Efficient Exploration in Average-Reward Constrained Reinforcement Learning: Achieving Near-Optimal Regret With Posterior Sampling

Danil Provodin [1 2] **Maurits Kaptein** [1] **Mykola Pechenizkiy** [1 3]

Abstract

We present a new algorithm based on posterior sampling for learning in Constrained Markov Decision Processes (CMDP) in the infinite-horizon undiscounted setting. The algorithm achieves near-optimal regret bounds while being advantageous empirically compared to the existing algorithms. Our main theoretical result is a Bayesian regret bound for each cost component of $\tilde{O}(DS\sqrt{AT})$ for any communicating CMDP with S states, A actions, and diameter D. This regret bound matches the lower bound in order of time horizon T and is the best-known regret bound for communicating CMDPs achieved by a computationally tractable algorithm. Empirical results show that our posterior sampling algorithm outperforms the existing algorithms for constrained reinforcement learning.

1. Introduction

Reinforcement learning (RL) refers to the problem of learning by trial and error in sequential decision-making systems based on the scalar signal aiming to minimize the total cost accumulated over time. In many situations, however, the desired properties of the agent behavior are better described using constraints, as a single objective might not suffice to explain the real-life setting. For example, a robot should not only fulfill its task but should also control its wear and tear by limiting the torque exerted on its motors (Tessler et al., 2019); for telecommunication networks, it is necessary that the average end-to-end delay be limited, especially for voice traffic, while maximizing the throughput of the system (Altman, 1999); and autonomous driving vehicles should reach the destination in a time and fuel-efficient manner while obeying traffic rules (Le et al., 2019). A natural approach for handling such cases is specifying the problem using multiple objectives, where one objective is optimized subject to constraints on the others.

A typical way of formulating the constrained RL problem is a Constrained Markov Decision Process (CMDP) (Altman, 1999), which proceeds in discrete time steps. At each time step, the system occupies a *state*, and the decision maker chooses an *action* from the set of allowable actions. As a result of choosing the action, the decision maker receives a (possibly stochastic) vector of *costs*, and the system then transitions to the next state according to a fixed *state transition distribution*. In the reinforcement learning problem, the underlying state transition distributions and/or cost distributions are unknown and need to be learned from observations while aiming to minimize the total cost.

Learning in CMDPs has been a recurrent topic in the reinforcement learning literature, with numerous works addressing this challenge in episodic and discounted settings (see, e.g., Efroni et al. (2020); Brantley et al. (2020); Qiu et al. (2020); Liu et al. (2021); Kalagarla et al. (2023)). We consider the reinforcement learning problem in a more general *infinite-horizon average reward* setting. When decisions are made frequently so that the discount rate is very close to 1, the decision-makers may prefer to compare policies on the basis of their expected infinite-horizon average reward instead of the expected total discounted reward, and the objective becomes to achieve optimal long-term average performance under constraints. This criterion is especially relevant for inventory systems with frequent restocking decisions or queueing control theory, particularly when applied to controlling computer systems (Puterman, 1994).

Learning in CMDP in the *infinite-horizon average reward* setting appears to be more challenging because it depends on the limiting behavior of the underlying stochastic process, and approaches for analyzing this setting vary with the class structure of CMDPs. For instance, Singh et al. (2023) and Zheng & Ratliff (2020) consider a restricted class of *ergodic* CMDPs. In ergodic CMDPs, any policy will reach every state after a sufficient number of steps, making them self-explorative and easier to learn than general cases. Nevertheless, achieving near-optimal regret bounds is still non-trivial

[1]Eindhoven University of Technology, Eindhoven, The Netherlands [2]Jheronimus Academy of Data Science, 's-Hertogenbosch, The Netherlands [3]University of Jyväskylä, Jyväskylä, Finland. Correspondence to: Danil Provodin <d.provodin@tue.nl>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Table 1. Summary of work on provably efficient constrained RL in the infinite-horizon average reward setting. S and A represent the number of states and actions, m is the number of constraints, T is the total horizon, T_M is the mixing time, D is the diameter of CMDP, p represents transitions, and $sp(p)$ is the span of CMDP (defined in Section 2). $\tilde{O}$ hides logarithmic factors. The "Required knowledge" column denotes the information an algorithm requires as an input. The "Computation" column roughly denotes the time complexity with "Efficient" meaning an algorithm is designed to solve a problem using minimal resources, "Inefficient" – an algorithm consumes more time than necessary, and "Intractable" – an algorithm for which no known polynomial-time solution exists.

	Algorithm	Main Regret	Constraint violation	CMDP class	Required knowledge	Computation
frequentist	C-UCRL (Zheng & Ratliff, 2020)	$\tilde{O}(mSAT^{3/4})$	0	ergodic	safe policy π and p	efficient
	UCRL-CMDP (Singh et al., 2023)	$\tilde{O}(T_M\sqrt{SA}T^{2/3})$	$\tilde{O}(T_M\sqrt{SA}T^{2/3})$	ergodic	T	inefficient
	Alg. 3 (Chen et al., 2022)	$\tilde{O}(sp(p)(S^2AT^2)^{1/3})$	$\tilde{O}(sp(p)(S^2AT^2)^{1/3})$	weakly communicating	$sp(p), T$	inefficient
	Alg. 4 (Chen et al., 2022)	$\tilde{O}(sp(p)S\sqrt{AT})$	$\tilde{O}(sp(p)S\sqrt{AT})$	weakly communicating	$sp(p), T$	intractable
Bayesian	CMDP-PSRL (Agarwal et al., 2022)	$\tilde{O}(T_MS\sqrt{AT})$	$\tilde{O}(T_MS\sqrt{AT})$	ergodic	-	efficient
	PSConRL (this paper)	$\tilde{O}(DS\sqrt{AT})$	$\tilde{O}(DS\sqrt{AT})$	communicating	-	efficient
	lower bound (Singh et al., 2023)	$\Omega(\sqrt{DSAT})$	$\Omega(\sqrt{DSAT})$	-	-	-

under constraints, and the proposed algorithms only achieve suboptimal regret bounds: with UCRL-CMDP (Singh et al., 2023) achieving $\tilde{O}(T^{2/3})$ regret and cost violation bound and C-UCRL (Zheng & Ratliff, 2020) achieving $\tilde{O}(T^{3/4})$ regret bound with no cost violations. In contrast, Chen et al. (2022) consider a broad class of *weakly communicating* CMDPs, which allows more interesting practical scenarios. They propose two algorithms in this more general setting, albeit imposing impractical assumptions about knowledge of some problem-specific parameters. The first algorithm is computationally tractable but theoretically suboptimal, only achieving $\tilde{O}(T^{2/3})$ regret and cost violation bounds; the second is an intractable algorithm with near-optimal regret and cost violation bounds of $\tilde{O}(\sqrt{T})$. The main theoretical results for this setting are summarized in Table 1.

In this paper, we propose a practical *and* efficient algorithm based on the posterior sampling principle (Thompson, 1933). This principle involves maintaining a posterior distribution for the unknown parameters and guides the exploration by the variance of the distribution. The posterior sampling principle underpins many algorithms in reinforcement learning (Osband et al., 2013; Abbasi-Yadkori & Szepesvári, 2015; Agrawal & Jia, 2017; Ouyang et al., 2017).

Our main contribution is a posterior sampling-based algorithm (PSConRL), which achieves near-optimal Bayesian regret bounds while being computationally efficient. Drawing inspiration from the algorithmic design structure of Ouyang et al. (2017), the algorithm proceeds in episodes with two stopping criteria. At the beginning of every episode, it samples transition probability vectors from a posterior distribution for every state-action pair. The key idea of the algorithm is to switch to efficient exploration whenever the sampled transitions are infeasible, which we show to be necessary for communicating CMDPs. When sampled transitions are feasible, the algorithm solves for the optimal policy by utilizing a linear program (LP) in the space of occupancy measures that incorporates constraints directly (Altman, 1999). The optimal policy computed for the sampled CMDP is used throughout the episode. Under a Bayesian framework, we show that the expected regret and cost violation of our algorithm accumulated up to time T is bounded by $\tilde{O}(DS\sqrt{AT})$ for any communicating CMDP with S states, A actions, and diameter D.

