Eluder-based Regret for Stochastic Contextual MDPs

Orin Levy [*1] **Asaf Cassel** [*1] **Alon Cohen** [2 3] **Yishay Mansour** [*1 3]

Abstract

We present the E-UC^{3}RL algorithm for regret minimization in Stochastic Contextual Markov Decision Processes (CMDPs). The algorithm operates under the minimal assumptions of realizable function class and access to *offline* least squares and log loss regression oracles. Our algorithm is efficient (assuming efficient offline regression oracles) and enjoys a regret guarantee of $\widetilde{O}(H^3\sqrt{T|S||A|d_{\mathrm{E}}(\mathcal{P})\log(|\mathcal{F}||\mathcal{P}|/\delta)})$, with T being the number of episodes, S the state space, A the action space, H the horizon, $\mathcal{P}$ and $\mathcal{F}$ are finite function classes used to approximate the context-dependent dynamics and rewards, respectively, and $d_{\mathrm{E}}(\mathcal{P})$ is the Eluder dimension of $\mathcal{P}$ w.r.t the Hellinger distance. To the best of our knowledge, our algorithm is the first efficient and rate-optimal regret minimization algorithm for CMDPs that operates under the general offline function approximation setting. In addition, we extend the Eluder dimension to general bounded metrics which may be of independent interest.

1. Introduction

Reinforcement Learning (RL) is a field of machine learning that pertains to sequential decision making under uncertainty. At the heart of RL is the Markov Decision Process (MDP), a fundamental mathematical model that has been studied extensively. An agent repeatedly interacts with an MDP by observing its state $s \in S$ and selecting an action $a \in A$, which leads to a new state s' and an instantaneous reward that reflects the quality of the action taken. The agent's goal is to maximize her return during each episode of interaction with the MDP. MDPs can be applied to a wide range of real-life scenarios, including advertising, healthcare, games, robotics (see, e.g., Sutton & Barto, 2018; Mannor et al., 2022).

Many modern applications involve the presence of additional side information, or *context*, that impacts the environment. A naive approach to handling the context is to extend the state space of the environment to include it. However, this method increases the complexity of learning and policy representation. Contextual MDPs (CMDPs) offer a more efficient solution by keeping the state space small and treating the context as additional side-information that the agent observes at the start of each episode. Furthermore, there exists a mapping from each context to an MDP, and the optimal policy for a given context is the optimal policy of the corresponding MDP (Hallak et al., 2015). An example of a context is user information that remains constant throughout the episode. Such information may include the user's age and interests, and can deeply impact decision making. This feature makes CMDPs an excellent model for recommendation systems.

As is common in recent works, the aforementioned mapping from context to MDP is assumed to be taken from a known function class, and access to the function class is provided via an optimization oracle. A distinctive feature between works is whether they assume access to *online* or *offline* oracles. Intuitively, in both settings we have a function class $\mathcal{F} = \{f : X \to Y\}$, a loss $\ell : Y \times Y \to \mathbb{R}$, and a dataset[1] $\{(x_i, y_i)\}_{i=1}^n$. An offline oracle observes the entire data and needs to find $\hat{f}_\star \in \arg\min_{f\in\mathcal{F}} \sum_{i=1}^n \ell(f(x_i), y_i)$. An online oracle makes a sequence of predictions $f_1, \ldots, f_n$ where f_i can depend on data up to $i-1$, and its goal is to minimize regret, given by $\sum_{i=1}^n \ell(f_i(x_i), y_i) - \ell(\hat{f}_\star(x_i), y_i)$. The offline problem can potentially be easier to solve than the online problem. Moreover, practical deep RL applications typically work in the offline regime.

Previously, Modi & Tewari (2020) obtained $\widetilde{O}(\sqrt{T})$ regret for a generalized linear model (GLM). Foster et al. (2021) obtain $\widetilde{O}(\sqrt{T})$ regret for general function approximation and adversarially chosen contexts, assuming access to a much stronger online estimation oracle. However, they noted the challenge of implementing their methodol-

[*]Equal contribution [1]Balavatnick school of Computer Science, Tel Aviv University, Tel Aviv, Israel [2]School of Electrical Engeneering, Tel Aviv University, Tel Aviv, Israel [3]Google Research, Tel Aviv, Israel. Correspondence to: Orin Levy <orinlevy@mail.tau.ac.il>, Asaf Cassel <acassel@mail.tau.ac.il>, Alon Cohen <alonco@mail.tau.ac.il>, Yishay Mansour <mansour.yishay@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1]We think of X as the context and Y as the MDP.

ogy using offline oracles. It thus remained open whether, for stochastic contexts, we can restrict the access to offline oracles. Recently, Levy & Mansour (2023) gave an $\widetilde{O}(\sqrt{T}/p_{min})$ regret algorithm for stochastic contexts using offline least squares regression, where p_{min} is a minimum reachability parameter of the CMDP. This parameter can be arbitrarily small and in general CMDPs leads to an $\widetilde{O}(T^{3/4})$ regret guarantee. The question of whether the minimum reachability assumption can be obviated or replaced by a less restrictive assumption on the function class remained open.

In this work, we give the first $\widetilde{O}(\sqrt{T})$ regret algorithm for stochastic contexts using standard offline oracles, under the bounded Eluder dimension (Russo & Van Roy, 2013) assumption (more details in Section 2.3).

Contributions. We present the E-UC^{3}RL algorithm for stochastic CMDPs with *offline* regression oracles. Our algorithm is efficient (assuming efficient oracles) and enjoys an $\widetilde{O}\big(H^3\sqrt{T|S||A|d_{\mathrm{E}}(\mathcal{P})\log(|\mathcal{F}||\mathcal{P}|/\delta)}\big)$ regret bound with probability at least $1-\delta$, where S is the state space, A the action space, H the horizon, $\mathcal{P}$ and $\mathcal{F}$ are finite function classes used to approximate the context-dependent dynamics and rewards, respectively, and $d_{\mathrm{E}}(\mathcal{P})$ is the Eluder dimension with respect to Hellinger distance of the context-dependent dynamics function class $\mathcal{P}$. The algorithm builds on the "optimistic in expectation" approach of Levy & Mansour (2023) but modifies it with a log-loss oracle for the dynamics approximation and carefully chosen counterfactual reward bonuses. To that end we present an extension of the Eluder dimension to general bounded metrics (rather than the ℓ_2 norm considered by Russo & Van Roy (2013) and Osband & Van Roy (2014)). An additional key technical tool enabling our result is a multiplicative change of measure inequality for the value function. Both tools may be of separate interest.

Comparison with Levy & Mansour (2023). This work is most closely related to ours. It relies on a minimum reachability assumption and provides a regret bound of $\widetilde{O}(\sqrt{T}/p_{min})$ where p_{min} is the reachability parameter of the CMDP. This implies that any policy π will reach any state s with a probability of at least p_{min} for any context c, hence $p_{min} \leq 1/|S|$. As such, any policy inherently explores with probability p_{min}, significantly simplifying the exploration task. While the notion of minimum reachability is intuitive, it fails even for deterministic transition functions where $p_{min} = 0$. Moreover, it is impossible to estimate it online as we typically observe each context only once.

The primary focus of our work is to replace minimum reachability, which is a structural assumption about the true CMDP, with an assumption about the dynamics function class, which is chosen by the learner. This makes learning an effective exploratory policy non-trivial, necessitating innovative confidence bounds that capture the intricacies of the function class learning complexity. In our approach, we employ the Eluder dimension as the complexity measure. One can show that minimum reachability implies a bound on the Eluder dimension, but, the Eluder dimension can be much smaller.

The general function approximation literature. We stress that the role of the Eluder dimension in this work is to avoid direct dependence on the size of the context space, which could be prohibitively large, while also maintaining computational efficiency. This is unlike previous works on function approximation in RL (see, e.g., Jiang et al. (2017); Jin et al. (2021); Wu et al. (2023); Chen et al. (2022); Wang et al. (2020b); Dann et al. (2021); Liu et al. (2022)) that use an Eluder dimension to avoid dependence on $|S|, |A|$. These works are often computationally inefficient and require additional structural assumptions regarding the MDP, such as low Bellman-rank or Bellman completeness, or a much stronger optimization oracle.

Additional Related Work. Hallak et al. (2015) were the first to study regret guarantees in the the CMDP model. However, they assume a small context space, and their regret is linear in its size. Jiang et al. (2017) present OLIVE, a sample efficient algorithm for learning Contextual Decision Processes (CDP) under the low Bellman rank assumption. In contrast, we do not make any assumptions on the Bellman rank. Sun et al. (2019) use the Witness Rank to derive PAC bounds for model based learning of CDPs. Modi et al. (2018) present generalization bounds for learning *smooth* CMDPs and finite contextual linear combinations of MDPs. Modi & Tewari (2020) present a regret bound of $\widetilde{O}(\sqrt{T})$ for CMDPs under a Generalized Linear Model (GLM) assumption. Our function approximation framework is more general than smooth CMDPs or GLM.

Foster et al. (2021) present the Estimation to Decision (E2D) meta algorithm and apply it to obtain $\widetilde{O}(\sqrt{T})$ regret for adversarial Contextual RL. Later, Xie et al. (2022) show sample complexity bounds for online reinforcement learning using online oracle, that can be also applied to CMDPs. Levy et al. (2023) obtained similar results using their OMG-CMDP! algorithm. However, these works assume access to online estimation oracles and their bounds scale with the oracle's regret. In contrast, we use substantially weaker offline regression oracles. It is not clear to us whether Foster et al.'s Inverse Gap Minimization (IGM) technique or Levy et al.'s convex optimization with log-barrier method can be applied to CMDPs with offline regression oracles.

Levy & Mansour (2022) study the sample complexity of learning CMDPs using function approximation and provide the first general and efficient reduction from CMDPs to offline supervised learning. However, their sample complexity scales as ϵ^{-8}, and thus they cannot achieve the optimal $\sqrt{T}$

rate for the regret. Levy & Mansour (2023), previously mentioned here in relation to upper bounds, also showed an $\Omega(\sqrt{TH|S||A|\log(|\mathcal{G}|/|S|)/\log(|A|)})$ regret lower bound for the general setting of offline function approximation with $|\mathcal{G}|$, the size of the function class used to approximate the rewards in each state.

More broadly, CMDPs are a natural extension of the extensively studied Contextual Multi-Armed Bandit (CMAB) model. CMABs augment the Multi-Arm Bandit (MAB) model with a context that determines the rewards (Lattimore & Szepesvári, 2020; Slivkins, 2019). Langford & Zhang (2007); Agarwal et al. (2014) use an optimization oracle to derive an optimal regret bound that depends on the size of the policy class they compete against. Regression based approaches were presented in Agarwal et al. (2012); Foster & Rakhlin (2020); Foster et al. (2018); Foster & Krishnamurthy (2021); Simchi-Levi & Xu (2021); Zhang (2022). Most closely related to our work, Xu & Zeevi (2020) present the first optimistic algorithm for CMAB. They assume access to a least-squares regression oracle and achieve $\widetilde{O}(\sqrt{T|A|\log|\mathcal{F}|})$ regret, where $\mathcal{F}$ is a finite and realizable function class, used to approximate the rewards. Extending their techniques to CMDPs necessitates accounting for the context-dependent dynamics whose interplay with the rewards significantly complicates decision making. This is the main challenge both in our work and in Levy & Mansour (2023).

The Eluder dimension was introduced by Russo & Van Roy (2013) and applied to derive sublinear regret for MABs and CMABs. Osband & Van Roy (2014) showed an application of the Eluder dimension to derive a regret bound for model-based reinforcement learning and Wen & Van Roy (2017) for deterministic systems. Wang et al. (2020a) use it to derive a regret bound for value function approximation. Jin et al. (2021) present the Bellman-Eluder dimension and use it to develop sample-efficient algorithms for a family of RL problems where both the Bellman rank and the Eluder dimension are low. Ayoub et al. (2020) apply the Eluder dimension to derive a regret bound for tabular episodic RL using targeted value regression. We, on the other hand, extend the Eluder dimension to general bounded metrics and apply it to contextual RL.

2. Preliminaries

2.1. Episodic Loop-Free Markov Decision Process (MDP)

An MDP is defined by a tuple (S, A, P, r, s_0, H), where S and A are finite sets describing the state and action spaces, respectively; $s_0 \in S$ is the unique start state; $H \in \mathbb{N}$ is the horizon; $P : S \times A \times S \to [0, 1]$ defines the probability of transitioning to state s' given that we start at state s and perform action a; and $r(s, a)$ is the expected reward of performing action a at state s. An episode is a sequence of H interactions where at step h, if the environment is at state s_h and the agent performs action a_h then (regardless of past history) the environment transitions to state $s_{h+1} \sim P(\cdot \mid s_h, a_h)$ and the agent receives reward $R(s_h, a_h) \in [0, 1]$, sampled independently from a distribution $\mathcal{D}_{s_h,a_h}$ that satisfies $r(s_h, a_h) = \mathbb{E}_{\mathcal{D}_{s_h,a_h}}[R(s_h, a_h)]$.

For technical convenience and without loss of generality, we assume that the state space and accompanying transition probabilities have a loop-free (or layered) structure. Concretely, we assume that the state space can be decomposed into $H + 1$ disjoint subsets (layers) $S_0, S_1, \ldots, S_{H-1}, S_H$ such that transitions are only possible between consecutive layers, i.e., for $h' \neq h + 1$ we have $P(s_{h'}|s_h, a) = 0$ for all $s_{h'} \in S_{h'}, s_h \in S_h, a \in A$. In addition, $S_H = \{s_H\}$, meaning there is a unique final state with reward 0. We note that this assumption can always be satisfied by increasing the size of the state space by a factor of H.

A *deterministic stationary policy* $\pi : S \to A$ is a mapping from states to actions. Given a policy π and MDP $M = (S, A, P, r, s_0, H)$, the $h \in [H - 1]$ stage value function of a state $s \in S_h$ is defined as

$$V^{\pi}_{M,h}(s) = \mathbb{E}_{\pi,M}\left[\sum_{k=h}^{H-1} r(s_k, a_k)\middle| s_h = s\right].$$

For brevity, when $h = 0$ we denote $V^{\pi}_{M,0}(s_0) := V^{\pi}_M(s_0)$, which is the expected cumulative reward under policy π and its measure of performance. A policy $\pi^{\star}_M$ is *optimal* for MDP M if it satisfies that $\pi^{\star}_M \in \arg\max_{\pi:S\to A}\{V^{\pi}_M(s_0)\}$. It is well known that such a policy is optimal even among the class of stochastic and history dependent policies (see, e.g., Puterman, 2014; Sutton & Barto, 2018; Mannor et al., 2022).

2.2. Problem Setup: Stochastic Contextual Markov Decision Process (CMDP)

A *CMDP* is defined by a tuple $(\mathcal{C}, S, A, \mathcal{M})$ where $\mathcal{C}$ is the context space, S the state space and A the action space. The mapping $\mathcal{M}$ maps a context $c \in \mathcal{C}$ to an MDP $\mathcal{M}(c) = (S, A, P^c_{\star}, r^c_{\star}, s_0, H)$, where $r^c_{\star}(s, a) = \mathbb{E}[R^c_{\star}(s, a)|c, s, a]$, $R^c_{\star}(s, a) \sim \mathcal{D}_{c,s,a}$. We assume that $R^c_{\star}(s, a) \in [0, 1]$.

