Online Adversarial MDPs with Off-Policy Feedback and Known Transitions

Francesco Bacchiocchi[*]
Politecnico di Milano
francesco.bacchiocchi@polimi.it

Francesco Emanuele Stradi[*]
Politecnico di Milano
francescoemanuele.stradi@polimi.it

Matteo Papini
Universitat Pompeu Fabra
matteo.papini@upf.edu

Alberto Maria Metelli
Politecnico di Milano
albertomaria.metelli@polimi.it

Nicola Gatti
Politecnico di Milano
nicola.gatti@polimi.it

Abstract

In this paper, we face the challenge of online learning in *adversarial* Markov decision processes with known transitions and *off-policy* feedback. In this setting, the learner chooses a policy, but, differently from the traditional *on-policy* setting, the environment is explored by means of a different, fixed, and possibly unknown policy (named *colleague*'s policy), whose losses are revealed to the learner. The off-policy feedback presents an additional technical issue that is not present in traditional exploration-exploitation trade-off problems: the learner is charged with the regret of its chosen policy (w.r.t. a *comparator* policy) but it observes only the losses suffered by the colleague's policy. We first show that the state-of-the-art *optimistic* algorithms might suffer regret bounds which depend on the dissimilarity between the learner's policy and the colleague's one, which is guaranteed to be finite only under a uniform-coverage assumption of the colleague's policy. Contrariwise, we propose novel algorithms that, by employing *pessimistic* estimators—commonly adopted in the off-line reinforcement learning literature—ensure sublinear regret bounds depending on the more desirable dissimilarity between any comparator policy and the colleague's policy, even when the latter is unknown.

1 Introduction

Reinforcement learning (RL) has emerged as a powerful paradigm for training intelligent agents to make optimal decisions in complex and uncertain environments (Sutton and Barto, 2018). Within RL research, there has been a growing interest in online learning applied to *adversarial* Markov Decision Processes (MDPs, Even-Dar et al., 2009; Neu et al., 2010). This framework relaxes traditional stochastic and stationary assumptions to represent dynamic environments and address real-world decision-making scenarios that are constantly changing, misspecified, or corrupted. This is achieved by introducing an adversary that chooses the reward in a potentially arbitrary way, while the transitions are still stochastic.

One of the primary challenges in online RL lies in balancing exploration and exploitation (Sutton and Barto, 2018). Agents must *explore* the environment to discover new information and learn

[*]Equal Contribution.

16th European Workshop on Reinforcement Learning (EWRL 2023).

from it, while also *exploiting* the knowledge they have already acquired to make optimal decisions. Finding the right balance is crucial to ensure that the agent neither becomes overly conservative and fails to explore potentially rewarding options, nor fails to exploit actions it confidently knows to be good. To tackle this challenge, both in the stochastic and adversarial MDP setting, a large variety of algorithms have been developed leveraging several techniques, often inspired by the bandit literature (Lattimore and Szepesvári, 2020). To effectively navigate the exploration-exploitation trade-off, the great majority of algorithms rely on the principle of *optimism* in the face of uncertainty. Some examples are Upper Confidence Bound (UCB) algorithms for the stochastic-reward setting (e.g., Jaksch et al., 2010; Azar et al., 2017) and optimistic versions of mirror descent for the adversarial case (Jin et al., 2019). All these approaches consider *on-policy feedback*, meaning that the learner observes the trajectory and loss (or reward) generated by playing their own currently elected policy.

In this work, we consider a different form of feedback, that is *off-policy* feedback. In such a setting, the learner observes the trajectory and losses (or rewards) generated by playing a different policy, known in literature as *behavior policy*. We can think of the latter as being played by another agent that plays in parallel to our learning agent in the same environment, or against the same adversary. We will refer to this extra agent as the *colleague*. The behavior policy is fixed during the learning process and can be either known or unknown to the learner. In this setting, the learner faces a different challenge compared to the exploration-exploitation trade-off that characterizes the more common on-policy setting. Indeed, since the environment is explored by the fixed colleague's policy, the learner has no control over exploration. Hence, it should exploit available information as much as possible. At the same time, the learner should avoid over-exploitation of promising but under-explored decisions, that might lead it to risky or uncertain regions of the environment. Contrary to the on-policy setting, samples from these uncertain regions might never be collected by the colleague's policy (Levine et al., 2020). Intuitively, in such a scenario, optimistic approaches should be avoided, as they would precisely encourage exploration of the most uncertain regions of the environment, taking great risk without gaining any information in return.

Off-policy feedback has been widely investigated in the RL literature for the case of *stochastic* rewards, particularly in the setting of offline RL (Levine et al., 2020). In this setting, the learning agent has no direct access to the environment, only to a dataset of past interactions produced by an expert or by previous versions of the decision system itself, or a combination of different sources. For simplicity, it is common to consider the dataset as generated by a single, fixed behavior policy, corresponding to the notion of colleague considered here. Although classic offline RL algorithms like FQI (Ernst et al., 2005) are essentially pure exploitation, they are only guaranteed to efficiently find a near-optimal policy under strong assumptions on the behavior policy (Munos and Szepesvári, 2008). More recent algorithms are based on the principle of *pessimism* in the face of uncertainty, mirroring the on-policy principle in reverse. Intuitively, employing a pessimistic estimator keeps the agent away from regions that are too uncertain. Indeed, several pessimistic algorithms (Xiao et al., 2021; Rashidinejad et al., 2021; Jin et al., 2021; Zanette et al., 2021; Uehara and Sun, 2022; Cheng et al., 2022) have been proven to be efficient under significantly weaker assumptions on the behavior policy. However, the difficulty in establishing a meaningful notion of optimality for this setting (Xiao et al., 2021) leaves the debate open on whether or not pessimism is the ultimate approach to offline RL (a detailed survey of related work can be found in Appendix A).

Much less has been said on off-policy feedback in the adversarial setting, which is naturally online. In fact, off-policy learning with adversarial rewards has been first considered by Gabbianelli et al. (2022) for multi-armed-bandits (i.e., without states), and for linear contextual bandits with *i.i.d.* contexts (i.e., without dynamics). Their main motivation was theoretical: to study off-policy learning in a setting where the uncertainty due to the partial feedback is clearly decoupled from the inherent uncertainty of the environment, which takes the form of an arbitrary adversary. They showed how, in this setting, pessimism is crucial to achieve comparator-dependent regret bounds that scale with a notion of dissimilarity between the behavior policy and the comparator. However, they also hinted to potential *real-world applications*. Elaborating on their example, let us consider the context of big-tech companies, which consist of multiple semi-autonomous departments. Frequently, multiple departments are responsible for similar tasks, such as sales or procurement. In many cases, these departments make decisions autonomously while observing feedback related to larger, similar departments or, in some cases, feedback related to the broader macro-area they are assigned to. In such a scenario, the design of online algorithms able to achieve good learning performances while relying on parallel feedback is of paramount importance.

Given the theoretical appeal and the potential applications of off-policy adversarial learning, we believe it is of great interest to consider it in the context of dynamical systems. As a natural intermediate step between bandits and the full RL problem, we consider here MDPs with adversarial rewards, *known* stochastic transitions, and a potentially unknown behavior policy.

Original contributions. In this paper, we investigate the problem of online learning with *off-policy* feedback in adversarial Markov decision processes with known transitions. Precisely, we consider the case of episodic MDPs, where, at each episode $t \in [T]$, the agent plays a policy π_t over the horizon L and then observes the feedback generated by a colleague's policy π_C. We first show that state-of-the-art *optimistic* online learning algorithms for MDPs with adversarial losses might achieve a sublinear regret in T with constants that may be arbitrarily large. In particular, we show that the multiplicative factors in the regret bound of UOB-REPS (Jin et al., 2019) depend on the dissimilarity between the (occupancy measures of the) agent's learning dynamics π_t and colleague's policy π_C, even when π_C and the transition functions are known. Remarkably, we propose two *pessimistic* algorithms, namely P-REPS and P-REPS+, which do not suffer from such a weakness, with constants independent of the learning dynamics. Precisely, P-REPS works in the setting where the *colleague's policy is known*, and, by employing a pessimistic biased estimator, guarantees a sublinear regret with high probability which depends on the dissimilarity between the (occupancy measures of the) comparator policy π^* and the colleague's one π_C. Finally, we show that P-REPS+ achieves similar results even when the *colleague's policy is unknown*. This work answers the question raised by Gabbianelli et al. (2022), that is, whether it is possible to optimally learn in an online off-policy setting when the environment is a Markov decision process.

Paper structure. In Section 2, the problem formulation with the necessary notation is reported. In Section 3, the focus is on optimistic algorithms, precisely on a simplified version for known transitions of UOB-REPS (Jin et al., 2019), and its theoretical guarantees are provided. In Section 4, we show that employing pessimism leads to an improvement in the regret upper bound. In Section 5, we show that similar guarantees w.r.t Section 4 can be obtained when the colleague's policy is not known. Finally, in Section 6, conclusions and possible future works are reported.

2 Problem Formulation

In the following section, we present a comprehensive overview of the problem formulation, the underlying assumptions, and the performance measures employed in this work.

