Online Learning with Off-Policy Feedback in Adversarial MDPs

Francesco Bacchiocchi* , **Francesco Emanuele Stradi*** , **Matteo Papini** , **Alberto Maria Metelli** and **Nicola Gatti**

Politecnico di Milano

{francesco.bacchiocchi, francescoemanuele.stradi, matteo.papini, albertomaria.metelli, nicola.gatti}@polimi.it

Abstract

In this paper, we face the challenge of online learning in *adversarial* Markov decision processes with *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). The off-policy feedback presents an additional issue that is not present in traditional settings: the learner is charged with the regret of its chosen policy but it observes only the rewards gained by the colleague's policy. First, we present a lower-bound for the setting we propose, which shows that the optimal dependency of the sublinear regret is w.r.t. the dissimilarity between the optimal policy in hindsight and the colleague's policy. Then, we propose novel algorithms that, by employing *pessimistic* estimators—commonly adopted in the offline reinforcement learning literature—ensure sublinear regret bounds depending on the desired dissimilarity, even when the colleague's policy 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 [Even-Dar *et al.*, 2009; Neu *et al.*, 2014]. 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 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 [Jaksch *et al.*, 2010; Azar *et al.*, 2017] and optimistic versions of mirror descent for the adversarial case [Jin *et al.*, 2020]. All these approaches consider *on-policy feedback*, meaning that the learner observes the trajectory and loss (or reward) generated by playing their own currently selected 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 the 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 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. 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.*, 2022; 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.

Much less has been said on off-policy feedback in the adversarial setting, which is naturally online (in an offline scenario, one cannot expect the adversary to behave the same at training and evaluation time). In fact, off-policy learning with adversarial rewards has been first considered by [Gabbianelli *et al.*, 2023] for multi-armed-bandit (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 achieving 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, largely 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. As another example, imagine an online service where financial reward does not come immediately, but after a long time by building fidelity. The users are not providing immediate "satisfaction" feedback either, but they (or the users who belong to the same user base) are giving such feedback to a more established competitor through online reviews. In this case, the "colleague" is the competitor, the adversary is the user base, the colleague's policy is unknown, and, in the short run, the only observation is the feedback given to the colleague. Intuitively, it is still possible to improve your actions by observing the competitor. Doing so online, rather than by analyzing historical data, allows to quickly adapt to nonstationary phenomena like boredom and hype, and to the changing taste of the user base.

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.

1.1 Related Works

In the following, we present the most relevant works from the literature, dividing the discussion into offline RL, online RL, and off-policy feedback in online learning scenarios.

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 (e.g., [Xiao *et al.*, 2021; Rashidinejad *et al.*, 2022; 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.*, 2022], 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 meaningfulness 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, 2016] in MDPs is extensive, as investigated in several notable works [Auer *et al.*, 2008; Even-Dar *et al.*, 2009; Neu *et al.*, 2014]. [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, 2019a] 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, 2019b], 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.*, 2020]. 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.*, 2023], 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.

1.2 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 $t \in [T] \coloneqq \{1, \dots, T\}$, the agent plays a policy π_t over the horizon L and then observes the feedback generated by a colleague's policy π_C. First, we present a minimax regret lower bound that depends on a dissimilarity measure between the (occupancy measure of the) best policy in hindsight π^* and the policy π_C of the colleague. We propose two *pessimistic* algorithms, called P-REPS and P-REPS+. P-REPS works in the setting where the *colleague's policy is known*. By employing a pessimistically biased estimator, P-REPS guarantees sublinear regret with a high probability that depends on the dissimilarity between π^* and π_C. In terms of expected regret, P-REPS matches the lower bound. Finally, we show that P-REPS+ achieves similar regret guarantees compared to P-REPS, even when the *colleague's policy is unknown*. Our work answers positively to the question raised by [Gabbianelli *et al.*, 2023], whether it is possible to optimally learn in an adversarial off-policy setting when the environment is a Markov decision process.

1.3 Paper Structure

The paper is structured as follows.