We consider a *stochastic* CMDP, meaning, the context is stochastic. Formally, we assume that there is an unknown distribution $\mathcal{D}$ over the context space $\mathcal{C}$, and for each episode a context is sampled i.i.d. from $\mathcal{D}$. For mathematical convenience, we assume that the context space $\mathcal{C}$ is finite but potentially very large. Our results do not depend on the size of the context space and can be further extended to infinite context spaces.

A deterministic *context-dependent policy* $\pi : \mathcal{C} \times S \to A$

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

Interaction protocol. The interaction between the agent and the environment is defined as follows. In each episode $t = 1, 2, ..., T$ the agent:

(i) Observes context $c_t \sim \mathcal{D}$;
(ii) Chooses a policy π_t (based on c_t and the observed history);
(iii) Observes trajectory $(c_t, s_0^t, a_0^t, r_0^t, \ldots, s_H^t)$ generated by playing π_t in $\mathcal{M}(c_t)$.

Our goal is to minimize the regret, defined as

$$\mathcal{R}_T := \sum_{t=1}^{T} V_{\mathcal{M}(c_t)}^{\pi^\star(c_t;\cdot)}(s_0) - V_{\mathcal{M}(c_t)}^{\pi_t(c_t;\cdot)}(s_0),$$

where $\pi^\star \in \Pi_{\mathcal{C}}$ is an optimal context-dependent policy. We aim to derive regret bounds that are independent of the context space size $|\mathcal{C}|$. For that purpose, we make function approximation assumptions in Section 2.4, which rely on the following definition of Eluder dimension.

2.3. Metric Eluder Dimension

We extend the notion of Eluder dimension, given by Osband & Van Roy (2014), to general bounded metrics. Let $\mathcal{X}$ be a set and $(\mathcal{Y}, D)$ a bounded metric space. Let $\mathcal{P} \subseteq \{\mathcal{X} \to \mathcal{Y}\}$ be a set of functions from $\mathcal{X}$ to $\mathcal{Y}$. We say that $x \in \mathcal{X}$ is $(D, \epsilon)-$dependent of $x_1, \ldots, x_n$ if and only if for any $P, P' \in \mathcal{P}$ it holds that

$$\sum_{i=1}^{n} D^2(P(x_i), P'(x_i)) \leq \epsilon^2 \implies D^2(P(x), P'(x)) \leq \epsilon^2.$$

We say that $x \in \mathcal{X}$ is $(D, \epsilon)-$independent of $x_1, \ldots, x_n$ if it is not $(D, \epsilon)-$dependent.

Definition 2.1 (Metric-Eluder Dimension)**.** We say that $d := d_{\mathrm{E}}(\mathcal{P}, D, \epsilon)$ is the $(D, \epsilon)-$Eluder dimension of a class $\mathcal{P}$ if d is the maximum length of sequences $x_1, \ldots, x_d$ and $\epsilon_1', \ldots, \epsilon_d'$ such that for all $1 \leq i \leq d$, x_i is $(D, \epsilon_i')-$independent of its prefix $x_1, \ldots, x_{i-1}$ and $\epsilon_i' \geq \epsilon$.

This quantity roughly corresponds to the number of queries required to ϵ identify a function in $\mathcal{P}$. The utility of this definition is summarized in the following result, which is a straightforward adaptation of Proposition 6 in Osband & Van Roy (2014) (proof in Appendix A). For any $\mathcal{P}' \subseteq \mathcal{P}$, define its radius at $x \in \mathcal{X}$ as $w_{\mathcal{P}'}(x) = \sup_{P, P' \in \mathcal{P}'} D(P(x), P'(x))$.

Lemma 2.2. *For any $t \in [T], h \in [H]$ let $x_h^t \in \mathcal{X}$ and $P_t \in \mathcal{P}$ be arbitrary. Define the confidence sets with parameter β as*

$$\mathcal{P}_t = \left\{ P \in \mathcal{P} \ : \ \sum_{i=0}^{t-1} \sum_{h=0}^{H-1} D^2(P(x_h^i), P_t(x_h^i)) \leq \beta \right\}.$$

We have that

$$\sum_{t=1}^{T} \sum_{h=0}^{H-1} \left(w_{\mathcal{P}_t}(x_h^t)\right)^2 \leq 6 d_{\mathrm{E}}(\mathcal{P}, D, T^{-1/2}) \beta \log T.$$

2.4. Function Class Assumptions

We note that, without further assumptions, the regret may scale linearly in the size of the context space (Hallak et al., 2015). Even worse, if the context space contains more than T contexts, and the distribution over the contexts is uniform, the regret may scale linearly in T. We overcome this limitation by imposing the following function approximation assumptions, that extend similar notions in the Contextual Multi-Armed Bandits literature (Agarwal et al., 2012; Russo & Van Roy, 2013; Foster et al., 2018; Foster & Krishnamurthy, 2021; Simchi-Levi & Xu, 2021) to CMDPs.

Realizable reward function approximation. Our algorithm gets as input a finite function class $\mathcal{F} \subseteq \mathcal{C} \times S \times A \to [0, 1]$ such that there exists $f_\star \in \mathcal{F}$ that satisfies $f_\star(c, s, a) = r_\star^c(s, a)$ for all $c \in \mathcal{C}$ and $(s, a) \in S \times A$.

Realizable dynamics function approximation. Our algorithm gets as input a finite function class $\mathcal{P} \subseteq S \times (S \times A \times \mathcal{C}) \to [0, 1]$ such that $P_\star \in \mathcal{P}$, and every function $P \in \mathcal{P}$ represents valid transition probabilities, i.e., satisfies $\sum_{s' \in S} P(s' \mid s, a, c) = 1$ for all $c \in \mathcal{C}$ and $(s, a) \in S \times A$. For convenience, we denote $P(s' \mid s, a, c) = P^c(s' \mid s, a)$, for all $P \in \mathcal{P}$.

Offline regression oracles. Given a data set $D = \{(c_i, s_i, a_i, s_i', r_i)\}_{i=1}^n$, we assume access to offline oracles that solve the optimization problems:

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (f(c_i, s_i, a_i) - r_i)^2,$$

(Least Squares Regression (LSR))

$$\widehat{P} \in \arg\min_{P \in \mathcal{P}} \sum_{i=1}^{n} \log \frac{1}{P^{c_i}(s_i' \mid s_i, a_i)}.$$

(Log Loss Regression (LLR))

Notice that the above problems can always be solved by iterating over the function class. However, since we consider strongly convex loss functions, there are function classes where these optimization problems can be solved efficiently. One particular example is the class of linear functions.

Eluder Dimension (w.r.t the Squared Hellinger distance). As shown by Foster et al. (2021), the log-loss oracle provides generalization guarantees with respect to the squared Helligner distance.

Definition 2.3 (Squared Hellinger Distance)**.** For any two distributions $\mathbb{P}$, $\mathbb{Q}$ over a discrete support X, the Squared Hellinger Distance is defined as

$$D_H^2(\mathbb{P},\mathbb{Q}) := \sum_{x\in X}\left(\sqrt{\mathbb{P}(x)}-\sqrt{\mathbb{Q}(x)}\right)^2.$$

The Hellinger distance is a bounded metric. Thus, we assume a known upper bound $d_{\mathcal{P}}$ of $d_{\mathrm{E}}(\mathcal{P}, D_H, T^{-1/2})$, the Eluder dimension of $\mathcal{P}$ with respect to Hellinger distance.

Clearly, for a finite class the Eluder dimension is at most the number of functions. For classes of discrete distributions, where the minimum probability is $p > 0$, one can bound the Eluder dimension w.r.t. the Hellinger distance by d_2/p, where d_2 is the (standard) Eluder dimension w.r.t. ℓ_2. (See Lemma A.4).

3. Algorithm and Main Result

We present Eluder Upper Counterfactual Confidence for Contextual Reinforcement Learning (E-UC^{3}RL), given in Algorithm 1. At each episode t, the algorithm estimates the reward and dynamics using the regression oracles. It then constructs an optimistic CMDP using reward bonuses and plays its optimal policy. The reward bonuses are inspired by the notion of *counterfactual confidence*, suggested by Xu & Zeevi (2020) for CMABs. The original idea was to calculate the confidence bounds using the counterfactual actions of past policies given the current context. Levy & Mansour (2023) adapted this approach to CMDPs using the minimum reachability assumption, without which, it becomes crucial to also consider counterfactual states. Notice that the states are stochastically generated by the MDP in response to the agent's played actions. This makes counterfactual state computation impossible without access to the true dynamics. Instead, we consider the counterfactual probabilities of a state-action pair and evaluate this quantity using the estimated dynamics. These probabilities are typically referred to as *occupancy measures* (Zimin & Neu, 2013). Concretely, for any non-contextual policy π and dynamics P, let $q_h(s, a \mid \pi, P)$ denote the probability of reaching state $s \in S$ and performing action $a \in A$ at time $h \in [H]$ of an episode generated using policy π and dynamics P. Note that, given π and P, the occupancy measure of any state-action pair can be computed efficiently using a standard planning algorithm.

At round t and $(s, a, h, c)-$tuple the cumulative occupancy measure of past policies, i.e., $\sum_{i=1}^{t-1} q_h(s, a|\pi_i^c, P_\star^c)$ is a good indicator for the quality of the estimated dynamics and rewards $\widehat{f}_t$ and $\widehat{P}_t$. Thus we would ideally choose bonuses inversely proportional to this quantity. Since $P_\star$ is unknown, it is natural to replace it in $q_h(\cdot)$ with the most recent estimate $\widehat{P}_t$. However, the instability of the oracle estimates $\widehat{P}_t$ means that $q_h(s, a, |\pi_i^c, \widehat{P}_t^c)$ can change arbitrarily with t, which may lead to overly large bonuses. We resolve this instability using the Eluder dimension assumption, which allows us to replace $\widehat{P}_t$ with $\widehat{P}_i$ in $q_h(\cdot)$, thus stabilizing the occupancy measure estimate of π_i as $q_h(s, a, |\pi_i^c, \widehat{P}_i^c)$ for all t. Finally, since our bonuses are based on past context-dependent policies, we first have to compute $\pi_k(c_t; \cdot)$ for all $k \in [t-1]$, which is the purpose of our internal loop (Algorithm 1).

Algorithm 1 Eluder Upper Counterfactual Confidence for Contextual RL (E-UC^{3}RL)

inputs: MDP parameters: S, A, s_0, H;
tuning parameters β_r, β_P.
for round $t = 1, \dots, T$ **do**
compute using the LSR oracle:

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

computed using the LLR oracle:

$$\widehat{P}_t \in \arg\min_{P\in\mathcal{P}} \sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\log\left(\frac{1}{P^{c_i}(s_{h+1}^i|s_h^i, a_h^i)}\right)$$

observe a fresh context $c_t \sim \mathcal{D}$.
for $k = 1, 2, \dots, t$ **do**
compute for all $h \in [H]$ and $(s, a) \in S_h \times A$:

$$\widehat{r}_k^{c_t}(s,a) = \widehat{f}_k(c_t, s, a) + b_k^{\beta_r}(c_t, s, a, h) + Hb_k^{\beta_P}(c_t, s, a, h)$$

where

$$b_k^{\beta}(c, s, a, h) = \min\left(1, \frac{\beta/2}{1+\sum_{i=1}^{k-1} q_h(s, a|\pi_i(c;\cdot), \widehat{P}_i^c)}\right).$$

define $\widehat{\mathcal{M}}_k(c_t) = (S, A, \widehat{P}_k^{c_t}, \widehat{r}_k^{c_t}, s_0, H)$.
compute using a planning algorithm:

$$\pi_k(c_t;\cdot) \in \arg\max_{\pi:S\to A} V^{\pi}_{\widehat{\mathcal{M}}_k(c_t)}(s_0).$$

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

The following is our main result for Algorithm 1. We sketch its proof in Section 4, and defer the complete proof to Appendix B.5.

Theorem 3.1 (E-UC^{3}RL regret bound)**.** *For any $T > 1$ and*

$\delta \in (0,1)$, *suppose we run Algorithm 1 with parameters*

$$\beta_r = \sqrt{\frac{504TH^2 d_{\mathcal{P}} \log^2(64T^4 H|\mathcal{F}||\mathcal{P}|/\delta^2)}{|S||A|\log(T+1)}},$$

$$\beta_P = \sqrt{\frac{2029TH^2 d_{\mathcal{P}} \log^2(8TH|\mathcal{P}|/\delta)}{|S||A|\log(T+1)}},$$

and $d_{\mathcal{P}} \geq d_{\mathrm{E}}(\mathcal{P}, D_H, T^{-1/2})$. *Then, with probability at least* $1-\delta$ *it holds that*

$$\mathcal{R}_T(\textit{E-UC}^3\textit{RL}) \leq \widetilde{O}(H^3\sqrt{T|S||A|d_{\mathcal{P}}\left(\log(|\mathcal{F}||\mathcal{P}|/\delta)\right)}).$$

We remark that using covering numbers analysis (Shalev-Shwartz & Ben-David, 2014), our result naturally generalizes to infinite function classes as well as context spaces. In addition, when comparing our regret upper bound to the lower bound of Levy & Mansour (2023), there is an apparent gap of $H^{2.5}$ and $d_{\mathcal{P}}$ factors. We leave this gap for future research.

Computational efficiency of E-UC^{3}RL. The algorithm calls each oracle T times, making it oracle-efficient (since it's oracle-call complexity is in $poly(T)$). Aside from simple arithmetic operations, each of the $T(T+1)/2$ iterations of the internal loop call one MDP planning procedure and calculate the related occupancy measure. Both of these can be implemented efficiently using dynamic programming. Overall, excluding the oracle's computation time, the runtime complexity of our algorithm is in $\mathrm{poly}(T, |S|, |A|, H)$. Hence, if both the LSR and LLR oracles are computationally efficient then E-UC^{3}RL is also computationally efficient.

4. Analysis

Our analysis consists of four main steps:

(i) Establish an upper bound on the expected regret of the square and log loss regression oracles;
(ii) Construct confidence bounds over the expected value of any context-dependent policy for both dynamics and rewards;
(iii) Define the optimistic approximated CMDP and establish optimism lemmas;
(iv) Combine the above to derive a high probability regret bound.

In what follows, we present the main claims of our analysis, deferring the proofs to Appendix B. Before beginning, we discuss some of the challenges and present a key technical result, the value change of measure lemma (Lemma 4.1).

A Key Technical Challenge

Our goal is to derive computable and reliable confidence bounds over the expected value of any policy. The difficulty is that the offline regression oracles have regret guarantees only with respect to the trajectories' distributions, which are related to the true context-dependent dynamics $P_\star$. Hence, a main technical challenge is to translate the oracle's regret to a guarantee with respect to the estimated context-dependent dynamics $\widehat{P}_t$. Notice that the confidence bounds are computable only if stated in terms of $\widehat{P}_t$. Following ideas from Foster et al. (2021), we solve this issue using a multiplicative value change of measure that is based on the Hellinger distance. Concretely, the following change of measure lemma allows us to measure the value difference caused by the use of approximated transition probabilities in terms of the expected cumulative Hellinger distance (proof in Appendix B.1).

