Learning Adversarial Markov Decision Processes with Bandit Feedback and Unknown Transition

Chi Jin [1] **Tiancheng Jin** [2] **Haipeng Luo** [2] **Suvrit Sra** [3] **Tiancheng Yu** [3]

Abstract

We consider the task of learning in episodic finite-horizon Markov decision processes with an unknown transition function, bandit feedback, and adversarial losses. We propose an efficient algorithm that achieves $\tilde{\mathcal{O}}(L|X|\sqrt{|A|T})$ regret with high probability, where L is the horizon, $|X|$ the number of states, $|A|$ the number of actions, and T the number of episodes. To our knowledge, our algorithm is the first to ensure $\tilde{\mathcal{O}}(\sqrt{T})$ regret in this challenging setting; in fact it achieves the same regret as (Rosenberg & Mansour, 2019a) who consider the easier setting with full-information. Our key contributions are two-fold: a tighter confidence set for the transition function; and an optimistic loss estimator that is inversely weighted by an *upper occupancy bound*.

1. Introduction

Reinforcement learning studies the problem where a learner interacts with the environment sequentially and aims to improve her strategy over time. The environment dynamics are usually modeled as a Markov Decision Process (MDP) with a fixed and unknown transition function. We consider a general setting where the interaction proceeds in episodes with a fixed horizon. Within each episode the learner sequentially observes her current state, selects an action, suffers and observes the loss corresponding to the chosen state-action pair, and then transits to the next state according to the underlying transition function.[1] The goal of the learner is to minimize her regret: the difference between her total loss and the total loss of an optimal fixed policy.

The majority of the literature in learning MDPs assumes stationary losses, that is, the losses observed for a specific state-action pair follow a fixed and unknown distribution. To better capture applications with non-stationary or even adversarial losses, the works (Even-Dar et al., 2009; Yu et al., 2009) are among the first to study the problem of learning adversarial MDPs, where the losses can change arbitrarily between episodes. There are several follow-ups in this direction, such as (Yu et al., 2009; Neu et al., 2010; 2012; Zimin & Neu, 2013; Dekel & Hazan, 2013; Rosenberg & Mansour, 2019a). See Section 1.1 for more related work.

For an MDP with $|X|$ states, $|A|$ actions, T episodes, and L steps in each episode, the best existing result is the work (Rosenberg & Mansour, 2019a), which achieves $\tilde{\mathcal{O}}(L|X|\sqrt{|A|T})$ regret, assuming a fixed and unknown transition function, adversarial losses, but importantly *full-information* feedback: i.e., the loss for *every* state-action pair is revealed at the end of each episode. On the other hand, with the more natural and standard *bandit feedback* (where only the loss for each visited state-action pair is revealed), a later work by the same authors (Rosenberg & Mansour, 2019b) achieves regret $\tilde{\mathcal{O}}(L^{3/2}|X||A|^{1/4}T^{3/4})$, which has a much worse dependence on the number of episodes T compared to the full-information setting.

Our main contribution significantly improves on (Rosenberg & Mansour, 2019b). In particular, we propose an efficient algorithm that achieves $\tilde{\mathcal{O}}(L|X|\sqrt{|A|T})$ regret in the same setting with bandit feedback, an unknown transition function, and adversarial losses. Although our regret bound still exhibits a gap compared to the best existing lower bound $\Omega(L\sqrt{|X||A|T})$ (Jin et al., 2018), to the best of our knowledge, for this challenging setting our result is the first to achieve $\tilde{\mathcal{O}}(\sqrt{T})$ regret. Importantly, this also matches the regret upper bound of Rosenberg & Mansour (2019a), who consider the easier setting with full-information feedback.

Our algorithm builds on the UC-O-REPS algorithm (Rosenberg & Mansour, 2019a;b)—we also construct confidence sets to handle the unknown transition function, and apply Online Mirror Descent over the space of occupancy measures (see Section 2.1) to handle adversarial losses. The first key difference and challenge is that with bandit feedback, to apply Online Mirror Descent we must construct good loss

[1]Princeton University [2]University of Southern California [3]Massachusetts Institute of Technology. Correspondence to: Tiancheng Jin <tiancheng.jin@usc.edu>, Tiancheng Yu <yutc@mit.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

[1]As in previous work (Rosenberg & Mansour, 2019a;b), throughout we use the term "losses" instead of "rewards" to be consistent with the adversarial online learning literature. One can translate between losses and rewards by simply taking negation.

estimators since the loss function is not completely revealed. However, the most natural approach of building unbiased loss estimators via inverse probability requires knowledge of the transition function, and is thus infeasible in our setting.

We address this key challenge by proposing a novel biased and optimistic loss estimator (Section 3.3). Specifically, instead of inversely weighting the observation by the probability of visiting the corresponding state-action pair (which is unknown), we use the *maximum probability among all plausible transition functions* specified by a confidence set, which we call *upper occupancy bound*. This idea resembles the optimistic principle of using *upper confidence bounds* for many other problems of learning with bandit feedback, such as stochastic multi-armed bandits (Auer et al., 2002a), stochastic linear bandits (Chu et al., 2011; Abbasi-Yadkori et al., 2011), and reinforcement learning with stochastic losses (Jaksch et al., 2010; Azar et al., 2017; Jin et al., 2018). However, as far as we know, applying optimism in constructing loss estimators for an adversarial setting is new.

The second key difference of our algorithm from UC-O-REPS (Section 3.1) lies in a new confidence set for the transition function. Specifically, for each state-action pair, the confidence set used in UC-O-REPS and previous works such as (Jaksch et al., 2010; Azar et al., 2017) imposes a total variation constraint on the transition probability, while our proposed confidence set imposes *an independent constraint on the transition probability for each next state,* and is strictly tighter. Indeed, with the former we can only prove an $\tilde{\mathcal{O}}(L|X|^2\sqrt{|A|T})$ regret, while with the latter we improve it to $\tilde{\mathcal{O}}(L|X|\sqrt{|A|T})$. Analyzing the non-trivial interplay between our optimistic loss estimators and the new confidence set is one of our key technical contributions.