- In Section 2, we provide the problem formulation with the necessary notation. Precisely, we focus on the learner-environment interaction and on the performance measures used to evaluate the proposed algorithms.
- In Section 3, we provide a lower-bound for the proposed setting. Specifically, we show that any algorithm must suffer a regret which depends on the dissimilarity between the optimal occupancy and the one of the colleague.
- In Section 4, we present the algorithms tailored for the *off-policy* feedback and we prove their theoretical guarantees.

  In Section 4.1, we focus on the scenarios where the colleague's policy is known. In such a setting, we propose P-REPS (Algorithm 2) and we show that it achieves sublinear regret with high probability. Furthermore, we show that P-REPS matches the lower bound presented in Section 3 in terms of cumulative expected regret.

  In Section 4.2, we focus on the scenarios where the colleague's policy is *not* known to the learner. In such a case, we propose P-REPS+ (Algorithm 3) which attains similar guarantees w.r.t to P-REPS in terms of regret upper bound.
- In Section 5, we summarize our findings and discuss some possible future works.

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 process (MDP) is a 4-tuple $M = (X, A, P, \{r_t\}_{t=1}^T)$ where:

- X and A are the finite set of states and actions, respectively. By the loop-free property, X is partitioned into $L+1$ layers $X_0, \dots, X_L$ such that the first and the last layers are singletons, that is, $X_0 = \{x_0\}$ and $X_L = \{x_L\}$;
- T is the number of episodes, with $t \in [T]$ indexing a specific episode;
- $P : X \times A \times X \to [0, 1]$ 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 $k \in \{0, \dots, L-1\}$;
- $\{r_t\}_{t=1}^T$ is the sequence of rewards for each episode $t \in [T]$, and we assume $r_t \in [0, 1]^{|X \times A|}$. We refer to the reward of a specific state-action pair $(x, a) \in X \times A$ for a specific episode $t \in [T]$ as $r_t(x, a)$. Rewards 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 replicating 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 selecting action $a \in A$ in state $x \in X$.

We study an *off-policy* online setting, following [Gabbianelli *et al.*, 2023]. In this setting, there is an external fixed policy π_C which is played in parallel to the learner. The feedback received by the learner at the end of each episode is the

Algorithm 1 Learner-Environment Interaction

for $t \in [T]$ **do**
  The environment chooses r_t *adversarially*
  The learner chooses a policy $\pi_t : X \times A \to [0, 1]$
  The state is initialized as x_0 for both the agent and the colleague
  for $k \in \{0, \dots, L-1\}$ **do**
    The colleague plays $a'_k \sim \pi_C(\cdot|x'_k)$
    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 $\{r_t(x'_k, a'_k)\}_{k=0}^{L-1}$ but it gains $\{r_t(x_k, a_k)\}_{k=0}^{L-1}$
end for

one pertaining to π_C. Algorithm 1 describes the interaction between the learner and environment in an off-policy adversarial MDP.

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 $k \in \{0, \dots, L-1\}$,

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

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

$$\sum_{x \in X_k} \sum_{a \in A} q^{P,\pi}(x, a) = 1, \tag{2}$$

for each state-action pair $(x, a) \in X_k \times A$ and for each layer $k \in \{1, \dots, L\}$. In addition, we have

$$\sum_{a \in A} q^{P,\pi}(x, a) = \sum_{x' \in X_{k(x)-1}} \sum_{a' \in A} P(x|x', a') q^{P,\pi}(x', a'), \tag{3}$$

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 those satisfying Equation (2) and Equation (3). Note that any valid occupancy measure q induces a policy π^q defined as

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

such that $q^{P,\pi^q} = q$.

