Model-free Reinforcement Learning in Infinite-horizon Average-reward Markov Decision Processes

Chen-Yu Wei [1] **Mehdi Jafarnia-Jahromi** [1] **Haipeng Luo** [1] **Hiteshi Sharma** [1] **Rahul Jain** [1]

Abstract

Model-free reinforcement learning is known to be memory and computation efficient and more amendable to large scale problems. In this paper, two model-free algorithms are introduced for learning *infinite-horizon average-reward* Markov Decision Processes (MDPs). The first algorithm reduces the problem to the discounted-reward version and achieves $\mathcal{O}(T^{2/3})$ regret after T steps, under the minimal assumption of weakly communicating MDPs. To our knowledge, this is the first model-free algorithm for general MDPs in this setting. The second algorithm makes use of recent advances in adaptive algorithms for adversarial multi-armed bandits and improves the regret to $\mathcal{O}(\sqrt{T})$, albeit with a stronger ergodic assumption. This result significantly improves over the $\mathcal{O}(T^{3/4})$ regret achieved by the only existing model-free algorithm by Abbasi-Yadkori et al. (2019a) for ergodic MDPs in the infinite-horizon average-reward setting.

1. Introduction

Reinforcement learning (RL) refers to the problem of an agent interacting with an unknown environment with the goal of maximizing its cumulative reward through time. The environment is usually modeled as a Markov Decision Process (MDP) with an unknown transition kernel and/or an unknown reward function. The fundamental trade-off between exploration and exploitation is the key challenge for RL: should the agent exploit the available information to optimize the immediate performance, or should it explore the poorly understood states and actions to gather more information to improve future performance?

There are two broad classes of RL algorithms: *model-based* and *model-free*. Model-based algorithms maintain an estimate of the underlying MDP and use that to determine a policy during the learning process. Examples include UCRL2 (Jaksch et al., 2010), REGAL (Bartlett & Tewari, 2009), PSRL (Ouyang et al., 2017b), SCAL (Fruit et al., 2018b), UCBVI (Azar et al., 2017), EBF (Zhang & Ji, 2019) and EULER (Zanette & Brunskill, 2019). Model-based algorithms are well-known for their sample efficiency. However, there are two general disadvantages of model-based algorithms: First, model-based algorithms require large memory to store the estimate of the model parameters. Second, it is hard to extend model-based approaches to non-parametric settings, e.g., continuous state MDPs.

Model-free algorithms, on the other hand, try to resolve these issues by directly maintaining an estimate of the optimal Q-value function or the optimal policy. Examples include Q-learning (Watkins, 1989), Delayed Q-learning (Strehl et al., 2006), TRPO (Schulman et al., 2015), DQN (Mnih et al., 2013), A3C (Mnih et al., 2016), and more. Model-free algorithms are not only computation and memory efficient, but also easier to be extended to large scale problems by incorporating function approximation.

It was believed that model-free algorithms are less sample-efficient compared to model-based algorithms. However, recently Jin et al. (2018) showed that (model-free) Q-learning algorithm with UCB exploration achieves a nearly-optimal regret bound, implying the possibility of designing algorithms with advantages of both model-free and model-based methods. Jin et al. (2018) addressed the problem for episodic finite-horizon MDPs. Following this work, Dong et al. (2019) extended the result to the infinite-horizon discounted-reward setting.

However, Q-learning based model-free algorithms with low regret for *infinite-horizon average-reward* MDPs, an equally heavily-studied setting in the RL literature, remains unknown. Designing such algorithms has proven to be rather challenging since the Q-value function estimate may grow unbounded over time and it is hard to control its magnitude in a way that guarantees efficient learning. Moreover, techniques such as backward induction in the finite-horizon setting or contraction mapping in the infinite-horizon discounted setting can not be applied to the infinite-horizon

[1]University of Southern California. Correspondence to: Chen-Yu Wei <chenyu.wei@usc.edu>, Mehdi Jafarnia-Jahromi <mjafarni@usc.edu>.

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

Table 1. Regret comparisons for RL algorithms in infinite-horizon average-reward MDPs with S states, A actions, and T steps. D is the diameter of the MDP, $\mathrm{sp}(v^*) \le D$ is the span of the optimal value function, $\mathbb{V}^\star_{s,a} := \mathrm{Var}_{s' \sim p(\cdot|s,a)}[v^*(s')] \le \mathrm{sp}(v^*)^2$ is the variance of the optimal value function, t_{mix} is the mixing time (Def 5.1), t_{hit} is the hitting time (Def 5.2), and $\rho \le t_{\text{hit}}$ is some distribution mismatch coefficient (Eq. (5)). For more concrete definition of these parameters, see Sections 3-5.

	Algorithm	**Regret**	**Comment**
Model-based	REGAL (Bartlett & Tewari, 2009)	$\widetilde{\mathcal{O}}(\mathrm{sp}(v^*)\sqrt{SAT})$	no efficient implementation
	UCRL2 (Jaksch et al., 2010)	$\widetilde{\mathcal{O}}(DS\sqrt{AT})$	-
	PSRL (Ouyang et al., 2017b)	$\widetilde{\mathcal{O}}(\mathrm{sp}(v^*)S\sqrt{AT})$	Bayesian regret
	OSP (Ortner, 2018)	$\widetilde{\mathcal{O}}(\sqrt{t_{\text{mix}}SAT})$	ergodic assumption and no efficient implementation
	SCAL (Fruit et al., 2018b)	$\widetilde{\mathcal{O}}(\mathrm{sp}(v^*)S\sqrt{AT})$	-
	KL-UCRL (Talebi & Maillard, 2018)	$\widetilde{\mathcal{O}}(\sqrt{S\sum_{s,a}\mathbb{V}^\star_{s,a}T})$	-
	UCRL2B (Fruit et al., 2019)	$\widetilde{\mathcal{O}}(S\sqrt{DAT})$	-
	EBF (Zhang & Ji, 2019)	$\widetilde{\mathcal{O}}(\sqrt{DSAT})$	no efficient implementation
Model-free	POLITEX(Abbasi-Yadkori et al., 2019a)	$t^3_{\text{mix}}t_{\text{hit}}\sqrt{SA}T^{\frac{3}{4}}$	ergodic assumption
	Optimistic Q-learning (**this work**)	$\widetilde{\mathcal{O}}(\mathrm{sp}(v^*)(SA)^{\frac{1}{3}}T^{\frac{2}{3}})$	-
	MDP-OOMD (**this work**)	$\widetilde{\mathcal{O}}(\sqrt{t^3_{\text{mix}}\rho AT})$	ergodic assumption
	lower bound (Jaksch et al., 2010)	$\Omega(\sqrt{DSAT})$	-

average-reward setting.

