A Sample-Efficient Algorithm for Episodic Finite-Horizon MDP with Constraints

Krishna C. Kalagarla, Rahul Jain, Pierluigi Nuzzo
Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles
{kalagarl,rahul.jain,nuzzo}@usc.edu

Abstract

Constrained Markov decision processes (CMDPs) formalize sequential decision-making problems whose objective is to minimize a cost function while satisfying constraints on various cost functions. In this paper, we consider the setting of episodic fixed-horizon CMDPs. We propose an online algorithm which leverages the linear programming formulation of repeated optimistic planning for finite-horizon CMDP to provide a probably approximately correctness (PAC) guarantee on the number of episodes needed to ensure a near optimal policy, i.e., with resulting objective value close to that of the optimal value and satisfying the constraints within low tolerance, with high probability. The number of episodes needed is shown to have linear dependence on the sizes of the state and action spaces and quadratic dependence on the time horizon and an upper bound on the number of possible successor states for a state-action pair. Therefore, if the upper bound on the number of possible successor states is much smaller than the size of the state space, the number of needed episodes becomes linear in the sizes of the state and action spaces and quadratic in the time horizon.

Introduction

Markov decision processes (MDPs) (Puterman 1994) offer a natural framework to express sequential decision-making problems and reason about autonomous system behaviors. However, the single cost objective of a traditional MDP formulation may fall short of fully capturing problems with multiple conflicting objectives and additional constraints that must be satisfied. Consider, for example, an autonomous car that is required to reach a destination at the earliest, but also satisfy a set of safety requirements and fuel consumption constraints, while keeping a desired comfort level (Le, Voloshin, and Yue 2019). The framework of constrained MDPs (CMDPs) (Altman 1999) extended MDPs by considering additional constraints on the expected long-term performance of a policy. The objective in a CMDP is to minimize the expected cumulative cost while satisfying the additional constraints. In this paper, we consider episodic finite-horizon CMDPs, where an agent interacts with a CMDP repeatedly in episodes of fixed length, a setting that can model a large number of repetitive tasks such as goods delivery or customer service.

We address the problem of online learning of CMDPs with unknown transition probabilities, by requiring only observed trajectories rather than sampling the transition function for any state-action pair from a generative model, which may not always be available. An important question which arises in online learning is the exploration-exploitation dilemma, i.e., the trade-off between exploration, to gain more information about the model, and exploitation, to minimize the cost. In this respect, the performance of learning algorithms is commonly evaluated in terms of (i) regret, i.e., the difference between the cumulative cost of the agent and that of the optimal policy in hindsight, and (ii) sample complexity, i.e., the number of steps for which the learning agent may not play a near-optimal policy. We consider a policy to be near-optimal if the expected cumulative cost is close to the optimal and the constraints are satisfied within a small tolerance. In this paper, we address sample efficiency by proposing an algorithm that provides probably approximately correctness (PAC) guarantees.

Our algorithm leverages the concept of *optimism in the face of uncertainty* (Lai and Robbins 1985; Auer, Jaksch, and Ortner 2009) to balance exploration and exploitation. The learning agent repeatedly defines a set of statistically plausible transition models given the observations made so far. It then chooses an optimistic transition probability model and an optimistic policy with respect to the given constrained MDP problem. This planning step is formulated as a linear programming (LP) problem in occupancy measures, whose solution gives the desired optimistic policy. This policy is executed for multiple episodes until a state-action pair has been visited sufficiently often. The total visitation counts are then updated and these steps are repeated. We show that the number of episodes in which the learning agent plays an ϵ-suboptimal policy is upper bounded by $\tilde{\mathcal{O}}\big(\frac{|\mathcal{S}||\mathcal{A}|C^2H^2}{\epsilon^2}\log_2\frac{1}{\delta}\big)$ with probability at least $1-\delta$, where C is an upper bound on the number of possible successor states for a state-action pair, $|\mathcal{S}|$ and $|\mathcal{A}|$ are the state and action space sizes, respectively, and H is the time horizon.

Contribution. In this paper, we present one of the first online algorithms with PAC guarantees for episodic constrained MDPs with unknown transition probabilities. We build on a previous result which provides a probably approximately correct (PAC) algorithm for unconstrained episodic MDPs (Dann and Brunskill 2015). However, differ-

ently from planning based on the Bellman optimality equations (Dann and Brunskill 2015), we address the presence of constraints by formulating an optimistic planning problem as an LP in occupancy measures. Consequently, our formulation leverages a novel construction for the set of plausible transition models and results in a sample complexity that is quadratic in the time horizon H, thus improving on the cubic bounds previously obtained with regret-based formulations (e.g., see (Efroni, Mannor, and Pirotta 2020)).

Related Work. Significant research has been reported on efficient learning for unconstrained MDPs. Algorithms like UCBVI (Azar, Osband, and Munos 2017), UBEV (Dann, Lattimore, and Brunskill 2017), EULER (Zanette and Brunskill 2019) and EULER-GP (Efroni et al. 2019) focus on regret analysis for episodic finite-horizon MDPs. PAC algorithms for unconstrained MDPs are addressed by (Dann and Brunskill 2015; Brafman and Tennenholtz 2002; Strehl, Li, and Littman 2009). While these algorithms are model-based reinforcement learning (RL) algorithms, model-free algorithms such as UCB-H and UCB-B (Jin et al. 2018) have also been shown to be sample efficient.

Sample-efficient exploration in CMDPs has recently started to receive more attention. Regret analysis for multiple model-based and model-free algorithms (Efroni, Mannor, and Pirotta 2020) has been performed in the setting of episodic CMDPs with stochastic cost functions and unknown transition probabilities. Our work addresses PAC complexity, and is therefore complementary to the result by (Efroni, Mannor, and Pirotta 2020). Regret analysis for constrained MDPs has also been studied in the settings of average cost (Singh, Gupta, and Shroff 2020), adversarial cost with tabular MDPs (Qiu et al. 2020), and adversarial cost with linear MDPs (Ding et al. 2020).

Still in the context of constrained MDPs, the C-UCRL algorithm (Zheng and Ratliff 2020) has shown to have sublinear regret and satisfy the constraints even while learning, albeit in the setting of known transition probabilities and unknown cost functions. A regret-optimal algorithm for constrained MDPs with concave objectives and convex and hard constraints (knapsacks) has also been studied (Brantley et al. 2020), dealing with problems with a fixed budget such that the learning is stopped as soon as the budget is consumed. Several of these regret algorithms can be modified following an idea from (Jin et al. 2018) to provide PAC guarantees for constrained MDP with time-horizon dependence of at least H^3. However, this procedure is impractical as it entails saving an extremely large number of policies and uniformly sampling them to get the PAC optimal policy. Finally, regret or PAC analysis have not been addressed so far in the context of policy optimization and Lagrangian-based efforts on constrained MDPs (Borkar 2005; Achiam et al. 2017; Tessler, Mankowitz, and Mannor 2018; Miryoosefi et al. 2019).