2.1 Adversarial Markov Decision Processes

An *adversarial episodic loop-free* Markov decision processes (MDPs) is a 4-tuple $M = (X, A, P, \{\ell_t\}_{t=1}^T)$ where:

- T is the number of episodes, with $t \in [T]$ indexing a specific episode. By the loop-free property, X is partitioned into L layers $X_0, \dots, X_L$ such that the first and the last layers are singletons, *i.e.*, $X_0 = \{x_0\}$ and $X_L = \{x_L\}$;
- $P : X \times A \times X \to [0, 1]$ is the transition function, where we denote by $P(x'|x, a)$ the probability of moving from state $x \in X$ to $x' \in X$ by taking action $a \in A$. By the loop-free property, it holds that $P(x'|x, a) > 0$ only if $x' \in X_{k+1}$ and $x \in X_k$ for some $k \in [0 \dots L-1]$;
- $\{\ell_t\}_{t=1}^T$ is a sequence of vectors describing the losses at each episode $t \in [T]$, namely $\ell_t \in [0, 1]^{|X \times A|}$. We refer to the loss of a specific state-action pair $x \in X, a \in A$ for a specific episode $t \in [T]$ as $\ell_t(x, a)$. Losses are chosen by an *adversary*, that is, no statistical assumptions are made.

Notice that any episodic MDP with horizon L that is *not* loop-free can be cast into a loop-free MDP by suitably duplicating the state space L times, *i.e.*, a state x is mapped to a set of new states (x, k), where $k \in [0 \dots L]$. The learner chooses a *policy* $\pi : X \times A \to [0, 1]$ at each episode, defining a probability distribution over actions at each state. For ease of notation, we denote by $\pi(\cdot|x)$ the action probability distribution for a state $x \in X$, with $\pi(a|x)$ denoting the probability of action $a \in A$.

We study an *off-policy* online setting, following Gabbianelli et al. (2022). In this setting, there is an external fixed policy π_C which is played in parallel to the learner. The feedback received by the

Algorithm 1 Learner-Environment Interaction

for $t \in [T]$ **do**
  The environment choses ℓ_t *adversarially*
  The learner chooses a policy $\pi_t : X \times A \to [0, 1]$
  The state is initialized to x_0
  for $k = 0, \dots, L-1$ **do**
    The colleague plays $a'_k \sim \pi_C(\cdot|x'_k)$ while the learner plays $a_k \sim \pi_t(\cdot|x_k)$
    The environment evolves to $x'_{k+1} \sim P(\cdot|x'_k, a'_k)$ for the colleague and to $x_{k+1} \sim P(\cdot|x_k, a_k)$ for the learner
  end for
  The learner observes $\{x'_k, a'_k\}_{k=0}^{L-1}$ and $\{\ell_t(x'_k, a'_k)\}_{k=0}^{L-1}$ but it suffers $\{\ell_t(x_k, a_k)\}_{k=0}^{L-1}$,
end for

learner at the end of each episode is the one pertaining to π_C. Algorithm 1 describes the interaction between the learner and environment in an off-policy adversarial MDP.

Remark 2.1. *In the second part of the paper we will deal with MDPs where the agent receives rewards in place of losses, namely $r_t(x, a) \ \ \forall t \in [T], x \in X, a \in A$. It is straightforward to check, taking $r_t(x, a) := 1 - \ell_t(x, a)$, that the two settings are equivalent.*

2.2 Occupancy Measures

We introduce the notion of *occupancy measure* (Zimin and Neu, 2013). Given a transition function P and a policy π, the occupancy measure $q^{P,\pi} \in [0, 1]^{X \times A}$ induced by P and π is such that, for every $x \in X_k$, $a \in A$, with $k \in [0 \dots L-1]$, it holds:

$$q^{P,\pi}(x, a) = \mathbb{P}[x_k = x, a_k = a|P, \pi], \qquad \text{and} \qquad q^{P,\pi}(x) = \sum_{a \in A} q^{P,\pi}(x, a). \tag{1}$$

It is straightforward to see that the occupancy measure $q^{P,\pi}$ of any policy π satisfies:

$$\sum_{a \in A} q^{P,\pi}(x, a) = \sum_{x' \in X_{k(x)-1}} \sum_{a' \in A} P(x|x', a') q^{P,\pi}(x', a'), \tag{2}$$

where $k(x)$ is the layer of state x (*i.e.*, $x \in X_k$). We denote by $\Delta(P)$ the space of valid occupancy measures induced by transition function P for any policy π, that are precisely those satisfying Equation (2). Note that any valid occupancy measure q induces a policy π^q defined as:

$$\pi^q(a|x) = \frac{q(x, a)}{q(x)}.$$

2.3 Cumulative Regret

We introduce the notion of cumulative regret over T rounds. We formally define the cumulative regret with respect to any fixed comparator policy π^* (and the associated occupancy q^{P,π^*}) as follows.

Definition 2.1 (Cumulative Regret)**.** *The cumulative regret over T rounds is defined as follows:*

$$R_T := \sum_{t=1}^{T} \ell_t^\top q^{P,\pi_t} - \sum_{t=1}^{T} \ell_t^\top q^{P,\pi^*}. \tag{3}$$

In traditional online learning settings, an algorithm presents good performance when its regret is sublinear in T, namely $R_T = o(T)$. In our setting, this property is not sufficient. Indeed, the regret necessarily depends on some dissimilarity measure between the comparator's policy π^* and the colleague's one π_C, and the larger the dissimilarity measure, the larger the regret (a formal definition of dissimilarity measure is provided in the following sections). In order for our algorithm to present a good performance, we need such a dissimilarity to be constant independently of the learning dynamics, thus only depending on π^* and π_C. For the sake of notation, we will refer to q^{P,π_t} using q_t, thus omitting the dependency on P and π, to q^{P,π^*} using q^*, and to q^{P,π_C} using q^{π_C}.

3 The Failure of Optimism vs. Uncertainty

In this section, we examine the limitations of optimism in our setting. Specifically, we investigate a simplified version, for MDPs with known transitions, of the *Upper Occupancy Bound Relative Entropy Policy Search* algorithm (UOB-REPS, Jin et al., 2019). This algorithm is commonly considered state-of-the-art in the standard online adversarial-MDP literature. Indeed, it guarantees the best regret upper bounds (in comparison to the existing literature, not in relation to the lower bound) in high probability. However, we show that, in our setting, this algorithm might fail to achieve sublinear regret bounds with constants only depending on the dissimilarity between π^* and π_C. Indeed, the regret scales with the dissimilarity between the learning dynamics π_t and π_C.

3.1 Algorithm

Algorithm 2 provides the pseudocode of UOB-REPS with known transitions. The initialization of the occupancy measure is uniform over the state-action space; thus, the policy π_1 is the one induced by the uniform occupancy measure (Line 1). We underline that a uniform occupancy measure does not ensure a uniform policy. Then, for every episode $t \in [T]$, the policy chosen by the algorithm is executed, while the feedback received by the learner, in terms of both losses and path, is the one collected by the policy π_C (Line 4). Once the losses are gathered, the algorithm builds an optimistic biased estimator as follows:

$$\widehat{\ell}_t(x,a) = \frac{\ell_t(x,a)}{q^{\pi_C}(x,a)+\gamma}\mathbb{1}\{x_{k(x)} = x, a_{k(x)} = a\} \qquad \forall (x,a) \in X \times A. \tag{4}$$

We underline the two main properties of the aforementioned estimator. First, the knowledge of both the colleague's policy π_C and the transition function P allows inferring q^{π_C}, namely the occupancy measure "played" by the colleague. Second, to have better regret guarantees, a constant factor $\gamma > 0$ is added to the denominator of the estimator, leading to an underestimate of the true loss (Line 6). Finally, Algorithm 2 updates the occupancy measure employing an Online Mirror Descent (OMD) update (Line 9) as follows:

$$q_{t+1} = \operatorname*{arg\,min}_{q \in \Delta(P)} \ \eta\langle q, \widehat{\ell}_t\rangle + D(q||q_t), \tag{5}$$

where $D(\cdot||\cdot)$ is the Bregman divergence, defined as the unnormalized KL-divergence. Precisely,

$$D(q||q') = \sum_{x,a} q(x,a)\ln\frac{q(x,a)}{q'(x,a)} - \sum_{x,a}(q(x,a) - q'(x,a)). \tag{6}$$

This update can be computed efficiently in the equivalent two-step version (see Jin et al. (2019)):

$$\widetilde{q}_{t+1}(x,a) = q_t(x,a)e^{-\eta\widehat{\ell}_t(x,a)}, \qquad q_{t+1} = \operatorname*{arg\,min}_{q \in \Delta(P)} D(q\|\widetilde{q}_{t+1}). \tag{7}$$

Precisely, the first update leading to $\widetilde{q}_{t+1}$ can be computed in closed form, while the projection is a convex problem with linear constraints that can be solved in polynomial time.

3.2 Regret Upper Bound

We state and discuss the main theoretical result concerning Algorithm 2. Precisely, UOB-REPS with off-policy feedback attains the following regret bound.

