Optimistic Policy Optimization with Bandit Feedback

Yonathan Efroni [*1] **Lior Shani** [*1] **Aviv Rosenberg** [2] **Shie Mannor** [1]

Abstract

Policy optimization methods are one of the most widely used classes of Reinforcement Learning (RL) algorithms. Yet, so far, such methods have been mostly analyzed from an optimization perspective, without addressing the problem of exploration, or by making strong assumptions on the interaction with the environment. In this paper we consider model-based RL in the tabular finite-horizon MDP setting with unknown transitions and bandit feedback. For this setting, we propose an optimistic policy optimization algorithm for which we establish $\tilde{O}(\sqrt{S^2AH^4K})$ regret for stochastic rewards. Furthermore, we prove $\tilde{O}(\sqrt{S^2AH^4}K^{2/3})$ regret for adversarial rewards. Interestingly, this result matches previous bounds derived for the bandit feedback case, yet with known transitions. To the best of our knowledge, the two results are the first sub-linear regret bounds obtained for policy optimization algorithms with unknown transitions and bandit feedback.

1. Introduction

Policy Optimization (PO) is among the most widely used methods in Reinforcement Learning (RL) (Peters & Schaal, 2006; 2008; Deisenroth & Rasmussen, 2011; Lillicrap et al., 2015; Levine et al., 2016; Gu et al., 2017). Unlike value-based approaches, e.g., Q-learning, these types of methods directly optimize the policy by incrementally changing it. Furthermore, PO methods span wide variety of popular algorithms such as policy-gradient algorithms (Sutton et al., 2000), natural policy gradient (Kakade, 2002), trust region policy optimization (TRPO) (Schulman et al., 2015) and soft actor-critic (Haarnoja et al., 2018).

[*]Equal contribution [1]Technion - Israel Institute of Technology, Haifa, Israel [2]Tel Aviv University, Tel Aviv, Israel. Correspondence to: Lior Shani <shanlior@gmail.com>, Yonathan Efroni <jonathan.efroni@gmail.com>.

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

Due to their popularity, there is a rich literature that provides different types of theoretical guarantees for different PO methods (Scherrer & Geist, 2014; Abbasi-Yadkori et al., 2019; Agarwal et al., 2019; Liu et al., 2019; Bhandari & Russo, 2019; Shani et al., 2019; Wei et al., 2019) for both the approximate and tabular settings. However, previous results, concerned with regret or PAC bounds for the RL setting when the model is unknown and only bandit feedback is given, provide guarantees which critically depend on 'concentrability coefficients' (Kakade & Langford, 2002; Munos, 2003; Scherrer, 2014) or on a unichain MDP assumption (Abbasi-Yadkori et al., 2019). However, these coefficients might be infinite and are usually small only for highly stochastic domains, while the unichain assumption is often very restrictive.

Recently, Cai et al. (2019) established an $\tilde{O}(\sqrt{K})$ regret bound for an optimistic PO method in the case of an unknown model and assuming full-information feedback on adversarially chosen instantaneous costs, where K is the number of episodes seen by the agent. In this work, we eliminate the full-information assumption on the cost, as in most practical settings only bandit feedback on the cost is given, i.e., the cost is observed through interacting with the environment. Specifically, we establish regret bounds for an optimistic PO method in the case of an unknown model and bandit feedback on the instantaneous cost in two regimes:

1. For stochastic cost, we establish an $\tilde{O}(\sqrt{S^2AH^4K})$ regret bound for a PO method (Section 6).

2. For adversarially chosen cost, we establish an $\tilde{O}(\sqrt{S^2AH^4}K^{2/3})$ regret bound for a PO method. The regret bound matches the $\tilde{O}\big(K^{2/3}\big)$ upper bound obtained by Neu et al. (2010a) for PO methods which have an access to the true model and observe bandit adversarial cost feedback (Section 7).

2. Preliminaries

Stochastic MDPs. A finite horizon stochastic Markov Decision Process (MDP) $\mathcal{M}$ is defined by a tuple $(\mathcal{S}, \mathcal{A}, H, \{p_h\}_{h=1}^H, \{c_h\}_{h=1}^H)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces with cardinality S and A, respectively, and $H \in \mathbb{N}$ is the horizon of the MDP. On time

Table 1. Comparison of our bounds with several state-of-the-art bounds for policy-based RL and occupancy measure RL in tabular finite-horizon MDPs. The time complexity of the algorithms is per episode; S and A are the sizes of the state and action sets, respectively; H is the horizon of the MDP; K is the total number of episodes; Env. describes the environment of the algorithm: stochastic (Sto) or adversarial (Adv); Policy based describes if an algorithm is based on policy updates or on occupancy measure updates. Costs and model terms describes how optimism is used in the estimators: For costs, a bonus term (Bonus) or an importance sampling estimator (IS). For transition model: a bonus term (Bonus) or a confidence interval of models (CI); The update procedure describes how the optimization problem is solved, using a state-wise closed-form solution (Closed form), or by solving an optimization problem over the entire state-action space (Optimization). The algorithms proposed in this paper are highlighted in gray. The other algorithms are OMD-BP (Neu et al., 2010b), UC-O-REPS (Rosenberg & Mansour, 2019a), OPPO (Cai et al., 2019) and UOB-REPS (Jin et al., 2019). (**) represents the different setting of the average cost criterion.

Algorithm	Regret	Env.	Bandit Feedback	Unknown Model	Policy Based	Costs	Model	Update Procedure
POMD	$\tilde{O}(\sqrt{S^2AH^4K})$	Sto.	✓	✓	✓	Bonus	Bonus	Closed form
OMDP-BP (**)	$\tilde{O}(K^{2/3})$	Adv.	✓	✗	✓	IS	-	Closed form
UC-O-REPS	$\tilde{O}(\sqrt{S^2AH^4K})$	Adv.	✗	✓	✗	-	CI	Optimization
OPPO	$\tilde{O}(\sqrt{S^3A^3H^4K})$	Adv.	✗	✓	✓	-	Bonus	Closed form
UOB-REPS	$\tilde{O}(\sqrt{S^2AH^4K})$	Adv.	✓	✓	✗	IS	CI	Optimization
POMD	$\tilde{O}(\sqrt{S^2AH^4K^{2/3}})$	Adv.	✓	✓	✓	IS	CI	Closed form

step h, and state s, the agent performs an action a, transitions to the next state s' according to a time-dependent transition function $p_h(s' \mid s, a)$, and suffers a random cost $C_h(s,a) \in [0,1]$ drawn i.i.d from a distribution with expectation $c_h(s,a)$.

