Efficient Rate Optimal Regret for Adversarial Contextual MDPs Using Online Function Approximation

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

Abstract

We present the OMG-CMDP! algorithm for regret minimization in adversarial Contextual MDPs. The algorithm operates under the minimal assumptions of realizable function class and access to online least squares and log loss regression oracles. Our algorithm is efficient (assuming efficient online regression oracles), simple and robust to approximation errors. It enjoys an $\widetilde{O}(H^{2.5}\sqrt{T|S||A|(\mathcal{R}_{TH}(\mathcal{O}) + H\log(\delta^{-1}))})$ regret guarantee, with T being the number of episodes, S the state space, A the action space, H the horizon and $\mathcal{R}_{TH}(\mathcal{O}) = \mathcal{R}_{TH}(\mathcal{O}^{\mathcal{F}}_{\text{sq}}) + \mathcal{R}_{TH}(\mathcal{O}^{\mathcal{P}}_{\text{log}})$ is the sum of the square and log-loss regression oracles' regret, used to approximate the context-dependent rewards and dynamics, respectively. To the best of our knowledge, our algorithm is the first efficient rate optimal regret minimization algorithm for adversarial CMDPs that operates under the minimal standard assumption of online function approximation.

1. Introduction

Reinforcement Learning (RL) is a framework for sequential decision making in unknown environments. A Markov Decision Process (MDP) is the mathematical model behind RL environments. In the classical episodic RL setting, in each episode an agent repeatedly interacts with the MDP for H steps, observing the current state and then choosing an action. Subsequently, the agent receives a reward and the process transitions to the next state. The agent's goal is to choose actions as to maximize the cumulative return. This model characterizes many applications including online advertising, robotics, games and healthcare, and has been extensively studied over the past three decades (see e.g., Sutton & Barto, 2018; Mannor et al., 2022).

However, the classical MDP model cannot efficiently capture the influence of additional side information on the environment. Consider, for example, a medical trial that tests the effect of a new medication for a disease. The reaction of a patient to the treatment, which is modeled by the environment, is deeply influenced by the patient's medical history, which includes age, weight and background diseases she or he might have. We refer to such external information, which is unaffected by the agent's decisions, i.e., treatment choice, as the patient's *context*. A standard MDP can encode the context as part of the state. However, as no two patients are identical, this can significantly increase the size of the state space, and hence the complexity of learning and even the complexity of representing a single policy. Instead, Contextual MDPs (CMDPs), keep the state space small and treat the context as side information, revealed to the agent at the start of each episode. Additionally, there is an unknown mapping from each context to an MDP, thus an optimal policy maps each context to an optimal policy of the related MDP.

As previously alluded to, the context space is often prohibitively large, prompting the use of function approximation. This has previously been studied in the context of Contextual Multi Armed Bandits (CMAB; Foster & Rakhlin, 2020; Simchi-Levi & Xu, 2021), and more recently in CMDPs (Foster et al., 2021b; Levy & Mansour, 2022b). Concretely, realizable function approximation implies that the agent is provided with a function class of mappings from contexts to MDPs, and the goal is to obtain learning guarantees in terms of the class' complexity rather than the size of the context space. Realizability further implies that the true mapping resides in the class.

In recent years, CMDPs have gained much interest. There are two major lines of works. The distinctive feature between these two lines is whether the context is *stochastic* or *adversarially chosen*. Hallak et al. (2015) were the first to study CMDPs assuming an adversarial context. Modi &

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

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Tewari (2020) considered adversarial contexts and a Generalized Linear Model as a function class and gave a $\sqrt{T}$ regret guarantee. Foster et al. (2021b) consider general function classes as part of their Estimation to Decision (E2D) framework. They assume an access to an online estimation oracle that finds the best fit over the function class given past contexts and observations. (A detailed discussion and comparison with this work appears in the sequel.) For stochastic CMDPs, Levy & Mansour (2022b) presented sublinear regret under the minimum reachability assumption, using access to *offline* least squares regression oracles. They also showed an $\Omega(\sqrt{TH|S||A|\log(|\mathcal{G}|/|S|)/\log(|A|)})$ regret lower bound where $|\mathcal{G}|$ is the size of the function class used to approximate the rewards. Later, Levy et al. (2022) showed regret of $\widetilde{O}(\sqrt{T})$ assuming an access to offline regression oracles without the reachability assumption. Clearly, adversarial CMDPs generalize stochastic CMDPs and require the use of online function approximation.

Contributions. We study regret in adversarial CMDPs under a natural online function approximation setting. We present the OMG-CMDP! algorithm, and prove that its regret is, with probability at least $1-\delta$, bounded as

$$H^{2.5}\sqrt{T|S||A|\left(\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{F}}_{\text{sq}})+\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{P}}_{\log})+H\log\frac{1}{\delta}\right)}$$

(up to poly-logarithmic factors of $T, |S|, |A|, H$) where S is the state space, A the action space, H the horizon and $\mathcal{P}$ and $\mathcal{F}$ are function classes used to approximate the context dependent rewards and dynamics, respectively. We assume access to online log loss and least squares regression oracles for the dynamics and rewards approximation, and $\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{P}}_{\log})$, $\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{F}}_{\text{sq}})$ denote their regret guarantees, respectively. Our algorithm performs only $2T$ oracle calls and its running time is in $\text{poly}(|S|, |A|, H, T)$ assuming an efficient online regression oracles. The main advantage of our technique is its simplicity. We present an intuitive algorithm which operates under the standard online function approximation assumptions. In addition, at each round, our played policy can be approximated efficiently using standard convex optimization algorithms.

Comparison with Foster et al. (2021b). This work is most related to ours. They obtained $\sqrt{T\,\mathcal{R}_T(\mathcal{O}_{\text{est}})}$ regret by applying their Estimation to Decision (E2D) meta algorithm to adversarial CMDPs, where $\mathcal{O}_{\text{est}}$ is an online estimation oracle that operates over a class of CMDPs. Their work, being very general, leaves their oracle implementation non-concrete and generic, consequently giving rise to relatively complex algorithmic machinery. Specifically, at each round t their algorithm outputs a distribution p_t over context-dependent policies using an Inverse Gap Weighting (IGW) technique. To make this computationally tractable, they compute an approximate Policy Cover (PC) that serves as p_t's support, and analyze the delicate interplay between the approximation error and the IGW technique. In contrast, we use a standard online function approximation oracles for both the dynamics and rewards given related function classes, providing a clear model-based representation of the learned CMDP. Our approach yields a natural and intuitive algorithm, only requiring an approximate solution of a convex optimization problem. Thus, it can be implemented efficiently.

Additional Related Literature. Sample complexity bounds for Contextual Decision Processes (CDPs) have been studied under various assumptions. Jiang et al. (2017) present OLIVE, a sample efficient algorithm for learning Contextual Decision Processes (CDP) under the low Bellman rank assumption and later Sun et al. (2019) show PAC bounds for model based learning of CDPs using the Witness Rank.

Modi et al. (2018) present generalization bounds for learning *smooth* CMDPs and finite contextual linear combinations of MDPs. We, in contrast, consider the regret of adversarial CMDPs.

Levy & Mansour (2022a) studied the sample complexity of learning stochastic CMDPs using a standard ERM oracle, and provided the first general and efficient reduction from Stochastic CMDPs to offline supervised learning.

CMDPs naturally extend the well-studied 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, and give an optimal regret bound that depends on the size of the policy class they compete against. Foster et al. (2021a) present instance-dependent regret bounds for stochastic CMAB assuming access to a function class $\mathcal{F}$ for the rewards approximation. Regression based approaches appear in Agarwal et al. (2012); Foster & Rakhlin (2020); Foster et al. (2018); Foster & Krishnamurthy (2021); Simchi-Levi & Xu (2021); Xu & Zeevi (2020) for both stochastic and adversarial CMABs. Foster & Rakhlin (2020) assume access to an online least-squares regression oracle, and prove an optimal regret bound for adversarial CMABs using the IGW technique.

2. Preliminaries: Episodic Markov Decision Process (MDP)

A (tabular) MDP is defined by a tuple (S, A, P, r, s_0, H), where S and A are finite state and action spaces respectively; $s_0 \in S$ is the unique start state; $H \in \mathbb{N}$ is the horizon; $P : S \times A \times S \to [0, 1]$ is the dynamics which defines the probability of transitioning to state s' given that

we start at state s and play 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 plays action a_h then 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 $r(s_h, a_h) = \mathbb{E}_{\mathcal{D}_{s_h,a_h}}[R(s_h, a_h)]$.

A *stochastic and non-stationary policy* $\pi = (\pi_h : S \to \Delta(A))_{h\in[H]}$ defines for each time step $h \in [H]$ a mapping from states to a distribution over 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. Let $\pi^{\star}_M \in \arg\max_{\pi}\{V^{\pi}_M(s_0)\}$ denote an optimal policy for MDP M.

Furthermore, we consider the notation of *occupancy measures* (see, e.g., Zimin & Neu, 2013). 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. Let $\mu(P) \subseteq [0,1]^{H\times S\times A}$ denote the set of all occupancy measures defined by the dynamics P and any stochastic policy π. $\mu(P)$ is defined as follows. Each $q \in \mu(P)$ satisfies the following three requirements altogether.

(i) $q \in [0,1]^{H\times S\times A}$ and $\forall h \in [H]$, $q_h \in \Delta(S \times A)$;

(ii) For all $s \in S$, $\sum_{a\in A} q_0(s,a) = \mathbb{I}[s = s_0]$; and

(iii) For all $h \in [H-1]$ and $s \in S$, $\sum_{a\in A} q_{h+1}(s,a) = \sum_{(s',a')\in S\times A} P(s \mid s', a') q_h(s', a')$.

π^q is the policy associated with occupancy measure $q \in \mu(P)$ and is defined as follows for all $h \in [H]$ and state-action pair $(s,a) \in S \times A$,

$$\pi^q_h(a|s) = \frac{q_h(s,a)}{\sum_{a'\in A} q_h(s,a')}.$$

If $\sum_{a'\in A} q_h(s,a') = 0$ then $\pi^q_h(a|s) := 1/|A|$. In addition, note that $\mu(P)$ is a convex set. See, e.g., Rosenberg & Mansour (2019) for more details.

3. Problem Setup: Adversarial Contextual MDP

Following the notations of Levy & Mansour (2022b), 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 $r^c_\star(s,a) = \mathbb{E}[R^c_\star(s,a) \mid c, s, a]$, $R^c_\star(s,a) \sim \mathcal{D}_{c,s,a}$. We assume that $R^c_\star(s,a) \in [0,1]$.
For mathematical convenience, we assume the contexts space $\mathcal{C}$ is finite but potentially very large and we do not want to depend on its size (in both our regret bound and computation run-time). Our results are naturally extended to infinite contexts space.

We consider an *adversarial* CMDP where in each episode the context can be chosen in a completely arbitrary manner, possibly by an adversary.

A stochastic and non-stationary *context-dependent policy* $\pi = (\pi^c)_{c\in\mathcal{C}}$ maps a context $c \in \mathcal{C}$ to a policy $\pi^c = (\pi^c_h : S \to \Delta(A))_{h\in[H]}$.

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

(i) An adversary chooses a context $c_t \in \mathcal{C}$;
(ii) The agent chooses a policy $\pi^{c_t}_t$;
(iii) The agent observes a trajectory generated by playing $\pi^{c_t}_t$ in $\mathcal{M}(c_t)$, denoted as $\sigma^t = (c_t, s^t_0, a^t_0, r^t_0, \ldots, s^t_{H-1}, a^t_{H-1}, r^t_{H-1}, s^t_H)$.

Our goal is to minimize the regret, defined as

$$\mathcal{R}_T := \sum_{t=1}^{T} V^{\pi^{c_t}_\star}_{\mathcal{M}(c_t)}(s_0) - V^{\pi^{c_t}_t}_{\mathcal{M}(c_t)}(s_0), \tag{1}$$

where $\pi_\star$ is an optimal context-dependent policy.

3.1. Assumptions

In this setting, without further assumptions, the regret may scale linearly in TH. We overcome this limitation by imposing the following minimal online function approximation assumptions, which extend similar notions in the CMAB literature (see, e.g., Agarwal et al., 2012; Foster et al., 2018; Foster & Rakhlin, 2020; Foster & Krishnamurthy, 2021) to CMDPs. We assume access to realizable function classes $\mathcal{F}$ and $\mathcal{P}$ that serve to approximate the context-dependent rewards and dynamics respectively. Realizability means that the true rewards and dynamics belong to the appropriate function class, and access is via online regression oracles (more details later).

Online regression oracle. We consider a standard online regression oracle with respect to a given loss function ℓ. The oracle performs real-valued online regression where the examples are chosen from some subspace $\mathcal{Z}$, with respect to a loss function ℓ and has a regret guarantee relative to the function class $\mathcal{F}$. We consider the following

online scenario. For every round $t = 1, ..., T$: (1) An adversary (possibly adaptive) chooses input $z_t \in \mathcal{Z}$. (2) The oracle observes z_t and returns a prediction $\hat{y}_t \in [0, 1]$. (3) The adversary chooses an outcome $y_t \in [0, 1]$. We follow Foster & Rakhlin (2020) and model the online oracle as a sequence of mappings $\mathcal{O}_\ell^t : \mathcal{Z} \times (\mathcal{Z}, \mathbb{R})^{t-1} \to [0, 1]$, where $\hat{y}_t = \mathcal{O}_\ell^t(z_t; (z_1, y_1), ..., (z_{t-1}, y_{t-1}))$. Each oracle implementation induces a mapping $\hat{f}_t(z) = \mathcal{O}_\ell(z; (z_1, y_1), \ldots, (z_{t-1}, y_{t-1}))$ which is the prediction for the input z at round t. (See Section 2.1 in Foster & Rakhlin, 2020). The oracles' regret guarantee with respect to the function class $\mathcal{F}$ is as follows.

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(z_t), y_t) \leq \mathcal{R}_T(\mathcal{O}_\ell^{\mathcal{F}}).$$

In the following, we consider the online regression oracle with respect to the square loss i.e., $\ell_{sq}(\hat{y}, y) = (\hat{y} - y)^2$ for the rewards approximation, and the log loss i.e., $\ell_{\log}(\hat{y}, y) = \log(y/\hat{y})$ for the dynamics approximation. For finite function classes $\mathcal{F}$ and $\mathcal{P}$, it is known that the regret of these oracles is logarithmic in the function class size (Cesa-Bianchi & Lugosi, 2006; Foster & Rakhlin, 2020; Foster et al., 2021b). Meaning, $\mathcal{R}_T(\mathcal{O}_{\text{sq}}^{\mathcal{F}}) = O(\log(|\mathcal{F}|))$ and $\mathcal{R}_T(\mathcal{O}_{\log}^{\mathcal{P}}) = O(\log(|\mathcal{P}|))$.[1] The above optimization problems can always be solved by iterating over the function class. But, since we consider strongly convex loss functions, there are function classes where these optimization problems can be solved efficiently. An obvious example is the class of linear functions.

