at least $1-\delta$. ■

Corollary 69 *Under the conditions of Theorem 66, for any $\delta\in(0,1/2]$, the expected regret is lower bounded by $\Omega(\sqrt{TH|S||A|\ln(N)/\ln(|A|)})$ with probability 1.*

Proof Take any $\delta\in(0,1/2]$ and consider the results of Theorem 66.

Let G denote the good event in which every state s was visited at least $\frac{T}{2M}$ times. In the proof of Theorem 66 we showed that $\mathbb{P}[G]\geq 1-\delta$. Hence, the corollary follows by total expectation low when conditioning on the event G.

■

Appendix F. Auxiliary Lemmas

Lemma 70 (Value-difference, corollary 1 Efroni et al. (2020)) *Let M, M' be any H-finite horizon MDP. Then, for any two policies π, π' the following holds*

$$
\begin{aligned}
&V_1^{\pi,M}(s)-V_1^{\pi',M'}(s)=\\
&\qquad=\sum_{h=1}^{H-1}\mathbb{E}\left[\langle Q_h^{\pi,M}(s_h,\cdot),\pi_h(\cdot|s_h)-\pi'_h(\cdot|s_h)\rangle|s_1=s,\pi',M'\right]\\
&\qquad\quad+\sum_{h=1}^{H-1}\mathbb{E}\left[c_h(s_h,a_h)-c'_h(s_h,a_h)+(p_h(\cdot|s_h,a_h)-p'_h(\cdot|s_h,a_h))V_{h+1}^{\pi,M}|s_h=s,\pi',M'\right].
\end{aligned}
$$

Bellow we present two additional **well-known** versions of the above value-difference lemma. We present those versions mainly for our convenience in using them, and provide proofs for completeness.

Lemma 71 (value difference where the dynamics is P) *Let π be a deterministic policy. Let $M=(S,A,P,r,s_0,H)$ and $\overline{M}=(S,A,\overline{P},\overline{r},s_0,H)$ be two H-finite horizon MDPs. Then, for any $s\in S$ and $h\in[H-1]$ it holds that*

$$
\begin{aligned}
&V_{M,h}^{\pi}(s)-V_{\overline{M},h}^{\pi}(s)=\\
&\mathbb{E}_{\pi,P}\left[\sum_{h'=h}^{H-1}\left(r(s_{h'},a_{h'})-\overline{r}(s_{h'},a_{h'})+\sum_{s'\in S}(P(s'|s_{h'},a_{h'})-\overline{P}(s'|s_{h'},a_{h'}))V_{\overline{M},h+1(s')}^{\pi}\right)\Big|s_h=s\right].
\end{aligned}
$$

Remark 72 *Since from the final state (i.e., the state at time H) there are no transitions or rewards, for completeness we define*

$$V^{\pi}_{M,H}(s) = V^{\pi}_{\overline{M},H}(s) = 0, \;\; \forall s \in S.$$

Proof We prove the lemma by backwards induction on h.

Base case: $h = H-1$. By Bellman equations for every state $s \in S$ the following holds.

$$\begin{aligned}
V^{\pi}_{M,H-1}(s) - V^{\pi}_{\overline{M},H-1}(s) =& r(s,\pi(s)) + \underbrace{\mathop{\mathbb{E}}_{s'\sim P(\cdot|s,\pi(s))}\left[V^{\pi}_{M,H}(s')\right]}_{=0} \\
&- \left(\overline{r}(s,\pi(s)) + \underbrace{\mathop{\mathbb{E}}_{s'\sim \overline{P}(\cdot|s,\pi(s))}\left[V^{\pi}_{\overline{M},H}(s')\right]}_{=0}\right) \\
=& r(s,\pi(s)) - \overline{r}(s,\pi(s)) \\
\underbrace{=}_{(1)}& \mathbb{E}_{\pi,P}[r(s_{H-1},\pi(s_{H-1})) - \overline{r}(s_{H-1},\pi(s_{H-1}))|s_{H-1}=s] \\
\underbrace{=}_{(2)}& \mathbb{E}_{\pi,P}\Big[r(s_{H-1},\pi(s_{H-1})) - \overline{r}(s_{H-1},\pi(s_{H-1})) \\
&+ \sum_{s'\in S}(P(s'|s,\pi(s)) - \overline{P}(s'|s,\pi(s)))\cdot V^{\pi}_{\overline{M},H}(s')\Big|s_{H-1}=s\Big],
\end{aligned}$$

where equality (1) is since given that $s_{H-1} = s$, the expectation over π and P has no effect on both terms. (2) is since $V^{\pi}_{\overline{M},H}(s) = 0$ for all $s' \in S$.

