Offline Oracle-Efficient Learning for Contextual MDPs via Layerwise Exploration-Exploitation Tradeoff

Jian Qian
MIT EECS
jianqian@mit.edu

Haichen Hu
MIT LIDS
huhc@mit.edu

David Simchi-Levi
MIT IDSS
dslevi@mit.edu

Abstract

Motivated by the recent discovery of a statistical and computational reduction from contextual bandits to offline regression [36], we address the general (stochastic) Contextual Markov Decision Process (CMDP) problem with horizon H (as known as CMDP with H layers). In this paper, we introduce a reduction from CMDPs to offline density estimation under the realizability assumption, i.e., a model class $\mathcal{M}$ containing the true underlying CMDP is provided in advance. We develop an efficient, statistically near-optimal algorithm requiring only $O(H \log T)$ calls to an offline density estimation algorithm (or oracle) across all T rounds of interaction. This number can be further reduced to $O(H \log\log T)$ if T is known in advance. Our results mark the first efficient and near-optimal reduction from CMDPs to offline density estimation without imposing any structural assumptions on the model class. A notable feature of our algorithm is the design of a layerwise exploration-exploitation tradeoff tailored to address the layerwise structure of CMDPs. Additionally, our algorithm is versatile and applicable to pure exploration tasks in reward-free reinforcement learning.

1 Introduction

Markov Decision Processes (MDPs) model the long-term interaction between a learner and the environment and are used in diverse areas such as inventory management, recommendation systems, advertising, and healthcare [35, 37]. The Contextual MDP (CMDP) extends MDPs by incorporating external factors, known as *contexts*, such as gender, age, location, or device in customer interactions, or lab data and medical history in healthcare [19, 33]. In an H-layer CMDP, the learner receives an instantaneous reward at each step over H steps and aims to maximize the cumulative reward (return). For T rounds of interaction, the learner's performance is measured by *regret*, which is the difference between the total return obtained and that of an optimal policy.

In the special case of contextual bandits (one-layer CMDPs), a decade of research has led to algorithms with optimal regret bounds and efficient implementations with access to an offline regression algorithm (also termed as an *offline regression oracle*) [13, 2, 3, 15, 14, 36, 40]. Most notably, Simchi-Levi and Xu [36] demonstrates an offline-oracle-based algorithm FALCON that achieves optimal regret for general (stochastic) contextual bandits with access to an offline regression oracle (e.g., the Empirical Risk Minimization (ERM) oracle). Moreover, the algorithm is efficient given the output of the offline oracle (also referred to as offline oracle-efficient) and requires only $O(\log\log T)$ calls to the oracle across all T rounds if T is known. These properties are highly desirable in practice since they reduce the computational problem of contextual bandits to the classical problem of offline regression with little overhead. However, to the best of our knowledge, no algorithm with these properties is available in the literature for general (stochastic) CMDPs. So, in this paper, we study the following question: *Is there an offline-oracle-efficient algorithm that achieves the optimal regret for general (stochastic) CMDPs with only $O(H \log\log T)$ number of oracle calls?*

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Table 1: Algorithms' performance with general finite model class $\mathcal{M}$ and i.i.d. contexts. The optimal rate here refers to $\widetilde{O}(\text{poly}(H, S, A)\sqrt{T \log |\mathcal{M}|})$. All algorithms assume realizability, so it is omitted from the table. The reachability assumption and the varying representation assumption are very stringent for tabular CMDP, for details we refer to Appendix B.

Algorithm	Regret rate	Computational complexity	Assumption
E2D [16]	Optimal	$O(T)$ online oracle calls	No
OMG-CMDP! [26]	Optimal	$O(T)$ online oracle calls	No
RM-UCDD [24]	Suboptimal	$O(T)$ offline oracle calls	Reachability
CMDP-VR [11]	Optimal	$O(T)$ offline oracle calls	Varying Rep.
LOLIPOP (**this work**)	Optimal	$O(\log T)$ offline oracle calls	No

Several works have provided partial results for this question. Foster et al. [16] provides a general reduction from interactive decision making to online density estimation and has CMDP as an application. The proposed E2D algorithm achieves optimal regret but is online-oracle-efficient (as opposed to offline-oracle-efficient) since it requires access to an online density estimation algorithm. Foster et al. [18] provides a further reduction from online density estimation to offline density estimation, with the caveat that the reduction itself is inefficient. A similar online-oracle-efficient algorithm is developed by Levy et al. [26]. A separate thread of optimism-based algorithms for CMDPs extending the UCCB algorithm for contextual bandits [40] is studied by Levy and Mansour [24], Deng et al. [11] with either assumption on the reachability of the CMDP or the representation structure of the CMDP (see Appendix B for more details). Last but not least, the algorithms proposed by Foster et al. [16, 18], Levy and Mansour [24], Deng et al. [11] all require $O(T)$ number of oracle calls to the online/offline oracle respectively.

In this work, we present an affirmative answer to the question by introducing the algorithm of LOLIPOP (Algorithm 1). For S number of states, A number of actions, and a given model class $\mathcal{M}$ where the underlying true model lies, the algorithm achieves the regret guarantee of $\widetilde{O}(\text{poly}(H, S, A)\sqrt{T \log |\mathcal{M}|})$. This regret guarantee is minimax optimal up to $\text{poly}(H, S, A)$ factor [24]. The LOLIPOP algorithm assumes access to a Maximum Likelihood Estimation (MLE) oracle and is offline-oracle efficient. The results can be generalized to general offline density estimation oracles. The most notable technical features are: (1) It generalizes the FALCON algorithm by Simchi-Levi and Xu [36] to adapt to the multi-layer structure of a CMDP. More specifically, the FALCON algorithm is divided into $O(\log \log T)$ epochs, each corresponding to an oracle call, a fixed randomized policy. However, it is known for the MDPs that the learner has to switch its randomized policy at least $\widetilde{\Omega}(H)$ times to achieve sublinear regret [44]. Indeed, we further divide each epoch into H segments, each with an oracle call, a new randomized policy for layerwise exploration-exploitation tradeoff. (2) In each segment, the exploration-exploitation tradeoff is done through Inverse Gap Weighting (IGW) on estimated regret for a set of explorative policies. The idea of running IGW on such a policy cover is proposed by Foster et al. [16]. However, their policy cover is designed for H-layer exploration-exploration tradeoff and only works with strong online estimation oracles. In contrast, we refine the estimation of the occupancy measure layerwise by introducing the *trusted occupancy measures*. This refinement enables our algorithm to work with offline oracles. (3) Many other policy cover-based methods [12, 29, 30, 5] are developed for exploration tasks. Most notably, Mhammedi et al. [29] clips the occupancy measures on states with low reachability. Our approach takes a step forward to clip all transitions with low reachability to compute the trusted occupancy measures.

Besides all the above novelties, the LOLIPOP algorithm is versatile and applicable to the pure exploration task of reward-free reinforcement learning for CMDPs. Concretely, it obtains near-optimal sample complexity of $O\big(H^7 S^4 A^3 \log(|\mathcal{M}|/\delta)/\varepsilon^2\big)$ with only $O(H)$ number of oracle calls. Both the sample complexity bound and computational efficiency result for reward-free learning for stochastic CMDPs are new to the best of our knowledge.

2 Preliminaries

We defer the standard notation and related works to Appendices A and B.

2.1 Problem Setup

A Contextual Makovian Decision Process (CMDP) is defined by the tuple $(\mathcal{C}, M = \{M(c)\}_{c\in\mathcal{C}}, \mathcal{S}, \mathcal{A}, s^1)$, where $\mathcal{C}$ is the contextual space, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space and $s^1 \in \mathcal{S}$ is a fixed starting state independent of the context. We focus on tabular CMDPs which assumes $S = |\mathcal{S}|, A = |\mathcal{A}| < \infty$. For any context $c \in \mathcal{C}$, $M(c) = \{P_M^h(c), R_M^h(c)\}_{h=1}^H$ consists of H-layers of probability transition kernel $\{P_M^h(c)\}_{h=1}^H$ and reward distributions $\{R_M^h(c)\}_{h=1}^H$, where $P_M^h(c)$ and $R_M^h(c)$ are specified by $P_M^h(\cdot \mid s, a; c) \in \Delta(\mathcal{S})$ and $R_M^h(s, a; c) \in \Delta([0,1])$ for all $h \in [H]$ and $s, a \in \mathcal{S} \times \mathcal{A}$. For simplicity, we also denote $M = \{P_M^h, R_M^h\}_{h=1}^H$, where $P_M^h = \{P_M^h(c)\}_{c\in\mathcal{C}}$ and $R_M^h = \{R_M^h(c)\}_{c\in\mathcal{C}}$. Let Π_{RNS} denote the set of all randomized, non-stationary policies, where any $\pi = (\pi^1, \ldots, \pi^H) \in \Pi_{\mathrm{RNS}}$ has $\pi^h : \mathcal{S} \to \Delta(\mathcal{A})$ for any $h \in [H]$. We use T to denote the total number of rounds, H to denote the horizon (the total number of layers). Let $M_\star = \{P_\star^h, R_\star^h\}_{h\in[H]}$ be the underlying true CMDP the learner interact with. The interactive protocal proceeds in T rounds, where for each round t, the t-th trajectory is generated as:

- A context c_t is draw i.i.d. from an unknown distribution $\mathcal{D}$ and $s_t^1 = s^1$.
- The learner chooses the policy π_t based on the context.
- For $h = 1, \ldots, H$:
  - The action is drawn from the randomized policy $a_t^h \sim \pi_t^h(s_t^h)$.
  - The reward and the next state is drawn respectively from the reward distribution and the transition kernel, i.e., $r_t^h \sim R_\star^h(s_t^h, a_t^h; c_t)$ and $s_t^{h+1} \sim P_\star^h(\cdot \mid s_t^h, a_t^h; c_t)$.

Without lose of generality, throughout the paper, we assume that the total reward $0 \le \sum_{h=1}^H r^h \le 1$ almost surely. For any model M, context c and policy π, we use $M(\pi, c)$ to denote the distribution of the trajectory $c_1, \pi_1, s_1^1, a_1^1, r_1^1, \ldots, s_1^H, a_1^H, r_1^H$ given $M_\star = M$, $c_1 = c$, and $\pi_1 = \pi$. Also denote the probability and the expectation under $M(\pi, c)$ to be $\mathbb{P}^{M,\pi,c}(\cdot)$ and $\mathbb{E}^{M,\pi,c}[\cdot]$ respectively. Given any policy π, state s and action a, we define the action value function $Q_\star^h(s, a; \pi, c)$ at layer h and the value function $V_\star^h(s; \pi, c)$ at layer h under context c and policy π as

$$Q_\star^h(s, a; \pi, c) = \textstyle\sum_{j=h}^H \mathbb{E}^{M_\star,\pi,c}[r_1^j \mid s_1^h, a_1^h = s, a] \quad \text{and} \quad V_\star^h(s; \pi, c) = \max_{a\in\mathcal{A}} Q_\star^h(s, a; \pi, c).$$

We denote the optimal policy under context c as $\pi_{\star,c}$ and abbreviate its value function as $V_\star^h(\cdot; c)$. For $h = 1$, we further simply the notation by denoting $V_\star^1(c) = V_\star^1(s^1; c)$ and $V_\star^1(\pi, c) = V_\star^1(s^1; \pi, c)$. The regret of policy π under context c and the total regret[1] of the learner are defined as

$$\mathrm{reg}(\pi, c) = V_\star^1(c) - V_\star^1(\pi, c) \quad \text{and} \quad \mathrm{Reg}(T) := \textstyle\sum_{t=1}^T \mathbb{E}_t[\mathrm{reg}(\pi_t, c_t)],$$

where $\mathbb{E}_t[\cdot]$ is the conditional expectation given the interaction up to round t.

Assumption 2.1 (Realizability)**.** *The learner is given a model class $\mathcal{M}$ where the true underlying model $M_\star$ lies, that is, $M_\star \in \mathcal{M}$.*

2.2 Offline Densitiy Estimation Oracles

For any model class $\mathcal{M}$, a general offline density estimation oracle associated with $\mathcal{M}$, denoted by $\mathrm{OffDE}_{\mathcal{M}}$, is defined as an algorithm that generates a predictor $\widehat{M}$ based on the input data and $\mathcal{M}$. In this paper, we measure the performance of the predictor in terms of the squared Hellinger distance, which is defined for any two distributions $\mathbb{P}$ and $\mathbb{Q}$ for any common dominating measure ν[2] by

$$D_{\mathrm{H}}^2(\mathbb{P}, \mathbb{Q}) := \tfrac{1}{2} \int \left(\sqrt{\tfrac{d\mathbb{P}}{d\nu}} - \sqrt{\tfrac{d\mathbb{Q}}{d\nu}} \right)^2 d\nu,$$

Using squared Hellinger distance for reinforcement learning is popularized by Foster et al. [16], and we adopt such a divergence for our purpose as well. Concretely, we are interested in the following statistical guarantee.

Definition 2.1 (Offline density estimation oracle)**.** *Let p be a map from a context to a distribution on the set of policies Π_{RNS}, that is, for any $c \in \mathcal{C}$, $p(c) \in \Delta(\Pi_{\mathrm{RNS}})$. Given n training trajectories $(c_i, \pi_i, s_i^1, a_i^1, r_i^1, \ldots, s_i^H, a_i^H, r_i^H)$ i.i.d. drawn according to $c_i \sim \mathcal{D}$, $\pi_i \sim p(c_i)$ and*

[1] The regret we defined here is conventionally known as the pseudo-regret in the literature. The conventional regret defined as $\sum_{t=1}^T \mathrm{reg}(\pi_t, c_t)$ can be bounded by the pseudo-regret up to an additional $O(\sqrt{T \log(1/\delta)})$ term with a standard concentration argument, which we omit here for simplicity.

[2] The value is independent of the choice of ν.

$s_i^1, a_i^1, r_i^1, \dots, s_i^H, a_i^H, r_i^H$ *be the trajectory sampled according to* $M_\star(\pi_i, c_i)$*. The offline density estimation oracle* $\mathrm{OffDE}_{\mathcal{M}}$ *returns a predictor* $\widehat{M}$*. For any* $\delta \in (0, 1/2)$*, with probability at least* $1 - \delta$*, we have*

$$\mathbb{E}_{c\sim\mathcal{D},\pi\sim p(c)}\left[D_{\mathrm{H}}^2\left(\widehat{M}(\pi, c), M_\star(\pi, c)\right)\right] \le \mathcal{E}_{\mathcal{M},\delta}(n).$$

The Maximum Likelihood Estimation estimator $\mathsf{MLE}_{\mathcal{M}}$ is an example of an offline density estimation oracle that achieves $\mathcal{E}_{\mathcal{M},\delta}(n) \lesssim \log(|\mathcal{M}|/\delta)/n$ (see more details in Appendix C) and can be implemented using ERM on the log loss. Moreover, this implementation can be efficient for cases like the multinomial logit model [32, Example 2.4] and [1].

3 Main Results and Algorithm

In this section, we present our main results and introduce the algorithm of LOLIPOP (Algorithm 1). First, we give an overview of the algorithm. Then, we discuss the theoretical guarantees obtained by this algorithm. Finally, we introduce the algorithm's different components with corresponding guarantees. All proofs are deferred to Appendix D.

3.1 Main Results

Overview of Algorithm 1. The algorithm proceeds with epochs. The total number of T rounds is divided into N epochs. For an epoch schedule $0 = \tau_0 < \tau_1 < \cdots < \tau_N = T/H$ to be specified later, the m-th epoch will last $H(\tau_m - \tau_{m-1})$ rounds. Furthermore, each epoch is evenly divided into H segments, each consisting of $\tau_m - \tau_{m-1}$ rounds. During the h-th segment in m-th epoch, a kernel $p_m^h : \mathcal{C} \to \Delta(\Pi)$ will be specified to determine the policy. More specifically, upon receiving the context c_t, a policy π_t will be sampled from $p_m^h(c_t)$ and executed. After collecting the trajectories $\{c_t, \pi_t\} \cup \{s_t^j, a_t^j, r_t^j\}_{j\in[H]}$ in the h-th segment of the m-th epoch for $\tau_{m-1}H + (\tau_m - \tau_{m-1})(h - 1) + 1 \le t \le \tau_{m-1}H + (\tau_m - \tau_{m-1})h$, the offline density estimation oracle $\mathrm{OffDE}_{\mathcal{M}}$ is called with these trajectories as input. Denote the output $\widehat{M}_m^h$, we will only be interested in the h-th layer of this output, which we denote by $\{\widehat{P}_m^h, \widehat{R}_m^h\}$. Then the collections of estimators $\widehat{M}_m = \{\widehat{P}_m^h, \widehat{R}_m^h\}_{h=1}^H$ will be used for the next epoch.