A stochastic policy $\pi : \mathcal{S} \times [H] \to \Delta_A$ is a mapping from states and time-step indices to a distribution over actions, i.e., $\Delta_A = \{\pi \in \mathbb{R}^A : \sum_a \pi(a) = 1, \pi(a) \geq 0\}$. The performance of a policy π when starting from state s at time h is measured by its value function, which is defined as

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{h'=h}^{H} c_{h'}(s_{h'}, a_{h'}) \mid s_h = s, \pi, p\right], \qquad (2.1)$$

where the expectation is with respect to the randomness of the transition function, the cost function and the policy. The Q-function of a policy given the state action pair (s,a) at time-step h is defined by

$$Q_h^\pi(s,a) = \mathbb{E}\left[\sum_{h'=h}^{H} c_{h'}(s_{h'}, a_{h'}) \mid s_h = s, a_h = a, \pi, p\right]. \qquad (2.2)$$

The two satisfy the following relation:

$$\begin{aligned} Q_h^\pi(s,a) &= c_h(s,a) + p_h(\cdot \mid s,a) V_{h+1}^\pi, \\ V_h^\pi(s) &= \langle Q_h^\pi(s,\cdot), \pi_h(\cdot \mid s) \rangle, \end{aligned} \qquad (2.3)$$

with $p_h(\cdot \mid s,a)V = \sum_{s'} p_h(s' \mid s,a) V(s')$ for $V \in \mathbb{R}^S$, and $\langle \cdot, \cdot \rangle$ is the dot product.

An optimal policy π^* minimizes the value for all states s and time-steps h simultaneously (Puterman, 2014), and its corresponding optimal value is denoted by $V_h^*(s) = \min_\pi V_h^\pi(s)$, for all $h \in [H]$. We consider an agent that repeatedly interacts with an MDP in a sequence of K episodes such that the starting state at the k-th episode, s_1^k, is initialized by a fixed state s_1[*]. The agent does not have access to the model, and the costs are received by bandit feedback, i.e., the agent only observes the costs of encountered state-action pairs. At the beginning of the k-th episode, the agent chooses a policy π_k and samples a trajectory $\{s_h^k, a_h^k, C_h^k(s_h^k, a_h^k)\}_{h=1}^H$ by interacting with the stochastic MDP using this policy, where (s_h^k, a_h^k) are the state and action at the h-th time-step of the k-th episode. The performance of the agent for stochastic MDPs is measured by its *regret* relatively to the value of the optimal policy, defined as $\mathrm{Regret}(K') = \sum_{k=1}^{K'} V_1^{\pi_k}(s_1^k) - V_1^*(s_1^k)$ for all $K' \in [K]$, and π_k is the policy of the agent at the k-th episode.

Adversarial MDPs. Unlike stochastic MDP, in adversarial MDP, we let the cost to be determined by an adversary at the beginning of every episode, whereas the transition function is fixed. Thus, we denote the MDP at the k-th episode by $\mathcal{M}^k = (\mathcal{S}, \mathcal{A}, H, \{p_h\}_{h=1}^H, \{c_h^k\}_{h=1}^H)$. As in (2.1), (2.2), we define the value function and Q-function of a policy π at the k-th episode by

$$V_h^{k,\pi}(s) = \mathbb{E}\left[\sum_{h'=h}^{H} c_{h'}^k(s_{h'}, a_{h'}) \mid s_h = s, \pi, p\right],$$

$$Q_h^{k,\pi}(s,a) = \mathbb{E}\left[\sum_{h'=h}^{H} c_{h'}^k(s_{h'}, a_{h'}) \mid s_h = s, a_h = a, \pi, p\right].$$

[*] for simplicity we fix the initial state, but the results hold when it is drawn from a fixed distribution.

Notably, $V_h^{k,\pi}$ and $Q_h^{k,\pi}$ satisfy the relations in relation (2.3).

We consider an agent which repeatedly interacts with an adversarial MDP in a sequence of K episodes. Each episode starts from a fixed initial state, $s_1^k = s_1$. As in the stochastic case, at the beginning of the k-th episode, the agent chooses a policy π_k and samples a trajectory $\{s_h^k, a_h^k, c_h^k(s_h^k, a_h^k)\}_{h=1}^H$ by interacting with the adversarial MDP. In this case, the performance of the agent is measured by its *regret* relatively to the value of the best policy in hindsight. The objective is to minimize $\mathrm{Regret}(K') = \max_\pi \sum_{k=1}^{K'} V_1^{k,\pi_k}(s_1) - V_1^{k,\pi}(s_1)$ for all $K' \in [K]$.

Notations and Definitions. The filtration $\mathcal{F}_k$ includes all events (states, actions, and costs) until the end of the k-th episode, including the initial state of the $k+1$ episode. We denote by $n_h^k(s,a)$, the number of times that the agent has visited state-action pair (s,a) at the h-th step, and by $\bar{X}_k$, the empirical average of a random variable X. Both quantities are based on experience gathered until the end of the k^{th} episode and are $\mathcal{F}_k$ measurable. We also define the probability to visit the state-action pair (s,a) at the k-th episode at time-step h by $w_h^k(s,a) = \Pr\big(s_h^k = s, a_h^k = a \mid s_1^k, \pi_k, p\big)$. Since π_k is $\mathcal{F}_{k-1}$ measurable, so is $w_h^k(s,a)$. In what follows, we refer to $w_h^k(s,a)$ as the *state-action occupancy measure*. Furthermore, we define the state visitation frequency of a policy π in state s given a transition model p as $d_h^\pi(s;p) = \mathbb{E}[\mathbb{1}\{s_h = s\} \mid s_1, \pi, p]$. By the two definitions, it holds that $w_h^k(s,a) = d_h^{\pi_k}(s;p)\pi_h^k(a \mid s)$.