Preliminaries

In this section, we introduce preliminary concepts from finite-horizon MDPs and CMDPs.

Notation. We denote the set of natural numbers by $\mathbb{N}$ and use $h \in \{1, \dots, H\}$ and $k \in \mathbb{N}$ to denote a time step inside an episode and a phase index, respectively. The indicator function $\mathbb{I}(s = s_1)$ evaluates to 1 when $s = s_1$ and 0 otherwise. The probability simplex over set S is denoted by Δ_S. For functions $f, p : \mathcal{S} \to \mathbb{R}$ and $\mathcal{S}$ a finite set, we write $p(\cdot)f = \sum_{s \in \mathcal{S}} p(s)f(s)$. Finally, we adopt the notation $\tilde{\mathcal{O}}$ which is similar to the usual $\mathcal{O}$ notation but ignores logarithmic factors.

Finite-Horizon MDPs. We consider an episodic finite-horizon MDP (Puterman 1994), which can be formally defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, s_1, p, c)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the finite state and action spaces, respectively. The agent interacts with the environment in episodes of length H and each episode starts with the same initial state s_1. The non-stationary transition probability is denoted by p where $p_h(s'|s, a)$ is the the probability of transitioning to state s' upon taking action a at state s at time step h. Further, we denote by $Succ(s, a)$ the set of possible successor states of state s and action a. The maximum number of possible successor states is denoted by $C = \max_{s,a} |Succ(s, a)|$. The non-stationary cost of taking action a in state s at step $h \in \{1, \dots, H\}$ is a random variable $C_h(s, a) \in [0, 1]$, with mean $c_h(s, a)$. Finally, we set $c = (c_1, \dots, c_H)$.

A non-stationary randomized policy $\pi = (\pi_1, \dots, \pi_H) \in \Pi$, where $\pi_i : \mathcal{S} \to \Delta_{\mathcal{A}}$, maps each state to a probability simplex over the action space. We denote by $a_h \sim \pi_h(s_h)$ the action taken at time step h at state s_h according to policy π. For a state $s \in \mathcal{S}$ and time step $h \in \{1, \dots, H\}$, the value function of a non-stationary randomized policy, $V_h^\pi(s; c, p)$, is defined as:

$$V_h^\pi(s; c, p) = \mathbb{E}\left[\sum_{i=h}^{H} c_i(s_i, a_i) | s_h = s, \pi, p\right],$$

where the expectation is over the environment and policy randomness. We omit π, c, p when they are clear from the context. Similarly, for a state $s \in \mathcal{S}$, an action $a \in \mathcal{A}$, and time step $h \in \{1, \dots, H\}$, the Q-value function is defined as $Q_h^\pi(s, a; c, p) =$

$$= c_h(s, a) + \mathbb{E}\left[\sum_{i=h+1}^{H} c_i(s_i, a_i) | s_h = s, a_h = a, \pi, p\right].$$

There always exists an optimal non-stationary deterministic policy π^* (Puterman 1994) such that $V_h^{\pi^*}(s) = V_h^*(s) = \inf_\pi V_h^\pi(s)$ and $Q_h^{\pi^*}(s, a) = Q_h^*(s, a) = \inf_\pi Q_h^\pi(s, a)$. The Bellman optimality equations (Puterman 1994) enable us to compute the optimal policy by backward induction:

$$V_h^*(s) = \min_{a \in \mathcal{A}} \left[c_h(s, a) + p_h(\cdot|s, a)V_{h+1}^*\right],$$
$$Q_h^*(s, a) = c_h(s, a) + p_h(\cdot|s, a)V_{h+1}^*,$$

where $V_{H+1}^*(s) = 0$ and $V_h^*(s) = \min_{a \in \mathcal{A}} Q_h^*(s, a)$. The optimal policy π^* is thus greedy with respect to Q_h^*.

Finite-Horizon Constrained MDPs. A finite-horizon constrained MDP is a finite-horizon MDP along with additional I constraints (Altman 1999) expressed by pairs

of constraint cost functions and thresholds, $\{d_i, l_i\}_{i=1}^I$. The cost of taking action a in state s at time step $h \in \{1, \ldots, H\}$ with respect to the i^{th} constraint cost function is a random variable $D_{i,h}(s,a) \in [0,1]$, with mean $d_{i,h}(s,a)$. The total expected cost of an episode under policy π with respect to cost functions $c, d_i, i \in \{1, \ldots, I\}$, is the respective value function from the initial state s_1, i.e., $V_1^\pi(s_1; c), V_1^\pi(s_1; d_i), i \in \{1, \ldots, I\}$, respectively, by definition. The objective of a CMDP is to find a policy which minimizes the total expected objective cost under the constraint that the total expected constraint costs are below the respective desired thresholds. Formally,

$$\begin{aligned} \pi^* \in \underset{\pi \in \Pi}{\text{argmin}} \quad & V_1^\pi(s_1; c, p) \\ \text{s.t.} \quad & V_1^\pi(s_1; d_i, p) \leq l_i \quad \forall i \in \{1, \ldots, I\}. \end{aligned} \tag{1}$$

The optimal value is $V^* = V_1^{\pi^*}(s_1; c, p)$. The optimal policy may be randomized (Altman 1999), i.e., an optimal deterministic policy may not exist as in the case of the finite-horizon MDP. Further, the Bellman optimality equations do not hold due to the constraints. Thus, we cannot leverage backward induction to find an optimal policy. A linear programming approach has been shown (Altman 1999) to find an optimal policy.

Linear Programming for CMDPs. Occupancy measures (Altman 1999) allow formulating the optimization problem (1) as a linear program (LP). Occupancy measure q^π of a policy π in a finite-horizon MDP is defined as the expected number of visits to a state-action pair (s,a) in an episode at time step h. Formally,

$$\begin{aligned} q_h^\pi(s,a;p) &= \mathbb{E}\left[\mathbb{I}\{s_h = s, a_h = a\} | s_1 = s_1, \pi, p\right] \\ &= P\left[s_h = s, a_h = a | s_1 = s_1, \pi, p\right]. \end{aligned}$$