Throughout this paper, we will adopt the following convention for the free variables m, π, c, h, s, a. They will be used to denote an epoch index in $[N]$, a policy in Π_{RNS}, a context in $\mathcal{C}$, a layer index in $[H]$, a state in $\mathcal{S}$, and an action in $\mathcal{A}$ respectively.

Before we dive into the details of the algorithm, we highlight first the theoretical guarantees obtained.

Theorem 3.1. *If T is known, then by choosing the epoch schedule $\tau_m = 2(T/H)^{1-2^{-m}}$ for $m \ge 1$ and the offline density estimation oracle $\mathrm{OffDE}_{\mathcal{M}} = \mathsf{MLE}_{\mathcal{M}}$, the outputs $\{\pi_t\}_{t\in[T]}$ of Algorithm 1 satisfies that with probability at least $1 - \delta$,*

$$\mathrm{Reg}(T) \lesssim \sqrt{H^7 S^4 A^3 T \cdot \log(|\mathcal{M}| \log\log T/\delta) \log\log T}$$

with only $O(H \log\log T)$ number of oracle calls to the $\mathsf{MLE}_{\mathcal{M}}$ oracle for $\delta \in (0, 1/2)$.

Theorem 3.2. *If T is not known, then by choosing the epoch schedule $\tau_m = 2^m$ for $m \ge 1$ and the offline density estimation oracle $\mathrm{OffDE}_{\mathcal{M}} = \mathsf{MLE}_{\mathcal{M}}$, the outputs $\{\pi_t\}_{t\in[T]}$ of Algorithm 1 satisfies that with probability at least $1 - \delta$,*

$$\mathrm{Reg}(T) \lesssim \sqrt{H^7 S^4 A^3 T \cdot \log(|\mathcal{M}| \log T/\delta)}$$

with $O(H \log T)$ number of oracle calls to the $\mathsf{MLE}_{\mathcal{M}}$ oracle for $\delta \in (0, 1/2)$.

The theorems above show that Algorithm 1 with both epoch schedules achieve near-optimal statistical complexity that a matches the lower bound of $\Omega(\sqrt{HSAT \log |\mathcal{M}| / \log A})$ proven by Levy and Mansour [24] up to a $\mathrm{poly}(H, S, A)$ factor.

Computational efficiency. Consider the epoch schedule $\tau_m = 2^m$ for $m \in \mathbb{N}$ as discussed in Theorem 3.2. For any unknown T, our algorithm operates over $O(\log T)$ epochs, making one oracle call per epoch. Thus, the computational complexity is $O(\log T)$ oracle calls over T rounds, with an additional per-round cost of $O(\mathrm{poly}(H, S, A, \log T))$. This offers potential advantages over existing algorithms that achieve near-optimal rates without assumptions beyond realizability. The E2D

Algorithm 1 Layerwise pOLicy cover Inverse gaP weighting with trusted OccuPancy measures (LOLIPOP)

```
Require: epoch schedule $0 = \tau_{-1} = \tau_0 < \tau_1 < \cdots < \tau_N = T/H$, confidence parameter $\delta \in (0, 1/2)$, model class $\mathcal{M}$, offline oracle $\mathrm{OffDE}_{\mathcal{M}}$.
Initialize: $\widehat{M}_0 = \{\widehat{P}_0^h, \widehat{R}_0\}_{h=1}^H$, where $\widehat{P}_0^h$ is any transtion kernel and $\widehat{R}_0$ is constantly $0$.
for epoch $m = 1, 2, \cdots, N$ do
    Let $\mathcal{E}_m = \mathcal{E}_{\mathcal{M},\delta/2N^2}(\tau_{m-1} - \tau_{m-2})$, $\gamma_m = \sqrt{H^6 S^4 A^3 / \mathcal{E}_m}$ and $\eta_m = \gamma_m / 720 e H^5 S^3 A^2$.
    for segment $h = 1, \ldots, H$ do
        for round $t = \tau_{m-1}H + (\tau_m - \tau_{m-1})(j-1) + 1, \cdots, \tau_{m-1}H + (\tau_m - \tau_{m-1})h$ do
            Observe context $c_t \in \mathcal{C}$ from the environment.
            for $s, a \in \mathcal{S} \times \mathcal{A}$ do
                Compute
                    $\pi_{m,c_t}^{h,s,a} = \arg\max_{\pi} \frac{\widetilde{d}_m^h(s, a; \pi, c_t)}{SA + \eta_m \cdot \widehat{\mathrm{reg}}_{m-1}(\pi, c_t)},$
                where $\widetilde{d}_m^h$ is the trusted occupancy measure defined as in Definition 3.1.
            Let the policy cover $\Pi_m^h(c_t) = \{\widehat{\pi}_{m-1,c_t}\} \cup \{\pi_{m,c_t}^{h,s,a}\}_{s,a \in \mathcal{S} \times \mathcal{A}}$.
            Define $p_m^h(c_t)$ to be the Inverse Gap Weighting distribution on the policy cover $\Pi_{m,c_t}^h$
                $p_m^h(c_t, \pi) = \frac{1}{\lambda_{m,c_t}^h + \eta_m \cdot \widehat{\mathrm{reg}}_{m-1}(\pi, c_t)}, \quad \forall \pi \in \Pi_m^h(c_t),$    (1)
            where $\lambda_{m,c_t}^h$ is the constant that normalize the distribution.
            Sample and execute $\pi_t \sim p_m^h(c_t)$ and observe the trajectory $c_t, \pi_t, \{s_t^j, a_t^j, r_t^j\}_{j \in [H]}$.
        Run $\mathrm{OffDE}_{\mathcal{M}}$ with the input trajectories $\{c_t, \pi_t, \{s_t^j, a_t^j, r_t^j\}_{j \in [H]}\}_{t:m(t)=m}$ and obtain the $h$-th layer estimator $\widehat{P}_m^h$ and $\widehat{R}_m^h$.
```

algorithm [16], for instance, requires $O(T)$ calls to an online density estimation oracle, involving significantly more calls to a more complex oracle for a general model class $\mathcal{M}$. On the other hand, the Version Space Averaging + E2D algorithm [18] requires $O(T)$ calls to an offline density estimation oracle and incurs a computational cost scaling with $O(|\mathcal{M}|)$ per round. Compared to our algorithm, this results in far more oracle calls and considerably higher computational costs per round.

If the total number of rounds T is known to the learner, we can further reduce the computational cost of LOLIPOP. For any $T \in \mathbb{N}$, consider the epoch schedule $\tau_m = 2(T/H)^{1-2^{-m}}$ as in Theorem 3.1, similar to Simchi-Levi and Xu [36]. In this scenario, LOLIPOP will run in $O(\log \log T)$ epochs, making only $O(\log \log T)$ oracle calls over T rounds while still maintaining a slightly worse regret guarantee.

Lower bound on switching cost. There is a lower bound on the switching cost of the scale $\Omega(\log \log T)$ [44], where the switching cost is the number of switches in the learner's randomized policy. Thus, any learner that only switches its randomized policy after an oracle call will need more than $\Omega(\log \log T)$ number of oracle calls.

Technical challenge. The main technical challenge is to accurately estimate the occupancy measures for all layers. Naively, the upper bound of divergence between the true occupancy measure and the estimated occupancy measure accumulates exponentially with respect to the number of layers because, naively, the divergence at each layer is upper bounded by the summation of divergences from previous steps. This phenomenon is unavoidable when estimating all the occupancy measures from one dataset generated from a single policy. To avoid this exponential divergence, we apply two methods. First, we turn to layerwise design. Specifically, we generate for occupancy measures at each layer a new dataset from a different policy. This alleviates the exponential accumulation of divergence. Second, we turn to multiplicative guarantees between true occupancy measures and the approximated occupancy measures, i.e., they are equivalent up to a small constant. To achieve this, we construct the trusted occupancy measure (see Definition 3.1), which discards the rarely visited state-action pairs. We then use the trusted occupancy measures to guide exploration.

$$\begin{tikzcd}
\left\{\widehat{P}^h_{m-1}, \widehat{R}^h_{m-1}\right\}_{h\in[H]} \arrow[r] \arrow[rrr, bend left] \arrow[rrrrr, bend left] & \Pi^1_m, p^1_m \arrow[d] & & \Pi^2_m, p^2_m \arrow[d] & & \ldots\ldots \arrow[d] \\
& \widehat{P}^1_m, \widehat{R}^1_m \arrow[r] & \widetilde{\mathcal{T}}^1_m, \widetilde{d}^2_m \arrow[ur] & \widehat{P}^2_m, \widehat{R}^2_m \arrow[r] & \widetilde{\mathcal{T}}^2_m, \widetilde{d}^3_m \arrow[ur] & \widehat{P}^H_m, \widehat{R}^H_m
\end{tikzcd}$$

Figure 1: The dependence graph of the construction. The estimation $\widehat{M}_{m-1} = \{\widehat{P}^h_{m-1}, \widehat{R}^h_{m-1}\}_{h\in[H]}$ from the previous round provides the optimal policy $\widehat{\pi}_{m-1}$ (Line 9) and the regret estimation $\widehat{\text{reg}}_{m-1}$ (Line 8) for the construction of Π^h_m, p^h_m. The estimation $\widehat{P}^h_m, \widehat{R}^h_m$ is generated by calling the oracle $\text{OffDE}_{\mathcal{M}}$ on the trajectories collected with policy kernel p^h_m (Line 12). The trusted transitions and trusted occupancy measures $\widetilde{\mathcal{T}}^h_m, \widetilde{d}^{h+1}_m$ are computed from $\widetilde{d}^h_m, \widehat{P}^h_m$ (Eqs. (2) and (3)). The policy cover Π^h_m is the union of $\widehat{\pi}_{m-1,\cdot}$ and the policies $\{\pi^{h,s,a}_{m,\cdot}\}_{s,a\in\mathcal{S}\times\mathcal{A}}$ calculated in Line 8 which requires $\widetilde{\mathcal{T}}^{h-1}_m, \widetilde{d}^h_m$. The policy kernel p^h_m is the inverse gap weighting on Π^h_m (Line 10).

3.2 Detailed Construction and Guarantees of Each Component

In this section, we explain in detail our construction and the guarantees of each component along the dependence graph (Figure 1). We first introduce how the estimators $\{\widehat{P}^h_{m-1}, \widehat{R}^h_{m-1}\}_{h\in[H]}$ from the previous epoch are used in the new epoch. Then we proceed to introduce how to construct $p^h_m(c_t)$ given $\Pi^h_m(c_t)$. Next, we introduce how are $\widehat{P}^h_m, \widehat{R}^h_m$ obtained given p^h_m. Subsequently, we present the definition of the set of trusted transitions $\widetilde{\mathcal{T}}^h_m$ and trusted occupancy measure $\widetilde{d}^h_m$. Finally, we present how the policy cover $\Pi^h_m(c_t)$ is constructed.

During epoch m, we will be using the estimators $\{\widehat{P}^h_{m-1}, \widehat{R}^h_{m-1}\}_{h=1}^H$ from the previous epoch for regret estimation. More specifically, for π, c, h, s, we define the value functions with respect to the model $\{\widehat{P}^h_{m-1}(c), \widehat{R}^h_{m-1}(c)\}_{h=1}^H$ as $\widehat{V}^h_{m-1}(s;\pi,c)$. The optimal value function is denoted by $\widehat{V}^1_{m-1}(s;c) = \max_\pi \widehat{V}^1_{m-1}(s;\pi,c)$. For $h = 1$, we further simply the notation by denoting $\widehat{V}^1_{m-1}(c) = \widehat{V}^1_{m-1}(s^1;c)$ and $\widehat{V}^1_{m-1}(\pi,c) = \widehat{V}^1_{m-1}(s^1;\pi,c)$. Also denote the optimal policy under context c by $\widehat{\pi}_{m-1,c} = \arg\max_\pi \widehat{V}^1_{m-1}(\pi,c)$. Thus, the regret is estimated to be

$$\widehat{\text{reg}}_{m-1}(\pi,c) = \widehat{V}^1_{m-1}(c) - \widehat{V}^1_{m-1}(\pi,c).$$

At round t, let $m(t)$ and $h(t)$ be the epoch in which the segment round t lies. We note that during each epoch m and segment h, all of the notions $\Pi^h_m(c_t)$, $\widehat{P}^h_m(c_t)$, $\widehat{R}^h_m(c_t)$, $\widetilde{\mathcal{T}}^h_m(c_t)$, $\widetilde{d}^h_m(\cdot,\cdot;\cdot,c_t)$, $p^h_m(c_t)$, and $\pi^{h,s,a}_{m,c_t}$ will not depend on the specific time step t, but only the context c_t. Thus, we will use $\Pi^h_m(c)$ to denote the policy cover if $c_t = c$ when $m(t), h(t) = m, h$. Similar conventions regarding the context c apply to the notations $\widehat{P}^h_m, \widehat{R}^h_m, \widetilde{\mathcal{T}}^h_m, \widetilde{d}^h_m, p^h_m, \pi^{h,s,a}_{m,\cdot}$. Under any context c, the policy cover Π^h_m will include $\widehat{\pi}_{m-1,c}$ and has no more than $SA + 1$ policies. These two properties together guarantee that the Inverse Gap Weighting [14] randomized policy $p^h_m(c)$ (Line 10) satisfies the following guarantee on the estimated regret.

Lemma 3.1. *For any m, h, c, the definition of the randomized policy $p^h_m(c)$ is well defined, i.e., there exist $\lambda^h_{m,c} \in [0, 2SA]$ such that $\sum_{\pi\in\Pi^h_m(c)} p^h_{m,c}(\pi) = 1$. Furthermore, we have the estimated regret is bounded by $\mathbb{E}_{\pi\sim p^h_m(c)}\left[\widehat{\text{reg}}_{m-1}(\pi,c)\right] \lesssim \sqrt{H^4S^4A^3\cdot\mathcal{E}_m}$.*

The choice of $\lambda^h_{m,c}$ here is for compactness of presentation. It can be chosen to be $2SA$ for suboptimal arms and collect the probability remained to the optimal arm [36], which is computationally efficient. The computation for the policy $\pi^{h(t),s,a}_{m(t),c_t}$ for any t, s, a can be computed in $\text{poly}(H, S, A, \log T)$ time by formulating it as a linear fractional programming problem. We defer the details to Appendix G.

Since p^h_m maps $\mathcal{C}$ to randomized policies, it is thus a policy kernel. This means the trajectories generated in epoch m and segment h follow an i.i.d. distribution as described in the definition of Definition 2.1. By applying the guarantee from Definition 2.1, we have the following guarantee on $\widehat{P}^h_m, \widehat{R}^h_m$.

Lemma 3.2. *For any m, h, and $c \sim \mathcal{D}, \pi \sim p_m^h(c)$, we have with probability at least $1 - \frac{\delta}{2N^2}$, that*

$$\mathbb{E}_{c,\pi}\left[\mathbb{E}^{M_\star,\pi,c}\left[D_{\mathrm{H}}^2\left(\widehat{P}_m^h(s_1^h, a_1^h; c), P_\star^h(s_1^h, a_1^h; c)\right) + D_{\mathrm{H}}^2\left(\widehat{R}_m^h(s_1^h, a_1^h; c), R_\star^h(s_1^h, a_1^h; c)\right)\right]\right] \leq \mathcal{E}_m.$$

If the offline density estimation oracle is chosen to be the Maximum Likelihood Estimation oracle $\mathsf{MLE}_{\mathcal{M}}$, we will obtain $\mathcal{E}_m \lesssim \log(T\mathcal{M}/\delta)/(\tau_{m-1} - \tau_{m-2})$.