We use $\tilde{O}(X)$ to refer to a quantity that depends on X up to a poly-log expression of a quantity at most polynomial in S, A, K, H and δ^{-1}. Similarly, $\lesssim$ represents $\leq$ up to numerical constans or poly-log factors. We define $X \vee Y := \max\{X, Y\}$.

Mirror Descent. The mirror descent (MD) algorithm (Beck & Teboulle, 2003) is a proximal convex optimization method that minimizes a linear approximation of the objective together with a proximity term, defined in terms of a Bregman divergence between the old and new solution estimates. In our analysis we choose the Bregman divergence to be the Kullback–Leibler (KL) divergence, d_{KL}. If $\{f_k\}_{k=1}^K$ is a sequence of convex functions $f_k : \mathbb{R}^d \to \mathbb{R}$, and C is a constraints set, the k-th iterate of MD is the following:

$$x_{k+1} \in \arg\min_{x \in C} \{t_K \langle \nabla f_k(x_k), x - x_k \rangle + d_{KL}(x || x_k)\},$$

where t_K is a stepsize. In our case, C is the unit simplex Δ, and thus the optimization problem has a closed-form solution,

$$\forall i \in [d],\ x_{k+1}(i) = \frac{x_k(i)\exp(-t_K \nabla_i f_k(x_k))}{\sum_j x_k(j)\exp(-t_K \nabla_j f_k(x_k))}.$$

The MD algorithm ensures $\mathrm{Regret}(K') = \sum_{k=1}^{K'} f(x_k) - \min_x f(x) \in O(\sqrt{K})$ for all $K' \in [K]$.

3. Related Work

Approximate Policy Optimization: A large body of work addresses the convergence properties of policy optimization algorithms from an optimization perspective. In Kakade & Langford (2002), the authors analyzed the Conservative Policy Iteration (CPI) algorithm, an RL variant of the Frank-Wolfe algorithm (Scherrer & Geist, 2014; Vieillard et al., 2019), and showed it approximately converges to the global optimal solution. Recently, Liu et al. (2019) established the convergence of TRPO when neural networks are being used as the function approximators. Furthermore, Shani et al. (2019) showed that TRPO (Schulman et al., 2015) is in fact a natural RL adaptation of the MD algorithm, and established convergence guarantees. In (Agarwal et al., 2019), the authors obtained convergence results for policy gradient based algorithms. However, all of the aforementioned works rely on the strong assumption of a finite concentrability coefficient, i.e., $\max_{\pi,s,h} d_h^{\pi^*}(s;p)/d_h^\pi(s;p) < \infty$. This assumption bypasses the need to address exploration (Kakade & Langford, 2002), and leads to global guarantees based on the local nature of the policy gradients (Scherrer & Geist, 2014).

Mirror Descent in Adversarial Reinforcement Learning: There are two different methodologies for using MD updates in RL. The first and more practical one, is using MD-like updates directly on the policy. The second is based on optimizing over the space of state-action occupancy measures, that is, visitation frequencies for state-action pairs. An occupancy measure represents a policy implicitly. For convenience, previous results for regret minimization using MD approaches are summarized in Table 1.

Following the policy optimization approach, and assuming bandit feedback and known dynamics, Neu et al. (2010b) (OMDP-BF) established $\tilde{O}(K^{2/3})$ regret for the average reward criteria. Alternatively, by assuming full information on the reward functions, unknown dynamics and further assuming both the reward and transition dynamics are linear in some d-dimensional features, Cai et al. (2019) established $\tilde{O}(\sqrt{d^3H^4K})$ regret for their OPPO algorithm. The tabular case is a specific setting of the latter for $d = SA$.

Instead of directly optimizing the policy, Zimin & Neu (2013) proposed optimizing over the space of state-action occupancy measures with the Relative Entropy Policy Search

(O-REPS) algorithm. The O-REPS algorithm implicitly learns a policy by solving an MD optimization problem on the primal linear programming formulation of the MDP (Altman, 1999; Neu et al., 2017). Considering full information and unknown transitions, Rosenberg & Mansour (2019b) obtained $\tilde{O}(\sqrt{S^2AH^4K})$ regret for their UC-O-REPS algorithm. Later, Rosenberg & Mansour (2019a) extended the algorithm to bandit feedback and obtained a regret of $\tilde{O}(K^{3/4})$. Recently, by considering an optimistically biased importance sampling estimator, Jin et al. (2019) established $\tilde{O}(\sqrt{S^2AH^4K})$ for their UOB-REPS algorithm[†]. The O-REPS variants' updates constitute solving a convex optimization problem with HS^2A variables on each episode, instead of the closed form solution updates of the direct policy optimization variants.

Value-based Regret Minimization in Episodic RL: As opposed to Policy-based methods, there is an extensive literature about regret minimization in episodic MDPs using value-based methods. The works of (Azar et al., 2017; Dann et al., 2017; Jin et al., 2018; Zanette & Brunskill, 2019; Efroni et al., 2019) use the optimism in face of uncertainty principle to achieve near-optimal regret bounds. Jin et al. (2018) also establish a lower bound of $\Omega(\sqrt{SAH^3K})$.

4. Mirror Descent for MDPs

Algorithm 1 POMD with Known Model

Require: t_K, π_1 is the uniform policy.
  for $k = 1, .., K$ **do**
    # Policy Evaluation
    for $\forall h = H, H-1, .., 1$ **do**
      for $\forall s, a \in \mathcal{S} \times \mathcal{A}$ **do**
        $Q_h^{\pi_k}(s,a) = c_h(s,a) + p_h(\cdot \mid s,a) V_{h+1}^{\pi_k}$
      end for
    end for
    # Policy Improvement
    for $\forall s, a, h \in \mathcal{S} \times \mathcal{A} \times [H]$ **do**
      $\pi_h^{k+1}(a|s) = \frac{\pi_h^k(a|s)\exp\left(-t_K Q_h^{\pi_k}(s,a)\right)}{\sum_{a'} \pi_h^k(a'|s)\exp\left(-t_K Q_h^{\pi_k}(s,a')\right)}$
    end for
  end for

The empirical success of TRPO (Schulman et al., 2015) and SAC (Haarnoja et al., 2018) had motivated recent study of MD-like update rules for solving MDPs (Geist et al., 2019) when the model of the environment is known. Although not explicitly discussed in (Geist et al., 2019), such an algorithm can also provide guarantees – by similar proof technique – for the case where the cost function is adversarially chosen on each episode.