A closely related study by Agarwal et al. (2022) analyzes the long-term average Bellman error in constrained optimization to address potential infeasibility issues of posterior sampling. They achieve the Bayesian regret and cost violation bounds of $\tilde{O}(T_MS\sqrt{AT})$, where T_M is the mixing time.[1] However, they focus on the ergodic CMDP structure, and, as detailed in Sections 3.2, their method cannot be applied to communicating CMDPs.

Thus, the main result of the paper shows that near-optimal Bayesian regret bounds are achievable in constrained reinforcement learning. To the best of our knowledge, this is the first work to obtain a computationally tractable algorithm with near-optimal regret bounds for the infinite-horizon average reward setting when underlying CMDP is communicating. Additionally, simulation results demonstrate that our algorithm significantly outperforms existing approaches for three CMDP benchmarks.

[1] Mixing time can be arbitrarily loose compared to the diameter, e.g., $T_M \sim O(D^S)$ for some problem instances (Bartlett & Tewari, 2009).

The rest of the paper is organized as follows. Section 2 is devoted to the methodological setup and contains the problem formulation. The PSConRL algorithm is introduced in Section 3. Analysis of the algorithm is presented in Section 4, which is followed by numerical experiments in Section 5. Section 6 briefly reviews the previous related work. Finally, we conclude with Section 7.

2. Problem formulation

2.1. Constrained Markov Decision Processes

A constrained MDP model is defined as a tuple $M = (\mathcal{S}, \mathcal{A}, p, \boldsymbol{c}, \tau)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $p : \mathcal{S} \times \mathcal{A} \to \Delta^{\mathcal{S}}$ is the transition function, with $\Delta^{\mathcal{S}}$ indicating simplex over $\mathcal{S}$, $\boldsymbol{c} : \mathcal{S} \times \mathcal{A} \to [0,1]^{m+1}$ is the cost vector function, and $\tau \in [0,1]^m$ is a cost threshold. In general, CMDP is an MDP with multiple cost functions $(c_0, c_1, \ldots, c_m)$, one of which, c_0, is used to set the optimization objective, while the others, $(c_1, \ldots, c_m)$, are used to restrict what policies can do. A stationary policy π is a mapping from state space $\mathcal{S}$ to a probability distribution on the action space $\mathcal{A}$, $\pi : \mathcal{S} \to \Delta^{\mathcal{A}}$, which does not change over time. Let $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$, where $|\cdot|$ denotes the cardinality.

For transitions p and a scalar cost function c, a stationary policy π induces a Markov chain, and the expected infinite-horizon average cost (*loss*) for state $s \in \mathcal{S}$ is defined as

$$J^{\pi}(s; c, p) = \overline{\lim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_p^{\pi} \left[c(s_t, a_t) | s_0 = s \right], \quad (1)$$

where $\mathbb{E}_p^{\pi}$ is the expectation under the probability measure $\mathbb{P}_p^{\pi}$ over the set of infinitely long state-action trajectories. $\mathbb{P}_p^{\pi}$ is induced by policy π, transition function p, and the initial state s. Given some fixed initial state s and $\tau_1, \ldots, \tau_m \in \mathbb{R}$, the CMDP optimization problem is to find a policy π that minimizes $J^{\pi}(s; c_0, p)$ subject to the constraints $J^{\pi}(s; c_i, p) \le \tau_i, i = 1, \ldots, m$:

$$\min_{\pi} J^{\pi}(s; c_0, p) \text{ s.t. } J^{\pi}(s; c_i, p) \le \tau_i,\ i = 1, \ldots, m\,. \quad (2)$$

Communicating CMDPs. To control the regret vector (defined below), we consider the subclass of communicating CMDPs. Formally, define the diameter of CMDP with transitions p as the minimum time required to go from one state to another in the CMDP using some stationary policy:

$$D(p) = \max_{s \neq s'} \min_{\pi : \mathcal{S} \to \Delta^{\mathcal{A}}} T_{s \to s'}^{\pi},$$

where $T_{s\to s'}^{\pi}$ is the expected number of steps to reach state s' when starting from state s and using policy π. CMDP is communicating if and only if it has a finite diameter, that is to say, for every pair of states s and s' there exists a stationary policy under which s' is accessible from s in at most $D(p)$ steps, for some finite $D(p) \ge 0$.

We define Ω_* to be the set of all transitions p such that the CMDP with transition probabilities p is communicating, and there exists a number D such that $D(p) \le D$. We will focus on CMDPs with transition probabilities in set Ω_*.

Next, by (Puterman, 1994)[Theorem 8.2.6], for scalar cost function c, transitions p that corresponds to communicating CMDP, and stationary policy π, there exists a bias function $v(s; c, p)$ satisfying the *Bellman equation* for all $s \in \mathcal{S}$:

$$J^{\pi}(s; c, p) + v^{\pi}(s; c, p) \quad (3)$$
$$= \sum_{a\in\mathcal{A}} \pi(a|s) \left[c(s, a) + \sum_{s'\in\mathcal{S}} p(s'|s, a) v^{\pi}(s'; c, p) \right].$$

If v satisfies the Bellman equation, v plus any constant also satisfies the Bellman equation. Furthermore, the loss of the optimal stationary policy π_* does not depend on the initial state, i.e., $J^{\pi_*}(s; c, p) = J^{\pi_*}(c, p)$, as presented in (Puterman, 1994)[Theorem 8.3.2]. Without loss of generality, let $\min_{s\in\mathcal{S}} v^{\pi_*}(s; c_i, p) = 0$, for $i = 1, \ldots m$, and define the span of the MDP as $sp(p) = \max_{1\le i\le m} \max_{s\in\mathcal{S}} v^{\pi_*}(s; c_i, p)$. Note, if $D(p) \le D$, then $sp(p) \le D$ as well (Bartlett & Tewari, 2009).

Linear programming for solving CMDPs. When CMDP is known, an optimal policy for (2) can be obtained by solving the following linear program (LP)(Altman, 1999):

$$\min_{\mu} \sum_{s,a} \mu(s, a) c_0(s, a), \quad (4)$$

$$\text{s.t.} \quad \sum_{s,a} \mu(s, a) c_i(s, a) \le \tau_i, \quad i = 1, \ldots, m, \quad (5)$$

$$\sum_{a} \mu(s, a) = \sum_{s',a} \mu(s', a) p(s', a, s), \quad \forall s \in \mathcal{S}, \quad (6)$$

$$\mu(s, a) \ge 0, \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}, \quad \sum_{s,a} \mu(s, a) = 1, \quad (7)$$

where the decision variable $\mu(s, a)$ is occupancy measure (fraction of visits to (s, a)). Given the optimal solution for LP (4)-(7), $\mu_*(s, a)$, one can construct the optimal stationary policy $\pi_*(a|s)$ for (2) by choosing action a in state s with probability $\frac{\mu_*(s,a)}{\sum_{a'} \mu_*(s,a')}$.

Given the above definitions and results, we can now define the reinforcement learning problem studied in this paper.

2.2. The reinforcement learning problem

We study the reinforcement learning problem where an agent interacts with a communicating CMDP $M = (\mathcal{S}, \mathcal{A}, p_*, \boldsymbol{c}, \tau)$. We assume that the agent has complete knowledge of $\mathcal{S}, \mathcal{A}$, and the cost function $\boldsymbol{c}$, but not the transitions p_* or the diameter D. This assumption is common

for RL literature (Bartlett & Tewari, 2009; Agrawal & Jia, 2017; Osband & Van Roy, 2017; Kalagarla et al., 2023) and is without loss of generality because the complexity of learning the cost and reward functions is dominated by the complexity of learning the transition probability.