Induction step: we assume the induction hypothesis holds for all $k \in [h+1, H]$ and prove for h. Consider the following derivation for any state $s \in S$.

$$\begin{aligned}
& V^{\pi}_{M,h}(s) - V^{\pi}_{\overline{M},h}(s) \\
\underbrace{=}_{(1)}& r(s,\pi(s)) + \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,\pi(s))}\left[V^{\pi}_{M,h+1}(s')\right] - \left(\overline{r}(s,\pi(s)) + \mathop{\mathbb{E}}_{s'\sim \overline{P}(\cdot|s,\pi(s))}\left[V^{\pi}_{\overline{M},h+1}(s')\right]\right) \\
\underbrace{=}_{(2)}& r(s,\pi(s)) - \overline{r}(s,\pi(s)) + \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,\pi(s))}\left[V^{\pi}_{M,h+1}(s')\right] - \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,\pi(s))}\left[V^{\pi}_{\overline{M},h+1}(s')\right] \\
&+ \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,\pi(s))}\left[V^{\pi}_{\overline{M},h+1}(s')\right] - \mathop{\mathbb{E}}_{s'\sim \overline{P}(\cdot|s,\pi(s))}\left[V^{\pi}_{\overline{M},h+1}(s')\right] \\
\underbrace{=}_{(3)}& r(s,\pi(s)) - \overline{r}(s,\pi(s)) + \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,\pi(s))}\left[V^{\pi}_{M,h+1}(s') - V^{\pi}_{\overline{M},h+1}(s')\right] \\
&+ \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,\pi(s))}\left[V^{\pi}_{\overline{M},h+1}(s')\right] - \mathop{\mathbb{E}}_{s'\sim \overline{P}(\cdot|s,\pi(s))}\left[V^{\pi}_{\overline{M},h+1}(s')\right] \\
\underbrace{=}_{(4)}& r(s,\pi(s)) - \overline{r}(s,\pi(s)) + \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,\pi(s))}\Big[\mathbb{E}_{\pi,P}\Big[\sum_{h'=h+1}\Big(r(s_{h'},a_{h'}) - \overline{r}(s_{h'},a_{h'}) \\
&+ \sum_{s''\in S}(P(s''|s_{h'},a_{h'}) - \overline{P}(s''|s_{h'},a_{h'})V^{\pi}_{\overline{M},h'+1}(s'')\Big)\Big|s_{h+1}=s'\Big]\Big] \\
&+ \sum_{s''\in S}(P(s''|s,\pi(s)) - \overline{P}(s''|s,\pi(s))V^{\pi}_{\overline{M},h+1}(s'')
\end{aligned}$$

$$\underbrace{=}_{(5)} \mathbb{E}_{\pi,P}\left[\sum_{h'=h}^{H-1}\left(r(s_{h'},a_{h'})-\overline{r}(s_{h'},a_{h'})+\sum_{s'\in S}\left(P(s'|s_{h'},a_{h'})-\overline{P}(s'|s_{h'},a_{h'})\right)V^{\pi}_{\overline{M},h'+1}(s')\right)\Big|s_h=s\right].$$

Here, (1) is by Bellman equations for the value functions. (2) is by adding and subtracting $\mathbb{E}_{s\sim P(\cdot|s,\pi(s))}\left[V^{\pi}_{\overline{M},h+1}(s')\right]$ and re-organizing. (3) is by linearity of expectation. (4) is by the induction hypothesis. (5) is since the expectation over $s'\sim P(\cdot|s,\pi(s))$ translates to the expectation induced by π and P when applied on s_{h+1} given that $s_h=s$. Hence, given that $s_h=s$ and since π is deterministic, taking expectation over π and P has no influence on both $r(s_h,a_h)=r(s,\pi(s))$ and $\overline{r}(s_h,a_h)=\overline{r}(s,\pi(s))$. ■