Lemma 4.1 (Value change of measure)**.** *Let* $r : S \times A \to [0,1]$ *be a bounded expected rewards function. Let* $P_\star$ *and* $\widehat{P}$ *denote two dynamics and consider the MDPs* $M = (S, A, P_\star, r, s_0, H)$ *and* $\widehat{M} = (S, A, \widehat{P}, r, s_0, H)$. *Then, for any policy* π *it holds that*

$$V^\pi_{\widehat{M}}(s) \leq 3V^\pi_M(s) + 9H^2 \mathop{\mathbb{E}}_{P_\star,\pi}\left[\sum_{h=0}^{H-1} D^2_H(\widehat{P}(\cdot|s_h,a_h), P_\star(\cdot|s_h,a_h)) \,\middle|\, s_0 = s\right].$$

Notice that this bound is loose when the reward function is not small. However, it is significantly tighter than standard results when the rewards are small. For instance, later in the analysis we consider the reward $r^c(s,a) = (\widehat{f}_t(c,s,a) - f_\star(c,s,a))^2$ that is the squared reward approximation error. Letting $\widehat{\mathcal{M}} = (S, A, \widehat{P}, r^c, s_0, H)$ and $\mathcal{M} = (S, A, P_\star, r^c, s_0, H)$, Lemma 4.1 implies that the expected reward approximation error with respect to $\widehat{P}$ is at most a constant multiple of the expected reward and dynamics approximation errors with respect to $P_\star$. In contrast, a standard change of measure replaces the squared Hellinger distance with Total Variation (TV) whose cumulative error scales as $\sqrt{T}$.

Step 1: Establishing Oracle Guarantees

The regret guarantees of the least-squares oracle were established in Levy & Mansour (2023, Lemma B.10), stated in the Appendix as Lemma B.5. The following corollary bounds the cumulative expected least-squares loss of the sequence of the oracle's predictions (proof in Appendix B.2). In the following, we denote the expected squared error at round t over a trajectory generated by π where the context is c as $\mathcal{E}^t_{\mathrm{sq}}(\pi, c)$. Formally,

$$\mathcal{E}^t_{\mathrm{sq}}(\pi, c) := \mathop{\mathbb{E}}_{\pi(c;\cdot),P^c_\star}\left[\sum_{h=0}^{H-1}\left(\widehat{f}_t(c,s_h,a_h) - f_\star(c,s_h,a_h)\right)^2 \,\middle|\, s_0\right].$$

Corollary 4.2 (Reward approximation bound)**.** *Let $\widehat{f}_t \in \mathcal{F}$ be the least squares minimizer of Line 3 in Algorithm 1. For any $\delta \in (0,1)$ it holds with probability at least $1-\delta$,*

$$\mathbb{E}_c\left[\sum_{i=1}^{t-1} \mathcal{E}^t_{\mathrm{sq}}(\pi_i, c)\right] \leq 68H\log(2T^3|\mathcal{F}|/\delta), \qquad \forall t \geq 1.$$

Next, we analyze the expected regret of the dynamic's log-loss oracle in terms of the Hellinger distance. The following result is a straightforward application of Lemma A.14 in Foster et al. (2021) (proof in Appendix B.2). To that end, we denote the expected squared Hellinger distance at round t over a trajectory generated by π where the context is c as $\mathcal{E}^t_{\mathrm{H}}(\pi, c)$. Formally,

$$\mathcal{E}^t_{\mathrm{H}}(\pi, c) := \mathop{\mathbb{E}}_{\pi(c;\cdot), P^c_\star}\left[\sum_{h=0}^{H-1} D^2_H(P^c_\star(\cdot|s_h,a_h), \widehat{P}^c_t(\cdot|s_h,a_h)) \,\middle|\, s_0\right].$$

Corollary 4.3 (Dynamics approximation bound)**.** *Let $\widehat{P}_t \in \mathcal{P}$ be the log loss minimizer of Line 4 in Algorithm 1. For any $\delta \in (0,1)$ it holds that with probability at least $1-\delta$,*

$$\mathbb{E}_c\left[\sum_{i=1}^{t-1} \mathcal{E}^t_{\mathrm{H}}(\pi_i, c)\right] \leq 2H\log(TH|\mathcal{P}|/\delta), \qquad \forall t \geq 1.$$

Lastly, we apply a variant of Corollary 4.3 together with the shrinking confidence bound guarantee in Lemma 2.2 to derive the following bound (proof in Appendix B.2).

Lemma 4.4 (Stability error of log-loss oracle)**.** *Let $\widehat{P}_i \in \mathcal{P}$ denote the log-loss minimizer at round $i \in [T]$. For any $\delta \in (0,1)$ it holds with probability at least $1-\delta$ that*

$$\mathbb{E}_c\left[\sum_{i=1}^{t-1} \mathcal{E}^i_{\mathrm{H}}(\pi_i, c)\right] \leq 112Hd_{\mathcal{P}}\log^2(2TH|\mathcal{P}|/\delta), \ \forall t \geq 1,$$

where $d_{\mathcal{P}} \geq d_{\mathrm{E}}(\mathcal{P}, D_H, T^{-1/2})$, the Eluder dimension of $\mathcal{P}$ at scale $T^{-1/2}$.

One of the novel contributions of our work is the use of the Eluder dimension to bound the stability error due to the log-loss oracle. This allows us to choose stable bonuses that yield valid and computable confidence bounds.

Step 2: Constructing Confidence Bounds

Our main goal in this subsection is to upper bound w.h.p the expected value difference between the true and the empirical CMDPs, for any context-dependent policy π (Corollary 4.7). For that purpose, we derive confidence bounds over the rewards and dynamics approximation. Let π_t denote the context-dependent policy selected at round t. For any $h \in [H]$ and state-action pair $(s,a) \in S_h \times A$, context $c \in \mathcal{C}$, and round $t \geq 1$, we define the reward bonuses

$$\begin{aligned} b^R_{t,h}(c,s,a) &:= b^{\beta_r}_t(c,s,a,h), \\ b^P_{t,h}(c,s,a) &:= Hb^{\beta_P}_t(c,s,a,h), \end{aligned} \tag{1}$$

where $b^\beta_t(c,s,a,h)$ is defined in Line 7. $b^R_{t,h}$ is the bonus related to the rewards approximation error, and $b^P_{t,h}$ is the bonus related to that of the dynamics. We remark that these bonuses differ only in constant terms ($H\beta_P$ versus β_r), and are identical to the bonus terms defined in Algorithm 1 (Line 7). We use these bonuses in our optimistic construction to account for the approximation errors in the rewards and dynamics, respectively. Next, for any context $c \in \mathcal{C}$ and functions $f \in \mathcal{F}, P \in \mathcal{P}$ we define the MDP $\mathcal{M}^{(f,P)}(c) = (S, A, P^c, f(c,\cdot,\cdot), s_0, H)$. The following results derive confidence bounds for the dynamics and rewards approximation in terms of the reward bonuses and approximation errors (proofs in Appendix B.3).

Lemma 4.5 (Confidence bound for rewards approximation w.r.t the approximated dynamics)**.** *Let $P_\star$ and $f_\star$ be the true context-dependent dynamics and rewards. Let $\widehat{P}_t$ and $\widehat{f}_t$ be the approximated context-dependent dynamics and rewards at round t. Then, for any $t \geq 1$, and context-dependent policy $\pi \in \Pi_{\mathcal{C}}$,*

$$\begin{aligned} \left|\mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}^{(f_\star,\widehat{P}_t)}(c)}(s_0)\right] - \mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{f}_t,\widehat{P}_t)}(c)}(s_0)\right]\right| &\leq \frac{H}{2\beta_r} \\ &+ \frac{3}{2\beta_r}\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathcal{E}^t_{\mathrm{sq}}(\pi_i,c)\right] + \frac{9H^2}{2\beta_r}\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathcal{E}^i_{\mathrm{H}}(\pi_i,c)\right] \\ &+ \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s\in S^c_h}\sum_{a\in A} q_h(s,a|\pi(c;\cdot),\widehat{P}^c_t)\cdot b^R_{t,h}(c,s,a)\right]. \end{aligned}$$

Lemma 4.6 (Confidence bound for dynamics approximation w.r.t the true rewards $f_\star$)**.** *Let $P_\star$ and $f_\star$ be the true context-dependent dynamics and rewards. Let $\widehat{P}_t$ be the approximated context-dependent dynamics at round t. Then, for any $t \geq 1$, and context-dependent policy $\pi \in \Pi_{\mathcal{C}}$,*

$$\begin{aligned} \left|\mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}^{(f_\star,P_\star)}(c)}(s_0)\right] - \mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}^{(f_\star,\widehat{P}_t)}(c)}(s_0)\right]\right| &\leq \frac{H^2}{2\beta_P} \\ &+ \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s\in S^c_h}\sum_{a\in A} q_h(s,a|\pi(c;\cdot),\widehat{P}^c_t)\cdot b^P_{t,h}(c,s,a)\right] \\ &+ \frac{6H}{\beta_P}\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathcal{E}^t_{\mathrm{H}}(\pi_i,c)\right] + \frac{18H^3}{\beta_P}\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathcal{E}^i_{\mathrm{H}}(\pi_i,c)\right]. \end{aligned}$$

The proofs of Lemmas 4.5 and 4.6 are similar and can be found in Appendix B.3. By applying the high probability approximation bounds in Corollaries 4.2 and 4.3

and Lemma 4.4 to Lemmas 4.5 and 4.6, we obtain the desired high probability confidence bound on the expected value approximation (proof in Appendix B.3).

Corollary 4.7. *Under the terms of Lemmas 4.5 and 4.6, the following holds with probability at least $1 - 3\delta/4$ simultaneously for all $t \geq 1$ and $\pi \in \Pi_{\mathcal{C}}$.*

$$\left|\mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}^{(f_\star,P_\star)}(c)}(s_0)\right] - \mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{f}_t,\widehat{P}_t)}(c)}(s_0)\right]\right| \leq$$
$$\mathbb{E}_c\Big[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c) b^R_{t,h}(c,s_h,a_h)\Big]$$
$$+\mathbb{E}_c\Big[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c) b^P_{t,h}(c,s_h,a_h)\Big]$$
$$+\frac{2029H^4 d_{\mathcal{P}}}{\beta_P}\log^2(8TH|\mathcal{P}|/\delta)$$
$$+\frac{504H^3 d_{\mathcal{P}}}{\beta_r}\log^2(64T^4H|\mathcal{F}||\mathcal{P}|/\delta^2).$$

Step 3: Establishing Optimism Lemmas

In this subsection, we use the results of step 2 (Corollary 4.7) to establish the properties of the optimistic approximated CMDP $\widehat{\mathcal{M}}_t$. Namely, that its optimal value is higher than that of the true CMDP $\mathcal{M}$ (Lemma 4.8), but also that the optimistic value of π_t is not significantly higher than its true value (Lemma 4.9), in expectation over the context. Combining both results and applying Azuma's inequality yields the regret bound.

We begin by defining the optimistic-in-expectation context-dependent reward function at every round $t \geq 1$ as $\widehat{r}^c_t(s,a) := \widehat{f}_t(c,s,a) + b^R_{t,h}(c,s,a) + b^P_{t,h}(c,s,a)$, where the bonuses are defined in Equation (1), and we note that $\widehat{r}^c_t(s,a) \in [0, H+2]$ for all $h \in [H]$ and $(c,s,a) \in \mathcal{C} \times S_h \times A$. The approximated optimistic-in-expectation CMDP at round t is defined as $(\mathcal{C}, S, A, \widehat{\mathcal{M}}_t)$ where for any context $c \in \mathcal{C}$ we define $\widehat{\mathcal{M}}_t(c) := (S, A, \widehat{P}^c, \widehat{r}^c, s_0, H)$. We also recall that $\mathcal{M}(c) = \mathcal{M}^{(f_\star,P_\star)}(c)$ is the true CMDP. The next two Lemmas establish the properties of the optimistic CMDP (proofs in Appendix B.4).

Lemma 4.8 (Optimism in expectation)**.** *Let $\pi_\star$ be an optimal context-dependent policy for $\mathcal{M}$. Under the good event of Corollary 4.7, for any $t \geq 1$ it holds that*

$$\mathbb{E}_c\left[V^{\pi_\star(c;\cdot)}_{\mathcal{M}(c)}(s_0)\right] \leq \mathbb{E}_c\left[V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)\right]$$
$$+\frac{2029H^4 d_{\mathcal{P}}}{\beta_P}\log^2(8TH|\mathcal{P}|/\delta)$$
$$+\frac{504H^3 d_{\mathcal{P}}}{\beta_r}\log^2\big(64T^4H|\mathcal{F}||\mathcal{P}|/\delta^2\big).$$

Lemma 4.9 (The cost of approximation)**.** *Under the good event of Corollary 4.7, we have that for every $t \geq 1$*

$$\mathbb{E}_c\left[V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)\right] \leq \mathbb{E}_c\left[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)\right]$$
$$+2\sum_{h=0}^{H-1}\mathbb{E}_c\left[\mathop{\mathbb{E}}_{\pi_t(c;\cdot),\widehat{P}_t^c}\left[b^R_{t,h}(c,s_h,a_h) + b^R_{t,h}(c,s_h,a_h)\right]\right]$$
$$+\frac{2029H^4 d_{\mathcal{P}}}{\beta_P}\log^2(8TH|\mathcal{P}|/\delta)$$
$$+\frac{504H^3 d_{\mathcal{P}}}{\beta_r}\log^2(64T^4H|\mathcal{F}||\mathcal{P}|/\delta^2).$$

Step 4: Deriving the Regret Bound

Using the above results, we derive Theorem 3.1 as follows. Summing Lemmas 4.8 and 4.9 over $1 \leq t \leq T$ bounds a notion of expected regret. Next, we use a standard algebraic argument (Lemma B.18) to bound the expected cumulative bonuses as

$$\sum_{t=1}^{T}\sum_{h=0}^{H-1}\mathbb{E}_c\left[\mathop{\mathbb{E}}_{\pi_t(c;\cdot),\widehat{P}_t^c}\left[b^R_{t,h}(c,s_h,a_h) + b^R_{t,h}(c,s_h,a_h)\right]\right]$$
$$\leq H|S||A|(\beta_r + H\beta_P)\log(T+1).$$

By plugging in our choice of β_P and β_r we bound the expected regret by

$$271H^3\sqrt{T|S||A|d_{\mathcal{P}}\log(T+1)\log^2(18T^4H|\mathcal{F}||\mathcal{P}|/\delta^2)}.$$

We derive the high probability result by using Corollary 4.7 and applying Azuma's inequality. The complete proof is in Appendix B.5.

5. Discussion and Conclusion

In this paper, we make a step forward in understanding RL with offline function approximation. We consider the tabular CMDP setting, under the offline function approximation assumption, and obtain a rate-optimal regret bound. To obtain our result, we extend the Eluder dimension presented by Russo & Van Roy (2013) to a general bounded metric, rather than only the ℓ_2 norm. This result may be of separate interest. Further, by applying the Metric Eluder dimension with Hellinger distance, we obtain our algorithm EUC^{3}RL, the first efficient algorithm for regret minimization in Contextual MDPs that uses offline regression oracles. We note that our algorithm requires a known bound on the Eluder dimension and its regret depends on it. Obtaining an efficient algorithm that has rate-optimal regret using offline oracles but without dependence on the Eluder dimension is an important open question for future research. Extending our technique to RL with rich observations is also an interesting direction for future research.