We focus on a Bayesian framework for the unknown parameter p_*. That is, at the beginning of the interaction, the actual transition probabilities p_* are randomly generated from the prior distribution f_1. The agent can use past observations to learn the underlying CMDP model and decide future actions. The goal is to minimize the total cost $\sum_{t=1}^{T} c_0(s_t, a_t)$ while violating constraints as little as possible, or equivalently, minimize the total regret for the main cost component and auxiliary cost components over a time horizon T, defined as

$$BR_+(T; c_0) = \mathbb{E}\left[\sum_{t=1}^{T} \left(c_0(s_t, a_t) - J^{\pi_*}(c_0; p_*)\right)_+\right],$$

$$BR_+(T; c_i) = \mathbb{E}\left[\sum_{t=1}^{T} \left(c_i(s_t, a_t) - \tau_i\right)_+\right], i = 1, \ldots, m,$$

where $s_t, a_t, t = 1, \ldots, T$, are generated by the agent, $J^{\pi_*}(c_0; p_*)$ is the optimal loss of the CMDP M, and $[x]_+ := \max\{0, x\}$. The above expectation is with respect to the prior distribution f_1, the randomness in the state transitions, and the randomized policy.

2.3. Assumptions

We introduce two mild assumptions that are common in reinforcement learning literature.

Assumption 2.1. The support of the prior distribution f_1 is a subset of Ω_*. That is, the CMDP M is communicating and $D(p_*) \leq D$.

This type of assumption is common for the Bayesian framework (see, e.g., (Ouyang et al., 2017; Agarwal et al., 2022)) and is not overly restrictive (Bartlett & Tewari, 2009; Chen et al., 2022). In the experiments section, we provide a practical justification for this assumption and show that it can be supported by choosing Dirichlet distribution as a prior.

Assumption 2.2. There exists $\gamma > 0$ and unknown policy $\bar{\pi}(\cdot|s) \in \Delta^{\mathcal{A}}$ such that $J^{\bar{\pi}}(c_i, p_*) \leq \tau_i - \gamma$ for all $i \in \{1, \ldots, m\}$, and without loss of generality, we assume under such policy $\bar{\pi}$, the Markov chain resulting from the CMDP is irreducible and aperiodic.

The first part of the assumption is standard in constrained reinforcement learning (see, e.g., (Efroni et al., 2020; Ding et al., 2021)) and is mild as we do not require the knowledge of such policy. The second part is without loss of generality due to Puterman (1994)[Proposition 8.3.1] and Puterman (1994)[Proposition 8.5.8]. By imposing this assumption, we can control the sensitivity of problem (2) to the deviation between the true and sampled transitions. Later, we will use this assumption to guarantee that the minimization problem in Eq. (2) becomes feasible under the sampled transitions.

3. PSConRL: Learning algorithm for constrained reinforcement learning

In this section, we propose the Posterior Sampling for Constrained Reinforcement Learning (PSConRL) algorithm. Our algorithm is based on an intuitive idea of constructing an adaptive exploration mechanism to address the feasibility issues. It maintains posteriors for the transition function and combines the steps of solving LP through the lens of occupancy measure with the construction of exploration MDPs (whenever LP is infeasible). Below, we describe the main components of our algorithm, which is summarized in Algorithm 1.

Bayes rule. At each timestep t, given history h_t, the agent can compute posterior distribution f_t given by $f_t(\mathcal{P}) = \mathbb{P}(p_* \in \mathcal{P}|h_t)$ for any set $\mathcal{P}$. Upon applying action a_t and observing a new state s_{t+1}, the posterior distribution at $t+1$ can be updated according to Bayes' rule as

$$f_{t+1}(dp) = \frac{p(s_{t+1}|s_t, a_t) f_t(dp)}{\int p'(s_{t+1}|s_t, a_t) f_t(dp')}. \tag{8}$$

The key challenge of posterior sampling is that neither problem in Eq. (2) nor LP (4)-(7) are guaranteed to be feasible under the sampled transitions $p_t \sim f_t$, and it is unclear how the agent should proceed if LP (4)-(7) is infeasible. As we show in Lemma 4.4, after sufficient exploration, LP (4)-(7) becomes feasible with high probability (when each state-action pair is visited $\sqrt{T/A}$ times). Therefore, whenever LP (4)-(7) is infeasible, the agent switches to efficient exploration by constructing shortest path policies for a set of MDPs described below.

Reduction to a set of exploration MDPs. To facilitate efficient exploration, we introduce a set of MDPs, denoted as $\{(\mathcal{S}, \mathcal{A}, p_t, c_s)\}_{s \in \mathcal{S}}$. Each MDP in this set retains the original state and action spaces, with the transition function $p_t \sim f_t$ and a state-dependent cost function c_s, defined as

$$c_s(s', a) = \begin{cases} 1, & \text{if } s' \neq s; \\ 0, & \text{otherwise.} \end{cases}$$

Consider a specific target state $\bar{s}$ and its corresponding MDP $M_{\bar{s}}^t = (\mathcal{S}, \mathcal{A}, p_t, c_{\bar{s}})$. Note, MDP $M_{\bar{s}}^t$ is communicating with a scalar cost function, and, from MDP theory, we know that there exists an optimal policy $\pi_{\bar{s}}^t$ that satisfies the Bellman optimality equation:

$$J^*(c_{\bar{s}}, p_t) + v^*(s; c_{\bar{s}}, p_t) \tag{9}$$
$$= \min_{a \in \mathcal{A}} \left\{ c_{\bar{s}}(s, a) + \sum_{s' \in \mathcal{S}} p_t(s'|s, a) v^*(s'; c_{\bar{s}}, p_t) \right\}, \forall s \in \mathcal{S}.$$

In essence, the optimal policy $\pi_{\bar{s}}^t$ corresponds to a policy that efficiently guides the agent through the MDP toward the target state $\bar{s}$, thereby enabling efficient exploration. The formalization of this intuitive concept will be presented in Section 4.

3.1. Algorithm description

PSConRL begins with a prior distribution over transitions f_1 and proceeds in episodes. Let $N_t(s,a)$ denote the number of visits to (s,a) before time t and $N_t(s)$ denote the number of visits to s. We use two stopping criteria of Ouyang et al. (2017) for episode construction. The rounds $t = 1, ..., T$ are broken into consecutive episodes as follows: the k-th episode begins at the round t_k immediately after the end of $(k-1)$-th episode and ends at the first round t such that (i) $N_t(s,a) \geq 2N_{t_k}(s,a)$ or (ii) $t \leq t_k + T_{k-1}$ for some state-action pair (s,a), where $T_k = t_{k+1} - t_k$ is the length of episode k. The first criterion is the doubling trick of Jaksch et al. (2010) and ensures the algorithm has visited some state-action pair (s,a) at least the same number of times it had visited this pair (s,a) before episodes k started. The second criterion controls the growth rate of episode length and is believed to be necessary under the Bayesian setting (Ouyang et al., 2017).

At the beginning of episode k, a parameter p_k is sampled from the posterior distribution f_{t_k}, where t_k is the start of the k-th episode. During each episode k, actions are generated from the optimal stationary policy π_k for the sampled parameter p_k, which is observed either by solving LP (4)-(7) (if it is feasible) or by recovering the shortest path policy for a state with minimum visitations to it. Using Assumption 2.2, we will show that eventually, after $O(\sqrt{T})$ steps, the sampled CMDP becomes feasible, and the algorithm will effectively compute π_k by solving LP (4)-(7).

Remark 3.1. Note that PSConRL only requires the knowledge of $\mathcal{S}$, $\mathcal{A}$, $\boldsymbol{c}$, and the prior distribution f_1. It does not require the knowledge of the horizon T, or the bias span $sp(p)$ as in Singh et al. (2023) and Chen et al. (2022).

3.2. Importance of additional exploration for communicating CMDPs

In this subsection, we highlight the importance of reducing the problem to the exploration MDPs within our algorithm. In contrast to our approach, PSRL-CMDP (Agarwal et al., 2022) exclusively solves LP (4)-(7) for the optimal solution, and in cases when the optimal solution is infeasible, they opt to disregard constraints and proceed with the unconstrained problem. They argue that, eventually, the LP becomes feasible due to the self-exploratory properties of ergodic CMDPs. Unfortunately, this argument does not hold for communicating CMDPs, as demonstrated by the following example.