Finally, we remark that our proposed upper occupancy bounds can be computed efficiently via backward dynamic programming and solving some linear programs greedily, and thus our algorithm can be implemented efficiently.

1.1. Related Work

Stochastic losses. Learning MDPs with stochastic losses and bandit feedback is relatively well-studied for the tabular case (that is, finite number of states and actions). For example, in the episodic setting, using our notation,[2] the UCRL2 algorithm of Jaksch et al. (2010) achieves $\tilde{\mathcal{O}}(\sqrt{L^3|X|^2|A|T})$ regret, and the UCBVI algorithm of Azar et al. (2017) achieves the optimal bound $\tilde{\mathcal{O}}(L\sqrt{|X||A|T})$, both of which are model-based algorithms and construct confidence sets for both the transition function and the loss function. The recent work (Jin et al., 2018) achieves a suboptimal bound $\tilde{\mathcal{O}}(\sqrt{L^3|X||A|T})$ via an optimistic Q-learning algorithm that is model-free. Besides the episodic setting, other setups such as discounted losses or infinite-horizon average-loss setting have also been heavily studied; see for example (Ouyang et al., 2017; Fruit et al., 2018; Zhang & Ji, 2019; Wei et al., 2019; Wang et al., 2019) for some recent works.

Adversarial losses. Based on whether the transition function is known and whether the feedback is full-information or bandit, we discuss four categories separately.

Known transition and full-information feedback. Early works on adversarial MDPs assume a known transition function and full-information feedback. For example, Even-Dar et al. (2009) propose the algorithm MDP-E and prove a regret bound of $\tilde{\mathcal{O}}(\tau^2\sqrt{T\ln|A|})$ where τ is the mixing time of the MDP; another work (Yu et al., 2009) achieves $\tilde{\mathcal{O}}(T^{2/3})$ regret. Both of these consider a continuous setting (as opposed to the episodic setting that we study). Later Zimin & Neu (2013) consider the episodic setting and propose the O-REPS algorithm which applies Online Mirror Descent over the space of occupancy measures, a key component adopted by (Rosenberg & Mansour, 2019a) and our work. O-REPS achieves the optimal regret $\tilde{\mathcal{O}}(L\sqrt{T\ln(|X||A|)})$ in this setting.

Known transition and bandit feedback. Several works consider the harder bandit feedback model while still assuming known transitions. The work (Neu et al., 2010) achieves regret $\tilde{\mathcal{O}}(L^2\sqrt{T|A|}/\alpha)$, assuming that all states are reachable with some probability α under all policies. Later, Neu et al. (2014) eliminates the dependence on α but only achieves $\tilde{\mathcal{O}}(T^{2/3})$ regret. The O-REPS algorithm of (Zimin & Neu, 2013) again achieves the optimal regret $\tilde{\mathcal{O}}(\sqrt{L|X||A|T})$. Another line of works (Arora et al., 2012; Dekel & Hazan, 2013) assumes deterministic transitions for a continuous setting without some unichain structure, which is known to be harder and suffers $\Omega(T^{2/3})$ regret (Dekel et al., 2014).

Unknown transition and full-information feedback. To deal with unknown transitions, Neu et al. (2012) propose the Follow the Perturbed Optimistic Policy algorithm and achieve $\tilde{\mathcal{O}}(L|X||A|\sqrt{T})$ regret. Combining the idea of confidence sets and Online Mirror Descent, the UC-O-REPS algorithm of (Rosenberg & Mansour, 2019a) improves the regret to $\tilde{\mathcal{O}}(L|X|\sqrt{|A|T})$. We note that this work also studies general convex performance criteria, which we do not consider.

Unknown transition and bandit feedback. This is the setting considered in our work. The only previous work we are aware of (Rosenberg & Mansour, 2019b) achieves a regret bound of $\tilde{\mathcal{O}}(T^{3/4})$, or $\tilde{\mathcal{O}}(\sqrt{T}/\alpha)$ under the strong assumption that under any policy, all states are reachable with probability α that could be arbitrarily small in gen-

[2] We warn the reader that in some of these cited papers, the notation $|X|$ or T might be defined differently (often L times smaller for $|X|$ and L times larger for T). We have translated the bounds based on Table 1 of (Jin et al., 2018) using our notation defined in Section 2.

eral. Our algorithm achieves $\tilde{\mathcal{O}}(\sqrt{T})$ regret without this assumption by using a different loss estimator and by using a tighter confidence set. We also note that the lower bound of $\Omega(L\sqrt{|X||A|T})$ (Jin et al., 2018) still applies.

Adversarial transition functions. There exist a few works that consider both time-varying transition functions and time-varying losses (Yu & Mannor, 2009; Cheung et al., 2019; Lykouris et al., 2019). Most recently, Lykouris et al. (2019) consider a stochastic problem with C episodes arbitrarily corrupted and obtain $\tilde{\mathcal{O}}(C\sqrt{T}+C^2)$ regret (ignoring dependence on other parameters). This bound is of order $\tilde{\mathcal{O}}(\sqrt{T})$ only when C is a constant, and is vacuous whenever $C=\Omega(\sqrt{T})$. In comparison, our bound is always $\tilde{\mathcal{O}}(\sqrt{T})$ no matter how much corruption there is in the losses, but our algorithm cannot handle changing transition functions.

2. Problem Formulation

An adversarial Markov decision process is defined by a tuple $(X, A, P, \{\ell_t\}_{t=1}^T)$, where X is the finite state space, A is the finite action space, $P : X \times A \times X \to [0,1]$ is the transition function, with $P(x'|x,a)$ being the probability of transferring to state x' when executing action a in state x, and $\ell_t : X \times A \to [0,1]$ is the loss function for episode t.

In this work, we consider an episodic setting with finite horizons and assume that the MDP has a layered structure, satisfying the following conditions:

- The state space X consists of $L+1$ layers $X_0, \ldots, X_L$ such that $X = \bigcup_{k=0}^{L} X_k$ and $X_i \cap X_j = \emptyset$ for $i \neq j$.
- X_0 and X_L are singletons, that is, $X_0 = \{x_0\}$ and $X_L = \{x_L\}$.
- Transitions are possible only between consecutive layers. In other words, if $P(x'|x,a) > 0$, then $x' \in X_{k+1}$ and $x \in X_k$ for some k.