Acknowledgements

We would like to thank the reviewers for their helpful comments.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396 and grant agreement No. 101078075). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work received additional support from the Israel Science Foundation (ISF, grant numbers 993/17 and 2549/19), Tel Aviv University Center for AI and Data Science (TAD), the Yandex Initiative for Machine Learning at Tel Aviv University, the Len Blavatnik and the Blavatnik Family Foundation, and by the Israeli VATAT data science scholarship.

AC is supported by the Israeli Science Foundation (ISF) grant no. 2250/22.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Agarwal, A., Dudík, M., Kale, S., Langford, J., and Schapire, R. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, pp. 19–26. PMLR, 2012.

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pp. 1638–1646. PMLR, 2014.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. Model-based reinforcement learning with value-targeted regression. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 463–474. PMLR, 13–18 Jul 2020.

Chen, Z., Li, C. J., Yuan, A., Gu, Q., and Jordan, M. I. A general framework for sample-efficient function approximation in reinforcement learning. *arXiv preprint arXiv:2209.15634*, 2022.

Dann, C., Mohri, M., Zhang, T., and Zimmert, J. A provably efficient model-free posterior sampling method for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12040–12051, 2021.

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

Foster, D., Agarwal, A., Dudik, M., Luo, H., and Schapire, R. Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, pp. 1539–1548. PMLR, 2018.

Foster, D. J. and Krishnamurthy, A. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. *Advances in Neural Information Processing Systems*, 34:18907–18919, 2021.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Hallak, A., Di Castro, D., and Mannor, S. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 13406–13418. Curran Associates, Inc., 2021.

Langford, J. and Zhang, T. The epoch-greedy algorithm for multi-armed bandits with side information. In Platt, J., Koller, D., Singer, Y., and Roweis, S. (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

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

Levy, O. and Mansour, Y. Learning efficiently function approximation for contextual MDP. *arXiv preprint arXiv:2203.00995*, 2022.

Levy, O. and Mansour, Y. Optimism in face of a context: Regret guarantees for stochastic contextual mdp. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8510–8517, 2023.

Levy, O., Cohen, A., Cassel, A. B., and Mansour, Y. Efficient rate optimal regret for adversarial contextual mdps using online function approximation. In *ICML*, 2023.

Liu, Q., Chung, A., Szepesvári, C., and Jin, C. When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, pp. 5175–5220. PMLR, 2022.

Mannor, S., Mansour, Y., and Tamar, A. *Reinforcement Learning: Foundations*. Online manuscript; https://sites.google.com/view/rlfoundations/home, 2022. accessed March-05-2023.

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

Modi, A., Jiang, N., Singh, S., and Tewari, A. Markov decision processes with continuous side information. In *Algorithmic Learning Theory*, pp. 597–618. PMLR, 2018.

Osband, I. and Van Roy, B. Model-based reinforcement learning and the eluder dimension. *Advances in Neural Information Processing Systems*, 27, 2014.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Rosenberg, A., Cohen, A., Mansour, Y., and Kaplan, H. Near-optimal regret bounds for stochastic shortest path. In *International Conference on Machine Learning*, pp. 8210–8219. PMLR, 2020.

Russo, D. and Van Roy, B. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

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

Shani, L., Efroni, Y., Rosenberg, A., and Mannor, S. Optimistic policy optimization with bandit feedback. In *ICML*, 2020.

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

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

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pp. 2898–2933. PMLR, 2019.

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

Wang, R., Salakhutdinov, R. R., and Yang, L. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6123–6135. Curran Associates, Inc., 2020a.

Wang, R., Salakhutdinov, R. R., and Yang, L. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33:6123–6135, 2020b.

Wen, Z. and Van Roy, B. Efficient reinforcement learning in deterministic systems with value function generalization. *Mathematics of Operations Research*, 42(3): 762–782, 2017.

Wu, Y., He, J., and Gu, Q. Uniform-PAC guarantees for model-based RL with bounded eluder dimension. In Evans, R. J. and Shpitser, I. (eds.), *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2304–2313. PMLR, 31 Jul–04 Aug 2023.

Xie, T., Foster, D. J., Bai, Y., Jiang, N., and Kakade, S. M. The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157*, 2022.

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

Zhang, T. Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857, 2022.

Zimin, A. and Neu, G. Online learning in episodic markovian decision processes by relative entropy policy search. *Advances in neural information processing systems*, 26, 2013.

A. Metric Eluder Dimension

Our goal is to prove Lemma 2.2, a variant of Proposition 6 in Osband & Van Roy (2014). Recall that for any $\mathcal{P}' \subseteq \mathcal{P}$, its radius at $x \in \mathcal{X}$ is

$$w_{\mathcal{P}'}(x) = \sup_{P, P' \in \mathcal{P}'} D(P(x), P'(x)).$$

Now, for any $t \in [T], h \in [H]$ let $x_h^t \in \mathcal{X}$ and $P_t \in \mathcal{P}$ be arbitrary. Define the confidence sets with parameter β as

$$\mathcal{P}_t = \left\{ P \in \mathcal{P} \ : \ \sum_{i=0}^{t-1} \sum_{h=0}^{H-1} D^2(P(x_h^i), P_t(x_h^i)) \leq \beta \right\}.$$

Lemma A.1 (Restatement of Lemma 2.2)**.** *Suppose that $\beta \geq H$. We have that*

$$\sum_{t=1}^{T} \sum_{h=0}^{H-1} (w_{\mathcal{P}_t}(x_h^t))^2 \leq 6 d_{\mathrm{E}}(\mathcal{P}, D, T^{-1/2}) \beta \log T.$$

We first need the following result, which is a direct adaptation of Lemma 1 in (Osband & Van Roy, 2014) (see proof below for completeness).

Lemma A.2. *We have that for any $\epsilon > 0$*

$$\sum_{t=1}^{T} \sum_{h=0}^{H-1} \mathbb{I}[w_{\mathcal{P}_t}(x_h^t) > \epsilon] \leq \left(\frac{4\beta}{\epsilon^2} + H \right) d_{\mathrm{E}}(\mathcal{P}, D, \epsilon).$$

Proof of Lemma 2.2. To reduce notation, write $w_{t,h} = w_{\mathcal{P}_t}(x_h^t)$, and $d = d_{\mathrm{E}}(\mathcal{P}, D, T^{-1/2})$. Next, reorder the sequence $(w_{1,0}, \ldots, w_{1,H-1}, \ldots, w_{T,0}, \ldots, w_{T,H-1}) \to (w_{i_1}, \ldots, w_{i_{TH}})$ where $w_{i_1} \geq w_{i_2} \geq \ldots \geq w_{i_{TH}}$. Then we have that

$$\begin{aligned}
\sum_{t=1}^{T} \sum_{h=0}^{H-1} (w_{\mathcal{P}_t}(x_h^t))^2 = \sum_{s=1}^{TH} w_{i_s}^2 &= \sum_{s=1}^{TH} w_{i_s}^2 \mathbb{I}[w_{i_s} \leq T^{-1/2}] + \sum_{s=1}^{TH} w_{i_s}^2 \mathbb{I}[w_{i_s} > T^{-1/2}] \\
&\leq \sum_{s=1}^{TH} \frac{1}{T} + \sum_{s=1}^{TH} w_{i_s}^2 \mathbb{I}[w_{i_s} > T^{-1/2}] \\
&= H + \sum_{s=1}^{T} w_{i_s}^2 \mathbb{I}[w_{i_s} > T^{-1/2}].
\end{aligned}$$

Now, let $\epsilon \geq T^{-1/2}$ and suppose that $w_{i_s} > \epsilon$. Since w_{i_s} are ordered in descending order, this implies that $\sum_{t=1}^{T} \sum_{h=0}^{H-1} \mathbb{I}(w_{\mathcal{P}_t}(x_h^t) > \epsilon) \geq s$. On the other hand, by Lemma A.2, we have

$$\sum_{t=1}^{T} \sum_{h=0}^{H-1} \mathbb{I}(w_{\mathcal{P}_t}(x_h^t) > \epsilon) \leq \left(\frac{4\beta}{\epsilon^2} + H \right) d_{\mathrm{E}}(\mathcal{P}, D, \epsilon) \leq \left(\frac{4\beta}{\epsilon^2} + H \right) d,$$

where the second transition used that $d_{\mathrm{E}}(\mathcal{P}, D, \epsilon')$ is non-increasing in ϵ'. We conclude that $s \leq \big(4\beta/\epsilon^2 + H\big)d$, and changing sides gives that $\epsilon^2 \leq 4\beta/(s - dH)$. This implies that

$$w_{i_s} > T^{-1/2} \implies w_{i_s}^2 \leq \frac{4\beta d}{s - dH}.$$

We thus have

$$\begin{aligned}
\sum_{s=1}^{TH} w_{i_s}^2 \mathbb{I}(w_{i_s} > T^{-1/2}) &\leq 1 + dH + \sum_{s=dH+2}^{T} w_{i_s}^2 \mathbb{I}(w_{i_s} > T^{-1/2}) && (w_{i_s} \leq 1) \\
&\leq 1 + dH + \sum_{s=dH+2}^{T} \frac{4\beta d}{s - dH} \\
&\leq 1 + dH + 4\beta d \left(\int_{t=1}^{T} \frac{1}{t} dt \right) \\
&= 1 + dH + 4\beta d \log T \\
&\leq 6 d\beta \log T. && (\beta \geq H)
\end{aligned}$$

□

Proof of Lemma A.2. The proof follows similarly to Lemma 1 in Osband & Van Roy (2014). We begin by showing that if $w_{\mathcal{P}_t}(x_h^t) > \epsilon$ then x_h^t is $(D, \epsilon)-$dependent on fewer than $4\beta/\epsilon^2$ disjoint sub-sequences of $(x_0^1, \ldots, x_{H-1}^1, \ldots, x_0^{t-1}, \ldots, x_{H-1}^{t-1})$. To see this, note that if $w_{\mathcal{P}_t}(x_h^t) > \epsilon$ there are $P, P' \in \mathcal{P}_t$ such that

$$D(P(x_h^t), P'(x_h^t)) > \epsilon.$$

Now, suppose that x_h^t is $(D, \epsilon)-$dependent on a sub-sequence $(x_{h_1}^{t_1}, \ldots, x_{h_n}^{t_n})$ of $(x_0^1, \ldots, x_{H-1}^1, \ldots, x_0^{t-1}, \ldots, x_{H-1}^{t-1})$. Since $D(P(x_h^t), P'(x_h^t)) \geq \epsilon$, the dependence implies that

$$\sum_{j=1}^{n} D^2(P(x_{h_j}^{t_j}), P'(x_{h_j}^{t_j})) > \epsilon^2.$$

We conclude that, if x_h^t is $\epsilon-$dependent on K disjoint sub-sequences of $(x_0^1, \ldots, x_{H-1}^1, \ldots, x_0^{t-1}, \ldots, x_{H-1}^{t-1})$ then

$$\sum_{i=1}^{t-1} \sum_{h=0}^{H-1} D^2(P(x_h^i), P'(x_h^i)) > K\epsilon^2.$$

On the other hand, since $P, P' \in \mathcal{P}_t$, we have

$$\begin{aligned}
\sum_{i=1}^{t-1} \sum_{h=0}^{H-1} D^2(P(x_h^i), P'(x_h^i)) &\leq \sum_{i=1}^{t-1} \sum_{h=0}^{H-1} \left[D(P(x_h^i), P_t(x_h^i)) + D(P_t(x_h^i), P'(x_h^i)) \right]^2 && \text{(triangle inequality for } D) \\
&\leq 2 \sum_{i=1}^{t-1} \sum_{h=0}^{H-1} D^2(P(x_h^i), P_t(x_h^i)) + 2 \sum_{i=1}^{t-1} \sum_{h=0}^{H-1} D^2(P_t(x_h^i), P'(x_h^i)) \\
& && ((a+b)^2 \leq 2(a^2 + b^2)) \\
&\leq 4\beta.
\end{aligned}$$

It follows that $K < 4\beta/\epsilon^2$.

Next, denote $d := d_{\mathrm{E}}(\mathcal{P}, D, \epsilon)$. We show that in any action sequence $(y_1, \ldots, y_\tau)$, there is some element y_j that is $(D, \epsilon)-$dependent on at least $\tau/d - 1$ disjoint sub-sequences of $(y_1, \ldots, y_{j-1})$. To show this, we will construct $K = \lceil (\tau/d) - 1 \rceil$ disjoint sub-sequences $B_1, \ldots, B_K$. First, let $B_m = (y_m)$ for $m = 1, \ldots, K$. If y_{K+1} is $(D, \epsilon)-$dependent on each sub-sequence $B_1, \ldots, B_K$, the claim is established. Otherwise, append y_{K+1} to a sub-sequence B_m that y_{K+1} is $(D, \epsilon)-$independent of. Repeat this process for elements with indices $i > K + 1$ until y_i is $\epsilon-$dependent on all sub-sequences. Suppose in contradiction that the process terminated without finding the desired y_i. By the definition of the Eluder dimension, each B_m must satisfy $|B_m| \leq d$ and thus $\sum_{m=1}^{K} |B_m| \leq Kd$. On the other hand

$$Kd = \lceil (\tau/d) - 1 \rceil d < \tau = \sum_{m=1}^{K} |B_m| \leq Kd,$$

where the last equality follows since all elements of $(y_1, \dots, y_\tau)$ where placed. This is a contradiction, and thus the process must terminate successfully.