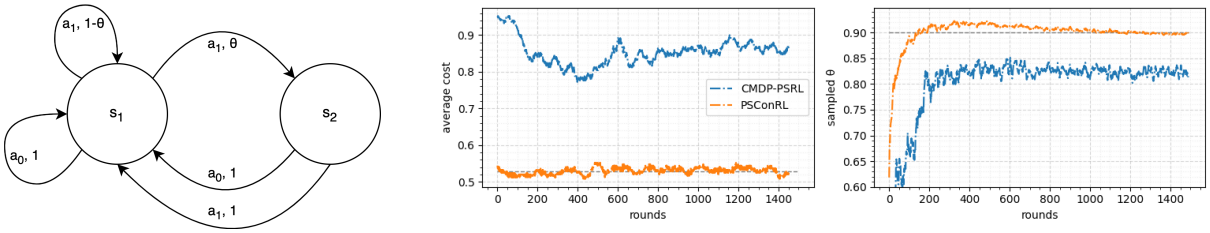
Algorithm 1 Posterior Sampling for Constrained Reinforcement Learning (PSConRL)

Input: f_1
Initialization: $t \leftarrow 1$, $t_k \leftarrow 0$, $\pi_0(\cdot) \leftarrow \frac{1}{|\mathcal{A}|}$
for episodes $k = 1, 2, \ldots$ **do**
  $T_{k-1} \leftarrow t - t_k$
  $t_k \leftarrow t$
  Generate $p_k(\cdot|s,a) \sim f_{t_k}$
  if LP (4)-(7) is feasible under $p_k(\cdot|s,a)$ **then**
    Compute $\pi_k(\cdot)$ by solving LP (4)-(7)
  else
    Select s with minimum number of visits $N_{t_k}(s)$
    Compute $\pi_k(\cdot)$ by solving Eq. (9) for MDP M_s
  end if
  repeat
    Apply action $a_t = \pi_k(s_t)$
    Observe new state s_{t+1}
    Update counter $N_t(s_t, a_t)$
    Update f_{t+1} according to Eq. (8)
    $t \leftarrow t + 1$
  until $t \leq t_k + T_{k-1}$ and $N_t(s,a) \leq 2N_{t_k}(s,a)$ for some $(s,a) \in \mathcal{S} \times \mathcal{A}$
end for

Example 3.2. Consider a two-state $\mathcal{S} = \{s_0, s_1\}$, two-action $\mathcal{A} = \{a_0, a_1\}$ CMDP in which the controlled transition probabilities $p_*(s_0, a_1, s_1) = \theta$ and $p_*(s_0, a_1, s_0) = 1 - \theta$ are unknown, while remaining probabilities are $p_*(s_0, a_0, s_1) = 1, p_*(s_1, \cdot, s_1) = 1$ and known. See Figure 1(a) for illustration. Assume that $r(s_0, \cdot) = 1$, $c(s_0, \cdot) = 1$ and $r(s_1, \cdot) = 0$, $c(s_1, \cdot) = 0$, i.e., reward and cost depend only upon the current state. Further, let $\theta = 0.9$ and the average cost threshold $\tau = 0.5275$ (the lowest possible budget that corresponds to a feasible problem). Note that this CMDP is not ergodic because starting from s_0 and utilizing a policy that chooses action a_0 will never visit state s_1. Also, such a policy would clearly correspond to an optimal solution in case there is no budget constraint.

For two algorithms, PSConRL (ours) and PSRL-CMDP (Agarwal et al., 2022), we demonstrate their performance through simulations on this toy CMDP. For both algorithms, we set the prior distribution to $Beta(1,1)$ with the parameters of the distribution being the number of visitations to (s_0, a_0) and (s_0, a_1) state-action pairs. At each timestep, we sample the plausible parameter $\tilde{\theta}$. Whenever the sampled CMDP is infeasible, we utilize the optimal policy for the unconstrained problem for PSRL-CMDP (policy that chooses action a_0 all the time) and the shortest path policy according to Eq. (9) for PSConRL (policy that chooses action a_1 all the time). Note that the sampled CMDP is infeasible every time $\tilde{\theta} < \theta$, due to the choice of cost threshold.

Figure 1(b) demonstrates the results of the experiment.



(a) Symbolic illustration of Example 3.2.

(b) Simulation results for Example 3.2.

Figure 1. CMDP illustration and results of the experiments for Example 3.2, with $\theta = 0.9$ and the average cost threshold $\tau = 0.5275$. Figure 1(a) represents the CMDP in symbolic form. Figure 1(b) presents average cost (left), and realizations of $\tilde{\theta}$ (right). Results are averaged over 5 runs.

Specifically, we present the average cost (left) and realizations of $\tilde{\theta}$ (right). Taking a closer look at the average cost subplot (left), we can see that PSConRL consistently fluctuates around the cost threshold and, overall, satisfies the constraint of the problem, whereas PSRL-CMDP severely violates the constraint. Moving to the right subplot, it is evident that PSConRL successfully learns the true value of parameter θ, while PSRL-CMDP fails to do so.

A series of assumptions in (Agarwal et al., 2022) makes a Markov chain induced by any policy aperiodic, recurrent, and irreducible. Such favorable properties make any CMDP self-exploratory, meaning that for a sampled CMDP, a policy that solely maximizes the main reward (regardless of constraints) will sufficiently explore the environment and, eventually, collect enough information to find the true optimal solution. However, this does not hold for communicating CMDPs and necessitates additional exploration to ensure feasibility. As such, a more involved theoretical analysis is required to address this issue for communicating CMDPs.

4. Regret bound

We now provide our main result for the PSConRL algorithm for learning in CMDPs.

Theorem 4.1. *For any communicating CMDP M with S states, A actions, under Assumptions 2.1 and 2.2, for $T \geq \Omega((D/\gamma)^4 S^2 A \log^2(2AT))$, the Bayesian regret for main and auxiliary cost components of Algorithm 1 are bounded:*

$$BR_+(T; c_i) \leq O\left(DS\sqrt{AT\log(AT)}\right), i = 0, \ldots, m.$$

Here $O(\cdot)$ notation hides only the absolute constant.E

Remark 4.2. The regret bound closely matches the theoretical lower bound of $\Omega(\sqrt{DSAT})$. Also, the provided bound matches the best bound for the undiscounted setting without constraints. We emphasize that the $O(\sqrt{DS})$ gap between lower and upper bounds remains an open question for the undiscounted setting with and without constraints.

The full proof of Theorem 4.1 is presented in the Appendix A.1. In the remainder of this section, we introduce three lemmas that are pivotal to our analysis and present a proof sketch for Theorem 4.1.

4.1. Key lemmas

A key property of posterior sampling is that conditioned on the information at time t, the transition functions p_* and p_t have the same distribution if p_t is sampled from the posterior distribution at time t (Osband et al., 2013). Since the PSConRL algorithm samples p_k at the stopping time t_k, we use the stopping time version of the posterior sampling property stated as follows.

Lemma 4.3 (Posterior sampling lemma; adapted from Lemma 1 of (Jafarnia-Jahromi et al., 2021))**.** *Let t_k be a stopping time with respect to the filtration $(\mathcal{F}_t)_{t=1}^{\infty}$, and p_k be the sample drawn from the posterior distribution at time t_k. Then, for any measurable function g and any $\mathcal{F}_{t_k}$-measurable random variable X, we have*

$$\mathbb{E}\left[g(p_k, X)\right] = \mathbb{E}\left[g(p_*, X)\right].$$

Recall that in every episode k, PSConRL runs either an optimal loss policy by solving LP (4)-(7) for the sampled transitions or computes the optimal stationary policy for a fixed finite MDP. In Lemma 4.4, we show that problem (2) becomes feasible under sampled transitions after sufficient exploration of every state-action pair, i.e., there exists a policy that satisfies constraints in problem (2) and Algorithm 1 will effectively find an optimal solution for LP (4)-(7).

We address the feasibility issues by using the deviation bound between sampled and true transitions and the limiting matrix properties of the resulting Markov chains. Unlike optimistic algorithms (Singh et al., 2023; Chen et al., 2022), which optimize over a confidence set of plausible transitions, Lemma 4.4 introduces a computationally efficient approach to deal with feasibility issues.