Reward function approximation. We assume that the learner has access to a class of reward functions $\mathcal{F} \subseteq \mathcal{C} \times S \times A \to [0, 1]$, each function $f \in \mathcal{F}$ maps context $c \in \mathcal{C}$, state $s \in S$ and an action $a \in A$ to a (approximate) reward $r \in [0, 1]$. We use $\mathcal{F}$ to approximate the context-dependent rewards function of any state $s \in S$ and action $a \in A$ using an online least-squares regression (OLSR) oracle under the following realizability assumption.

Assumption 1. *There exists a function $f_\star \in \mathcal{F}$ such that for all t and $(s, a) \in S \times A$, $f_\star(c_t, s, a) = r_\star^{c_t}(s, a)$.*

Assumption 2 (Square Loss Oracle Regret). *The oracle $\mathcal{O}_{\text{sq}}^{\mathcal{F}}$ guarantees that for every sequence of trajectories $\{\sigma^t\}_{t=1}^T$, regret is bounded as*

$$\sum_{t=1}^{T} \sum_{h=0}^{H-1} (\hat{f}_t(c_t, s_h^t, a_h^t) - r_h^t)^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \sum_{h=0}^{H-1} (f(c_t, s_h^t, a_h^t) - r_h^t)^2 \leq \mathcal{R}_{TH}(\mathcal{O}_{\text{sq}}^{\mathcal{F}}).$$

Dynamics function approximation. For the unknown context-dependent dynamics case, our algorithm gets as input a function class $\mathcal{P} \subseteq S \times A \times S \times \mathcal{C} \to [0, 1]$, where every function $P \in \mathcal{P}$ satisfies $\sum_{s' \in S} P(s' \mid s, a, c) = 1$ for all $c \in \mathcal{C}$ and $(s, a) \in S \times A$. We use $\mathcal{P}$ to approximate the context-dependent dynamics using an online log-loss regression (OLLR) oracle under the following realizability assumption. For any $P \in \mathcal{P}$ we denote $P^c(s' \mid s, a) := P(s' \mid s, a, c)$.

Assumption 3 (Dynamics Realizability). *There exists a function $P \in \mathcal{P}$ such that for all t, and every $(s, a, s') \in S \times A \times S$, $P(s' \mid s, a, c_t) = P_\star^{c_t}(s' \mid s, a)$.*

Assumption 4 (Log Loss Oracle Regret). *Given a function class $\mathcal{P}$ of context-dependent transition probabilities function, the oracle $\mathcal{O}_{\log}^{\mathcal{P}}$ guarantees that for every sequence of trajectories $\{(c_t; s_0^t, a_0^t, r_0^t, \ldots, s_H^t)\}_{t=1}^T$, regret is bounded as*

$$\sum_{t=1}^{T} \sum_{h=0}^{H-1} \log \frac{1}{\widehat{P}_t^{c_t}(s_{h+1}^t \mid s_h^t, a_h^t)} - \inf_{P \in \mathcal{P}} \sum_{t=1}^{T} \sum_{h=0}^{H-1} \log \frac{1}{P^{c_t}(s_{h+1}^t \mid s_h^t, a_h^t)} \leq \mathcal{R}_{TH}(\mathcal{O}_{\log}^{\mathcal{P}}).$$

Notations. For an event E we denote by $\mathbb{I}[E]$ the indicator function which returns 1 if E holds and 0 otherwise. We denote expectation by $\mathbb{E}[\cdot]$. We also use the following abbreviations for the oracles regrets. We denote $\mathcal{R}^{\log} := \mathcal{R}_{TH}(\mathcal{O}_{\log}^{\mathcal{P}})$ and $\mathcal{R}^{\text{sq}} := \mathcal{R}_{TH}(\mathcal{O}_{\text{sq}}^{\mathcal{F}})$. Note that our oracles are called T times, each time we feed them with H examples. For that reason their regret bounds depend on TH. This in contrast to previous works in which the oracle gets only one example in each call. When using the notation $\widetilde{O}(\cdot)$ we omit poly-logarithmic factors of $|S|, |A|, H, T$. For a vector $x \in \mathbb{R}^d$ we denote by $\|x\|_1 := \sum_{i=1}^d |x_i|$ the ℓ_1 norm of x.

4. Algorithm and Main Result

We present the *Occupancy Measures approximated reGularization algorithm for regret minimization in adversarial CMDP* (OMG-CMDP!; Algorithm 1). At each round $t = 1, 2, \ldots, T$, we first approximate the rewards and dynamics, using the online oracles, based on the observed

[1] We remark that in the case that $\mathcal{F}$ or $\mathcal{P}$ are non-convex, some implementations of the oracles might return a function in the convex hull of $\mathcal{F}$ and $\mathcal{P}$, respectively. Such an implementation is, for instance, Vovk's aggregation algorithm (Cesa-Bianchi & Lugosi, 2006).

trajectories up to round $t-1$. We denote by $\hat{f}_t$ the approximated rewards function, and by $\widehat{P}_t$ the approximated dynamics at round t. After observing the current context c_t, we solve the optimization problem in Equation (2) over the set of occupancy measures defined by $\widehat{P}_t^{c_t}$. For the optimal solution $\hat{q}^t$, we derive the appropriate policy $\pi_t^{c_t}$ and run it to generate a trajectory. We use the observed trajectory to update the oracles.

Algorithm 1 Occupancy Measures approximated reGularization for adversarial CMDP (OMG-CMDP!)

inputs:

- MDP parameters: S, A, s_0, H.
- Function classes $\mathcal{F}$ for rewards approximation and $\mathcal{P}$ for dynamics approximation.
- Confidence parameter $\delta \in (0,1)$ and tuning parameter γ.
- OLSR oracle $\mathcal{O}_{\text{sq}}^{\mathcal{F}}$ and OLLR oracle $\mathcal{O}_{\log}^{\mathcal{P}}$.

for round $t = 1, \dots, T$ **do**

approximate rewards $\hat{f}_t = \mathcal{O}_{\text{sq}}^{\mathcal{F}}$ and dynamics $\widehat{P}_t = \mathcal{O}_{\log}^{\mathcal{P}}$.

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

solve

$$\hat{q}^t = \arg\max_{q \in \mu(\widehat{P}_t^{c_t})} \sum_{\substack{(h,s,a)\in \\ [H]\times S\times A}} q_h(s,a) \cdot \hat{f}_t(c_t,s,a) + \frac{1}{\gamma} \sum_{\substack{(h,s,a)\in \\ [H]\times S\times A}} \log(q_h(s,a)). \tag{2}$$

derive policy as follows, for all $h \in [H]$ and $(s,a) \in S \times A$.

$$\pi_{t,h}^{c_t}(a \mid s) = \frac{\hat{q}_h^t(s,a)}{\sum_{a' \in A} \hat{q}_h^t(s,a')}.$$

play $\pi_t^{c_t}$ and observe a trajectory

$$\sigma^t = (c_t; s_0^t, a_0^t, r_0^t, \dots, s_H^t).$$

update $\mathcal{O}_{\text{sq}}^{\mathcal{F}}$ using $\{((c_t, s_h^t, a_h^t), r_h^t)\}_{h=0}^{H-1}$.

update $\mathcal{O}_{\log}^{\mathcal{P}}$ using $\{(c_t, s_h^t, a_h^t, s_{h+1}^t)\}_{h=0}^{H-1}$.

end for

A key step in our algorithm is reflected in the optimization problem (Equation (2)). We solve the maximization problem over the set of occupancy measures of the approximated dynamics $\widehat{P}_t^{c_t}$. We remark that the maximization problem in Equation (2) is a strictly concave maximization problem over the set of occupancy measures of $\widehat{P}_t^{c_t}$. Hence it has a single global maximum that can be approximated efficiently.

Note that usually in regret-minimization RL literature one optimizes over the empirical dynamics with additional bonuses to the rewards that promote exploration (optimism). The magnitude of the bonuses is derived directly from the size of the confidence intervals over the dynamics and rewards approximation. In our case, however, the contexts are adversarially chosen hence it is impossible to define such confidence intervals. Nevertheless, the use of log-barrier regularization provides the necessary trade-off between exploration and exploitation, in a manner resembling optimistic approaches, thus replacing the need for the aforementioned bonuses. The application of log-barrier regularization differentiates us from previous works in contextual RL literature.

Our main result is the regret guarantee of Algorithm 1, stated in the following theorem.

Theorem 5 (Regret Bound). *Let $\delta \in (0,1)$. For $\gamma = \sqrt{\frac{|S||A|T}{31H^3(2\mathcal{R}^{\text{sq}}+\mathcal{R}^{\log}+18H\log(2H/\delta))}}$, with probability at least $1-\delta$ it holds that $\mathcal{R}_T$(OMG-CMDP!) is bounded by*

$$\widetilde{O}\left(H^{2.5}\sqrt{T|S||A|\left(\mathcal{R}^{\text{sq}} + \mathcal{R}^{\log} + H\log\delta^{-1}\right)}\right).$$

For finite function classes $\mathcal{F}$ and $\mathcal{P}$ the following corollary is immediately implied by Theorem 5.

Corollary 6. *Let $\mathcal{F}$ and $\mathcal{P}$ be finite function classes for the rewards and dynamics approximation, respectively. Let $\delta \in (0,1)$. There exists oracle implementations such that for an appropriate choice of γ, with probability at least $1-\delta$, $\mathcal{R}_T$(OMG-CMDP!) is bounded by*

$$\widetilde{O}\left(H^{2.5}\sqrt{T|S||A|\left(\log|\mathcal{F}| + \log|\mathcal{P}| + H\log\delta^{-1}\right)}\right).$$

5. Analysis

Our regret analysis consists of two main conceptual steps. The first step is to derive concentration bounds on our oracles' regret (Section 5.1). Next, note that the regret defined in Equation (1) can be decomposed as follows,

$$\mathcal{R}_T = \sum_{t=1}^{T} V_{\mathcal{M}(c_t)}^{\pi_\star^{c_t}}(s_0) - V_{\widehat{\mathcal{M}}_t(c_t)}^{\pi_\star^{c_t}}(s_0) \tag{3}$$

$$+ \sum_{t=1}^{T} V_{\widehat{\mathcal{M}}_t(c_t)}^{\pi_\star^{c_t}}(s_0) - V_{\widehat{\mathcal{M}}_t(c_t)}^{\pi_t^{c_t}}(s_0) \tag{4}$$

$$+ \sum_{t=1}^{T} V_{\widehat{\mathcal{M}}_t(c_t)}^{\pi_t^{c_t}}(s_0) - V_{\mathcal{M}(c_t)}^{\pi_t^{c_t}}(s_0), \tag{5}$$

where Equation (3) is the cumulative difference between the value of the true optimal policy $\pi_\star$ on the true and approximated MDPs of the context c_t at round t. Equation (4)

is the cumulative difference between the value of $\pi_\star$ and π_t on the approximated model at round t, for the context c_t. Equation (5) is the cumulative difference between the value of the selected policy π_t on the approximated and true MDPs at round t for the context c_t. In Section 5.2 we upper bound the three value-difference sums in terms of the oracles' expected regret. By applying concentration bounds (Lemmas 8 and 9), we bound the oracles' expected regret by the empirical one, thus bounding the regret of our algorithm with high probability.

Throughout the analysis we consider the cumulative error caused by the dynamics approximation in terms of the Squared Hellinger distance between the true and approximated dynamics.

Definition 7 (Squared Hellinger Distance). For any two distributions $\mathbb{P}$, $\mathbb{Q}$ over a discrete support X we define the Squared Hellinger Distance as

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

A useful property of the squared Hellinger distance is that for any two distributions $\mathbb{P}$ and $\mathbb{Q}$ it holds that

$$\|\mathbb{P}-\mathbb{Q}\|_1^2 \le 4D_H^2(\mathbb{P},\mathbb{Q}). \tag{6}$$

We bound the cumulative value differences (Equations (3) to (5)) using the following quantities:

(1) The cumulative expected least-squares loss over each round t, i.e.,

$$\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}\left(\hat{f}_t(c_t,s_h,a_h)-f_\star(c_t,s_h,a_h)\right)^2\,\middle|\,s_0\right].$$

For abbreviation, we denote the above by $\mathcal{E}_T(\ell_{\mathrm{sq}})$.

(2) The cumulative expected squared Hellinger distance over each round t, i.e.,

$$\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}D_H^2(P_\star^{c_t}(\cdot|s_h,a_h),\widehat{P}_t^{c_t}(\cdot|s_h,a_h))\,\middle|\,s_0\right].$$

For abbreviation, we denote the above by $\mathcal{E}_T(D_H^2)$.

5.1. Oracle Concentration Bounds

The following lemma bounds the expected regret of the online least squares regression oracle by its realized regret.

Lemma 8 (Concentration of OLSR regret). *Under Assumptions 1 and 2, for any $\delta\in(0,1)$, the following holds with probability at least $1-\delta/2$.*

$$\mathcal{E}_T(\ell_{\mathrm{sq}}) \le \mathcal{R}^{\mathrm{sq}} + 16H\log(2/\delta).$$

We analyze the expected regret of the log loss regression oracle in terms of the Hellinger distance. The following lemma is an immediate implication of Lemma A.14 in (Foster et al., 2021b).

Lemma 9 (Concentration of LLR regret w.r.t Hellinger distance). *Under Assumptions 3 and 4, for any $\delta\in(0,1)$, with probability at least $1-\delta/2$ it holds that*

$$\mathcal{E}_T(D_H^2) \le \mathcal{R}^{\log} + 2H\log(2H/\delta).$$

See Appendix A.1 for full proofs.

5.2. Value-Difference Bounds

In this subsection we bound the three value difference sums (Equations (3) to (5)). For that purpose, we represent the value function in terms of the *occupancy measures* (Zimin & Neu, 2013). Recall that occupancy measures are defined as follows. 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. Thus, the value function of any policy π with respect to the MDP (S,A,P,r,s_0,H) can be presented using the occupancy measures as follows.

$$V_M^\pi(s_0) = \sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} q_h(s,a\mid\pi,P)\cdot r(s,a). \tag{7}$$

In the analysis, we use the following notations, to denote a specific occupancy measure of interest. For all $(s,a,h)\in S\times A\times[H]$ we denote:

(i) $\hat{q}_h^t(s,a) := q_h(s,a\mid\pi_t^{c_t},\widehat{P}_t^{c_t})$;

(ii) $q_h^t(s,a) := q_h(s,a\mid\pi_t^{c_t},P_\star^{c_t})$;

(iii) $\hat{q}_h^{t,\star}(s,a) := q_h(s,a\mid\pi_\star^{c_t},\widehat{P}_t^{c_t})$.

We now have all the required tools to state our cumulative value difference bounds. We first bound the value difference caused by the CMDP approximation error, for the true optimal context-dependent policy $\pi_\star$.