2.3 Cumulative Regret

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

Definition 2.1 (Cumulative Regret)**.** *Given a set of policies $\{\pi_t\}_{t=1}^T$ specifying the learner's strategy at each round, we define the cumulative regret over T rounds as follows:*

$$R_T(\{\pi_t\}_{t=1}^T) := \max_{q \in \Delta(P)} \sum_{t=1}^T \langle r_t, q \rangle - \sum_{t=1}^T \langle r_t, q^{P,\pi_t} \rangle. \tag{4}$$

Furthermore, we define the expected cumulative regret as $\mathbb{E}\left[R_T(\{\pi_t\}_{t=1}^T)\right]$, where the expectation is taken over the randomness of the environment. In traditional online learning settings, an algorithm presents good performance when its regret is sublinear in T, namely $R_T(\{\pi_t\}_{t=1}^T) = o(T)$. In our setting, this property is not sufficient. Indeed, the regret necessarily depends on some dissimilarity measure between the optimal 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). To achieve 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}. Furthermore, from here on, we omit the dependence on the policies selected by the learner $\{\pi_t\}_{t=1}^T$ in the formulation of the cumulative regret, referring to $R_T(\{\pi_t\}_{t=1}^T)$ as R_T whenever it is clear from the context.

3 Lower Bound

As a first step, we introduce a negative result for online off-policy settings. Such a negative result rules out the possibility for any algorithm to suffer a regret bound independent of the dissimilarity measure between the optimal occupancy and the colleague's one. Formally, the following holds.

Theorem 3.1. *For any policies selected by the learner $\{\pi_t\}_{t=1}^T$ and any colleague's policy π_C, there exists an instance such that, for some absolute constant $C > 0$:*

$$\mathbb{E}[R_T] \geq C \sqrt{T \sum_{(x,a) \in X \times A} \frac{q^*(x, a)}{q^{\pi_C}(x, a)}},$$

where q^* is the occupancy measure of the best policy in hindsight, i.e., $\pi^* \in \arg\max_\pi \sum_{t=1}^T \langle r_t, q^{P,\pi} \rangle$.

The negative result presented in Theorem 3.1 captures the intuition that, when the optimal occupancy is not well covered by the colleague's one, any algorithm may suffer arbitrarily large regret. To prove Theorem 3.1, we consider two stochastic bandit instances, each with two available arms. In the two instances, the first arms have the same distribution, while the distributions of the second arms are slightly different. Thus, if the colleague's policy provides few samples from the second arm to the learner, it would fail to distinguish efficiently between the two.

In the following, we denote the dissimilarity between the optimal occupancy q^* with respect to the colleague's occupancy q^{π_C} as *coverage ratio* and we formally define it as follows:

$$\mathcal{D}(\pi^*, \pi_C) \coloneqq \sum_{(x,a)\in X\times A} \frac{q^*(x,a)}{q^{\pi_C}(x,a)}.$$

Consequently, from now on, we aim at designing no-regret algorithms whose regret bound scales with such a parameter $\mathcal{D}(\pi^*, \pi_C)$. As a final remark, we observe that the result presented in Theorem 3.1 is in line with the one achieved in offline off-policy settings. Specifically, [Rashidinejad *et al.*, 2022] show that no algorithm can achieve an expected suboptimality gap independent of the coverage ratio between the optimal occupancy measure and the behavior one.

4 Algorithms

Online learning algorithms are generally developed in order to properly manage the exploration-exploitation trade-off. This is necessary in order to deal with the partial observability of the environment, namely, to reduce the uncertainty about the environment while keeping the regret small. Nevertheless, when the feedback is *off-policy*, the learner does not gain knowledge choosing explorative policies, thus, employing exploration techniques may be counterproductive. In the appendix (see App. B), we study the regret of UOB-REPS [Jin *et al.*, 2020] with known transitions (assuming the knowledge of the transitions, the algorithm almost reduces to [Zimin and Neu, 2013]) in the off-policy setting. To the best of our knowledge, UOB-REPS is considered the state-of-the-art (optimistic) algorithm to solve adversarial MDPs. Following the standard regret analysis, UOB-REPS achieves sublinear regret which scales with the suboptimal constant:

$$\sup_{t\in[T]} \sum_{(x,a)\in X\times A} \frac{q_t(x,a)}{q^{\pi_C}(x,a)}.$$

The latter result is coherent with [Gabbianelli *et al.*, 2023], where a similar result holds employing a non-pessimistic algorithm such as EXP3. Although this result cannot rule out the employment of optimism in off-policy settings, we take it as an indication to focus on pessimism instead.

4.1 Pessimistic Algorithm with Known Policy

As previously observed, optimistic algorithms are designed to steer the learning dynamics towards less explored state-action pairs. However, when the feedback is related to an independent policy, exploring may be useless, as the learner does not gain any knowledge about the rewards achieved at each episode.