It is easy to see that the occupancy measure q^π of a policy π satisfies the following properties, expressing non-negativity and flow conservation, respectively:

$$\begin{aligned} q_h^\pi(s,a) &\geq 0, \quad \forall\, (s,a,h) \in \mathcal{S} \times \mathcal{A} \times \{1, \ldots, H\}, \\ q_1^\pi(s,a) &= \pi_1(a|s)\mathbb{I}(s = s_1), \quad \forall\, (s,a) \in \mathcal{S} \times \mathcal{A}, \\ \sum_a q_h^\pi(s,a) &= \sum_{s',a'} p_{h-1}(s|s',a') q_{h-1}^\pi(s',a'), \\ &\forall s \in \mathcal{S}, h \in \{2, \ldots, H\}, \end{aligned}$$

where $\mathbb{I}(s = s_1)$ is the initial state distribution. The space of the occupancy measures satisfying the above constraints is denoted by $\Delta(\mathcal{M})$. A policy π generates an occupancy measure $q \in \Delta(\mathcal{M})$ if

$$\pi_h(a|s) = \frac{q_h(s,a)}{\sum_b q_h(s,b)}, \ \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times \{1, \ldots, H\}. \tag{2}$$

Thus, there exists a unique generating policy for all occupancy measures in $\Delta(\mathcal{M})$ and *vice versa*. Further, the total expected cost of an episode under policy π with respect to cost function c can be expressed in terms of the occupancy measure as follows:

$$V_1^\pi(s_1; c, p) = \sum_{h,s,a} q_h^\pi(s,a;p) c_h(s,a).$$

The optimization problem (1) can then be reformulated as a linear program (Altman 1999; Zimin and Neu 2013) as follows:

$$\begin{aligned} q^* \in \underset{q \in \Delta(\mathcal{M})}{\text{argmin}} \quad & \sum_{h,s,a} q_h(s,a) c_h(s,a), \\ \text{s.t.} \quad & \sum_{h,s,a} q_h(s,a) d_{i,h}(s,a) \leq l_i, \quad \forall i \in \{1, \ldots, I\}. \end{aligned}$$

The optimal policy π^* can be obtained from q^* following (2).

The Learning Problem

We consider the setting where an agent repeatedly interacts with a finite-horizon CMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, s_1, p, c, \{d_i, l_i\}_{i=1}^I)$ with stationary transition probability (i.e., $p_h = p, \forall h \in \{1, \ldots, H\}$) in episodes of fixed length H, starting from the same initial state s_1. For simplicity of analysis,[1] we assume that the cost functions $c, \{d_i\}_{i=1}^I$ are known to the learning agent, but the transition probability p is unknown. The agent estimates the transition probability in an online manner by observing the trajectories over multiple episodes.

The main objective is to design an online learning algorithm such that, for given $\epsilon \in (0,1]$, $\delta \in (0,1)$ and CMDP $\mathcal{M}$, the number of episodes for which the agent follows an ϵ-suboptimal policy is bounded above by a polynomial (up to logarithmic factors) in the relevant quantities $(|\mathcal{S}|, |\mathcal{A}|, H, \frac{1}{\epsilon}, \frac{1}{\delta})$ with high probability, i.e., with probability at least $1 - \delta$ (PAC guarantee). A policy π is said to be ϵ-optimal if the total expected objective cost of an episode under policy π is within ϵ of the optimal value, i.e., $V_1^\pi(s_1; c, p) \leq V^* + \epsilon$, and the constraints are satisfied within an ϵ tolerance, i.e., $V_1^\pi(s_1; d_i, p) \leq l_i + \epsilon, \forall\, i \in \{1, \ldots, I\}$. We make the following assumption of feasibility.

Assumption 1 *The given CMDP $\mathcal{M}$ is feasible, i.e., there exists a policy π such that the constraints are satisfied.*

The UC-CFH Algorithm

Algorithm Description. We consider an adaptation of the model-based algorithm UCFH (Dann and Brunskill 2015) to the setting of CMDPs, which we call Upper-Confidence Constrained Fixed-Horizon episodic reinforcement learning (UC-CFH) algorithm. The algorithm leverages the approach of *optimism in the face of uncertainty* (Auer, Jaksch, and Ortner 2009) to balance exploration and exploitation.

The algorithm operates in phases indexed by k and whose length is not fixed but, instead, depends on the observations made until the current episode. Each phase consists of three stages: planning, policy execution, and update of the visitation counts.

[1] The complexity of learning the transition probability dominates the complexity of learning the cost functions (Auer and Ortner 2005). The algorithm can be readily extended to the setting of unknown cost functions by using an optimistic lower bound of the cost function obtained from its empirical estimate in place of the known cost function.

For each phase k, UC−CFH defines a set of plausible transition models based on the number of visits to state-action pairs (s,a) and transition tuples (s,a,s') so far. A policy π^k is chosen by solving an optimistic planning problem, which is expressed as an LP problem (lines 13-16 in Algorithm 1). The planning problem (CONSTRAINEDEXTENDEDLP in the algorithm) is detailed below.

The algorithm maintains two types of visitation counts. Counts $v(s,a)$ and $v(s,a,s')$ are the number of visits to state-action pairs (s,a) and transition tuples (s,a,s'), respectively, since the last update of state-action pair (s,a). Counts $n(s,a)$ and $n(s,a,s')$ are the total number of visits to state-action pairs (s,a) and transition tuples (s,a,s'), respectively, before the update of state-action pair (s,a). These visitation counts are all initialized to zero.

During the policy execution stage of phase k (lines 18-27 in Algorithm 1), the agent executes the current policy π^k, observes the tuples (s_t,a_t,s_{t+1}), and updates the respective visitation counts $v(s_t,a_t)$ and $v(s_t,a_t,s_{t+1})$. This policy π^k is executed until a state action pair (s,a) has been visited often enough since the last update of (s,a), i.e., $v(s,a)$ is large enough (lines 26-27 in Algorithm 1).

In the next stage of phase k (lines 29-33 in Algorithm 1), the visitation counts $n(s,a)$ and $n(s,a,s')$ corresponding to the sufficiently visited state-action pair (s,a) are updated as $n(s,a) = n(s,a) + v(s,a)$, $n(s,a,s') = n(s,a,s') + v(s,a,s')$ and visitation counts $v(s,a), v(s,a,s')$ are reset to 0. This iteration of planning-execution-update describes a phase of the algorithm.

Optimistic Planning. At the start of each phase k, UC−CFH estimates the true transition model by its empirical average as

$$\bar{p}^k(s'|s,a) = \frac{n^k(s,a,s')}{\max\{1, n^k(s,a)\}}, \quad \forall (s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}.$$

The algorithm further defines confidence intervals for the transition probabilities of the CMDP, such that the true transition probabilities lie in them with high probability. Formally, for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, we define:

$$B_p^k(s,a) = \{\tilde{p}(.|s,a) \in \Delta_{\mathcal{S}} : \forall s' \in \mathcal{S}$$
$$|\tilde{p}(s'|s,a) - \bar{p}^k(s'|s,a)| \le \beta_p^k(s,a,s')\},$$

where the size of the confidence intervals $\beta_p^k(s,a,s')$ is built using the empirical Bernstein inequality (Maurer and Pontil 2009) and, for any $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$, is defined as

$$\beta_p^k(s,a,s') = \sqrt{\frac{2\bar{p}^k(s'|s,a)(1-\bar{p}^k(s'|s,a))\ln\frac{4}{\delta'}}{\max(1,n^k(s,a))}} + \\ + \frac{7\ln\frac{4}{\delta'}}{3\max(1,n^k(s,a)-1)},$$

where δ' is as defined in the algorithm and $\bar{p}^k(s'|s,a)(1-\bar{p}^k(s'|s,a))$ is the variance associated with the empirical estimate $\bar{p}^k(s'|s,a)$.