Lemma 10 (The cost of approximation for $\pi_\star$). *The following holds for any choice of parameter $\hat{\gamma}>0$.*

$$\begin{aligned}\sum_{t=1}^{T} V_{\mathcal{M}(c_t)}^{\pi_\star^{c_t}}(s_0) - V_{\widehat{\mathcal{M}}_t(c_t)}^{\pi_\star^{c_t}}(s_0) &\le \sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\frac{\hat{q}_h^{t,\star}(s,a)}{\hat{\gamma}\cdot\hat{q}_h^t(s,a)} \\ &\quad + 2\hat{\gamma}\cdot\mathcal{E}_T(\ell_{\mathrm{sq}}) + 29\hat{\gamma}H^4\cdot\mathcal{E}_T(D_H^2).\end{aligned}$$

Proof sketch. Fix $t \in [T]$ and apply the value difference lemma (Lemma 23 in the Appendix) to obtain

$$\begin{aligned} &V^{\pi_\star^{c_t}}_{\mathcal{M}(c_t)}(s_0) - V^{\pi_\star^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) \\ &\quad \le \sum_{h,s,a} \hat{q}_h^{t,\star}(s,a) \cdot (f_\star(c_t,s,a) - \hat{f}_t(c_t,s,a)) \\ &\qquad + H \sum_{h,s,a} \hat{q}_h^{t,\star}(s,a) \cdot \|P_\star^{c_t}(\cdot|s,a) - \widehat{P}_t^{c_t}(\cdot|s,a)\|_1. \end{aligned}$$

We then multiply each term in both sums by $\sqrt{\frac{\hat{\gamma}\cdot\hat{q}_h^t(s,a)}{\hat{\gamma}\cdot\hat{q}_h^t(s,a)}}$ and apply the arithmetic-geometric means (AM-GM) inequality to change the occupancy measure form $\hat{q}^{t,\star}$ to $\hat{q}^t$ and add a dependency in $\hat{\gamma}$. We obtain that the previous is bounded as

$$\begin{aligned} &\le \sum_{h,s,a} \frac{\hat{q}_h^{t,\star}(s,a)}{\hat{\gamma}\cdot\hat{q}_h^t(s,a)} \\ &\quad + \frac{\hat{\gamma}}{2} \sum_{h,s,a} \hat{q}_h^t(s,a) \cdot (f_\star(c_t,s,a) - \hat{f}_t(c_t,s,a))^2 \\ &\quad + \frac{\hat{\gamma} H^2}{2} \sum_{h,s,a} \hat{q}_h^t(s,a) \cdot \|P_\star^{c_t}(\cdot|s,a) - \widehat{P}_t^{c_t}(\cdot|s,a)\|_1^2. \end{aligned}$$

Lastly we apply Equation (6) to bound the squared ℓ_1 norm with the squared Hellinger distance, and then the occupancy measure change (Corollary 22 in the Appendix) to replace $\hat{q}^t$ with q^t. We obtain that the latter is bounded by

$$\le \sum_{h,s,a} \frac{\hat{q}_h^{t,\star}(s,a)}{\hat{\gamma}\cdot\hat{q}_h^t(s,a)} + 2\hat{\gamma}\cdot\mathcal{E}_T(\ell_{\mathrm{sq}}) + 29\hat{\gamma}H^4\cdot\mathcal{E}_T(D_H^2).$$

The lemma follows by summing over each $t \in [T]$. For more details see Lemma 26 and Corollary 27 in the Appendix. ■

Next, we bound the cumulative value difference between $\pi_\star$ and π_t on the approximated model in round t.

Lemma 11 (Suboptimality of $\pi_\star$ in $\widehat{\mathcal{M}}_t$). *It holds that*

$$\begin{aligned} &\sum_{t=1}^T V^{\pi_\star^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) - V^{\pi_t^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) \\ &\qquad \le \frac{TH|S||A|}{\gamma} - \sum_{t=1}^T \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \frac{\hat{q}_h^{t,\star}(s,a)}{\gamma\cdot\hat{q}_h^t(s,a)}. \end{aligned}$$

Proof. For every round $t \in [T]$, consider the first order derivative of the concave objective function in Equation (2) and denote it by $\hat{L}_t'(q;c_t)$. i.e.,

$$\hat{L}_t'(q;c_t) = \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \left(\hat{f}_t(c_t,s,a) + \frac{1}{\gamma\cdot q_h(s,a)} \right).$$

Let $\pi_\star = (\pi_\star^c)_{c\in\mathcal{C}}$ denote an optimal context-dependent policy for the true CMDP. For every round t, the occupancy measure $\hat{q}_h^{t,\star}(s,a) := q_h(s,a|\pi_\star^{c_t}, \widehat{P}_t^{c_t})$ is a feasible solution for the maximization problem in Equation (2), since $\hat{q}^{t,\star} \in \mu(\widehat{P}_t^{c_t})$. Since $\hat{q}^t$ is the optimal solution, by first order optimality conditions for concave functions (Boyd, Boyd, and Vandenberghe, 2004) it holds that

$$\begin{aligned} &\sum_{h=0}^{H-1} \sum_{s,a} \hat{q}_h^{t,\star}(s,a) \cdot \left(\hat{f}_t(c_t,s,a) + \frac{1}{\gamma\hat{q}_h^t(s,a)} \right) \\ &\quad - \sum_{h=0}^{H-1} \sum_{s,a} \hat{q}_h^t(s,a) \cdot \left(\hat{f}_t(c_t,s,a) + \frac{1}{\gamma\hat{q}_h^t(s,a)} \right) \le 0, \end{aligned}$$

which implies that

$$\begin{aligned} &\sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} (\hat{q}_h^{t,\star}(s,a) - \hat{q}_h^t(s,a)) \hat{f}_t(c_t,s,a) \\ &\quad \le \frac{H|S||A|}{\gamma} - \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \frac{\hat{q}_h^{t,\star}(s,a)}{\gamma\cdot\hat{q}_h^t(s,a)}. \end{aligned} \tag{8}$$

By the value representation using occupancy measures (Equation (7)), for every round $t \in [T]$ it holds that,

$$\begin{aligned} &V^{\pi_\star^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) - V^{\pi_t^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) \\ &\quad = \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} (\hat{q}_h^{t,\star}(s,a) - \hat{q}_h^t(s,a)) \cdot \hat{f}_t(c_t,s,a). \end{aligned} \tag{9}$$

Hence, the lemma follows by combining Equations (8) and (9) and summing over each round $t \in [T]$. ■

Note that for the choice in $\hat{\gamma} = \gamma$ for Lemma 10, the term $\sum_{t=1}^T \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \frac{\hat{q}_h^{t,\star}(s,a)}{\hat{\gamma}\cdot\hat{q}_h^t(s,a)}$ is canceled with the second term in RHS of Lemma 11.

Lastly, we bound the value difference caused by the approximation, for the selected policy π_t.

Lemma 12 (The cost of approximation for π_t). *The following holds for any two parameters $p_1, p_2 > 0$.*

$$\begin{aligned} &\sum_{t=1}^T V^{\pi_t^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) - V^{\pi_t^{c_t}}_{\mathcal{M}(c_t)}(s_0) \\ &\qquad \le \frac{TH}{2p_1} + \frac{p_1}{2}\cdot\mathcal{E}_T(\ell_{\mathrm{sq}}) + \frac{TH}{2p_2} + 2p_2\cdot\mathcal{E}_T(D_H^2). \end{aligned}$$

To prove the lemma, we use the value difference lemma and AM-GM with the parameters p_1, p_2 similarly to showm for Lemma 10.

5.3. Regret Bound

We obtain Theorem 5 by combining the results of Lemmas 10 to 12 and applying our concentration bounds stated in Lemmas 8 and 9. See proof sketch bellow. (For detailed proof see Appendix A.2.3).

Proof sketch of Theorem 5. We start with the regret decomposition to the three sums.

$$\begin{aligned}
&\mathcal{R}_T(\text{OMG-CMDP!}) \\
&= \sum_{t=1}^{T} V^{\pi_\star^{c_t}}_{\mathcal{M}(c_t)}(s_0) - V^{\pi_\star^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) \\
&\quad + \sum_{t=1}^{T} V^{\pi_\star^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) - V^{\pi_t^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) \\
&\quad + \sum_{t=1}^{T} V^{\pi_t^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) - V^{\pi_t^{c_t}}_{\mathcal{M}(c_t)}(s_0).
\end{aligned}$$

We bound the the first term by Lemma 10, the second by Lemma 11 and the last by Lemma 12. This yields the following bound.

$$\begin{aligned}
\mathcal{R}_T &\le 2\gamma \cdot \mathcal{E}_T(\ell_{\text{sq}}) + 29\gamma H^4 \cdot \mathcal{E}_T(D_H^2) + \frac{H|S||A|T}{\gamma} \\
&\quad + \frac{TH}{2p_1} + \frac{p_1}{2} \cdot \mathcal{E}_T(\ell_{\text{sq}}) + \frac{TH}{2p_2} + 2p_2 \cdot \mathcal{E}_T(D_H^2).
\end{aligned}$$

By Lemmas 8 and 9, the following holds with probability at least $1-\delta$.

$$\begin{aligned}
\mathcal{R}_T &\le 2\gamma\left(2 \cdot \mathcal{R}^{\text{sq}} + 16H\log(2/\delta)\right) \\
&\quad + 29\gamma H^4\left(\mathcal{R}^{\log} + 2H\log(2H/\delta)\right) \\
&\quad + \frac{H|S||A|T}{\gamma} \\
&\quad + \frac{TH}{2p_1} + \frac{p_1}{2}\left(2\mathcal{R}^{\text{sq}} + 16H\log(2/\delta)\right) \\
&\quad + \frac{TH}{2p_2} + 2p_2\left(\mathcal{R}^{\log} + 2H\log(2H/\delta)\right) \\
&= \widetilde{O}\left(H^{2.5}\sqrt{T|S||A|\left(\mathcal{R}^{\text{sq}} + \mathcal{R}^{\log} + H\log\delta^{-1}\right)}\right)
\end{aligned}$$

where the last identity is by our choice of γ, and for $\hat{\gamma} = \gamma$, $p_1 = \sqrt{\frac{TH}{2\mathcal{R}^{\text{sq}}+16H\log(2/\delta)}}$ and $p_2 = \sqrt{\frac{TH}{4(\mathcal{R}^{\log}+2H\log(2H/\delta))}}$. Since both good events hold with probability at least $1-\delta$ we obtain the theorem. ■

6. Approximated Solution

The objective of the optimization problem in Equation (2) is a sum of a self-concordant barrier function (the log function) and a linear function. Hence, the optimal solution for the problem can be approximated efficiently using interior-point convex optimization algorithms such as Newton's Method. These algorithms return an ϵ-approximated solution and have a running time of $O(\text{poly}(d)\log\epsilon^{-1})$, where $d = H|S||A|$ is the dimension of the problem (Nesterov & Nemirovskii, 1994).

Suppose that in each round t we derive the policy $\pi_t^{c_t}$ using, instead of the optimal solution, an occupancy measure $\hat{q}^t$ that yields an ϵ-approximation to the objective of the optimization problem in Equation (2). The following analysis shows that for $\epsilon = \frac{1}{16\gamma T}$, we obtain a similar regret guarantee. In addition, by our choice of γ, the running time complexity of the optimization algorithm is $\text{poly}(|S|, |A|, H, \log(T))$. We start by bounding the difference between the optimal and the approximated iterates. (See proof in Lemma 33 in the Appendix.)

Lemma 13 (Iterates' difference). *For every round t let*

$$\begin{aligned}
\hat{L}_t(q; c_t) &= \sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} q_h(s,a)\cdot \hat{f}_t(c_t,s,a) \\
&\quad + \frac{1}{\gamma}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\log(q_h(s,a)).
\end{aligned}$$

Denote the objective of the optimization problem in Equation (2). Let $\widetilde{q} \in \arg\max_{q\in\mu(\widehat{P}_t^{c_t})} \hat{L}_t(q; c_t)$. Let $q \in \mu(\widehat{P}_t^{c_t})$ and suppose that $\hat{L}_t(\widetilde{q}; c_t) - \hat{L}_t(q; c_t) \le \epsilon$. Then,

$$\sum_{s\in S}\sum_{a\in A}\sum_{h=0}^{H-1}\left(\frac{q_h(s,a)}{\widetilde{q}_h(s,a)} - 1\right)^2 \le 4\epsilon\gamma.$$

Using Lemma 13, we modify the bound of Lemma 11 and obtain the following corollary. (See Lemma 34 in the Appendix for full proof.)

Corollary 14. *For every round $t \in [T]$ and a context $c_t \in \mathcal{C}$, let $\widetilde{q}^t$ be the optimal solution to the maximization problem in Equation (2). Suppose that $\hat{q}_h^t \in \mu(\widehat{P}_t^{c_t})$ satisfies $\hat{L}_t(\widetilde{q}^t; c_t) - \hat{L}_t(\hat{q}^t; c_t) \le \epsilon$, and $\epsilon\gamma \le 1/16$. Then,*

$$\begin{aligned}
\sum_{t=1}^{T} V^{\pi_\star^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) - V^{\pi_t^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) &\le \frac{H|S||A|T}{\gamma} \\
&\quad - \sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\frac{\hat{q}_h^{t,\star}(s,a)}{2\gamma\cdot\hat{q}_h^t(s,a)} + 2T\sqrt{\epsilon\gamma H}.
\end{aligned}$$

By replacing Lemma 11 with Corollary 14 we derive the following regret bound using similar steps to those shown in Section 5.3. (See Appendix B for full proofs.)

Theorem 15 (Regret bound). *For any $\delta \in (0,1)$, let $\gamma = \sqrt{\frac{|S||A|T}{62H^3(2\mathcal{R}^{\text{sq}}+\mathcal{R}^{\log}+18H\log(2H/\delta))}}$. Suppose that at*

each round t we have an ϵ-approximation to the optimal solution of Equation (2) *for* $\epsilon = \frac{1}{16\gamma T}$. *Then, with probability at least* $1-\delta$, $\mathcal{R}_T$(*Approx OMG-CMDP!*) *is bounded as*

$$\widetilde{O}\left(H^{2.5}\sqrt{T|S||A|\left(\mathcal{R}^{\text{sq}}+\mathcal{R}^{\log}+H\log\delta^{-1}\right)}\right).$$

7. Discussion

In this paper we provide the first efficient reduction from Adversarial CMDPs to online regression. The novelty of our approach is the use of concave optimization with log-barrier regularization over occupancy measures. This technique might prove useful in other settings of function approximation with a small underlying state space, e.g., block or rich observation MDPs. We note that there is an H^2 gap between our regret upper bound and the lower bound of Levy & Mansour, 2022b. We leave closing this gap as an open problem for future research.

Acknowledgements

AC is supported by the Israeli Science Foundation (ISF) grant no. 2250/22.
This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), by the Israeli 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.

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.

Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Cohen, A., Efroni, Y., Mansour, Y., and Rosenberg, A. Minimax regret for stochastic shortest path. *Advances in Neural Information Processing Systems*, 34:28350–28361, 2021.

Efroni, Y., Shani, L., Rosenberg, A., and Mannor, S. Optimistic policy optimization with bandit feedback. *arXiv preprint arXiv:2002.08243*, 2020.

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., Rakhlin, A., Simchi-Levi, D., and Xu, Y. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. In *Conference on Learning Theory*, pp. 2059–2059. PMLR, 2021a.

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*, 2021b.

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.

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*, 2022a.

Levy, O. and Mansour, Y. Optimism in face of a context: Regret guarantees for stochastic contextual MDP. *arXiv preprint arXiv:2207.11126. (To appear in AAAI 2023)*, 2022b.