In this paper, we make significant progress in this direction and propose two model-free algorithms for learning infinite-horizon average-reward MDPs. The first algorithm, Optimistic Q-learning (Section 4), achieves a regret bound of $\widetilde{\mathcal{O}}(T^{2/3})$ with high probability for the broad class of weakly communicating MDPs.[1] This is the first model-free algorithm in this setting under only the minimal weakly communicating assumption. The key idea of this algorithm is to artificially introduce a discount factor for the reward, to avoid the aforementioned unbounded Q-value estimate issue, and to trade-off this effect with the approximation introduced by the discount factor. We remark that this is very different from the R-learning algorithm of (Schwartz, 1993), which is a variant of Q-learning with no discount factor for the infinite-horizon average-reward setting.

The second algorithm, MDP-OOMD (Section 5), attains an improved regret bound of $\widetilde{\mathcal{O}}(\sqrt{T})$ for the more restricted class of *ergodic* MDPs. This algorithm maintains an instance of a multi-armed bandit algorithm at each state to learn the best action. Importantly, the multi-armed bandit algorithm needs to ensure several key properties to achieve our claimed regret bound, and to this end we make use of the recent advances for adaptive adversarial bandit algorithms from (Wei & Luo, 2018) in a novel way.

[1]Throughout the paper, we use the notation $\widetilde{\mathcal{O}}(\cdot)$ to suppress log terms.

To the best of our knowledge, the only existing model-free algorithm for this setting is the POLITEX algorithm (Abbasi-Yadkori et al., 2019a;b), which achieves $\widetilde{\mathcal{O}}(T^{3/4})$ regret for ergodic MDPs only. Both of our algorithms enjoy a better bound compared to POLITEX, and the first algorithm even removes the ergodic assumption completely.[2]

For comparisons with other existing model-based approaches for this problem, see Table 1. We also conduct experiments comparing our two algorithms. Details are deferred to Appendix D due to space constraints.

2. Related Work

We review the related literature with regret guarantees for learning MDPs with finite state and action spaces (there are many other works on asymptotic convergence or sample complexity, a different focus compared to our work). Three common settings have been studied: 1) finite-horizon episodic setting, 2) infinite-horizon discounted setting, and 3) infinite-horizon average-reward setting. For the first two settings, previous works have designed efficient algorithms with regret bound or sample complexity that is (almost) information-theoretically optimal, using either model-based approaches such as (Azar et al., 2017), or model-free approaches such as (Jin et al., 2018; Dong et al., 2019).

[2]POLITEX is studied in a more general setup with function approximation though. See the end of Section 5.1 for more comparisons.

For the infinite-horizon average-reward setting, many model-based algorithms have been proposed, such as (Auer & Ortner, 2007; Jaksch et al., 2010; Ouyang et al., 2017b; Agrawal & Jia, 2017; Talebi & Maillard, 2018; Fruit et al., 2018a;b). These algorithms either conduct *posterior sampling* or follow the *optimism in face of uncertainty* principle to build an MDP model estimate and then plan according to the estimate (hence model-based). They all achieve $\tilde{\mathcal{O}}(\sqrt{T})$ regret, but the dependence on other parameters are suboptimal. Recent works made progress toward obtaining the optimal bound (Ortner, 2018; Zhang & Ji, 2019); however, their algorithms are not computationally efficient – the time complexity scales exponentially in the number of states. On the other hand, except for the naive approach of combining Q-learning with ϵ-greedy exploration (which is known to suffer regret exponential in some parameters (Osband et al., 2014)), the only existing model-free algorithm for this setting is POLITEX, which only works for ergodic MDPs.

Two additional works are closely related to our second algorithm MDP-OOMD: (Neu et al., 2013) and (Wang, 2017). They all belong to *policy optimization* method where the learner tries to learn the parameter of the optimal policy directly. Their settings are quite different from ours and the results are not comparable. We defer more detailed comparisons with these two works to the end of Section 5.1.

3. Preliminaries

An infinite-horizon average-reward Markov Decision Process (MDP) can be described by $(\mathcal{S}, \mathcal{A}, r, p)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function and $p : \mathcal{S}^2 \times \mathcal{A} \to [0, 1]$ is the transition probability such that $p(s'|s, a) := \mathbb{P}(s_{t+1} = s' \mid s_t = s, a_t = a)$ for $s_t \in \mathcal{S}, a_t \in \mathcal{A}$ and $t = 1, 2, 3, \cdots$. We assume that $\mathcal{S}$ and $\mathcal{A}$ are finite sets with cardinalities S and A, respectively. The average reward per stage of a deterministic/stationary policy $\pi : \mathcal{S} \to \mathcal{A}$ starting from state s is defined as

$$J^\pi(s) := \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} r(s_t, \pi(s_t)) \;\middle|\; s_1 = s\right]$$

where s_{t+1} is drawn from $p(\cdot|s_t, \pi(s_t))$. Let $J^*(s) := \max_{\pi\in\mathcal{A}^S} J^\pi(s)$. A policy π^* is said to be optimal if it satisfies $J^{\pi^*}(s) = J^*(s)$ for all $s \in \mathcal{S}$.

We consider two standard classes of MDPs in this paper: (1) *weakly communicating* MDPs defined in Section 4 and (2) *ergodic* MDPs defined in Section 5. The weakly communicating assumption is weaker than the ergodic assumption, and is in fact known to be necessary for learning infinite-horizon MDPs with low regret (Bartlett & Tewari, 2009).

Standard MDP theory (Puterman, 2014) shows that for these two classes, there exist $q^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ (unique up to an additive constant) and unique $J^* \in [0, 1]$ such that $J^*(s) = J^*$ for all $s \in \mathcal{S}$ and the following Bellman equation holds:

$$J^* + q^*(s, a) = r(s, a) + \mathbb{E}_{s'\sim p(\cdot|s,a)}[v^*(s')], \quad (1)$$

where $v^*(s) := \max_{a\in\mathcal{A}} q^*(s, a)$. The optimal policy is then obtained by $\pi^*(s) = \operatorname{argmax}_a q^*(s, a)$.

We consider a learning problem where $\mathcal{S}, \mathcal{A}$ and the reward function r are known to the agent, but not the transition probability p (so one cannot directly solve the Bellman equation). The knowledge of the reward function is a typical assumption as in (Bartlett & Tewari, 2009; Gopalan & Mannor, 2015; Ouyang et al., 2017b), and can be removed at the expense of a constant factor for the regret bound.