These assumptions were made in previous work (Neu et al., 2012; Zimin & Neu, 2013; Rosenberg & Mansour, 2019a) as well. They are not necessary but greatly simplify notation and analysis. Such a setup is sometimes referred to as the loop-free stochastic shortest path problem in the literature. It is clear that this is a strict generalization of the episodic setting studied in (Azar et al., 2017; Jin et al., 2018) for example, where the number of states is the same for each layer (except for the first and the last one).[3] We also point out that our algorithms and results can be easily modified to deal with a more general setup where the first layer has multiple states and in each episode the initial state is decided adversarially, as in (Jin et al., 2018) (details omitted).

Protocol 1 Learner-Environment Interaction

Parameters: state space X and action space A (known to the learner), unknown transition function P
for $t = 1$ **to** T **do**
  adversary decides a loss function $\ell_t : X \times A \to [0,1]$
  learner decides a policy π_t and starts in state x_0
  for $k = 0$ **to** $L-1$ **do**
    learner selects action $a_k \sim \pi_t(\cdot|x_k)$
    learner observes loss $\ell_t(x_k, a_k)$
    environment draws a new state $x_{k+1} \sim P(\cdot|x_k, a_k)$
    learner observes state x_{k+1}
  end for
end for

The interaction between the learner and the environment is presented in Protocol 1. Ahead of time, the environment decides an MDP, and only the state space X with its layer structure and the action space A are known to the learner. The interaction proceeds in T episodes. In episode t, the adversary decides the loss function ℓ_t, which can depend on the learner's algorithm and the randomness before episode t. Simultaneously, the learner starts from state x_0 and decides a stochastic policy $\pi_t : X \times A \to [0,1]$, where $\pi_t(a|x)$ is the probability of taking action a at a given state x, so that $\sum_{a \in A} \pi_t(a|x) = 1$ for every state x. Then, the learner executes this policy in the MDP, generating L state-action pairs $(x_0, a_0), \ldots, (x_{L-1}, a_{L-1})$.[4] Specifically, for each $k = 0, \ldots, L-1$, action a_k is drawn from $\pi_t(\cdot|x_k)$ and the next state x_{k+1} is drawn from $P(\cdot|x_k, a_k)$.

Importantly, instead of observing the loss function ℓ_t at the end of episode t (Rosenberg & Mansour, 2019a), in our setting the learner only observes the loss for each visited state-action pair: $\ell_t(x_0, a_0), \ldots, \ell_t(x_{L-1}, a_{L-1})$. That is, we consider the more challenging setting with bandit feedback.

For any given policy π, we denote its expected loss in episode t by

$$\mathbb{E}\left[\sum_{k=0}^{L-1} \ell_t(x_k, a_k) \middle| P, \pi\right],$$

where the notation $\mathbb{E}[\cdot|P,\pi]$ emphasizes that the state-action pairs $(x_0, a_0), \ldots, (x_{L-1}, a_{L-1})$ are random variables generated according to the transition function P and a stochastic policy π. The total loss over T episodes for any fixed policy

[3] In addition, some of these works (such as (Azar et al., 2017)) also assume that the states have the same name for different layers, and the transition between the layers remains the same. Our setup does not make this assumption and is closer to that of (Jin et al., 2018). We also refer the reader to footnote 2 of (Jin et al., 2018) for how to translate regret bounds between settings with and without this extra assumption.

[4] Formally, the notation $(x_0, a_0), \ldots, (x_{L-1}, a_{L-1})$ should have a t dependence. Throughout the paper we omit this dependence for conciseness as it is clear from the context.

π is thus

$$L_T(\pi) = \sum_{t=1}^{T} \mathbb{E}\left[\sum_{k=0}^{L-1} \ell_t(x_k, a_k) \middle| P, \pi\right],$$

while the total loss of the learner is

$$L_T = \sum_{t=1}^{T} \mathbb{E}\left[\sum_{k=0}^{L-1} \ell_t(x_k, a_k) \middle| P, \pi_t\right].$$

The goal of the learner is to minimize the regret, defined as

$$R_T = L_T - \min_{\pi} L_T(\pi)$$

where π ranges over all stochastic policies.

Notation. We use $k(x)$ to denote the index of the layer to which state x belongs, and $\mathbb{I}\{\cdot\}$ to denote the indicator function whose value is 1 if the input holds true and 0 otherwise. Let $o_t = \{(x_k, a_k, \ell_t(x_k, a_k))\}_{k=0}^{L-1}$ be the observation of the learner in episode t, and $\mathcal{F}_t$ be the σ-algebra generated by $(o_1, \ldots, o_{t-1})$. Also let $\mathbb{E}_t[\cdot]$ be a shorthand of $\mathbb{E}[\cdot|\mathcal{F}_t]$.

2.1. Occupancy Measures

Solving the problem with techniques from online learning requires introducing the concept of *occupancy measures* (Altman, 1999; Neu et al., 2012). Specifically, the occupancy measure $q^{P,\pi} : X \times A \times X \to [0, 1]$ associated with a stochastic policy π and a transition function P is defined as

$$q^{P,\pi}(x, a, x') = \Pr\left[x_k = x, a_k = a, x_{k+1} = x' \mid P, \pi\right],$$

where $k = k(x)$ is the index of the layer to which x belongs. In other words, $q^{P,\pi}(x, a, x')$ is the marginal probability of encountering the triple (x, a, x') when executing policy π in a MDP with transition function P.

Clearly, an occupancy measure q satisfies the following two properties. First, due to the loop-free structure, each layer is visited exactly once and thus for every $k = 0, \ldots, L-1$,

$$\sum_{x \in X_k} \sum_{a \in A} \sum_{x' \in X_{k+1}} q(x, a, x') = 1. \tag{1}$$

Second, the probability of entering a state when coming from the previous layer is exactly the probability of leaving from that state to the next layer (except for x_0 and x_L). Therefore, for every $k = 1, \ldots, L-1$ and every state $x \in X_k$, we have

$$\sum_{x' \in X_{k-1}} \sum_{a \in A} q(x', a, x) = \sum_{x' \in X_{k+1}} \sum_{a \in A} q(x, a, x'). \tag{2}$$

It turns out that these two properties suffice for any function $q : X \times A \times A \to [0, 1]$ to be an occupancy measure associated with some transition function and some policy.