Theorem 3.1. *With probability at least $1-3\delta$, Algorithm 2 attains, for any valid comparator's occupancy measure $q^* \in \Delta(P)$, the regret bound:*

$$R_T \leq \gamma \sum_{t,x,a} \frac{q_t(x,a)}{q^{\pi_C}(x,a)} + \eta \sup_{t,x,a}\frac{q_t(x,a)}{q^{\pi_C}(x,a)}\left(|X||A|T + \frac{L^2}{2\gamma}\ln\left(\frac{L}{\delta}\right)\right) + \mathcal{O}\left(\frac{1}{\eta} + \frac{1}{\gamma} + \sqrt{T}\right).$$

In particular, setting $\eta = \gamma = \mathcal{O}\left(1/\sqrt{T}\right)$, we have:

$$R_T \leq \mathcal{O}\left(\sup_{t,x,a}\frac{q_t(x,a)}{q^{\pi_C}(x,a)}\sqrt{T}\right).$$

Algorithm 2 UOB-REPS with Known Transitions

Require: state space X, action space A, transition function P, episode number T, colleague's policy π_C

1: For all $k \in [0, \dots, L-1], x \in X_k, a \in A$, initialize the occupancy as

$$q_1(x,a) = \frac{1}{|X_k||A|}$$

and the policy as $\pi_1 = \pi^{q_1}$

2: **for** $t \in [T]$ **do**

3: Execute policy π_t for L steps and obtain the trajectory generated by π_C, namely (x_k, a_k) and

4: losses $\ell_t(x_k, a_k)$ for $k \in [0, \dots, L-1]$

5: **for** $(x,a) \in X \times A$ **do**

6:

$$\widehat{\ell}_t(x,a) = \frac{\ell_t(x,a)}{q^{\pi_C}(x,a) + \gamma} \mathbb{1}\{x_{k(x)} = x, a_{k(x)} = a\}$$

7: **end for**

8: Update occupancy measure:

9:

$$q_{t+1} = \operatorname*{arg\,min}_{q \in \Delta(P)} \ \eta \langle q, \widehat{\ell}_t \rangle + D(q||q_t)$$

10: Update policy $\pi_{t+1} = \pi^{q_{t+1}}$

11: **end for**

The above result is intuitive and self-explanatory. Algorithm 2 attains sublinear regret even with off-policy feedback. Nevertheless, the factor $\sup_{t,x,a} \frac{q_t(x,a)}{q^{\pi_C}(x,a)}$ could potentially amplify the regret arbitrarily, as the learning dynamics q_t may be unpredictable. Such a result leads to the need for developing new tools which take into account the impossibility of receiving on-policy feedback during the learning process.

4 Pessimistic Algorithm with Known Policy

As previously observed, the problem of optimistic estimators is that they are designed to manage the *exploration-exploitation* trade-off. Precisely, the learning dynamics are driven by the need for exploring new areas of the decision space in order to gather as much information as possible. Contrariwise, when the feedback is related to an independent policy, the incentive to explore must be limited, since, in principle, more exploration does not result in a better understanding of the optimization problem.

To overcome this issue, we employ pessimistic estimators of the rewards. Indeed, in a standard online learning framework, pessimism would result in a sub-optimal solution to the exploration-exploitation dilemma. Nevertheless, it does help in off-policy scenarios, limiting the exploration of the learning agent and guaranteeing regret bounds which depend on the constant factor

$$\mathcal{D}(\pi^*, \pi_C) \coloneqq \sum_{x,a} \frac{q^*(x,a)}{q^{\pi_C}(x,a)},$$

for any comparator policy π^*. For the sake of simplicity, in the formulation of pessimistic estimators used from here on, we will deal with MDPs where the learner receives rewards in place of losses. As previously argued, this change does not affect the generality of the results.

4.1 Algorithm

Algorithm 3 provides the pseudocode of our *Pessimistic Relative Entropy Search* (P-REPS). More specifically, Algorithm 3 is a pessimistic variant of UOB-REPS (Jin et al., 2019) with known transitions (Algorithm 2). In the following, we describe the algorithm and remark the main differences compared to the original, optimistic version by Jin et al. (2019).

The occupancy measure is initially set to be uniform over the state-action space and the policy π_1 is the one induced by q^{π_1} (Line 1). We remark that, in the case of *off-policy* feedback, during each episode $t \in [T]$, we only receive the rewards and observe the trajectory of the colleague's policy π_C

Algorithm 3 Pessimistic Relative Entropy Policy Search (P-REPS)

Require: state space X, action space A, transition function P, episode number T, colleague's policy π_C

1: For all $k \in [0, \dots, L-1], x \in X_k, a \in A$, initialize occupancy

$$q_1(x,a) = \frac{1}{|X_k||A|}$$

and $\pi_1 = \pi^{q_1}$.

2: **for** $t \in [T]$ **do**

3: Execute policy π_t for L steps and obtain trajectory based on π_C, namely (x_k, a_k) for

4: $k \in [0..L-1]$ and rewards $r_t(x_k, a_k)$

5: **for** $(x,a) \in X \times A$ **do**

6:

$$\widehat{\ell}_t(x,a) = 1 - \frac{r_t(x,a)}{q^{\pi_C}(x,a) + \gamma} \mathbb{1}\{x_{k(x)} = x, a_{k(x)} = a\}$$

7: **end for**

8: Update occupancy measure:

$$\widetilde{q}_{t+1}(x,a) = \frac{q_t(x,a) e^{-\eta \widehat{\ell}_t(x,a)}}{\sum_{x' \in X_{k(x)}, a' \in A} q_t(x',a') e^{-\eta \widehat{\ell}_t(x',a')}}$$

$$q_{t+1} = \arg\min_{q \in \Delta(P)} D(q \| \widetilde{q}_{t+1})$$

9: Update policy $\pi_{t+1} = \pi^{q_{t+1}}$

10: **end for**

(Line 4), while our own policy is executed. Once the rewards are gathered, the algorithm builds a pessimistic estimator as follows:

$$\widehat{r}_t(x,a) = \frac{r_t(x,a)}{q^{\pi_C}(x,a) + \gamma} \mathbb{1}\{x_{k(x)} = x, a_{k(x)} = a\} \qquad \forall (x,a) \in X \times A. \tag{8}$$

Similarly to Algorithm 2, we add a constant factor $\gamma > 0$ in the denominator of the resulting biased estimator. We remark that, since we employ the rewards in place of losses, this choice leads to an underestimate of the reward received. Moreover, for the sake of coherence with respect to the online learning literature, we turn this reward estimate into a loss one (Line 6). Namely, we define:

$$\widehat{\ell}_t(x,a) = 1 - \widehat{r}_t(x,a) \qquad \forall (x,a) \in X \times A. \tag{9}$$

Differently from UOB-REPS, Algorithm 3 updates the occupancy measure by employing a normalized version of OMD (Line 8) as follows:

$$\widetilde{q}_{t+1}(x,a) = \frac{q_t(x,a) e^{-\eta \widehat{\ell}_t(x,a)}}{\sum_{x' \in X_{k(x)}, a' \in A} q_t(x',a') e^{-\eta \widehat{\ell}_t(x',a')}}, \quad q_{t+1} = \arg\min_{q \in \Delta(P)} D(q \| \widetilde{q}_{t+1}). \tag{10}$$

The peculiarity of this update lies in its first unconstrained step. Specifically, the standard unconstrained optimization update $q_t(x,a) e^{-\eta \widehat{\ell}_t(x,a)}$ is normalized over the state-action space. This technical adjustment is necessary to partially bridge the gap between lazy updates, as discussed by Neu (2015) and Gabbianelli et al. (2022), where the decision point is optimized independently from the projection, and greedy updates, which are more common in the online adversarial MDPs literature.

Finally, we remark that the computational complexity of the projection step is the same as in Algorithm 2. In particular, the projection is again a convex optimization problem with linear constraints, which can be solved in polynomial time. Therefore, the projection step in Algorithm 3 can be efficiently computed.

4.2 Regret Upper Bound

In this section, we state the theoretical guarantees provided by Algorithm 3 in terms of regret upper bounds. Precisely, P-REPS attains the following regret bound when the feedback is off-policy.

Algorithm 4 Pessimistic Relative Entropy Policy Search with unknown colleague policy (P-REPS+)

Require: state space X, action space A, transition function P, number of episodes T.