Lemma 4.4 (Feasibility lemma)**.** *If $N_{t_k}(s,a) \geq \sqrt{T/A}$, $\|p_k(\cdot|s,a) - p_*(\cdot|s,a)\|_1 \leq \sqrt{\frac{14S\log(2ATt_k)}{\max\{1, N_{t_k}(s,a)\}}}$ for all (s,a), and $\gamma \geq D\sqrt{\frac{14SA^{1/2}\log(2AT^2)}{\sqrt{T}}}$ there exists policy π, which satisfies $J^{\pi}(c_i, p_k) \leq \tau_i$ for all $i \in \{1, \ldots, m\}$.*

Next, in Lemma 4.5, we prove that PSConRL explores the environment efficiently, whenever LP (4)-(7) is infeasible, and requires $O(DS\sqrt{AT})$ to visit each state-action pair $\sqrt{T/A}$ times.

Lemma 4.5 (Exploration lemma)**.** *Define set* $\mathcal{G} = \{p \in \Omega_* : \exists \pi \text{ s.t. } J^\pi(c_i; p) \le \tau_i, \forall i \in \{1, \dots, m\}\}$. *Whenever* π_k *is computed as an optimal policy for Eq. (9), i.e.,* $p_k \notin \mathcal{G}$*, the average number of timesteps to visit each state-action pair* $\sqrt{T/A}$ *times is bounded by* $2DS\sqrt{AT}+1$*. Formally,*

$$\sum_{k:t_k \le T} \mathbb{E}\left[\sum_{t=t_k}^{t_{k+1}-1} \mathbb{I}\left\{\exists (s,a) : N_{t_k}(s,a) < \sqrt{T}\right\} \mid p_k \notin \mathcal{G}\right] \le 2DS\sqrt{AT} + 1.$$

Lemma 4.5 plays a crucial role in facilitating efficient exploration. It ensures that the deviation between sampled and true transitions becomes sufficiently small, thereby satisfying the conditions outlined in Lemma 4.4. Importantly, our exploration mechanism requires overall $O(\sqrt{T})$ steps, whereas the existing approaches designate $O(T^{2/3})$ steps for exploration in constrained problems and only achieve suboptimal regret of $\tilde{O}(T^{2/3})$, e.g., UCRL-CMDP (Singh et al., 2023) and Alg. 3 (Chen et al., 2022). Only Alg. 4 (Chen et al., 2022) allocates $O(\sqrt{T})$ steps for exploration, leading to near-optimal regret bound. However, this exploration scheme renders their algorithm intractable.

The Feasibility and Exploration lemmas form one of the main novel components of the analysis of Theorem 4.1.

4.2. Proof Sketch of Theorem 4.1

Below, we show a proof sketch of the main theorem for the main regret component. The proof for auxiliary cost components is deferred to the Appendix A.1.

We decompose the total regret into the sum of episodic regrets conditioned on the event that the sampled CMDP is feasible:

$$\begin{aligned} BR_+(T; c_0) &= \mathbb{E}\left[\sum_{t=1}^{T} (c_0(s_t,a_t) - J^{\pi_*})_+\right] \\ &= \sum_{k=1}^{K_T} \mathbb{E}\left[\sum_t [c_0(s_t,a_t) - J^{\pi_*}] \,|p_k \in \mathcal{G}\right] \mathbb{P}(p_k \in \mathcal{G}) \\ &+ \sum_{k=1}^{K_T} \mathbb{E}\left[\sum_t [c_0(s_t,a_t) - J^{\pi_*} | \, p_k \notin \mathcal{G}\right] \mathbb{P}(p_k \notin \mathcal{G}), \end{aligned}$$

where $J^{\pi_*} = J^{\pi_*}(c_0; p_*)$ is the optimal loss of CMDP M, K_T is the number of episodes, and $\mathcal{G}$ is defined in the statement of Lemma 4.5.

For the first term, conditioned on the good event, $\{p_k \in \mathcal{G}\}$, the sampled CMDP is feasible, and the standard analysis of Ouyang et al. (2017) can be applied. Lemma A.2 shows that this term can be bounded by $(D+1)\sqrt{2SAT\log(T)} + 49DS\sqrt{AT\log(AT)}$.

As for the second term, we further decompose it conditioned on two events: $A_1 = \{p_k \notin \mathcal{G} \wedge N_{t_k}(s,a) \ge \sqrt{T/A}, \forall s,a\}$ and $A_2 = \{p_k \notin \mathcal{G} \wedge \exists (s,a) : N_{t_k}(s,a) < \sqrt{T/A}\}$. Using the Feasibility lemma, we then show that $\mathbb{P}(A_1)$ is bounded by $2/15Tt_k$ for each k, and the total regret corresponding to event A_1 is negligible.

Next, conditioned on A_2, we can utilize the Exploration lemma and show that $\sum_k \mathbb{E}\left[\sum_t [c_0(s_t,a_t) - J^{\pi_*} | \, A_2\right] \mathbb{P}(A_2) < 2DS\sqrt{AT}+1$, due to the efficient exploration property of our algorithm.

Putting all bounds together, we obtain the resulting regret bound of:

$$BR_+(T; c_0) \le O\left(DS\sqrt{AT\log(AT)}\right).$$

5. Simulation results

In this section, we evaluate the performance of PSConRL. The source code of the experiments can be found at `https://github.com/danilprov/cmdp`.

We present PSConRL using Dirichlet priors with parameters $[0.1, \dots, 0.1]$. The Dirichlet distribution is a convenient choice for maintaining posteriors for the transition probability vectors $p(s,a)$ since it is a conjugate prior for categorical and multinomial distributions. Moreover, Dirichlet prior is proven to be highly effective for any underlying MDP in unconstrained problems (Osband & Van Roy, 2017).

We employ three algorithms as baselines: C-UCRL (Zheng & Ratliff, 2020), UCRL-CMDP (Singh et al., 2023), and FHA (Alg. 3) from (Chen et al., 2022). Both Alg. 4 of (Chen et al., 2022) and PSRL-CMDP of (Agarwal et al., 2022) are omitted from the empirical analysis due to their practical inapplicability. For additional information about the baselines, see Appendix B.1.

We run our experiments on three gridworld environments: Marsrover 4x4, Marsrover 8x8 (Zheng & Ratliff, 2020), and Box (Leike et al., 2017). To enable fair comparison, all algorithms were extended to the unknown reward/costs and unknown probability transitions setting (see Appendix B for more experimental details). Figure 2 illustrates the simulation results of all algorithms across three benchmark environments. The top row shows the cumulative regret of the main cost component. The bottom row presents the cumulative constraint violation.

We first analyze the behavior of the algorithm on Marsrover environments (left and middle columns). The cumulative regret (top row) shows that PSConRL consistently outperforms all three algorithms. Looking at the cumulative