Lemma 1 (Rosenberg & Mansour (2019a)). *If a function $q : X \times A \times X \to [0, 1]$ satisfies conditions (1) and (2), then it is a valid occupancy measure associated with the following induced transition function P^q and induced policy π^q:*

$$P^q(x'|x, a) = \frac{q(x, a, x')}{\sum_{y \in X_{k(x)+1}} q(x, a, y)},$$

$$\pi^q(a|x) = \frac{\sum_{x' \in X_{k(x)+1}} q(x, a, x')}{\sum_{b \in A} \sum_{x' \in X_{k(x)+1}} q(x, b, x')}.$$

We denote by Δ the set of valid occupancy measures, that is, the subset of $[0, 1]^{X \times A \times X}$ satisfying conditions (1) and (2). For a fixed transition function P, we denote by $\Delta(P) \subset \Delta$ the set of occupancy measures whose induced transition function P^q is exactly P. Similarly, we denote by $\Delta(\mathcal{P}) \subset \Delta$ the set of occupancy measures whose induced transition function P^q belongs to a set of transition functions $\mathcal{P}$.

With the concept of occupancy measure, we can reduce the problem of learning a policy to the problem of learning an occupancy measure and apply online linear optimization techniques. Specifically, with slight abuse of notation, for an occupancy measure q we define

$$q(x, a) = \sum_{x' \in X_{k(x)+1}} q(x, a, x')$$

for all $x \neq x_L$ and $a \in A$, which is the probability of visiting state-action pair (x, a). Then the expected loss of following a policy π for episode t can be rewritten as

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=0}^{L-1} \ell_t(x_k, a_k) \middle| P, \pi\right] \\
&= \sum_{k=0}^{L-1} \sum_{x \in X_k} \sum_{a \in A} q^{P,\pi}(x, a) \ell_t(x, a) \\
&= \sum_{x \in X \setminus \{x_L\}, a \in A} q^{P,\pi}(x, a) \ell_t(x, a) \triangleq \langle q^{P,\pi}, \ell_t \rangle,
\end{aligned}$$

and accordingly the regret of the learner can be rewritten as

$$R_T = L_T - \min_{\pi} L_T(\pi) = \sum_{t=1}^{T} \langle q^{P,\pi_t} - q^*, \ell_t \rangle, \tag{3}$$

where $q^* \in \operatorname{argmin}_{q \in \Delta(P)} \sum_{t=1}^{T} \langle q, \ell_t \rangle$ is the optimal occupancy measure in $\Delta(P)$.

On the other hand, assume for a moment that the set $\Delta(P)$ were known and the loss function ℓ_t was revealed at the end of episode t. Consider an online linear optimization problem (see (Hazan et al., 2016) for example) with decision set $\Delta(P)$ and linear loss parameterized by ℓ_t at time t. In

other words, at each time t, the learner proposes $q_t \in \Delta(P)$ and suffers loss $\langle q_t, \ell_t \rangle$. The regret of this problem is

$$\sum_{t=1}^{T} \langle q_t - q^*, \ell_t \rangle. \tag{4}$$

Therefore, if in the original problem, we set $\pi_t = \pi^{q_t}$, then the two regret measures Eq. (3) and Eq. (4) are exactly the same by Lemma 1 and we have thus reduced the problem to an instance of online linear optimization.

It remains to address the issues that $\Delta(P)$ is unknown and that we have only partial information on ℓ_t. The first issue can be addressed by constructing a confidence set $\mathcal{P}$ based on observations and replacing $\Delta(P)$ with $\Delta(\mathcal{P})$, and the second issue is addressed by constructing loss estimators with reasonably small bias and variance. For both issues, we propose new solutions compared to (Rosenberg & Mansour, 2019b).

Note that importantly, the above reduction does not reduce the problem to an instance of the well-studied bandit linear optimization (Abernethy et al., 2008) where the quantity $\langle q_t, \ell_t \rangle$ (or a sample with this mean) is observed. Indeed, roughly speaking, what we observed in our setting are samples with mean $\langle q^{P,\pi^{q_t}}, \ell_t \rangle$. These two are different when we do not know P and have to operate over the set $\Delta(\mathcal{P})$.

3. Algorithm

The complete pseudocode of our algorithm, UOB-REPS, is presented in Algorithm 2. The three key components of our algorithm are: 1) maintaining a confidence set of the transition function, 2) using Online Mirror Descent to update the occupancy measure, and 3) constructing loss estimators, each described in detail below.

3.1. Confidence Sets

The idea of maintaining a confidence set of the transition function P dates back to (Burnetas & Katehakis, 1997). Specifically, the algorithm maintains counters to record the number of visits of each state-action pair (x, a) and each state-action-state triple (x, a, x'). To reduce the computational complexity, a doubling epoch schedule is deployed, so that a new epoch starts whenever there exists a state-action whose counter is doubled compared to its initial value at the beginning of the epoch. For epoch $i > 1$, let $N_i(x, a)$ and $M_i(x'|x, a)$ be the initial values of the counters, that is, the total number of visits of pair (x, a) and triple (x, a, x') before epoch i. Then the empirical transition function for this epoch is defined as

$$\bar{P}_i(x'|x, a) = \frac{M_i(x'|x, a)}{\max\{1, N_i(x, a)\}}.$$

Most previous works (such as (Jaksch et al., 2010; Azar et al., 2017; Rosenberg & Mansour, 2019b)) construct a confidence set which includes all transition functions with bounded total variation compared to $\bar{P}_i(\cdot|x, a)$ for each (x, a) pair. However, to ensure lower bias for our loss estimators, we propose a tighter confidence set which includes all transition functions with bounded distance compared to $\bar{P}_i(x'|x, a)$ for *each triple* (x, a, x'). More specifically, the confidence set for epoch i is defined as[5]

$$\mathcal{P}_i = \Big\{ \widehat{P} : \Big| \widehat{P}(x'|x, a) - \bar{P}_i(x'|x, a) \Big| \le \epsilon_i(x'|x, a), \\ \forall (x, a, x') \in X_k \times A \times X_{k+1}, k = 0, \ldots, L-1 \Big\}, \tag{5}$$

where the confidence width $\epsilon_i(x'|x, a)$ is defined as

$$2\sqrt{\frac{\bar{P}_i(x'|x, a) \ln\left(\frac{T|X||A|}{\delta}\right)}{\max\{1, N_i(x, a) - 1\}}} + \frac{14 \ln\left(\frac{T|X||A|}{\delta}\right)}{3 \max\{1, N_i(x, a) - 1\}} \tag{6}$$

for some confidence parameter $\delta \in (0, 1)$. For the first epoch ($i = 1$), $\mathcal{P}_i$ is simply the set of all transition functions so that $\Delta(\mathcal{P}_i) = \Delta$.[6]