Policy Optimization by Mirror Descent (POMD) (see Algorithm 1) is conceptually similar to the Policy Iteration (PI) algorithm (Puterman, 2014). It alternates between two stages: (i) policy evaluation, and (ii) policy improvement. Furthermore, much alike PI, POMD updates its policy on the entire state space, given the evaluated Q-function. However, as oppose to PI, the policy improvement stage is 'soft'. Instead of updating according to the greedy policy, the algorithm applies soft update that keeps the next policy 'close' to the current one due to the KL-divergence term.

Similarly to standard analysis of the MD algorithm, Geist et al. (2019) established $\tilde{O}(\sqrt{K})$ bound on the regret of Algorithm 1. In the next sections, we apply the same approach to problems with unknown model and bandit feedback.

5. Extended Value Difference Lemma

The analysis of both stochastic and adversarial cases is built upon a central lemma which we now review. The lemma is a variant of (Cai et al., 2019)[Lemma 4.2], which generalizes classical value difference lemmas. Rewriting it in the following form, enables us to establish our results (proof in Appendix D).

Lemma 1 (Extended Value Difference)**.** *Let π, π' be two policies, and $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \{p_h\}_{h=1}^H, \{c_h\}_{h=1}^H)$ and $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, \{p'_h\}_{h=1}^H, \{c'_h\}_{h=1}^H)$ be two MDPs. Let $\hat{Q}_h^{\pi,\mathcal{M}}(s,a)$ be an approximation of the Q-function of policy π on the MDP $\mathcal{M}$ for all h, s, a, and let $\hat{V}_h^{\pi,\mathcal{M}}(s) = \left\langle \hat{Q}_h^{\pi,\mathcal{M}}(s,\cdot), \pi_h(\cdot \mid s) \right\rangle$. Then,*

$$\hat{V}_1^{\pi,\mathcal{M}}(s_1) - V_1^{\pi',\mathcal{M}'}(s_1) =$$
$$\sum_{h=1}^{H} \mathbb{E}\left[\left\langle \hat{Q}_h^{\pi,\mathcal{M}}(s_h,\cdot), \pi_h(\cdot \mid s_h) - \pi'_h(\cdot \mid s_h) \right\rangle \mid s_1, \pi', p' \right] +$$
$$\sum_{h=1}^{H} \mathbb{E}\left[\hat{Q}_h^{\pi,\mathcal{M}}(s_h,a_h) - c'_h(s_h,a_h) - p'_h(\cdot|s_h,a_h)\hat{V}_{h+1}^{\pi,\mathcal{M}} \mid s_1, \pi', p' \right]$$

where $V_1^{\pi',\mathcal{M}'}$ is the value function of π' in the MDP $\mathcal{M}'$.

This lemma generalizes existing value difference lemmas. For example, in (Kearns & Singh, 2002; Dann et al., 2017) the term $V_1^{\pi,\mathcal{M}}(s) - V_1^{\pi,\mathcal{M}'}(s)$ is analyzed, whereas in (Kakade & Langford, 2002) the term $V_1^{\pi,\mathcal{M}}(s) - V_1^{\pi',\mathcal{M}}(s)$ is analyzed. In next sections, we will demonstrate how Lemma 1 results in a simple analysis of the POMD algorithm. Importantly, the resulting regret bounds do not depend on concentrability coefficients (see Section 3) nor on any other structural assumptions.

[†]Note that in Jin et al. (2019), the regret of UOB-REPS is $\tilde{O}(\sqrt{S^2AH^2K})$. However, this is due to the loop-free assumption. To remove this assumption, one needs to multiply the number of states by a factor of H.

6. Policy Optimization in Stochastic MDPs

We are now ready to analyze the optimistic version of POMD for stochastic environments (see Algorithm 2). Instead of using the known model as in POMD, in Algorithm 2 we use the empirical model to estimate the Q-function of an empirical optimistic MDP, with the empirical transition function $\bar{p}$ and an optimistic cost function $\hat{c}$. The empirical transition function $\bar{p}$ and empirical cost function $\bar{c}$ are computed by averaging the observed transitions and costs, respectively, that is,

$$\bar{p}_h^k(s' \mid s,a) = \frac{\sum_{k'=1}^k \mathbb{1}\left(s_h^{k'}=s, a_h^{k'}=a, s_{h+1}^{k'}=s'\right)}{\sum_{k'=1}^k \mathbb{1}\left(s_h^{k'}=s, a_h^{k'}=a\right) \vee 1}$$
$$\bar{c}_h^k(s,a) = \frac{\sum_{k'=1}^k C_h^{k'}(s,a)\mathbb{1}\left(s_h^{k'}=s, a_h^{k'}=a\right)}{\sum_{k'=1}^k \mathbb{1}\left(s_h^{k'}=s, a_h^{k'}=a\right) \vee 1},$$

for every s, a, s', h, k.

Algorithm 2 Optimistic POMD for Stochastic MDPs

Require: t_K, π_1 is the uniform policy.
  for $k = 1, ..., K$ **do**
    Rollout a trajectory by acting π_k
    # Policy Evaluation
    $\forall s \in \mathcal{S},\ V_{H+1}^k(s) = 0$
    for $\forall h = H, .., 1$ **do**
      for $\forall s, a \in \mathcal{S} \times \mathcal{A}$ **do**
        $\hat{c}_h^{k-1}(s,a) = \bar{c}_h^{k-1}(s,a) - b_h^{k-1}(s,a)$, Eq. (6.1)
        $Q_h^k(s,a) = \hat{c}_h^{k-1}(s,a) + \bar{p}_h^{k-1}(\cdot|s,a)V_{h+1}^k$
        $Q_h^k(s,a) = \max\{Q_h^k(s,a), 0\}$
      end for
      for $\forall s \in \mathcal{S}$ **do**
        $V_h^k(s) = \langle Q_h^k(s,\cdot), \pi_h^k(\cdot \mid s)\rangle$
      end for
    end for
    # Policy Improvement
    for $\forall h, s, a \in [H] \times \mathcal{S} \times \mathcal{A}$ **do**
      $\pi_h^{k+1}(a|s) = \frac{\pi_h^k(a|s)\exp\left(-t_K Q_h^k(s,a)\right)}{\sum_{a'} \pi_h^k(a'|s)\exp\left(-t_K Q_h^k(s,a')\right)}$
    end for
    Update counters and empirical model, $n_k, \bar{c}^k, \bar{p}^k$
  end for