For all (x, a) initialize counters $N(x, a) = 0$, for all $k \in [0, \dots, L-1], x \in X_k, a \in A$, initialize the occupancy

$$q_1(x, a) = \frac{1}{|X_k||A|}$$

and initialize the policy $\pi_1 = \pi^{q_1}$

for $t \in [T]$ **do**

Execute policy π_t for L steps and obtain the trajectory generated by π_C, namely (x_k, a_k) and rewards $r_t(x_k, a_k)$ for $k \in [0..L-1]$

for $k \in [0..L-1]$ **do**

Update counters:

$$N(x_k, a_k) \leftarrow N(x_k, a_k) + 1$$

end for

for $(x, a) \in X \times A$ **do**

$$\widehat{q}_t^{\,\pi_C}(x, a) \leftarrow \frac{N(x, a)}{t}$$

$$\widehat{\ell}_t(x, a) = 1 - \frac{r_t(x, a)}{\widehat{q}_t^{\,\pi_C}(x, a) + \gamma_t} \mathbb{1}\{x_{k(x)} = x, a_{k(x)} = a\}$$

end for

Update occupancy measure:

$$\widetilde{q}_{t+1}(x, a) = \frac{q_t(x, a) e^{-\eta \widehat{\ell}_t(x,a)}}{\sum_{x' \in X_{k(x)}, a' \in A} q_t(x', a') e^{-\eta \widehat{\ell}_t(x', a')}}$$

$$q_{t+1} = \arg\min_{q \in \Delta(P)} D(q \| \widetilde{q}_{t+1})$$

Update policy $\pi_{t+1} = \pi^{q_{t+1}}$

end for

Theorem 4.1. *With probability at least $1 - 2\delta$, Algorithm 3 with $\gamma = \eta/2$ attains, for any valid comparator occupancy measure $q^* \in \Delta(P)$, the regret bound:*

$$R_T \leq \frac{L \ln(|X||A|)}{\eta} + \mathcal{D}(\pi^*, \pi_C) \left(\sqrt{2T \ln\left(\frac{|X||A|}{\delta}\right)} + \gamma T \right) + L \sqrt{2T \ln\left(\frac{1}{\delta}\right)}.$$

In particular, setting $\eta = 2\gamma = \mathcal{O}\left(1/\sqrt{T}\right)$,

$$R_T \leq \mathcal{O}\left(\mathcal{D}(\pi^*, \pi_C) \sqrt{T}\right).$$

We compare this regret bound with that attained by Algorithm 2. The main difference is the constant factor which affects the regret bound. More precisely, differently from the regret bound of Algorithm 2, the multiplicative factor in the regret bound of Algorithm 3 is independent of the learning dynamics of the learner. Furthermore, the constant factor $\mathcal{D}(\pi^*, \pi_C)$ implies that, the closer the colleague's policy is to the one of the comparator, the better the bound. When the colleague's policy is equivalent to the comparator, namely when $q^* = q^{\pi_C}$, the regret bound is almost equivalent to the well-known ones for the standard on-policy (bandit) feedback.

5 Pessimistic Algorithm with Unknown Policy

We now investigate off-policy feedback in settings where the colleague's policy is *not* known, and we show that, with a slight modification to Algorithm 3, we achieve similar regret guarantees. To address the uncertainty arising from the unknown policy of the colleague, we employ a time-varying reward estimator, similarly to what Gabbianelli et al. (2022) did for multi-armed bandits. This allows us to

introduce an additional level of pessimism in the estimates compared to that used in Algorithm 3. This extra pessimism helps us to deal with the uncertainty introduced by the unknown policy and achieve the desired regret guarantees.

5.1 Algorithm

Algorithm 4 provides the pseudocode of our *Pessimistic Relative Entropy Search with an unknown colleague's policy* (P-REPS+). The initialization and the interaction with the environment strictly follow the one of Algorithm 3 (Line 1- 4), except that a counter for each state-action pair (x, a) is initialized as $N(x, a) = 0$. Once the rewards are collected, the algorithm builds a pessimistic estimator (Line 9) as:

$$\widehat{r}_t(x,a) = \frac{r_t(x,a)}{\widehat{q}_t^{\pi_C}(x,a) + \gamma_t} \mathbb{1}\{x_{k(x)} = x, a_{k(x)} = a\} \qquad \forall (x,a) \in X \times A, \tag{11}$$

where $\widehat{q}_t^{\pi_C}(x,a)$ is the empirical mean of the occupancy measure for every state-action pair (x, a). We observe that, in Algorithm 4, the pessimistic factor γ_t is time-dependent. This is because, when the colleague's policy is *not* known, the pessimistic factor γ_t should incorporate the uncertainty related to the empirical estimate of the occupancy measure $\widehat{q}_t^{\pi_C}$. Specifically, using Hoeffding's inequality, it can be shown that, with a failure probability of $\delta_t \in [0, 1]$, it holds that:

$$|\widehat{q}_t^{\pi_C}(x,a) - q^{\pi_C}(x,a)| \leq \epsilon_t \coloneqq \sqrt{\frac{\ln\left(|X||A|/\delta_t\right)}{2t}} \quad \forall (x,a) \in X \times A, \tag{12}$$

with $t \geq 1$. Thus, the time-dependent pessism factor is set as $\gamma_t = \epsilon_t + \gamma$, where the γ is the same as in Algorithm 3. The intuition behind this choice stems from the idea of introducing additional pessimism in the biased estimator. This is done to address the uncertainty that arises from the empirical mean estimation of the occupancy measure $\widehat{q}_t^{\pi_C}$. To be coherent with the online learning literature, we turn the rewards estimates into losses (Line 9) as usual:

$$\widehat{\ell}_t(x,a) = 1 - \widehat{r}_t(x,a) \;\; \forall (x,a) \in X \times A.$$

Then, for each state-action pair (x, a) along the path traversed by the colleague, the counters are updated accordingly (Line 6). Finally, Algorithm 4 updates the occupancy measure employing a normalized version of OMD as done in Algorithm 3 (Line 11).

5.2 Regret Upper Bound

P-REPS attains the following regret bound when the learner has off-policy feedback and the colleague policy is *not* known.

Theorem 5.1. *With probability at least* $1 - 3\delta$*, Algorithm 4 for* $\gamma_t = \epsilon_t + \gamma = \epsilon_t + \eta/2$ *with* $\epsilon_t = \sqrt{\frac{\ln(|X||A|T/\delta)}{2t}}$ *(*i.e.*,* $\delta_t = \delta/T$*) attains, for any valid comparator occupancy measure* $q^* \in \Delta(P)$*:*

$$R_T \leq \frac{L \ln(|X||A|)}{\eta} + L\sqrt{2T \ln\left(\frac{1}{\delta}\right)} + \mathcal{D}\left(\pi^*, \pi_C\right)\left(4\sqrt{2T \ln\left(\frac{1}{\delta}\right)} + \ln\left(\frac{T|X||A|}{\delta}\right)\sqrt{T} + \gamma T\right)$$

In particular, setting $\eta = 2\gamma = \mathcal{O}\left(1/\sqrt{T}\right)$*,*

$$R_T \leq \tilde{\mathcal{O}}\left(\mathcal{D}\left(\pi^*, \pi_C\right)\sqrt{T}\right).$$

Theorem 5.1 shows that the dependency on the constant factor $\mathcal{D}\left(\pi^*, \pi_C\right)$ is still achievable when the colleague's policy is not known beforehand, by paying an additional $\mathcal{O}(\ln T)$ factor to deal with the uncertainty in the estimation of q^{π_C}.

6 Conclusions and Future Works

In this paper, we study the problem of online learning with *off-policy* feedback in adversarial Markov decision processes with known transitions. We first show that state-of-the-art *optimistic* algorithms

might achieve sublinear regret which depends on the maximum dissimilarity between the occupancy measure of the policies chosen during the learning process and the one of the colleague. Then, we propose two *pessimistic* algorithms. P-REPS works in the setting where the *colleague's policy is known* and achieves sublinear regret which depends on the dissimilarity between the comparator occupancy measure and the one of the colleague, while P-REPS+ guarantees similar results when the *colleague's policy is unknown*, employing an estimator with additional pessimistic bias.

In the future, we aim to extend our results to encompass settings with unknown transition functions, similarly to what was done by Jin et al. (2019) for on-policy feedback. Specifically, we should investigate whether a pessimistic estimator is sufficient to achieve comparator-dependent regret bounds or if additional techniques for dealing with uncertain transitions must be incorporated.

References

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL `https://proceedings.neurips.cc/paper/2008/file/e4a6222cdb5b34375400904f03d8e6a5-Paper.pdf`.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning, 2017.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 3852–3878. PMLR, 2022.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *J. Mach. Learn. Res.*, 6:503–556, 2005.

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Germano Gabbianelli, Matteo Papini, and Gergely Neu. Online learning with off-policy feedback, 2022.

Elad Hazan. Introduction to online convex optimization. *CoRR*, abs/1909.05207, 2019. URL `http://arxiv.org/abs/1909.05207`.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. volume 11, pages 1563–1600, 2010.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial mdps with bandit feedback and unknown transition, 2019. URL `https://arxiv.org/abs/1912.01192`.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5084–5096. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/jin21e.html`.

Michail G. Lagoudakis and Ronald Parr. Least-squares policy iteration. *J. Mach. Learn. Res.*, 4: 1107–1149, 2003.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems, 2020.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *J. Mach. Learn. Res.*, 9:815–857, 2008.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits, 2015.

Gergely Neu, Andras Antos, András György, and Csaba Szepesvári. Online markov decision processes under bandit feedback. *Advances in Neural Information Processing Systems*, 23, 2010.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism, 2021.

Aviv Rosenberg and Yishay Mansour. Online stochastic shortest path with bandit feedback and unknown transition function. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a. URL https://proceedings.neurips.cc/paper/2019/file/a0872cc5b5ca4cc25076f3d868e1bdf8-Paper.pdf.

Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial Markov decision processes. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5478–5486. PMLR, 09–15 Jun 2019b. URL https://proceedings.mlr.press/v97/rosenberg19a.html.

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

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *ICLR*. OpenReview.net, 2022.

Chenjun Xiao, Yifan Wu, Jincheng Mei, Bo Dai, Tor Lattimore, Lihong Li, Csaba Szepesvari, and Dale Schuurmans. On the optimality of batch policy optimization algorithms. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11362–11371. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/xiao21b.html.