Levy, O., Cassel, A. B., Cohen, A., and Mansour, Y. Eluder-based regret for stochastic contextual mdps. *CoRR*, abs/2211.14932, 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.

Nesterov, Y. and Nemirovskii, A. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

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

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

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.

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

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. Proofs

In the following analysis, we represent the value function in terms of the *occupancy measures.* (See e.g., Puterman (2014); Zimin & Neu (2013)). The occupancy measures are defined as follows. 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. Thus, the value function of any policy π with respect to the MDP (S, A, P, r, s_0, H) can be presented as follows.

$$V_M^{\pi}(s_0) = \sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} q_h(s, a|\pi, P) \cdot r(s, a). \tag{10}$$

Throughout the analysis we consider the cumulative error caused by the dynamics approximation in terms of the Squared Hellinger distance between the true and approximated dynamics.

Definition (Squared Hellinger Distance, Definition 7). For any two distributions $\mathbb{P}$, $\mathbb{Q}$ over a discrete support X we define the Squared Hellinger Distance as

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

A useful property of the squared Hellinger distance is that for any two distributions $\mathbb{P}$ and $\mathbb{Q}$ it holds that $\|\mathbb{P} - \mathbb{Q}\|_1^2 \leq 4D_H^2(\mathbb{P}, \mathbb{Q})$.

Notations. In the analysis, we use the following notations, to denote the following occupancy measures. For all $(s, a, h) \in S \times A \times [H]$ we denote

- $\hat{q}_h^t(s, a) := q_h(s, a|\pi_t^{c_t}, \widehat{P}_t^{c_t})$,
- $q_h^t(s, a) := q_h(s, a|\pi_t^{c_t}, P_\star^{c_t})$,
- $\hat{q}_h^{t,\star}(s, a) := q_h(s, a|\pi_\star^{c_t}, \widehat{P}_t^{c_t})$.

A.1. Oracle Concentration Bounds

The following states our concentration bounds, in terms of our regression oracles regret.

A.1.1. Least Squares Regression Oracle for Rewards Approximation

In the following, we use Freedman's concentration inequality.

Lemma 16 (Freedman's inequality (see e.g., Agarwal et al., 2014; Cohen et al., 2021)). *Let $\{Z_t\}_{t \geq 1}$ be a real-valued martingale difference sequence adapted to a filtration $\{F_t\}_{t \geq 0}$ and let $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot|F_t]$. If $|Z_t| \leq R$ almost surely, then for any $T \in \mathbb{N}$ and $\eta \in (0, 1/R)$ it holds with probability at least $1 - \delta$ that,*

$$\sum_{t=1}^{T} Z_t \leq \eta \sum_{t=1}^{T} \mathbb{E}_{t-1}[Z_t^2] + \frac{\log(1/\delta)}{\eta}.$$

Lemma (concentration of LSR regret, restatement of Lemma 8). *Under Assumption 1 and Assumption 2, for any $\delta \in (0, 1)$, the following holds with probability at least $1 - \delta$.*

$$\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi_t^{c_t}, P_\star^{c_t}} \left[\sum_{h=0}^{H-1} \left(\hat{f}_t(c_t, s_h, a_h) - f_\star(c_t, s_h, a_h) \right)^2 \middle| s_0 \right] \leq 2 \cdot \mathcal{R}_{TH}(\mathcal{O}_{\text{sq}}^{\mathcal{F}}) + 16H \log(1/\delta).$$

Proof. Let us define a filtration $F_{t-1} = (\sigma^1, \ldots, \sigma^{t-1}, c_t)$. Then,

$$\begin{aligned} Z_t =& \mathbb{E}\left[\sum_{h=0}^{H-1} (\hat{f}_t(c_t, s_h^t, a_h^t) - r_h^t)^2 - (f_\star(c_t, s_h^t, a_h^t) - r_h^t)^2 \middle| F_{t-1} \right] \\ & - \sum_{h=0}^{H-1} (\hat{f}_t(c_t, s_h^t, a_h^t) - r_h^t)^2 - (f_\star(c_t, s_h^t, a_h^t) - r_h^t)^2 \end{aligned}$$

defines a martingale difference sequence for that filtration. We first prove the following auxiliary claim.

Claim 17. *The followings hold for all $t \in [T]$.*

1. $|Z_t| \leq 2H$.
2. $\mathbb{E}\left[\sum_{h=0}^{H-1}(\hat{f}_t(c_t, s_h^t, a_h^t) - r_h^t)^2 - (f_\star(c_t, s_h^t, a_h^t) - r_h^t)^2 | F_{t-1}\right] =$
   $\mathbb{E}\left[\sum_{h=0}^{H-1}(\hat{f}_t(c_t, s_h^t, a_h^t) - f_\star(c_t, s_h^t, a_h^t))^2 \middle| F_{t-1}\right] =$
   $\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} q_h^t(s,a) \cdot \left(\hat{f}_t(c_t, s, a) - f_\star(c_t, s, a)\right)^2$.
3. $\mathbb{E}[Z_t^2 | F_{t-1}] \leq 4H \cdot \mathbb{E}\left[\sum_{h=0}^{H-1}(\hat{f}_t(c_t, s_h^t, a_h^t) - f_\star(c_t, s_h^t, a_h^t))^2 \middle| F_{t-1}\right]$.

Proof. The first property is immediate. For the second property, we have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{h=0}^{H-1}(\hat{f}_t(c_t, s_h^t, a_h^t) - r_h^t)^2 - (f_\star(c_t, s_h^t, a_h^t) - r_h^t)^2 | F_{t-1}\right] \\
=&\mathbb{E}\left[\sum_{h=0}^{H-1}(\hat{f}_t(c_t, s_h^t, a_h^t) - f_\star(c_t, s_h^t, a_h^t))(\hat{f}_t(c_t, s_h^t, a_h^t) + f_\star(c_t, s_h^t, a_h^t) - 2r_h^t) | F_{t-1}\right] \\
=&\mathbb{E}\left[\sum_{h=0}^{H-1}(\hat{f}_t(c_t, s_h^t, a_h^t) - f_\star(c_t, s_h^t, a_h^t))(\hat{f}_t(c_t, s_h^t, a_h^t) + f_\star(c_t, s_h^t, a_h^t) - 2\mathbb{E}[r_h^t | c_t, s_h^t, a_h^t]) | F_{t-1}\right] \\
=&\mathbb{E}\left[\sum_{h=0}^{H-1}(\hat{f}_t(c_t, s_h^t, a_h^t) - f_\star(c_t, s_h^t, a_h^t))(\hat{f}_t(c_t, s_h^t, a_h^t) + f_\star(c_t, s_h^t, a_h^t) - 2f_\star(c_t, s_h^t, a_h^t)) | F_{t-1}\right] \\
=&\mathbb{E}\left[\sum_{h=0}^{H-1}(\hat{f}_t(c_t, s_h^t, a_h^t) - f_\star(c_t, s_h^t, a_h^t))^2 | F_{t-1}\right],
\end{aligned}
$$

where in the second and third equalities we used that $\mathbb{E}[r_h^t | c_t, s_h^t, a_h^t] = f_\star(c_t, s_h^t, a_h^t)$ and that $\hat{f}_t(c_t, s_h^t, a_h^t)$ and r_h^t are independent given s_h^t, a_h^t and the filtration F_{t-1}.

For the third property, consider the following derivation.

$$
\begin{aligned}
&\mathbb{E}[Z_t^2 | F_{t-1}] \\
=&\mathbb{E}\left[\left(\sum_{h=0}^{H-1}(\hat{f}_t(c_t, s_h^t, a_h^t) - r_h^t)^2 - (f_\star(c_t, s_h^t, a_h^t) - r_h^t)^2\right)^2 \middle| F_{t-1}\right] \\
&- \mathbb{E}^2\left[\sum_{h=0}^{H-1}(\hat{f}_t(c_t, s_h^t, a_h^t) - r_h^t)^2 - (f_\star(c_t, s_h^t, a_h^t) - r_h^t)^2 \middle| F_{t-1}\right] \\
\leq&\mathbb{E}\left[\left(\sum_{h=0}^{H-1}(\hat{f}_t(c_t, s_h^t, a_h^t) - r_h^t)^2 - (f_\star(c_t, s_h^t, a_h^t) - r_h^t)^2\right)^2 \middle| F_{t-1}\right] \\
\leq&H \cdot \mathbb{E}\left[\sum_{h=0}^{H-1}\left((\hat{f}_t(c_t, s_h^t, a_h^t) - r_h^t)^2 - (f_\star(c_t, s_h^t, a_h^t) - r_h^t)^2\right)^2 \middle| F_{t-1}\right] \\
=&H \cdot \mathbb{E}\left[\sum_{h=0}^{H-1}(\hat{f}_t(c_t, s_h^t, a_h^t) - f_\star(c_t, s_h^t, a_h^t))^2(\hat{f}_t(c_t, s_h^t, a_h^t) + f_\star(c_t, s_h^t, a_h^t) - 2r_h^t)^2 \middle| F_{t-1}\right] \\
\leq&4H \cdot \mathbb{E}\left[\sum_{h=0}^{H-1}(\hat{f}_t(c_t, s_h^t, a_h^t) - f_\star(c_t, s_h^t, a_h^t))^2 \middle| F_{t-1}\right].
\end{aligned}
$$

■

We now back to the proof of the lemma. By Lemma 16 and Claim 17, for $\eta \in (0, 1/2H)$ with probability at least $1-\delta$ it holds that

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left[\sum_{h=0}^{H-1} (\hat{f}_t(c_t, s_h^t, a_h^t) - r_h^t)^2 - (f_\star(c_t, s_h^t, a_h^t) - r_h^t)^2 \middle| F_{t-1}\right] \\
&- \sum_{h=0}^{H-1} (\hat{f}_t(c_t, s_h^t, a_h^t) - r_h^t)^2 - (f_\star(c_t, s_h^t, a_h^t) - r_h^t)^2 \\
=&\sum_{t=1}^{T} Z_t \\
\leq&\eta \sum_{t=1}^{T} \mathbb{E}_{t-1}[Z_t^2] + \frac{\log(1/\delta)}{\eta} && \text{(By Lemma 16)} \\
\leq&\eta \cdot 4H \cdot \sum_{t=1}^{T} \mathbb{E}\left[\sum_{h=0}^{h-1} (\hat{f}_t(c_t, s_h^t, a_h^t) - f_\star(c_t, s_h^t, a_h^t))^2 \middle| F_{t-1}\right] + \frac{\log(1/\delta)}{\eta}. && \text{(By Claim 17)}
\end{aligned}$$

The latter implies that

$$\begin{aligned}
&(1 - \eta \cdot 4 \cdot H) \cdot \sum_{t=1}^{T} \mathbb{E}\left[\sum_{h=0}^{H-1} (\hat{f}_t(c_t, s_h^t, a_h^t) - f_\star(c_t, s_h^t, a_h^t))^2 \middle| F_{t-1}\right] \\
&\leq \sum_{t=1}^{T} \sum_{h=0}^{H-1} (\hat{f}_t(c_t, s_h^t, a_h^t) - r_h^t)^2 - (f_\star(c_t, s_h^t, a_h^t) - r_h^t)^2 + \frac{\log(1/\delta)}{\eta}.
\end{aligned}$$

For $\eta = \frac{1}{8H} \in (0, 1/2H)$ we obtain

$$\begin{aligned}
&\frac{1}{2} \cdot \sum_{t=1}^{T} \mathbb{E}\left[\sum_{h=0}^{H-1} (\hat{f}_t(c_t, s_h^t, a_h^t) - f_\star(c_t, s_h^t, a_h^t))^2 \middle| F_{t-1}\right] \\
&\leq \sum_{t=1}^{T} \sum_{h=0}^{H-1} (\hat{f}_t(c_t, s_h^t, a_h^t) - r_h^t)^2 - (f_\star(c_t, s_h^t, a_h^t) - r_h^t)^2 + 8H \log(1/\delta).
\end{aligned}$$

Thus, when combine the above with part 2 of Claim 17 we obtain,

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} q_h^t(s, a) \cdot \left(\hat{f}_t(c_t, s, a) - f_\star(c_t, s, a)\right)^2 \\
&\leq 2 \cdot \sum_{t=1}^{T} \sum_{h=0}^{H-1} (\hat{f}_t(c_t, s_h^t, a_h^t) - r_h^t)^2 - (f_\star(c_t, s_h^t, a_h^t) - r_h^t)^2 + 16H \log(1/\delta).
\end{aligned} \tag{11}$$

By the oracle guarantees (Assumption 2),

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{h=0}^{H-1} (\hat{f}_t(c_t, s_h^t, a_h^t) - r_h^t)^2 - (f_\star(c_t, s_h^t, a_h^t) - r_h^t)^2 \\
&\leq \sum_{t=1}^{T} \sum_{h=0}^{H-1} (\hat{f}_t(c_t, s_h^t, a_h^t) - r_h^t)^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \sum_{h=0}^{H-1} (f(c_t, s_h^t, a_h^t) - r_h^t)^2 \\
&\leq \mathcal{R}_{TH}(\mathcal{O}_{\text{sq}}^{\mathcal{F}}).
\end{aligned} \tag{12}$$

By combining Equations (11) and (12) we obtain the lemma as,

$$\sum_{t=1}^{T} \sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} q_h^t(s, a) \cdot \left(\hat{f}_t(c_t, s, a) - f_\star(c_t, s, a)\right)^2 \leq 2 \cdot \mathcal{R}_{TH}(\mathcal{O}_{\text{sq}}^{\mathcal{F}}) + 16H \log(1/\delta),$$

which using the fact that $q_h^t(s,a) := q_h(s,a|\pi_t^{c_t}, P_\star^{c_t})$ implies

$$\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi_t^{c_t}, P_\star^{c_t}} \left[\sum_{h=0}^{H-1} \left(\hat{f}_t(c_t, s_h, a_h) - f_\star(c_t, s_h, a_h) \right)^2 \middle| s_0 \right] \leq 2 \cdot \mathcal{R}_{TH}(\mathcal{O}_{\text{sq}}^{\mathcal{F}}) + 16H \log(1/\delta).$$

■

A.1.2. Log-Loss Regression Oracle for Dynamics Approximation

To bound the oracle regret, we use the following lemma.