The optimistic cost function $\hat{c}$ is obtained by adding a bonus term which drives the algorithm to explore, i.e., $\hat{c}_h^{k-1}(s,a) = \bar{c}_h^{k-1}(s,a) - b_h^{k-1}(s,a)$, and we set

$$b_h^{k-1}(s,a) = b_h^{c,k-1}(s,a) + b_h^{p,k-1}(s,a). \tag{6.1}$$

The two bonus terms compensate on the lack of knowledge of the true costs and transition model, and are

$$b_h^{c,k-1}(s,a) = \tilde{O}\left(\frac{1}{\sqrt{n_h^{k-1}(s,a)}}\right),$$
$$b_h^{p,k-1}(s,a) = \tilde{O}\left(\frac{\sqrt{S}}{\sqrt{n_h^{k-1}(s,a)}}\right). \tag{6.2}$$

The following theorem bounds the regret of Algorithm 2. A full proof is found in Appendix B.2.

Theorem 1. *For any $K' \in [K]$, setting $t_K = \tilde{O}\left(H^{-1}K^{-1/2}\right)$ the regret of Algorithm 2 is bounded by*

$$\textit{Regret}(K') \le \tilde{O}\left(\sqrt{S^2AH^4K}\right).$$

Proof Sketch. We start by decomposing the regret into three terms according to Lemma 1, and then bound each term separately to get our final regret bound. For any π,

$$\begin{aligned}
\mathrm{Regret}(K') &= \sum_{k=1}^{K'} V_1^{\pi_k}(s_1^k) - V_1^{\pi}(s_1^k) \\
&= \sum_{k=1}^{K'} V_1^{\pi_k}(s_1^k) - V_1^k(s_1^k) + \sum_{k=1}^{K'} V_1^k(s_1^k) - V_1^{\pi}(s_1^k) \\
&= \underbrace{\sum_k V_1^{\pi_k}(s_1) - V_1^k(s_1)}_{(i)} \\
&+ \underbrace{\sum_{k,h} \mathbb{E}\left[\left\langle Q_h^k(s_h,\cdot), \pi_h^k(\cdot \mid s_h) - \pi_h(\cdot \mid s_h)\right\rangle \mid s_1, \pi, p\right]}_{(ii)} \\
&+ \underbrace{\sum_{k,h} \mathbb{E}\left[Q_h^k(s_h,a_h) - c_h(s_h,a_h) - p_h(\cdot|s_h,a_h)V_{h+1}^k \mid s_1, \pi, p\right]}_{(iii)}
\end{aligned}$$

Term (i): Bias of V^k. Term (i) is the bias between the estimated and true value of π_k, V^k and V^{π_k}, respectively. Applying Lemma 1, while using $\mathbb{E}[X(s_h,a_h) \mid s_1, \pi_k, p] = \mathbb{E}\left[X(s_h^k, a_h^k) \mid \mathcal{F}_{k-1}\right]$ for any $\mathcal{F}_{k-1}$-measurable function $X \in \mathbb{R}^{S\times A}$, we bound Term (i) by

$$\sum_{k,h} \mathbb{E}\left[\Delta c_h^{k-1}(s_h^k,a_h^k) + H\left\|\Delta p_h^{k-1}(\cdot|s_h^k,a_h^k)\right\|_1 \mid \mathcal{F}_{k-1}\right]$$
$$+ \sum_{k,h} \mathbb{E}\left[b_h^{c,k-1}(s_h^k,a_h^k) + b_h^{p,k-1}(s_h^k,a_h^k) \mid \mathcal{F}_{k-1}\right].$$

Here $\Delta c_h^{k-1}(s,a) = c_h(s,a) - \bar{c}_h^{k-1}(s,a)$ and $\Delta p_h^{k-1}(\cdot \mid s,a) = p_h(\cdot \mid s,a) - \bar{p}_h^{k-1}(\cdot \mid s,a)$, are the differences

between the true cost and transition model to the empirical cost and transition model. Applying Hoeffding's bound and L_1 deviation bound (Weissman et al., 2003) we get that w.h.p. for any s, a

$$\Delta c_h(s,a) \le \tilde{O}\left(\frac{1}{\sqrt{n_h^{k-1}(s,a)}}\right) = b_h^r(s,a),$$

$$\|\Delta p_h(\cdot \mid s,a)\|_1 \le \tilde{O}\left(\frac{\sqrt{S}}{\sqrt{n_h^{k-1}(s,a)}}\right) = b_h^p(s,a).$$

Thus, w.h.p., we get

$$(i) \lesssim \sum_{k=1}^{K'}\sum_{h=1}^{H}\mathbb{E}\left[\frac{H\sqrt{S}}{\sqrt{n_h^{k-1}(s_h^k,a_h^k)}} \mid \mathcal{F}_{k-1}\right],$$

which can be bounded by $\tilde{O}\left(\sqrt{S^2AH^4K}\right)$ using standard techniques (e.g., Dann et al. (2017)).

Term (ii): OMD Analysis. Term (ii) is the linear approximation used in MD optimization procedure. We bound it using an analysis of OMD. By applying usual OMD analysis (see Lemma 16) we have that for any policy π and s, h,

$$\sum_{k=1}^{K}\left\langle Q_h^k(\cdot \mid s), \pi_h^k(\cdot \mid s) - \pi_h(\cdot \mid s)\right\rangle \le \frac{\log A}{t_K} + \frac{t_K}{2}\sum_{k=1}^{K}\sum_{a}\pi_h^k(a \mid s)(Q_h^k(s,a))^2.$$

We plug this back to Term (ii) and use the fact that $0 \le Q_h^k(s,a) \le H$, to obtain

$$\text{Term (ii)} = \sum_{h=1}^{H}\mathbb{E}\left[\sum_{k=1}^{K'}\left\langle Q_h^k(s_h,\cdot), \pi_h^k(\cdot|s_h) - \pi_h(\cdot|s_h)\right\rangle \mid s_1,\pi,p\right] \le \frac{H\log A}{t_K} + \frac{t_K H^3 K}{2}.$$

By choosing $t_K = \sqrt{2\log A/(H^2K)}$, we obtain Term (ii) $\le \sqrt{2H^4K\log A}$.