Specifically, the learning protocol is as follows. An agent starts at an arbitrary state $s_1 \in \mathcal{S}$. At each time step $t = 1, 2, 3, \cdots$, the agent observes state $s_t \in \mathcal{S}$ and takes action $a_t \in \mathcal{A}$ which is a function of the history $s_1, a_1, s_2, a_2, \cdots, s_{t-1}, a_{t-1}, s_t$. The environment then determines the next state by drawing s_{t+1} according to $p(\cdot|s_t, a_t)$. The performance of a learning algorithm is evaluated through the notion of *cumulative regret*, defined as the difference between the total reward of the optimal policy and that of the algorithm:

$$R_T := \sum_{t=1}^{T} \Big(J^* - r(s_t, a_t)\Big).$$

Since $r \in [0, 1]$ (and subsequently $J^* \in [0, 1]$), the regret can at worst grow linearly with T. If a learning algorithm achieves sub-linear regret, then R_T/T goes to zero, i.e., the average reward of the algorithm converges to the optimal per stage reward J^*. The best existing regret bound is $\tilde{\mathcal{O}}(\sqrt{DSAT})$ achieved by a model-based algorithm (Zhang & Ji, 2019) (where D is the diameter of the MDP) and it matches the lower bound of (Jaksch et al., 2010).

Throughout the paper, we assume that T is known. When it is unknown, one can simply apply the standard doubling trick to obtain the same regret bound up to a constant (see e.g., (Shalev-Shwartz, 2011, Section 2.3.1)).

4. Optimistic Q-Learning

In this section, we introduce our first algorithm, OPTIMISTIC Q-LEARNING (see Algorithm 1 for pseudocode). The algorithm works for any weakly communicating MDPs. An MDP is weakly communicating if its state space $\mathcal{S}$ can be partitioned into two subsets: in the first subset, all states are transient under any stationary policy; in the second subset, every two states are accessible from each other under some stationary policy. It is well-known that the weakly communicating condition is necessary for ensuring low regret in this setting (Bartlett & Tewari, 2009).

Algorithm 1 OPTIMISTIC Q-LEARNING

Parameters: $H \geq 2$, confidence level $\delta \in (0,1)$
Initialization: $\gamma = 1 - \frac{1}{H}, \quad \forall s : \hat{V}_1(s) = H$
$\forall s, a : Q_1(s,a) = \hat{Q}_1(s,a) = H, \quad n_1(s,a) = 0$
Define: $\forall \tau, \alpha_\tau = \frac{H+1}{H+\tau}, b_\tau = 4\,\mathrm{sp}(v^*)\sqrt{\frac{H}{\tau}\ln\frac{2T}{\delta}}$
for $t = 1, \ldots, T$ **do**
Take action

$$a_t = \mathrm{argmax}_{a \in \mathcal{A}}\, \hat{Q}_t(s_t, a). \tag{2}$$

Observe s_{t+1}.
Update:

$$\begin{aligned} n_{t+1}(s_t,a_t) &\leftarrow n_t(s_t,a_t) + 1 \\ \tau &\leftarrow n_{t+1}(s_t,a_t) \\ Q_{t+1}(s_t,a_t) &\leftarrow (1-\alpha_\tau) Q_t(s_t,a_t) \\ &\quad + \alpha_\tau \left[r(s_t,a_t) + \gamma \hat{V}_t(s_{t+1}) + b_\tau \right] \\ \hat{Q}_{t+1}(s_t,a_t) &\leftarrow \min\left\{ \hat{Q}_t(s_t,a_t), Q_{t+1}(s_t,a_t) \right\} \\ \hat{V}_{t+1}(s_t) &\leftarrow \max_{a \in \mathcal{A}} \hat{Q}_{t+1}(s_t,a). \end{aligned} \tag{3}$$

(All other entries of $n_{t+1}, Q_{t+1}, \hat{Q}_{t+1}, \hat{V}_{t+1}$ remain the same as those in $n_t, Q_t, \hat{Q}_t, \hat{V}_t$.)

Define $\mathrm{sp}(v^*) = \max_s v^*(s) - \min_s v^*(s)$ to be the span of the value function, which is known to be bounded for weakly communicating MDPs. In particular, it is bounded by the diameter of the MDP (see (Lattimore & Szepesvári, 2018, Lemma 38.1)). We assume that $\mathrm{sp}(v^*)$ is known and use it to set the parameters. However, in the case when it is unknown, we can replace $\mathrm{sp}(v^*)$ with any upper bound of it (e.g. the diameter) in both the algorithm and the analysis.

The key idea of Algorithm 1 is to solve the undiscounted problem via learning a discounted MDP (with the same states, actions, reward function, and transition), for some discount factor γ (defined in terms of a parameter H). Define V^* and Q^* to be the optimal value-function and Q-function of the discounted MDP, satisfying the Bellman equation:

$$\begin{aligned} \forall (s,a), \quad Q^*(s,a) &= r(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)}[V^*(s')] \\ \forall s, \quad V^*(s) &= \max_{a \in \mathcal{A}} Q^*(s,a). \end{aligned}$$

The way we learn this discounted MDP is essentially the same as the algorithm of Dong et al. (2019), which itself is based on the idea from (Jin et al., 2018). Specifically, the algorithm maintains an estimate $\hat{V}_t$ for the optimal value function V^* and $\hat{Q}_t$ for the optimal Q-function Q^*, which itself is a clipped version of another estimate Q_t. Each time the algorithm takes a greedy action with the maximum estimated Q value (Eq. (2)). After seeing the next state, the algorithm makes a stochastic update of Q_t based on the Bellman equation, importantly with an extra *bonus* term b_τ and a carefully chosen step size α_τ (Eq.(3)). Here, τ is the number of times the current state-action pair has been visited, and the bonus term b_τ scales as $\mathcal{O}(\sqrt{H/\tau})$, which encourages exploration since it shrinks every time a state-action pair is executed. The choice of the step size α_τ is also crucial as pointed out in (Jin et al., 2018) and determines a certain effective period of the history for the current update.

While the algorithmic idea is similar to (Dong et al., 2019), we emphasize that our analysis is different and novel:

- First, Dong et al. (2019) analyze the sample complexity of their algorithm while we analyze the regret.
- Second, we need to deal with the approximation effect due to the difference between the discounted MDP and the original undiscounted one (Lemma 2).
- Finally, part of our analysis improves over that of (Dong et al., 2019) (specifically our Lemma 3). Following the original analysis of (Dong et al., 2019) would lead to a worse bound here.