Algorithm 1 UC−CFH: Upper-Confidence Constrained Fixed-Horizon Episodic Reinforcement Learning Algorithm

1: **Input**: Desired tolerance $\epsilon \in (0,1]$, failure tolerance $\delta \in (0,1)$, fixed-horizon MDP $\mathcal{M}$
2: **Result**: With probability at least $1-\delta$, ϵ-optimal policy
3:
4: $k := 1, \quad w_{min} := \frac{\epsilon}{4H|\mathcal{S}||\mathcal{A}|}, \quad \delta' := \frac{\delta}{2N_{\max}C}$;
5: $N_{max} := |\mathcal{S}||\mathcal{A}|\log_2\frac{|\mathcal{S}|H}{w_{min}}$;
6:
7: $m := \frac{2304C^2H^2}{\epsilon^2}(\log_2\log_2 H)^2\log_2^2\frac{8H^2|\mathcal{S}|^2|\mathcal{A}|}{\epsilon}\ln\frac{4}{\delta'}$;
8: $n(s,a) = v(s,a) = n(s,a,s') := 0$,
9: $\forall s \in \mathcal{S}, a \in \mathcal{A}, s' \in Succ(s,a)$;
10:
11: **while** True **do**
12:
13: $\bar{p}(s'|s,a) := \frac{n(s,a,s')}{\max\{1,n(s,a,s')\}}$,
14: $\forall s \in \mathcal{S}, a \in \mathcal{A}, s' \in Succ(s,a)$;
15:
16: $\pi^k := \text{CONSTRAINEDEXTENDEDLP}(\bar{p}, n)$;
17:
18: **repeat**
19: **for** $t = 0$ to $H-1$ **do**
20: $a_t \sim \pi_h^k(s_t)$;
21: $s_{t+1} \sim p(.|s_t,a_t)$;
22: $v(s_t,a_t) := v(s_t,a_t)+1$;
23: $v(s_t,a_t,s_{t+1}) := v(s_t,a_t,s_{t+1})+1$;
24: **end for**
25: **until** there is $(s,a) \in \mathcal{S}\times\mathcal{A}$,
26: s.t. $v(s,a) \ge \max\{mw_{min}, n(s,a)\}$ and
27: $n(s,a) < |\mathcal{S}|mH$
28:
29: $n(s,a) := n(s,a) + v(s,a)$;
30: $n(s,a,s') := n(s,a,s') + v(s,a,s')$,
31: $v(s,a) = v(s,a,s') := 0$,
32: $\forall s' \in Succ(s,a)$;
33: $k := k+1$;
34:
35: **end while**

Given the confidence intervals B_p^k, the algorithm then computes a policy π^k by performing optimistic planning. Given a confidence set of possible transition models, it selects an optimistic transition probability model and optimistic policy with respect to the given constrained MDP problem. This can be expressed as the following optimization problem:

$$(\tilde{p}^k, \pi^k) = \underset{\pi\in\Pi, \tilde{p}\in B_p^k}{\text{argmin}} \quad V_1^\pi(s_1; c, \tilde{p}) \qquad (3)$$
$$\text{s.t.} \quad V_1^\pi(s_1; d_i, \tilde{p}) \le l_i \quad \forall i \in \{1,\dots,I\}.$$

We allow time-dependent transitions, i.e., choosing different transition models at different time steps of an episode, even if the true CMDP has stationary transition probability. This does not affect the theoretical guarantees, since the true transition probability still lies in the confidence sets with high probability.

These confidence intervals differ from the ones considered in UCFH (Dann and Brunskill 2015), which have an additional condition that the standard deviation associated with a transition model, i.e., $\sqrt{\tilde{p}(1-\tilde{p})}$, be close to that of the empirical estimate $\sqrt{\bar{p}(1-\bar{p})}$. We remove this condition to be able to express the optimistic planning problem (3) as a linear program. However, this causes the PAC bound to have a quadratic dependence on C instead of a linear dependence.

CONSTRAINEDEXTENDEDLP Algorithm. Problem (3) can be expressed as an extended LP by leveraging the state-action-state occupancy measure $z^\pi(s,a,s';p)$, defined as $z^\pi_h(s,a,s';p) = p_h(s'|s,a)q^\pi_h(s,a;p)$, to express the confidence intervals of the transition probabilities(Efroni, Mannor, and Pirotta 2020). The extended LP over z can be formulated as follows:

$$\min_{z\geq 0} \sum_{h,s,a,s'} z_h(s,a,s')c_h(s,a),$$

$$\sum_{h,s,a,s'} z_h(s,a,s')d_{i,h}(s,a) \leq l_i, \quad \forall i \in \{1,\dots,I\}$$

$$\sum_{a,s'} z_h(s,a,s') = \sum_{s',a'} z_{h-1}(s',a',s), \ \forall s \in \mathcal{S}, \forall h \in \{2,\dots,H\}$$

$$\sum_{a,s'} z_1(s,a,s') = \mathbb{I}(s=s_1), \quad \forall s \in \mathcal{S}$$

$$z_h(s,a,s') - (\bar{p}^k(s'|s,a) + \beta^k_p(s,a,s')) \sum_y z_h(s,a,y) \leq 0,$$

$$\forall (s,a,s',h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times\{1,\dots,H\}$$

$$- z_h(s,a,s') + (\bar{p}^k(s'|s,a) - \beta^k_p(s,a,s')) \sum_y z_h(s,a,y) \leq 0,$$

$$\forall (s,a,s',h) \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}\times\{1,\dots,H\}.$$

The last two constraints of the above LP encode the condition that the transition probability must lie in the desired confidence interval. The desired policy π^k and the chosen transition probabilities are recovered from the computed occupancy measures as:

$$\pi^k_h(a|s) = \frac{\sum_{s'} z_h(s,a,s')}{\sum_{a,s'} z_h(s,a,s')}$$

and

$$\tilde{p}^k_h(s'|s,a) = \frac{z_h(s,a,s')}{\sum_{s'} z_h(s,a,s')}.$$

The above planning routine, referred to as CONSTRAINEDEXTENDEDLP in the algorithm, was also used in the context of adversarial MDPs (Jin and Luo 2019; Rosenberg and Mansour 2019). The following theorem establishes the PAC guarantee for the algorithm `UC-CFH`.