As a result, we employ pessimistic estimators of the rewards. Such estimators leverage the situations where the dissimilarity between the optimal occupancy in hindsight and the colleague's one is small, and allow us to achieve the desired dependence on the coverage ratio in the regret bound.

Algorithm. Algorithm 2 provides the pseudocode of our *Pessimistic Relative Entropy Search* (P-REPS). More specifically, Algorithm 2 is a pessimistic variant of UOB-REPS [Jin *et al.*, 2020] with known transitions (Algorithm 4). In the following, we describe the algorithm and remark the main differences compared to the original, optimistic version by [Jin *et al.*, 2020].

The policy is initially set to be uniform over the action space for each state and the occupancy q^{π_1} is the one induced by 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 (Line 3), while our own policy is executed. Once the rewards are gathered, the algorithm builds a pessimistic estimator as follows:

$$\widehat{r}_t(x,a) = \frac{r_t(x,a)}{q^{\pi_C}(x,a)+\gamma}\mathbb{1}\{x_{k(x)} = x, a_{k(x)} = a\}, \quad (5)$$

for each $(x,a) \in X \times A$. In Equation (5), we add a constant factor $\gamma > 0$ in the denominator of the resulting biased estimator. We remark that this choice leads to an underestimate of the rewards (Line 5).

Differently from UOB-REPS, Algorithm 2 updates the occupancy measure by employing a normalized version of OMD (Line 7) as follows:

$$\widetilde{q}_{t+1}(x,a) = \frac{q_t(x,a)e^{\eta\widehat{r}_t(x,a)}}{\sum_{x'\in X_{k(x)},a'\in A} q_t(x',a')e^{\eta\widehat{r}_t(x',a')}}, \quad (6)$$

$$q_{t+1} = \arg\min_{q\in\Delta(P)} D(q\|\widetilde{q}_{t+1}), \quad (7)$$

where $D(\cdot||\cdot)$ is the KL-divergence. The peculiarity of this update lies in its first unconstrained step. Specifically, the standard unconstrained optimization update $q_t(x,a)e^{\eta\widehat{r}_t(x,a)}$ 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.*, 2023], where the decision point is optimized independently from the projection, and greedy updates, which are more common in the online adversarial MDP literature. We remark that the computational complexity of the projection step (see Line 7) is the same as in UOB-REPS. In particular, the projection is still a convex optimization problem with linear constraints, which can be solved in polynomial time. Therefore, the projection step in Algorithm 2 can be efficiently computed.

Regret upper bound. P-REPS attains the following regret upper bound when the feedback is off-policy.

Theorem 4.1. *With probability at least $1-2\delta$, Algorithm 2 with $\eta = 2\gamma$ attains the following regret bound:*

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

which, setting

$$\eta = 2\gamma = \sqrt{\frac{L\ln(|X||A|)}{T}}$$

leads to,

$$R_T \le \mathcal{O}\left(L\mathcal{D}(\pi^*, \pi_C)\sqrt{\ln\left(\frac{|X||A|}{\delta}\right)T}\right).$$

Algorithm 2 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 $k \in \{0, \dots, L-1\}$ and $x \in X_k$, initialize policy

$$\pi_1(a|x) = \frac{1}{|A|}$$

and initialize the occupancy $q_1 = q^{\pi_1}$

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 $k \in \{0, \dots, L-1\}$ and rewards $r_t(x_k, a_k)$

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

5:

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

6: **end for**

7: Update occupancy measure:

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

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

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

9: **end for**

Furthermore, by means of Theorem 4.1, we can bound the expected cumulative regret as follows.