We now state the main regret guarantee of Algorithm 1.

Theorem 1. *If the MDP is weakly communicating, Algorithm 1 with* $H = \min\left\{ \sqrt{\frac{\mathrm{sp}(v^*)T}{SA}}, \left(\frac{T}{SA\ln\frac{4T}{\delta}}\right)^{\frac{1}{3}} \right\}$ *ensures that with probability at least* $1-\delta$, R_T *is of order*

$$\mathcal{O}\left(\sqrt{\mathrm{sp}(v^*)SAT} + \mathrm{sp}(v^*)\left(T^{\frac{2}{3}} \left(SA\ln\tfrac{T}{\delta}\right)^{\frac{1}{3}} + \sqrt{T\ln\tfrac{1}{\delta}} \right) \right).$$

Our regret bound scales as $\widetilde{\mathcal{O}}(T^{2/3})$ and is suboptimal compared to model-based approaches with $\widetilde{\mathcal{O}}(\sqrt{T})$ regret (such as UCRL2) that matches the information-theoretic lower bound (Jaksch et al., 2010). However, this is the first model-free algorithm with sub-linear regret (under only the weakly communicating condition), and how to achieve $\widetilde{\mathcal{O}}(\sqrt{T})$ regret via model-free algorithms remains unknown. Also note that our bound depends on $\mathrm{sp}(v^*)$ instead of the potentially much larger diameter of the MDP. To our knowledge, existing approaches that achieve $\mathrm{sp}(v^*)$ dependence are all model-based (Bartlett & Tewari, 2009; Ouyang et al., 2017b; Fruit et al., 2018b) and use very different arguments.

4.1. Proof sketch of Theorem 1

The proof starts by decomposing the regret as

$$\begin{aligned}R_T &= \sum_{t=1}^{T}(J^* - r(s_t,a_t)) \\ &= \sum_{t=1}^{T}(J^* - (1-\gamma)V^*(s_t)) \\ &\quad + \sum_{t=1}^{T}(V^*(s_t) - Q^*(s_t,a_t)) \\ &\quad + \sum_{t=1}^{T}(Q^*(s_t,a_t) - \gamma V^*(s_t) - r(s_t,a_t)).\end{aligned}$$

Each of these three terms are handled through Lemmas 2, 3 and 4 whose proofs are deferred to the appendix. Plugging in $\gamma = 1 - \frac{1}{H}$ and picking the optimal H finish the proof. One can see that the $\widetilde{\mathcal{O}}(T^{2/3})$ regret comes from the bound $\frac{T}{H}$ from the first term and the bound $\sqrt{HT}$ from the second.

Lemma 2. *The optimal value function V^* of the discounted MDP satisfies*

1. $|J^* - (1-\gamma)V^*(s)| \le (1-\gamma)\,\mathrm{sp}(v^*)$, $\forall s \in \mathcal{S}$,
2. $\mathrm{sp}(V^*) \le 2\,\mathrm{sp}(v^*)$.

This lemma shows that the difference between the optimal value in the discounted setting (scaled by $1-\gamma$) and that of the undiscounted setting is small as long as γ is close to 1. The proof is by combining the Bellman equation of the these two settings and direct calculations.

Lemma 3. *With probability at least $1-\delta$, we have*

$$\begin{aligned}&\sum_{t=1}^{T}(V^*(s_t) - Q^*(s_t,a_t)) \\ &\le 4HSA + 24\,\mathrm{sp}(v^*)\sqrt{HSAT\ln\tfrac{2T}{\delta}}.\end{aligned}$$

This lemma is one of our key technical contributions. To prove this lemma one can write

$$\begin{aligned}&\sum_{t=1}^{T}(V^*(s_t) - Q^*(s_t,a_t)) \\ &= \sum_{t=1}^{T}(V^*(s_t) - \hat{V}_t(s_t)) + \sum_{t=1}^{T}(\hat{Q}_t(s_t,a_t) - Q^*(s_t,a_t)),\end{aligned}$$

using the fact that $\hat{V}_t(s_t) = \hat{Q}_t(s_t,a_t)$ by the greedy policy. The main part of the proof is to show that the second summation can in fact be bounded as $\sum_{t=2}^{T+1}(\hat{V}_t(s_t) - V^*(s_t))$ plus a small sub-linear term, which cancels with the first summation.

Lemma 4. *With probability at least $1-\delta$,*

$$\begin{aligned}&\sum_{t=1}^{T}(Q^*(s_t,a_t) - \gamma V^*(s_t) - r(s_t,a_t)) \\ &\le 2\,\mathrm{sp}(v^*)\sqrt{2T\ln\tfrac{1}{\delta}} + 2\,\mathrm{sp}(v^*).\end{aligned}$$

This lemma is proven via Bellman equation for the discounted setting and Azuma's inequality.

5. $\tilde{\mathcal{O}}(\sqrt{T})$ Regret for Ergodic MDPs

In this section, we propose another model-free algorithm that achieves $\tilde{\mathcal{O}}(\sqrt{T})$ regret bound for *ergodic* MDPs, a sub-class of weakly communicating MDPs. An MDP is ergodic if for any stationary policy π, the induced Markov chain is *irreducible* and *aperiodic*. Learning ergodic MDPs is arguably easier than the general case because the MDP is explorative by itself. However, achieving $\tilde{\mathcal{O}}(\sqrt{T})$ regret bound in this case with model-free methods is still highly non-trivial and we are not aware of any such result in the literature. Below, we first introduce a few useful properties of ergodic MDPs, all of which can be found in (Puterman, 2014).

We use *randomized policies* in this approach. A randomized policy π maps every state s to a distribution over actions $\pi(\cdot|s) \in \Delta_A$, where $\Delta_A = \{x \in \mathbb{R}_+^A : \sum_a x(a) = 1\}$. In an ergodic MDP, any policy π induces a Markov chain with a unique stationary distribution $\mu^\pi \in \Delta_S$ satisfying $(\mu^\pi)^\top P^\pi = (\mu^\pi)^\top$, where $P^\pi \in \mathbb{R}^{S\times S}$ is the induced transition matrix defined as $P^\pi(s,s') = \sum_a \pi(a|s)p(s'|s,a)$. We denote the stationary distribution of the optimal policy π^* by μ^*.