Now, let

$$\tau = \sum_{t=1}^{T} \sum_{h=0}^{H-1} \mathbb{I}[w_{\mathcal{P}_t}(x_h^t) > \epsilon],$$

and take $(y_1, \dots, y_\tau)$ to be the sub-sequence $(x_{h_1}^{t_1}, \dots, x_{h_\tau}^{t_\tau})$ of elements x_h^t for which $w_{\mathcal{P}_t}(x_h^t) > \epsilon$. Then there exists $x_{h_j}^{t_j}$ that depends on at least $\tau/d - 1$ disjoint sub-sequences $(x_{h_1}^{t_1}, \dots, x_{h_\tau}^{t_\tau})$. Notice that at most $H-1$ of these sub-sequences are not sub-sequences of $(x_0^1, \dots, x_{H-1}^1, \dots, x_0^{t_j-1}, \dots, x_{H-1}^{t_j-1})$. We conclude that $x_{h_j}^{t_j}$ depends on at least $(\tau/d) - H$ disjoint sub-sequences of $(x_0^1, \dots, x_{H-1}^1, \dots, x_0^{t_j-1}, \dots x_{H-1}^{t_j-1})$. On the other hand, since $w_{\mathcal{P}_{t_j}}(x_{h_j}^{t_j}) > \epsilon$ then by the first part of the proof it depends on fewer than $4\beta/\epsilon^2$ such disjoint sub-sequences, thus

$$\frac{\tau}{d} - H \leq \frac{4\beta}{\epsilon^2},$$

so $\tau \leq (4\beta/\epsilon^2 + H)d$ as desired. □

A.1. Upper bound using the ℓ_2 Eluder dimension

Proposition A.3. *Assume that P, Q are two distributions over finite domain $\mathcal{X}$, such that for all $x \in \mathcal{X}$ it holds that $P(x), Q(x) \geq p$. Then,*

$$\|P - Q\|_2^2 \leq 4D_H^2(P, Q) \leq \frac{1}{p}\|P - Q\|_2^2$$

Lemma A.4. *Let function class $\mathcal{P}$, and assume $P(x) \geq p$ holds for some $p > 0$ for every $P \in \mathcal{P}$ and x in the domain. Let $\epsilon \in (0, 1]$ and assume the the Eluder dimension w.r.t the ℓ_2 at scale $\sqrt{p}\epsilon$ is d. Then, the Eluder dimension w.r.t the Squared Hellinger Distance at scale ϵ is upper bounded by d/p.*

Proof. Let d be the Eluder dimension w.r.t the ℓ_2-norm at scale $\sqrt{p}\epsilon$ of the class $\mathcal{P}$. Let $k > \lceil 1/p \rceil$ and $m = dk$. Also let $x_1, x_2, \dots, x_{m+1}$ be an arbitrary sequence. We show that there must exist $j \in [m+1]$ such that x_j is ϵ dependent on its prefix in Hellinger distance, thus the Hellinger Eluder dimension at scale ϵ is at most m. As in Lemma A.2, there exists n and $\{I_j\}_{j \in [k]}$ disjoint subsequences such that $\cup_{j \in [k]} I_j = [n-1]$, and x_n is $\sqrt{p}\epsilon$ dependent in ℓ_2 on each $I_j, j \in [k]$.

Now, let $P, P' \in \mathcal{P}$ such that $D_H^2(P(x_n), P'(x_n)) \geq \epsilon^2$. Then, by Proposition A.3 we have that $\|P(x_n) - P'(x_n)\|_2^2 \geq 4p\epsilon^2$. Because of the dependence, for all $j \in [k]$ we have that $\sum_{\tau \in I_j} \|P(x_\tau) - P'(x_\tau)\|_2^2 \geq 4p\epsilon^2$. We conclude, using Proposition A.3 again, that

$$\sum_{\tau=1}^{n-1} D_H^2(P(x_\tau), P'(x_\tau)) \geq \frac{1}{4} \sum_{\tau=1}^{n-1} \|P(x_\tau) - P'(x_\tau)\|_2^2 = \frac{1}{4} \sum_{j \in [k]} \sum_{\tau \in I_j} \|P(x_\tau) - P'(x_\tau)\|_2^2 \geq \sum_{j \in [k]} p\epsilon^2 = pk\epsilon^2 > \epsilon^2,$$

thus x_n is ϵ dependent on its prefix in Hellinger distance. We conclude that no sequence of length $m+1$ can have all elements ϵ independent of their prefix in Hellinger distance. Thus, the Hellinger Eluder at scale ϵ is at most $m \approx d/p$.

□

B. Proofs

B.1. Multiplicative Value Change of Measure

First, we give the following Bernstein type tail bound (see e.g., Rosenberg et al., 2020, Lemma D.4).

Lemma B.1. *Let $\{X_t\}_{t \geq 1}$ be a sequence of random variables with expectation adapted to a filtration $\mathcal{F}_t$. Suppose that $0 \leq X_t \leq 1$ almost surely. Then with probability at least $1 - \delta$*

$$\sum_{t=1}^{T} \mathbb{E}[X_t \mid \mathcal{F}_{t-1}] \leq 2 \sum_{t=1}^{T} X_t + 4 \log \frac{2}{\delta}$$

Recall the Helligner distance given in Definition 2.3. The following change of measure result is due to (Foster et al., 2021).

Lemma B.2 (Lemma A.11 in Foster et al., 2021)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures on $(\mathcal{X}, \mathrm{F})$. For all $h : \mathcal{X} \to \mathbb{R}$ with $0 \leq h(X) \leq R$ almost surely under $\mathbb{P}$ and $\mathbb{Q}$, we have*

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

In particular,

$$\mathbb{E}_{\mathbb{P}}[h(X)] \leq 3\mathbb{E}_{\mathbb{Q}}[h(X)] + 4RD_H^2(\mathbb{P}, \mathbb{Q}).$$

Next, we need the following refinement of the previous result.

Corollary B.3. *For any $\beta \geq 1$,*

$$\mathbb{E}_{\mathbb{P}}[h(X)] \leq (1 + 1/\beta)\mathbb{E}_{\mathbb{Q}}[h(X)] + 3\beta RD_H^2(\mathbb{P}, \mathbb{Q}).$$

Proof. Let $\eta \in (0, 1)$. Consider the following derivation.

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}[h(X)] - \mathbb{E}_{\mathbb{Q}}[h(X)] &\leq \sqrt{2R(\mathbb{E}_{\mathbb{P}}[h(X)] + \mathbb{E}_{\mathbb{Q}}[h(X)]) \cdot D_H^2(\mathbb{P}, \mathbb{Q})} \\
&\leq \eta(\mathbb{E}_{\mathbb{P}}[h(X)] + \mathbb{E}_{\mathbb{Q}}[h(X)]) + \frac{R}{2\eta} D_H^2(\mathbb{P}, \mathbb{Q}).
\end{aligned}$$

The above implies

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}[h(X)] &\leq \frac{1+\eta}{1-\eta}\mathbb{E}_{\mathbb{Q}}[h(X)] + \frac{R}{2\eta(1-\eta)} D_H^2(\mathbb{P}, \mathbb{Q}) \\
&= \left(1 + \frac{1}{\beta}\right)\mathbb{E}_{\mathbb{Q}}[h(X)] + 3R\frac{(2\beta+1)^2}{2\beta} D_H^2(\mathbb{P}, \mathbb{Q}) && \text{(Plug } \eta = \tfrac{1}{2\beta+1} \text{ for all } \beta \in (0, \infty).) \\
&\leq \left(1 + \frac{1}{\beta}\right)\mathbb{E}_{\mathbb{Q}}[h(X)] + 3R\beta D_H^2(\mathbb{P}, \mathbb{Q}). && \text{(For any } \beta \geq 1)
\end{aligned}$$

□

Lemma B.4 (restatement of Lemma 4.1)**.** *Let $r : S \times A \to [0, 1]$ be a bounded expected rewards function. Let $P_\star$ and $\widehat{P}$ denote two dynamics and consider the MDPs $M = (S, A, P_\star, r, s_0, H)$ and $\widehat{M} = (S, A, \widehat{P}, r, s_0, H)$. Then, for any policy π we have*

$$V_{\widehat{M}}^{\pi}(s) \leq 3V_M^{\pi}(s) + 9H^2 \mathop{\mathbb{E}}_{P_\star, \pi}\left[\sum_{h=0}^{H-1} D_H^2(\widehat{P}(\cdot|s_h, a_h), P_\star(\cdot|s_h, a_h)) \middle| s_0 = s\right].$$

Proof. We first prove by backwards induction that for all $h \in [H-1]$ the following holds.

$$V_{\widehat{M},h}^{\pi}(s) \leq \left(1 + \frac{1}{H}\right)^{H-h}\left[V_{M,h}^{\pi}(s) + \mathop{\mathbb{E}}_{P_\star, \pi}\left[\sum_{h'=h}^{H-1} 3H^2 D_H^2(\widehat{P}(\cdot|s_{h'}, a_{h'}), P_\star(\cdot|s_{h'}, a_{h'})) \middle| s_h = s\right]\right].$$

The base case, $h = H - 1$ is immediate since $V_{\widehat{M},h}^{\pi}(s) = V_{M,h}^{\pi}(s)$. Now, we assume that the above holds for $h + 1$ and prove that it holds for h. To see this, we have that

$$\begin{aligned}
V_{\widehat{M},h}^{\pi}(s) &= \mathop{\mathbb{E}}_{a \sim \pi(\cdot|s)}\left[r(s, a) + \mathbb{E}_{s' \sim \widehat{P}(\cdot|s,a)}\left[V_{\widehat{M},h+1}^{\pi}(s')\right]\right] && \text{(By Bellman's equations)} \\
&\leq \mathop{\mathbb{E}}_{a \sim \pi(\cdot|s)}\left[r(s, a) + \left(1 + \frac{1}{H}\right)\mathbb{E}_{s' \sim P_\star(\cdot|s,a)}\left[V_{\widehat{M},h+1}^{\pi}(s')\right] + 3H^2 D_H^2(\widehat{P}(\cdot|s, a), P_\star(\cdot|s, a))\right] && \text{(Corollary B.3)} \\
&\leq \mathop{\mathbb{E}}_{a \sim \pi(\cdot|s)}\left[r(s, a) + 3H^2 D_H^2(\widehat{P}(\cdot|s, a), P_\star(\cdot|s, a))\right] && \text{(Induction hypothesis)}
\end{aligned}$$

$$
\begin{aligned}
&+ \mathop{\mathbb{E}}_{a\sim\pi(\cdot|s)}\left[\left(1+\frac{1}{H}\right)^{H-h} \mathop{\mathbb{E}}_{s'\sim P_\star(\cdot|s,a)}\left[V^\pi_{M,h+1}(s')\right]\right] \\
&+ \mathop{\mathbb{E}}_{a\sim\pi(\cdot|s)}\left[\left(1+\frac{1}{H}\right)^{H-h} \mathop{\mathbb{E}}_{s'\sim P_\star(\cdot|s,a)}\left[\mathbb{E}\left[\sum_{h'=h+1}^{H-1} 3H^2 D_H^2(\widehat{P}(\cdot|s_{h'},a_{h'}), P_\star(\cdot|s_{h'},a_{h'}))\middle| s_{h+1}=s'\right]\right]\right] \\
&\leq \left(1+\frac{1}{H}\right)^{H-h} \mathop{\mathbb{E}}_{a\sim\pi(\cdot|s)}\left[r(s,a)+\mathop{\mathbb{E}}_{s'\sim P_\star(\cdot|s,a)}\left[V^\pi_{M,h+1}(s')\right]\right] && (r, D_H^2 \geq 0) \\
&+ \left(1+\frac{1}{H}\right)^{H-h} \mathop{\mathbb{E}}_{P_\star,\pi}\left[\sum_{h'=h}^{H-1} 3H^2 D_H^2(\widehat{P}(\cdot|s_{h'},a_{h'}), P_\star(\cdot|s_{h'},a_{h'}))\middle| s_h=s\right] \\
&= \left(1+\frac{1}{H}\right)^{H-h}\left[V^\pi_{M,h}(s)+\mathop{\mathbb{E}}_{P_\star,\pi}\left[\sum_{h'=h}^{H-1} 3H^2 D_H^2(\widehat{P}(\cdot|s_{h'},a_{h'}), P_\star(\cdot|s_{h'},a_{h'}))\middle| s_h=s\right]\right], && \text{(By Bellman's equations)}
\end{aligned}
$$

as desired. Plugging in $h=0$ and using that $\left(1+\frac{1}{H}\right)^H \leq 3$ concludes the proof. □

B.2. Oracle Bounds (Step 1)

Reward oracle.

Lemma B.5 (Lemma B.10 in Levy & Mansour, 2023)**.** *For any $\delta \in (0,1)$, with probability at least $1-\delta$ we have*

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

simultaneously, for all $t \geq 2$ and any fixed sequence of functions $f_1, f_2, \ldots \in \mathcal{F}$.

Corollary B.6 (restatement of Corollary 4.2)**.** *Let $\widehat{f}_t \in \mathcal{F}$ be the least squares minimizer in Algorithm 1. For any $\delta \in (0,1)$ it holds that with probability at least $1-\delta$ we have*

$$
\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathop{\mathbb{E}}_{\pi_i(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1}\left(\widehat{f}_t(c,s_h,a_h)-f_\star(c,s_h,a_h)\right)^2\middle| s_0\right]\right] \leq 68H\log(2T^3|\mathcal{F}|/\delta),
$$

simultaneously, for all $t \geq 1$.

Proof. Recall that for all $t \geq 2$, $\widehat{f}_t$ is the least square minimizer at round t. Hence, by our assumption that $f_\star \in \mathcal{F}$

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

Thus the corollary immediately follows by Lemma B.5. □

Dynamics oracle. Recall the Hellinger distance given in Definition 2.3. The following lemma by (Foster et al., 2021) upper bounds the expected cumulative Hellinger Distance in terms of the log-loss. Let $\mathcal{X}$ be a set and $\mathcal{Y}$ be a finite set. Let $x^{(t)}, y^{(t)}$, $t \geq 1$, be a sequence of random variables that satisfy $y^{(t)} \sim g_\star(\cdot|x^{(t)})$, where $g_\star : \mathcal{Y}\times\mathcal{X} \to \mathbb{R}_+$ maps $x \in \mathcal{X}$ to the density of $y \in \mathcal{Y}$. Define $\mathcal{H}^{(t)} = (x^{(1)}, y^{(1)}, \ldots, x^{(t)}, y^{(t)})$ and let $\mathcal{G}^{(t)} = \sigma(\mathcal{H}^{(t)})$. Next, for a random variable Z, we define $\mathbb{E}_t Z := \mathbb{E}[Z|\mathcal{G}_t]$.

Lemma B.7 (Lemma A.14 from Foster et al., 2021)**.** *Let $g : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}_+$ be a mapping from $\mathcal{X}$ to densities over $\mathcal{Y}$. Consider a sequence of $\{0,1\}$-valued random variables $(\mathbb{I}_t)_{t \le T}$ where $\mathbb{I}_t$ is $F^{(t-1)}$-measurable. For any $\delta \in (0,1)$ we have that with probability at least $1-\delta$,*

$$\sum_{t=1}^{T} \mathbb{E}_{t-1}\left[D_H^2(g(\cdot|x^{(t)}), g_\star(\cdot|x^{(t)}))\right]\mathbb{I}_t \le \sum_{t=1}^{T}\left(\log\frac{1}{g(y^{(t)}|x^{(t)})} - \log\frac{1}{g_\star(y^{(t)})|x^{(t)}}\right)\mathbb{I}_t + 2\log(1/\delta).$$

Additionally, with probability at least $1-\delta$,

$$\sum_{t=1}^{T} D_H^2(g(\cdot|x^{(t)}), g_\star(\cdot|x^{(t)})) \le \sum_{t=1}^{T}\left(\log\frac{1}{g(y^{(t)}|x^{(t)})} - \log\frac{1}{g_\star(y^{(t)})|x^{(t)}}\right) + 2\log(1/\delta).$$

Using the above lemma, we bound the realized and expected cumulative Hellinger distance between the approximated and true dynamics, by the actual regret of the log-loss regression oracle (and constant terms), with high probability.

Lemma B.8 (Concentration of log-loss oracle)**.** *For any $\delta \in (0,1)$ it holds that with probability at least $1-\delta$ we have*

$$\begin{aligned}
&\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathop{\mathbb{E}}_{\pi_i(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1} D_H^2(P_\star^c(\cdot|s_h,a_h), P^c(\cdot|s_h,a_h))\middle| s_0\right]\right] \\
&\le \sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\log\left(\frac{1}{P^{c_i}(s_{h+1}^i|s_h^i,a_h^i)}\right) - \sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\log\left(\frac{1}{P_\star^{c_i}(s_{h+1}^i|s_h^i,a_h^i)}\right) + 2H\log(TH|\mathcal{P}|/\delta).
\end{aligned}$$

simultaneously, for all $t \ge 1$ and $P \in \mathcal{P}$.