The most involved part of our construction concerns the idea of *trusted transitions* and *trusted occupancy measures*. This construction eliminates the parts of transitions that are too scarcely visited. The purpose will be clear in the guarantees (Lemmas 3.3 and 3.4) subsequent to the definitions.

Definition 3.1. *For any m, π, c, h, s, a, we define iteratively the* ***trusted occupancy measures*** *$\widetilde{d}_m^h(s; \pi, c)$, $\widetilde{d}_m^h(s, a; \pi, c)$ and the set of* ***trusted transitions*** *$\widetilde{\mathcal{T}}_m^h(c)$ at layer h as the following:*

$$\begin{aligned} \widetilde{d}_m^1(s; \pi, c) &= \mathbb{1}(s = s^1), \qquad \widetilde{d}_m^1(s, a; \pi, c) = \mathbb{1}(s = s^1)\pi^1(s, a), \\ \widetilde{d}_m^h(s; \pi, c) &:= \sum\nolimits_{s', a', s \in \widetilde{\mathcal{T}}_m^{h-1}(c)} \widetilde{d}_m^{h-1}(s', a'; \pi, c)\widehat{P}_m^{h-1}(s|s', a'; c), \\ \widetilde{d}_m^h(s, a; \pi, c) &:= \widetilde{d}_m^h(s; \pi, c)\pi^h(s, a). \end{aligned} \tag{2}$$

For any m, h, c, the set of trusted transitions are defined as the set of transtions

$$\widetilde{\mathcal{T}}_m^h(c) \triangleq \left\{(s, a, s') \middle| \max_\pi \frac{\widetilde{d}_m^h(s, a; \pi, c)}{SA + \eta_m \cdot \widehat{\mathrm{reg}}_{m-1}(\pi, c)} \cdot \widehat{P}_m^h(s'|s, a; c) \geq \frac{1}{\zeta_m}\right\}, \tag{3}$$

where $\zeta_m = \frac{\gamma_m}{8eH(H+1)^2}$. Notice that to define $\widetilde{d}_m^h(s; \pi, c)$ and $\widetilde{d}_m^h(s, a; \pi, c)$, we ***only*** *need $\{\widetilde{\mathcal{T}}_m^j(c), \widehat{P}_m^j(c)\}_{j \in [h-1]}$. Thus the two definitions are iteratively well-defined. Meanwhile, we also define the* ***observable occupancy measures*** *as the occupancy measures of the true model going through only the trusted transitions, i.e.,*

$$\begin{aligned} d_m^1(s; \pi, c) &= \mathbb{1}(s = s^1), \qquad d_m^1(s, a; \pi, c) = \mathbb{1}(s = s^1)\pi^1(s, a), \\ d_m^h(s; \pi, c) &:= \sum\nolimits_{s', a', s \in \widetilde{\mathcal{T}}_m^{h-1}(c)} d_m^{h-1}(s', a'; \pi, c)P_\star^{h-1}(s|s', a'; c), \\ d_m^h(s, a; \pi, c) &:= d_m^h(s; \pi, c)\pi_h(s, a). \end{aligned}$$

The computation of the set of trusted transitions need not enumerate all policies. The trusted transition set can be computed in $\mathrm{poly}(H, S, A, \log T)$ time by formulating it as a linear fractional programming problem. We defer the details to Appendix G.

Define the estimated occupancy measures $\widehat{d}_m^h(s; \pi, c) := \mathbb{E}^{\widehat{M}_m,\pi,c}[\mathbb{1}(s_1^h = s)]$ and $\widehat{d}_m^h(s, a; \pi, c) := \mathbb{E}^{\widehat{M}_m,\pi,c}[\mathbb{1}(s_1^h, a_1^h = s, a)]$. The trusted occupancy measure, though it eliminates rarely visited transitions, remains a valid estimate for all policies because the divergence between the estimated occupancy measure and itself is bounded. Specifically, we have the following lemma.

Lemma 3.3. *For all m, π, h, s, a, under any context c, we have*

$$\widehat{d}_m^h(s, a; \pi, c) - \widetilde{d}_m^h(s, a; \pi, c) \leq 32e\sqrt{H^4 S^2 A \cdot \mathcal{E}_m} + \widehat{\mathrm{reg}}_{m-1}(\pi; c)/(90HSA).$$

The next guarantee is the key to our analysis and is the most non-trivial guarantee of our construction. The following lemma states that if, for a context c, the Hellinger divergence between $\widehat{P}$ and $P_\star^h$ at layer h is small for all $h \in [H]$, then the trusted occupancy measure is upper bounded by a scaling of the observable occupancy measure.

Lemma 3.4. *For any m and c, assume for all $h \in [H]$,*

$$\mathbb{E}_{\pi \sim p_m^h(c)}\mathbb{E}^{M_\star,\pi,c}\left[D_{\mathrm{H}}^2\left(\widehat{P}_m^h(s_1^h, a_1^h; c), P_\star^h(s_1^h, a_1^h; c)\right)\right] \leq H/\gamma_m.$$

Then for the same m, c and all π, h, s, a, we have

$$\widetilde{d}_m^h(s, a; \pi, c) \leq (1 + 1/H)^{2(h-1)} d_m^h(s, a; \pi, c).$$

Since the trusted occupancy measure is upper bounded up to scaling by the observable occupancy measure, then the state-action pairs with large trusted occupancy measures are guaranteed to be visited often in the true dynamics as well. This enables more accurate planning and is thus crucial to our analysis.

Finally, we state the coverage guarantee achieved by the construction of Π^h_m and p^h_m. Concretely, we upper bound the trusted occupancy measure $\widetilde{d}^h_m(\cdot,\cdot;\pi,\cdot)$ of any policy π by the trusted occupancy measure induced by policy kernel p^h_m.

Lemma 3.5. *For any* m, π, c, h, s, a, *we have*

$$\begin{aligned}&\widetilde{d}^h_m(s,a;\pi,c)\cdot D_{\mathrm{H}}(P^h_\star(s,a;c),\widehat{P}^h_m(s,a;c))\\ &\qquad\le \frac{\gamma_m}{e^2H}\cdot p^h_m(c,\pi^{h,s,a}_{m,c})\widetilde{d}^h_m(s,a;\pi^{h,s,a}_{m,c},c)\cdot D^2_{\mathrm{H}}(P^h_\star(s,a;c),\widehat{P}^h_m(s,a;c))+\mathcal{E}'_m,\end{aligned}\tag{4}$$

where $\mathcal{E}'_m = \left(2e^2\sqrt{\frac{\mathcal{E}_m}{H^4S^2A}}+\frac{1}{720H^4S^3A^2}\widehat{\mathrm{reg}}_{m-1}(\pi,c)\right)\widetilde{d}^h_m(s,a;\pi,c)$ *,and* $p^h_m(c,\pi^{h,s,a}_{m,c})$ *is the probability of* $p^h_m(c)$ *on* $\pi^{h,s,a}_{m,c}$. *The guarantee Eq. (4) also holds replacing* $D_{\mathrm{H}}(P^h_\star(s,a;c),\widehat{P}^h_m(s,a;c))$ *with* $D_{\mathrm{H}}(R^h_\star(s,a;c),\widehat{R}^h_m(s,a;c))$ *on both sides of the inequality.*

4 Regret Analysis

In this section, we prodive a proof sketch of the regret analysis. Detailed proofs are deferred to Appendix E. We first aggregate the component-wise guarantees (Lemmas 3.1 to 3.5) from Section 3 to present the following epoch-wise guarantee.

Lemma 4.1. *For any* m, *any policies* $\{\pi_c\}_{c\in\mathcal{C}}$, *and* $\delta\in(0,1/2)$, *with probability at least* $1-\delta/M$,

$$\mathbb{E}_{c\sim\mathcal{D}}\left[\left|\widehat{V}^1_m(\pi_c,c)-V^1_\star(\pi_c,c)\right|\right]\le\frac{1}{20}\mathbb{E}_{c\sim\mathcal{D}}\left[\widehat{\mathrm{reg}}_{m-1}(\pi_c,c)\right]+77e\sqrt{H^6S^4A^3\cdot\mathcal{E}_m}.$$

Proof sketch of Lemma 4.1. For simplicity, in this proof sketch, we assume the true reward distribution is known[3]. For this, we first apply the celebrated local simulation lemma (Lemma C.2) in reinforcement learning to relate the divergence of the value functions to the stepwise divergences as the following. Under any context c,

$$\left|\widehat{V}^1_m(\pi_c,c)-V^1_\star(\pi_c,c)\right|\le\sum\nolimits_{h,s,a}\widehat{d}^h_m(s,a;\pi_c,c)D_{\mathrm{H}}\left(\widehat{P}^h_m(s,a;c),P^h_\star(s,a;c)\right).$$

Then we can exchange the estimated occupancy measure $\widehat{d}^h_m(s,a;\pi_c,c)$ by the trusted occupancy measure through Lemma 3.3, that is,

$$\begin{aligned}\widehat{d}^h_m(s,a;\pi_c,c)D_{\mathrm{H}}\left(\widehat{P}^h_m(s,a;c),P^h_\star(s,a;c)\right)&\le\widetilde{d}^h_m(s,a;\pi_c,c)D_{\mathrm{H}}\left(\widehat{P}^h_m(s,a;c),P^h_\star(s,a;c)\right)\\&\quad+32e\sqrt{H^4S^2A\cdot\mathcal{E}_m}+\widehat{\mathrm{reg}}_{m-1}(\pi;c)/(90HSA).\end{aligned}$$

Then by coverage guarantee Lemma 3.5, for any h,s,a, we can bound

$$\begin{aligned}&\widetilde{d}^h_m(s,a;\pi_c,c)D_{\mathrm{H}}\left(\widehat{P}^h_m(s,a;c),P^h_\star(s,a;c)\right)\\&\qquad\le\frac{\gamma_m}{e^2H}\cdot p^h_m(c,\pi^{h,s,a}_{m,c})\widetilde{d}^h_m(s,a;\pi^{h,s,a}_{m,c},c)\cdot D^2_{\mathrm{H}}(P^h_\star(s,a;c),\widehat{P}^h_m(s,a;c))+\mathcal{E}'_m.\end{aligned}$$

If the assumption in Lemma 3.4 is satisfied, then by Lemma 3.4 and the definition of p^h_m, we have

$$\begin{aligned}&\sum_{h,s,a}\frac{\gamma_m}{e^2H}\cdot p^h_m(c,\pi^{h,s,a}_{m,c})\widetilde{d}^h_m(s,a;\pi^{h,s,a}_{m,c},c)\cdot D^2_{\mathrm{H}}(P^h_\star(s,a;c),\widehat{P}^h_m(s,a;c))\\&\quad\le\sum_{h,s,a}\frac{\gamma_m}{H}\cdot p^h_m(c,\pi^{h,s,a}_{m,c})d^h_m(s,a;\pi^{h,s,a}_{m,c},c)\cdot D^2_{\mathrm{H}}(P^h_\star(s,a;c),\widehat{P}^h_m(s,a;c))\\&\quad\le\frac{\gamma_m}{H}\cdot\sum_h\mathbb{E}_{\pi\sim p^h_m(c)}\left[\mathbb{E}^{M_\star,\pi,c}\left[D^2_{\mathrm{H}}\left(\widehat{P}^h_m(s^h_1,a^h_1;c),P^h_\star(s^h_1,a^h_1;c)\right)\right]\right].\end{aligned}$$

[3] The uncertainty in reward distribution is not the main hardness in this problem. Specifically, our proof extends to offline oracles with squared loss guarantees between the mean rewards for divergence between the reward distributions as in [8].

If the assumptions in Lemma 3.4 are not satisfied, we have similar control as well (see the full proof Appendix E for details). Altogether with taking expectation on c, by the offline density estimation bound from Lemma 3.2, we have

$$\mathbb{E}_{c\sim\mathcal{D}}\left[\left|\widehat{V}_m^1(\pi_c, c) - V_\star^1(\pi_c, c)\right|\right] \le \mathbb{E}\big[\widehat{\mathrm{reg}}_{m-1}(\pi_c, c)\big]/40 + 39e\sqrt{H^6 S^4 A^3 \cdot \mathcal{E}_m}. \qquad \square$$

A revised version of the regret analysis (Lemma E.1) in Simchi-Levi and Xu [36], which relates the epoch-wise guarantee to the regret estimation error, can be found in Appendix E. Combining Lemmas 4.1 and E.1, we arrive at the following general regret guarantee.

Theorem 4.1. *The outputs $\{\pi_t\}_{t\in[T]}$ of Algorithm 1 satisfies with probability at least $1-\delta$ that*

$$\mathrm{Reg}(T) \lesssim \sum\nolimits_{m=1}^{N} (\tau_m - \tau_{m-1}) \cdot \sqrt{H^8 S^4 A^3 \cdot \mathcal{E}_m}$$

for $\delta \in (0, 1/2)$.

5 Extension: Reward-free Reinforcement Learning for CMDPs

In this section, we introduce the application of Algorithm 1 in the task of reward-free reinforcement learning in (stochastic) CMDPs. All proofs in this section will be deferred to Appendix F.

Reward-free reinforcement learning [22]. Reward-free reinforcement learning aims to efficiently explore the environment without relying on observed rewards. By doing so, it aims to enable the computation of a nearly optimal policy for any given reward function, utilizing only the trajectory data collected during exploration and without needing further interaction with the environment. This approach holds particular significance in scenarios where reward functions are refined over multiple iterations to encourage specific behaviors through trial and error, such as in constrained RL formulations. In such cases, repeatedly applying the same RL algorithm with varying reward functions can prove to be highly inefficient regarding sample usage, underscoring the efficiency of reward-free reinforcement learning.

Problem formulation. The major differences between the regret minimization setting (Section 2.1) and the reward-free reinforcement learning are that in the latter, no reward signals are observed during the interaction, and the goal of the latter is to output a CMDP prediction $\widehat{M}$ whose value functions are close to the underlying true CMDP $M_\star$ for any reward distributions. To accommodate such a change, for any model $M = \{P_M^h, R_M^h\}_{h\in[H]}$ and reward distribution $R = \{R^h\}_{h\in[H]}$, we define $M(\cdot; R) = \{P_M^h, R^h\}_{h\in[H]} = \{P_M^h(c), R^h(c)\}_{h\in[H], c\in\mathcal{C}}$ to be model M with the reward distribution part replaced by R. Thus in the reward-free reinforcement learning setting, the underlying true model satisfies $M_\star = M_\star(\cdot; 0)$ where 0 is used to denote the reward distribution that is constantly 0.

For any model M reward distribution R, context c and policy π, we use $M(\pi, c; R)$ to denote the distribution of the trajectory $c_1, \pi_1, s_1^1, a_1^1, r_1^1, \ldots, s_1^H, a_1^H, r_1^H$ given $M_\star = M(\cdot; R)$, $c_1 = c$, and $\pi_1 = \pi$. Also denote the probability and the expectation under $M(\pi, c; R)$ to be $\mathbb{P}^{M,\pi,c,R}(\cdot)$ and $\mathbb{E}^{M,\pi,c,R}[\cdot]$ respectively. Given reward distribution R, any policy π, state s and action a, we define the action value function $Q_\star^h(s, a; \pi, c, R)$ at layer h and the value function $V_\star^1(s; \pi, c, R)$ at layer h under context c and policy π as

$$Q_\star^h(s, a; \pi, c, R) = \sum_{j=h}^{H} \mathbb{E}^{M_\star,\pi,c,R}[r_1^j \mid s_1^h, a_1^h = s, a] \;\text{ and }\; V_\star^h(s; \pi, c, R) = \max_{a\in\mathcal{A}} Q_\star^h(s, a; \pi, c, R).$$

We denote the optimal policy with reward distribution R under context c as $\pi_{\star,c,R}$ and abbreviate its value function as $V_\star^h(\cdot; c, R)$. For $h = 1$, we denote $V_\star^1(c, R) = V_\star^1(s^1; c, R)$ and $V_\star^1(\pi, c, R) = V_\star^1(s^1; \pi, c, R)$. We also denote V_M^h as the value functions when $M_\star = M$.