For ergodic MDPs, the long-term average reward J^π of any fixed policy π is independent of the starting state and can be written as $J^\pi = (\mu^\pi)^\top r^\pi$ where $r^\pi \in [0,1]^S$ is such that $r^\pi(s) := \sum_a \pi(a|s)r(s,a)$. For any policy π, the following Bellman equation has a solution $q^\pi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ that is unique up to an additive constant:

$$J^\pi + q^\pi(s,a) = r(s,a) + \mathbb{E}_{s'\sim p(\cdot|s,a)}[v^\pi(s')],$$

where $v^\pi(s) = \sum_a \pi(a|s)q^\pi(s,a)$. In this section, we impose an extra constraint: $\sum_s \mu^\pi(s)v^\pi(s) = 0$ so that q^π is indeed unique. In this case, it can be shown that v^π has the following form:

$$v^\pi(s) = \sum_{t=0}^{\infty}\left(\mathbf{e}_s^\top (P^\pi)^t - (\mu^\pi)^\top\right) r^\pi \tag{4}$$

where $\mathbf{e}_s$ is the basis vector with 1 in coordinate s.

Furthermore, ergodic MDPs have finite *mixing time* and *hitting time*, defined as follows.

Algorithm 2 MDP-OOMD

Define: episode length $B = 16 t_{\text{mix}} t_{\text{hit}} (\log_2 T)^2$ and number of episodes $K = T/B$
Initialize: $\pi_1'(a|s) = \pi_1(a|s) = \frac{1}{A}, \forall s, a.$
for $k = 1, 2, \ldots, K$ **do**
  for $t = (k-1)B + 1, \ldots, kB$ **do**
    Draw $a_t \sim \pi_k(\cdot|s_t)$ and observe s_{t+1}.
  Define trajectory $\mathcal{T}_k = (s_{(k-1)B+1}, a_{(k-1)B+1}, \ldots, s_{kB}, a_{kB})$.
  for *all* $s \in \mathcal{S}$ **do**
    $\widehat{\beta}_k(s, \cdot) = \textsc{EstimateQ}(\mathcal{T}_k, \pi_k, s)$.
    $(\pi_{k+1}'(\cdot|s), \pi_{k+1}(\cdot|s)) = \textsc{OomdUpdate}(\pi_k'(\cdot|s), \widehat{\beta}_k(s, \cdot))$.

Definition 5.1 ((Levin & Peres, 2017; Wang, 2017)). *The mixing time of an ergodic MDP is defined as*

$$t_{mix} := \max_{\pi} \min \left\{ t \geq 1 \;\middle|\; \|(P^\pi)^t(s, \cdot) - \mu^\pi\|_1 \leq \frac{1}{4}, \forall s \right\},$$

that is, the maximum time required for any policy starting at any initial state to make the state distribution $\frac{1}{4}$-close (in ℓ_1 norm) to the stationary distribution.

Definition 5.2. *The hitting time of an ergodic MDP is defined as*

$$t_{hit} := \max_{\pi} \max_{s} \frac{1}{\mu^\pi(s)},$$

that is, the maximum inverse stationary probability of visiting any state under any policy.

Our regret bound also depends on the following *distribution mismatch coefficient*:

$$\rho := \max_{\pi} \sum_{s} \frac{\mu^*(s)}{\mu^\pi(s)} \tag{5}$$

which has been used in previous work (Kakade & Langford, 2002; Agarwal et al., 2019). Clearly, one has $\rho \leq t_{\text{hit}} \sum_s \mu^*(s) = t_{\text{hit}}$. Note that these quantities are all parameters of the MDP only and are considered as finite constants compared to the horizon T. We thus assume that T is large enough so that t_{mix} and t_{hit} are both smaller than $T/4$. Also, we assume that these quantities are known to the algorithm.

5.1. Policy Optimization via Optimistic OMD

The key to get $\widetilde{\mathcal{O}}(\sqrt{T})$ bound is to learn the optimal policy π^* directly, by reducing the problem to solving an adversarial multi-armed bandit (MAB) (Auer et al., 2002) instance at each individual state.

The details of our algorithm MDP-OOMD is shown in Algorithm 2. It proceeds in episodes, and maintains an independent copy of a specific MAB algorithm for each state. At the beginning of episode k, each MAB algorithm outputs an action distribution $\pi_k(\cdot|s)$ for the corresponding state s, which together induces a policy π_k. The learner then executes policy π_k throughout episode k. At the end of the episode, for every state s we feed a reward estimator $\widehat{\beta}_k(s, \cdot) \in \mathbb{R}^A$ to the corresponding MAB algorithm, where $\widehat{\beta}_k$ is constructed using the samples collected in episode

Algorithm 3 EstimateQ

Input: $\mathcal{T}, \pi, s$

$\mathcal{T}$: a state-action trajectory from t_1 to t_2 $(s_{t_1}, a_{t_1}, \ldots, s_{t_2}, a_{t_2})$
π : a policy used to sample the trajectory $\mathcal{T}$
s : target state

Define: $N = 4 t_{\text{mix}} \log_2 T$ (window length minus 1)
Initialize: $\tau \leftarrow t_1$, $i \leftarrow 0$
while $\tau \leq t_2 - N$ **do**
  if $s_\tau = s$ **then**
    $i \leftarrow i + 1$
    Let $R = \sum_{t=\tau}^{\tau+N} r(s_t, a_t)$.
    Let $y_i(a) = \frac{R}{\pi(a|s)} \mathbf{1}[a_\tau = a], \forall a.$ $(y_i \in \mathbb{R}^A)$
    $\tau \leftarrow \tau + 2N$
  else
    $\tau \leftarrow \tau + 1$
if $i \neq 0$ **then**
  return $\frac{1}{i} \sum_{j=1}^{i} y_j$.
else
  return $\mathbf{0}$.

Algorithm 4 OomdUpdate

Input: $\pi' \in \Delta_A, \widehat{\beta} \in \mathbb{R}^A$
Define:
Regularizer $\psi(x) = \frac{1}{\eta} \sum_{a=1}^{A} \log \frac{1}{x(a)}$, for $x \in \mathbb{R}_+^A$
Bregman divergence associated with ψ:

$$D_\psi(x, x') = \psi(x) - \psi(x') - \langle \nabla \psi(x'), x - x' \rangle$$

Update:

$$\pi'_{next} = \operatorname*{argmax}_{\pi \in \Delta_A} \left\{ \langle \pi, \widehat{\beta} \rangle - D_\psi(\pi, \pi') \right\} \tag{6}$$

$$\pi_{next} = \operatorname*{argmax}_{\pi \in \Delta_A} \left\{ \langle \pi, \widehat{\beta} \rangle - D_\psi(\pi, \pi'_{next}) \right\} \tag{7}$$

return $(\pi'_{next}, \pi_{next})$.

k (see Algorithm 3). Finally all MAB algorithms update their distributions and output π_{k+1} for the next episode (Algorithm 4).