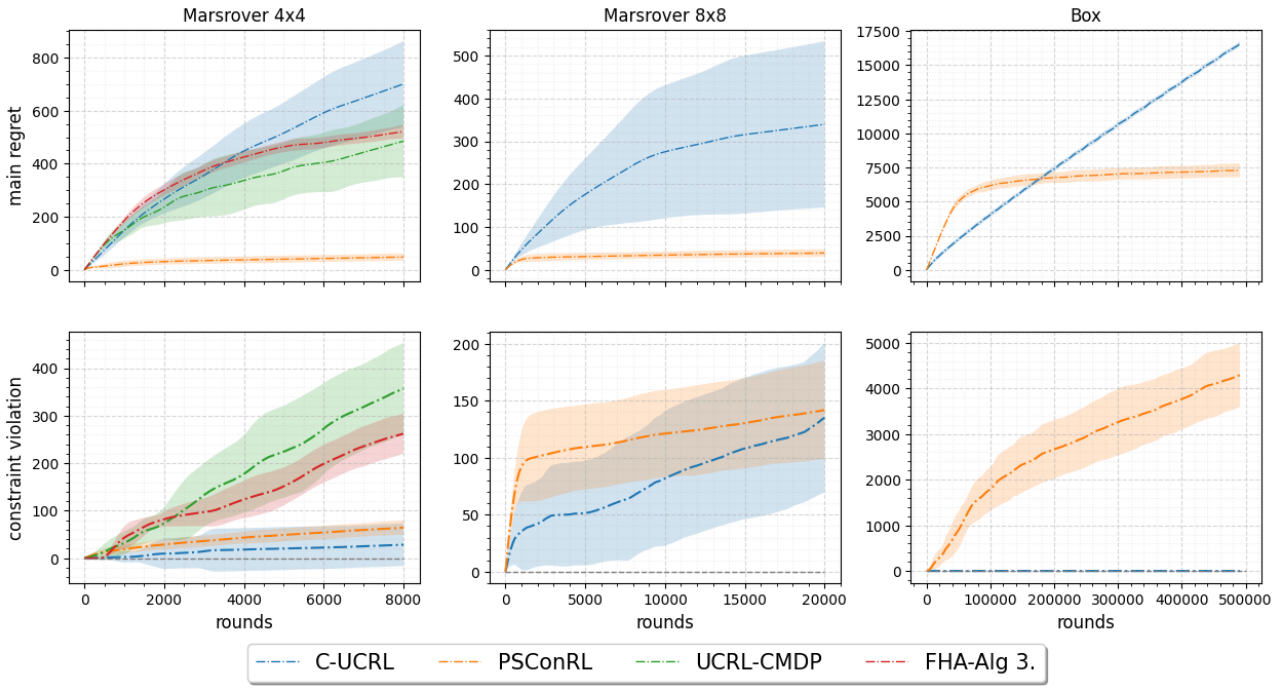


Figure 2. The main regret and constraint violation of the algorithms as a function of the horizon for Marsrover 4x4 (left column), Marsrover 8x8 (middle column), and Box (right column). **(Top row)** shows the cumulative regret of the main cost component. **(Bottom row)** shows the cumulative constraint violation. Results are averaged over 50 runs for Marsrover 4x4 and over 30 runs for Marsrover 8x8 and Box. Results for UCRL-CMDP and FHA (Alg. 3) are averaged over 10 runs for Marsrover 4x4.

constraint violation (bottom row), we see that PSConRL is comparable with C-UCRL, the only algorithm that addresses safe exploration. In the Box example (right column), PSConRL significantly outperforms C-UCRL, which incurs near-linear regret. We note that exploration is relatively costly in this benchmark compared to Marsrover environments (see the difference on the x and y-axes in the top row), which suggests that C-UCRL might be impractical in (at least some) problems where exploration is non-trivial. In Figure 5, we further elaborate on the average performance of the algorithms interpreting regret behavior.

We note the pronounced computational inefficiency of UCRL-CMDP and FHA (Alg. 3) algorithms. UCRL-CMDP involves optimization not only across the space of occupancy measures but also across the set of plausible CMDPs, resulting in an exhausting non-linear program. In the most favorable scenario, it requires $O((S^2A)^4)$ operations per episode. On the other hand, FHA (Alg. 3) maintains linearity in its main optimization program but necessitates solving it at each timestep, leading to time complexity of $O((SAT^{1/3})^2)$ per episode. Although both algorithms enjoy polynomial time complexity, the undesirable dependence on problem parameters makes them impractical even for moderate-sized problems. Due to these substantial drawbacks, we have limited their implementation to the Marsrover 4x4 environment.

Additionally, we would like to point out that in these examples, CMDPs are fixed and not generated from the Dirichlet prior. Therefore, we conjecture that PSConRL has the same regret bounds under a non-Bayesian setting.

6. Related work

Several algorithms based on the *optimism in the face of uncertainty* (OFU) principle have been proposed for constrained RL problems. For the episodic setting, both Efroni et al. (2020) and Brantley et al. (2020) consider sample efficient exploration utilizing a double optimism principle. As previously mentioned, Singh et al. (2023) and Chen et al. (2022) study OFU-based algorithms in the infinite-horizon average reward setting. It is worth mentioning that OFU-based algorithms often involve optimization across a set of plausible models (see, e.g., (Efroni et al., 2020; Singh et al., 2023)), which makes them computationally less appealing.

Another line of closely related works investigates safe RL, addressing constrained reinforcement learning problems with constant or zero constrain violation guarantees. Several algorithms were proposed in episodic setting (Liu et al., 2021; Wei et al., 2022; Kalagarla et al., 2023), with Kalagarla et al. (2023) focusing on posterior sampling algorithm for safe reinforcement learning. In the infinite-horizon average reward setting, the safe RL problem was previously

analyzed in (Zheng & Ratliff, 2020; Chen et al., 2022). Notably, safe RL algorithms often assume that the transition model and/or safe policy are known.

Among other related work, Lagrangian relaxation is a widely adopted technique for solving CMDPs. The works of (Achiam et al., 2017; Tessler et al., 2019) present constrained policy optimization approaches that demonstrate prominent successes in artificial environments. However, these approaches are notoriously sample-inefficient and lack theoretical guarantees. More scalable versions of the Lagrangian-based methods were proposed in (Chow et al., 2018; Qiu et al., 2020; Chen et al., 2021; Provodin et al., 2022). In general, the Lagrangian relaxation method can achieve high performance, but it is sensitive to the initialization of the Lagrangian multipliers and learning rate.

7. Conclusion

In this paper, we introduced the PSConRL algorithm for efficient exploration in constrained reinforcement learning under the infinite-horizon average reward criterion. Our algorithm achieves near-optimal Bayesian regret bounds for each cost component while being computationally efficient and easy to implement. By addressing these aspects, PSConRL fills a crucial gap in provably efficient constrained RL.

PSConRL leverages LP solutions to determine optimal policies and incorporates efficient exploration whenever the sampled CMDP is infeasible. As demonstrated in Section 3.2, the empirical comparison between PSConRL and CMDP-PSLR highlights that the exploration step is not merely a technical requirement for proofs but is indeed essential for effective learning in communicating CMDPs.

Finally, we validated our approach using simulations on three gridworld domains and showed that PSConRL quickly converges to the optimal policy even when CMDPs are not sampled from Dirichlet priors, consistently outperforming existing algorithms. Our insights suggest that the use of posterior sampling might be of great value for designing a computationally efficient algorithm with near-optimal frequentist regret bounds. Exploring this direction further is a promising avenue for future work. We also believe that this superior performance extends beyond the scope of gridworld domains to real-life applications.

Acknowledgements

This project is partially financed by the Dutch Research Council (NWO) and the ICAI initiative in collaboration with KPN, the Netherlands. The authors thank Pratik Gajane and Thiago D. Simão for discussions on earlier drafts of the paper.

Impact Statement

Our work focuses on the theoretical foundations of constrained reinforcement learning, emphasizing practical relevance and applicability. We believe that understanding the theoretical foundations is essential and, when coupled with addressing practically relevant issues, can directly guide the principled and effective application of these methods to real-life problems.

We believe that constraints represent a pivotal limitation in extending RL to real-life problems, such as nuclear fusion, medical treatment, and advertising. Consequently, the potential impact of our work extends to developing the foundations contributing to safe and responsible AI in the context of constrained reinforcement leanring.

References

Abbasi-Yadkori, Y. and Szepesvári, C. Bayesian optimal control of smoothly parameterized systems. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, 2015.

Achiam, J., Held, D., Tamar, A., and Abbeel, P. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Agarwal, M., Bai, Q., and Aggarwal, V. Regret guarantees for model-based reinforcement learning with long-term average constraints. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, 2022.

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

Altman, E. Constrained markov decision processes, 1999.

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

Brantley, K., Dudik, M., Lykouris, T., Miryoosefi, S., Simchowitz, M., Slivkins, A., and Sun, W. Constrained episodic reinforcement learning in concave-convex and knapsack settings. In *Advances in Neural Information Processing Systems*, 2020.

Chen, L., Jain, R., and Luo, H. Learning infinite-horizon average-reward Markov decision process with constraints. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Chen, Y., Dong, J., and Wang, Z. A primal-dual approach to constrained markov decision processes, 2021.

Chow, Y., Nachum, O., Duenez-Guzman, E., and Ghavamzadeh, M. A lyapunov-based approach to safe reinforcement learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018.

Ding, D., Wei, X., Yang, Z., Wang, Z., and Jovanovic, M. R. Provably efficient safe exploration via primal-dual policy optimization. In *The 24th International Conference on Artificial Intelligence and Statistics*, 2021.

Efroni, Y., Mannor, S., and Pirotta, M. Exploration-exploitation in constrained mdps, 2020.