Assumption 5.1 (Realizability for reward-free RL)**.** *Suppose the learner is given a model class $\mathcal{M}$ that contains the underlying true model $M_\star$. Assume all models $M \in \mathcal{M}$ have 0 reward.*

For a given $\varepsilon, \delta > 0$ and a model class $\mathcal{M}$, the goal of the learner is to output a model $\widehat{M}$ at the end of the interaction such that for any reward distribution R and set of policies $\{\pi_c\}_{c\in\mathcal{C}}$, the model satisfies

$$\mathbb{E}_{c\sim\mathcal{D}}\left[\left|V_\star^1(\pi_c, c, R) - V_{\widehat{M}}^1(\pi_c, c, R)\right|\right] \le \varepsilon \tag{5}$$

with probability at least $1-\delta$. An algorithm that achieves this objective is called (ε, δ)-learns the model class $\mathcal{M}$. Then we have the following guarantee from Algorithm 1.

Theorem 5.1. *If we choose $\tau_1 = T/(2H)$ and $\tau_2 = T/H$, the outputs $\widehat{M}_2$ of Algorithm 1 satisfies the reward-free objective Eq. (5) with probability at least $1-\delta$, with T at most bounded by*

$$T \le O\big(H^7 S^4 A^3 \log(|\mathcal{M}|/\delta)/\varepsilon^2\big)$$

for $\delta \in (0, 1/2)$. Moreover, the algorithm requires $O(H)$ number of oracle calls to the $\mathsf{MLE}_{\mathcal{M}}$ oracle.

The proof follows a similar argument of Lemma 4.1. In addition, we have a matching lower bound up to a $\mathrm{poly}(H, S, A)$ factor adapted from the non-contextual lower bound from Jin et al. [22].

Theorem 5.2. *Fix $\varepsilon \le 1$, $\delta \le 1/2$, $H, A \ge 2$. Suppose $S \ge L \log A$ for a large enough universal constant L and $K \ge 0$ large enough. Then, there exists a CMDP class $\mathcal{M}$ with $|\mathcal{S}| = S$, $|\mathcal{A}| = A$, $|\mathcal{M}| = K$, and horizon H and a distribution μ on $\mathcal{M}$ such that any algorithm Alg that $(\varepsilon/24, \delta)$-learns the class $\mathcal{M}$ satisfies $\mathbb{E}_{M\sim\mu}\mathbb{E}_{M,\mathsf{Alg}}[T] \gtrsim \log|\mathcal{M}|/\varepsilon^2$, where T is the number of trajectories required by the algorithm Alg to achieve $(\varepsilon/24, \delta)$ accuracy and $\mathbb{E}_{M,\mathsf{Alg}}[\cdot]$ is the expectation under the interaction between the algorithm Alg and model M.*

6 Discussion

Extension to low rank CMDPs. Low rank MDPs represent a significant extension to tabular MDPs, as explored in various studies [6, 28, 21, 4]. Linear MDPs are typically the first step beyond tabular MDPs. Extending our approach to linear CMDPs would be a substantial achievement. The primary challenge lies in identifying the trusted transitions within linear CMDPs. The current construction for tabular CMDPs does not readily apply here because it does not utilize the low-rank structure.

Extension to model-free learning. Our approach is model-based. However, model-free methods are often more practical for real-world applications. The main challenge lies in effectively balancing exploration and exploitation using only the value functions, as opposed to our method which depends on the occupancy measure.

More efficient oracles. In this paper, we focus on offline density estimation oracles due to the necessity of a small Hellinger distance between the estimated model and the true model for our approach. An offline regression oracle would only provide a 2-norm distance guarantee, which is inadequate for our purposes. Nevertheless, it is interesting to explore whether a reduction from CMDPs to offline regression could be feasible.

Acknowledgements

We thank Alexander Rakhlin, Dylan Foster, Yanjun Han, Zak Mhammedi, and anonymous reviewers for useful discussions. JQ acknowledges support from ARO through award W911NF-21-1-0328 and from the Simons Foundation and NSF through award DMS-2031883.

References

[1] Alekh Agarwal. Selective sampling algorithms for cost-sensitive multiclass prediction. In *International Conference on Machine Learning*, pages 1220–1228. PMLR, 2013.

[2] Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert E Schapire. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, 2012.

[3] 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, 2014.

[4] Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. FLAMBE: Structural complexity and representation learning of low rank MDPs. *Neural Information Processing Systems (NeurIPS)*, 2020.

[5] Philip Amortila, Dylan J. Foster, and Akshay Krishnamurthy. Scalable online exploration via coverability, 2024.

[6] Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1):33–57, 1996.

[7] Abraham Charnes and William W Cooper. Programming with linear fractional functionals. *Naval Research logistics quarterly*, 9(3-4):181–186, 1962.

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

[9] Jinglin Chen, Aditya Modi, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. On the statistical efficiency of reward-free exploration in non-linear rl. *Advances in Neural Information Processing Systems*, 35:20960–20973, 2022.

[10] Yuan Cheng, Ruiquan Huang, Jing Yang, and Yingbin Liang. Improved sample complexity for reward-free reinforcement learning under low-rank mdps. *arXiv preprint arXiv:2303.10859*, 2023.

[11] Junze Deng, Yuan Cheng, Shaofeng Zou, and Yingbin Liang. Sample complexity characterization for linear contextual mdps. *arXiv preprint arXiv:2402.02700*, 2024.

[12] Simon S Du, Yuping Luo, Ruosong Wang, and Hanrui Zhang. Provably efficient Q-learning with function approximation via distribution shift error checking oracle. In *Advances in Neural Information Processing Systems*, pages 8060–8070, 2019.

[13] Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 169–178. AUAI Press, 2011.

[14] Dylan J Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. *International Conference on Machine Learning (ICML)*, 2020.

[15] Dylan J Foster, Alekh Agarwal, Miroslav Dudík, Haipeng Luo, and Robert E. Schapire. Practical contextual bandits with regression oracles. *International Conference on Machine Learning*, 2018.

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

[17] Dylan J Foster, Alexander Rakhlin, Ayush Sekhari, and Karthik Sridharan. On the complexity of adversarial decision making. *Advances in Neural Information Processing Systems*, 35: 35404–35417, 2022.

[18] Dylan J Foster, Yanjun Han, Jian Qian, and Alexander Rakhlin. Online estimation via offline estimation: An information-theoretic framework. *arXiv preprint arXiv:2404.10122*, 2024.

[19] Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual markov decision processes, 2015.

[20] Pihe Hu, Yu Chen, and Longbo Huang. Towards minimax optimal reward-free reinforcement learning in linear mdps. In *The Eleventh International Conference on Learning Representations*, 2022.

[21] 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, 2017.

[22] Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR, 2020.

[23] Yin Tat Lee and Aaron Sidford. Solving linear programs with sqrt(rank) linear system solves. *arXiv preprint arXiv:1910.08033*, 2019.

[24] Orin Levy and Yishay Mansour. Optimism in face of a context: Regret guarantees for stochastic contextual mdp, 2023.

[25] Orin Levy, Asaf Cassel, Alon Cohen, and Yishay Mansour. Eluder-based regret for stochastic contextual mdps. *arXiv preprint arXiv:2211.14932*, 2022.

[26] Orin Levy, Alon Cohen, Asaf Cassel, and Yishay Mansour. Efficient rate optimal regret for adversarial contextual mdps using online function approximation. In *International Conference on Machine Learning*, pages 19287–19314. PMLR, 2023.

[27] Gen Li, Yuling Yan, Yuxin Chen, and Jianqing Fan. Optimal reward-agnostic exploration in reinforcement learning. 2023.

[28] Francisco S Melo and M Isabel Ribeiro. Q-learning with linear function approximation. In *International Conference on Computational Learning Theory*, pages 308–322. Springer, 2007.

[29] Zakaria Mhammedi, Dylan J Foster, and Alexander Rakhlin. Representation learning with multi-step inverse kinematics: An efficient and optimal approach to rich-observation rl. In *International Conference on Machine Learning*, pages 24659–24700. PMLR, 2023.

[30] Zakaria Mhammedi, Adam Block, Dylan J. Foster, and Alexander Rakhlin. Efficient model-free exploration in low-rank mdps, 2024.

[31] Sobhan Miryoosefi and Chi Jin. A simple reward-free approach to constrained reinforcement learning. In *International Conference on Machine Learning*, pages 15666–15698. PMLR, 2022.

[32] Aditya Modi and Ambuj Tewari. Contextual markov decision processes using generalized linear models. *arXiv preprint arXiv:1903.06187*, 2019.

[33] 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.

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

[35] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[36] 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.

[37] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[38] Andrew Wagenmaker, Max Simchowitz, and Kevin Jamieson. Beyond no regret: Instance-dependent pac reinforcement learning, 2022.

[39] Ruosong Wang, Simon S Du, Lin Yang, and Russ R Salakhutdinov. On reward-free reinforcement learning with linear function approximation. *Advances in neural information processing systems*, 33:17816–17826, 2020.

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

[41] Yunbei Xu and Assaf Zeevi. Upper counterfactual confidence bounds: a new optimism principle for contextual bandits, 2024.

[42] Xuezhou Zhang, Yuda Song, Masatoshi Uehara, Mengdi Wang, Alekh Agarwal, and Wen Sun. Efficient reinforcement learning in block mdps: A model-free representation learning approach. In *International Conference on Machine Learning*, pages 26517–26547. PMLR, 2022.

[43] Zihan Zhang, Simon S Du, and Xiangyang Ji. Nearly minimax optimal reward-free reinforcement learning. *arXiv preprint arXiv:2010.05901*, 2020.

[44] Zihan Zhang, Yuhang Jiang, Yuan Zhou, and Xiangyang Ji. Near-optimal regret bounds for multi-batch reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 24586–24596, 2022.

A Notation

For any integer n, we use $[n]$ to denote the set $\{1, \dots, n\}$. For any set $\mathcal{X}$, we use $\Delta(\mathcal{X})$ to denote the set of all distributions on the set $\mathcal{X}$. We define $O(\cdot)$, $\Omega(\cdot)$, $\Theta(\cdot)$, $\widetilde{O}(\cdot)$, $\widetilde{\Omega}(\cdot)$, $\widetilde{\Theta}(\cdot)$ following standard non-asymptotic big-oh notation. We use the binary relation $x \lesssim y$ to indicate that $x \le O(y)$. $\mathbb{1}(\mathcal{E})$ is an indicator function of event $\mathcal{E}$.

B Related Works

Contextual bandits and contextual MDPs. The SquareCB [14] obtains optimal regret for contextual bandits with access to an online regression oracle. This is extended to the CMDPs by Foster et al. [16] with the E2D algorithm. However, the algorithm requires $O(T)$ called to an online density estimation oracle. Compared to our algorithm, it necessitates significantly more calls to an oracle that is harder to implement for a general model class $\mathcal{M}$. After the FALCON algorithm [36] establishes the reduction from contextual bandits to offline regression, Xu and Zeevi [41] proposed the UCCB algorithm, which is less oracle-efficient in terms of oracle calls but adopts the prevalent "optimism in the face of uncertainty" principle and is thus easier to generalize. More specifically, the UCCB algorithm is extended to CMDPs by Levy and Mansour [24], Deng et al. [11] with assumptions on the model class. The RM-UCDD algorithm proposed by Levy and Mansour [24] requires the model class to have a minimum reachability $p_{\min}$ to all states under any policy and the regret guarantees scale with $O(\mathrm{poly}(H, S, A) \cdot (1/p_{\min})\sqrt{T \log |\mathcal{M}|})$. This assumption precludes model classes with small reachability, which frequently happens in practice [4]. The CMDP-VR algorithm proposed by Deng et al. [11] assumes a varying representation assumption on the model class instead. The assumption asserts that any model $M = \{P^h, R^h\}_{h\in[H]} \in \mathcal{M}$ satisfies for any c, h, s, a, s', $P^h(s'|s, a; c) = \langle \phi^h(s, a; c), \mu^h(s')\rangle$ for the known feature vector $\phi^h(s, a; c) \in \mathbb{R}^d$ and an unknown vector $\mu^h(s') \in \mathbb{R}^d$ which does not depend on the context c. This assumption is stringent because canonically, the feature vector for CMDPs will be chosen to be the unit vector indexed by s, a in $\mathbb{R}^{SA}$, i.e., $\phi^h(s, a; c) = e_{s,a} \in \mathbb{R}^{SA}$. Then, the requirement of $\mu^h(s')$ not depending on the context forces P^h to not depend on the context as well. This reduces the CMDP to an MDP. While it is possible to complicate the feature vector to not reduce to an MDP, this would result in a higher dimension in the feature vector space, which will be reflected in the regret bounds obtained $(\widetilde{O}(\mathrm{poly}(H, d)\sqrt{T \log |\mathcal{M}|})$. Another significant disadvantage compared with our algorithm is that the RM-UCDD and CMDP-VR algorithm requires $O(T)$ number of oracle calls. Other structural assumptions on the model class, such as a generalized linear structure or bounded eluder dimension, have been explored by Modi and Tewari [32], Levy et al. [25].

Reward-free reinforcement learning. Reward-free reinforcement learning aims to efficiently explore the environment without relying on observed rewards. By doing so, it aims to enable the computation of a nearly optimal policy for any given reward function, utilizing only the trajectory data collected during exploration and without needing further interaction with the environment. This framework is proposed by [22] and has been extensively studied for MDPs with various assumptions [43, 4, 39, 42, 9, 31, 38, 20, 10, 27, 30, 5]. However, to the best of our knowledge, we are the first to study the reward-free reinforcement learning setting for stochastic CMDPs. We provide a near-optimal sample complexity upper bound and a matching lower bound up to a $\mathrm{poly}(H, S, A)$ factor with only $O(H)$ number of oracle calls. Nevertheless, the upper bound is obtained by adjusting the exploration-exploitation, highlighting the flexibility of our algorithm.

C Technical Tools

C.1 Maximum Likelihood Estimation for Density Estimation

Example C.1 (MLE for finite model class)**.** Let $\mathcal{M}$ be a finite model class and the MLE estimator $\widehat{M}$ be defined by

$$\widehat{M} = \arg\max_{M\in\mathcal{M}} \prod_{i=1}^{n} \mathbb{P}^{M,\pi_i,c_i}\left(\{s_i^h, a_i^h, r_i^h\}_{h\in[H]}\right).$$

For any $\delta \in (0, 1/2)$, we have with probability at least $1 - \delta$,

$$\mathbb{E}_{c\sim\mathcal{D},\pi\sim p(c)}\left[D^2_{\mathrm{H}}\left(\widehat{M}(\pi, c), M_\star(\pi, c)\right)\right] \lesssim \frac{\log(|\mathcal{M}|/\delta)}{n}.$$

◁

Proof of Example C.1. For any $\delta \in (0, 1/2)$, by Lemma C.2 of Foster et al. [18], we have with probability at least $1 - \delta/2$,

$$\sum_{i=1}^{n} D_{\mathrm{H}}^{2}\Big(\widehat{M}(\pi_i, c_i), M_\star(\pi_i, c_i)\Big) \lesssim \log(|\mathcal{M}|/\delta).$$

Then by Lemma A.3 of Foster et al. [16], we have with probability at least $1 - \delta/2$,

$$\mathbb{E}_{c\sim\mathcal{D},\pi\sim p(c)}\Big[D_{\mathrm{H}}^{2}\Big(\widehat{M}(\pi, c), M_\star(\pi, c)\Big)\Big] \lesssim \sum_{i=1}^{n} D_{\mathrm{H}}^{2}\Big(\widehat{M}(\pi_i, c_i), M_\star(\pi_i, c_i)\Big) + \log(|\mathcal{M}|/\delta).$$

Then by union bound, we obtain the desired result. □

C.2 Information Theory

Lemma C.1 (Lemma B.4 of Foster et al. [17]). *Let $\mathbb{P}$ and $\mathbb{Q}$ be two distributions on space χ. Let $h : \chi \to R$ be a function. Then we have:*

$$|\mathbb{E}_{\mathbb{P}}[h] - \mathbb{E}_{\mathbb{Q}}[h]| \leq \sqrt{2^{-1}(\mathbb{E}_{\mathbb{P}}[h^2] + \mathbb{E}_{\mathbb{Q}}[h^2])D_H^2(\mathbb{P}, \mathbb{Q})}.$$

C.3 Reinforcement Learning

Lemma C.2 (Lemma F.3 of [16]). *Let $M = \{P_M^h, R_M^h\}_{h\in[H]}$ and $\overline{M} = \{P_{\overline{M}}^h, R_{\overline{M}}^h\}_{h\in[H]}$ be two CMDPs. For any policy π and context c, we have*

$$\begin{aligned}
&V_M^1(\pi, c) - V_{\overline{M}}^1(\pi, c) \\
&= \sum_{h=1}^{H} \mathbb{E}^{\overline{M},\pi,c}\big[\big(P_M^h(s_1^{h+1}|s_1^h, a_1^h; c) - P_{\overline{M}}^h(s_1^{h+1}|s_1^h, a_1^h; c)\big)V_M^{h+1}(s_1^{h+1}; \pi, c)\big] \\
&\quad + \sum_{h=1}^{H} \mathbb{E}^{\overline{M},\pi,c}\Big[\mathbb{E}_{r^h\sim R_M(s_1^h,a_1^h;c)}\big[r^h\big] - \mathbb{E}_{r^h\sim R_{\overline{M}}(s_1^h,a_1^h;c)}\big[r^h\big]\Big] \\
&\leq \sum_{h=1}^{H}\sum_{s,a} \mathbb{E}^{\overline{M},\pi,c}\big[\mathbb{1}(s_1^h, a_1^h = s, a)\big]\big(D_{\mathrm{H}}\big(P_M^h(s, a; c), P_{\overline{M}}^h(s, a; c)\big) + D_{\mathrm{H}}\big(R_M^h(s, a; c), R_{\overline{M}}^h(s, a; c)\big)\big).
\end{aligned}$$

D Proofs from Section 3

In this section, we present the proofs for Lemmas 3.1 to 3.5.