Corollary 4.1. *Algorithm 2 with*

$$\eta = 2\gamma = \sqrt{\frac{L \ln(|X||A|)}{T \mathcal{D}(\pi^*, \pi_C)}}$$

attains the following expected regret bound:

$$\mathbb{E}[R_T] \leq \mathcal{O}\left(\sqrt{LT \mathcal{D}(\pi^*, \pi_C) \ln(|X||A|)}\right).$$

We observe that Algorithm 2 matches the lower bound presented in Section 3 in terms of expected regret, confirming that employing pessimistic estimators is crucial when the feedback is *off-policy*. This result is coherent with [Gabbianelli *et al.*, 2023], where similar regret guarantees hold in the multi-armed bandit settings. On the other hand, in the high-probability version, the regret upper bound scales linearly with the coverage ratio $\mathcal{D}(\pi^*, \pi_C)$, leaving open the question of whether such a dependence is unavoidable when considering the regret bound with high probability or if a square-root dependence can be achieved with a better analysis. Notably, we remark that Theorem 4.1 still holds if the comparator occupancy is not optimal in hindsight, but any feasible occupancy measure. In such a case, the dissimilarity $\mathcal{D}(\pi^*, \pi_C)$ is between the comparator's and the colleague's occupancy. The same reasoning will be valid for the theoretical results of Section 4.2.

4.2 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 2, we achieve similar regret guarantees. To address the uncertainty arising from the unknown policy of the colleague, we employ a time-varying reward estimator. This allows us to introduce an additional level of pessimism in the estimates compared to that used in Algorithm 2. This extra pessimism helps us to deal with the uncertainty introduced by the unknown policy and achieve the desired regret guarantees.

Algorithm. Algorithm 3 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 2 (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, for each state-action pair (x, a) along the path traversed by the colleague, the counters are updated accordingly (Line 6). Then, the algorithm builds a pessimistic estimator (Line 9) as:

$$\widehat{r}_t(x,a) = \frac{r_t(x,a)}{\widehat{q}_t^{\pi_C}(x,a) + \gamma_t} \mathbb{1}\{x_{k(x)} = x, a_{k(x)} = a\}, \quad (8)$$

for every $(x, a) \in X \times A$, 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 3, 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:

$$|\widehat{q}_t^{\pi_C}(x,a) - q^{\pi_C}(x,a)| \leq \epsilon_t \coloneqq \sqrt{\frac{\ln(2|X||A|/\delta_t)}{2t}}, \quad (9)$$

for each $(x, a) \in X \times A$, with $t \geq 1$. Thus, the time-dependent pessimistic factor is set as $\gamma_t = \epsilon_t + \gamma$, where the γ is the same as in Algorithm 2. 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 estimation of the occupancy measure $\widehat{q}_t^{\pi_C}$. Finally, Algorithm 3 updates the occupancy measure employing a normalized version of OMD as done in Algorithm 2 (Line 11).

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 4.2. *With probability at least* $1 - 3\delta$*, Algorithm 3 with*

$$\gamma_t = \epsilon_t + \gamma = \epsilon_t + \eta/2 = \sqrt{\frac{\ln(2|X||A|/\delta_t)}{2t}} + \eta/2$$

and $\delta_t = \delta/T$ *attains the following regret bound:*

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

$$\times \left(4\sqrt{2T \ln\left(\frac{1}{\delta}\right)} + \ln\left(\frac{2T|X||A|}{\delta}\right)\sqrt{T} + \gamma T \right).$$

In particular, setting

$$\eta = 2\gamma = \sqrt{\frac{L \ln(|X||A|)}{T}},$$

we have:

$$R_T \leq \widetilde{\mathcal{O}}\left(L\mathcal{D}\left(\pi^*, \pi_C\right) \ln\left(\frac{|X||A|}{\delta}\right)\sqrt{T}\right).$$

Theorem 4.2 shows that the dependency on the constant factor $\mathcal{D}(\pi^*, \pi_C)$ 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}. Finally, by means of Theorem 4.2 we can bound the expected cumulative regret as follows.

Corollary 4.2. *Algorithm 3 with*

$$\eta = 2\gamma = \sqrt{\frac{L \ln(|X||A|)}{T}}$$

attains the following expected regret bound:

$$\mathbb{E}\left[R_T\right] \leq \widetilde{\mathcal{O}}\left(\mathcal{D}\left(\pi^*, \pi_C\right) \ln\left(|X||A|\right)\sqrt{LT}\right).$$

The result presented in Corollary 4.2 is in line with the one provided by [Gabbianelli *et al.*, 2023] when the colleague's policy is *not* known. Indeed, similarly to the multi-armed bandit case, the expected regret bound has a linear dependency on the dissimilarity $\mathcal{D}(\pi^*, \pi_C)$. We leave as an open problem whether the optimal $\sqrt{\mathcal{D}(\pi^*, \pi_C)}$ dependency can be achieved when the colleague's policy is unknown.