Lemma 18 (Lemma A.14 in Foster et al., 2021b). *Consider a sequence of $\{0,1\}$-valued random variables $(\mathbb{I}_t)_{t\leq T}$ where $\mathbb{I}_t$ is $F^{(t-1)}$-measurable. For any $\delta \in (0,1)$ we have that with probability at least $1-\delta$,*

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

Lemma (concentration of LLR regret w.r.t Hellinger distance, restatement of Lemma 9). *Under Assumption 3 and Assumption 4, for any $\delta \in (0,1)$, with probability at least $1-\delta$ it holds that*

$$\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi_t^{c_t}, P_\star^{c_t}} \left[\sum_{h=0}^{H-1} D_H^2(P_\star^{c_t}(\cdot|s_h,a_h), \widehat{P}_t^{c_t}(\cdot|s_h,a_h)) \middle| s_0 \right] \leq \mathcal{R}_{TH}(\mathcal{O}_{\log}^{\mathcal{P}}) + 2H\log(H/\delta).$$

Proof. Recall $q_h^t(s,a) := q_h(s,a|\pi_t^{c_t}, P_\star^{c_t})$.

$$\begin{aligned}
&\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi_t^{c_t}, P_\star^{c_t}} \left[\sum_{h=0}^{H-1} D_H^2(P_\star^{c_t}(\cdot|s_h,a_h), \widehat{P}_t^{c_t}(\cdot|s_h,a_h)) \middle| s_0 \right] \\
&= \sum_{t=1}^{T} \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} q_h^t(s,a) \cdot D_H^2(P_\star^{c_t}(\cdot|s,a), \widehat{P}_t^{c_t}(\cdot|s,a)) \\
&= \sum_{h=0}^{H-1} \sum_{t=1}^{T} \sum_{s\in S} \sum_{a\in A} q_h^t(s,a) \cdot D_H^2(P_\star^{c_t}(\cdot|s,a), \widehat{P}_t^{c_t}(\cdot|s,a)) \\
&\underbrace{=}_{(i)} \sum_{h=0}^{H-1} \sum_{t=1}^{T} \mathbb{E}\left[D_H^2(P_\star^{c_t}(\cdot|s_h,a_h), \widehat{P}_t^{c_t}(\cdot|s_h,a_h)) \middle| \mathbb{H}_{t-1}, c_t \right] \\
&\leq \sum_{t=1}^{T} \sum_{h=0}^{H-1} \log\left(\frac{1}{\widehat{P}^{c_t}(s_{h+1}^t|s_h^t,a_h^t)} \right) - \sum_{t=1}^{T} \sum_{h=0}^{H-1} \log\left(\frac{1}{P_\star^{c_t}(s_{h+1}^t|s_h^t,a_h^t)} \right) + 2H\log(H/\delta) \\
&\hfill \text{(By Lemma 18, holds w.p. at least } 1-\delta) \\
&\leq \sum_{t=1}^{T} \sum_{h=0}^{H-1} \log\left(\frac{1}{\widehat{P}^{c_t}(s_{h+1}^t|s_h^t,a_h^t)} \right) - \inf_{P\in\mathcal{P}} \left\{ \sum_{t=1}^{T} \sum_{h=0}^{H-1} \log\left(\frac{1}{P^{c_t}(s_{h+1}^t|s_h^t,a_h^t)} \right) \right\} + 2H\log(H/\delta) \quad \text{(By realizability)} \\
&\leq \mathcal{R}_{TH}(\mathcal{O}_{\log}^{\mathcal{P}}) + 2H\log(H/\delta).
\end{aligned}$$

The filtration used in (i) is over the history up to time t, $\mathbb{H}_{t-1} = (\sigma_1, \ldots, \sigma_{t-1})$ and the context in time t, c_t. ■

A.2. Regret Analysis

Recall the Hellinger distance given in Definition 7. The following change of measure result is due to (Foster et al., 2021b).

Lemma 19 (Lemma A.11 in Foster et al., 2021b). *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 20. *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

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

■

In the following regret analysis, we use the value change of measure with respect to the Hellinger distance, introduced by Levy et al. (2022).

Lemma 21 (Lemma 3 in Levy et al., 2022). *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

$$\begin{aligned}V_{\widehat{M},h}^{\pi}(s) &= \mathop{\mathbb{E}}_{a \sim \pi(\cdot|s)}\left[r(s, a) + \mathbb{E}_{s' \sim \widehat{P}(\cdot|s,a)}\left[V_{\widehat{M},h+1}^{\pi}(s')\right]\right] && \text{(By Bellman's equations)} \\ &\leq \mathop{\mathbb{E}}_{a \sim \pi(\cdot|s)}\left[r(s, a) + \left(1 + \frac{1}{H}\right)\mathbb{E}_{s' \sim P_\star(\cdot|s,a)}\left[V_{\widehat{M},h+1}^{\pi}(s')\right] + 3H^2 D_H^2(\widehat{P}(\cdot|s, a), P_\star(\cdot|s, a))\right] && \text{(Corollary 20)} \\ &\leq \mathop{\mathbb{E}}_{a \sim \pi(\cdot|s)}\left[r(s, a) + 3H^2 D_H^2(\widehat{P}(\cdot|s, a), P_\star(\cdot|s, a))\right] && \text{(Induction hypothesis)} \\ &+ \mathop{\mathbb{E}}_{a \sim \pi(\cdot|s)}\left[\left(1 + \frac{1}{H}\right)^{H-h} \mathop{\mathbb{E}}_{s' \sim P_\star(\cdot|s,a)}\left[V_{M,h+1}^{\pi}(s')\right]\right]\end{aligned}$$

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

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

This change of measure lemma upper bounds the value difference caused by the use of an approximated dynamics, instead of the true one, in terms of the expected Hellinger distance across a trajectory. This bound might seem very loose as a value difference bound, however, when the rewards are very small, it yields a significantly tighter then the standard bounds. In Lemmas 24 and 25 we apply the value change of measure lemma where the rewards function are the squared loss of the rewards approximation, and the squared L_1 loss of the dynamics approximation. As these rewards are very small, Lemma 3 implies that those expected approximation errors with respect to the approximated dynamics $\widehat{P}$ are at most a small constant multiple of these expected errors where the expectation is with respect to the true dynamics $P_\star$. Thus, the lemma helps us to translated the expected errors from the approximated measures to the true measures.

The following is an immediate corollary of Lemma 21 and Equation (10), which we will use in our analysis.

Corollary 22 (Occupancy measures change). *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*

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

In addition, we use the following version on the Value Difference Lemma introduced by Efroni et al. (2020).

Lemma 23 (Value-difference, Corollary 1; Efroni et al., 2020). *Let M, M' be any H-finite horizon MDPs. Then, for any two policies π, π' the following holds*

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

We are now have all the required tolls for the regret analysis.

A.2.1. Probability Measure Transitions

In the following, we present probability transition measures when applies to the cumulative approximation error for both the rewards and dynamics.

Lemma 24 (probabilities transition for rewards). *The following holds.*

$$\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\hat{q}_h^t(s,a)\cdot(\hat{f}_t(c_t,s,a)-f_\star(c_t,s,a))^2$$
$$\leq 3\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}(\hat{f}_t(c_t,s_h,a_h)-f_\star(c_t,s_h,a_h))^2\Big|s_0\right]$$
$$+9H^2\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}D_H^2(P_\star^{c_t}(\cdot|s_h,a_h),\widehat{P}_t^{c_t}(\cdot|s_h,a_h))\Big|s_0\right].$$

Proof. For any context $c\in\mathcal{C}$ and function $\hat{f}_t\in\mathcal{F}$ we have that $\tilde{r}^c(s,a):=(\hat{f}_t(c,s,a)-f_\star(c,s,a))^2$ is bounded in $[0,1]$. Recall $\hat{q}_h^t(s,a)=q_h(s,a|\pi_t^{c_t},\widehat{P}_t^{c_t})$ and $q_h^t(s,a):=q_h(s,a|\pi_t^{c_t},P_\star^{c_t})$. Hence, by Corollary 22, the following holds.

$$\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\hat{q}_h^t(s,a)\cdot(\hat{f}_t(c_t,s,a)-f_\star(c_t,s,a))^2\leq 3\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S_h}\sum_{a\in A}q_h^t(s,a)(\hat{f}_t(c_t,s,a)-f_\star(c_t,s,a))^2$$
$$+9H^2\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S_h}\sum_{a\in A}q_h^t(s,a)\cdot D_H^2(P_\star^{c_t}(\cdot|s_h,a_h),\widehat{P}_t^{c_t}(\cdot|s_h,a_h)).$$

Thus, the lemma follows. ■

Lemma 25 (probability transition for dynamics). *The following holds.*

$$\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\hat{q}_h^t(s,a)\cdot\left(\sum_{s'\in S}\left|\widehat{P}_t^{c_t}(s'|s,a)-P_\star^{c_t}(s'|s,a)\right|\right)^2$$
$$\leq 48H^2\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}D_H^2(P_\star^{c_t}(\cdot|s_h,a_h),\widehat{P}_t^{c_t}(\cdot|s_h,a_h))\Big|s_0\right].$$

where $\hat{q}_h^t(s,a):=q_h(s,a|\pi_t^{c_t},\widehat{P}_t^{c_t})$ *and* $q_h^t(s,a):=q_h(s,a|\pi_t^{c_t},P_\star^{c_t})$.

Proof. For any context $c\in\mathcal{C}$ and context-dependent dynamics $\widehat{P}_t\in\mathcal{P}$ we have that

$$\tilde{r}^c(s,a):=\left(\sum_{s'\in S}\left|\widehat{P}_t^{c}(s'|s,a)-P_\star^{c}(s'|s,a)\right|\right)^2$$

is bounded in $[0,4]$. Hence, by Corollary 22, the following holds.

$$\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\hat{q}_h^t(s,a)\cdot\left(\sum_{s'\in S}\left|\widehat{P}_t^{c_t}(s'|s,a)-P_\star^{c_t}(s'|s,a)\right|\right)^2$$
$$\leq 3\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S_h}\sum_{a\in A}q_h^t(s,a)\left(\sum_{s'\in S}\left|\widehat{P}_t^{c_t}(s'|s,a)-P_\star^{c_t}(s'|s,a)\right|\right)^2$$
$$+36H^2\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S_h}\sum_{a\in A}q_h^t(s,a)\cdot D_H^2(P_\star^{c_t}(\cdot|s_h,a_h),\widehat{P}_t^{c_t}(\cdot|s_h,a_h))$$
$$\leq 12\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S_h}\sum_{a\in A}q_h^t(s,a)D_H^2(P_\star^{c_t}(\cdot|s_h,a_h),\widehat{P}_t^{c_t}(\cdot|s_h,a_h)) \qquad (\|\cdot\|_1^2\leq 4D_H^2)$$
$$+36H^2\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S_h}\sum_{a\in A}q_h^t(s,a)\cdot D_H^2(P_\star^{c_t}(\cdot|s_h,a_h),\widehat{P}_t^{c_t}(\cdot|s_h,a_h))$$

$$\leq 48H^2 \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi_t^{c_t}, P_\star^{c_t}} \left[\sum_{h=0}^{H-1} D_H^2(P_\star^{c_t}(\cdot|s_h,a_h), \widehat{P}_t^{c_t}(\cdot|s_h,a_h)) \middle| s_0 \right].$$

Thus, the lemma follows. ■

A.2.2. Value Difference Bounds

In the following we derive three value difference bounds, which will be used to bound the regret.

Lemma 26. *The following holds for any $\hat{\gamma} > 0$.*

$$\begin{aligned}
\sum_{t=1}^{T} V_{\mathcal{M}(c_t)}^{\pi_\star^{c_t}}(s_0) - V_{\widehat{\mathcal{M}}_t(c_t)}^{\pi_\star^{c_t}}(s_0) \leq & \sum_{t=1}^{T} \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \frac{\hat{q}_h^{t,\star}(s,a)}{\hat{\gamma}\cdot \hat{q}_h^t(s,a)} \\
& + \frac{\hat{\gamma}}{2} \sum_{t=1}^{T} \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \hat{q}_h^t(s,a) \cdot \left(\hat{f}_t(c_t,s,a) - f_\star(c_t,s,a)\right)^2 \\
& + \frac{\hat{\gamma}\cdot H^2}{2} \sum_{t=1}^{T} \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \hat{q}_h^t(s,a) \cdot \left(\sum_{s'\in S} \left| \widehat{P}_t^{c_t}(s'|s,a) - P_\star^{c_t}(s'|s,a) \right| \right)^2,
\end{aligned}$$

where $\hat{q}_h^t(s,a) := q_h(s,a|\pi_t^{c_t}, \widehat{P}_t^{c_t})$ and $\hat{q}_h^{t,\star}(s,a) := q_h(s,a|\pi_\star^{c_t}, \widehat{P}_t^{c_t})$ is the occupancy measure defined by an optimal context-dependent policy of the true CMDP $\pi_\star = (\pi_\star^c)_{c\in\mathcal{C}}$. In addition, $\widehat{\mathcal{M}}_t(c) := (S, A, \widehat{P}_t^c, \hat{f}_t(c;\cdot,\cdot), s_0, H)$.

Proof. Consider the following derivation.

$$\begin{aligned}
& \sum_{t=1}^{T} V_{\mathcal{M}(c_t)}^{\pi_\star^{c_t}}(s_0) - V_{\widehat{\mathcal{M}}_t(c_t)}^{\pi_\star^{c_t}}(s_0) \\
= & \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi_\star^{c_t}, \widehat{P}_t^{c_t}} \left[\sum_{h=0}^{H-1} \left(\left(\hat{f}_t(c_t,s_h,a_h) - f_\star(c_t,s_h,a_h)\right) \right.\right. && \text{(Value Difference, Lemma 23)} \\
& \left.\left. + \sum_{s'\in S} \left(\widehat{P}_t^{c_t}(s'|s_h^t,a_h^t) - P_\star^{c_t}(s'|s_h^t,a_h^t)\right) \cdot V_{\mathcal{M}(c_t),h+1}^{\pi_\star^{c_t}}(s') \right) \middle| s_0 \right] \\
= & \sum_{t=1}^{T} \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \hat{q}_h^{t,\star}(s,a) \left(f_\star(c_t,s,a) - \hat{f}_t(c_t,s,a)\right) \\
& + \sum_{t=1}^{T} \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \hat{q}_h^{t,\star}(s,a) \sum_{s'\in S} \left(P_\star^{c_t}(s'|s,a) - \widehat{P}_t^{c_t}(s'|s,a)\right) V_{\mathcal{M}(c_t),h+1}^{\pi_\star^{c_t}}(s') \\
\leq & \sum_{t=1}^{T} \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \hat{q}_h^{t,\star}(s,a) \left(f_\star(c_t,s,a) - \hat{f}_t(c_t,s,a)\right) \\
& + H \cdot \sum_{t=1}^{T} \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \hat{q}_h^{t,\star}(s,a) \sum_{s'\in S} \left|P_\star^{c_t}(s'|s,a) - \widehat{P}_t^{c_t}(s'|s,a)\right| \\
= & \sum_{t=1}^{T} \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \hat{q}_h^{t,\star}(s,a) \cdot \sqrt{\frac{\hat{\gamma}\cdot\hat{q}_h^t(s,a)}{\hat{\gamma}\cdot\hat{q}_h^t(s,a)}} \left(f_\star(c_t,s,a) - \hat{f}_t(c_t,s,a)\right) \\
& + H \cdot \sum_{t=1}^{T} \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \hat{q}_h^{t,\star}(s,a) \cdot \sqrt{\frac{\hat{\gamma}\cdot\hat{q}_h^t(s,a)}{\hat{\gamma}\cdot\hat{q}_h^t(s,a)}} \sum_{s'\in S} \left|P_\star^{c_t}(s'|s,a) - \widehat{P}_t^{c_t}(s'|s,a)\right| \\
= & \sum_{t=1}^{T} \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \sqrt{\frac{\hat{q}_h^{t,\star}(s,a)}{\hat{\gamma}\cdot\hat{q}_h^t(s,a)}} \cdot \sqrt{\hat{q}_h^{t,\star}(s,a)\cdot\hat{\gamma}\cdot\hat{q}_h^t(s,a)} \left(f_\star(c_t,s,a) - \hat{f}_t(c_t,s,a)\right)
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\sqrt{\frac{\hat{q}_h^{t,\star}(s,a)}{\hat{\gamma}\cdot\hat{q}_h^t(s,a)}}\cdot\sqrt{\hat{q}_h^{t,\star}(s,a)\cdot\hat{\gamma}\cdot\hat{q}_h^t(s,a)}\cdot H\cdot\left(\sum_{s'\in S}\left|P_\star^{c_t}(s'|s,a)-\widehat{P}_t^{c_t}(s'|s,a)\right|\right)\\
\leq&\frac{1}{2}\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\left(\frac{\hat{q}_h^{t,\star}(s,a)}{\hat{\gamma}\cdot\hat{q}_h^t(s,a)}+\hat{q}_h^{t,\star}(s,a)\cdot\hat{\gamma}\cdot\hat{q}_h^t(s,a)\cdot\left(\hat{f}_t(c_t,s,a)-f_\star(c_t,s,a)\right)^2\right)\\
&\qquad\qquad\text{(Since for all } a,b\in\mathbb{R},\ ab\leq\tfrac{1}{2}(a^2+b^2) \text{ by AM-GM)}\\
&+\frac{1}{2}\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\left(\frac{\hat{q}_h^{t,\star}(s,a)}{\hat{\gamma}\cdot\hat{q}_h^t(s,a)}+\hat{q}_h^{t,\star}(s,a)\cdot\hat{\gamma}\cdot\hat{q}_h^t(s,a)\cdot H^2\cdot\left(\sum_{s'\in S}\left|P_\star^{c_t}(s'|s,a)-\widehat{P}_t^{c_t}(s'|s,a)\right|\right)^2\right)\\
\leq&\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\frac{\hat{q}_h^{t,\star}(s,a)}{\hat{\gamma}\cdot\hat{q}_h^t(s,a)}\\
&+\frac{\hat{\gamma}}{2}\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\hat{q}_h^t(s,a)\cdot\left(\hat{f}_t(c_t,s,a)-f_\star(c_t,s,a)\right)^2\\
&+\frac{\hat{\gamma}\cdot H^2}{2}\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\hat{q}_h^t(s,a)\cdot\left(\sum_{s'\in S}\left|\widehat{P}_t^{c_t}(s'|s,a)-P_\star^{c_t}(s'|s,a)\right|\right)^2,
\end{aligned}
$$