Proof of Lemma 3.1. We fix an arbitrary context c throughout the proof. Let $u(\lambda) := \sum_{\pi\in\Pi_m^h} 1/(\lambda + \eta_m \cdot \widehat{\mathrm{reg}}_{m-1}(\pi))$. Since for any $h \in [H]$, $\widehat{\pi}_{m-1,c} \in \Pi_m^h$, then $u(\lambda) \geq 1/(\lambda + \eta_m \cdot \widehat{\mathrm{reg}}_{m-1}(\widehat{\pi}_{m-1,c})) = 1/\lambda$. On the other hand, $u(\lambda) \leq (SA + 1)/\lambda$. Moreover, $u(\lambda)$ is clearly monotonically decreasing with $u(0) = \infty$ and $u(SA + 1) \leq 1$. Thus there exists $\lambda_{m,c}^h \in (0, SA + 1]$ such that $u(\lambda_{m,c}^h) = 1$ as we desired.

Now we have the regret is bounded by

$$\mathbb{E}_{\pi\sim p_m^h(c)}\big[\widehat{\mathrm{reg}}_{m-1}(\pi; c)\big] = \sum_{\pi\in\Pi_{m,c}^h} \frac{\widehat{\mathrm{reg}}_{m-1}(\pi; c)}{\lambda_{m,c}^h + \eta_m \cdot \widehat{\mathrm{reg}}_{m-1}(\pi; c)} \leq \sum_{\pi\in\Pi_m^h(c)} \frac{1}{\eta_m} \leq \frac{2SA}{\eta_m}.$$

Finally, we recall $\eta_m = \gamma_m/(720e^3H^5S^3A^2)$ and $\gamma_m = \sqrt{\frac{H^6S^4A^3}{\mathcal{E}_m}}$ plug this bound in, then we have

$$\mathbb{E}_{\pi\sim p_m^h(c)}\big[\widehat{\mathrm{reg}}_{m-1}(\pi; c)\big] \leq 1440e^3 \cdot \sqrt{H^4S^4A^3 \cdot \mathcal{E}_m}.$$

□

Proof of Lemma 3.2. By definition of Definition 2.1. □

Proof of Lemma 3.3. For any m, π, c, h, s, a, by the definition of $\widetilde{d}^h_m(s,a;\pi,c)$, we the difference between $\widehat{d}^h_m(s,a;\pi,c)$ and $\widetilde{d}^h_m(s,a;\pi,c)$ are the parts of occupancy measures that do not go through the trusted transitions, i.e.,

$$
\begin{aligned}
&\widehat{d}^h_m(s,a;\pi,c) - \widetilde{d}^h_m(s,a;\pi,c) \\
&\qquad = \sum_{j=1}^{h-1} \sum_{(s^j,a^j,s^{j+1}) \notin \widetilde{\mathcal{T}}^j_m(c)} \widetilde{d}^j_m(s^j,a^j;\pi,c)\widehat{P}^j_m(s^{j+1}|s^j,a^j;c)\widehat{P}^{j+1:h}_{m+1}(s|s^{j+1};\pi,c)\pi^h(s,a),
\end{aligned}
$$

where $\widehat{P}^{j+1:h}_{m+1}(s|s^{j+1};\pi,c)$ is the estimated transition probability from s^{j+1} at step $j+1$ to s at step h under policy π and context c. Then since $(s^j,a^j,s^{j+1}) \notin \widetilde{\mathcal{T}}^j_m(c)$, we have

$$
\begin{aligned}
&\sum_{j=1}^{h-1} \sum_{(s^j,a^j,s^{j+1}) \notin \widetilde{\mathcal{T}}^j_m(c)} \widetilde{d}^j_m(s^j,a^j;\pi,c)\widehat{P}^j_m(s^{j+1}|s^j,a^j;c)\widehat{P}^{j+1:h}_{m+1}(s|s^{j+1};\pi,c)\pi^h(s,a) \\
&\le \sum_{j=1}^{h-1} \sum_{(s^j,a^j,s^{j+1}) \notin \widetilde{\mathcal{T}}^j_m(c)} \widetilde{d}^j_m(s^j,a^j;\pi,c)\widehat{P}^j_m(s^{j+1}|s^j,a^j;c) \\
&\le \sum_{j=1}^{h-1} \sum_{(s^j,a^j,s^{j+1}) \notin \widetilde{\mathcal{T}}^j_m(c)} \frac{SA + \eta_m \widehat{\mathrm{reg}}_{m-1}(\pi;c)}{\zeta_m} \\
&\le \frac{hS^2A(SA + \eta_m \widehat{\mathrm{reg}}_{m-1}(\pi;c))}{\zeta_m}.
\end{aligned}
$$

Recall the choice of $\zeta_m = \frac{\gamma_m}{8eH(H+1)^2}$, $\eta_m = \frac{\gamma_m}{720e^3H^5S^3A^2}$ and $\gamma_m = \sqrt{\frac{H^6S^4A^3}{\mathcal{E}_m}}$, we have

$$
\widehat{d}^h_m(s,a;\pi,c) - \widetilde{d}^h_m(s,a;\pi,c) \le 32e\sqrt{H^4S^2A \cdot \mathcal{E}_m} + \frac{1}{90HSA}\widehat{\mathrm{reg}}_{m-1}(\pi;c).
$$

□

Proof of Lemma 3.4. We prove iteratively between the objective

$$
\widetilde{d}^h_m(s,a;\pi,c) \le \left(1 + \frac{1}{H}\right)^{2(h-1)} d^h_m(s,a;\pi,c)
$$

and the following claim: For any h and any $(s,a,s') \in \widetilde{\mathcal{T}}^h_m(c)$, we have

$$
\widehat{P}^h_m(s'|s,a;c) \le \left(1 + \frac{1}{H}\right)^2 P^h_\star(s'|s,a;c).
$$

First, we show that for any $h \in [H]$,

$$
\begin{aligned}
&\forall \pi, s, a,\ \widetilde{d}^h_m(s,a;\pi,c) \le \left(1 + \frac{1}{H}\right)^{2(h-1)} d^h_m(s,a;\pi,c) \\
&\qquad\qquad \Rightarrow\ \forall (s,a,s') \in \widetilde{\mathcal{T}}^h_m(c),\ \widehat{P}^h_m(s'|s,a;c) \le \left(1 + \frac{1}{H}\right)^2 P^h_\star(s'|s,a;c). \qquad (6)
\end{aligned}
$$

For this, we note by Lemma C.1 that for any h, s, a, s',

$$
\begin{aligned}
&\widehat{P}^h_m(s'|s,a;c) \qquad (7) \\
&\le P^h_\star(s'|s,a;c) + \sqrt{2^{-1}(\widehat{P}^h_m(s'|s,a;c) + P^h_\star(s'|s,a;c))D^2_{\mathrm{H}}\Big(\mathrm{Ber}(\widehat{P}^h_m(s'|s,a;c)), \mathrm{Ber}(P^h_\star(s'|s,a;c))\Big)} \\
&\le P^h_\star(s'|s,a;c) + \sqrt{2^{-1}(\widehat{P}^h_m(s'|s,a;c) + P^h_\star(s'|s,a;c))D^2_{\mathrm{H}}\Big(\widehat{P}^h_m(s,a;c), P^h_\star(s,a;c)\Big)}, \qquad (8)
\end{aligned}
$$

where the second inequality is by data-processing inequality [34]. Then by AM-GM, we have

$$\sqrt{2^{-1}(\widehat{P}_m^h(s'|s,a;c)+P_\star^h(s'|s,a;c))D_{\mathrm{H}}^2\Big(\widehat{P}_m^h(s,a;c),P_\star^h(s,a;c)\Big)}$$
$$\le \frac{1}{4H}(\widehat{P}_m^h(s'|s,a;c)+P_\star^h(s'|s,a;c))+H\cdot D_{\mathrm{H}}^2\Big(\widehat{P}_m^h(s,a;c),P_\star^h(s,a;c)\Big).$$

Plug the above back into Eq. (8) and reorganize, we obtain

$$\widehat{P}_m^h(s'|s,a;c)\le\left(1+\frac{1}{H}\right)\cdot P_\star^h(s'|s,a;c)+(H+1)D_{\mathrm{H}}^2\Big(\widehat{P}_m^h(s,a;c),P_\star^h(s,a;c)\Big).$$

Then multiply both sides by $\widetilde{d}_m^h(s,a;\pi,c)$, we have

$$\widetilde{d}_m^h(s,a;\pi,c)\left(\widehat{P}_m^h(s'|s,a;c)-\left(1+\frac{1}{H}\right)P_\star^h(s'|s,a;c)\right)\le(H+1)\widetilde{d}_m^h(s,a;\pi,c)D_{\mathrm{H}}^2\Big(\widehat{P}_m^h(s,a;c),P_\star^h(s,a;c)\Big).$$

Meanwhile, by the definition of $\pi_{m,c}^{h,s,a}$, we have

$$\widetilde{d}_m^h(s,a;\pi,c)\le\frac{\widetilde{d}_m^h(s,a;\pi_{m,c}^{h,s,a},c)}{SA+\eta_m\cdot\widehat{\mathrm{reg}}_{m-1}(\pi_{m,c}^{h,s,a},c)}\cdot(SA+\eta_m\cdot\widehat{\mathrm{reg}}_{m-1}(\pi,c)).$$

Then by the induction hypothesis, we have

$$\frac{\widetilde{d}_m^h(s,a;\pi_{m,c}^{h,s,a},c)}{SA+\eta_m\cdot\widehat{\mathrm{reg}}_{m-1}(\pi_{m,c}^{h,s,a},c)}\cdot(SA+\eta_m\cdot\widehat{\mathrm{reg}}_{m-1}(\pi,c))$$
$$\le\frac{e^2d_m^h(s,a;\pi_{m,c}^{h,s,a},c)}{SA+\eta_m\cdot\widehat{\mathrm{reg}}_{m-1}(\pi_{m,c}^{h,s,a},c)}\cdot(SA+\eta_m\cdot\widehat{\mathrm{reg}}_{m-1}(\pi,c)).$$

Thus we have futher by the definition of $p_m^h(c)$ and the assumption that $\mathbb{E}_{\pi\sim p_m^h(c)}\mathbb{E}^{M_\star,\pi,c}\Big[D_{\mathrm{H}}^2\Big(\widehat{P}_m^h(s_1^h,a_1^h;c),P_\star^h(s_1^h,a_1^h;c)\Big)\Big]\le H/\gamma_m$,

$$\begin{aligned}&(H+1)\widetilde{d}_m^h(s,a;\pi,c)D_{\mathrm{H}}^2\Big(\widehat{P}_m^h(s,a;c),P_\star^h(s,a;c)\Big)\\&\quad\le e^2(H+1)(SA+\eta_m\cdot\widehat{\mathrm{reg}}_{m-1}(\pi,c))\frac{d_m^h(s,a;\pi_{m,c}^{h,s,a},c)}{SA+\eta_m\cdot\widehat{\mathrm{reg}}_{m-1}(\pi_{m,c}^{h,s,a},c)}\cdot D_{\mathrm{H}}^2\Big(\widehat{P}_m^h(s,a;c),P_\star^h(s,a;c)\Big)\\&\quad\le e^2(H+1)(SA+\eta_m\cdot\widehat{\mathrm{reg}}_{m-1}(\pi,c))p_m^h(c,\pi_{m,c}^{h,s,a})d_m^h(s,a;\pi_{m,c}^{h,s,a},c)D_{\mathrm{H}}^2\Big(P_\star^h(s,a;c),\widehat{P}_m^h(s,a;c)\Big)\\&\quad\le e^2(H+1)(SA+\eta_m\cdot\widehat{\mathrm{reg}}_{m-1}(\pi,c))\mathbb{E}_{\pi\sim p_m^h(c)}\mathbb{E}^{M_\star,\pi,c}\Big[D_{\mathrm{H}}^2\Big(\widehat{P}_m^h(s_1^h,a_1^h;c),P_\star^h(s_1^h,a_1^h;c)\Big)\Big]\\&\quad\le\frac{e^2H(H+1)}{\gamma_m}(SA+\eta_m\cdot\widehat{\mathrm{reg}}_{m-1}(\pi,c))\\&\quad=\frac{1}{(H+1)\zeta_m}(SA+\eta_m\cdot\widehat{\mathrm{reg}}_{m-1}(\pi,c)),\end{aligned}$$

where the last equality is by the definition of ζ_m. In all, we have shown that

$$\widetilde{d}_m^h(s,a;\pi,c)\left(\widehat{P}_m^h(s'|s,a;c)-\left(1+\frac{1}{H}\right)P_\star^h(s'|s,a;c)\right)\le\frac{1}{(H+1)\zeta_m}(SA+\eta_m\cdot\widehat{\mathrm{reg}}_{m-1}(\pi,c)).$$

Now we prove by contradiction, if for any $(s,a,s')\in\widetilde{\mathcal{T}}_m^h(c)$, the reverse inequality is true, i.e., $\widehat{P}_m^h(s'|s,a;c)>\left(1+\frac{1}{H}\right)^2P_\star^h(s'|s,a;c)$. Then for any π,

$$\frac{1}{H+1}\widetilde{d}_m^h(s,a;\pi,c)\widehat{P}_m^h(s'|s,a;c)<\widetilde{d}_m^h(s,a;\pi,c)\left(\widehat{P}_m^h(s'|s,a;c)-\left(1+\frac{1}{H}\right)P_\star^h(s'|s,a;c)\right)$$
$$\le\frac{1}{(H+1)\zeta_m}(SA+\eta_m\cdot\widehat{\mathrm{reg}}_{m-1}(\pi,c)).$$

This contradicts the definition of $\widetilde{\mathcal{T}}_m^h(c)$.

For the other direction of Eq. (6), we prove for all $h \in [H]$,

$$\begin{cases} \widehat{P}_m^h(s'|s,a;c) \le \left(1+\frac{1}{H}\right)^2 P_\star^h(s'|s,a;c) & \forall (s,a,s') \in \widetilde{\mathcal{T}}_m^h, \\ \widetilde{d}_m^h(s,a;\pi,c) \le \left(1+\frac{1}{H}\right)^{2(h-1)} d_m^h(s,a;\pi,c) & \forall \pi, s, a. \end{cases}$$

$$\Rightarrow \quad \widetilde{d}_m^{h+1}(s,a;\pi,c) \le \left(1+\frac{1}{H}\right)^{2h} d_m^{h+1}(s,a;\pi,c), \;\; \forall \pi, s, a. \tag{9}$$

This direction is straightforward since

$$\begin{aligned} \widetilde{d}_m^{h+1}(s,a;\pi,c) &= \sum_{(s',a',s)\in\widetilde{\mathcal{T}}_m^h} \widetilde{d}_m^h(s,a;\pi,c)\widehat{P}_m^h(s'|s,a;c)\pi^{h+1}(s,a;c) \\ &\le \left(1+\frac{1}{H}\right)^{2h} \sum_{(s',a',s)\in\widetilde{\mathcal{T}}_m^h} d_m^h(s,a;\pi,c)P_\star^h(s'|s,a;c)\pi_{h+1}(s,a;c) \\ &= \left(1+\frac{1}{H}\right)^{2h} d_m^{h+1}(s,a;\pi,c). \end{aligned}$$

With the two derivations Eq. (6) and Eq. (9), along with the fact that the initial argument of Eq. (6) holds by defintion for $h = 1$. Thus we conclude the proof.