Lemma 73 (value difference where the dynamics is $\overline{P}$) *Let π be a deterministic policy. Let $M=(S,A,P,r,s_0,H)$ and $\overline{M}=(S,A,\overline{P},\overline{r},s_0,H)$ be two H-finite horizon MDPs. Then, for any $s\in S$ and $h\in[H-1]$ it holds that*

$$V^{\pi}_{M,h}(s)-V^{\pi}_{\overline{M},h}(s)=$$
$$\mathbb{E}_{\pi,\overline{P}}\left[\sum_{h'=h}^{H-1}\left(r(s_{h'},a_{h'})-\overline{r}(s_{h'},a_{h'})+\sum_{s'\in S}(P(s'|s_{h'},a_{h'})-\overline{P}(s'|s_{h'},a_{h'}))V^{\pi}_{M,h+1(s')}\right)\Big|s_h=s\right].$$

Proof We prove the lemma by backwards induction on h.

Base case: $h=H-1$. By Bellman equations we have for every state $s\in S$:

$$\begin{aligned}V^{\pi}_{M,H-1}(s)-V^{\pi}_{\overline{M},H-1}(s)=&r(s,\pi(s))+\underbrace{\mathbb{E}_{s'\sim P(\cdot|s,\pi(s))}\left[V^{\pi}_{M,H}(s')\right]}_{=0}\\&-\left(\overline{r}(s,\pi(s))+\underbrace{\mathbb{E}_{s'\sim\overline{P}(\cdot|s,\pi(s))}\left[V^{\pi}_{\overline{M},H}(s')\right]}_{=0}\right)\\=&r(s,\pi(s))-\overline{r}_{H-1}(s,\pi(s))\\\underbrace{=}_{(1)}&\mathbb{E}_{\pi,\overline{P}}[r(s_{H-1},\pi(s_{H-1}))-\overline{r}_{H-1}(s_{H-1},\pi(s_{H-1}))|s_{H-1}=s]\\\underbrace{=}_{(2)}&\mathbb{E}_{\pi,\overline{P}}\Big[r(s_{H-1},\pi(s_{H-1}))-\overline{r}_{H-1}(s_{H-1},\pi(s_{H-1}))\\&+\sum_{s'\in S}(P(s'|s,\pi(s))-\overline{P}(s'|s,\pi(s)))\cdot V^{\pi}_{M,H}(s')\Big|s_{H-1}=s\Big],\end{aligned}$$

where equality (1) is since given that $s_{H-1}=s$, the expectation over π and $\overline{P}$ has no effect on both terms. (2) is since $V^{\pi}_{M,H}(s)=0$ for all $s'\in S$.

Induction step: we assume correctness for all $k\in[h+1,H]$ and prove for h. Consider the following derivation for any state $s\in S$.

$$\begin{aligned}&V^{\pi}_{M,h}(s)-V^{\pi}_{\overline{M},h}(s)\\\underbrace{=}_{(1)}&r(s,\pi(s))+\mathbb{E}_{s'\sim P(\cdot|s,\pi(s))}\left[V^{\pi}_{M,h+1}(s')\right]-\left(\overline{r}(s,\pi(s))+\mathbb{E}_{s'\sim\overline{P}(\cdot|s,\pi(s))}\left[V^{\pi}_{\overline{M},h+1}(s')\right]\right)\\\underbrace{=}_{(2)}&r(s,\pi(s))-\overline{r}(s,\pi(s))+\mathbb{E}_{s'\sim P(\cdot|s,\pi(s))}\left[V^{\pi}_{M,h+1}(s')\right]-\mathbb{E}_{s'\sim\overline{P}(\cdot|s,\pi(s))}\left[V^{\pi}_{M,h+1}(s')\right]\\&+\mathbb{E}_{s'\sim\overline{P}(\cdot|s,\pi(s))}\left[V^{\pi}_{M,h+1}(s')\right]-\mathbb{E}_{s'\sim\overline{P}(\cdot|s,\pi(s))}\left[V^{\pi}_{\overline{M},h+1}(s')\right]\end{aligned}$$