The reward estimator $\widehat{\beta}_k(s,\cdot)$ is an almost unbiased estimator for

$$\beta^{\pi_k}(s,\cdot) := q^{\pi_k}(s,\cdot) + NJ^{\pi_k} \tag{8}$$

with negligible bias (N is defined in Algorithm 3). The term NJ^{π_k} is the same for all actions and thus the corresponding MAB algorithm is trying to learn the best action at state s in terms of the average of Q-value functions $q^{\pi_1}(s,\cdot),\ldots,q^{\pi_K}(s,\cdot)$. To construct the reward estimator for state s, the sub-routine EstimateQ collects non-overlapping intervals of length $N+1=\widetilde{\mathcal{O}}(t_{\text{mix}})$ that start from state s, and use the standard inverse-propensity scoring to construct an estimator y_i for interval i (Line 5). In fact, to reduce the correlation among the non-overlapping intervals, we also make sure that these intervals are at least N steps apart from each other (Line 6). The final estimator $\widehat{\beta}_k(s,\cdot)$ is simply the average of all estimators y_i over these disjoint intervals. This averaging is important for reducing variance as explained later (see also Lemma 6).

The MAB algorithm we use is *optimistic online mirror descent* (OOMD) (Rakhlin & Sridharan, 2013) with *log-barrier* as the regularizer, analyzed in depth in (Wei & Luo, 2018). Here, optimism refers to something different from the optimistic exploration in Section 4. It corresponds to the fact that after a standard mirror descent update (Eq. (6)), the algorithm further makes a similar update using an optimistic prediction of the next reward vector, which in our case is simply the previous reward estimator (Eq. (7)). We refer the reader to (Wei & Luo, 2018) for more details, but point out that the optimistic prediction we use here is new.

It is clear that each MAB algorithm faces a non-stochastic problem (since π_k is changing over time) and thus it is important to deploy an adversarial MAB algorithm. The standard algorithm for adversarial MAB is Exp3 (Auer et al., 2002), which was also used for solving adversarial MDPs (Neu et al., 2013) (more comparisons with this to follow). However, there are several important reasons for our choice of the recently developed OOMD with log-barrier:

- First, the log-barrier regularizer produces a more exploratory distribution compared to Exp3 (as noticed in e.g. (Agarwal et al., 2017)), so we do not need an explicit exploration over the actions, which significantly simplifies the analysis compared to (Neu et al., 2013).
- Second, log-barrier regularizer provides more *stable* updates compared to Exp3 in the sense that $\pi_k(a|s)$ and $\pi_{k-1}(a|s)$ are within a multiplicative factor of each other (see Lemma 7). This implies that the corresponding policies and their Q-value functions are also stable, which is critical for our analysis.
- Finally, the optimistic prediction of OOMD, together with our particular reward estimator from EstimateQ, provides a variance reduction effect that leads to a better regret bound in terms of ρ instead of t_{hit}. See Lemma 8 and Lemma 9.

The regret guarantee of our algorithm is shown below.

Theorem 5. *For ergodic MDPs, with an appropriate chosen learning rate η for Algorithm 4, MDP-OOMD achieves*

$$\mathbb{E}[R_T] = \widetilde{\mathcal{O}}\left(\sqrt{t_{mix}^3 \rho A T}\right).$$

Note that in this bound, the dependence on the number of states S is hidden in ρ, since $\rho \geq \sum_s \frac{\mu^*(s)}{\mu^*(s)} = S$. Compared to the bound of Algorithm 1 or some other model-based algorithms such as UCRL2, this bound has an extra dependence on t_{mix}, a potentially large constant. As far as we know, all existing mirror-descent-based algorithms for the average-reward setting has the same issue (such as (Neu et al., 2013; Wang, 2017; Abbasi-Yadkori et al., 2019a)). The role of t_{mix} in our analysis is almost the same as that of $1/(1-\gamma)$ in the discounted setting (γ is the discount factor). Specifically, a small t_{mix} ensures 1) a short trajectory needed to approximate the Q-function with expected trajectory reward (in view of Eq. (12)) and 2) an upper bound for the magnitude of $q(s,a)$ and $v(s)$ (Lemma 14). For the discounted setting these are ensured by the discount factor already.

Comparisons. Neu et al. (2013) considered learning ergodic MDPs with *known* transition kernel and *adversarial* rewards, a setting incomparable to ours. Their algorithm maintains a copy of Exp3 for each state, but the reward estimators fed to these algorithms are constructed using the knowledge of the transition kernel and are very different from ours. They proved a regret bound of order $\widetilde{\mathcal{O}}\left(\sqrt{t_{\text{mix}}^3 t_{\text{hit}} A T}\right)$, which is worse than ours since $\rho \leq t_{\text{hit}}$.

In another recent work, (Wang, 2017) considered learning ergodic MDPs under the assumption that the learner is provided with a generative model (an oracle that takes in a state-action pair and output a sample of the next state). They derived a sample-complexity bound of order $\widetilde{\mathcal{O}}\left(\frac{t_{\text{mix}}^2 \tau^2 SA}{\epsilon^2}\right)$ for finding an ϵ-optimal policy, where $\tau = \max\left\{\max_s \left(\frac{\mu^*(s)}{1/S}\right)^2, \max_{s',\pi}\left(\frac{1/S}{\mu^\pi(s')}\right)^2\right\}$, which is at least $\max_\pi \max_{s,s'} \frac{\mu^*(s)}{\mu^\pi(s')}$ by AM-GM inequality. This result is again incomparable to ours, but we point out that our distribution mismatch coefficient ρ is always bounded by τS, while τ can be much larger than ρ on the other hand.

Finally, Abbasi-Yadkori et al. (2019a) considers a more general setting with function approximation, and their algorithm Politex maintains a copy of the standard exponential

weight algorithm for each state, very similar to (Neu et al., 2013). When specified to our tabular setting, one can verify (according to their Theorem 5.2) that POLITEX achieves $t_{\text{mix}}^3 t_{\text{hit}} \sqrt{SA} T^{\frac{3}{4}}$ regret, which is significantly worse than ours in terms of all parameters.