By the empirical Bernstein inequality and union bounds, one can show the following (see Appendix B.1 for the proof):

Lemma 2. *With probability at least* $1 - 4\delta$*, we have* $P \in \mathcal{P}_i$ *for all* i*.*

Moreover, ignoring constants one can further show that our confidence bound is strictly tighter than those used in (Rosenberg & Mansour, 2019a;b), which is important for getting our final regret bound (more discussions to follow in Section 4).

3.2. Online Mirror Descent (OMD)

The OMD component of our algorithm is the same as (Rosenberg & Mansour, 2019b). As discussed in Section 2.1, our problem is closely related to an online linear optimization problem over some occupancy measure space. In particular, our algorithm maintains an occupancy measure $\widehat{q}_t$ for episode t and executes the induced policy $\pi_t = \pi^{\widehat{q}_t}$. We apply Online Mirror Descent, a standard algorithmic framework to tackle online learning problems, to update the occupancy measure as

$$\widehat{q}_{t+1} = \operatorname*{argmin}_{q \in \Delta(\mathcal{P}_i)} \eta \langle q, \widehat{\ell}_t \rangle + D(q \,\|\, \widehat{q}_t) \tag{7}$$

[5] It is understood that in the definition of the confidence set (Eq. (5)), there is also an implicit constraint on $\widehat{P}(\cdot|x, a)$ being a valid distribution over the states in $X_{k(x)+1}$, for each (x, a) pair. This is omitted for conciseness.

[6] To represent $\mathcal{P}_1$ in the form of Eq. (5), one can simply let $\bar{P}_1(\cdot|x, a)$ be any distribution and $\epsilon_1(x'|x, a) = 1$.

Algorithm 2 Upper Occupancy Bound Relative Entropy Policy Search (UOB-REPS)

Input: state space X, action space A, episode number T, learning rate η, exploration parameter γ, and confidence parameter δ

Initialization:
Initialize epoch index $i = 1$ and confidence set $\mathcal{P}_1$ as the set of all transition functions.
For all $k = 0, \dots, L-1$ and all $(x, a, x') \in X_k \times A \times X_{k+1}$, initialize counters

$$N_0(x,a) = N_1(x,a) = M_0(x'|x,a) = M_1(x'|x,a) = 0$$

and occupancy measure

$$\widehat{q}_1(x,a,x') = \frac{1}{|X_k||A||X_{k+1}|}.$$

Initialize policy $\pi_1 = \pi^{\widehat{q}_1}$.

for $t = 1$ **to** T **do**
Execute policy π_t for L steps and obtain trajectory $x_k, a_k, \ell_t(x_k, a_k)$ for $k = 0, \dots, L-1$.
Compute upper occupancy bound for each k:

$$u_t(x_k, a_k) = \text{COMP-UOB}(\pi_t, x_k, a_k, \mathcal{P}_i).$$

Construct loss estimators for all (x, a):

$$\widehat{\ell}_t(x,a) = \frac{\ell_t(x,a)}{u_t(x,a) + \gamma} \mathbb{I}\{x_{k(x)} = x, a_{k(x)} = a\}.$$

Update counters: for each k,

$$N_i(x_k, a_k) \leftarrow N_i(x_k, a_k) + 1,$$
$$M_i(x_{k+1}|x_k, a_k) \leftarrow M_i(x_{k+1}|x_k, a_k) + 1.$$

if $\exists k,\ N_i(x_k, a_k) \geq \max\{1, 2N_{i-1}(x_k, a_k)\}$ **then**
Increase epoch index $i \leftarrow i + 1$.
Initialize new counters: for all (x, a, x'),

$$N_i(x,a) = N_{i-1}(x,a), M_i(x'|x,a) = M_{i-1}(x'|x,a).$$

Update confidence set $\mathcal{P}_i$ based on Eq. (5).
end if

Update occupancy measure (D defined in Eq. (8)):

$$\widehat{q}_{t+1} = \operatorname*{argmin}_{q \in \Delta(\mathcal{P}_i)} \eta \langle q, \widehat{\ell}_t \rangle + D(q \parallel \widehat{q}_t).$$

Update policy $\pi_{t+1} = \pi^{\widehat{q}_{t+1}}$.
end for

Algorithm 3 COMP-UOB

Input: a policy π_t, a state-action pair (x, a) and a confidence set $\mathcal{P}$ of the form

$$\left\{\widehat{P} : \left|\widehat{P}(x'|x,a) - \bar{P}(x'|x,a)\right| \leq \epsilon(x'|x,a),\ \forall (x,a,x')\right\}$$

Initialize: for all $\tilde{x} \in X_{k(x)}$, set $f(\tilde{x}) = \mathbb{I}\{\tilde{x} = x\}$.

for $k = k(x) - 1$ **to** 0 **do**
for all $\tilde{x} \in X_k$ **do**
Compute $f(\tilde{x})$ based on Eq. (10):

$$f(\tilde{x}) = \sum_{a \in A} \pi_t(a|\tilde{x}) \cdot \text{GREEDY}\left(f, \bar{P}(\cdot|\tilde{x}, a), \epsilon(\cdot|\tilde{x}, a)\right)$$

(see Appendix A.2 for the procedure GREEDY).
end for
end for
Return: $\pi_t(a|x) f(x_0)$.