Proof. Fix some $t \ge 1$ and $P \in \mathcal{P}$. We have with probability at least $1 - \frac{\delta}{T|\mathcal{P}|}$ that

$$\begin{aligned}
&\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathop{\mathbb{E}}_{\pi_i(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1} D_H^2(P_\star^c(\cdot|s_h,a_h), P^c(\cdot|s_h,a_h))\middle| s_0\right]\right] \\
&= \sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\mathbb{E}_c\left[\mathop{\mathbb{E}}_{\pi_i(c;\cdot),P_\star^c}\left[D_H^2(P_\star^c(\cdot|s_h,a_h), P^c(\cdot|s_h,a_h))\middle| s_0\right]\right] \\
&\underbrace{=}_{(i)} \sum_{h=0}^{H-1}\sum_{i=1}^{t-1}\mathbb{E}\left[D_H^2(P_\star^{c_i}(\cdot|s_h^i,a_h^i), P^{c_i}(\cdot|s_h^i,a_h^i))\middle| s_0,\pi_i\right] \\
&\le \sum_{h=0}^{H-1}\left(\sum_{i=1}^{t-1}\log\left(\frac{P_\star^{c_i}(s_{h+1}^i|s_h^i,a_h^i)}{P^{c_i}(s_{h+1}^i|s_h^i,a_h^i)}\right) + 2\log(HT|\mathcal{P}|/\delta)\right) && \text{(Foster et al., 2021, Lemma A.14)} \\
&= \sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\log\left(\frac{1}{P^{c_i}(s_{h+1}^i|s_h^i,a_h^i)}\right) - \sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\log\left(\frac{1}{P_\star^{c_i}(s_{h+1}^i|s_h^i,a_h^i)}\right) + 2H\log(HT|\mathcal{P}|/\delta)
\end{aligned}$$

The filtration used in (i) is over the history up to time t, $\mathbb{H}_{t-1} = (\sigma^1, \dots, \sigma^{t-1})$. Now, by taking a union bound over every $t = 1 \dots, T$ and $P \in \mathcal{P}$, we obtain the lemma. □

Lemma B.9 (Realized log-loss error)**.** *For any $\delta \in (0,1)$ it holds that with probability at least $1-\delta$ we have*

$$\begin{aligned}
&\sum_{i=1}^{t-1}\sum_{h=0}^{H-1} D_H^2(P_\star^{c_i}(\cdot|s_h^i,a_h^i), P^{c_i}(\cdot|s_h^i,a_h^i)) \\
&\le \sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\log\left(\frac{1}{P^{c_i}(s_{h+1}^i|s_h^i,a_h^i)}\right) - \sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\log\left(\frac{1}{P_\star^{c_i}(s_{h+1}^i|s_h^i,a_h^i)}\right) + 2H\log(TH|\mathcal{P}|/\delta).
\end{aligned}$$

simultaneously, for all $t \ge 1$ and $P \in \mathcal{P}$.

Proof. Fix some $t \geq 1$ and $P \in \mathcal{P}$. We have with probability at least $1 - \frac{\delta}{T|\mathcal{P}|}$ that

$$\begin{aligned}
&\sum_{i=1}^{t-1}\sum_{h=0}^{H-1} D_H^2(P_\star^{c_i}(\cdot|s_h^i,a_h^i), P^{c_i}(\cdot|s_h^i,a_h^i)) \\
&= \sum_{h=0}^{H-1}\sum_{i=1}^{t-1} D_H^2(P_\star^{c_i}(\cdot|s_h^i,a_h^i), P^{c_i}(\cdot|s_h^i,a_h^i)) \\
&\leq \sum_{h=0}^{H-1}\left(\sum_{i=1}^{t-1}\log\left(\frac{P_\star^{c_i}(s_{h+1}^i|s_h^i,a_h^i)}{P^{c_i}(s_{h+1}^i|s_h^i,a_h^i)}\right) + 2\log(HT|\mathcal{P}|/\delta)\right) && \text{(Foster et al., 2021, Lemma A.14)} \\
&= \sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\log\left(\frac{1}{P^{c_i}(s_{h+1}^i|s_h^i,a_h^i)}\right) - \sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\log\left(\frac{1}{P_\star^{c_i}(s_{h+1}^i|s_h^i,a_h^i)}\right) + 2H\log(HT|\mathcal{P}|/\delta).
\end{aligned}$$

The filtration used in (i) is over the history up to time t, $\mathbb{H}_{t-1} = (\sigma^1, \ldots, \sigma^{t-1})$. Now, by taking a union bound over every $t = 1.\ldots, T$ and $P \in \mathcal{P}$, we obtain the lemma. □

Corollary B.10 (restatement of Corollary 4.3)**.** *Let $\widehat{P}_t \in \mathcal{P}$ be the log loss minimizer in Algorithm 1. For any $\delta \in (0,1)$ it holds that with probability at least $1-\delta$ we have*

$$\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathop{\mathbb{E}}_{\pi_i(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1} D_H^2(P_\star^c(\cdot|s_h,a_h), \widehat{P}_t^c(\cdot|s_h,a_h))\middle| s_0\right]\right] \leq 2H\log(TH|\mathcal{P}|/\delta), \qquad \forall 1 \leq t \leq T.$$

Proof. By our assumption that $P_\star \in \mathcal{P}$, and $\widehat{P}_t$ is the log loss minimizer at time t, it holds that

$$\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\log\left(\frac{1}{\widehat{P}_t^{c_i}(s_{h+1}^i|s_h^i,a_h^i)}\right) - \sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\log\left(\frac{1}{P_\star^{c_i}(s_{h+1}^i|s_h^i,a_h^i)}\right) \leq 0.$$

Thus, the corollary immediately follows by Lemma B.8. □

Lemma B.11 (Stability error of log-loss oracle, restatement of Lemma 4.4)**.** *Let $\widehat{P}_i \in \mathcal{P}$ denote the log-loss minimizer at round $i \in [T]$. For any $\delta \in (0,1)$ it holds with probability at least $1-\delta$ that*

$$\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathbb{E}_{\pi_i(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1} D_H^2(P_\star^c(\cdot|s_h,a_h), \widehat{P}_i^c(s_h,a_h))\middle| s_0\right]\right] \leq 112Hd_{\mathcal{P}}\log^2(2TH|\mathcal{P}|/\delta)$$

simultaneously, for all $t \geq 1$, where $d_{\mathcal{P}} \geq d_{\mathrm{E}}(\mathcal{P}, D_H, T^{-1/2})$, the Eluder dimension of $\mathcal{P}$ at scale $T^{-1/2}$.

Proof. Let $\beta = 2H\log(2TH|\mathcal{P}|/\delta)$ and define

$$\mathcal{P}_t = \left\{P \in \mathcal{P} \;:\; \sum_{i=1}^{t-1}\sum_{h=0}^{H-1} D_H^2(P^{c_i}(\cdot|s_h^i,a_h^i), \widehat{P}_t^{c_i}(\cdot|s_h^i,a_h^i)) \leq \beta\right\}.$$

Now, suppose that Lemma B.9 holds with $\delta/2$. Then, since $\widehat{P}_t$ is the log-loss minimizer, we have that $P_\star \in \mathcal{P}_t$ for all $1 \leq t \leq T$. Next, recalling that

$$w_{\mathcal{P}_t}(s,a,c) = \sup_{P,P' \in \mathcal{P}_t} D_H(P^c(\cdot|s,a), P'^c(\cdot|s,a)),$$

thus, we have that

$$\sum_{i=1}^{t-1}\sum_{h=0}^{H-1} D_H^2(P_\star^{c_i}(\cdot|s_h^i,a_h^i), \widehat{P}_i^{c_i}(\cdot|s_h^i,a_h^i) \leq \sum_{i=1}^{t-1}\sum_{h=0}^{H-1}(w_{\mathcal{P}_i}(s_h^i,a_h^i,c_i))^2.$$

Applying Lemma 2.2 we get that

$$\sum_{i=1}^{t-1}\sum_{h=0}^{H-1} D_H^2(P_\star^{c_i}(\cdot|s_h^i,a_h^i),\widehat{P}_i^{c_i}(\cdot|s_h^i,a_h^i) \le 24 d_{\mathcal{P}} H\beta \log T$$
$$\le 48 H d_{\mathcal{P}} \log^2(2TH|\mathcal{P}|/\delta).$$

Finally, we apply Lemma B.1 to get that with probability at least $1-\delta/2$

$$\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathbb{E}_{\pi_i(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1} D_H^2(P_\star^c(\cdot|s_h,a_h),\widehat{P}_i^c(s_h,a_h))\Big|s_0\right]\right]$$
$$=\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\mathbb{E}\Big[D_H^2(P_\star^{c_i}(\cdot|s_h^i,a_h^i),\widehat{P}_i^{c_i}(\cdot|s_h^i,a_h^i)\Big|\pi_i,s_0\Big]$$
$$\le 2\sum_{i=1}^{t-1}\sum_{h=0}^{H-1} D_H^2(P_\star^{c_i}(\cdot|s_h^i,a_h^i),\widehat{P}_i^{c_i}(\cdot|s_h^i,a_h^i)+16H\log\frac{2TH}{\delta}.$$

Taking a union bound and combining the last two inequalities concludes the proof. □

B.3. Confidence Bounds (Step 2)

In the following analysis, we use an occupancy measures-based representation of the value function. Recall the definition of the *occupancy measures* (Zimin & Neu, 2013). For any non-contextual policy π and dynamics P, let $q_h(s,a|\pi,P)$ denote the probability of reaching state $s\in S$ and performing action $a\in A$ at time $h\in[H]$ of an episode generated using policy π and dynamics P.

Using this notation, the value function of any policy π with respect to the MDP (S,A,P,r,s_0,H) can be represented as follows.

$$V_M^\pi(s_0)=\sum_{h=0}^{H-1}\sum_{s\in S_h}\sum_{a\in A} q_h(s,a|\pi,P)\cdot r(s,a). \tag{2}$$

Thus, the following is an immediate corollary of Lemma 4.1.

Corollary B.12. *For any (non-contextual) policy π, two dynamics P and $\widehat{P}$, and rewards function r that is bounded in $[0,1]$ it holds that*

$$\sum_{h=0}^{H-1}\sum_{s\in S_h}\sum_{a\in A} q_h(s,a|\pi,\widehat{P})\cdot r(s,a)\le 3\sum_{h=0}^{H-1}\sum_{s\in S_h}\sum_{a\in A} q_h(s,a|\pi,P)\cdot r(s,a)$$
$$+9H^2\sum_{h=0}^{H-1}\sum_{s\in S_h}\sum_{a\in A} q_h(s,a|\pi,P)\cdot D_H^2(P(\cdot|s,a),\widehat{P}(\cdot|s,a)).$$

We are now ready to prove the confidence bounds. Recall the reward bonuses $b_{t,h}^R, b_{t,h}^P$ defined in Equation (1).

Lemma B.13 (restatement of Lemma 4.5)**.** *Let $P_\star$ and $f_\star$ be the true context dependent dynamics and rewards. Let $\widehat{P}_t$ and $\widehat{f}_t$ be the approximated context-dependent dynamics and rewards at round t. Then, for any $t\ge 1$, and context-dependent policy $\pi\in\Pi_{\mathcal{C}}$ the following holds.*

$$\left|\mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}(f_\star,\widehat{P}_t)(c)}(s_0)\right]-\mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{f}_t,\widehat{P}_t)(c)}(s_0)\right]\right|$$
$$\le \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c)b_{t,h}^R(c,s_h,a_h)\right]$$
$$+\frac{3}{2\beta_r}\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathbb{E}_{\pi_i(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1}\Big(f_\star(c,s_h,a_h)-\widehat{f}_t(c,s_h,a_h)\Big)^2\Bigg|s_0\right]\right]$$

$$+ \frac{9H^2}{2\beta_r}\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathop{\mathbb{E}}_{\pi_i(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1} D_H^2(P_\star^c(\cdot|s_h,a_h),\widehat{P}_i^c(\cdot|s_h,a_h))\middle|s_0\right]\right]$$
$$+ \frac{H}{2\beta_r}.$$

Proof. We have that

$$\begin{aligned}
&\left|\mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}(f_\star,\widehat{P}_t)(c)}(s_0)\right] - \mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{f}_t,\widehat{P}_t)(c)}(s_0)\right]\right| \\
&= \left|\mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}(f_\star,\widehat{P}_t)(c)}(s_0) - V^{\pi(c;\cdot)}_{\mathcal{M}(\widehat{f}_t,\widehat{P}_t)(c)}(s_0)\right]\right| && \text{(By linearity of expectation)} \\
&= \left|\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c)\cdot\left(\widehat{f}_t(c,s_h,a_h) - f_\star(c,s_h,a_h)\right)\right]\right| && \text{(Equation (2))} \\
&\le \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c)\cdot\left|\widehat{f}_t(c,s_h,a_h) - f_\star(c,s_h,a_h)\right|\right] && \text{(Triangle ineq.)} \\
&= \mathbb{E}_c\left[\sum_{h=0}^{H-1}\min\left\{1,\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c)\cdot\left|\widehat{f}_t(c,s_h,a_h) - f_\star(c,s_h,a_h)\right|\right\}\right] \\
&\le \mathbb{E}_c\left[\sum_{h=0}^{H-1}\min\left\{1,\sum_{s_h\in S_h^c}\sum_{a_h\in A}\frac{\beta_r}{2}\frac{q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c)}{1+\sum_{i=1}^{t-1}q_h(s_h,a_h|\pi_i(c;\cdot),\widehat{P}_i^c)}\right.\right. && \text{(AM-GM)} \\
&\left.\left.+\frac{1}{2\beta_r}q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c)\left(1+\sum_{i=1}^{t-1}q_h(s_h,a_h|\pi_i(c;\cdot),\widehat{P}_i^c)\right)\left(f_\star(c,s_h,a_h)-\widehat{f}_t(c,s_h,a_h)\right)^2\right\}\right] \\
&\le \mathbb{E}_c\left[\sum_{h=0}^{H-1}\min\left\{1,\sum_{s_h\in S_h^c}\sum_{a_h\in A}\frac{\beta_r}{2}\frac{q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c)}{1+\sum_{i=1}^{t-1}q_h(s_h,a_h|\pi_i(c;\cdot),\widehat{P}_i^c)}\right\}\right] \\
&+\frac{H}{2\beta_r}+\frac{1}{2\beta_r}\mathbb{E}_c\left[\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi_i(c;\cdot),\widehat{P}_i^c)\left(f_\star(c,s_h,a_h)-\widehat{f}_t(c,s_h,a_h)\right)^2\right] \\
&\le \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_i^c)\min\left\{1,\frac{\beta_r/2}{1+\sum_{i=1}^{t-1}q_h(s_h,a_h|\pi_i(c;\cdot),\widehat{P}_i^c)}\right\}\right] \\
&+\frac{3}{2\beta_r}\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathop{\mathbb{E}}_{\pi_i(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1}\left(f_\star(c,s_h,a_h)-\widehat{f}_t(c,s_h,a_h)\right)^2\middle|s_0\right]\right] && \text{(Corollary B.12)} \\
&+\frac{9H^2}{2\beta_r}\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathop{\mathbb{E}}_{\pi_i(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1}D_H^2(P_\star^c(\cdot|s_h,a_h),\widehat{P}_i^c(\cdot|s_h,a_h))\middle|s_0\right]\right]+\frac{H}{2\beta_r},
\end{aligned}$$

and the lemma follows by $b^R_{t,h}$ definition. □

Lemma B.14 (restatement of Lemma 4.6)**.** *Let $P_\star$ and $f_\star$ be the true context dependent dynamics and rewards. Let $\widehat{P}_t$ be the approximated context-dependent dynamics at round t. Then, for any $t \ge 1$, and context-dependent policy $\pi \in \Pi_{\mathcal{C}}$ we have*

$$\begin{aligned}
\left|\mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}(f_\star,P_\star)(c)}(s_0)\right] - \mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}(f_\star,\widehat{P}_t)(c)}(s_0)\right]\right| &\le \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s\in S_h^c}\sum_{a\in A} q_h(s,a|\pi(c;\cdot),\widehat{P}_t^c)\cdot b^P_{t,h}(c,s,a)\right] \\
&+\frac{6H}{\beta_P}\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathop{\mathbb{E}}_{\pi_i(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1}D_H^2(P_\star^c(\cdot|s_h,a_h),\widehat{P}_t^c(\cdot|s_h,a_h))\middle|s_0\right]\right]
\end{aligned}$$

$$+ \frac{18H^3}{\beta_P}\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathop{\mathbb{E}}_{\pi_i(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1} D_H^2(P_\star^c(\cdot|s_h,a_h),\widehat{P}_i^c(\cdot|s_h,a_h))\middle|s_0\right]\right]$$
$$+ \frac{H^2}{2\beta_P}.$$