5.2. Proof sketch of Theorem 5

We first decompose the regret as follows:

$$\begin{aligned} R_T &= \sum_{t=1}^{T} J^* - r(s_t, a_t) \\ &= B \sum_{k=1}^{K} (J^* - J^{\pi_k}) + \sum_{k=1}^{K} \sum_{t \in \mathcal{I}_k} (J^{\pi_k} - r(s_t, a_t)), \quad (9) \end{aligned}$$

where $\mathcal{I}_k := \{(k-1)B + 1, \ldots, kB\}$ is the set of time steps for episode k. Using the *reward difference lemma* (Lemma 15 in the appendix), the first term of Eq. (9) can be written as

$$B \sum_s \mu^*(s) \left[\sum_{k=1}^{K} \sum_a (\pi^*(a|s) - \pi_k(a|s)) q^{\pi_k}(s, a) \right],$$

where the term in the square bracket can be recognized as exactly the regret of the MAB algorithm for state s and is analyzed in Lemma 8 of Section 5.3. Combining the regret of all MAB algorithms, Lemma 9 then shows that in expectation the first term of Eq. (9) is at most

$$\widetilde{\mathcal{O}} \left(\frac{BA}{\eta} + \frac{\eta T N^3 \rho}{B} + \eta^3 T N^6 \right). \quad (10)$$

On the other hand, the expectation of the second term in Eq.(9) can be further written as

$$\begin{aligned} &\mathbb{E}\left[\sum_{k=1}^{K} \sum_{t \in \mathcal{I}_k} (J^{\pi_k} - r(s_t, a_t)) \right] \\ &= \mathbb{E}\left[\sum_{k=1}^{K} \sum_{t \in \mathcal{I}_k} (\mathbb{E}_{s' \sim p(\cdot|s_t, a_t)}[v^{\pi_k}(s')] - q^{\pi_k}(s_t, a_t)) \right] \\ &\qquad \text{(Bellman equation)} \\ &= \mathbb{E}\left[\sum_{k=1}^{K} \sum_{t \in \mathcal{I}_k} (\mathbb{E}_{s' \sim p(\cdot|s_t, a_t)}[v^{\pi_k}(s')] - v^{\pi_k}(s_{t+1})) \right] \\ &\quad + \mathbb{E}\left[\sum_{k=1}^{K} \sum_{t \in \mathcal{I}_k} (v^{\pi_k}(s_t) - q^{\pi_k}(s_t, a_t)) \right] \\ &\quad + \mathbb{E}\left[\sum_{k=1}^{K} \sum_{t \in \mathcal{I}_k} (v^{\pi_k}(s_{t+1}) - v^{\pi_k}(s_t)) \right] \\ &= \mathbb{E}\left[\sum_{k=1}^{K} (v^{\pi_k}(s_{kB+1}) - v^{\pi_k}(s_{(k-1)B+1})) \right] \\ &\qquad \text{(the first two terms above are zero)} \\ &= \mathbb{E}\left[\sum_{k=1}^{K-1} (v^{\pi_k}(s_{kB+1}) - v^{\pi_{k+1}}(s_{kB+1})) \right] \\ &\quad + \mathbb{E}\left[v^{\pi_K}(s_{KB+1}) - v^{\pi_1}(s_1) \right]. \quad (11) \end{aligned}$$

The first term in the last expression can be bounded by $\mathcal{O}(\eta N^3 K) = \mathcal{O}(\eta N^3 T/B)$ due to the stability of OOMDUPDATE (Lemma 7) and the second term is at most $\mathcal{O}(t_{\text{mix}})$ according to Lemma 14 in the appendix.

Combining these facts with $N = \widetilde{\mathcal{O}}(t_{\text{mix}})$, $B = \widetilde{\mathcal{O}}(t_{\text{mix}} t_{\text{hit}})$, Eq. (9) and Eq. (10) and choosing the optimal η, we arrive at

$$\begin{aligned} \mathbb{E}[R_T] &= \widetilde{\mathcal{O}} \left(\frac{BA}{\eta} + \eta \frac{t_{\text{mix}}^3 \rho T}{B} + \eta^3 t_{\text{mix}}^6 T \right) \\ &= \widetilde{\mathcal{O}} \left(\sqrt{t_{\text{mix}}^3 \rho A T} + \left(t_{\text{mix}}^3 t_{\text{hit}} A \right)^{\frac{3}{4}} T^{\frac{1}{4}} + t_{\text{mix}}^2 t_{\text{hit}} A \right). \end{aligned}$$

5.3. Auxiliary Lemmas

To analyze the regret, we establish several useful lemmas, whose proofs can be found in the Appendix. First, we show that $\widehat{\beta}_k(s, a)$ is an almost unbiased estimator for $\beta^{\pi_k}(s, a)$.

Lemma 6. *Let $\mathbb{E}_k[x]$ denote the expectation of a random variable x conditioned on all history before episode k. Then for any k, s, a (recall β defined in Eq. (8)),*

$$\left| \mathbb{E}_k \left[\widehat{\beta}_k(s, a) \right] - \beta^{\pi_k}(s, a) \right| \le \mathcal{O}\left(\frac{1}{T} \right), \quad (12)$$

$$\mathbb{E}_k \left[\left(\widehat{\beta}_k(s, a) - \beta^{\pi_k}(s, a) \right)^2 \right] \le \mathcal{O}\left(\frac{N^3 \log T}{B \pi_k(a|s) \mu^{\pi_k}(s)} \right). \quad (13)$$

The next lemma shows that in OOMD, π_k and π_{k-1} are close in a strong sense, which further implies the stability for several other related quantities.

Lemma 7. *For any k, s, a,*

$$\begin{aligned}
|\pi_k(a|s) - \pi_{k-1}(a|s)| &\le \mathcal{O}(\eta N \pi_{k-1}(a|s)), \qquad (14)\\
|J^{\pi_k} - J^{\pi_{k-1}}| &\le \mathcal{O}(\eta N^2),\\
|v^{\pi_k}(s) - v^{\pi_{k-1}}(s)| &\le \mathcal{O}(\eta N^3),\\
|q^{\pi_k}(s,a) - q^{\pi_{k-1}}(s,a)| &\le \mathcal{O}(\eta N^3),\\
|\beta^{\pi_k}(s,a) - \beta^{\pi_{k-1}}(s,a)| &\le \mathcal{O}(\eta N^3).
\end{aligned}$$

The next lemma shows the regret bound of OOMD based on an analysis similar to (Wei & Luo, 2018).

Lemma 8. *For a specific state s, we have*

$$\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}(\pi^*(a|s) - \pi_k(a|s))\widehat{\beta}_k(s,a)\right] \le \mathcal{O}\left(\frac{A\ln T}{\eta} + \eta\mathbb{E}\left[\sum_{k=1}^{K}\sum_{a}\pi_k(a|s)^2\left(\widehat{\beta}_k(s,a) - \widehat{\beta}_{k-1}(s,a)\right)^2\right]\right),$$

where we define $\widehat{\beta}_0(s,a) = 0$ for all s and a.