where i is the index of the epoch to which episode $t+1$ belongs, $\eta > 0$ is some learning rate, $\widehat{\ell}_t$ is some loss estimator for ℓ_t, and $D(\cdot\|\cdot)$ is a Bregman divergence. Following (Rosenberg & Mansour, 2019a;b), we use the unnormalized KL-divergence as the Bregman divergence:

$$D(q \parallel q') = \sum_{x,a,x'} q(x,a,x') \ln \frac{q(x,a,x')}{q'(x,a,x')} - \sum_{x,a,x'} \left(q(x,a,x') - q'(x,a,x')\right). \tag{8}$$

Note that as pointed out earlier, ideally one would use $\Delta(P)$ as the constraint set in the OMD update, but since P is unknown, using $\Delta(\mathcal{P}_i)$ in place of it is a natural idea. Also note that the update can be implemented efficiently, similarly to Rosenberg & Mansour (2019a) (see Appendix A.1 for details).

3.3. Loss Estimators

A common technique to deal with partial information in adversarial online learning problems (such as adversarial multi-armed bandits (Auer et al., 2002b)) is to construct loss estimators based on observations. In particular, inverse importance-weighted estimators are widely applicable. For our problem, with a trajectory $x_0, a_0, \dots, x_{L-1}, a_{L-1}$ for episode t, a common importance-weighted estimator for $\ell_t(x,a)$ would be

$$\frac{\ell_t(x,a)}{q^{P,\pi_t}(x,a)} \mathbb{I}\left\{x_{k(x)} = x, a_{k(x)} = a\right\}.$$

Clearly this is an unbiased estimator for $\ell_t(x,a)$. Indeed, the conditional expectation $\mathbb{E}_t[\mathbb{I}\left\{x_{k(x)} = x, a_{k(x)} = a\right\}]$ is

exactly $q^{P,\pi_t}(x,a)$ since the latter is exactly the probability of visiting (x,a) when executing policy π_t in a MDP with transition function P.

The issue of this standard estimator is that we cannot compute $q^{P,\pi_t}(x,a)$ since P is unknown. To address this issue, Rosenberg & Mansour (2019b) directly use $\widehat{q}_t(x,a)$ in place of $q^{P,\pi_t}(x,a)$, leading to an estimator that could be either an overestimate or an underestimate, and they can only show $\tilde{\mathcal{O}}(T^{3/4})$ regret with this approach.

Instead, since we have a confidence set $\mathcal{P}_i$ that contains P with high probability (where i is the index of the epoch to which t belongs), we propose to replace $q^{P,\pi_t}(x,a)$ with an *upper occupancy bound* defined as

$$u_t(x,a) = \max_{\widehat{P}\in\mathcal{P}_i} q^{\widehat{P},\pi_t}(x,a),$$

that is, the largest possible probability of visiting (x,a) among all the plausible environments. In addition, we also adopt the idea of *implicit exploration* from (Neu, 2015) to further increase the denominator by some fixed amount $\gamma > 0$. Our final estimator for $\ell_t(x,a)$ is

$$\widehat{\ell}_t(x,a) = \frac{\ell_t(x,a)}{u_t(x,a)+\gamma}\mathbb{I}\left\{x_{k(x)} = x, a_{k(x)} = a\right\}.$$

The implicit exploration is important for several technical reasons such as obtaining a high probability regret bound, the key motivation of the work (Neu, 2015) for multi-armed bandits.

Clearly, $\widehat{\ell}_t(x,a)$ is a biased estimator and in particular is underestimating $\ell_t(x,a)$ with high probability (since by definition $q^{P,\pi_t}(x,a) \le u_t(x,a)$ if $P \in \mathcal{P}_i$). The idea of using underestimates for adversarial learning with bandit feedback can be seen as an optimism principle which encourages exploration, and appears in previous work such as (Allenberg et al., 2006; Neu, 2015) in different forms and for different purposes. A key part of our analysis is to show that the bias introduced by these estimators is reasonably small, which eventually leads to a better regret bound compared to (Rosenberg & Mansour, 2019b).

Computing upper occupancy bound efficiently. It remains to discuss how to compute $u_t(x,a)$ efficiently. First note that

$$u_t(x,a) = \pi_t(a|x) \max_{\widehat{P}\in\mathcal{P}_i} q^{\widehat{P},\pi_t}(x) \tag{9}$$

where once again we slightly abuse the notation and define $q(x) = \sum_{a'\in A} q(x,a')$ for any occupancy measure q, which is the marginal probability of visiting state x under the associated policy and transition function. Further define

$$f(\tilde{x}) = \max_{\widehat{P}\in\mathcal{P}_i} \Pr\left[x_{k(x)} = x \;\middle|\; x_{k(\tilde{x})} = \tilde{x}, \widehat{P}, \pi_t\right],$$

for any $\tilde{x}$ with $k(\tilde{x}) \le k(x)$, which is the maximum probability of visiting x starting from state $\tilde{x}$, under policy π_t and among all plausible transition functions in $\mathcal{P}_i$. Clearly one has $u_t(x,a) = \pi_t(a|x)f(x_0)$, and also $f(\tilde{x}) = \mathbb{I}\{\tilde{x} = x\}$ for all $\tilde{x}$ in the same layer as x. Moreover, since the confidence set $\mathcal{P}_i$ imposes an independent constraint on $\widehat{P}(\cdot|x,a)$ for each different pair (x,a), we have the following recursive relation:

$$f(\tilde{x}) = \sum_{a\in A}\pi_t(a|\tilde{x})\left(\max_{\widehat{P}(\cdot|\tilde{x},a)}\sum_{x'\in X_{k(\tilde{x})+1}}\widehat{P}(x'|\tilde{x},a)f(x')\right) \tag{10}$$

where the maximization is over the constraint that $\widehat{P}(\cdot|\tilde{x},a)$ is a valid distribution over $X_{k(\tilde{x})+1}$ and also

$$\left|\widehat{P}(x'|\tilde{x},a) - \bar{P}_i(x'|\tilde{x},a)\right| \le \epsilon_i(x'|\tilde{x},a), \forall x' \in X_{k(\tilde{x})+1}.$$

This optimization can be solved efficiently via a greedy approach after sorting the values of $f(x')$ for all $x' \in X_{k(\tilde{x})+1}$ (see Appendix A.2 for details). This suggests computing $u_t(x,a)$ via backward dynamic programming from layer $k(x)$ down to layer 0, detailed in Algorithm 3.