Andrea Zanette, Martin J. Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. In *NeurIPS*, pages 13626–13640, 2021.

Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf.

A Related Work

Off-Policy Reinforcement Learning. Off-policy feedback has been largely studied in the offline, or "batch" RL literature (Lagoudakis and Parr, 2003; Ernst et al., 2005). In such a setting, the learner cannot interact with the environment and has instead only access to a fixed dataset collected by a behavior policy (Levine et al., 2020). Recently, the pessimistic approach has gathered a lot of interest in this area, especially on the theoretical side (Xiao et al., 2021; Rashidinejad et al., 2021; Jin et al., 2021; Zanette et al., 2021; Uehara and Sun, 2022; Cheng et al., 2022). Precisely, pessimistic offline RL methods avoid the strong requirement of the behavior policy covering the whole space of reachable states and actions, which is often unfeasible in practice, and only require coverage of optimal decisions. This leads to regret bounds which depend on the *partial* coverage with respect to the optimal (or a different comparator) policy (Rashidinejad et al., 2021), rather than the *uniform* coverage over policy space that is required, for instance, by FQI (Munos and Szepesvári, 2008). The minimax sample complexity rate for this problem is $\mathcal{O}(\epsilon^{-2})$, corresponding to $\mathcal{O}(\sqrt{T})$ regret. However, the meaningfullness of minimax optimality in this setting is debated, since greedy and even optimistic algorithms, besides pessimistic ones, have been shown to attain it. At the same time, instance-dependent optimality as defined in the online setting is not attainable in the offline setting. See Xiao et al. (2021) for an extensive discussion. The same authors have proposed a "weighted" notion of minimax optimality that justifies the use of pessimism. However, comparator-dependent or "partial coverage" bounds remain the main theoretical appeal of pessimistic algorithms.

Online Learning in MDPs. The body of research focusing on online learning problems (Cesa-Bianchi and Lugosi, 2006; Hazan, 2019) in MDPs is extensive, as investigated in several notable works (Auer et al., 2008; Even-Dar et al., 2009; Neu et al., 2010). Azar et al. (2017) study the challenge of optimal exploration in episodic MDPs where transitions are unknown and losses are stochastic, and only bandit (partial, on-policy) feedback is available. Their algorithm matches the $\Omega(\sqrt{L|X||A|T})$ lower bound for this setting (Jaksch et al., 2010), where T represents the number of episodes, L the episode length, $|X|$ the number of states, and $|A|$ the number of actions. Instead, Rosenberg and Mansour (2019b) consider the online learning problem in episodic MDPs with adversarial losses and unknown transitions, but with full-information feedback. They propose an online-learning algorithm with a regret upper bound of $\tilde{\mathcal{O}}(L|X|\sqrt{|A|T})$. The same scenario is explored by Rosenberg and Mansour (2019a), albeit with the more challenging bandit feedback, leading to a $\tilde{\mathcal{O}}(T^{3/4})$ regret upper bound, which was subsequently improved to $\tilde{\mathcal{O}}(L|X|\sqrt{|A|T})$ by Jin et al. (2019). Matching the $\Omega(L\sqrt{|X||A|T})$ lower bound for this setting is still an open problem. The most similar setting to the one considered in this paper is the one from Zimin and Neu (2013): adversarial losses, known transitions and (on-policy) bandit feedback. Their algorithm matches a $\Omega(\sqrt{L|X||A|T})$ lower bound up to logarithmic factors.

Online Learning with Off-Policy Feedback. Off-policy settings are quite novel in the online (adversarial) learning literature. To the best of our knowledge, the main existing contribution is by Gabbianelli et al. (2022), who investigate the setting where the learner observes the rewards sampled following a behavior policy in multi-armed bandit and linear contextual bandit problems. Off-policy feedback in adversarial MDPs is uncharted territory.

B Optimistic Algorithm

In this section, we report the omitted proofs related to the optimistic algorithm. We first show how the regret can be decomposed. Then, we proceed bounding each term.

B.1 Regret Decomposition

The regret term can be easily decomposed, similarly to Jin et al. (2019), as follows:

$$\begin{aligned} R_T &= \sum_{t=1}^{T} \ell_t^\top q_t - \sum_{t=1}^{T} \ell_t^\top q^* \\ &= \underbrace{\sum_{t=1}^{T} \langle q_t, \ell_t - \widehat{\ell}_t \rangle}_{①} + \underbrace{\sum_{t=1}^{T} \langle q_t - q^*, \widehat{\ell}_t \rangle}_{②} + \underbrace{\sum_{t=1}^{T} \langle q^*, \widehat{\ell}_t - \ell_t \rangle}_{③}. \end{aligned}$$

We underline that the first and the last terms are related to the distance between the optimistic estimator and the real loss vector, while the second one depends on the learning dynamic of Online Mirror Descent.

Bound on ①. We start bounding the first term of the regret bound.

Lemma B.1. *With probability at least $1-\delta$, Algorithm 2 attains:*

$$\sum_{t=1}^{T} \langle q_t, \ell_t - \widehat{\ell}_t \rangle \leq \gamma \sum_{t,x,a} \frac{q_t(x,a)}{q^{\pi_C}(x,a)} + L\sqrt{2T \ln \frac{1}{\delta}}.$$

Proof. We decompose ① as follows:

$$\begin{aligned} ① &= \sum_{t=1}^{T} \langle q_t, \ell_t - \widehat{\ell}_t \rangle \\ &= \sum_{t=1}^{T} \langle q_t, \ell_t - \mathbb{E}[\widehat{\ell}_t] \rangle + \sum_{t=1}^{T} \langle q_t, \mathbb{E}[\widehat{\ell}_t] - \widehat{\ell}_t \rangle \\ &\leq \sum_{t=1}^{T} \langle q_t, \ell_t - \mathbb{E}[\widehat{\ell}_t] \rangle + L\sqrt{2T \ln \frac{1}{\delta}}, \end{aligned}$$

where the last step inequality holds with probability at least $1-\delta$ due to Azuma-Hoeffding.

We then focus on the right term, and we rewrite it as:

$$\begin{aligned} \sum_{t=1}^{T} \langle q_t, \ell_t - \mathbb{E}[\widehat{\ell}_t] \rangle &= \sum_{t,x,a} q_t(x,a)\ell_t(x,a) \left(1 - \frac{\mathbb{E}_t\left[\mathbb{1}\left\{x_{k(x)} = x, a_{k(x)} = a\right\}\right]}{q^{\pi_C}(x,a) + \gamma}\right) \\ &= \sum_{t,x,a} q_t(x,a)\ell_t(x,a) \left(1 - \frac{q^{\pi_C}}{q^{\pi_C}(x,a) + \gamma}\right) \\ &= \sum_{t,x,a} \frac{q_t(x,a)\ell_t(x,a)}{q^{\pi_C}(x,a) + \gamma} \left(q^{\pi_C}(x,a) + \gamma - q^{\pi_C}(x,a)\right) \\ &= \sum_{t,x,a} \frac{q_t(x,a)\ell_t(x,a)}{q^{\pi_C}(x,a) + \gamma} \gamma \\ &\leq \gamma \sum_{t,x,a} \frac{q_t(x,a)}{q^{\pi_C}(x,a)}. \end{aligned}$$

Putting everything together we obtain, with probability at least $1-\delta$:

$$
\begin{aligned}
① &= \sum_{t=1}^{T} \langle q_t, \ell_t - \widehat{\ell}_t \rangle \\
&\leq \gamma \sum_{t,x,a} \frac{q_t(x,a)}{q^{\pi_C}(x,a)} + L\sqrt{2T \ln \frac{1}{\delta}},
\end{aligned}
$$

which concludes the proof. □

Bound on ②. We proceed bounding the second term of the regret. We start stating a useful lemma which directly follows from Lemma 11 (Jin et al., 2019).

Lemma B.2. *For any sequence of functions $\alpha_1, \dots \alpha_T$ such that α_t is $\mathcal{F}_t-$ measurable for all $t \in [T]$, then with probability of at least $1-\delta$ we have:*

$$
\sum_{t=1}^{T} \sum_{x,a} \alpha_t(x,a) \left(\widehat{\ell}_t(x,a) - \ell_t(x,a) \right) \leq L \ln \left(\frac{L}{\delta} \right).
$$

The previous result is intuitive. Since the estimator employed by Algorithm 2 is optimistic, the sum over T of the difference between the estimator and real loss vector can be bounded by a constant factor, independent from T.

Now we are ready to prove the bound of the second term of the regret. The proof follows the standard one for Online Mirror Descent with entropic regularizer and then focus on bounding the biased estimator.