Proof. The following holds for any $t \geq 1$ and a context-dependent policy $\pi \in \Pi_{\mathcal{C}}$.

$$\left|\mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(f_\star,P_\star)(c)}(s_0)] - \mathbb{E}_c[V^{\pi(c;\cdot)}_{\mathcal{M}(f_\star,\widehat{P}_t)(c)}(s_0)]\right|$$
$$= \left|\mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}(f_\star,P_\star)(c)}(s_0) - V^{\pi(c;\cdot)}_{\mathcal{M}(f_\star,\widehat{P}_t)(c)}(s_0)\right]\right| \quad \text{(By linearity of expectation)}$$
$$= \left|\mathbb{E}_c\left[\mathbb{E}_{\pi(c;\cdot),\widehat{P}_t^c}\left[\sum_{h=0}^{H-1}\sum_{s'\in S}(P_\star^c(s'|s_h,a_h) - \widehat{P}_t^c(s'|s_h,a_h))V^{\pi(c;\cdot)}_{\mathcal{M}(f_\star,P_\star),h+1}(s')\right]\middle|s_0\right]\right| \quad \text{(Lemma C.2)}$$
$$= \left|\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c)\sum_{s'\in S}(P_\star^c(s'|s_h,a_h) - \widehat{P}_t^c(s'|s_h,a_h))V^{\pi(c;\cdot)}_{\mathcal{M}(f_\star,P_\star),h+1}(s')\right]\right| \quad \text{(Equation (2))}$$
$$\leq \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c)\left|\sum_{s'\in S}(P_\star^c(s'|s_h,a_h) - \widehat{P}_t^c(s'|s_h,a_h))V^{\pi(c;\cdot)}_{\mathcal{M}(f_\star,P_\star),h+1}(s')\right|\right]$$
(Triangle inequality)
$$\leq H\mathbb{E}_c\left[\sum_{h=0}^{H-1}\min\left\{1, \sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c)\sum_{s'\in S}\left|P_\star^c(s'|s_h,a_h) - \widehat{P}_t^c(s'|s_h,a_h)\right|\right\}\right]$$
($f_\star \in [0,1]$, $V^{\pi(c;\cdot)}_{\mathcal{M}(f_\star,P_\star),h}(s') \in [0,H]$ for all $h \in [H]$ and $s' \in S$)
$$\leq H\mathbb{E}_c\left[\sum_{h=0}^{H-1}\min\left\{1, \sum_{s_h\in S_h^c}\sum_{a_h\in A}\left(\frac{\beta_P}{2}\frac{q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c)}{1+\sum_{i=1}^{t-1}q_h(s_h,a_h|\pi_i(c;\cdot),\widehat{P}_i^c)}\right.\right.\right. \quad \text{(AM-GM)}$$
$$\left.\left.\left. + \frac{q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c)}{2\beta_P}(1+\sum_{i=1}^{t-1}q_h(s_h,a_h|\pi_i(c;\cdot),\widehat{P}_i^c))\left(\sum_{s'\in S}\left|P_\star^c(s'|s_h,a_h) - \widehat{P}_t^c(s'|s_h,a_h)\right|\right)^2\right)\right\}\right]$$
$$\leq \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c)H\min\left\{1, \frac{\beta_P/2}{1+\sum_{i=1}^{t-1}q_h(s_h,a_h|\pi_i(c;\cdot),\widehat{P}_i^c)}\right\}\right] + \frac{H^2}{2\beta_P}$$
$$+ \frac{H}{2\beta_P}\mathbb{E}_c\left[\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi_i(c;\cdot),\widehat{P}_i^c)\left(\sum_{s'\in S}\left|P_\star^c(s'|s_h,a_h) - \widehat{P}_t^c(s'|s_h,a_h)\right|\right)^2\right]$$
$$\leq \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c)H\min\left\{1, \frac{\beta_P/2}{1+\sum_{i=1}^{t-1}q_h(s_h,a_h|\pi_i(c;\cdot),\widehat{P}_i^c)}\right\}\right] + \frac{H^2}{2\beta_P}$$
$$+ \frac{3H}{2\beta_P}\mathbb{E}_c\left[\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi_i(c;\cdot),P_\star^c)\left(\sum_{s'\in S}\left|P_\star^c(s'|s_h,a_h) - \widehat{P}_t^c(s'|s_h,a_h)\right|\right)^2\right]$$
$$+ \frac{18H^3}{\beta_P}\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathop{\mathbb{E}}_{\pi_i(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1} D_H^2(P_\star^c(\cdot|s_h,a_h),\widehat{P}_i^c(\cdot|s_h,a_h))\middle|s_0\right]\right] \quad \text{(Corollary B.12)}$$
$$\leq \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c)H\min\left\{1, \frac{\beta_P/2}{1+\sum_{i=1}^{t-1}q_h(s_h,a_h|\pi_i(c;\cdot),\widehat{P}_i^c)}\right\}\right] + \frac{H^2}{2\beta_P}$$

$$+ \frac{6H}{\beta_P}\mathbb{E}_c\left[\sum_{i=1}^{t-1}\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi_i(c;\cdot),P_\star^c)D_H^2(P_\star^c(\cdot|s_h,a_h),\widehat{P}_t^c(\cdot|s_h,a_h))\right] \qquad (\mathrm{TV}^2 \le 4D_H^2)$$

$$+ \frac{18H^3}{\beta_P}\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathop{\mathbb{E}}_{\pi_i(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1} D_H^2(P_\star^c(\cdot|s_h,a_h),\widehat{P}_i^c(\cdot|s_h,a_h))\middle| s_0\right]\right]$$

$$= \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c)b_{t,h}^P(c,s_h,a_h)\right] + \frac{H^2}{2\beta_P}$$

$$+ \frac{6H}{\beta_P}\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathop{\mathbb{E}}_{\pi_i(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1} D_H^2(P_\star^c(\cdot|s_h,a_h),\widehat{P}_t^c(\cdot|s_h,a_h))\middle| s_0\right]\right]$$

$$+ \frac{18H^3}{\beta_P}\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathop{\mathbb{E}}_{\pi_i(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1} D_H^2(P_\star^c(\cdot|s_h,a_h),\widehat{P}_i^c(\cdot|s_h,a_h))\middle| s_0\right]\right].$$

□

Corollary B.15 (restatement of Corollary 4.7)**.** *Under the terms of Lemmas 4.5 and 4.6, the following holds with probability at least* $1-3\delta/4$ *simultaneously for all* $t\ge 1$ *and* $\pi\in\Pi_{\mathcal{C}}$*:*

$$\begin{aligned}
&\left|\mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}^{(f_\star,P_\star)}(c)}(s_0)\right] - \mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{f}_t,\widehat{P}_t)}(c)}(s_0)\right]\right| \\
&\le \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c)\big(b_{t,h}^R(c,s_h,a_h)+b_{t,h}^P(c,s_h,a_h)\big)\right] \\
&+ \frac{2029H^4 d_{\mathcal{P}}}{\beta_P}\log^2(8TH|\mathcal{P}|/\delta) + \frac{504H^3 d_{\mathcal{P}}}{\beta_r}\log^2(64T^4H|\mathcal{F}||\mathcal{P}|/\delta^2).
\end{aligned}$$

Proof. We begin by taking a union bound on the events of Corollaries 4.2 and 4.3 and Lemma 4.4 to get that with probability at least $1-3\delta/4$, simultaneously for all $t\ge 1$

$$\begin{aligned}
&\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathop{\mathbb{E}}_{\pi_i(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1}\Big(\widehat{f}_t(c,s_h,a_h)-f_\star(c,s_h,a_h)\Big)^2\middle| s_0\right]\right] \le 68H\log(8T^3|\mathcal{F}|/\delta) \\
&\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathop{\mathbb{E}}_{\pi_i(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1} D_H^2(P_\star^c(\cdot|s_h,a_h),\widehat{P}_t^c(\cdot|s_h,a_h))\middle| s_0\right]\right] \le 2H\log(4TH|\mathcal{P}|/\delta) \qquad (3) \\
&\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathop{\mathbb{E}}_{\pi_i(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1} D_H^2(P_\star^c(\cdot|s_h,a_h),\widehat{P}_i^c(\cdot|s_h,a_h))\middle| s_0\right]\right] \le 112Hd_{\mathcal{P}}\log^2(8TH|\mathcal{P}|/\delta).
\end{aligned}$$

Assuming this event holds, we get that for all $t\ge 1$ and context-dependent policy $\pi\in\Pi_{\mathcal{C}}$.

$$\begin{aligned}
&\left|\mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}^{(f_\star,P_\star)}(c)}(s_0)\right] - \mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{f}_t,\widehat{P}_t)}(c)}(s_0)\right]\right| \\
&\le \left|\mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}^{(f_\star,P_\star)}(c)}(s_0)\right] - \mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}^{(f_\star,\widehat{P}_t)}(c)}(s_0)\right]\right| \qquad \text{(By triangle inequality)} \\
&\quad + \left|\mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}^{(f_\star,\widehat{P}_t)}(c)}(s_0)\right] - \mathbb{E}_c\left[V^{\pi(c;\cdot)}_{\mathcal{M}^{(\widehat{f}_t,\widehat{P}_t)}(c)}(s_0)\right]\right| \\
&\le \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c)\big(b_{t,h}^R(c,s_h,a_h)+b_{t,h}^P(c,s_h,a_h)\big)\right] \qquad \text{(Lemmas 4.5 and 4.6)} \\
&\quad + \frac{3}{2\beta_r}\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathop{\mathbb{E}}_{\pi_i(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1}\Big(f_\star(c,s_h,a_h)-\widehat{f}_t(c,s_h,a_h)\Big)^2\middle| s_0\right]\right]
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{6H}{\beta_P}\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathop{\mathbb{E}}_{\pi_i(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1} D_H^2(P_\star^c(\cdot|s_h,a_h),\widehat{P}_t^c(\cdot|s_h,a_h))\middle| s_0\right]\right] \\
&+ \left(\frac{9H^2}{2\beta_r}+\frac{18H^3}{\beta_P}\right)\mathbb{E}_c\left[\sum_{i=1}^{t-1}\mathop{\mathbb{E}}_{\pi_i(c;\cdot),P_\star^c}\left[\sum_{h=0}^{H-1} D_H^2(P_\star^c(\cdot|s_h,a_h),\widehat{P}_i^c(\cdot|s_h,a_h))\middle| s_0\right]\right] \\
&+ \frac{H}{2\beta_r}+\frac{H^2}{2\beta_P} \\
\leq\ & \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c)\big(b_{t,h}^R(c,s_h,a_h)+b_{t,h}^P(c,s_h,a_h)\big)\right] && \text{(Equation (3))} \\
&+ \frac{3}{2\beta_r}68H\log(8T^3|\mathcal{F}|/\delta)+\frac{6H}{\beta_P}2H\log(4TH|\mathcal{P}|/\delta) \\
&+ \left(\frac{9H^2}{2\beta_r}+\frac{18H^3}{\beta_P}\right)112Hd_{\mathcal{P}}\log^2(8TH|\mathcal{P}|/\delta) \\
&+ \frac{H}{2\beta_r}+\frac{H^2}{2\beta_P} \\
\leq\ & \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c)\big(b_{t,h}^R(c,s_h,a_h)+b_{t,h}^P(c,s_h,a_h)\big)\right] \\
&+ \frac{2029H^4d_{\mathcal{P}}}{\beta_P}\log^2(8TH|\mathcal{P}|/\delta)+\frac{504H^3d_{\mathcal{P}}}{\beta_r}\log^2(64T^4H|\mathcal{F}||\mathcal{P}|/\delta^2),
\end{aligned}
$$

as stated. □

B.4. Establishing Optimism Lemmas (Step 3)

Lemma B.16 (restatement of Lemma 4.8)**.** *Let $\pi_\star$ be an optimal context-dependent policy for $\mathcal{M}$. Under the good event of Corollary 4.7, we have that for any $t\geq 1$*

$$
\begin{aligned}
\mathbb{E}_c\left[V_{\mathcal{M}(c)}^{\pi_\star(c;\cdot)}\right] \leq & \mathbb{E}_c\left[V_{\widehat{\mathcal{M}}_t(c)}^{\pi_t(c;\cdot)}(s_0)\right]+\frac{2029H^4d_{\mathcal{P}}}{\beta_P}\log^2(8TH|\mathcal{P}|/\delta) \\
&+\frac{504H^3d_{\mathcal{P}}}{\beta_r}\log^2(64T^4H|\mathcal{F}||\mathcal{P}|/\delta^2).
\end{aligned}
$$

Proof. Fix any round $t\geq 1$ consider the following derivation.