Jafarnia-Jahromi, M., Chen, L., Jain, R., and Luo, H. Online learning for stochastic shortest path model via posterior sampling, 2021.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 2010.

Kalagarla, K. C., Jain, R., and Nuzzo, P. Safe posterior sampling for constrained mdps with bounded constraint violation, 2023.

Le, H., Voloshin, C., and Yue, Y. Batch policy learning under constraints. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., Orseau, L., and Legg, S. Ai safety gridworlds, 2017.

Liu, T., Zhou, R., Kalathil, D., Kumar, P., and Tian, C. Learning policies with zero or bounded constraint violation for constrained mdps. In *Advances in Neural Information Processing Systems*, 2021.

Osband, I. and Van Roy, B. Why is posterior sampling better than optimism for reinforcement learning? In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Osband, I., Russo, D., and Van Roy, B. (more) efficient reinforcement learning via posterior sampling. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, 2013.

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

Provodin, D., Gajane, P., Pechenizkiy, M., and Kaptein, M. An empirical evaluation of posterior sampling for constrained reinforcement learning, 2022.

Puterman, M. L. Markov decision processes: Discrete stochastic dynamic programming, 1994.

Qiu, S., Wei, X., Yang, Z., Ye, J., and Wang, Z. Upper confidence primal-dual reinforcement learning for cmdp with adversarial loss. In *Advances in Neural Information Processing Systems*, 2020.

Singh, R., Gupta, A., and Shroff, N. B. Learning in constrained markov decision processes. *IEEE Transactions on Control of Network Systems*, 2023.

Tessler, C., Mankowitz, D. J., and Mannor, S. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2019.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 1933.

Wei, H., Liu, X., and Ying, L. A provably-efficient model-free algorithm for infinite-horizon average-reward constrained markov decision processes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Zheng, L. and Ratliff, L. Constrained upper confidence reinforcement learning. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, 2020.

A. Omitted details for Section 4

A.1. Proof of Theorem 4.1

Bounding regret of the main cost component. To analyze the performance of PSConRL over T time steps, define $K_T = \arg\max\{k : t_k \leq T\}$, number of episodes of PSConRL until time T. By Ouyang et al. (2017)[Lemma 1], K_T is upper-bounded by $\sqrt{2SAT\log(T)}$. Using the tower rule, we can decompose the total regret into the sum of episodic regrets conditioned on the good event that the sampled CMDP is feasible:

$$\begin{aligned} BR_+(T;c_0) &= \mathbb{E}\left[\sum_{t=1}^{T}\left(c_0(s_t,a_t) - J^{\pi_*}(c_0;p_*)\right)_+\right] = \sum_{k=1}^{K_T}\mathbb{E}\left[R_{0,k}\right] \\ &= \sum_{k=1}^{K_T}\mathbb{E}\left[R_{0,k}|p_k \notin \mathcal{G}\right]\mathbb{P}\left(p_k \notin \mathcal{G}\right) + \sum_{k=1}^{K_T}\mathbb{E}\left[R_{0,k}|p_k \in \mathcal{G}\right]\mathbb{P}\left(p_k \in \mathcal{G}\right), \end{aligned} \tag{10}$$

where $R_{0,k} = \sum_{t=t_k}^{t_{k+1}-1}\left[c_0(s_t,a_t) - J^{\pi_*}(c_0;p_*)\right]_+$, $J^{\pi_*}(c_0;p_*)$ is the optimal loss of CMDP M, and $\mathcal{G}$ is defined in the statement of Lemma 4.5.

Define two events $A_1 = \{p_k \notin \mathcal{G} \wedge N_{t_k}(s,a) \geq \sqrt{T/A}, \forall s,a\}$ and $A_2 = \{p_k \notin \mathcal{G} \wedge \exists(s,a) : N_{t_k}(s,a) < \sqrt{T/A}\}$. Then, the first term of (10) can be further decomposed as

$$\sum_{k=1}^{K_T}\mathbb{E}\left[R_{0,k}|p_k \notin \mathcal{G}\right]\mathbb{P}\left(p_k \notin \mathcal{G}\right) = \sum_{k=1}^{K_T}\mathbb{E}\left[R_{0,k}|A_1\right]\mathbb{P}\left(A_1\right) + \sum_{k=1}^{K_T}\mathbb{E}\left[R_{0,k}|A_2\right]\mathbb{P}\left(A_2\right).$$

First, we bound $\sum_{k=1}^{K_T}\mathbb{E}\left[R_{0,k}|A_1\right]\mathbb{P}\left(A_1\right)$. Let $\bar{p}_k(s'|s,a) = \frac{N_{t_k}(s,a,s')}{N_{t_k}(s,a)}$ be the empirical mean for the transition probability at the beginning of episode k, where $N_{t_k}(s,a,s')$ is the number of visits to (s,a,s'). Define the confidence set

$$B_k = \left\{p : \left\|\bar{p}_k(\cdot|s,a) - p(\cdot|s,a)\right\|_1 \leq \beta_k\right\},$$

where $\beta_k = \sqrt{\frac{14S\log(2AT t_k)}{\max\{1,N_{t_k}(s,a)\}}}$.

Now, we observe that $\{A_1\} \subseteq \{\left\|p_k(\cdot|s,a) - p_*(\cdot|s,a)\right\|_1 > \beta_k\}$, otherwise, by Lemma 4.4, problem (2) would be feasible under p_k, and therefore $p_k \in \mathcal{G}$ which contradicts to $p_k \notin \mathcal{G}$. Next, we note that B_k is $\mathcal{F}_{t_k}$-measurable which allows us to use Lemma 4.3. Setting $\delta = 1/T$ in Lemma A.7 implies that $\mathbb{P}\left(\left\|p_k(\cdot|s,a) - p_*(\cdot|s,a)\right\|_1 > \beta_k\right)$ can be bounded by $\frac{2}{15Tt_k^6}$. Indeed,

$$\mathbb{P}\left(\left\|p_k(\cdot|s,a) - p_*(\cdot|s,a)\right\|_1 > \beta_k\right) \leq \mathbb{P}\left(p_* \notin B_k\right) + \mathbb{P}\left(p_k \notin B_k\right) = 2\mathbb{P}\left(p_* \notin B_k\right) \leq \frac{2}{15Tt_k^6},$$

where the last equality follows from Lemma 4.3 and the last inequality is due to Lemma A.7.

Finally, we have

$$\sum_{k=1}^{K_T}\mathbb{E}\left[R_{0,k}|A_1\right]\mathbb{P}\left(A_1\right) \leq \sum_{k=1}^{K_T}\frac{2(t_{k+1}-t_k)}{15Tt_k^6} \leq \frac{2}{15}\sum_{k=1}^{\infty}k^{-6} \leq 1.$$

To bound the term $\sum_{k=1}^{K_T}\mathbb{E}\left[R_{0,k}|A_2\right]\mathbb{P}\left(A_2\right)$, we rewrite it as

$$\begin{aligned} \sum_{k=1}^{K_T}\mathbb{E}\left[R_{0,k} \mid A_2\right]\mathbb{P}\left(A_2\right) &= \sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\mathbb{E}\left[\left(c_0(s_t,a_t) - J^{\pi_*}(c_0;p_*)\right)|A_2\right]\mathbb{P}\left(A_2\right) \\ &\leq \sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\mathbb{P}\left(A_2\right) \leq \sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\mathbb{P}\left(\exists(s,a) : N_{t_k}(s,a) < \sqrt{T/A} \mid p_k \notin \mathcal{G}\right), \end{aligned}$$

where the first inequality holds because $|(c_0(s_t,a_t)-J^{\pi_*}(c_0;p_*)|\leq 1$ and the last inequality is by $\mathbb{P}(A\wedge B)=\mathbb{P}(A|B)\mathbb{P}(B)$. Then, by Lemma 4.5, we obtain

$$\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\mathbb{P}\left(\exists(s,a):N_{t_k}(s,a)<\sqrt{T/A}\,|\,p_k\notin\mathcal{G}\right)\leq 2DS\sqrt{AT}+1.$$

For the second term of (10), conditioned on the good event, $\{p_k\in\mathcal{G}\}$, the sampled CMDP is feasible, and the standard analysis of Ouyang et al. (2017) can be applied. Lemma A.2 shows that this term can be bounded by $(D+1)\sqrt{2SAT\log(T)}+49DS\sqrt{AT\log(AT)}$.