Term (iii): Optimism. We choose our exploration bonuses in Eq. (6.2) such that Term (iii) is non-positive. Specifically, we choose the bonus such that $Q_h^k(s,a) - c_h(s,a) - p_h(\cdot|s,a)V_{h+1}^k \le 0$ for any s, a, which implies that Term(iii) ≤ 0. □

Remark 6.1. *The choice of the bonus term $b_h^{p,k}(s,a)$ is smaller than in (Cai et al., 2019) by a factor of $\sqrt{S}$. This translates to an improved regret bound by this factor, although (Cai et al., 2019) assumes full-information feedback on the cost function.*

Remark 6.2 (Bonus vs. Optimistic Model)**.** *Instead of using the additive exploration bonus b^p – which compensate on the lack of knowledge of transition model – one can use an optimistic model approach, as in UCRL2 (Jaksch et al., 2010). Following analogous analysis as of Theorem 1 one can establish the same guarantee $\tilde{O}(\sqrt{S^2AH^4K})$. However, the additive bonus approach results in an algorithm with reduced computational cost.*

Remark 6.3 (Optimism of POMD)**.** *Unlike value-based algorithms (e.g., Jaksch et al. (2010)) V^k, the value-function by which POMD improves upon, is not necessarily optimistic relatively to V^*. Instead, it is optimistic relatively to the value of π_k, i.e., $V^k \le V^{\pi_k}$.*

7. Policy Optimization in Adversarial MDPs

Algorithm 3 Optimistic POMD for Adversarial MDPs

Require: t_K, γ, π_1 is the uniform policy.
  for $k = 1, ..., K$ **do**
    Rollout a trajectory by acting π_k
    for all h, s **do**
      Compute $u_h^k(s)$ by $\pi_k, \mathcal{P}^{k-1}$, Eq. (7.1)
    end for
    # Policy Evaluation
    $\forall s \in \mathcal{S},\ V_{H+1}^k(s) = 0$
    for $\forall h = H, .., 1$ **do**
      for $\forall s, a \in \mathcal{S}\times\mathcal{A}$ **do**
        $\hat{c}_h^k(s,a) = \frac{c_h^k(s,a)\mathbb{1}\{s=s_h^k, a=a_h^k\}}{u_h^k(s)\pi_h^k(a|s)+\gamma}$
        $\hat{p}_h^k(\cdot|s,a) \in \underset{\hat{p}_h(\cdot|s,a)\in\mathcal{P}_h^{k-1}(s,a)}{\arg\min}\ \hat{p}_h(\cdot|s,a)V_{h+1}^k$
        $Q_h^k(s,a) = \hat{c}_h^k(s,a) + \hat{p}_h^k(\cdot|s,a)V_{h+1}^k$
      end for
      for $\forall s \in \mathcal{S}$ **do**
        $V_h^k(s) = \langle Q_h^k(s,\cdot), \pi_h^k(\cdot \mid s)\rangle$
      end for
    end for
    # Policy Improvement
    for $\forall h, s, a \in [H]\times\mathcal{S}\times\mathcal{A}$ **do**
      $\pi_h^{k+1}(a|s) = \frac{\pi_h^k(a|s)\exp\left(-t_K Q_h^k(s,a)\right)}{\sum_{a'}\pi_h^k(a'|s)\exp\left(-t_K Q_h^k(s,a')\right)}$
    end for
    Update counters and model, $n_k, \bar{p}^k$
  end for

In this section, we turn to analyze an optimistic version of POMD for adversarial environments (Algorithm 3). Similarly to the stochastic case, Algorithm 3 follows the POMD scheme, and alternates between policy evaluation, and, soft policy improvement, based on MD-like updates.

Unlike POMD for stochastic environments, the policy evaluation stage of Algorithm 3 uses different estimates of the instantaneous cost and model. The instantaneous cost is evaluated by a biased importance-sampling estimator, originally suggested by (Neu, 2015), and recently generalized to adversarial RL settings by (Jin et al., 2019),

$$\hat{c}_h^k(s,a) = \frac{c_h^k(s,a)\mathbb{1}\{s=s_h^k, a=a_h^k\}}{u_h^k(s)\pi_h^k(a\mid s)+\gamma},$$
$$\text{where } u_h^k(s) = \max_{\hat{p}\in\mathcal{P}^{k-1}} d_h^{\pi_k}(s;\hat{p}). \tag{7.1}$$

Here $\mathcal{P}^{k-1}$ is the set of transition functions obtained by using confidence intervals around the empirical model (see Appendix C.1.2).

In Algorithm 3 of Jin et al. (2019), the authors suggest a computationally efficient dynamic programming based approach for calculating $u_h^k(s)$ for all h, s. The motivation for such an estimate lies in the EXP3 algorithm (Auer et al., 2002) for adversarial bandits, which uses an unbiased importance-sampling estimator $\hat{c}(a) = \frac{c^k(a)\mathbb{1}\{a=a^k\}}{\pi^k(a)}$. Later, Neu (2015) showed that an optimistically biased estimator $\hat{c}(a) = \frac{c^k(a)\mathbb{1}\{a=a^k\}}{\pi^k(a)+\gamma}$ that motivates exploration can also be used in this setting. Generalizing the latter estimator to the adversarial RL setting requires to use the estimator $\hat{c}_h^k(s,a) = \frac{c_h^k(s,a)\mathbb{1}\{s=s_h^k,a=a_h^k\}}{d_h^{\pi_k}(s;p)\pi_h^k(a|s)+\gamma}$. However, since the model is unknown, Jin et al. (2019) uses $u_h^k(s)$ as an upper bound on $d_h^{\pi_k}(s;p)$ which further drives exploration.

Instead of using the empirical model and subtracting a bonus term, Algorithm 3 uses an optimistic model (Jaksch et al., 2010) for the policy evaluation stage. The model by which Q^k is evaluated is the one which results in the smallest loss among possible models,

$$\hat{p}_h^k(\cdot|s,a) \in \operatorname*{arg\,min}_{\hat{p}_h(\cdot|s,a)\in\mathcal{P}_h^{k-1}(s,a)} \hat{p}_h(\cdot|s,a)V_{h+1}^k.$$

The solution to this optimization problem can be computed efficiently (see, e.g., Jaksch et al. (2010)).