4. Analysis

In this section, we analyze the regret of our algorithm and prove the following theorem.

Theorem 3. *With probability at least* $1-9\delta$, UOB-REPS *with* $\eta = \gamma = \sqrt{\frac{L\ln(L|X||A|/\delta)}{T|X||A|}}$ *ensures:*

$$R_T = \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right).$$

The proof starts with decomposing the regret into four different terms. Specifically, by Eq. (3) the regret can be written as $R_T = \sum_{t=1}^T\langle q_t - q^*, \ell_t\rangle$ where we define $q_t = q^{P,\pi_t}$ and $q^* \in \operatorname{argmin}_{q\in\Delta(P)}\sum_{t=1}^T\langle q,\ell_t\rangle$. We then add and subtract three terms and decompose the regret as

$$R_T = \underbrace{\sum_{t=1}^T\langle q_t - \widehat{q}_t, \ell_t\rangle}_{\textsc{Error}} + \underbrace{\sum_{t=1}^T\left\langle \widehat{q}_t, \ell_t - \widehat{\ell}_t\right\rangle}_{\textsc{Bias}_1} + \underbrace{\sum_{t=1}^T\left\langle \widehat{q}_t - q^*, \widehat{\ell}_t\right\rangle}_{\textsc{Reg}} + \underbrace{\sum_{t=1}^T\left\langle q^*, \widehat{\ell}_t - \ell_t\right\rangle}_{\textsc{Bias}_2}.$$

Here, the first term ERROR measures the error of using $\widehat{q}_t$ to approximate q_t; the third term REG is the regret of the corresponding online linear optimization problem and is controlled by OMD; the second and the fourth terms $\textsc{Bias}_1$ and $\textsc{Bias}_2$ correspond to the bias of the loss estimators.

We bound ERROR and BIAS_1 in the rest of this section. Bounding REG and BIAS_2 is relatively standard and we defer the proofs to Appendix B.3. Combining all the bounds (specifically, Lemmas 5, 6, 12, and 14), applying a union bound, and plugging in the (optimal) values of η and γ prove Theorem 3.

Throughout the analysis we use i_t to denote the index of the epoch to which episode t belongs. Note that $\mathcal{P}_{i_t}$ and $\widehat{q}_t$ are both $\mathcal{F}_t$-measurable. We start by stating a key technical lemma which essentially describes how our new confidence set shrinks over time and is critical for bounding ERROR and BIAS_1 (see Appendix B.2 for the proof).

Lemma 4. *With probability at least $1-6\delta$, for any collection of transition functions $\{P_t^x\}_{x\in X}$ such that $P_t^x \in \mathcal{P}_{i_t}$ for all x, we have*

$$\sum_{t=1}^{T}\sum_{x\in X, a\in A}|q^{P_t^x,\pi_t}(x,a)-q_t(x,a)| = \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right).$$

Bounding ERROR. With the help of Lemma 4, we immediately obtain the following bound on ERROR.

Lemma 5. *With probability at least $1-6\delta$,* UOB-REPS *ensures* $\text{ERROR} = \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right)$.

Proof. Since all losses are in $[0,1]$, we have $\text{ERROR} \le \sum_{t=1}^{T}\sum_{x,a}|\widehat{q}_t(x,a)-q_t(x,a)| = \sum_{t=1}^{T}\sum_{x,a}|q^{P_t^x,\pi_t}(x,a)-q_t(x,a)|$, where we define $P_t^x = P^{\widehat{q}_t} \in \mathcal{P}_{i_t}$ for all x so that $\widehat{q}_t = q^{P_t,\pi_t}$ (by the definition of π_t and Lemma 1). Applying Lemma 4 finishes the proof. □

Note that in the proof above, we set P_t^x to be the same for all x. In fact, in this case our Lemma 4 is similar to (Rosenberg & Mansour, 2019a, Lemmas B.2 and B.3) and it also suffices to use their looser confidence bound. However, in the next application of Lemma 4 to bounding BIAS_1, it turns out to be critical to set P_t^x to be different for different x and also to use our tighter confidence bound.

Bounding BIAS_1. To bound the term $\text{BIAS}_1 = \sum_{t=1}^{T}\langle\widehat{q}_t, \ell_t - \widehat{\ell}_t\rangle$, we need to show that $\widehat{\ell}_t$ is not underestimating ℓ_t by too much, which, at a high-level, is also ensured due to the fact that the confidence set becomes more and more accurate for frequently visited state-action pairs.

Lemma 6. *With probability at least $1-7\delta$,* UOB-REPS *ensures*

$$\text{BIAS}_1 = \mathcal{O}\left(L|X|\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)} + \gamma|X||A|T\right).$$

Proof. First note that $\langle\widehat{q}_t, \widehat{\ell}_t\rangle$ is in $[0, L]$ because $P^{\widehat{q}_t} \in \mathcal{P}_{i_t}$ by the definition of $\widehat{q}_t$ and thus $\widehat{q}_t(x,a) \le u_t(x,a)$ by the definition of u_t, which implies

$$\sum_{x,a}\widehat{q}_t(x,a)\widehat{\ell}_t(x,a) \le \sum_{x,a}\mathbb{I}\{x_{k(x)}=x, a_{k(x)}=a\} = L.$$