Theorem 1 *For $\epsilon \in (0,1]$, $\delta \in (0,1)$, with probability at least $1-\delta$, algorithm* `UC-CFH` *yields at most $\tilde{\mathcal{O}}\big(\frac{|\mathcal{S}||\mathcal{A}|C^2H^2}{\epsilon^2}\log_2\frac{1}{\delta}\big)$ episodes with ϵ-suboptimal policies π^k, i.e., $V_1^{\pi^k}(s_1,c) - V^* > \epsilon$ or $V_1^{\pi^k}(s_1,d_i) - l_i > \epsilon$, for any $i \in \{1,\dots,I\}$.*

Thus, in the natural setting of a limited size of successor states, i.e., $C \ll |\mathcal{S}|$, the number of episodes needed by `UC-CFH` to obtain an ϵ-optimal policy with high probability has a linear dependence on the state and action space sizes $|\mathcal{S}|$ and $|\mathcal{A}|$, respectively, and quadratic dependence on the time horizon H.

PAC Analysis

For state-action pairs, we introduce a notion of *knownness*, to indicate how often the pair has been visited relative to its expected number of visits under a policy, and a notion of *importance*, to indicate the influence that the pair has on the total expected cost of a policy (Dann and Brunskill 2015). We consider a fine-grained categorization of *knownness* of state-action pairs, similar to the one by (Lattimore and Hutter 2012; Dann and Brunskill 2015), instead of a binary categorization (Brafman and Tennenholtz 2002; Strehl, Li, and Littman 2009). These notions are essential for the analysis of the algorithm.

We define the weight of a state-action pair (s,a) under policy π^k as its expected number of visits in an episode, i.e.,

$$w_k(s,a) = \sum_{t=1}^{H} P\left[s_t = s, a_t = a | \pi^k, s_1\right].$$

The *importance* ι_k of a state-action pair (s,a) with respect to policy π^k is an integer defined as its relative weight with respect to w_{min} on a logarithmic scale:

$$\iota_k(s,a) = \min\left\{z_i : z_i \geq \frac{w_k(s,a)}{w_{min}}\right\},$$

where $z_1 = 0, z_i = 2^{i-2}, \forall\, i \geq 2$. Similarly, the *knownness* κ_k of a state-action pair (s,a) is an integer defined as

$$\kappa_k(s,a) = \max\left\{z_i : z_i \leq \frac{n_k(s,a)}{m w_k(s,a)}\right\},$$

where $z_1 = 0, z_i = 2^{i-2}, \forall\, i \geq 2$, and the constant m is as defined in Algorithm 1. We then divide the (s,a)-pairs into categories as follows:

$$\mathcal{X}_{k,\kappa,\iota} = \{(s,a) \in \mathcal{X}_k : \kappa_k(s,a) = \kappa, \iota_k(s,a) = \iota\},$$
$$\bar{\mathcal{X}}_k = \mathcal{S}\times\mathcal{A}\backslash\mathcal{X}_k,$$

where $\mathcal{X}_k = \{(s,a) \in \mathcal{S}\times\mathcal{A} : \iota_k(s,a) > 0\}$ is the active set and $\bar{\mathcal{X}}_k$ is the inactive set, i.e., the set of state-action pairs that are unlikely to be visited under policy π^k. The idea is that the model estimated by the algorithm is accurate if only a small number of state-action pairs are in categories with low knownness, that is, they are important under the current policy but have not yet been sufficiently observed.

We therefore distinguish between phases k, where the condition $|\mathcal{X}_{k,\kappa,\iota}| \leq \kappa$ for all κ and ι holds, and phases where this does not hold. This condition ensures that the number of state-action pairs in categories with low knownness are small and there are more state-action pairs in categories with higher knownness. We will further prove that the policy is ϵ-optimal in episodes which satisfy this condition.

Proof of Theorem 1

The proof of Theorem 1 consists of the following parts, supported by technical lemmas that are postponed to the next subsection to improve readability. We first show in Lemma 2 that the true transition model is contained within the confidence sets for all phases with high probability, i.e., the true transition probability p belongs to B_p^k for all k with probability at least $1-\frac{\delta}{2}$.

We then use a result from (Dann and Brunskill 2015) to provide a high probability upper bound on the number of episodes for which the condition $\forall\, \kappa, \iota : |\mathcal{X}_{k,\kappa,\iota}| \leq \kappa$ is violated. This result is restated as Lemma 3, with minor modification to accommodate randomized policies instead of deterministic policies. By Lemma 3, the number of episodes with $|\mathcal{X}_{k,\kappa,\iota}| > \kappa$ for some κ, ι is bounded above by $6NE_{max}$, where $N = |\mathcal{S}||\mathcal{A}|m$ and $E_{max} = \log_2 \frac{H}{w_{min}} \log_2 |\mathcal{S}||\mathcal{A}|$ with probability at least $1-\frac{\delta}{2}$. The choice of m in Theorem 1 satisfies the condition on m in Lemma 3. We therefore have, with high probability, i.e., at least $1-\frac{\delta}{2}$, $|\mathcal{X}_{k,\kappa,\iota}| \leq \kappa$ for all κ, ι for the remaining episodes.

By union bound, for episodes beyond the first $6NE_{max}$, we can conclude that $|\mathcal{X}_{k,\kappa,\iota}| \leq \kappa$ for all κ, ι and $p \in B_p^k$ with probability at least $1-\delta$. Further, in Lemma 9, we show that, in episodes with $|\mathcal{X}_{k,\kappa,\iota}| \leq \kappa$ for all κ, ι, the optimistic expected total cost is ϵ-close to the true expected total cost. Therefore, the following hold:

$$|V_1^{\pi^k}(s_1, c) - \tilde{V}_1^{\pi^k}(s_1, c)| \leq \epsilon,$$

$$|V_1^{\pi^k}(s_1, d_i) - \tilde{V}_1^{\pi^k}(s_1, d_i)| \leq \epsilon, \quad \forall i \in \{1, \ldots, I\}.$$

We note that $\tilde{p}^k, \pi^k$ were obtained by solving the following optimization problem:

$$(\tilde{p}^k, \pi^k) = \underset{\pi \in \Pi, \tilde{p} \in B_p^k}{\operatorname{argmin}} \quad V_1^{\pi}(s_1; c, \tilde{p}) \tag{4}$$

$$\text{s.t.} \quad V_1^{\pi}(s_1; d_i, \tilde{p}) \leq l_i \quad \forall i \in \{1, \ldots, I\}.$$