5 Conclusions

In this paper, we answer several questions raised by [Gabbianelli *et al.*, 2023]. We study the problem of online learning with *off-policy* feedback in adversarial Markov decision processes. We first present a lower bound for the proposed setting which shows that the optimal regret bound depends on the coverage ratio between the optimal occupancy and the colleague's one. 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 optimal 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.

Algorithm 3 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 policy

$$\pi_1(a|x) = \frac{1}{|A|}$$

and initialize the occupancy $q_1 = q^{\pi_1}$

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

Set:

$$\gamma_t \coloneqq \gamma + \epsilon_t = \gamma + \sqrt{\frac{\ln\left(2|X||A|T/\delta\right)}{2t}}$$

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 $k \in \{0, \dots, L-1\}$

for $k \in \{0, \dots, L-1\}$ **do**

Update counters:

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

end for

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

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

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

end for

Update occupancy measure:

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

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

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

end for

Besides the potential online-learning applications discussed in Section 1, we believe that our extension of the *off-policy* setting to MDPs is of particular interest for the offline reinforcement learning community. In particular, our implementation of pessimism does not rely on explicit uncertainty quantification, which may be useful for developing new algorithms for offline RL.

As future work, we aim to extend our results to encompass settings with unknown transition functions. 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.

Acknowledgments

This paper is supported by FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, Investment 1.3, Line on Artificial Intelligence). The authors thank Gergely Neu for the helpful discussions.

Contribution Statement

Francesco Bacchiocchi and Francesco Emanuele Stradi contributed equally.

References

[Auer *et al.*, 2008] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

[Azar *et al.*, 2017] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning, (ICML)*, volume 70, pages 263–272. PMLR, 2017.

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

[Cheng *et al.*, 2022] Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning (ICML)*, volume 162, pages 3852–3878. PMLR, 2022.

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

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

[Gabbianelli *et al.*, 2023] Germano Gabbianelli, Gergely Neu, and Matteo Papini. Online learning with off-policy feedback. In *International Conference on Algorithmic Learning Theory (ALT)*, volume 201, pages 620–641. PMLR, 2023.

[Hazan, 2016] Elad Hazan. Introduction to online convex optimization. *Found. Trends Optim.*, 2(3-4):157–325, 2016.

[Jaksch *et al.*, 2010] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, 2010.

[Jin *et al.*, 2020] Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, pages 4860–4869. PMLR, 2020.

[Jin *et al.*, 2021] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pages 5084–5096. PMLR, 18–24 Jul 2021.

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

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

[Levine *et al.*, 2020] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *CoRR*, abs/2005.01643, 2020.

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

[Neu *et al.*, 2014] Gergely Neu, András György, Csaba Szepesvári, and András Antos. Online markov decision processes under bandit feedback. *IEEE Trans. Autom. Control.*, 59(3):676–691, 2014.

[Neu, 2015] Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 3168–3176, 2015.

[Rashidinejad *et al.*, 2022] Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *IEEE Trans. Inf. Theory*, 68(12):8156–8196, 2022.

[Rosenberg and Mansour, 2019a] Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, pages 5478–5486. PMLR, 09–15 Jun 2019.

[Rosenberg and Mansour, 2019b] Aviv Rosenberg and Yishay Mansour. Online stochastic shortest path with bandit feedback and unknown transition function. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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

[Uehara and Sun, 2022] Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *The Tenth International Conference on Learning Representations (ICLR)*. OpenReview.net, 2022.

[Xiao *et al.*, 2021] 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 *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pages 11362–11371. PMLR, 18–24 Jul 2021.

[Zanette *et al.*, 2021] Andrea Zanette, Martin J. Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pages 13626–13640, 2021.

[Zimin and Neu, 2013] Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.