$$
\begin{aligned}
\mathbb{E}_c\left[V_{\mathcal{M}(c)}^{\pi_\star(c;\cdot)}(s_0)\right] &= \mathbb{E}_c\left[V_{\mathcal{M}^{(f_\star,P_\star)}(c)}^{\pi_\star(c;\cdot)}(s_0)\right] \\
&\leq \mathbb{E}_c[V_{\mathcal{M}^{(\hat{f}_t,\widehat{P}_t)}(c)}^{\pi_\star(c;\cdot)}(s_0)] && \text{(By Corollary 4.7)} \\
&\quad+\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A} q_h(s_h,a_h|\pi(c;\cdot),\widehat{P}_t^c)\big(b_{t,h}^R(c,s_h,a_h)+b_{t,h}^P(c,s_h,a_h)\big)\right] \\
&\quad+\frac{2029H^4d_{\mathcal{P}}}{\beta_P}\log^2(8TH|\mathcal{P}|/\delta)+\frac{504H^3d_{\mathcal{P}}}{\beta_r}\log^2(64T^4H|\mathcal{F}||\mathcal{P}|/\delta^2) \\
&=\mathbb{E}_c[V_{\widehat{\mathcal{M}}_t(c)}^{\pi_\star(c;\cdot)}(s_0)]+\frac{2029H^4d_{\mathcal{P}}}{\beta_P}\log^2(8TH|\mathcal{P}|/\delta) && \text{(Equation (2))} \\
&\quad+\frac{504H^3d_{\mathcal{P}}}{\beta_r}\log^2(64T^4H|\mathcal{F}||\mathcal{P}|/\delta^2) \\
&\leq \mathbb{E}_c\left[V_{\widehat{\mathcal{M}}_t(c)}^{\pi_t(c;\cdot)}(s_0)\right]+\frac{2029H^4d_{\mathcal{P}}}{\beta_P}\log^2(8TH|\mathcal{P}|/\delta) && (\pi_t \text{ is optimal in } \widehat{\mathcal{M}}_t)
\end{aligned}
$$

$$+ \frac{504H^3 d_{\mathcal{P}}}{\beta_r} \log^2(64T^4 H|\mathcal{F}||\mathcal{P}|/\delta^2),$$

as the lemma states. □

Lemma B.17 (restatement of Lemma 4.9)**.** *Under the good event of Corollary 4.7, we have that for every $t \geq 1$*

$$\begin{aligned}\mathbb{E}_c\left[V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)\right] \leq \mathbb{E}_c\left[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)\right] &+ 2\sum_{h=0}^{H-1}\mathbb{E}_c\left[\mathop{\mathbb{E}}_{\pi_t(c;\cdot),\widehat{P}^c_t}\left[b^R_{t,h}(c,s_h,a_h) + b^P_{t,h}(c,s_h,a_h)\right]\right] \\ &+ \frac{2029H^4 d_{\mathcal{P}}}{\beta_P}\log^2(8TH|\mathcal{P}|/\delta) \\ &+ \frac{504H^3 d_{\mathcal{P}}}{\beta_r}\log^2(64T^4H|\mathcal{F}||\mathcal{P}|/\delta^2).\end{aligned}$$

Proof. For all $t \geq 1$ the following holds.

$$\begin{aligned}
&\mathbb{E}_c\left[V^{\pi_t(c;\cdot)}_{\widehat{\mathcal{M}}_t(c)}(s_0)\right] \\
&= \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s\in S^c_h}\sum_{a_h\in A} q_h(s,a|\pi_t(c;\cdot),\widehat{P}^c_t)\cdot\left(\widehat{f}_t(c,s_h,a_h) + b^R_{t,h}(c,s_h,a_h) + b^P_{t,h}(c,s_h,a_h)\right)\right] && \text{(Equation (2))} \\
&= \mathbb{E}_c\left[V^{\pi_t(c;\cdot)}_{\mathcal{M}(\widehat{f}_t,\widehat{P}_t)(c)}(s_0)\right] \\
&\quad + \mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S^c_h}\sum_{a_h\in A} q_h(s_h,a_h|\pi_t(c;\cdot),\widehat{P}^c_t)\big(b^R_{t,h}(c,s_h,a_h) + b^P_{t,h}(c,s_h,a_h)\big)\right] \\
&\leq \mathbb{E}_c\left[V^{\pi_t(c;\cdot)}_{\mathcal{M}(f_\star,P_\star)(c)}(s_0)\right] && \text{(Corollary 4.7)} \\
&\quad + 2\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S^c_h}\sum_{a_h\in A} q_h(s_h,a_h|\pi_t(c;\cdot),\widehat{P}^c_t)\big(b^R_{t,h}(c,s_h,a_h) + b^P_{t,h}(c,s_h,a_h)\big)\right] \\
&\quad + \frac{2029H^4 d_{\mathcal{P}}}{\beta_P}\log^2(8TH|\mathcal{P}|/\delta) + \frac{504H^3 d_{\mathcal{P}}}{\beta_r}\log^2(64T^4H|\mathcal{F}||\mathcal{P}|/\delta^2),
\end{aligned}$$

and the proof follows by writing the second term as an expectation over s_h, a_h when playing $\pi_t(c;\cdot)$ on the dynamics $\widehat{P}^c_t$. □

B.5. Regret Bound

We begin with a technical result that bounds the expected cumulative bonuses.

Lemma B.18. *Let $b^R_{t,h}(c,s_h,a_h), b^P_t(c,s_h,a_h)$ be the reward bonuses in Equation (1). We have that*

$$\sum_{t=1}^{T}\sum_{h=0}^{H-1}\mathbb{E}_c\Big[\mathbb{E}_{\pi_t(c;\cdot),\widehat{P}^c_t}\left[b^R_{t,h}(c,s_h,a_h) + b^P_t(c,s_h,a_h)\right]\Big] \leq H|S||A|(\beta_r + H\beta_P)\log(T+1).$$

Proof. First we bound $b^R_{t,h}(c,s_h,a_h), b^P_{t,h}(c,s_h,a_h)$ by the second term in the minimum to get that

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=0}^{H-1}\mathbb{E}_c\Big[\mathbb{E}_{\pi_t(c;\cdot),\widehat{P}^c_t}\left[b^R_{t,h}(c,s_h,a_h) + b^P_{t,h}(c,s_h,a_h)\right]\Big] \\
&= \mathbb{E}_c\left[\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s_h\in S^c_h}\sum_{a_h\in A} q_h(s_h,a_h|\pi_t(c;\cdot),\widehat{P}^c_t)\left[b^R_{t,h}(c,s_h,a_h) + b^P_{t,h}(c,s_h,a_h)\right]\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{2}(\beta_r + \beta_P H)\mathbb{E}_c\left[\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A}\frac{q_h(s_h,a_h|\pi_t(c;\cdot),\widehat{P}_t^c)}{1+\sum_{i=1}^{t-1}q_h(s_h,a_h|\pi_i(c;\cdot),\widehat{P}_i^c)}\right]\\
&= \frac{1}{2}(\beta_r + \beta_P H)\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A}\sum_{t=1}^{T}\frac{q_h(s_h,a_h|\pi_t(c;\cdot),\widehat{P}_t^c)}{1+\sum_{i=1}^{t-1}q_h(s_h,a_h|\pi_i(c;\cdot),\widehat{P}_i^c)}\right]\\
&\leq \frac{1}{2}(\beta_r + \beta_P H)\mathbb{E}_c\left[\sum_{h=0}^{H-1}\sum_{s_h\in S_h^c}\sum_{a_h\in A}2\log(T+1)\right]\\
&= H|S||A|(\beta_r+\beta_P H)\log(T+1),
\end{aligned}
$$

where the last inequality used Lemma C.1 with $x_t = q_h(s_h,a_h|\pi_t(c;\cdot),\widehat{P}_t^c)$. □

Using all the above, we derive our main result stated in the following theorem.

Theorem B.19 (E-UC^{3}RL regret bound, restatement of Theorem 3.1)**.** *For any $T > 1$ and $\delta \in (0,1)$, suppose we run Algorithm 1 with parameters*

$$
\beta_r = \sqrt{\frac{504TH^2 d_{\mathcal{P}}\log^2(64T^4H|\mathcal{F}||\mathcal{P}|/\delta^2)}{|S||A|\log(T+1)}},\quad \beta_P = \sqrt{\frac{2029TH^2 d_{\mathcal{P}}\log^2(8TH|\mathcal{P}|/\delta)}{|S||A|\log(T+1)}},
$$

and $d_{\mathcal{P}} \geq d_{\mathrm{E}}(\mathcal{P}, D_H, T^{-1/2})$. Then, with probability at least $1-\delta$

$$
\mathcal{R}_T(\textit{E-UC}^3\textit{RL}) \leq \widetilde{O}\Big(H^3\sqrt{T|S||A|d_{\mathcal{P}}\left(\log(|\mathcal{F}|/\delta)+\log(|\mathcal{P}|/\delta)\right)}\Big).
$$

Proof of Theorem 3.1. We prove a regret bound under the following good events. The first event is that of Corollary 4.7, which occurs with probability at least $1-3\delta/4$. The second event is that

$$
\sum_{t=1}^{T} V_{\mathcal{M}(c_t)}^{\pi^\star(c_t;\cdot)}(s_0) - V_{\mathcal{M}(c_t)}^{\pi_t(c_t;\cdot)}(s_0) \leq \sum_{t=1}^{T}\mathbb{E}_c\left[V_{\mathcal{M}(c)}^{\pi^\star(c;\cdot)}(s_0)\right] - \mathbb{E}_c\left[V_{\mathcal{M}(c)}^{\pi_t(c;\cdot)}(s_0)\right] + H\sqrt{T\log(8/\delta)}. \tag{4}
$$

By Azuma's inequality (where the filtration is the histories $\{\mathbb{H}_t\}_{t=1}^T$), the above holds with probability at least $1-\delta/4$. Taking a union bound, the good event holds with probability at least $1-\delta$. Hence, assume the good events hold, and consider the following derivation.

$$
\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}_c\left[V_{\mathcal{M}(c)}^{\pi^\star(c;\cdot)}(s_0)\right] - \mathbb{E}_c\left[V_{\mathcal{M}(c)}^{\pi_t(c;\cdot)}(s_0)\right]\\
=&\sum_{t=1}^{T}\mathbb{E}_c\left[V_{\mathcal{M}(c)}^{\pi^\star(c;\cdot)}(s_0)\right] - \mathbb{E}_c\left[V_{\widehat{\mathcal{M}}_t(c)}^{\pi_t(c;\cdot)}(s_0)\right]\\
&+\sum_{t=1}^{T}\mathbb{E}_c\left[V_{\widehat{\mathcal{M}}_t(c)}^{\pi_t(c;\cdot)}(s_0)\right] - \mathbb{E}_c\left[V_{\mathcal{M}(c)}^{\pi_t(c;\cdot)}(s_0)\right]\\
\leq& 2\sum_{t=1}^{T}\sum_{h=0}^{H-1}\mathbb{E}_c\left[\mathop{\mathbb{E}}_{\pi_t(c;\cdot),\widehat{P}_t^c}\left[b_{t,h}^R(c,s_h,a_h)+b_{t,h}^P(c,s_h,a_h)\right]\right] && \text{(Lemmas 4.8 and 4.9)}\\
&+2T\frac{2029H^4 d_{\mathcal{P}}}{\beta_P}\log^2(8TH|\mathcal{P}|/\delta)\\
&+2T\frac{504H^3 d_{\mathcal{P}}}{\beta_r}\log^2(64T^4H|\mathcal{F}||\mathcal{P}|/\delta^2)\\
\leq& 2H|S||A|(\beta_r+H\beta_P)\log(T+1) && \text{(Lemma B.18)}\\
&+2T\frac{2029H^4 d_{\mathcal{P}}}{\beta_P}\log^2(8TH|\mathcal{P}|/\delta)
\end{aligned}
$$

$$
\begin{aligned}
&+ 2T\frac{504H^3 d_{\mathcal{P}}}{\beta_r}\log^2(64T^4H|\mathcal{F}||\mathcal{P}|/\delta^2)\\
=&4\sqrt{504T|S||A|H^4 d_{\mathcal{P}}\log(T+1)\log^2(64T^4H|\mathcal{F}||\mathcal{P}|/\delta^2} && \text{(Plugging in } \beta_P, \beta_r)\\
&+ 4\sqrt{2029T|S||A|H^6 d_{\mathcal{P}}\log(T+1)\log^2(8TH|\mathcal{P}|/\delta}\\
\leq&4\sqrt{2029T|S||A|H^6 d_{\mathcal{P}}\log(T+1)\log^2(8TH|\mathcal{P}|/\delta}\\
&270H^3\sqrt{T|S||A|d_{\mathcal{P}}\log(T+1)\log^2(18T^4H|\mathcal{F}||\mathcal{P}|/\delta^2)}.
\end{aligned}
$$

Finally, we get that

$$
\begin{aligned}
\mathcal{R}_T(\text{E-UC}^3\text{RL}) =& \sum_{t=1}^{T} V^{\pi^\star(c_t;\cdot)}_{\mathcal{M}(c_t)}(s_0) - V^{\pi_t(c_t;\cdot)}_{\mathcal{M}(c_t)}(s_0)\\
\leq& \sum_{t=1}^{T}\Big(\mathbb{E}_c\left[V^{\pi^\star(c;\cdot)}_{\mathcal{M}(c)}(s_0)\right] - \mathbb{E}_c\left[V^{\pi_t(c;\cdot)}_{\mathcal{M}(c)}(s_0)\right]\Big) && \text{(Equation (4))}\\
&+ H\sqrt{T\log(8/\delta)}\\
\leq&271H^3\sqrt{T|S||A|d_{\mathcal{P}}\log(T+1)\log^2(18T^4H|\mathcal{F}||\mathcal{P}|/\delta^2)}\\
=&\widetilde{O}\Big(H^3\sqrt{T|S||A|d_{\mathcal{P}}\left(\log(|\mathcal{F}|/\delta)+\log(|\mathcal{P}|/\delta)\right)}\Big). && \square
\end{aligned}
$$

C. Auxiliary lemmas

Lemma C.1. *Let $S_t = \lambda + \sum_{k=1}^{t-1} x_t$. Suppose $x_t \in [0, \lambda]$ and , then*

$$
\sum_{t=1}^{T}\frac{x_t}{S_t} \leq 2\log(T+1).
$$

Proof. The following holds.

$$
\begin{aligned}
\sum_{t=1}^{T}\frac{x_t}{S_t} &= \sum_{t=1}^{T}\frac{S_{t+1}-S_t}{S_t}\\
&= \sum_{t=1}^{T}\frac{S_{t+1}}{S_t} - 1\\
&\leq 2\sum_{t=1}^{T}\log\frac{S_{t+1}}{S_t} && (1 \leq \tfrac{S_{t+1}}{S_t} \leq 2 \text{ since } x_t \leq \lambda)\\
&= 2\sum_{t=1}^{T}\log S_{t+1} - \log S_t\\
&= 2\log\frac{S_{T+1}}{S_1} && \text{(telescopic sum)}\\
&\leq 2\log(T+1). && \square
\end{aligned}
$$

Lemma C.2 (value-difference, Corollary 1 in Shani et al., 2020)**.** *Let M, M' be any H-finite horizon MDPs. Then, for any two policies π, π' the following holds*

$$
\begin{aligned}
V_0^{\pi,M}(s) - V_0^{\pi',M'}(s) =& \sum_{h=0}^{H-1}\mathbb{E}\left[\langle Q_h^{\pi,M}(s_h,\cdot), \pi_h(\cdot|s_h) - \pi'_h(\cdot|s_h)\rangle | s_0 = s, \pi', M'\right]\\
&+ \sum_{h=0}^{H-1}\mathbb{E}\left[r_h(s_h,a_h) - r'_h(s_h,a_h) + (p_h(\cdot|s_h,a_h) - p'_h(\cdot|s_h,a_h))V_{h+1}^{\pi,M}|s_h = s, \pi', M'\right].
\end{aligned}
$$