Lemma B.3. *With probability at least $1-\delta$, Algorithm 2 attains, for any valid comparator occupancy measure $q^* \in \Delta(P)$:*

$$
\sum_{t=1}^{T} \langle q_t - q^*, \widehat{\ell}_t \rangle \leq \frac{L \ln(|X||A|)}{\eta} + \eta \sup_{t,x,a} \frac{q_t(x,a)}{q^{\pi_C}(x,a)} \left(|X||A|T + \frac{L^2}{2\gamma} \ln \left(\frac{L}{\delta} \right) \right).
$$

Proof. We focus on bounding,

$$
\sum_{t=1}^{T} \langle q_t - q^*, \widehat{\ell}_t \rangle.
$$

We first apply Lemma 13 (Jin et al., 2019) with the uniform initialization over the state-action pairs (x,a), and we obtain,

$$
\sum_{t=1}^{T} \langle q_t - q^*, \widehat{\ell}_t \rangle \leq \frac{L \ln(|X||A|)}{\eta} + \eta \sum_{t,x,a} q_t(x,a) \widehat{\ell}_t(x,a)^2.
$$

We focus on the second term:

$$
\begin{aligned}
\eta \sum_{t,x,a} q_t(x,a) \widehat{\ell}_t(x,a)^2 &\leq \eta \sup_{t,x,a} \frac{q_t(x,a)}{q^{\pi_C}(x,a)} \sum_{t,x,a} q^{\pi_C}(x,a) \frac{\ell_t(x,a)}{q^{\pi_C}(x,a) + \gamma} \widehat{\ell}_t(x,a) \\
&\leq \eta \sup_{t,x,a} \frac{q_t(x,a)}{q^{\pi_C}(x,a)} \sum_{t,x,a} \widehat{\ell}_t(x,a) \\
&\leq \eta \sup_{t,x,a} \frac{q_t(x,a)}{q^{\pi_C}(x,a)} \left(\sum_{t,x,a} \ell_t(x,a) + \frac{L}{2\gamma} \ln \left(\frac{L}{\delta} \right) \right) \\
&\leq \eta \sup_{t,x,a} \frac{q_t(x,a)}{q^{\pi_C}(x,a)} \left(|X||A|T + \frac{L}{2\gamma} \ln \left(\frac{L}{\delta} \right) \right),
\end{aligned}
$$

where the third inequality holds for Lemma B.2, with probability at least $1-\delta$.

Thus, the final result is the following, with probability at least $1-\delta$:

$$\sum_{t=1}^{T}\langle q_t - q^*, \widehat{\ell}_t\rangle \leq \frac{L\ln(|X||A|)}{\eta} + \eta \sup_{t,x,a}\frac{q_t(x,a)}{q^{\pi_C}(x,a)}\left(|X||A|T + \frac{L^2}{2\gamma}\ln\left(\frac{L}{\delta}\right)\right).$$

□

Bound on ③. We are now ready to bound the last term of the regret decomposition.

Lemma B.4. *With probability at least $1-\delta$, Algorithm 2 attains, for any valid comparator occupancy measure $q^* \in \Delta(P)$:*

$$\sum_{t=1}^{T}\langle q^*, \widehat{\ell}_t - \ell_t\rangle \leq \frac{L}{2\gamma}\ln\left(\frac{L|X||A|}{\delta}\right).$$

Proof. Applying Lemma B.2 with $\alpha_t = 2\gamma\mathbb{1}(x,a)$ and a union bound over actions and states we get

$$\sum_{t=1}^{T}\sum_{x,a}\widehat{\ell}_t(x,a) \leq \sum_{t=1}^{T}\sum_{x,a}\ell_t(x,a) + \frac{1}{2\gamma}L\ln\left(\frac{L|X||A|}{\delta}\right),$$

with probability of at least $1-\delta$, for each state-action pair (x,a). Then, by means of the latter inequality we get:

$$③ = \sum_{t=1}^{T}\langle q^*, \widehat{\ell}_t - \ell_t\rangle \leq \frac{L}{2\gamma}\ln\left(\frac{L|X||A|}{\delta}\right),$$

which concludes the proof. □

B.2 Regret Bound

Once bounded each component resulted from the regret decomposition, the final result on the regret bound follows immediately. Indeed,

Theorem 3.1. *With probability at least $1-3\delta$, Algorithm 2 attains, for any valid comparator's occupancy measure $q^* \in \Delta(P)$, the regret bound:*

$$R_T \leq \gamma\sum_{t,x,a}\frac{q_t(x,a)}{q^{\pi_C}(x,a)} + \eta\sup_{t,x,a}\frac{q_t(x,a)}{q^{\pi_C}(x,a)}\left(|X||A|T + \frac{L^2}{2\gamma}\ln\left(\frac{L}{\delta}\right)\right) + \mathcal{O}\left(\frac{1}{\eta} + \frac{1}{\gamma} + \sqrt{T}\right).$$

In particular, setting $\eta = \gamma = \mathcal{O}\left(1/\sqrt{T}\right)$, we have:

$$R_T \leq \mathcal{O}\left(\sup_{t,x,a}\frac{q_t(x,a)}{q^{\pi_C}(x,a)}\sqrt{T}\right).$$

Proof. The result immediately follows from the regret decomposition and Lemmas B.1, B.3 and B.4 □

C Pessimistic Algorithm with Known Policy

In this section, we report the omitted proof related to the pessimistic algorithm, when the colleague's policy is known beforehand.

Theorem 4.1. *With probability at least $1-2\delta$, Algorithm 3 with $\gamma = \eta/2$ attains, for any valid comparator occupancy measure $q^* \in \Delta(P)$, the regret bound:*

$$R_T \leq \frac{L\ln(|X||A|)}{\eta} + \mathcal{D}(\pi^*, \pi_C)\left(\sqrt{2T\ln\left(\frac{|X||A|}{\delta}\right)} + \gamma T\right) + L\sqrt{2T\ln\left(\frac{1}{\delta}\right)}.$$

In particular, setting $\eta = 2\gamma = \mathcal{O}\left(1/\sqrt{T}\right)$,

$$R_T \leq \mathcal{O}\left(\mathcal{D}\left(\pi^*, \pi_C\right)\sqrt{T}\right).$$

Proof. Let $\gamma > 0$, we define $\widehat{r}_t(x,a) := \frac{r_t(x,a)}{q^{\pi_C}(x,a)+\gamma}\mathbb{1}\{x_{k(x)} = x, a_{k(x)} = a\}$, $\widetilde{r}_t(x,a) := \frac{r_t(x,a)}{q^{\pi_C}(x,a)}\mathbb{1}\{x_{k(x)} = x, a_{k(x)} = a\}$ and $\widetilde{\ell}_t(x,a) := 1 - \widetilde{r}_t(x,a)$. Clearly, $\widehat{r}_t(x,a) \leq \widetilde{r}_t(x,a)$ and $\widehat{\ell}_t(x,a) \geq \widetilde{\ell}_t(x,a)$.

We start employing the analysis of Lemma 13 (Jin et al., 2019) with our uniform initialization over state-action pairs, for any valid comparator occupancy measure $q^* \in \Delta(P)$:

$$\sum_{t=1}^{T}\langle q_t - q^*, \widehat{\ell}_t\rangle \leq \frac{L\ln(|X||A|)}{\eta} + \frac{1}{\eta}\sum_{t=1}^{T} D(q_t\|\widetilde{q}_{t+1}). \tag{13}$$

Let us focus on the last term,

$$\begin{aligned}
D(q_t\|\widetilde{q}_{t+1}) &= \sum_{(x,a)} q_t(x,a)\ln\left(\frac{q_t(x,a)}{\frac{q_t(x,a)e^{-\eta\widehat{\ell}_t(x,a)}}{\sum_{x'\in X_{k(x)},a'\in A} q_t(x',a')e^{-\eta\widehat{\ell}_t(x',a')}}}\right) \\
&= \sum_{(x,a)} q_t(x,a)\left(\eta\widehat{\ell}_t(x,a) + \ln\left(\sum_{x'\in X_{k(x)},a'\in A} q_t(x',a')e^{-\eta\widehat{\ell}_t(x',a')}\right)\right) \\
&= \sum_{k=0}^{L-1}\sum_{x\in X_k,a\in A} q_t(x,a)\left(\eta\widehat{\ell}_t(x,a) + \ln\left(\sum_{x'\in X_k,a'\in A} q_t(x',a')e^{-\eta\widehat{\ell}_t(x',a')}\right)\right) \\
&= \eta\sum_{k=0}^{L-1}\sum_{x\in X_k,a\in A} q_t(x,a)\widehat{\ell}_t(x,a) + \sum_{k=0}^{L-1}\ln\left(\sum_{x'\in X_k,a'\in A} q_t(x',a')e^{-\eta\widehat{\ell}_t(x',a')}\right) \\
&= \eta\sum_{k=0}^{L-1}\sum_{x\in X_k,a\in A} q_t(x,a)(1-\widehat{r}_t(x,a)) + \sum_{k=0}^{L-1}\ln\left(\sum_{x'\in X_k,a'\in A} q_t(x',a')e^{-\eta(1-\widehat{r}_t(x',a'))}\right) \\
&= -\eta\sum_{k=0}^{L-1}\sum_{x\in X_k,a\in A} q_t(x,a)\widehat{r}_t(x,a) + \sum_{k=0}^{L-1}\ln\left(\sum_{x'\in X_k,a'\in A} q_t(x',a')e^{\eta\widehat{r}_t(x',a')}\right).
\end{aligned}$$