□

Proof of Lemma 3.5. For any m, π, c, h, s, a, by AM-GM, we have

$$\begin{aligned} &\widetilde{d}_m^h(s,a;\pi,c) D_{\mathrm{H}}\left(P_\star^h(s,a;c), \widehat{P}_m^h(s,a;c)\right) \\ &\qquad \le \frac{2e^2H(SA+\eta_m\widehat{\mathrm{reg}}_{m-1}(\pi_{m,c}^{h,s,a};c))}{\gamma_m\widetilde{d}_m^h(s,a;\pi_{m,c}^{h,s,a},c)} \cdot (\widetilde{d}_m^h(s,a;\pi,c))^2 \\ &\qquad\quad + \frac{\gamma_m\widetilde{d}_m^h(s,a;\pi_{m,c}^{h,s,a};c)}{2e^2H(SA+\eta_m\widehat{\mathrm{reg}}_{m-1}(\pi_{m,c}^{h,s,a};c))} \cdot D_{\mathrm{H}}^2\left(P_\star^h(s,a;c), \widehat{P}_m^h(s,a;c)\right). \end{aligned} \tag{10}$$

By the definition of $\pi_{m,c}^{h,s,a}$, we have futher

$$\begin{aligned} &\frac{2e^2H(SA+\eta_m\widehat{\mathrm{reg}}_{m-1}(\pi_{m,c}^{h,s,a};c))}{\gamma_m\widetilde{d}_m^h(s,a;\pi_{m,c}^{h,s,a},c)} \cdot (\widetilde{d}_m^h(s,a;\pi,c))^2 \\ &\qquad \le \frac{2e^2H(SA+\eta_m\widehat{\mathrm{reg}}_{m-1}(\pi;c))}{\gamma_m\widetilde{d}_m^h(s,a;\pi,c)} \cdot (\widetilde{d}_m^h(s,a;\pi,c))^2 \\ &\qquad = \frac{2e^2HSA+2e^2H\eta_m\widehat{\mathrm{reg}}_{m-1}(\pi;c)}{\gamma_m} \cdot \widetilde{d}_m^h(s,a;\pi,c). \end{aligned} \tag{11}$$

Recall the choice of $\eta_m = \gamma_m/(720e^3H^5S^3A^2)$ and $\gamma_m = \sqrt{\frac{H^6S^4A^3}{\mathcal{E}_m}}$, we have

$$\frac{2e^2HSA+2e^2H\eta_m\widehat{\mathrm{reg}}_{m-1}(\pi;c)}{\gamma_m} \le 2e^2\sqrt{\frac{\mathcal{E}_m}{H^4S^2A}} + \frac{1}{720H^4S^3A^2}\widehat{\mathrm{reg}}_m(\pi,c). \tag{12}$$

Also by the definition of $p_m^h(c)$, we have

$$\begin{aligned} &\frac{\gamma_m\widetilde{d}_m^h(s,a;\pi_{m,c}^{h,s,a};c)}{2e^2H(SA+\eta_m\widehat{\mathrm{reg}}_{m-1}(\pi_{m,c}^{h,s,a};c))} \cdot D_{\mathrm{H}}^2\left(P_\star^h(s,a;c), \widehat{P}_m^h(s,a;c)\right) \\ &\qquad \le \frac{\gamma_m}{e^2H} \cdot p_m^h(c,\pi_{m,c}^{h,s,a})\widetilde{d}_m^h(s,a;\pi_{m,c}^{h,s,a},c)D_{\mathrm{H}}^2\left(P_\star^h(s,a;c), \widehat{P}_m^h(s,a;c)\right). \end{aligned} \tag{13}$$

Now we plug Eqs. (11) to (13) back into Eq. (10) to obtain that

$$\begin{aligned}
&\widetilde{d}^h_m(s,a;\pi,c) D_{\mathrm{H}}\Big(P^h_\star(s,a;c), \widehat{P}^h_m(s,a;c)\Big) \\
&\qquad\le \left(2e^2\sqrt{\frac{\mathcal{E}_m}{H^4S^2A}} + \frac{1}{720H^4S^3A^2}\widehat{\mathrm{reg}}_{m-1}(\pi,c)\right)\widetilde{d}^h_m(s,a;\pi,c) \\
&\qquad\quad + \frac{\gamma_m}{e^2H}\cdot p^h_m(c,\pi^{h,s,a}_{m,c})\widetilde{d}^h_m(s,a;\pi^{h,s,a}_{m,c},c) D^2_{\mathrm{H}}\Big(P^h_\star(s,a;c), \widehat{P}^h_m(s,a;c)\Big).
\end{aligned}$$

Similar bounds can be obtained replacing $D_{\mathrm{H}}\Big(P^h_\star(s,a;c), \widehat{P}^h_m(s,a;c)\Big)$ with $D_{\mathrm{H}}\Big(R^h_\star(s,a;c), \widehat{R}^h_m(s,a;c)\Big)$. □

E Proofs from Section 4

Proof of Lemma 4.1. For this we first apply the local simulation lemma (Lemma C.2) in reinforcement learning to relate the divergence of the value functions to the stepwise divergences as the following. Under any context c,

$$\left|\widehat{V}^1_m(\pi_c,c) - V^1_\star(\pi_c,c)\right| \le \sum_{h,s,a}\widehat{d}^h_m(s,a;\pi_c,c)\Big(D_{\mathrm{H}}\Big(\widehat{P}^h_m(s,a;c), P^h_\star(s,a;c)\Big) + D_{\mathrm{H}}\Big(\widehat{R}^h_m(s,a;c), R^h_\star(s,a;c)\Big)\Big).$$

Then we can exchange the estimated occupancy measure $\widehat{d}^h_m(s,a;\pi_c,c)$ by the trusted occupancy measure through Lemma 3.3, that is,

$$\begin{aligned}
&\sum_{h,s,a}\widehat{d}^h_m(s,a;\pi_c,c) D_{\mathrm{H}}\Big(\widehat{P}^h_m(s,a;c), P^h_\star(s,a;c)\Big) \\
&\qquad\qquad \le \sum_{h,s,a}\widetilde{d}^h_m(s,a;\pi_c,c) D_{\mathrm{H}}\Big(\widehat{P}^h_m(s,a;c), P^h_\star(s,a;c)\Big) \\
&\qquad\qquad\quad + \sum_{h,s,a}\left(32e\sqrt{H^4S^2A\cdot\mathcal{E}_m} + \frac{1}{90HSA}\widehat{\mathrm{reg}}_{m-1}(\pi;c)\right) \\
&\qquad\qquad \le \sum_{h,s,a}\widetilde{d}^h_m(s,a;\pi_c,c) D_{\mathrm{H}}\Big(\widehat{P}^h_m(s,a;c), P^h_\star(s,a;c)\Big) \\
&\qquad\qquad\quad + 32e\sqrt{H^6S^4A^3\cdot\mathcal{E}_m} + \frac{1}{90}\widehat{\mathrm{reg}}_{m-1}(\pi_c;c).
\end{aligned} \tag{14}$$

Then by coverage guarantee Lemma 3.5, for any h, s, a, we can bound

$$\begin{aligned}
&\sum_{h,s,a}\widetilde{d}^h_m(s,a;\pi_c,c) D_{\mathrm{H}}\Big(\widehat{P}^h_m(s,a;c), P^h_\star(s,a;c)\Big) \\
&\qquad \le \sum_{h,s,a}\frac{\gamma_m}{e^2H}\cdot p^h_m(c,\pi^{h,s,a}_{m,c})\widetilde{d}^h_m(s,a;\pi^{h,s,a}_{m,c},c)\cdot D^2_{\mathrm{H}}(P^h_\star(s,a;c), \widehat{P}^h_m(s,a;c)) \\
&\qquad\quad + \sum_{h,s,a}\left(2e^2\sqrt{\frac{\mathcal{E}_m}{H^4S^2A}} + \frac{1}{720H^4S^3A^2}\widehat{\mathrm{reg}}_{m-1}(\pi,c)\right)\widetilde{d}^h_m(s,a;\pi,c) \\
&\qquad \le \sum_{h,s,a}\frac{\gamma_m}{e^2H}\cdot p^h_m(c,\pi^{h,s,a}_{m,c})\widetilde{d}^h_m(s,a;\pi^{h,s,a}_{m,c},c)\cdot D^2_{\mathrm{H}}(P^h_\star(s,a;c), \widehat{P}^h_m(s,a;c)) \\
&\qquad\quad + 2e^2\sqrt{\frac{\mathcal{E}_m}{H^2A}} + \frac{1}{720H^3S^2A^2}\widehat{\mathrm{reg}}_{m-1}(\pi_c,c).
\end{aligned} \tag{15}$$

Suppose the assumption in Lemma 3.4 is satisfied, then by Lemma 3.4, we have

$$\begin{aligned}
&\sum_{h,s,a} \frac{\gamma_m}{e^2 H} \cdot p_m^h(c, \pi_{m,c}^{h,s,a}) \widetilde{d}_m^h(s, a; \pi_{m,c}^{h,s,a}, c) \cdot D_{\mathrm{H}}^2(P_\star^h(s, a; c), \widehat{P}_m^h(s, a; c)) \\
&\qquad \le \sum_{h,s,a} \frac{\gamma_m}{H} \cdot p_m^h(c, \pi_{m,c}^{h,s,a}) d_m^h(s, a; \pi_{m,c}^{h,s,a}, c) \cdot D_{\mathrm{H}}^2(P_\star^h(s, a; c), \widehat{P}_m^h(s, a; c)) \\
&\qquad \le \frac{\gamma_m}{H} \cdot \sum_h \mathbb{E}_{\pi \sim p_m^h(c)} \left[\mathbb{E}^{M_\star, \pi, c} \left[D_{\mathrm{H}}^2 \left(\widehat{P}_m^h(s_1^h, a_1^h; c), P_\star^h(s_1^h, a_1^h; c) \right) \right] \right]. \qquad (16)
\end{aligned}$$

The $D_{\mathrm{H}}\left(\widehat{P}_m^h(s_1^h, a_1^h; c), P_\star^h(s_1^h, a_1^h; c)\right)$ in Eqs. (15), (16) and (26) can be replaced with $D_{\mathrm{H}}\left(\widehat{R}_m^h(s_1^h, a_1^h; c), R_\star^h(s_1^h, a_1^h; c)\right)$ as well. If the assumption in Lemma 3.4 is not satisfied, then we there exists j such that

$$\mathbb{E}_{\pi \sim p_m^h(c)} \mathbb{E}^{M_\star, \pi, c} \left[D_{\mathrm{H}}^2 \left(\widehat{P}_m^h(s_1^j, a_1^j; c), P_\star^h(s_1^j, a_1^j; c) \right) \right] > H / \gamma_m.$$

This implies

$$\left| \widehat{V}_m^1(\pi_c, c) - V_\star^1(\pi_c, c) \right| \le 1 \le \frac{\gamma_m}{H} \cdot \mathbb{E}_{\pi \sim p_m^h(c)} \mathbb{E}^{M_\star, \pi, c} \left[D_{\mathrm{H}}^2 \left(\widehat{P}_m^h(s_1^j, a_1^j; c), P_\star^h(s_1^j, a_1^j; c) \right) \right].$$

Thus, altogether, no matter the assumption in Lemma 3.4 is satisfied or not, we have shown

$$\begin{aligned}
\left| \widehat{V}_m^1(\pi_c, c) - V_\star^1(\pi_c, c) \right| &\le 76e \sqrt{H^6 S^4 A^3 \cdot \mathcal{E}_m} + \frac{1}{20} \widehat{\mathrm{reg}}_{m-1}(\pi_c; c) \\
&\quad + \frac{2\gamma_m}{H} \cdot \sum_h \mathbb{E}_{\pi \sim p_m^h(c)} \left[\mathbb{E}^{M_\star, \pi, c} \left[D_{\mathrm{H}}^2 \left(\widehat{P}_m^h(s_1^h, a_1^h; c), P_\star^h(s_1^h, a_1^h; c) \right) \right] \right].
\end{aligned}$$

Taking expectation on c, we have

$$\begin{aligned}
&\mathbb{E}_{c \sim \mathcal{D}} \left[\left| \widehat{V}_m^1(\pi_c, c) - V_\star^1(\pi_c, c) \right| \right] \\
&\quad \le \frac{1}{20} \mathbb{E}_{c \sim \mathcal{D}} \left[\widehat{\mathrm{reg}}_{m-1}(\pi_c, c) \right] + 76e \sqrt{H^6 S^4 A^3 \cdot \mathcal{E}_m} \\
&\qquad + \frac{2\gamma_m}{H} \cdot \sum_h \mathbb{E}_{c \sim \mathcal{D}} \mathbb{E}_{\pi \sim p_m^h(c)} \left[\mathbb{E}^{M_\star, \pi, c} \left[D_{\mathrm{H}}^2 \left(\widehat{P}_m^h(s_1^h, a_1^h; c), P_\star^h(s_1^h, a_1^h; c) \right) \right] \right] \\
&\quad \le \frac{1}{20} \mathbb{E} \left[\widehat{\mathrm{reg}}_{m-1}(\pi_c, c) \right] + 76e \sqrt{H^6 S^4 A^3 \cdot \mathcal{E}_m} + \gamma_m \cdot \mathcal{E}_m \\
&\quad \le \frac{1}{20} \mathbb{E} \left[\widehat{\mathrm{reg}}_{m-1}(\pi_c, c) \right] + 77e \sqrt{H^6 S^4 A^3 \cdot \mathcal{E}_m},
\end{aligned}$$

where the second inequality is by the offline density estimation bound from Lemma 3.2 □

Lemma E.1 ([36]). *Let $\varepsilon_1, \ldots, \varepsilon_N$ be N positive values. Suppose for any $m > 0$ and an arbitrary policy set $\{\pi_c\}_{c \in \mathcal{C}}$, we have:*

$$\mathbb{E}_{c \sim \mathcal{D}} [|\widehat{V}_m^1(\pi_c, c) - V_\star^1(\pi_c, c)|] \le \frac{1}{20} \mathbb{E}_{c \sim \mathcal{D}} [\widehat{\mathrm{reg}}_{m-1}(\pi_c, c)] + \varepsilon_m. \qquad (17)$$

Then for any $m > 0$,

$$\mathbb{E}_{c \sim \mathcal{D}} [\mathrm{reg}(\pi_c, c)] \le \frac{10}{9} \cdot \mathbb{E}_{c \sim \mathcal{D}} [\widehat{\mathrm{reg}}_m(\pi_c, c)] + \delta_m, \qquad (18)$$

$$\mathbb{E}_{c \sim \mathcal{D}} [\widehat{\mathrm{reg}}_m(\pi_c, c)] \le \frac{9}{8} \cdot \mathbb{E}_{c \sim \mathcal{D}} [\mathrm{reg}(\pi_c, c)] + \delta_m, \qquad (19)$$

where $\delta_1 = 2\varepsilon_1 + \frac{1}{10}$ and $\delta_m = \frac{1}{9} \delta_{m-1} + \frac{20}{9} \varepsilon_m$ for any $m \ge 2$.