Finally, we state a key lemma for proving Theorem 5.

Lemma 9. MDP-OOMD *ensures*

$$\mathbb{E}\left[B\sum_{k=1}^{K}\sum_{s}\sum_{a}\mu^*(s)\left(\pi^*(a|s) - \pi_k(a|s)\right)q^{\pi_k}(s,a)\right] = \mathcal{O}\left(\frac{BA\ln T}{\eta} + \eta\frac{TN^3\rho}{B} + \eta^3 TN^6\right).$$

6. Conclusions

In this work we propose two model-free algorithms for learning infinite-horizon average-reward MDPs. They are based on different ideas: one reduces the problem to the discounted version, while the other optimizes the policy directly via a novel application of adaptive adversarial multi-armed bandit algorithms. The main open question is how to achieve the information-theoretically optimal regret bound via a model-free algorithm, if it is possible at all. We believe that the techniques we develop in this work would be useful in answering this question.

We also remark that to run our algorithms, prior knowledge on parameters such as $\mathrm{sp}(v^*)$, t_{hit}, and t_{mix} (or their upper bounds) is required. In practice, they can be viewed as hyperparameters and tuned with standard techniques; in theory, this kind of assumption is made in almost all previous works on average-reward MDPs, except for some attempts in (Bartlett & Tewari, 2009) (unfortunately, their algorithm is not computationally tractable). Thus, how to learn an average-reward MDP without knowing the problem-dependent quantities still largely remains open.

Acknowledgements

The authors would like to thank Csaba Szepesvari for pointing out the related works (Abbasi-Yadkori et al., 2019a;b), Mengxiao Zhang for helping us prove Lemma 6, Gergely Neu for clarifying the analysis in (Neu et al., 2013), and Ronan Fruit for discussions on a related open problem presented at ALT 2019. Support from NSF for MJ (award ECCS-1810447), HL (awards IIS-1755781 and IIS-1943607), HS (award CCF-1817212) and RJ (awards ECCS-1810447 and CCF-1817212) is gratefully acknowledged.

References

Abbasi-Yadkori, Y., Bartlett, P., Bhatia, K., Lazic, N., Szepesvari, C., and Weisz, G. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pp. 3692–3702, 2019a.

Abbasi-Yadkori, Y., Lazic, N., Szepesvari, C., and Weisz, G. Exploration-enhanced politex. *arXiv preprint arXiv:1908.10479*, 2019b.

Agarwal, A., Luo, H., Neyshabur, B., and Schapire, R. E. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pp. 12–38, 2017.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in Markov decision processes. *arXiv preprint arXiv:1908.00261*, 2019.

Agrawal, S. and Jia, R. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pp. 1184–1194, 2017.

Auer, P. and Ortner, R. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 49–56, 2007.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 263–272. JMLR. org, 2017.

Bartlett, P. L. and Tewari, A. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 35–42. AUAI Press, 2009.

Bubeck, S., Li, Y., Luo, H., and Wei, C.-Y. Improved path-length regret bounds for bandits. In *Conference On Learning Theory*, 2019.

Chiang, C.-K., Yang, T., Lee, C.-J., Mahdavi, M., Lu, C.-J., Jin, R., and Zhu, S. Online optimization with gradual variations. In *Conference on Learning Theory*, pp. 6–1, 2012.

Dong, K., Wang, Y., Chen, X., and Wang, L. Q-learning with ucb exploration is sample efficient for infinite-horizon mdp. *arXiv preprint arXiv:1901.09311*, 2019.

Fruit, R., Pirotta, M., and Lazaric, A. Near optimal exploration-exploitation in non-communicating Markov decision processes. In *Advances in Neural Information Processing Systems*, pp. 2994–3004, 2018a.

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pp. 1573–1581, 2018b.

Fruit, R., Pirotta, M., and Lazaric, A. Improved analysis of ucrl2b, 2019. Available at rlgammazero.github.io/docs/ucrl2b_improved.pdf.

Gopalan, A. and Mannor, S. Thompson sampling for learning parameterized Markov decision processes. In *Conference on Learning Theory*, pp. 861–898, 2015.

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 *Advances in Neural Information Processing Systems*, pp. 4863–4873, 2018.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, 2002.

Lattimore, T. and Szepesvári, C. Bandit algorithms. *Cambridge University Press*, 2018.

Levin, D. A. and Peres, Y. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937, 2016.

Neu, G., György, A., Szepesvári, C., and Antos, A. Online Markov decision processes under bandit feedback. *IEEE Transactions on Automatic Control*, 59:676–691, 2013.

Ortner, R. Regret bounds for reinforcement learning via Markov chain concentration. *arXiv preprint arXiv:1808.01813*, 2018.

Osband, I., Van Roy, B., and Wen, Z. Generalization and exploration via randomized value functions. *arXiv preprint arXiv:1402.0635*, 2014.

Ouyang, Y., Gagrani, M., and Jain, R. Learning-based control of unknown linear systems with thompson sampling. *arXiv preprint arXiv:1709.04047*, 2017a.

Ouyang, Y., Gagrani, M., Nayyar, A., and Jain, R. Learning unknown Markov decision processes: A Thompson sampling approach. In *Advances in Neural Information Processing Systems*, pp. 1333–1342, 2017b.

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

Rakhlin, A. and Sridharan, K. Online learning with predictable sequences. In *Conference on Learning Theory*, pp. 993–1019, 2013.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897, 2015.

Schwartz, A. A reinforcement learning method for maximizing undiscounted rewards. In *Proceedings of the tenth international conference on machine learning*, volume 298, pp. 298–305, 1993.

Shalev-Shwartz, S. Online learning and online convex optimization. *Machine Learning*, 4(2):107–194, 2011.

Strehl, A. L. and Littman, M. L. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Strehl, A. L., Li, L., Wiewiora, E., Langford, J., and Littman, M. L. PAC model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pp. 881–888. ACM, 2006.

Talebi, M. S. and Maillard, O.-A. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *Algorithmic Learning Theory*, pp. 770–805, 2018.

Wang, M. Primal-dual π learning: Sample complexity and sublinear run time for ergodic Markov decision problems. *arXiv preprint arXiv:1710.06100*, 2017.

Watkins, C. J. C. H. *Learning from delayed rewards*. Phd Thesis, King's College, Cambridge, 1989.

Wei, C.-Y. and Luo, H. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pp. 1263–1291, 2018.

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, 2019.

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