Applying Azuma's inequality we thus have with probability at least $1-\delta$, $\sum_{t=1}^{T}\langle\widehat{q}_t, \mathbb{E}_t[\widehat{\ell}_t] - \widehat{\ell}_t\rangle \le L\sqrt{2T\ln\frac{1}{\delta}}$. Therefore, we can bound BIAS_1 by $\sum_{t=1}^{T}\langle\widehat{q}_t, \ell_t - \mathbb{E}_t[\widehat{\ell}_t]\rangle + L\sqrt{2T\ln\frac{1}{\delta}}$ under this event. We then focus on the term $\sum_t\langle\widehat{q}_t, \ell_t - \mathbb{E}_t[\widehat{\ell}_t]\rangle$ and rewrite it as (by the definition of $\widehat{\ell}_t$)

$$\begin{aligned}
&\sum_{t,x,a}\widehat{q}_t(x,a)\ell_t(x,a)\left(1-\frac{\mathbb{E}_t[\mathbb{I}\{x_{k(x)}=x, a_{k(x)}=a\}]}{u_t(x,a)+\gamma}\right)\\
&= \sum_{t,x,a}\widehat{q}_t(x,a)\ell_t(x,a)\left(1-\frac{q_t(x,a)}{u_t(x,a)+\gamma}\right)\\
&= \sum_{t,x,a}\frac{\widehat{q}_t(x,a)}{u_t(x,a)+\gamma}\left(u_t(x,a)-q_t(x,a)+\gamma\right)\\
&\le \sum_{t,x,a}|u_t(x,a)-q_t(x,a)| + \gamma|X||A|T
\end{aligned}$$

where the last step is again due to $\widehat{q}_t(x,a) \le u_t(x,a)$. Finally, note that by Eq. (9), one has $u_t = q^{P_t^x,\pi_t}$ for $P_t^x = \text{argmax}_{\widehat{P}\in\mathcal{P}_{i_t}} q^{\widehat{P},\pi_t}(x)$ (which is $\mathcal{F}_t$-measurable and belongs to $\mathcal{P}_{i_t}$ clearly). Applying Lemma 4 together with a union bound then finishes the proof. □

We point out again that this is the only part that requires using our new confidence set. With the looser one used in previous work we can only show $\sum_{t,x,a}|u_t(x,a)-q_t(x,a)| = \mathcal{O}\left(L|X|^2\sqrt{|A|T\ln\left(\frac{T|X||A|}{\delta}\right)}\right)$, with an extra $|X|$ factor.

5. Conclusion

In this work, we propose the first efficient algorithm with $\tilde{\mathcal{O}}(\sqrt{T})$ regret for learning MDPs with unknown transition function, adversarial losses, and bandit feedback. Our main algorithmic contribution is to propose a tighter confidence bound together with a novel optimistic loss estimator based on upper occupancy bounds. One natural open problem in this direction is to close the gap between our regret upper bound $\tilde{\mathcal{O}}(L|X|\sqrt{|A|T})$ and the lower bound of $\Omega(L\sqrt{|X||A|T})$ (Jin et al., 2018), which exists even for the full-information setting.

Acknowledgments

HL is supported by NSF Awards IIS-1755781 and IIS-1943607. SS is partially supported by NSF-BIGDATA

Award IIS-1741341 and an NSF-CAREER grant Award IIS-1846088. TY is partially supported by NSF BIGDATA grant IIS-1741341.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pp. 2312–2320, 2011.

Abernethy, J. D., Hazan, E., and Rakhlin, A. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory*, pp. 263–274, 2008.

Allenberg, C., Auer, P., Györfi, L., and Ottucsák, G. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *Proceedings of the 17th international conference on Algorithmic Learning Theory*, pp. 229–243, 2006.

Altman, E. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

Arora, R., Dekel, O., and Tewari, A. Deterministic mdps with adversarial rewards and bandit feedback. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, pp. 93–101, 2012.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002a.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1), 2002b.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 263–272, 2017.

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 19–26, 2011.

Burnetas, A. N. and Katehakis, M. N. Optimal adaptive policies for markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. Reinforcement learning under drift. *arXiv preprint arXiv:1906.02922*, 2019.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pp. 208–214, 2011.

Dekel, O. and Hazan, E. Better rates for any adversarial deterministic mdp. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 675–683, 2013.

Dekel, O., Ding, J., Koren, T., and Peres, Y. Bandits with switching costs: $T^{2/3}$ regret. In *Proceedings of the 46th annual ACM symposium on Theory of computing*, pp. 459–467, 2014.

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

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1578–1586, 2018.

Hazan, E. et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 4868–4878, 2018.

Lykouris, T., Simchowitz, M., Slivkins, A., and Sun, W. Corruption robust exploration in episodic reinforcement learning. *arXiv preprint arXiv:1911.08689*, 2019.

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Neu, G. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 3168–3176, 2015.

Neu, G., György, A., and Szepesvári, C. The online loop-free stochastic shortest-path problem. In *Proceedings of the 23rd Annual Conference on Learning Theory*, pp. 231–243, 2010.

Neu, G., Gyorgy, A., and Szepesvari, C. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, pp. 805–813, 2012.

Neu, G., Antos, A., György, A., and Szepesvári, C. Online markov decision processes under bandit feedback. *IEEE Transactions on Automatic Control*, pp. 676 – 691, 2014.

Ouyang, Y., Gagrani, M., Nayyar, A., and Jain, R. Learning unknown markov decision processes: a thompson sampling approach. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 1333–1342, 2017.

Rosenberg, A. and Mansour, Y. Online convex optimization in adversarial Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5478–5486, 2019a.

Rosenberg, A. and Mansour, Y. Online stochastic shortest path with bandit feedback and unknown transition function. In *Advances in Neural Information Processing Systems*, 2019b.

Wang, Y., Dong, K., Chen, X., and Wang, L. Q-learning with ucb exploration is sample efficient for infinite-horizon mdp. In *International Conference on Learning Representations*, 2019.

Wei, C.-Y., Jafarnia-Jahromi, M., Luo, H., Sharma, H., and Jain, R. Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. *arXiv preprint arXiv:1910.07072*, 2019.

Yu, J. Y. and Mannor, S. Arbitrarily modulated markov decision processes. In *Proceedings of the 48h IEEE Conference on Decision and Control*, pp. 2946–2953, 2009.

Yu, J. Y., Mannor, S., and Shimkin, N. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research*, 34(3):737–757, 2009.

Zhang, Z. and Ji, X. Regret minimization for reinforcement learning by evaluating the optimal bias function. In *Advances in Neural Information Processing Systems*, 2019.

Zimin, A. and Neu, G. Online learning in episodic markovian decision processes by relative entropy policy search. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pp. 1583–1591, 2013.