Thus, for $p \in B_p^k$, we have,

$$\begin{aligned} V_1^{\pi^k}(s_1, c) - V^* &= V_1^{\pi^k}(s_1, c) - \tilde{V}_1^{\pi^k}(s_1, c) + \\ &\quad + \tilde{V}_1^{\pi^k}(s_1, c) - V^* \\ &\leq V_1^{\pi^k}(s_1, c) - \tilde{V}_1^{\pi^k}(s_1, c) \\ &\quad \text{(by (4), since } p \in B_p^k) \\ &\leq \epsilon \quad \text{(by Lemma 9).} \end{aligned}$$

Similarly, for all $i \in \{1, \ldots, I\}$, we obtain

$$\begin{aligned} V_1^{\pi^k}(s_1, d_i) - l_i &= V_1^{\pi^k}(s_1, d_i) - \tilde{V}_1^{\pi^k}(s_1, d_i) + \\ &\quad + \tilde{V}_1^{\pi^k}(s_1, d_i) - l_i \\ &\leq V_1^{\pi^k}(s_1, d_i) - \tilde{V}_1^{\pi^k}(s_1, d_i) \\ &\quad \text{(since } \pi^k \text{ satisfies the constraints of (4))} \\ &\leq \epsilon \quad \text{(by Lemma 9).} \end{aligned}$$

By putting the above inequalities together we have that, with probability at least $1-\delta$, UC-CFH has at most $6|\mathcal{S}||\mathcal{A}|m \log_2 \frac{H}{w_{min}} \log_2 |\mathcal{S}||\mathcal{A}|$ ϵ-suboptimal episodes.

Technical Lemmas

We state the main lemmas used in the proof of Theorem 1.

Capturing the true transition model with high probability. We first restate the lemma that provides an upper bound on the total number of phases in the algorithm UC-CFH from (Dann and Brunskill 2015).

Lemma 1 *The total number of phases in the algorithm is bounded above by* $N_{max} = |\mathcal{S}||\mathcal{A}| \log_2 \frac{|\mathcal{S}|H}{w_{min}}$.

The above result is used along with concentration results based on the empirical Bernstein inequality (Maurer and Pontil 2009) and union bounds to show that the true transition model is contained within the confidence sets for all phases with high probability.

Lemma 2 *The true transition probability is contained within the confidence intervals for all phases with high probability, i.e.,* $p \in B_p^k, \forall k$ *with probability at least* $1-\frac{\delta}{2}$.

The above lemma implies that the extended LP of the planning stage is feasible in all phases with high probability, since the true CMDP is feasible by Assumption 1.

Number of episodes which violate $|\mathcal{X}_{k,\kappa,\iota}| \leq \kappa, \forall \kappa, \iota$. We restate the following result from (Dann and Brunskill 2015), with minor modification to accommodate randomized policies instead of deterministic policies, to provide a high probability upper bound on the number of episodes for which $|\mathcal{X}_{k,\kappa,\iota}| \leq \kappa, \forall \kappa, \iota$ is violated.

Lemma 3 *Let E be the number of episodes for which there are κ, ι with $|\mathcal{X}_{k,\kappa,\iota}| > \kappa$, and let $m \geq \frac{6H^2}{\epsilon} \ln \frac{2E_{max}}{\delta}$. Then, we obtain*

$$P(E \leq 6NE_{max}) \geq 1 - \delta/2,$$

where $N = |\mathcal{S}||\mathcal{A}|m$ and $E_{max} = \log_2 \frac{H}{w_{min}} \log_2 |\mathcal{S}||\mathcal{A}|$.

Difference between true and optimistic total cost. We use the following value difference lemma (Efroni, Mannor, and Pirotta 2020) to express the difference in value functions of policy π at time step h with respect to MDPs of different transition probabilities $p, \tilde{p}$, i.e., $V_h^{\pi} - \tilde{V}_h^{\pi}$, in terms of the value functions beyond h, $\tilde{V}_t^{\pi}, t > h$, and difference in transition probabilities $(p_t - \tilde{p}_t), t > h$. In the following, We also use the shorthand notations $V_h^{\pi}(s; c)$ and $\tilde{V}_h^{\pi}(s; c)$ for $V_h^{\pi}(s; c, p)$ and $\tilde{V}_h^{\pi}(s; c, \tilde{p})$, respectively. The cost function c is omitted when clear from the context.

Lemma 4 *Consider the MDPs M and $\tilde{M}$ denoted by $(\mathcal{S}, \mathcal{A}, p, c)$ and $(\mathcal{S}, \mathcal{A}, \tilde{p}, c)$, respectively. Then, the difference in the values with respect to the same policy π for any s, h can be written as*

$$V_h^{\pi}(s) - \tilde{V}_h^{\pi}(s) = \tag{5}$$

$$\mathbb{E}\left[\sum_{i=h}^{H} (p_i(\cdot|s_i, a_i) - \tilde{p}_i(\cdot|s_i, a_i)) \tilde{V}_{h+1}^{\pi} | \pi, p, s_h = s\right].$$

Moreover, we prove the following lemma which is used to upper bound the difference in transition probability $|p - \tilde{p}|$ in (5) in terms of $\tilde{p}$ and visitation counts n.

Lemma 5 *Let $\bar{p}, \tilde{p}, p \in [0,1]$, $\delta \in (0,1)$ such that $p, \tilde{p} \in CI$, where*

$$CI := \Bigg\{ p' \in [0,1] : |p' - \bar{p}| \leq \sqrt{\frac{2\bar{p}(1-\bar{p}) \ln \frac{4}{\delta}}{\max(1,n)}} + \quad (6)$$
$$+ \frac{7 \ln \frac{4}{\delta}}{3 \max(1, n-1)} \Bigg\}.$$

Then, $|\tilde{p} - p| \leq 2\sqrt{2}\sqrt{\frac{\tilde{p} \ln \frac{4}{\delta}}{\max(1,n-1)}} + 5(\frac{\ln \frac{4}{\delta}}{\max(1,n-1)})^{\frac{3}{4}} +$ $+ \frac{21 \ln \frac{4}{\delta}}{\max(1,n-1)}$.

The above lemma is proved by viewing (6) as a quadratic inequality in terms of $\sqrt{\bar{p}}$ and solving for $\bar{p}$. The resulting inequality is then substituted back in the original inequality to get the desired result.

For each phase k, the true transition probability p belongs to the confidence set B_p^k with high probability and the optimistic transition model $\tilde{p}^k$ is chosen from the confidence set. Then, p and $\tilde{p}^k$ belong to CI for a suitable δ, by definition of B_p^k. By Lemma 5, $|p - \tilde{p}^k|$ can then be upper bounded in terms of $\tilde{p}^k$ and n as described above. The following lemma upper bounds the summand in (5), $(p - \tilde{p}_h)(\cdot|s,a)\tilde{V}_{h+1}$, which is the difference between the expected values of successor states in MDPs with true transition probability p and optimistic transition probability model $\tilde{p}$.