Proof of Lemma E.1. We present the proof here for completeness. By Eq. (17), we have that for all $m \geq 0$ and π_c,

$$\begin{aligned}
&\mathbb{E}_{c\sim\mathcal{D}}[\mathrm{reg}(\pi_c, c)] - \mathbb{E}_{c\sim\mathcal{D}}[\widehat{\mathrm{reg}}_m(\pi_c, c)] \\
&\qquad = \mathbb{E}_{c\sim\mathcal{D}}[V_\star^1(\pi_{\star,c}, c) - \widehat{V}_m^1(\pi_{\star,c}, c)] + \mathbb{E}_{c\sim\mathcal{D}}[\widehat{V}_m^1(\pi_{\star,c}, c) - \widehat{V}_m^1(\widehat{\pi}_{m,c}, c)] \\
&\qquad\quad + \mathbb{E}_{c\sim\mathcal{D}}[\widehat{V}_m^1(\pi_c, c) - V_\star^1(\pi_c, c)] \\
&\qquad \leq \frac{1}{20}\mathbb{E}_{c\sim\mathcal{D}}[\widehat{\mathrm{reg}}_{m-1}(\pi_{\star,c}, c)] + \frac{1}{20}\mathbb{E}_{c\sim\mathcal{D}}[\widehat{\mathrm{reg}}_{m-1}(\pi_c, c)] + 2\varepsilon_m. \qquad (20)
\end{aligned}$$

Symmetrically, we have

$$\begin{aligned}
&\mathbb{E}_{c\sim\mathcal{D}}[\widehat{\mathrm{reg}}_m(\pi_c, c)] - \mathbb{E}_{c\sim\mathcal{D}}[\mathrm{reg}(\pi_c, c)] \\
&\qquad \leq \frac{1}{20}\mathbb{E}_{c\sim\mathcal{D}}[\widehat{\mathrm{reg}}_{m-1}(\widehat{\pi}_{m,c}, c)] + \frac{1}{20}\mathbb{E}_{c\sim\mathcal{D}}[\widehat{\mathrm{reg}}_{m-1}(\pi_c, c)] + 2\varepsilon_m. \qquad (21)
\end{aligned}$$

Now we inductively show for $m = 1, 2, \ldots$ that Eq. (18) and Eq. (19) hold. For $m = 1$, since $\mathbb{E}_{c\sim\mathcal{D}}[\mathrm{reg}(\pi_c, c)], \mathbb{E}_{c\sim\mathcal{D}}[\widehat{\mathrm{reg}}_1(\pi_c, c)] \leq 1$ for all π, then we have from Eq. (20) and Eq. (21)

$$\begin{aligned}
&\mathbb{E}_{c\sim\mathcal{D}}[\mathrm{reg}(\pi_c, c)] \leq \mathbb{E}_{c\sim\mathcal{D}}[\widehat{\mathrm{reg}}_1(\pi_c, c)] + 2\varepsilon_1 + \frac{1}{10} \quad \text{and} \\
&\mathbb{E}_{c\sim\mathcal{D}}[\widehat{\mathrm{reg}}_1(\pi_c, c)] \leq \mathbb{E}_{c\sim\mathcal{D}}[\mathrm{reg}(\pi_c, c)] + 2\varepsilon_1 + \frac{1}{10}.
\end{aligned}$$

Hence we have shown Eq. (18) and Eq. (19) for $m = 1$ and $\delta_1 = 2\varepsilon_1 + \frac{1}{10}$. Now, suppose Eq. (18) and Eq. (19) holds for all $1, 2, \ldots, m-1$. Plugging Eq. (19) for $m-1$ into the right hand side of Eq. (20), we have

$$\begin{aligned}
&\mathbb{E}_{c\sim\mathcal{D}}[\mathrm{reg}(\pi_c, c)] - \mathbb{E}_{c\sim\mathcal{D}}[\widehat{\mathrm{reg}}_m(\pi_c, c)] \\
&\qquad \leq \frac{1}{20}\mathbb{E}_{c\sim\mathcal{D}}[\widehat{\mathrm{reg}}_{m-1}(\pi_{\star,c}, c)] + \frac{1}{20}\mathbb{E}_{c\sim\mathcal{D}}[\widehat{\mathrm{reg}}_{m-1}(\pi_c, c)] + 2\varepsilon_m \\
&\qquad \leq \frac{1}{20}\left(\frac{9}{8}\mathbb{E}_{c\sim\mathcal{D}}[\mathrm{reg}(\pi_{\star,c}, c)] + \delta_{m-1}\right) + \frac{1}{20}\left(\frac{9}{8}\mathbb{E}_{c\sim\mathcal{D}}[\mathrm{reg}(\pi_c, c)] + \delta_{m-1}\right) + 2\varepsilon_m \\
&\qquad = \frac{1}{20}\delta_{m-1} + \frac{1}{20}\left(\frac{9}{8}\mathbb{E}_{c\sim\mathcal{D}}[\mathrm{reg}(\pi_c, c)] + \delta_{m-1}\right) + 2\varepsilon_m,
\end{aligned}$$

where the last equality is by $\mathbb{E}_{c\sim\mathcal{D}}[\mathrm{reg}(\pi_{\star,c}, c)] = 0$. Thus reorganizing the terms, we have

$$\mathbb{E}_{c\sim\mathcal{D}}[\mathrm{reg}(\pi_c, c)] \leq \frac{10}{9}\mathbb{E}_{c\sim\mathcal{D}}[\widehat{\mathrm{reg}}_m(\pi_c, c)] + \delta_m,$$

where $\delta_m = \frac{1}{9}\delta_{m-1} + \frac{20}{9}\varepsilon_m$. Thus we have shown Eq. (18) for m. Then plugging Eq. (19) for $m-1$ into the right hand side of Eq. (21), we have

$$\begin{aligned}
&\mathbb{E}_{c\sim\mathcal{D}}[\widehat{\mathrm{reg}}_m(\pi_c, c)] - \mathbb{E}_{c\sim\mathcal{D}}[\mathrm{reg}(\pi_c, c)] \\
&\qquad \leq \frac{1}{20}\mathbb{E}_{c\sim\mathcal{D}}[\widehat{\mathrm{reg}}_{m-1}(\widehat{\pi}_{m,c}, c)] + \frac{1}{20}\mathbb{E}_{c\sim\mathcal{D}}[\widehat{\mathrm{reg}}_{m-1}(\pi_c, c)] + 2\varepsilon_m \\
&\qquad \leq \frac{1}{20}\left(\frac{9}{8}\mathbb{E}_{c\sim\mathcal{D}}[\mathrm{reg}(\widehat{\pi}_{m,c}, c)] + \delta_{m-1}\right) + \frac{1}{20}\left(\frac{9}{8}\mathbb{E}_{c\sim\mathcal{D}}[\mathrm{reg}(\pi_c, c)] + \delta_{m-1}\right) + 2\varepsilon_m.
\end{aligned}$$

Furthermore, by Eq. (18) for m we have $\mathbb{E}_{c\sim\mathcal{D}}[\mathrm{reg}(\widehat{\pi}_{m,c}, c)] \leq \frac{10}{9}\mathbb{E}_{c\sim\mathcal{D}}[\widehat{\mathrm{reg}}(\widehat{\pi}_{m,c}, c)] + \delta_m = \delta_m$. Plug this in the aforementioned inequality, we have

$$\mathbb{E}_{c\sim\mathcal{D}}[\widehat{\mathrm{reg}}_m(\pi_c, c)] - \mathbb{E}_{c\sim\mathcal{D}}[\mathrm{reg}(\pi_c, c)] \leq \frac{1}{20}\left(\frac{9}{8}\delta_m + \delta_{m-1}\right) + \frac{1}{20}\left(\frac{9}{8}\mathbb{E}_{c\sim\mathcal{D}}[\mathrm{reg}(\pi_c, c)] + \delta_{m-1}\right) + 2\varepsilon_m.$$

Reorganizing the terms, this in turn gives

$$\mathbb{E}_{c\sim\mathcal{D}}[\widehat{\mathrm{reg}}_m(\pi_c, c)] \leq \frac{9}{8}\mathbb{E}_{c\sim\mathcal{D}}[\mathrm{reg}(\pi_c, c)] + \delta_m,$$

where recall $\delta_m = \frac{1}{9}\delta_{m-1} + \frac{20}{9}\varepsilon_m$. This proves Eq. (19) for m which completes our induction.

□

Proof of Theorem 4.1. Let $\mathcal{E}_0 = 1$. Then we have by Lemma 4.1, Eq. (17) holds with $\varepsilon_m = L\sqrt{H^6 S^4 A^3 \mathcal{E}_m}$ for $L > 0$ large enough for all $m > 0$. Combining Lemmas 3.1 and E.1, we have

$$\mathbb{E}_{c\sim\mathcal{D},\pi\sim p_m^h(c)}[\mathrm{reg}(\pi, c)] \lesssim \mathbb{E}_{c\sim\mathcal{D},\pi\sim p_m^h(c)}[\widehat{\mathrm{reg}}_{m-1}(\pi_c, c)] + \sum_{i=0}^{m-1} \frac{1}{9^{m-i}} \sqrt{H^6 S^4 A^3 \cdot \mathcal{E}_i}$$

Then by Lemma 3.1, we have further

$$\mathbb{E}_{c\sim\mathcal{D},\pi\sim p_m^h(c)}[\widehat{\mathrm{reg}}_{m-1}(\pi_c, c)] \lesssim \sqrt{H^6 S^4 A^3 \cdot \mathcal{E}_m}.$$

Hence, we have

$$\mathbb{E}_{c\sim\mathcal{D},\pi\sim p_m^h(c)}[\mathrm{reg}(\pi, c)] \lesssim \sum_{i=0}^{m} \frac{1}{9^{m-i}} \sqrt{H^6 S^4 A^3 \cdot \mathcal{E}_i}.$$

In all, we can obtain the following regret bound with probability at least $1-\delta$,

$$\begin{aligned} \sum\nolimits_{t=1}^{T} \mathrm{Reg}(T) &= \sum_{t=1}^{T} \mathbb{E}_{c_t\sim\mathcal{D},\pi_t\sim p_{m(t)}^{h(t)}(c_t)}[\mathrm{reg}(\pi_t, c_t)] \\ &= \sum_{h,m} \mathbb{E}_{c\sim\mathcal{D},\pi\sim p_m^h(c)}[\mathrm{reg}(\pi, c)] \cdot (\tau_m - \tau_{m-1}) \\ &\lesssim \sum_{m=1}^{N} (\tau_m - \tau_{m-1}) \cdot \sqrt{H^8 S^4 A^3 \cdot \mathcal{E}_m}, \end{aligned}$$

where the last step takes a union bound on the offline oracle guarantees. □

Proof of Theorem 3.1. Without loss of generality, assume that $T/H > 1000$. Since we are choosing $\tau_m = 2(T/H)^{1-2^{-m}}$, we first note that since $\tau_m \le T/H$, we have $(T/H)^{2^{-m}} \ge 2$. Then we have

$$\begin{aligned} \tau_m - \tau_{m-1} &= 2(T/H)^{1-2^{-m}} - 2(T/H)^{1-2^{1-m}} \\ &= 2(T/H)^{1-2^{1-m}}(T^{2^{-m}} - 1) \\ &\ge 2(T/H)^{1-2^{1-m}} = \tau_{m-1}. \end{aligned}$$

This implies $\tau_m - \tau_{m-1} \ge \frac{1}{2}\tau_m$ for $m \in [N]$. Then by Theorem 4.1, we have with probability at least $1-\delta$,

$$\begin{aligned} \sum\nolimits_{t=1}^{T} \mathrm{Reg}(T) &\le \sum_{m=1}^{N} (\tau_m - \tau_{m-1}) \cdot \sqrt{H^8 S^4 A^3 \cdot \mathcal{E}_m} \\ &\lesssim \sqrt{H^8 S^4 A^3 \log(|\mathcal{M}|N/\delta)} \cdot \sum_{m=2}^{N} \frac{\tau_m - \tau_{m-1}}{\sqrt{\tau_{m-1} - \tau_{m-2}}} + \tau_1\sqrt{H^8 S^4 A^3} \\ &\lesssim \sqrt{H^8 S^4 A^3 \log(|\mathcal{M}|N/\delta)} \cdot \sum_{m=2}^{N} \frac{\tau_m}{\sqrt{\tau_{m-1}}} + \sqrt{H^7 S^4 A^3 T} \\ &\lesssim \sqrt{H^8 S^4 A^3 \log(|\mathcal{M}|N/\delta)} N \sqrt{T/H} = \sqrt{H^7 S^4 A^3 T \cdot \log(|\mathcal{M}| \log\log T/\delta) \log\log T}, \end{aligned}$$

where the last inequality follows from

$$\frac{\tau_m}{\sqrt{\tau_{m-1}}} \le \frac{\tau_m}{\sqrt{(\tau_{m-1}+1)/2}} \le \frac{2(T/H)^{1-2^{-m}}}{(T/H)^{\frac{1}{2}(1-2^{1-m})}} = 2\sqrt{T/H}$$

and $N = O(\log\log T)$. So the number of oracle calls are $O(H \log\log T)$.

□

Proof of Theorem 3.2. Since we are choosing $\tau_m = 2^m$, we have $N = O(\log T)$. So the number of oracle calls are $O(H \log T)$. Meanwhile, by Theorem 4.1, we have with probability at least $1-\delta$,

$$\begin{aligned}
\sum\nolimits_{t=1}^{T} \mathrm{Reg}(T) &\le \sum_{m=1}^{N} (\tau_m - \tau_{m-1}) \cdot \sqrt{H^8 S^4 A^3 \cdot \mathcal{E}_m} \\
&\lesssim \sqrt{H^8 S^4 A^3 \log(|\mathcal{M}|N/\delta)} \left(1 + \sum_{m=3}^{N} \frac{2^{m-1}}{\sqrt{2^{m-2}}} \right) \\
&\lesssim \sqrt{H^8 S^4 A^3 \log(|\mathcal{M}|N/\delta)} \cdot 2^{N/2} = \sqrt{H^7 S^4 A^3 T \cdot \log(|\mathcal{M}| \log T/\delta)}.
\end{aligned}$$

□

F Proofs from Section 5

Proof of Theorem 5.1. The reward distributions $\widehat{R}_m$ are 0 for $m = 0, 1$ since $\widehat{R}_0$ are set to be 0 and the model class $\mathcal{M}$ consists of models with constantly 0 reward. The regret estimations $\widehat{\mathrm{reg}}_{m-1}$ are all 0 for $m = 1, 2$. Thus we apply the component-wise guarantees (Lemmas 3.2 to 3.5) to $\widehat{P}_2, \widehat{R}_2$ where $\widehat{R}_2 = 0$. Concretely, from Lemma 3.2 we have with probability at least $1-\delta$,

$$\mathbb{E}_{c\sim\mathcal{D},\pi\sim p_2^h(c)} \left[\mathbb{E}^{M_\star,\pi,c} \left[D_{\mathrm{H}}^2 \left(\widehat{P}_2^h(s_1^h, a_1^h; c), P_\star^h(s_1^h, a_1^h; c) \right) \right] \right] \lesssim \mathcal{E}_2, \tag{22}$$

with the offline $\mathsf{MLE}_{\mathcal{M}}$ oracle. From Lemma 3.3, we have for any π, c, h, s, a,

$$\widehat{d}_2^h(s, a; \pi, c) - \widetilde{d}_2^h(s, a; \pi, c) \le 32e\sqrt{H^4 S^2 A \cdot \mathcal{E}_2}. \tag{23}$$

From Lemma 3.4, we have if for a context c and all $h \in [H]$,

$$\mathbb{E}_{\pi\sim p_2^h(c)} \mathbb{E}^{M_\star,\pi,c} \left[D_{\mathrm{H}}^2 \left(\widehat{P}_2^h(s_1^h, a_1^h; c), P_\star^h(s_1^h, a_1^h; c) \right) \right] \le H/\gamma_2.$$

Then for the same c and all π, h, s, a, we have

$$\widetilde{d}_2^h(s, a; \pi, c) \le (1 + 1/H)^{2(h-1)} d_2^h(s, a; \pi, c). \tag{24}$$

Finally, from Lemma 3.5, we have for any π, c, h, s, a,

$$\begin{aligned}
&\widetilde{d}_2^h(s, a; \pi, c) \cdot D_{\mathrm{H}}(P_\star^h(s, a; c), \widehat{P}_2^h(s, a; c)) \\
&\qquad \le \frac{\gamma_2}{e^2 H} \cdot p_2^h(c, \pi_{2,c}^{h,s,a}) \widetilde{d}_2^h(s, a; \pi_{2,c}^{h,s,a}, c) \cdot D_{\mathrm{H}}^2(P_\star^h(s, a; c), \widehat{P}_2^h(s, a; c)) \\
&\qquad\quad + 2e^2 \sqrt{\frac{\mathcal{E}_2}{H^4 S^2 A}} \cdot \widetilde{d}_2^h(s, a; \pi, c).
\end{aligned} \tag{25}$$