Remark 7.1 (Optimistic Model vs. Additive Exploration Bonus)**.** *Replacing the optimistic model with additive bonuses, we were able to establish $\tilde{O}(K^{3/4})$ regret bound. It is not clear whether this approach can attain a $\tilde{O}(K^{2/3})$ regret bound, as achieved when using an optimistic model.*

The following theorem bounds the regret of Algorithm 3. A full proof is found in Appendix C.2.

Theorem 2. *For any $K' \in [K]$, setting $\gamma = \tilde{O}(A^{-1/2}K^{-1/3})$ and $t_K = \tilde{O}(H^{-1}K^{-2/3})$, the regret of Algorithm 3 is bounded by*

$$\textit{Regret}(K') \le \tilde{O}\Big(H^2 S\sqrt{A}(K^{2/3} + SAK^{1/3})\Big).$$

Central to the analysis are the following claims, formally established in Appendix C. The first is proved in (Jin et al., 2019)[Lemma 11], based upon (Neu, 2015)[Lemma 1].

Claim 1 (Jin et al. (2019), Lemma 11)**.** *Let $\alpha^1, .., \alpha^{K'}$ be a sequence of $\mathcal{F}_{k-1}$ measurable functions such that $\alpha^k \in [0, 2\gamma]^{S\times A}$. Then, for any h and $K' \in [K]$, with high probability, $\sum_{k=1}^{K'}\sum_{s,a}\alpha^k(s,a)\big(\hat{c}_h^k(s,a) - c_h^k(s,a)\big) \le \tilde{O}(1)$.*

Claim 2. *Let $\alpha^1, .., \alpha^{K'}$ be a sequence of $\mathcal{F}_{k-1}$ measurable functions such that $\alpha^k \in [0, 2\gamma]$. For any s, h and $K' \in [K]$, with high probability, $\sum_{k=1}^{K'}\alpha^k\big(V_h^k(s) - V_h^{\pi_k}(s)\big) \le \tilde{O}(H)$.*

Claim 2 (see Lemma 7 in the appendix) allows us to derive improved upper bound on $\sum_{k=1}^{K'} V_h^k(s)$ which is crucial to derive the $\tilde{O}(K^{2/3})$ regret bound. Naively, we can bound $V_h^k(s)$ by recalling it is a value function of an MDP with costs bounded by $1/\gamma$. This leads to the naive bound

$$\sum_{k=1}^{K'} V_h^k(s) \le K'H/\gamma. \tag{7.2}$$

However, a tighter upper bound can be obtained by applying Claim 2 with $\alpha^k = 2\gamma$ for all $k \in [K']$. We have that

$$\sum_{k=1}^{K'} V_h^k(s) \le \sum_{k=1}^{K'} V_h^{\pi_k}(s) + \frac{H}{\gamma} \le HK' + \frac{H}{\gamma}, \tag{7.3}$$

where in the last relation we used the fact that for any s, h, $V_h^{\pi_k}(s) \le H$. In the following proof sketch we apply the later upper bound and demonstrate its importance.

Proof Sketch. We decompose the regret as in Theorem 1 to (i) Bias term, (ii) OMD term, and (iii) Optimism term. We bound both the Bias and Optimism terms in the appendix while relying on both Claim 1 and Claim 2.

Term (ii): OMD Analysis. Similarly to the stochastic case, we utilize the usual OMD analysis (Lemma 16), which ensures that for any policy π and s, h,

$$\begin{aligned}
&\sum_{k=1}^{K'}\big\langle Q_h^k(\cdot\mid s), \pi_h^k(\cdot\mid s) - \pi_h(\cdot\mid s)\big\rangle \\
&\quad\le \frac{\log A}{t_K} + \frac{t_K}{2}\sum_{k=1}^{K'}\sum_a \pi_h^k(a\mid s)(Q_h^k(s,a))^2 \\
&\quad\le \frac{\log A}{t_K} + \frac{t_K H}{2\gamma}\sum_{k=1}^{K'}\underbrace{\sum_a \pi_h^k(a\mid s)Q_h^k(s,a)}_{=V_h^k(s)} \\
&\quad\le \frac{\log A}{t_K} + \frac{t_K H}{2\gamma}(HK' + \frac{H}{\gamma}),
\end{aligned}$$

where the second relation holds since $0 \leq Q_h^k(s,a) \leq \frac{H}{\gamma}$, and the third relation holds by applying Eq. (7.3). Plugging this in Term (ii) we get

$$\text{Term (ii)} =$$
$$= \sum_{h=1}^{H} \mathbb{E}\left[\sum_{k=1}^{K'} \langle Q_h^k(s_h,\cdot), \pi_h^k(\cdot|s_h) - \pi_h(\cdot|s_h)\rangle \mid s_1, \pi, p\right]$$
$$\leq \frac{H \log A}{t_K} + \frac{t_K H^2}{2\gamma}(HK' + \frac{H}{\gamma}).$$

□

8. Discussion

On-policy vs. Off-policy. There are two prevalent approaches for policy optimization in practice, on-policy and off-policy. On-policy algorithms, e.g., TRPO (Schulman et al., 2015), update the policy based on data gathered following the current policy. This results in updating the policy only in observed states. However, in terms of theoretical guarantees, the convergence analysis of this approach requires the strong assumption of finite concentrability coefficient (Kakade & Langford, 2002; Scherrer & Geist, 2014; Agarwal et al., 2019; Liu et al., 2019; Shani et al., 2019). The assumption arises from the need to acquire global guarantees from the local nature of policy gradients.

The approach taken in this work, is fundamentally different than such on-policy approaches. In each episode, instead of updating the policy only at visited states, the policy is updated over the entire state space, by using all the historical data (in the form of the empirical model). Thus, the analyzed approach bears resemblance to off-policy algorithms, e.g., SAC (Haarnoja et al., 2018). There, the authors i) estimate the Q-function of the current policy by sampling from a buffer, which contains historical data, and ii) apply an MD-like policy update to states sampled from the buffer.