Lemma 6 *Let*

$|p(s'|s,a) - \tilde{p}_h(s'|s,a)| \leq c_1(s,a) + c_2(s,a)\sqrt{\tilde{p}_h(s'|s,a)}$, *for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$. Then, for any policy π,*

$$|(p - \tilde{p}_h)(\cdot|s,a)\tilde{V}_{h+1}| \leq c_1(s,a)|Succ(s,a)|\|\tilde{V}_{h+1}\|_\infty +$$
$$+ c_2(s,a)\sqrt{|Succ(s,a)|}\tilde{\sigma}_h(s,a),$$

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, where $\tilde{\sigma}_h^2$ is the local variance function defined as: $\tilde{\sigma}_h^2(s,a) =$

$$\mathbb{E}[(\tilde{V}_{h+1}(s_{h+1}) - \mathbb{E}(\tilde{V}_{h+1}(s_{h+1})|s_h = s, \tilde{p}, \pi))^2$$
$$|s_h = s, a_h = a, \tilde{p}].$$

We then consider a sequence of MDPs $\mathcal{M}^{(d)}$ which have the same transition probability p of the true MDP but different cost functions $c^{(d)}$, and a similar sequence of MDPs $\tilde{\mathcal{M}}^{(d)}$ with the same transition probability $\tilde{p}$. In both sequences, for $d = 0$, the cost function is the same as that of the original cost function, i.e., $c_h^{(0)}$,$\tilde{c}_h^{(0)} = c_h$, $1 \leq h \leq H$. The following cost functions are defined recursively as $c_h^{(2d+2)}(s,a), \tilde{c}_h^{(2d+2)}(s,a) = \tilde{\sigma}_h^{(d),2}(s)$, where $\tilde{\sigma}_h^{(d),2}$ is the local variance of the value function under policy π with respect to the costs $c^{(d)}$, defined as $\tilde{\sigma}_h^{(d),2}(s) =$

$$\mathbb{E}\left[(\tilde{V}_{h+1}^{(d)}(s_{h+1}) - \mathbb{E}(\tilde{V}_{h+1}^{(d)}(s_{h+1})|s_h = s, \tilde{p}, \pi))^2 | s_h = s, \pi, \tilde{p}\right].$$

We note that $c_h^{(d)}(s,a) \in \left[0, H^d\right]$, and use the notation $V^{(d)}$ and $\tilde{V}^{(d)}$ for value functions of $\mathcal{M}^{(d)}$ and $\tilde{\mathcal{M}}^{(d)}$, respectively.

We also use the following lemma (Dann and Brunskill 2015) to bound $\sum_{i=1}^H \mathbb{E}\left[\tilde{\sigma}_i^2(s_i)|s_h = s, \tilde{p}, \pi\right]$ in Lemma 8 by $O(H^2)$ instead of the trivial $O(H^3)$.

Lemma 7 *The variance of the value function defined as $\mathcal{V}_h^\pi(s) = \mathbb{E}\left[(\sum_{i=h}^H c_i(s_i,a_i) - V_i^\pi(s_i))^2 | s_h = s, \pi\right]$ satisfies a Bellman equation $\mathcal{V}_h(s) = \mathbb{E}\left[\mathcal{V}_h(s_{h+1})|s_h = s, \pi\right] + \sigma_h^2(s)$, which gives $\mathcal{V}_h^\pi(s) = \sum_{i=h}^H \mathbb{E}\left[\sigma_i^2(s_i)|s_h = s, \pi\right]$. Since $0 \leq \mathcal{V}_1 \leq H^2 c_{max}^2$, for all $s \in \mathcal{S}$, we have $0 \leq \sum_{i=1}^H \mathbb{E}\left[\sigma_i^2(s_i)|s_h = s, \pi\right] \leq H^2 c_{max}^2$.*

If $p, \tilde{p} \in B_p^k$, the condition of Lemma 6 holds true by Lemma 5 for suitable constants. Then, by utilizing Lemmas 4, 6, and 7, we have the following recursive relation relating $|V_1^{(d)}(s_1) - \tilde{V}_1^{(d)}(s_1)|$ with $|V_1^{(2d+2)}(s_1) - \tilde{V}_1^{(2d+2)}(s_1)|$ when the condition $|\mathcal{X}_{\kappa,\iota}| \leq \kappa$ for all $(\kappa, \iota) \in \mathcal{K} \times \mathcal{I}$ holds. With constants m and δ' as defined in Algorithm 1, the analysis follows by splitting the state action pairs by importance, i.e., $(s,a) \in \mathcal{X}$ and $(s,a) \notin \mathcal{X}$ and using the definitions of weight w, knownness κ, and importance ι.

Lemma 8 *Let $p, \tilde{p} \in B_p^k$. If $|\mathcal{X}_{\kappa,\iota}| \leq \kappa$ for all (κ, ι), then*

$$|V_1^{(d)}(s_1) - \tilde{V}_1^{(d)}(s_1)| := \Delta_d \leq$$
$$\hat{A}_d + \hat{B}_d^1 + \hat{B}_d^2 + \min\{\hat{C}_d, \hat{C}'_d + \hat{C}''\sqrt{\Delta_{2d+2}}\}, \text{where}$$
$$\hat{A}_d = \frac{\epsilon H^d}{4}, \quad \hat{B}_d^1 = 42CH^{d+1}\left(\frac{|\mathcal{K} \times \mathcal{I}| \ln \frac{4}{\delta'}}{m}\right),$$
$$\hat{B}_d^2 = 10CH^{d+5/4}\left(\frac{|\mathcal{K} \times \mathcal{I}| \ln \frac{4}{\delta'}}{m}\right)^{3/4},$$
$$\hat{C}'_d = \sqrt{\frac{16C|\mathcal{K} \times \mathcal{I}|}{m} \ln \frac{4}{\delta'} H^{2d+2}}, \quad \hat{C}_d = \hat{C}'_d\sqrt{H},$$
$$\text{and } \hat{C}'' = \sqrt{\frac{16C|\mathcal{K} \times \mathcal{I}|}{m} \ln \frac{4}{\delta'}}.$$

This recurrence relation is simplified to show in Lemma 9 that, in phases with $|\mathcal{X}_{k,\kappa,\iota}| \leq \kappa$ for all κ, ι, the optimistic total expected cost $\tilde{V}_1^{\pi^k}(s_1)$ is close to that of the true one, $V_1^{\pi^k}(s_1)$. This lemma plays an important role in the final theorem to show that the policy obtained after sufficiently large number of episodes is ϵ-optimal with respect to the objective and constraints.