Now we are ready to prove our claim following similar derivations from the proof of Lemma 4.1. Concretely, we have for any set of policies $\{\pi_c\}_{c\in\mathcal{C}}$ and any context c, by local simulation lemma (Lemma C.2)

$$\left| V_\star^1(\pi_c, c, R) - \widehat{V}_2^1(\pi_c, c, R) \right| \le \sum_{h,s,a} \widehat{d}_2^h(s, a; \pi_c, c) D_{\mathrm{H}} \left(\widehat{P}_2^h(s, a; c), P_\star^h(s, a; c) \right).$$

Then by Eq. (23), we have

$$\begin{aligned}
&\sum_{h,s,a} \widehat{d}_2^h(s, a; \pi_c, c) D_{\mathrm{H}} \left(\widehat{P}_2^h(s, a; c), P_\star^h(s, a; c) \right) \\
&\qquad \le \sum_{h,s,a} \widetilde{d}_2^h(s, a; \pi_c, c) D_{\mathrm{H}} \left(\widehat{P}_2^h(s, a; c), P_\star^h(s, a; c) \right) + 32e\sqrt{H^6 S^4 A^3 \cdot \mathcal{E}_2}
\end{aligned} \tag{26}$$

Case I: If for a context c and all $h \in [H]$,

$$\mathbb{E}_{\pi\sim p_2^h(c)} \mathbb{E}^{M_\star,\pi,c} \left[D_{\mathrm{H}}^2 \left(\widehat{P}_2^h(s_1^h, a_1^h; c), P_\star^h(s_1^h, a_1^h; c) \right) \right] \le H/\gamma_2.$$

Then by Eqs. (24) and (25),

$$
\begin{aligned}
&\sum_{h,s,a} \widetilde{d}_2^h(s,a;\pi_c,c) D_{\mathrm{H}}\Big(\widehat{P}_2^h(s,a;c), P_\star^h(s,a;c)\Big) \\
&\qquad\qquad \le \frac{\gamma_2}{e^2 H} \cdot p_2^h(c, \pi_{2,c}^{h,s,a}) \widetilde{d}_2^h(s,a;\pi_{2,c}^{h,s,a},c) \cdot D_{\mathrm{H}}^2(P_\star^h(s,a;c), \widehat{P}_2^h(s,a;c)) \\
&\qquad\qquad\quad + 2e^2 \sqrt{\frac{\mathcal{E}_2}{H^4 S^2 A}} \cdot \widetilde{d}_2^h(s,a;\pi_c,c) \\
&\qquad\qquad \le \frac{\gamma_2}{H} \cdot p_2^h(c, \pi_{2,c}^{h,s,a}) d_2^h(s,a;\pi_{2,c}^{h,s,a},c) \cdot D_{\mathrm{H}}^2(P_\star^h(s,a;c), \widehat{P}_2^h(s,a;c)) \\
&\qquad\qquad\quad + 2e^2 \sqrt{\frac{\mathcal{E}_2}{H^4 S^2 A}} \cdot \widetilde{d}_2^h(s,a;\pi_c,c)
\end{aligned}
$$

Thus we have

$$
\begin{aligned}
&\Big|V_\star^1(\pi_c,c,R) - \widehat{V}_2^1(\pi_c,c,R)\Big| \\
&\qquad \lesssim \sqrt{H^6 S^4 A^3 \cdot \mathcal{E}_2} + \frac{\gamma_2}{H} \cdot \sum_{h=1}^{H} \mathbb{E}_{\pi \sim p_2^h(c)} \mathbb{E}^{M_\star,\pi,c}\left[D_{\mathrm{H}}^2\Big(\widehat{P}_2^h(s_1^h,a_1^h;c), P_\star^h(s_1^h,a_1^h;c)\Big)\right].
\end{aligned}
$$

Case II: If for a context c there exists $j \in [H]$ such that

$$
\mathbb{E}_{\pi \sim p_2^h(c)} \mathbb{E}^{M_\star,\pi,c}\left[D_{\mathrm{H}}^2\Big(\widehat{P}_2^j(s_1^j,a_1^j;c), P_\star^h(s_1^j,a_1^j;c)\Big)\right] > H/\gamma_2.
$$

Then

$$
\begin{aligned}
&\Big|V_\star^1(\pi_c,c,R) - \widehat{V}_2^1(\pi_c,c,R)\Big| \\
&\qquad\qquad \le 1 \le \frac{\gamma_2}{H} \cdot \sum_{h=1}^{H} \mathbb{E}_{\pi \sim p_2^h(c)} \mathbb{E}^{M_\star,\pi,c}\left[D_{\mathrm{H}}^2\Big(\widehat{P}_2^h(s_1^h,a_1^h;c), P_\star^h(s_1^h,a_1^h;c)\Big)\right].
\end{aligned}
$$

Combine Case I and II, we have

$$
\begin{aligned}
&\Big|V_\star^1(\pi_c,c,R) - \widehat{V}_2^1(\pi_c,c,R)\Big| \\
&\qquad \lesssim \sqrt{H^6 S^4 A^3 \cdot \mathcal{E}_2} + \frac{\gamma_2}{H} \cdot \sum_{h=1}^{H} \mathbb{E}_{\pi \sim p_2^h(c)} \mathbb{E}^{M_\star,\pi,c}\left[D_{\mathrm{H}}^2\Big(\widehat{P}_2^h(s_1^h,a_1^h;c), P_\star^h(s_1^h,a_1^h;c)\Big)\right].
\end{aligned}
$$

Then take expectation with respect to c together with Eq. (22) we obtain

$$
\begin{aligned}
\mathbb{E}_{c\sim\mathcal{D}}\left[\Big|V_\star^1(\pi_c,c,R) - \widehat{V}_2^1(\pi_c,c,R)\Big|\right] &\lesssim \sqrt{H^6 S^4 A^3 \cdot \mathcal{E}_2} + \gamma_2 \mathcal{E}_2 \\
&\lesssim \sqrt{\frac{H^7 S^4 A^3 \log(|\mathcal{M}|/\delta)}{T}} \\
&\lesssim \varepsilon,
\end{aligned}
$$

where the second inequality is by $\mathcal{E}_2 = H \log(|\mathcal{M}|/\delta)/T$ and the last inequality holds when

$$
T \ge \Omega\bigg(\frac{H^7 S^4 A^3 \log(|\mathcal{M}|/\delta)}{\varepsilon^2}\bigg).
$$

Thus the reward-free objective Eq. (5) is satisfied with probability at least $1-\delta$ with

$$
T \le O\bigg(\frac{H^7 S^4 A^3 \log(|\mathcal{M}|/\delta)}{\varepsilon^2}\bigg).
$$

□

Lemma F.1 (Lemma D.2 of Jin et al. [22])**.** *Fix $\varepsilon \le 1$, $\delta \le 1/2$, $H, A \ge 2$, and suppose that $S \ge L \log A$ for a universal constant L. Then consider the trivial context space $\mathcal{C} = \{c_0\}$, there exists a model class $\mathcal{M} = \{M_j\}_{j \in \mathcal{J}}$, with $|\mathcal{S}| = S$, $|\mathcal{A}| = A$, $|\mathcal{J}| \le e^{SA}$, and horizon H and a distribution μ on $\mathcal{M}$ such that any algorithm* Alg *that $(\varepsilon/12, \delta)$-learns the class $\mathcal{M}$ satisfies*

$$\mathbb{E}_{M\sim\mu}\mathbb{E}_{M,\mathsf{Alg}}[T] \gtrsim \frac{SA}{\varepsilon^2},$$

where T is the number of trajectories required by the algorithm Alg *to achieve $(\varepsilon/12, \delta)$ accuracy and $\mathbb{E}_{M,\mathsf{Alg}}[\cdot]$ is the expectation under the interaction between the algorithm* Alg *and model M.*

Proof of Theorem 5.2. . Let $n = \log K/(SA)$, then we consider the context space $\mathcal{C} = \{c_1, ..., c_n\}$ and the i.i.d. distribution $\mathcal{D}$ on the context space being uniform. Denote the model class obtained from Lemma F.1 by $\overline{\mathcal{M}} = \{\overline{M}_j(c_0)\}_{j\in\mathcal{J}}$. Let $J = \{j_1, ..., j_n\} \in \mathcal{J}^n$ be an index. Then we consider the model class $\mathcal{M} = \{M_J\}_{J\in\mathcal{J}^n}$, where $M_J(c_i) = \overline{M}_{j_i}(c_0)$, that is, the model class $\mathcal{M}$ is on each context $c_i \in \mathcal{C}$ an independent $\overline{\mathcal{M}}$. We first have that the size of the model class is bounded by $|\mathcal{M}| = |\mathcal{J}|^n \le e^{nSA} \le K$. Then we have for any algorithm Alg that $(\varepsilon/24, \delta)$-learns the class $\mathcal{M}$, it must have $(\varepsilon/12, \delta)$-learns the class $\mathcal{M}(c_i) = \{M_J(c_i)\}_{J\in\mathcal{J}^n} = \{\overline{M}_j(c_0)\}_{j\in\mathcal{J}}$ for at least half of the contexts $c_i \in \mathcal{C}$ by Markov's inequality. This, in turn, combined with Lemma F.1 gives that there exists a distribution μ on the model class $\mathcal{M}$ such that

$$\mathbb{E}_{M\sim\mu}\mathbb{E}_{M,\mathsf{Alg}}[T] \gtrsim \frac{n}{2}\frac{SA}{\varepsilon^2} \gtrsim \frac{\log K}{\varepsilon^2},$$

where T is the number of trajectories required by the algorithm Alg to achieve $(\varepsilon/12, \delta)$ accuracy and $\mathbb{E}_{M,\mathsf{Alg}}[\cdot]$ is the expectation under the interaction between the algorithm Alg and model M. Thus concludes our proof.

□

G Computation

In this section, we show that for any m, c, h, s, a, the policy $\pi^{h,s,a}_{m,c}$ (Line 8) and the trusted transitions $\widetilde{\mathcal{T}}^h_m(c)$ (Definition 3.1) can be computed efficiently through linear programming. For simplicity, we fix a context c throughout this section and omit its dependence. We first note that if $\widetilde{\mathcal{T}}^j_m$ for $j \le h-1$ and $\pi^{h,s,a}_m$ are computed, then $\widetilde{\mathcal{T}}^h_m$ can be computed in $\mathrm{poly}(HSA)$ time. To see this, for any (s, a, s'), we have by the definition of $\pi^{h,s,a}_m$ that

$$\max_\pi \frac{\widetilde{d}^h_m(s, a; \pi)}{SA + \eta_m \cdot \widehat{\mathrm{reg}}_{m-1}(\pi)} = \frac{\widetilde{d}^h_m(s, a; \pi^{h,s,a}_m)}{SA + \eta_m \cdot \widehat{\mathrm{reg}}_{m-1}(\pi^{h,s,a}_m)}.$$

Then $(s, a, s') \in \widetilde{\mathcal{T}}^h_m$ iff

$$\frac{\widetilde{d}^h_m(s, a; \pi^{h,s,a}_m)}{SA + \eta_m \cdot \widehat{\mathrm{reg}}_{m-1}(\pi^{h,s,a}_m)}\widehat{P}^h_m(s'|s, a) \ge \frac{1}{\zeta_m},$$

where the left-hand side can be computed given $\widetilde{\mathcal{T}}^j_m$ for $j \le h-1$ in $\mathrm{poly}(HSA)$ time. Thus we only need to show how to compute $\pi^{h,s,a}_m$. Assume we want to compute $\pi^{\bar{h},\bar{s},\bar{a}}_m$ for $\bar{h} \in [H], \bar{s} \in \mathcal{S}, \bar{a} \in \mathcal{A}$, given $\widetilde{\mathcal{T}}^h_m$ for $h \le \bar{h} - 1$. Let $\bar{r}^h_m(s, a) = \mathbb{E}_{r^h\sim\widehat{R}^h_{m-1}}[r^h]$ be the mean reward for model $\widehat{M}_{m-1}$ for h, s, a. Let $\bar{\mathbf{r}}_m = (\bar{r}^h_m(s, a))_{h,s,a}$ be the vector of mean rewards for the model $\widehat{M}_{m-1}$. We consider the following linear fractional program with the following decision variables:

$$\mathbf{d}^{\bar{h}}_m = \begin{pmatrix} \widetilde{\mathbf{d}}^{\bar{h}}_m \\ \widehat{\mathbf{d}}_{m-1} \end{pmatrix}, \quad \text{where } \widetilde{\mathbf{d}}^{\bar{h}}_m = (\widetilde{d}_{h,s,a,m})_{h,s,a\in[\bar{h}]\times\mathcal{S}\times\mathcal{A}}, \ \widehat{\mathbf{d}}_{m-1} = (\widehat{d}_{h,s,a,m-1})_{h,s,a\in[H]\times\mathcal{S}\times\mathcal{A}}.$$

We will use the variable $\widetilde{\mathbf{d}}^{\bar{h}}_m$ to simulate the trusted occupancy measures and $\widehat{\mathbf{d}}_{m-1}$ to simulate the estimated occupancy measures from $\widehat{M}_{m-1}$ by linear constraints. Concretely, consider the following linear fractional program:

$$\max_{\mathbf{d}_m^{\bar{h}}} \ \frac{\widetilde{d}_{\bar{h},\bar{s},\bar{a},m}}{SA + \eta_m(\widehat{V}_{m-1}^1 - \langle \widehat{\mathbf{d}}_{m-1}, \bar{\mathbf{r}}_{m-1} \rangle)},$$

$$\text{subject to :} \ \begin{cases} \widehat{d}_{h,s,a,m-1} \geq 0, & \forall\, h,s,a \in [H] \times \mathcal{S} \times \mathcal{A}, \\ \sum_a \widehat{d}_{1,s,a,m-1} = \mathbb{1}(s = s^1), & \forall s \in \mathcal{S}, \\ \sum_{s,a} \widehat{d}_{h,s,a,m-1} \widehat{P}_{m-1}^h(s'|s,a) = \sum_a \widehat{d}_{h+1,s',a,m-1}, & \forall\, h, s' \in [H] \times \mathcal{S}, \\ \widetilde{d}_{h,s,a,m} \geq 0, & \forall\, h,s,a \in [\bar{h}] \times \mathcal{S} \times \mathcal{A} \\ \sum_a \widetilde{d}_{1,s,a,m} = \mathbb{1}(s = s^1), & \forall s \in \mathcal{S}, \\ \sum_{s,a,s' \in \widetilde{\mathcal{T}}_m^{h-1}} \widetilde{d}_{h-1,s,a,m} \widehat{P}_m^{h-1}(s'|s,a) = \sum_a \widetilde{d}_{h,s',a,m} & \forall\, h \leq \bar{h}, s' \in \mathcal{S}. \end{cases}$$

This is a linear fractional program of $HSA + \bar{h}SA$ decision variables with $HSA + HS + \bar{h}SA + \bar{h}S$ constraints. It is clear from the linear constraints that this program simulates the MDPs $\widehat{M}_{m-1}$ and $\widehat{M}_m$, and the program obtains the right objective. Then, for this linear fractional program, we apply the Charnes-Cooper transformation [7] to transform it into a linear program. After the transformation, we can apply existing tools for solving linear programs (e.g., Lee and Sidford [23]) to solve for $\widetilde{\mathbf{d}}_m^{\bar{h}}$ which encodes the policy $\pi_m^{\bar{h},\bar{s},\bar{a}}$ in $\mathrm{poly}(HSA)$ time.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading “NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

   Answer: [Yes] .

   Justification: The main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope. The claims are validated by detailed proofs.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes] .

Justification: The paper discusses the limitations of the work in the discussion section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes] .

   Justification: The paper provides detailed assumptions and proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA] .

Justification: This paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: This paper does not include experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not

including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA] .

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA] .

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA] .

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes] .

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This is a theoretical work. There is no societal impact on the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA] .

    Justification: This paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA] .

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: This paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.