Now, we use Equation (12) of Gabbianelli et al. (2022), setting $\gamma = \eta/2$:

$$\begin{aligned}
D(q_t\|\widetilde{q}_{t+1}) &\leq -\eta\sum_{k=0}^{L-1}\sum_{x\in X_k,a\in A} q_t(x,a)\widehat{r}_t(x,a) + \sum_{k=0}^{L-1}\ln\left(\sum_{x'\in X_k,a'\in A} q_t(x',a')e^{\eta\widehat{r}_t(x',a')}\right) \\
&\leq -\eta\sum_{k=0}^{L-1}\sum_{x\in X_k,a\in A} q_t(x,a)\widehat{r}_t(x,a) + \sum_{k=0}^{L-1}\ln\left(\sum_{x'\in X_k,a'\in A} q_t(x',a')\exp\left(\frac{\eta}{2\gamma}\ln(1+2\gamma\widetilde{r}_t(x',a'))\right)\right) \\
&\leq -\eta\sum_{k=0}^{L-1}\sum_{x\in X_k,a\in A} q_t(x,a)\widehat{r}_t(x,a) + \sum_{k=0}^{L-1}\ln\left(\sum_{x'\in X_k,a'\in A} q_t(x',a')(1+\eta\widetilde{r}_t(x',a'))\right) \\
&\leq -\eta\sum_{k=0}^{L-1}\sum_{x\in X_k,a\in A} q_t(x,a)\widehat{r}_t(x',a') + \eta\sum_{k=0}^{L-1}\sum_{x'\in X_k,a'\in A} q_t(x',a')\widetilde{r}_t(x',a') \\
&= -\eta\langle q_t, \widehat{r}_t - \widetilde{r}_t\rangle \\
&= \eta\langle q_t, \widehat{\ell}_t - \widetilde{\ell}_t\rangle.
\end{aligned}$$

Going back to Equation (13):

$$\sum_{t=1}^{T}\langle q_t - q^*, \widehat{\ell}_t\rangle \leq \frac{L\ln(|X||A|)}{\eta} + \sum_{t=1}^{T}\langle q_t, \widehat{\ell}_t - \widetilde{\ell}_t\rangle. \tag{14}$$

Now, let us come back to the quantity of interest. We get,

$$\begin{aligned}
R_T &= \sum_{t=1}^{T}\langle q_t - q^*, \ell_t\rangle \\
&\leq \sum_{t=1}^{T}\langle q_t - q^*, \widetilde{\ell}_t\rangle + L\sqrt{2T\ln\left(\frac{1}{\delta}\right)} \\
&= \sum_{t=1}^{T}\langle q_t - q^*, \widehat{\ell}_t\rangle + \sum_{t=1}^{T}\langle q_t - q^*, \widetilde{\ell}_t - \widehat{\ell}_t\rangle + L\sqrt{2T\ln\left(\frac{1}{\delta}\right)} \\
&= \sum_{t=1}^{T}\langle q_t - q^*, \widehat{\ell}_t\rangle + \sum_{t=1}^{T}\langle q_t, \widetilde{\ell}_t - \widehat{\ell}_t\rangle - \sum_{t=1}^{T}\langle q^*, \widetilde{\ell}_t - \widehat{\ell}_t\rangle + L\sqrt{2T\ln\left(\frac{1}{\delta}\right)},
\end{aligned}$$

where the second inequality holds with probability at least $1-\delta$ for Azuma-Hoeffding inequality. Now, we apply Equation (14):

$$\begin{aligned}
R_T &\leq \frac{L\ln(|X||A|)}{\eta} + \sum_{t=1}^{T}\langle q_t, \widehat{\ell}_t - \widetilde{\ell}_t\rangle + \sum_{t=1}^{T}\langle q_t, \widetilde{\ell}_t - \widehat{\ell}_t\rangle - \sum_{t=1}^{T}\langle q^*, \widetilde{\ell}_t - \widehat{\ell}_t\rangle + L\sqrt{2T\ln\left(\frac{1}{\delta}\right)} \\
&= \frac{L\ln(|X||A|)}{\eta} + \sum_{t=1}^{T}\langle q^*, \widehat{\ell}_t - \widetilde{\ell}_t\rangle + L\sqrt{2T\ln\left(\frac{1}{\delta}\right)} \\
&= \frac{L\ln(|X||A|)}{\eta} + \sum_{t=1}^{T}\langle q^*, \widetilde{r}_t - \widehat{r}_t\rangle + L\sqrt{2T\ln\left(\frac{1}{\delta}\right)} \\
&= \frac{L\ln(|X||A|)}{\eta} + \sum_{t=1}^{T}\langle q^*, \widetilde{r}_t - \widehat{r}_t \pm \mathbb{E}_t\left[\widetilde{r}_t - \widehat{r}_t\right]\rangle + L\sqrt{2T\ln\left(\frac{1}{\delta}\right)} \\
&= \frac{L\ln(|X||A|)}{\eta} + \sum_{t=1}^{T}\langle q^*, (\widetilde{r}_t - \widehat{r}_t) - \mathbb{E}\left[\widetilde{r}_t - \widehat{r}_t\right]\rangle + \sum_{t=1}^{T}\langle q^*, \mathbb{E}_t\left[\widetilde{r}_t - \widehat{r}_t\right]\rangle + L\sqrt{2T\ln\left(\frac{1}{\delta}\right)} \\
&= \frac{L\ln(|X||A|)}{\eta} + \langle q^*, \sum_{t=1}^{T}(\widetilde{r}_t - \widehat{r}_t) - \mathbb{E}_t\left[\widetilde{r}_t - \widehat{r}_t\right]\rangle + \gamma\sum_{t=1}^{T}\langle q^*, \frac{r_t}{q^{\pi_C}+\gamma}\rangle + L\sqrt{2T\ln\left(\frac{1}{\delta}\right)} \\
&\leq \frac{L\ln(|X||A|)}{\eta} + \langle q^*, 1/q^{\pi_C}\rangle\sqrt{2T\ln\left(\frac{|X||A|}{\delta}\right)} + \gamma\sum_{t=1}^{T}\langle q^*, \frac{r_t}{q^{\pi_C}+\gamma}\rangle + L\sqrt{2T\ln\left(\frac{1}{\delta}\right)},
\end{aligned}$$

where the last inequality follows from Azuma-Hoeffding for every state-action pair, noticing that $|\widetilde{r}_t(x,a) - \widehat{r}_t(x,a)| \leq 1/q^{\pi_C}(x,a)$, and applying a union bound over states and actions. The final result holds with probability $1-2\delta$.

□

D Pessimistic Algorithm with Unknown Policy

In this section, we report the omitted proof related to the pessimistic algorithm, when the colleague's policy is *not* known beforehand.

Theorem 5.1. *With probability at least* $1-3\delta$*, Algorithm 4 for* $\gamma_t = \epsilon_t + \gamma = \epsilon_t + \eta/2$ *with* $\epsilon_t = \sqrt{\frac{\ln(|X||A|T/\delta)}{2t}}$ *(*i.e.*,* $\delta_t = \delta/T$*) attains, for any valid comparator occupancy measure* $q^* \in \Delta(P)$*:*

$$R_T \leq \frac{L\ln(|X||A|)}{\eta} + L\sqrt{2T\ln\left(\frac{1}{\delta}\right)} + \mathcal{D}\left(\pi^*, \pi_C\right)\left(4\sqrt{2T\ln\left(\frac{1}{\delta}\right)} + \ln\left(\frac{T|X||A|}{\delta}\right)\sqrt{T} + \gamma T\right)$$

In particular, setting $\eta = 2\gamma = \mathcal{O}\left(1/\sqrt{T}\right)$*,*

$$R_T \leq \tilde{\mathcal{O}}\left(\mathcal{D}\left(\pi^*, \pi_C\right)\sqrt{T}\right).$$

Proof. Let $\gamma_t > 0\ \forall t \in [T]$, we define $\widehat{r}_t(x,a) := \frac{r_t(x,a)}{\widehat{q}_t^{\pi_C}(x,a)+\gamma_t}\mathbb{1}\{x_{k(x)} = x, a_{k(x)} = a\}$, $\widetilde{r}_t(x,a) := \frac{r_t(x,a)}{q^{\pi_C}(x,a)}\mathbb{1}\{x_{k(x)} = x, a_{k(x)} = a\}$ and $\widetilde{\ell}_t(x,a) := 1 - \widetilde{r}_t(x,a)$.