Lemma 9 *Let $p, \tilde{p} \in B_p^k$. If $|\mathcal{X}_{k,\kappa,\iota}| \leq \kappa$ for all κ, ι, $\epsilon \in (0,1]$, and*

$$m \geq \frac{2304C^2H^2}{\epsilon^2}(\log_2 \log_2 H)^2 \log_2^2 \frac{8H^2|\mathcal{S}|^2|\mathcal{A}|}{\epsilon} \ln \frac{4}{\delta'},$$

then $|V_1^{\pi^k}(s_1) - \tilde{V}_1^{\pi^k}(s_1)| \leq \epsilon$ holds.

Conclusions

We addressed the problem of finding approximately optimal policies for finite-horizon MDPs with constraints and unknown transition probability. We introduced the UC-CFH algorithm that is based on the optimism-in-the-face-of-uncertainty principle and offered, to the best of our knowledge, the first result in terms of provable PAC guarantees for both performance and constraint violations. Our PAC bound exhibits quadratic dependence on the horizon length. In the future, we plan to consider other types of constraints, e.g., chance or risk constraints, and extensions to the infinite-horizon setting.

Acknowledgments

This research was supported in part by the US National Science Foundation (NSF) under Awards 1839842 and 1846524, the Office of Naval Research (ONR) under Award N00014-20-1-2258, and the Defense Advanced Research Projects Agency (DARPA) under Award HR00112010003. The views, opinions, or findings contained in this article should not be interpreted as representing the official views or policies, either expressed or implied, by the US Government. Approved for public release; distribution is unlimited.

References

Achiam, J.; Held, D.; Tamar, A.; and Abbeel, P. 2017. Constrained Policy Optimization. *arXiv preprint arXiv:1705.10528* .

Altman, E. 1999. *Constrained Markov Decision Processes*, volume 7. CRC Press.

Auer, P.; Jaksch, T.; and Ortner, R. 2009. Near-optimal Regret Bounds for Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 89–96.

Auer, P.; and Ortner, R. 2005. Online Regret Bounds for a New Reinforcement Learning Algorithm. In *Proceedings 1st Austrian Cognitive Vision Workshop*.

Azar, M. G.; Osband, I.; and Munos, R. 2017. Minimax Regret Bounds for Reinforcement Learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 263–272.

Borkar, V. S. 2005. An Actor-Critic Algorithm for Constrained Markov Decision Processes. *Systems & control letters* 54(3): 207–213.

Brafman, R. I.; and Tennenholtz, M. 2002. R-max–A General Polynomial Time algorithm for Near-Optimal Reinforcement Learning. *Journal of Machine Learning Research* 3(Oct): 213–231.

Brantley, K.; Dudik, M.; Lykouris, T.; Miryoosefi, S.; Simchowitz, M.; Slivkins, A.; and Sun, W. 2020. Constrained Episodic Reinforcement Learning in Concave-Convex and Knapsack Settings. *arXiv preprint arXiv:2006.05051* .

Dann, C.; and Brunskill, E. 2015. Sample Complexity of Episodic Fixed-Horizon Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2818–2826.

Dann, C.; Lattimore, T.; and Brunskill, E. 2017. Unifying PAC and Regret: Uniform PAC Bounds for Episodic Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 5713–5723.

Ding, D.; Wei, X.; Yang, Z.; Wang, Z.; and Jovanović, M. R. 2020. Provably Efficient Safe Exploration via Primal-Dual Policy Optimization. *arXiv preprint arXiv:2003.00534* .

Efroni, Y.; Mannor, S.; and Pirotta, M. 2020. Exploration-Exploitation in Constrained MDPs. *arXiv preprint arXiv:2003.02189* .

Efroni, Y.; Merlis, N.; Ghavamzadeh, M.; and Mannor, S. 2019. Tight Regret Bounds for Model-Based Reinforcement Learning with Greedy Policies. In *Advances in Neural Information Processing Systems*, 12224–12234.

Jin, C.; Allen-Zhu, Z.; Bubeck, S.; and Jordan, M. I. 2018. Is Q-learning Provably Efficient? In *Advances in Neural Information Processing Systems*, 4863–4873.

Jin, T.; and Luo, H. 2019. Learning Adversarial MDPs with Bandit Feedback and Unknown Transition. *arXiv preprint arXiv:1912.01192* .

Lai, T. L.; and Robbins, H. 1985. Asymptotically Efficient Adaptive Allocation Rules. *Advances in applied mathematics* 6(1): 4–22.

Lattimore, T.; and Hutter, M. 2012. PAC Bounds for Discounted MDPs. In *International Conference on Algorithmic Learning Theory*, 320–334. Springer.

Le, H.; Voloshin, C.; and Yue, Y. 2019. Batch Policy Learning under Constraints. volume 97 of *Proceedings of Machine Learning Research*, 3703–3712.

Maurer, A.; and Pontil, M. 2009. Empirical Bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740* .

Miryoosefi, S.; Brantley, K.; Daume III, H.; Dudik, M.; and Schapire, R. E. 2019. Reinforcement Learning with Convex Constraints. In *Advances in Neural Information Processing Systems*, 14093–14102.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY, USA: John Wiley & Sons, Inc., 1st edition. ISBN 0471619779.

Qiu, S.; Wei, X.; Yang, Z.; Ye, J.; and Wang, Z. 2020. Upper Confidence Primal-Dual Reinforcement Learning for CMDP with Adversarial Loss.

Rosenberg, A.; and Mansour, Y. 2019. Online Convex Optimization in Adversarial Markov Decision Processes. In *International Conference on Machine Learning*, 5478–5486.

Singh, R.; Gupta, A.; and Shroff, N. B. 2020. Learning in Markov Decision Processes under Constraints. *arXiv preprint arXiv:2002.12435* .

Strehl, A. L.; Li, L.; and Littman, M. L. 2009. Reinforcement Learning in Finite MDPs: PAC Analysis. *Journal of Machine Learning Research* 10(11).

Tessler, C.; Mankowitz, D. J.; and Mannor, S. 2018. Reward Constrained Policy Optimization. *arXiv preprint arXiv:1805.11074* .

Zanette, A.; and Brunskill, E. 2019. Tighter Problem-Dependent Regret Bounds in Reinforcement Learning without Domain Knowledge using Value Function Bounds. volume 97 of *Proceedings of Machine Learning Research*, 7304–7312. PMLR.

Zheng, L.; and Ratliff, L. 2020. Constrained Upper Confidence Reinforcement Learning. *ArXiv* abs/2001.09377.

Zimin, A.; and Neu, G. 2013. Online Learning in Episodic Markovian Decision Processes by Relative Entropy Policy Search. In *Advances in Neural Information Processing Systems*, 1583–1591.