as stated. ■

Corollary 27 (restatement of Lemma 10). *The following holds for any $\hat{\gamma}>0$.*

$$
\begin{aligned}
\sum_{t=1}^{T}V_{\mathcal{M}(c_t)}^{\pi_\star^{c_t}}(s_0)-V_{\widehat{\mathcal{M}}_t(c_t)}^{\pi_\star^{c_t}}(s_0)\leq&\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\frac{\hat{q}_h^{t,\star}(s,a)}{\hat{\gamma}\cdot\hat{q}_h^t(s,a)}\\
&+2\hat{\gamma}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}(\hat{f}_t(c_t,s_h,a_h)-f_\star(c_t,s_h,a_h))^2\middle|s_0\right]\\
&+29\hat{\gamma}H^4\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}D_H^2(P_\star^{c_t}(\cdot|s_h,a_h),\widehat{P}_t^{c_t}(\cdot|s_h,a_h))\middle|s_0\right].
\end{aligned}
$$

Proof. Recall that $\hat{q}_h^t(s,a):=q_h(s,a|\pi_t^{c_t},\widehat{P}_t^{c_t})$ and $\hat{q}_h^{t,\star}(s,a):=q_h(s,a|\pi_\star^{c_t},\widehat{P}_t^{c_t})$. Consider the following derivation.

$$
\begin{aligned}
&\sum_{t=1}^{T}V_{\mathcal{M}(c_t)}^{\pi_\star^{c_t}}(s_0)-V_{\widehat{\mathcal{M}}_t(c_t)}^{\pi_\star^{c_t}}(s_0)\\
\leq&\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\frac{\hat{q}_h^{t,\star}(s,a)}{\hat{\gamma}\cdot\hat{q}_h^t(s,a)} && \text{(By Lemma 26)}\\
&+\frac{\hat{\gamma}}{2}\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\hat{q}_h^t(s,a)\cdot\left(\hat{f}_t(c_t,s,a)-f_\star(c_t,s,a)\right)^2\\
&+\frac{\hat{\gamma}\cdot H^2}{2}\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\hat{q}_h^t(s,a)\cdot\left(\sum_{s'\in S}\left|\widehat{P}_t^{c_t}(s'|s,a)-P_\star^{c_t}(s'|s,a)\right|\right)^2\\
\leq&\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\frac{\hat{q}_h^{t,\star}(s,a)}{\hat{\gamma}\cdot\hat{q}_h^t(s,a)}\\
&+\frac{3}{2}\hat{\gamma}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}(\hat{f}_t(c_t,s_h,a_h)-f_\star(c_t,s_h,a_h))^2\middle|s_0\right] && \text{(By Lemma 24)}\\
&+\frac{9}{2}\hat{\gamma}H^2\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}D_H^2(P_\star^{c_t}(\cdot|s_h,a_h),\widehat{P}_t^{c_t}(\cdot|s_h,a_h))\middle|s_0\right]
\end{aligned}
$$

$$+ \frac{48}{2}\hat{\gamma}H^4 \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi_t^{c_t}, P_\star^{c_t}} \left[\sum_{h=0}^{H-1} D_H^2(P_\star^{c_t}(\cdot|s_h,a_h), \widehat{P}_t^{c_t}(\cdot|s_h,a_h)) \middle| s_0 \right] \qquad \text{(By Lemma 25)}$$

$$\leq \sum_{t=1}^{T} \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \frac{\hat{q}_h^{t,\star}(s,a)}{\hat{\gamma}\cdot \hat{q}_h^t(s,a)}$$

$$+ 2\hat{\gamma} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi_t^{c_t}, P_\star^{c_t}} \left[\sum_{h=0}^{H-1} (\hat{f}_t(c_t,s_h,a_h) - f_\star(c_t,s_h,a_h))^2 \middle| s_0 \right]$$

$$+ 29\hat{\gamma}H^4 \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi_t^{c_t}, P_\star^{c_t}} \left[\sum_{h=0}^{H-1} D_H^2(P_\star^{c_t}(\cdot|s_h,a_h), \widehat{P}_t^{c_t}(\cdot|s_h,a_h)) \middle| s_0 \right].$$

■

Lemma 28 (restatement of Lemma 11). *For every round $t \in [T]$ and a context $c_t \in \mathcal{C}$, the optimal solution $\hat{q}^t$ for the maximization problem in Equation* (2) *satisfies the following,*

$$V^{\pi_\star^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) - V^{\pi_t^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) \leq \frac{H|S||A|}{\gamma} - \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \frac{\hat{q}_h^{t,\star}(s,a)}{\gamma \cdot \hat{q}_h^t(s,a)},$$

where $\hat{q}_h^t(s,a) := q_h(s,a|\pi_t^{c_t}, \widehat{P}_t^{c_t})$ and $\hat{q}_h^{t,\star}(s,a) := q_h(s,a|\pi_\star^{c_t}, \widehat{P}_t^{c_t})$ is the occupancy measure defined by an optimal context-dependent policy of the true CMDP $\pi_\star = (\pi_\star^c)_{c\in\mathcal{C}}$, recalling $\widehat{\mathcal{M}}_t(c) := (S, A, \widehat{P}_t^c, \hat{f}_t(c;\cdot,\cdot), s_0, H)$.

Proof. For every round $t \in [T]$ let $\hat{L}_t(q;c_t)$ denote the objective of the maximization problem in Equation (2) in round t, i.e.,

$$\hat{L}_t(q;c_t) = \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} q_h(s,a)\cdot \hat{f}_t(c_t,s,a) + \frac{1}{\gamma} \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \log(q_h(s,a)).$$

Thus the first-order derivation is

$$\hat{L}_t'(q;c_t) = \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \left(\hat{f}_t(c_t,s,a) + \frac{1}{\gamma\cdot q_h(s,a)} \right).$$

Let $\pi_\star = (\pi_\star^c)_{c\in\mathcal{C}}$ denote an optimal context-dependent policy for the true CMDP. For every round t, the occupancy measures $\hat{q}_h^{t,\star}(s,a) := q_h(s,a|\pi_\star^{c_t}, \widehat{P}_t^{c_t})$ is a feasible solution (since $\hat{q}^{t,\star} \in \mu(\widehat{P}_t^{c_t})$). Since $\hat{q}^t$ is the optimal solution, the following holds by first order optimality conditions.

$$\sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \left(\hat{f}_t(c_t,s,a) + \frac{1}{\gamma\cdot \hat{q}_h^t(s,a)} \right) \left(\hat{q}_h^{t,\star}(s,a) - \hat{q}_h^t(s,a) \right) \leq 0$$

$$\implies \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \hat{q}_h^{t,\star}(s,a)\cdot \left(\hat{f}_t(c_t,s,a) + \frac{1}{\gamma\cdot \hat{q}_h^t(s,a)} \right) - \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \hat{q}_h^t(s,a)\cdot \hat{f}_t(c_t,s,a) - \frac{H|S||A|}{\gamma} \leq 0$$

$$\implies \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \hat{q}_h^{t,\star}(s,a)\cdot \hat{f}_t(c_t,s,a) - \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \hat{q}_h^t(s,a)\cdot \hat{f}_t(c_t,s,a) \leq \frac{H|S||A|}{\gamma} - \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \frac{\hat{q}_h^{t,\star}(s,a)}{\gamma\cdot \hat{q}_h^t(s,a)}.$$

By definition we have for every round t,

$$V^{\pi_\star^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) - V^{\pi_t^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) = \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \hat{q}_h^{t,\star}(s,a)\cdot \hat{f}_t(c_t,s,a) - \sum_{h=0}^{H-1} \sum_{s\in S} \sum_{a\in A} \hat{q}_h^t(s,a)\cdot \hat{f}_t(c_t,s,a),$$

hence the lemma follows.

■

Lemma 29. *The following holds.*

$$\sum_{t=1}^{T} V^{\pi_t^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) - V^{\pi_t^{c_t}}_{\mathcal{M}(c_t)}(s_0) \leq \sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} q_h^t(s,a)\cdot(\hat{f}_t(c_t,s,a) - f_\star(c_t,s,a)) + H\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} q_h^t(s,a)\cdot\sum_{s'\in S}\left|\widehat{P}_t^{c_t}(s'|s,a) - P_\star^{c_t}(s'|s,a)\right|,$$

where $q_h^t(s,a) := q_h(s,a|\pi_t^{c_t}, P_\star^{c_t})$.

Proof. By the Value Difference Lemma, (Lemma 23), the following holds.

$$\begin{aligned}
&\sum_{t=1}^{T} V^{\pi_t^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) - V^{\pi_t^{c_t}}_{\mathcal{M}(c_t)}(s_0)\\
=&\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}\left(\hat{f}_t(c_t,s_h,a_h) - f_\star(c_t,s_h,a_h)\right) + \sum_{s'\in S}\left(\widehat{P}_t^{c_t}(s'|s_h^t,a_h^t) - P_\star^{c_t}(s'|s_h^t,a_h^t)\right)\cdot V^{\pi_t^{c_t}}_{\widehat{\mathcal{M}}_t(c_t),h+1}(s')\middle| s_0\right]\\
=&\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} q_h^t(s,a)\cdot(\hat{f}_t(c_t,s,a) - f_\star(c_t,s,a))\\
&+\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} q_h^t(s,a)\cdot\sum_{s'\in S}(\widehat{P}_t^{c_t}(s'|s,a) - P_\star^{c_t}(s'|s,a))\cdot V^{\pi_t^{c_t}}_{\widehat{\mathcal{M}}_t(c_t),h+1}(s')\\
\leq&\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} q_h^t(s,a)\cdot(\hat{f}_t(c_t,s,a) - f_\star(c_t,s,a))\\
&+H\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} q_h^t(s,a)\cdot\sum_{s'\in S}\left|\widehat{P}_t^{c_t}(s'|s,a) - P_\star^{c_t}(s'|s,a)\right|,
\end{aligned}$$

as stated. ■

Lemma 30. *The following holds for any parameter* $p_1 > 0$.

$$\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} q_h^t(s,a)(\hat{f}_t(c_t,s,a) - f_\star(c_t,s,a)) \leq \frac{TH}{2p_1} + \frac{p_1}{2}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}\left(\hat{f}_t(c_t,s_h,a_h) - f_\star(c_t,s_h,a_h)\right)^2\middle| s_0\right].$$

Proof. Consider the following derivation, where $q_h^t(s,a) := q_h(s,a|\pi_t^{c_t}, P_\star^{c_t})$.

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} q_h^t(s,a)\left(\hat{f}_t(c_t,s,a) - f_\star(c_t,s,a)\right)\\
=&\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\sqrt{\frac{q_h^t(s,a)}{p_1}}\cdot\sqrt{p_1\cdot q_h^t(s,a)}\left(\hat{f}_t(c_t,s,a) - f_\star(c_t,s,a)\right)\\
\leq&\frac{1}{2}\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\left(\frac{q_h^t(s,a)}{p_1} + p_1\cdot q_h^t(s,a)\left(\hat{f}_t(c_t,s,a) - f_\star(c_t,s,a)\right)^2\right)\\
&\qquad\text{(Since for all } a,b\in\mathbb{R},\ ab\leq\tfrac{1}{2}(a^2+b^2) \text{ by AM-GM)}\\
=&\frac{1}{2p_1}\sum_{t=1}^{T}\underbrace{\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} q_h^t(s,a)}_{\leq H} + \frac{p_1}{2}\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} q_h^t(s,a)\left(\hat{f}_t(c_t,s,a) - f_\star(c_t,s,a)\right)^2
\end{aligned}$$

$$\leq \frac{TH}{2p_1} + \frac{p_1}{2}\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} q_h^t(s,a)\left(\hat{f}_t(c_t,s,a) - f_\star(c_t,s,a)\right)^2$$
$$= \frac{TH}{2p_1} + \frac{p_1}{2}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}\left(\hat{f}_t(c_t,s_h,a_h) - f_\star(c_t,s_h,a_h)\right)^2 \middle| s_0\right].$$

■

Lemma 31. *The following holds for any parameter $p_2 > 0$.*

$$\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} q_h^t(s,a)\sum_{s'\in S}|\widehat{P}_t^{c_t}(s'|s,a) - P_\star^{c_t}(s'|s,a)| \leq \frac{TH}{2p_2} + 2p_2\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1} D_H^2(P_\star^{c_t}(\cdot|s_h,a_h),\widehat{P}_t^{c_t}(\cdot|s_h,a_h)) \middle| s_0\right].$$

Proof. Recall $q_h^t(s,a) := q_h(s,a|\pi_t^{c_t},P_\star^{c_t})$ and consider the following derivation.