The uniform updates of policy-based methods analyzed in this work are in stark contrast to value-based algorithms, such as in (Jin et al., 2018; Efroni et al., 2019), where only observed states are updated. It remains an important open question, whether such updates could also be implemented in a provable policy based algorithm. In the case of stochastic POMD, this may be achieved by using optimistic Q-function estimates, instead of estimating the model with UCB-bonus, similarly to (Jin et al., 2019). There, the authors keep the estimates optimistic with respect to the optimal Q-function, Q^*. However, in approximate policy optimization, the policy improvement is done with respect to Q^{π_k}, as described in Algorithm 1. Therefore, differently than in (Jin et al., 2019), such off-policy version would require learning an optimistic Q^{π_k} estimator, instead of Q^*.

Policy vs. State-Action Occupancy Optimization. In our work, we proposed algorithms which directly optimize the policy. In this scenario, the policy is updated independently at each time step h and state s. That is, an optimization problem is solved over the action space in each h, s. Therefore, this method requires solving HS optimization problems of size A, where each has *a closed form solution* in the tabular setting.

Alternatively, algorithms based on the O-REPS framework (Zimin & Neu, 2013), follow a different approach and optimize over the state-action occupancy measures instead of directly on policies. In the case of unknown transition model, taking such an approach requires solving a constrained convex optimization problem, later relaxed to a convex optimization problem with only non-negativity constraints (Rosenberg & Mansour, 2019b) of size HS^2A, in each episode. Unlike the policy optimization approach, this optimization problem *does not have a closed form solution*. Thus, the computational cost of optimizing over the state-action occupancy measures is much worse than the policy optimization one.

Another significant shortcoming in applying the O-REPS framework is the difficulty to scale it to the function approximation setting. Specifically, in case the state-action occupancy measure is represented by a non-linear function, it is unclear how to solve the constrained optimization problem as defined in (Rosenberg & Mansour, 2019b). Differently than the O-REPS framework, the policy optimization approach scales naturally to the function approximation setting, e.g., Haarnoja et al. (2018). In this important aspect, policy optimization is preferable.

Interestingly, our work establishes $\tilde{O}(\sqrt{K})$ regret when using POMD for the stochastic case, suggesting that policy-based methods are sufficient for solving stochastic MDPs, and thus preferable, compared to the O-REPS framework, as they also enjoy better computational properties. However, in the adversarial case, Jin et al. (2019) recently established $\tilde{O}(\sqrt{K})$ regret for the UOB-REPS algorithm, where the adversarial variant of POMD only obtains $\tilde{O}\left(K^{2/3}\right)$ regret. Hence, it is of importance to understand whether it is possible to bridge this gap between policy and occupancy measure based methods, or alternatively to show that this gap is in fact a true drawback of policy optimization methods in the adversarial case.

9. Acknowledgments

We thank the anonymous reviewers for providing us with very helpful comments. This work was partially funded by the Israel Science Foundation under ISF grant number 1380/16.

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, 2019.

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.

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

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, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 263–272, 2017.

Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*, 2019.

Dann, C., Lattimore, T., and Brunskill, E. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 5713–5723, 2017.

Deisenroth, M. and Rasmussen, C. E. Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on machine learning (ICML-11)*, pp. 465–472, 2011.

Efroni, Y., Merlis, N., Ghavamzadeh, M., and Mannor, S. Tight regret bounds for model-based reinforcement learning with greedy policies. In *Advances in Neural Information Processing Systems*, pp. 12203–12213, 2019.

Geist, M., Scherrer, B., and Pietquin, O. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pp. 2160–2169, 2019.

Gu, S., Holly, E., Lillicrap, T., and Levine, S. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 3389–3396. IEEE, 2017.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870, 2018.

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.

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. Learning adversarial markov decision processes with bandit feedback and unknown transition. *arXiv preprint arXiv:1912.01192*, 2019.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *ICML*, volume 2, pp. 267–274, 2002.

Kakade, S. M. A natural policy gradient. In *Advances in neural information processing systems*, pp. 1531–1538, 2002.

Kearns, M. and Singh, S. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2-3):209–232, 2002.

Levine, S., Finn, C., Darrell, T., and Abbeel, P. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Liu, B., Cai, Q., Yang, Z., and Wang, Z. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*, 2019.

Maurer, A. and Pontil, M. Empirical bernstein bounds and sample variance penalization. *stat*, 1050:21, 2009.

Munos, R. Error bounds for approximate policy iteration. In Fawcett, T. and Mishra, N. (eds.), *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pp. 560–567. AAAI Press, 2003.

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., Antos, A., György, A., and Szepesvári, C. Online markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems*, pp. 1804–1812, 2010a.

Neu, G., György, A., and Szepesvári, C. The online loop-free stochastic shortest-path problem. In *COLT*, volume 2010, pp. 231–243. Citeseer, 2010b.

Neu, G., Jonsson, A., and Gómez, V. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Orabona, F. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Peters, J. and Schaal, S. Policy gradient methods for robotics. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2219–2225. IEEE, 2006.

Peters, J. and Schaal, S. Reinforcement learning of motor skills with policy gradients. *Neural networks*, 21(4):682–697, 2008.

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

Rosenberg, A. and Mansour, Y. Online stochastic shortest path with bandit feedback and unknown transition function. In *Advances in Neural Information Processing Systems*, pp. 2209–2218, 2019a.

Rosenberg, A. and Mansour, Y. Online convex optimization in adversarial markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 5478–5486, 2019b.

Scherrer, B. Approximate policy iteration schemes: a comparison. In *International Conference on Machine Learning*, pp. 1314–1322, 2014.

Scherrer, B. and Geist, M. Local policy search in a convex space and conservative policy iteration as boosted policy search. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–50. Springer, 2014.

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.

Shani, L., Efroni, Y., and Mannor, S. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. *arXiv preprint arXiv:1909.02769*, 2019.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pp. 1057–1063, 2000.

Vieillard, N., Pietquin, O., and Geist, M. On connections between constrained optimization and reinforcement learning. *arXiv preprint arXiv:1910.08476*, 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.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

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*, pp. 7304–7312, 2019.

Zimin, A. and Neu, G. Online learning in episodic markovian decision processes by relative entropy policy search. In *Advances in neural information processing systems*, pp. 1583–1591, 2013.