We first notice that the analysis of Lemma 13 of Jin et al. (2019) still holds when the colleague's policy π_C is not known, namely, with our initialization over the state-action pairs (x,a) we have:

$$\sum_{t=1}^{T}\langle q_t - q^*, \widehat{\ell}_t\rangle \leq \frac{L\ln(|X||A|)}{\eta} + \frac{1}{\eta}\sum_{t=1}^{T} D(q_t\|\widetilde{q}_{t+1}). \tag{15}$$

To deal with the second terms we need to upperbound the error of the empirical mean on the occupancy measure $\widehat{q}_t^{\,\pi_C}$. Thus we define a tolerance parameter ϵ_t depending on the failure probability $\delta_t \in [0,1]$, as follows:

$$\epsilon_t = \sqrt{\frac{\ln\left(|X||A|/\delta_t\right)}{2t}}, \quad t \geq 1.$$

Furtnermore we let E_t be the event defined as follows:

$$E_t = \{|\widehat{q}_t^{\,\pi_C}(x,a) - q^{\pi_C}(x,a)| \leq \epsilon_t\ \forall (x,a) \in X \times A\}.$$

By Hoeffding inequality and applying a union bound we have that the event E_t holds with probability $1-\delta_t$. Then, by letting $\mathcal{E} := \bigcap_{t\in[T]} E_t$ we have that:

$$\mathbb{P}(\mathcal{E}) = \mathbb{P}\left(\bigcap_{t\in[T]} E_t\right) \geq 1 - \sum_{t=1}^{T}\delta_t = 1 - \delta$$

which holds by letting $\delta_t = \delta/T$. Moreover we let $\gamma_t = \epsilon_t + \eta/2 = \epsilon_t + \gamma$, where γ is the same as Algorithm 3, we have a similar result to (Gabbianelli et al., 2022), that is,

$$\widehat{r}_t(x,a) \leq \frac{1}{\eta}\ln\left(1 + \eta\tilde{r}_t(x,a)\right). \tag{16}$$

Proceeding as in the known policy setting, we focus on $D(q_t\|\widetilde{q}_{t+1})$ and we obtain:

$$D(q_t\|\widetilde{q}_{t+1}) = -\eta\sum_{k=0}^{L-1}\sum_{x\in X_k, a\in A} q_t(x,a)\widehat{r}_t(x,a) + \sum_{k=0}^{L-1}\ln\left(\sum_{x'\in X_k, a'\in A} q_t(x',a')e^{\eta\widehat{r}_t(x',a')}\right).$$

Now we apply Eq. (16) to have, under the event $\mathcal{E}$,

$$\begin{aligned}
D(q_t\|\widetilde{q}_{t+1}) &\le -\eta\sum_{k=0}^{L-1}\sum_{x\in X_k,a\in A} q_t(x,a)\widehat{r}_t(x,a) + \sum_{k=0}^{L-1}\ln\left(\sum_{x'\in X_k,a'\in A} q_t(x',a')e^{\eta\widehat{r}_t(x',a')}\right)\\
&\le -\eta\sum_{k=0}^{L-1}\sum_{x\in X_k,a\in A} q_t(x,a)\widehat{r}_t(x,a) + \sum_{k=0}^{L-1}\ln\left(\sum_{x'\in X_k,a'\in A} q_t(x',a')\exp\left(\frac{\eta}{2\gamma}\ln(1+2\gamma\widetilde{r}_t(x',a'))\right)\right)\\
&\le -\eta\sum_{k=0}^{L-1}\sum_{x\in X_k,a\in A} q_t(x,a)\widehat{r}_t(x,a) + \sum_{k=0}^{L-1}\ln\left(\sum_{x'\in X_k,a'\in A} q_t(x',a')(1+\eta\widetilde{r}_t(x',a'))\right)\\
&\le -\eta\sum_{k=0}^{L-1}\sum_{x\in X_k,a\in A} q_t(x,a)\widehat{r}_t(x',a') + \eta\sum_{k=0}^{L-1}\sum_{x'\in X_k,a'\in A} q_t(x',a')\widetilde{r}_t(x',a')\\
&= -\eta\langle q_t, \widehat{r}_t - \widetilde{r}_t\rangle\\
&= \eta\langle q_t, \widehat{\ell}_t - \widetilde{\ell}_t\rangle.
\end{aligned}$$

Now following the proof of Theorem 4.1, we obtain:

$$\begin{aligned}
R_T &\le \sum_{t=1}^{T}\langle q_t - q^*, \widehat{\ell}_t\rangle + \sum_{t=1}^{T}\langle q_t, \widetilde{\ell}_t - \widehat{\ell}_t\rangle - \sum_{t=1}^{T}\langle q^*, \widetilde{\ell}_t - \widehat{\ell}_t\rangle + L\sqrt{2T\ln\left(\frac{1}{\delta}\right)}\\
&\le \frac{L\ln(|X||A|)}{\eta} + \sum_{t=1}^{T}\langle q^*, \widetilde{r}_t - \widehat{r}_t \pm \mathbb{E}_t\left[\widetilde{r}_t - \widehat{r}_t\right]\rangle + L\sqrt{2T\ln\left(\frac{1}{\delta}\right)},
\end{aligned}$$

where the first inequality holds with probability at least $1-\delta$ for Azuma-Hoeffding inequality. Thus, we proceed bounding the following:

$$R_T \le \frac{L\ln(|X||A|)}{\eta} + \sum_{t=1}^{T}\langle q^*, (\widetilde{r}_t - \widehat{r}_t) - \mathbb{E}_t\left[\widetilde{r}_t - \widehat{r}_t\right]\rangle + \sum_{t=1}^{T}\langle q^*, \mathbb{E}_t\left[\widetilde{r}_t - \widehat{r}_t\right]\rangle + L\sqrt{2T\ln\left(\frac{1}{\delta}\right)}. \tag{17}$$

To bound the second term of the Inequality (17) we first observe:

$$\begin{aligned}
|\langle q^*, \widetilde{r}_t - \widehat{r}_t - \mathbb{E}_t\left[\widetilde{r}_t - \widehat{r}_t\right]|\rangle &\le 2|\langle q^*, \widetilde{r}_t - \widehat{r}_t\rangle|\\
&= 2\left|\langle q^*, \frac{\widehat{q}_t^{\,\pi_C} - q^{\pi_C} + \gamma_t}{q^{\pi_C}\left(\widehat{q}_t^{\,\pi_C} + \gamma_t\right)}\rangle\right|\\
&\le 2\left|\langle q^*, \frac{\gamma + 2\epsilon_t}{q^{\pi_C}\left(\gamma + \epsilon_t\right)}\rangle\right| \le 4\langle q^*, 1/q^{\pi_C}\rangle,
\end{aligned}$$

where we employed the fact that under the event $\mathcal{E}$ it holds $\widehat{q}_t^{\,\pi_C} \le q^{\pi_C} + \epsilon_t$ and the definition of $\gamma_t = \gamma + \epsilon_t$. Then by Azuma inequality we have with probability $1-\delta$:

$$\sum_{t=1}^{T}\langle q^*, (\widetilde{r}_t - \widehat{r}_t) - \mathbb{E}_t\left[\widetilde{r}_t - \widehat{r}_t\right]\rangle \le 4\sqrt{2T\ln\left(\frac{1}{\delta}\right)}\langle q, 1/q^{\pi_C}\rangle.$$

We know focus on bounding the third term of the Inequality (17), as follows:

$$\begin{aligned}
\sum_{t=1}^{T}\langle q^*, \mathbb{E}_t\left[\widetilde{r}_t - \widehat{r}_t\right]\rangle &= \sum_{t=1}^{T}\langle q^*, q^{\pi_C}\frac{\widehat{q}_t^{\,\pi_C} - q^{\pi_C} + \gamma_t}{q^{\pi_C}\left(\widehat{q}_t^{\,\pi_C} + \gamma_t\right)} r_t\rangle \\
&\leq \sum_{t=1}^{T}\langle q^*, \frac{\gamma_t + \epsilon_t}{\widehat{q}_t^{\,\pi_C} + \gamma_t} r_t\rangle \\
&\leq \sum_{t=1}^{T}\langle q^*, \frac{\gamma + 2\epsilon_t}{q^{\pi_C} + \gamma} r_t\rangle \\
&= \sum_{t=1}^{T}\langle q^*, \frac{\gamma}{q^{\pi_C} + \gamma} r_t\rangle + \sum_{t=1}^{T}\langle q^*, \frac{2\epsilon_t}{q^{\pi_C} + \gamma} r_t\rangle \\
&= \gamma\sum_{t=1}^{T}\langle q^*, \frac{r_t}{q^{\pi_C} + \gamma}\rangle + \sum_{t=1}^{T}\langle q^*, \frac{2\epsilon_t}{q^{\pi_C} + \gamma} r_t\rangle \\
&\leq \gamma\sum_{t=1}^{T}\langle q^*, \frac{r_t}{q^{\pi_C} + \gamma}\rangle + 2\langle q^*, \frac{1}{q^{\pi_C} + \gamma}\rangle\sum_{t=1}^{T}\epsilon_t \\
&\leq \gamma\sum_{t=1}^{T}\langle q^*, \frac{r_t}{q^{\pi_C} + \gamma}\rangle + 4\langle q^*, \frac{1}{q^{\pi_C} + \gamma}\rangle \ln\left(\frac{T|X||A|}{\delta}\right)\sqrt{T},
\end{aligned}$$

where the first steps follows from the definition of γ_t and $\widehat{q}_t^{\,\pi_C}$ and the last inequality follows from the fact that $\sum_{t\in[T]}\frac{1}{t} \leq 2\sqrt{T}$.

The final result holds with probability $1 - 3\delta$, namely,

$$\begin{aligned}
R_T \leq & \frac{L\ln(|X||A|)}{\eta} + 4\sqrt{2T\ln\left(\frac{1}{\delta}\right)}\langle q^*, 1/q^{\pi_C}\rangle + \gamma\sum_{t=1}^{T}\langle q^*, \frac{r_t}{q^{\pi_C} + \gamma}\rangle \\
& + 4\langle q^*, \frac{1}{q^{\pi_C} + \gamma}\rangle \ln\left(\frac{T|X||A|}{\delta}\right)\sqrt{T} + L\sqrt{2T\ln\left(\frac{1}{\delta}\right)}
\end{aligned}$$

□