$$\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} q_h^t(s,a)\sum_{s'\in S}\left|\widehat{P}_t^{c_t}(s'|s,a) - P_\star^{c_t}(s'|s,a)\right|$$
$$= \sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\sqrt{\frac{q_h^t(s,a)}{p_2}}\cdot\sqrt{p_2\cdot q_h^t(s,a)}\left(\sum_{s'\in S}\left|\widehat{P}_t^{c_t}(s'|s,a) - P_\star^{c_t}(s'|s,a)\right|\right)$$
$$\leq \frac{1}{2}\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\left(\frac{q_h^t(s,a)}{p_2} + p_2\cdot q_h^t(s,a)\left(\sum_{s'\in S}\left|\widehat{P}_t^{c_t}(s'|s,a) - P_\star^{c_t}(s'|s,a)\right|\right)^2\right)$$

(Since for all $a,b\in\mathbb{R}$, $ab \leq \frac{1}{2}(a^2+b^2)$ by AM-GM)

$$= \frac{1}{2p_2}\underbrace{\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} q_h^t(s,a)}_{\leq H} + \frac{p_2}{2}\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} q_h^t(s,a)\left(\sum_{s'\in S}\left|\widehat{P}_t^{c_t}(s'|s,a) - P_\star^{c_t}(s'|s,a)\right|\right)^2$$
$$\leq \frac{TH}{2p_2} + \frac{p_2}{2}\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} q_h^t(s,a)\left(\sum_{s'\in S}\left|\widehat{P}_t^{c_t}(s'|s,a) - P_\star^{c_t}(s'|s,a)\right|\right)^2$$
$$\leq \frac{TH}{2p_2} + 2p_2\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} q_h^t(s,a)\cdot D_H^2(\widehat{P}_t^{c_t}(\cdot|s,a),P_\star^{c_t}(\cdot|s,a)) \qquad (\|\cdot\|_1^2 \leq 4D_H^2)$$
$$= \frac{TH}{2p_2} + 2p_2\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1} D_H^2(P_\star^{c_t}(\cdot|s_h,a_h),\widehat{P}_t^{c_t}(\cdot|s_h,a_h)) \middle| s_0\right].$$

■

Corollary 32 (restatement of Lemma 12). *The following holds for any two parameters $p_1,p_2 > 0$.*

$$\sum_{t=1}^{T} V_{\widehat{\mathcal{M}}_t(c_t)}^{\pi_t^{c_t}}(s_0) - V_{\mathcal{M}(c_t)}^{\pi_t^{c_t}}(s_0) \leq \frac{TH}{2p_1} + \frac{p_1}{2}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}\left(\hat{f}_t(c_t,s_h,a_h) - f_\star(c_t,s_h,a_h)\right)^2 \middle| s_0\right] + \frac{TH}{2p_2} + 2p_2\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1} D_H^2(P_\star^{c_t}(\cdot|s_h,a_h),\widehat{P}_t^{c_t}(\cdot|s_h,a_h)) \middle| s_0\right].$$

Proof. Consider the following derivation.

$$\sum_{t=1}^{T} V_{\widehat{\mathcal{M}}_t(c_t)}^{\pi_t^{c_t}}(s_0) - V_{\mathcal{M}(c_t)}^{\pi_t^{c_t}}(s_0)$$

$$\leq \sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} q_h^t(s,a)\cdot(\hat{f}_t(c_t,s,a)-f_\star(c_t,s,a)) \qquad \text{(By Lemma 29)}$$

$$+ H\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} q_h^t(s,a)\cdot\sum_{s'\in S}\left|\widehat{P}_t^{c_t}(s'|s,a)-P_\star^{c_t}(s'|s,a)\right|$$

$$\leq \frac{TH}{2p_1}+\frac{p_1}{2}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}\left(\hat{f}_t(c_t,s_h,a_h)-f_\star(c_t,s_h,a_h)\right)^2\middle|s_0\right] \qquad \text{(By Lemma 30)}$$

$$+\frac{TH}{2p_2}+2p_2\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}D_H^2(P_\star^{c_t}(\cdot|s_h,a_h),\widehat{P}_t^{c_t}(\cdot|s_h,a_h))\middle|s_0\right]. \qquad \text{(By Lemma 31)}$$

■

A.2.3. Regret Bound

The following theorem states our main result, which is a regret bound for Algorithm 1.

Theorem (restatement of Theorem 5). *For any $\delta\in(0,1)$, let $\gamma=\sqrt{\frac{|S||A|T}{31H^3\left(2\mathcal{R}_{TH}(\mathcal{O}_{\text{sq}}^{\mathcal{F}})+\mathcal{R}_{TH}(\mathcal{O}_{\log}^{\mathcal{P}})+18H\log(2H/\delta)\right)}}$. Then, with probability at least $1-\delta$ it holds that*

$$\mathcal{R}_T(\textit{OMG-CMDP!})\leq\widetilde{O}\left(H^{2.5}\sqrt{T|S||A|\left(\mathcal{R}_{TH}(\mathcal{O}_{\text{sq}}^{\mathcal{F}})+\mathcal{R}_{TH}(\mathcal{O}_{\log}^{\mathcal{P}})+H\log\delta^{-1}\right)}\right).$$

Proof. By Lemma 8, with probability at least $1-\delta/2$, it holds that

$$\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}\left(\hat{f}_t(c_t,s_h,a_h)-f_\star(c_t,s_h,a_h)\right)^2\middle|s_0\right]\leq 2\mathcal{R}_{TH}(\mathcal{O}_{\text{sq}}^{\mathcal{F}})+16H\log(2/\delta).$$

By Lemma 9, with probability at least $1-\delta/2$, it holds that

$$\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}D_H^2(P_\star^{c_t}(\cdot|s_h,a_h),\widehat{P}_t^{c_t}(\cdot|s_h,a_h))\middle|s_0\right]\leq\mathcal{R}_{TH}(\mathcal{O}_{\log}^{\mathcal{P}})+2H\log(2H/\delta).$$

We prove a regret bound under those two good events.

$$\mathcal{R}_T(\text{OMG-CMDP!})$$

$$=\sum_{t=1}^{T}V_{\mathcal{M}(c_t)}^{\pi_\star^{c_t}}(s_0)-V_{\mathcal{M}(c_t)}^{\pi_t^{c_t}}(s_0)$$

$$=\sum_{t=1}^{T}V_{\mathcal{M}(c_t)}^{\pi_\star^{c_t}}(s_0)-V_{\widehat{\mathcal{M}}_t(c_t)}^{\pi_\star^{c_t}}(s_0)+\sum_{t=1}^{T}V_{\widehat{\mathcal{M}}_t(c_t)}^{\pi_\star^{c_t}}(s_0)-V_{\widehat{\mathcal{M}}_t(c_t)}^{\pi_t^{c_t}}(s_0)+\sum_{t=1}^{T}V_{\widehat{\mathcal{M}}_t(c_t)}^{\pi_t^{c_t}}(s_0)-V_{\mathcal{M}(c_t)}^{\pi_t^{c_t}}(s_0)$$

$$\leq\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\frac{\hat{q}_h^{t,\star}(s,a)}{\gamma\cdot\hat{q}_h^t(s,a)} \qquad \text{(By Corollary 27, for } \hat{\gamma}=\gamma.)$$

$$+2\gamma\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}(\hat{f}_t(c_t,s_h,a_h)-f_\star(c_t,s_h,a_h))^2\middle|s_0\right]$$

$$+29\gamma H^4\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}D_H^2(P_\star^{c_t}(\cdot|s_h,a_h),\widehat{P}_t^{c_t}(\cdot|s_h,a_h))\middle|s_0\right]$$

$$+\frac{H|S||A|T}{\gamma}-\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\frac{\hat{q}_h^{t,\star}(s,a)}{\gamma\cdot\hat{q}_h^t(s,a)} \qquad \text{(By Lemma 28, applied for each } t\in[T])$$

$$
\begin{aligned}
&+\frac{TH}{2p_1} && \text{(By Corollary 32)}\\
&+\frac{p_1}{2}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}\left(\hat{f}_t(c_t,s_h,a_h)-f_\star(c_t,s_h,a_h)\right)^2\middle|s_0\right]\\
&+\frac{TH}{2p_2}+2p_2\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}D_H^2(P_\star^{c_t}(\cdot|s_h,a_h),\widehat{P}_t^{c_t}(\cdot|s_h,a_h))\middle|s_0\right]\\
\le&2\gamma\left(2\cdot\mathcal{R}_{TH}(\mathcal{O}_{\text{sq}}^{\mathcal{F}})+16H\log(2/\delta)\right) && \text{(By the good events)}\\
&+29\gamma H^4\left(\mathcal{R}_{TH}(\mathcal{O}_{\log}^{\mathcal{P}})+2H\log(2H/\delta)\right)\\
&+\frac{H|S||A|T}{\gamma}\\
&+\frac{TH}{2p_1}+\frac{p_1}{2}\left(2\cdot\mathcal{R}_{TH}(\mathcal{O}_{\text{sq}}^{\mathcal{F}})+16H\log(2/\delta)\right)\\
&+\frac{TH}{2p_2}+2p_2\left(\mathcal{R}_{TH}(\mathcal{O}_{\log}^{\mathcal{P}})+2H\log(2H/\delta)\right)\\
\le&\gamma\cdot 31H^4\left(2\cdot\mathcal{R}_{TH}(\mathcal{O}_{\text{sq}}^{\mathcal{F}})+\mathcal{R}_{TH}(\mathcal{O}_{\log}^{\mathcal{P}})+18H\log(2H/\delta)\right)+\frac{H|S||A|T}{\gamma}\\
&+\frac{TH}{2p_1}+\frac{p_1}{2}\left(2\cdot\mathcal{R}_{TH}(\mathcal{O}_{\text{sq}}^{\mathcal{F}})+16H\log(2/\delta)\right)\\
&+\frac{TH}{2p_2}+2p_2\left(\mathcal{R}_{TH}(\mathcal{O}_{\log}^{\mathcal{P}})+2H\log(2H/\delta)\right)\\
=&2H^{2.5}\sqrt{31T|S||A|\left(2\cdot\mathcal{R}_{TH}(\mathcal{O}_{\text{sq}}^{\mathcal{F}})+\mathcal{R}_{TH}(\mathcal{O}_{\log}^{\mathcal{P}})+18H\log(2H/\delta)\right)}\\
&\qquad\qquad\text{(For }\gamma=\sqrt{\tfrac{|S||A|T}{31H^3\left(2\cdot\mathcal{R}_{TH}(\mathcal{O}_{\text{sq}}^{\mathcal{F}})+\mathcal{R}_{TH}(\mathcal{O}_{\log}^{\mathcal{P}})+18H\log(2H/\delta)\right)}}\text{)}\\
&+\sqrt{TH\left(2\cdot\mathcal{R}_{TH}(\mathcal{O}_{\text{sq}}^{\mathcal{F}})+16H\log(2/\delta)\right)} && \text{(For }p_1=\sqrt{\tfrac{TH}{2\mathcal{R}_{TH}(\mathcal{O}_{\text{sq}}^{\mathcal{F}})+16H\log(2/\delta)}}\text{)}\\
&+2\sqrt{TH\left(\mathcal{R}_{TH}(\mathcal{O}_{\log}^{\mathcal{P}})+2H\log(2H/\delta)\right)} && \text{(For }p_2=\sqrt{\tfrac{TH}{4\left(\mathcal{R}_{TH}(\mathcal{O}_{\log}^{\mathcal{P}})+2H\log(2H/\delta)\right)}}\text{)}\\
=&\widetilde{O}\left(H^{2.5}\sqrt{T|S||A|\left(\mathcal{R}_{TH}(\mathcal{O}_{\text{sq}}^{\mathcal{F}})+\mathcal{R}_{TH}(\mathcal{O}_{\log}^{\mathcal{P}})+H\log\delta^{-1}\right)}\right).
\end{aligned}
$$

Since the good events hold with probability at least $1-\delta$, so is the regret bound above. ■

B. Approximated Solution

The objective of optimization problem in Equation (2) is defined as a sum of a self-concordant barrier function (the log function) and a linear function. Hence, the optimal solution for the problem can be approximated efficiently using interior-point convex optimization algorithms such as Newton's Method. These algorithms return an ϵ-approximated solution and has running time of $O(poly(d)\log\epsilon^{-1})$, where d is the dimension of the problem (Nesterov & Nemirovskii, 1994).

Suppose that in each round t we derive the policy $\pi_t^{c_t}$ using, instead of the optimal solution, an occupancy measure $\hat{q}^t$ that yields an ϵ-approximation to the objective of the optimization problem in Equation (2). The following analysis shows that for $\epsilon=\frac{1}{16\gamma T}$, we obtain a similar regret guarantee. In addition, by our choice of γ, the running time complexity of the optimization algorithm is $\text{poly}(|S|,|A|,H,\log(T))$.

B.1. Regret Analysis

The following lemma bounds the difference between the optimal and the approximated iterates.

Lemma 33 (iterates difference, restatement of Lemma 13). *For every round $t \in [T]$ let*

$$\hat{L}_t(q; c_t) = \sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} q_h(s,a) \cdot \hat{f}_t(c_t, s, a) + \frac{1}{\gamma} \sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} \log(q_h(s,a)).$$

denote the objective of the optimization problem in Equation (2)*. Let $\widetilde{q} \in \arg\max_{q \in \mu(\widehat{P}_t^{c_t})} \hat{L}_t(q; c_t)$. Let $q \in \mu(\widehat{P}_t^{c_t})$ and suppose that $\hat{L}_t(\widetilde{q}; c_t) - \hat{L}_t(q; c_t) \leq \epsilon$. Then,*

$$\sum_{s \in S} \sum_{a \in A} \sum_{h=0}^{H-1} \left(\frac{q_h(s,a)}{\widetilde{q}_h(s,a)} - 1 \right)^2 \leq 4\epsilon\gamma.$$

Proof. Recall the Bregman divergence with respect to a θ-self-concordant barrier R as follows:

$$B_R(y||x) = R(y) - R(x) - \nabla R(x) \cdot (y - x).$$

We have the following lower bound on the Bregman divergence, where $\|x - y\|_x^2 = (x-y)^\top \nabla^2 R(x)(x-y)$.

$$B_R(y||x) \geq \rho\left(\|y - x\|_x\right) \text{ for } \rho(z) = z - \log(1+z).$$

We have that $\rho(z) \geq z^2/4$ for all $z \in [0,1]$, and $\|y - x\|_x \leq 1$. Thus,

$$B_R(y||x) \geq \rho\left(\|y - x\|_x\right) \geq \frac{1}{4}\|y - x\|_x^2.$$

Let $R(q) := -\frac{1}{\gamma} \sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} \log(q_h(s,a))$.
Recall that adding a linear function to the barrier function does not change the Bregman divergence. Since $\hat{L}$ is a result of adding a linear function to R we get that

$$B_R(q||\widetilde{q}) = B_{-\hat{L}}(q||\widetilde{q}) = -\hat{L}_t(q; c_t) + \hat{L}_t(\widetilde{q}; c_t) + \nabla \hat{L}_t(\widetilde{q}; c_t) \cdot (q - \widetilde{q}) \leq \epsilon + \nabla \hat{L}_t(\widetilde{q}; c_t) \cdot (q - \widetilde{q}) \leq \epsilon,$$

where the last step is by the first order optimality condition.
Furthermore, the Hessian is a diagonal matrix $\nabla^2 R$, where $\nabla^2 R(q)_{(s,a,h),(s,a,h)} = \frac{1}{\gamma q_h^2(s,a)}$. With us have that

$$\|q - \widetilde{q}\|_{\nabla^2 R(\widetilde{q})}^2 = \sum_{s \in S} \sum_{a \in A} \sum_{h=0}^{H-1} \frac{(q_h(s,a) - \widetilde{q}_h(s,a))^2}{\gamma \widetilde{q}_h^2(s,a)} = \frac{1}{\gamma} \sum_{s \in S} \sum_{a \in A} \sum_{h=0}^{H-1} \left(\frac{q_h(s,a)}{\widetilde{q}_h(s,a)} - 1 \right)^2.$$

we obtain that

$$\sum_{s \in S} \sum_{a \in A} \sum_{h=0}^{H-1} \left(\frac{q_h(s,a)}{\widetilde{q}_h(s,a)} - 1 \right)^2 \leq 4\epsilon\gamma. \quad \blacksquare$$

We use the above result to derive a bound on the value difference between $\pi_\star$ and π_t on the approximated model.