$$
\begin{aligned}
&\underbrace{=}_{(3)} r(s,\pi(s)) - \overline{r}(s,\pi(s)) + \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,\pi(s))}\left[V^{\pi}_{M,h+1}(s')\right] - \mathop{\mathbb{E}}_{s'\sim \overline{P}(\cdot|s,\pi(s))}\left[V^{\pi}_{M,h+1}(s')\right] \\
&\qquad + \mathop{\mathbb{E}}_{s'\sim \overline{P}(\cdot|s,\pi(s))}\left[V^{\pi}_{M,h+1}(s') - V^{\pi}_{\overline{M},h+1}(s')\right] \\
&\underbrace{=}_{(4)} r(s,\pi(s)) - \overline{r}(s,\pi(s)) + \sum_{s''\in S}\left(P(s''|s,\pi(s)) - \overline{P}(s''|s,\pi(s)\right)\cdot V^{\pi}_{M,h+1}(s'') \\
&\qquad + \mathop{\mathbb{E}}_{s'\sim \overline{P}(\cdot|s,\pi(s))}\Big[\mathbb{E}_{\pi,\overline{P}}\Big[\sum_{h'=h+1}\Big(r(s_{h'},a_{h'}) - \overline{r}(s_{h'},a_{h'}) \\
&\qquad + \sum_{s''\in S}\left(P(s''|s_{h'},a_{h'}) - \overline{P}(s''|s_{h'},a_{h'})\cdot V^{\pi}_{M,h'+1}(s'')\right)\Big|s_{h+1}=s'\Big]\Big] \\
&\underbrace{=}_{(5)} \mathop{\mathbb{E}}_{\pi,\overline{P}}\left[\sum_{h'=h}^{H-1}\left(r(s_{h'},a_{h'}) - \overline{r}(s_{h'},a_{h'}) + \sum_{s'\in S}\left(P(s'|s_{h'},a_{h'}) - \overline{P}(s'|s_{h'},a_{h'})\right)\cdot V^{\pi}_{M,h'+1}(s')\right)\Big|s_h=s\right].
\end{aligned}
$$

Here, (1) is by Bellman equations for the value functions. (2) is by adding and subtracting $\mathbb{E}_{s\sim\overline{P}(\cdot|s,\pi(s))}\left[V^{\pi}_{M,h+1}(s')\right]$ and re-organizing. (3) is by linearity of expectation. (4) is by the induction hypothesis. (5) is since the expectation over $s'\sim\overline{P}(\cdot|s,\pi(s))$ translates to the expectation induced by π and $\overline{P}$ when applied on s_{h+1} given that $s_h=s$. Hence given that $s_h=s$ and π is deterministic, taking expectation over π and $\overline{P}$ has no influence on both $r(s_h,a_h)=r(s,\pi(s))$ and $\overline{r}(s_h,a_h)=\overline{r}(s,\pi(s))$. ■

Lemma 74 (Bretagnolle Huber-Carol inequality) *Let X be a random variable taking values in $\{1,2,\ldots,k\}$ where $\mathbb{P}[X=i]=p_i$. Assume we sample X for n times and observe the value i in $\hat{n}_i$ outcomes. Then,*

$$
\mathbb{P}\left[\sum_{i=1}^{k}\left|\frac{\hat{n}_i}{n} - p_i\right| \geq \lambda\right] \leq 2^{k+1}\exp\left(-n\lambda^2/2\right).
$$

The above implies that w.p. at least $1-\delta$ we have

$$
\sum_{i=1}^{k}\left|\frac{\hat{n}_i}{n} - p_i\right| \geq \lambda
$$

for any

$$
\lambda \geq \sqrt{\frac{2}{n}\ln(1/\delta) + (|S|+1)\ln(2)}.
$$