Lemma 34 (restatement of Corollary 14). *For every round $t \in [T]$ and a context $c_t \in \mathcal{C}$, let $\widetilde{q}^t$ be the optimal solution to the maximization problem in Equation* (2)*. Suppose that $\hat{q}_h^t \in \mu(\widehat{P}_t^{c_t})$ satisfies $\hat{L}_t(\hat{q}^t; c_t) - \hat{L}_t(\widetilde{q}^t; c_t) \leq \epsilon$, and $\epsilon\gamma \leq 1/16$. Then we have that*

$$V_{\widehat{\mathcal{M}}_t(c_t)}^{\pi_\star^{c_t}}(s_0) - V_{\widehat{\mathcal{M}}_t(c_t)}^{\pi_t^{c_t}}(s_0) \leq \frac{H|S||A|}{\gamma} - \sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} \frac{\hat{q}_h^{t,\star}(s,a)}{2\gamma \cdot \hat{q}_h^t(s,a)} + 2\sqrt{\epsilon\gamma H},$$

where $\hat{q}_h^t(s,a) := q_h(s,a|\pi_t^{c_t}, \widehat{P}_t^{c_t})$, and $\hat{q}_h^{t,\star}(s,a) := q_h(s,a|\pi_\star^{c_t}, \widehat{P}_t^{c_t})$, $\widehat{\mathcal{M}}_t(c) := (S, A, \widehat{P}_t^c, \hat{f}_t(c,\cdot,\cdot), s_0, H)$ is the estimated CMDP at round t, $\pi_t^{c_t}$ is the policy induced by $\widehat{q}^t$, and $\pi_\star = (\pi_\star^c)_{c \in \mathcal{C}}$ is the optimal context-dependent policy of the true CMDP.

Proof. For every round $t \in [T]$ let $\hat{L}_t(q; c_t)$ denote the objective of the maximization problem in Equation (2) in round t, i.e.,

$$\hat{L}_t(q; c_t) = \sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} q_h(s,a) \cdot \hat{f}_t(c_t, s, a) + \frac{1}{\gamma} \sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} \log(q_h(s,a)).$$

Thus the first-order derivation is

$$\hat{L}'_t(q; c_t) = \sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} \left(\hat{f}_t(c_t, s, a) + \frac{1}{\gamma \cdot q_h(s,a)} \right).$$

Let $\pi_\star = (\pi_\star^c)_{c \in \mathcal{C}}$ denote an optimal context-dependent policy for the true CMDP. For every round t, the occupancy measures $\hat{q}_h^{t,\star}(s,a) := q_h(s, a | \pi_\star^{c_t}, \widehat{P}_t^{c_t})$ is a feasible solution (since $\hat{q}^{t,\star} \in \mu(\widehat{P}_t^{c_t})$). Since $\widetilde{q}^t$ is the optimal solution, the following holds by first order optimality conditions.

$$\begin{aligned}
&\sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} \left(\hat{f}_t(c_t, s, a) + \frac{1}{\gamma \cdot \widetilde{q}_h^t(s,a)} \right) \left(\hat{q}_h^{t,\star}(s,a) - \widetilde{q}_h^t(s,a) \right) \leq 0 \\
\implies &\sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} \hat{q}_h^{t,\star}(s,a) \cdot \left(\hat{f}_t(c_t, s, a) + \frac{1}{\gamma \cdot \widetilde{q}_h^t(s,a)} \right) - \sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} \widetilde{q}_h^t(s,a) \cdot \hat{f}_t(c_t, s, a) - \frac{H|S||A|}{\gamma} \leq 0 \\
\implies &\sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} \hat{q}_h^{t,\star}(s,a) \cdot \hat{f}_t(c_t, s, a) - \sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} \widetilde{q}_h^t(s,a) \cdot \hat{f}_t(c_t, s, a) \leq \frac{H|S||A|}{\gamma} - \sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} \frac{\hat{q}_h^{t,\star}(s,a)}{\gamma \cdot \widetilde{q}_h^t(s,a)}.
\end{aligned}$$

By definition we have for every round t,

$$\begin{aligned}
V^{\pi_\star^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) - V^{\widetilde{\pi}_t^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) &= \sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} \hat{q}_h^{t,\star}(s,a) \cdot \hat{f}_t(c_t, s, a) - \sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} \widetilde{q}_h^t(s,a) \cdot \hat{f}_t(c_t, s, a) \\
&\leq \frac{H|S||A|}{\gamma} - \sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} \frac{\hat{q}_h^{t,\star}(s,a)}{\gamma \cdot \widetilde{q}_h^t(s,a)} \\
&\leq \frac{H|S||A|}{\gamma} - (1 - 2\sqrt{\epsilon\gamma}) \sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} \frac{\hat{q}_h^{t,\star}(s,a)}{\gamma \cdot \hat{q}_h^t(s,a)} && \text{(By Lemma 33)} \\
&\leq \frac{H|S||A|}{\gamma} - \frac{1}{2\gamma} \sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} \frac{\hat{q}_h^{t,\star}(s,a)}{\hat{q}_h^t(s,a)}. && \text{(For } 2\sqrt{\epsilon\gamma} \leq \tfrac{1}{2}\text{)}
\end{aligned}$$

In addition we have that

$$\begin{aligned}
V^{\widetilde{\pi}_t^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) - V^{\pi_t^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) &= \sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} (\widetilde{q}_h^t(s,a) - \hat{q}_h^t(s,a)) \cdot \hat{f}_t(c_t, s, a) \\
&\leq \sqrt{\sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} \frac{(\widetilde{q}_h^t(s,a) - \hat{q}_h^t(s,a))^2}{(\widetilde{q}_h^t(s,a))^2}} \sqrt{\sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} \widetilde{q}_h^t(s,a)^2 \hat{f}_t(c_t, s, a)^2} \\
&\qquad \text{(By Cauchy–Schwarz inequality)} \\
&\leq 2\sqrt{\epsilon\gamma} \sqrt{\sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} \widetilde{q}_h^t(s,a)^2 \hat{f}_t(c_t, s, a)^2} && \text{(Lemma 33)} \\
&\leq 2\sqrt{\epsilon\gamma} \sqrt{\sum_{h=0}^{H-1} \sum_{s \in S} \sum_{a \in A} \widetilde{q}_h^t(s,a) \hat{f}_t(c_t, s, a)^2} && (\widetilde{q}^2 \leq \widetilde{q}) \\
&\leq 2\sqrt{\epsilon\gamma H}.
\end{aligned}$$

Combining both bounds concludes the proof.

■

We are now ready to derive our regret bound, by combining the above lemma with our previous value bounds. The proof is almost identical to the proof of Theorem 5.

Theorem (restatement of Theorem 15). *For any $\delta \in (0,1)$ let $\gamma = \sqrt{\frac{|S||A|T}{62H^3\left(2\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{F}}_{\text{sq}})+\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{P}}_{\log})+18H\log(2H/\delta)\right)}}$. In addition, suppose that at each round t we have an ϵ-approximation to the optimal solution of Equation (2) for $\epsilon = \frac{1}{16\gamma T}$. Then, with probability at least $1-\delta$,*

$$\mathcal{R}_T(\textit{Approximated OMG-CMDP!}) \leq \widetilde{O}\left(H^{2.5}\sqrt{T|S||A|\left(\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{F}}_{\text{sq}}) + \mathcal{R}_{TH}(\mathcal{O}^{\mathcal{P}}_{\log}) + H\log\delta^{-1}\right)}\right).$$

Proof. By Lemma 8, with probability at least $1-\delta/2$, it holds that

$$\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}\left(\hat{f}_t(c_t,s_h,a_h) - f_\star(c_t,s_h,a_h)\right)^2\middle| s_0\right] \leq 2\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{F}}_{\text{sq}}) + 16H\log(2/\delta).$$

By Lemma 9, with probability at least $1-\delta/2$, it holds that

$$\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}D_H^2(P_\star^{c_t}(\cdot|s_h,a_h),\widehat{P}_t^{c_t}(\cdot|s_h,a_h))\middle| s_0\right] \leq \mathcal{R}_{TH}(\mathcal{O}^{\mathcal{P}}_{\log}) + 2H\log(2H/\delta).$$

We prove a regret bound under those two good events.

$$\begin{aligned}
&\mathcal{R}_T(\text{Approximated OMG-CMDP!})\\
=&\sum_{t=1}^{T}V^{\pi_\star^{c_t}}_{\mathcal{M}(c_t)}(s_0) - V^{\pi_t^{c_t}}_{\mathcal{M}(c_t)}(s_0)\\
=&\sum_{t=1}^{T}V^{\pi_\star^{c_t}}_{\mathcal{M}(c_t)}(s_0) - V^{\pi_\star^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) + \sum_{t=1}^{T}V^{\pi_\star^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) - V^{\pi_t^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) + \sum_{t=1}^{T}V^{\pi_t^{c_t}}_{\widehat{\mathcal{M}}_t(c_t)}(s_0) - V^{\pi_t^{c_t}}_{\mathcal{M}(c_t)}(s_0)\\
\leq&\sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\frac{\hat{q}_h^{t,\star}(s,a)}{2\gamma\cdot\hat{q}_h^t(s,a)} && \text{(By Corollary 27, for } \hat{\gamma}=2\gamma.)\\
&+4\gamma\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}(\hat{f}_t(c_t,s_h,a_h) - f_\star(c_t,s_h,a_h))^2\middle| s_0\right]\\
&+58\gamma H^4\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}D_H^2(P_\star^{c_t}(\cdot|s_h,a_h),\widehat{P}_t^{c_t}(\cdot|s_h,a_h))\middle| s_0\right]\\
&+\frac{TH|S||A|}{\gamma} - \sum_{t=1}^{T}\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}\frac{\hat{q}_h^{t,\star}(s,a)}{2\gamma\cdot\hat{q}_h^t(s,a)} + 2T\sqrt{\epsilon\gamma H} && \text{(By Lemma 34)}\\
&+\frac{TH}{2p_1} && \text{(By Corollary 32)}\\
&+\frac{p_1}{2}\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}\left(\hat{f}_t(c_t,s_h,a_h) - f_\star(c_t,s_h,a_h)\right)^2\middle| s_0\right]\\
&+\frac{TH}{2p_2} + 2p_2\sum_{t=1}^{T}\mathop{\mathbb{E}}_{\pi_t^{c_t},P_\star^{c_t}}\left[\sum_{h=0}^{H-1}D_H^2(P_\star^{c_t}(\cdot|s_h,a_h),\widehat{P}_t^{c_t}(\cdot|s_h,a_h))\middle| s_0\right]\\
\leq&4\gamma\left(2\cdot\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{F}}_{\text{sq}}) + 16H\log(2/\delta)\right) && \text{(By the good events)}
\end{aligned}$$

$$
\begin{aligned}
&+ 58\gamma H^4 \left(\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{P}}_{\text{log}}) + 2H\log(2H/\delta)\right)\\
&+ \frac{H|S||A|T}{\gamma} + 2T\sqrt{\epsilon\gamma H}\\
&+ \frac{TH}{2p_1} + \frac{p_1}{2}\left(2\cdot\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{F}}_{\text{sq}}) + 16H\log(2/\delta)\right)\\
&+ \frac{TH}{2p_2} + 2p_2\left(\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{P}}_{\text{log}}) + 2H\log(2H/\delta)\right)\\
\le&\gamma\cdot 62H^4\left(2\cdot\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{F}}_{\text{sq}}) + \mathcal{R}_{TH}(\mathcal{O}^{\mathcal{P}}_{\text{log}}) + 18H\log(2H/\delta)\right) + \frac{H|S||A|T}{\gamma}\\
&+ 2T\sqrt{\epsilon\gamma H}\\
&+ \frac{TH}{2p_1} + \frac{p_1}{2}\left(2\cdot\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{F}}_{\text{sq}}) + 16H\log(2/\delta)\right)\\
&+ \frac{TH}{2p_2} + 2p_2\left(\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{P}}_{\text{log}}) + 2H\log(2H/\delta)\right)\\
=&2H^{2.5}\sqrt{62T|S||A|\left(2\cdot\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{F}}_{\text{sq}}) + \mathcal{R}_{TH}(\mathcal{O}^{\mathcal{P}}_{\text{log}}) + 18H\log(2H/\delta)\right)}\\
&\qquad\qquad\text{(For } \gamma = \sqrt{\tfrac{|S||A|T}{62H^3\left(2\cdot\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{F}}_{\text{sq}}) + \mathcal{R}_{TH}(\mathcal{O}^{\mathcal{P}}_{\text{log}}) + 18H\log(2H/\delta)\right)}}\text{)}\\
&+ 2T^{1.25}\epsilon^{0.5}\left(\frac{|S||A|}{62H\left(2\cdot\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{F}}_{\text{sq}}) + \mathcal{R}_{TH}(\mathcal{O}^{\mathcal{P}}_{\text{log}}) + 18H\log(2H/\delta)\right)}\right)^{1/4}\\
&+ \sqrt{TH\left(2\cdot\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{F}}_{\text{sq}}) + 16H\log(2/\delta)\right)} \qquad \text{(For } p_1 = \sqrt{\tfrac{TH}{2\cdot\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{F}}_{\text{sq}}) + 16H\log(2/\delta)}}\text{)}\\
&+ 2\sqrt{TH\left(\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{P}}_{\text{log}}) + 2H\log(2H/\delta)\right)} \qquad \text{(For } p_2 = \sqrt{\tfrac{TH}{4\left(\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{P}}_{\text{log}}) + 2H\log(2H/\delta)\right)}}\text{)}\\
=&\widetilde{O}\left(H^{2.5}\sqrt{T|S||A|\left(\mathcal{R}_{TH}(\mathcal{O}^{\mathcal{F}}_{\text{sq}}) + \mathcal{R}_{TH}(\mathcal{O}^{\mathcal{P}}_{\text{log}}) + H\log\delta^{-1}\right)}\right). \qquad \text{(For } \epsilon = \tfrac{1}{16\gamma T}\text{)}
\end{aligned}
$$

Since the good events hold with probability at least $1-\delta$, so is the regret bound above. ■