Putting all bounds together, we obtain the resulting regret bound of:

$$BR_+(T;c_0)\leq O\left(DS\sqrt{AT\log(AT)}\right).$$

Bounding regret of auxiliary cost components. Without loss of generality, fix the cost component c_i and its threshold τ_i for some i and focus on analyzing the i-th component regret. Similarly to the decomposition of the main component, we obtain:

$$\begin{aligned}BR_+(T;c_i)&=\mathbb{E}\left[\sum_{t=0}^{T}\left(c_i(s_t,a_t)-\tau_i\right)_+\right]=\sum_{k=1}^{K_T}\mathbb{E}\left[R_{i,k}\right]\\&=\sum_{k=1}^{K_T}\mathbb{E}\left[R_{i,k}|p_k\notin\mathcal{G}\right]\mathbb{P}\left(p_k\notin\mathcal{G}\right)+\sum_{k=1}^{K_T}\mathbb{E}\left[R_{i,k}|p_k\in\mathcal{G}\right]\mathbb{P}\left(p_k\in\mathcal{G}\right)\end{aligned}$$

where $R_{i,k}=\sum_{t=t_k}^{t_{k+1}-1}\left[c_i(s_t,a_t)-\tau_i\right]_+$.

The first term can be analyzed similarly to the main cost component and bounded by $2DS\sqrt{AT}+2$. The regret bound of the second term is the same as the regret bound of the analogous term of the main cost component. Its analysis is marginally different and provided in Lemma A.3. □

A.2. Proof of Feasibility lemma (Lemma 4.4)

Proof. Fix some $i\in\{1,\dots,m\}$. Further, we will omit index i and write c and τ instead of c_i and τ_i.

With slight abuse of notation, we rewrite the equation (3) in vector form:

$$J^{\pi,p,c}+v^{\pi,p}=c_\pi+P_\pi v^{\pi,p}. \tag{11}$$

Above, $J^{\pi,p,c}$, $v^{\pi,p}$, and c_π are S dimensional vectors of $J_s^{\pi,p,c}$, $v_s^{\pi,p}$, and $c_{s,\pi(s)}$ with $J_s^{\pi,p,c}=J^\pi(s;c,p)$, $v_s^{\pi,p}=v^\pi(s;p)$, and $c_{s,\pi(s)}=\sum_{a\in\mathcal{A}}\pi(a|s)c(s,a)$; and P_π is the transition matrix whose rows formed by the vectors $p_{s,\pi(s)}$, where $p_{s,\pi(s)}=\sum_{a\in\mathcal{A}}\pi(a|s)p(\cdot|s,a)$.

Let $P_{\bar\pi}^k$ be the transition matrix whose rows are formed by the vectors $p_{s,\bar\pi(s)}^k$, and $P_{\bar\pi}^*$ be the transition matrix whose rows are formed by the vectors $p_{s,\bar\pi(s)}^*$. Since $N_{t_k}(s,a)\geq\sqrt{T/A}$ for all (s,a), $\|p_k(\cdot|s,a)-p_*(\cdot|s,a)\|_1\leq\sqrt{\frac{14S\log(2ATt_k)}{\max\{1,N_{t_k}(s,a)\}}}$, and the span of the bias function $v^{\bar\pi,p_*}$ is at most D (by Assumption 2.1), we observe

$$\left(p_k(\cdot|s,a)-p_*(\cdot|s,a)\right)^\intercal v^{\bar\pi,p_*}\leq\|p_k(\cdot|s,a)-p_*(\cdot|s,a)\|_1\left\|v^{\bar\pi,p_*}\right\|_\infty\leq\delta D$$

where $\delta=\sqrt{\frac{14SA^{1/2}\log(2ATt_k)}{\sqrt{T}}}$. Above implies

$$\left(P_{\bar\pi}^k-P_{\bar\pi}^*\right)v^{\bar\pi,p_*}\leq\delta D\mathbf{1} \tag{12}$$

where $\mathbf{1}$ is the vector of all 1s.

Following (Agrawal & Jia, 2017), let $(P_{\bar\pi}^k)^*$ denote the limiting matrix for Markov chain with transition matrix $P_{\bar\pi}^k$. Observe that $P_{\bar\pi}^k$ is aperiodic and irreducible because of Assumption 2.2. This implies that $(P_{\bar\pi}^k)^*$ is of the form $\mathbf{1}\boldsymbol{q}^\intercal$ where $\boldsymbol{q}$ is the stationary distribution of $P_{\bar\pi}^k$ (refer to (A.4) in (Puterman, 1994)). Also, $(P_{\bar\pi}^k)^*P_{\bar\pi}^k=(P_{\bar\pi}^k)^*$ and $(P_{\bar\pi}^k)^*\mathbf{1}=\mathbf{1}$.

Therefore, the gain of policy $\bar{\pi}$

$$J^{\bar{\pi},p_k,c}\mathbf{1} = (c_{\bar{\pi}}^{\intercal}\boldsymbol{q})\mathbf{1} = (P_{\bar{\pi}}^k)^* c_{\bar{\pi}}$$

Now,

$$\begin{aligned}
J^{\bar{\pi},p_k,c}\mathbf{1} - J^{\bar{\pi},p_*,c}\mathbf{1} &= (P_{\bar{\pi}}^k)^* c_{\bar{\pi}} - J^{\bar{\pi},p_*,c}\mathbf{1} \\
&= (P_{\bar{\pi}}^k)^* c_{\bar{\pi}} - J^{\bar{\pi},p_*,c}\left((P_{\bar{\pi}}^k)^*\mathbf{1}\right) && \text{(using } (P_{\bar{\pi}}^k)^*\mathbf{1} = \mathbf{1}\text{)} \\
&= (P_{\bar{\pi}}^k)^*\left(c_{\bar{\pi}} - J^{\bar{\pi},p_*,c}\mathbf{1}\right) \\
&= (P_{\bar{\pi}}^k)^*\left(I - P_{\bar{\pi}}^*\right) v^{\bar{\pi},p_*} && \text{(using (11))} \\
&= (P_{\bar{\pi}}^k)^*\left(P_{\bar{\pi}}^k - P_{\bar{\pi}}^*\right) v^{\bar{\pi},p_*} && \text{(using } (P_{\bar{\pi}}^k)^* P_{\bar{\pi}}^k = (P_{\bar{\pi}}^k)^*\text{)} \\
&\leq D\delta\mathbf{1}. && \text{(using (12) and } (P_{\bar{\pi}}^k)^*\mathbf{1} = \mathbf{1}\text{)}
\end{aligned}$$

Then observing that $D\delta \leq \gamma$, we obtain

$$J^{\bar{\pi}}(c, p_k) - J^{\bar{\pi}}(c, p_*) \leq D\delta \leq \gamma.$$

Using Assumption 2.2 and rearranging the terms in the inequality above, it follows

$$J^{\bar{\pi}}(c, p_k) \leq J^{\bar{\pi}}(c, p_*) + \gamma \leq \tau - \gamma + \gamma \leq \tau.$$

□

A.3. Proof of Exploration lemma (Lemma 4.5)

Before providing the proof for the Exploration lemma, we first show that PSConRL requires at most D timesteps to reach a target state when LP (4)-(7) is infeasible.

Lemma A.1. *Fix some target state $\bar{s}$ and its corresponding MDP $M_{\bar{s}}$ and let $\pi_{\bar{s}}$ be a solution of Eq. (9). Then $T_{s\to\bar{s}}^{\pi_{\bar{s}}} \leq D$.*

Proof. For simplicity, assume that MDP $M_{\bar{s}}$ is aperiodic (we will consider the general case later). In such MDP, value iteration is known to converge, and we can find (J^*, v^*) that satisfy Eq. (9) and the corresponding optimal policy $\pi_{\bar{s}}^*$.

Assume that there exists some policy π and state s such that $T_{s\to\bar{s}}^{\pi_{\bar{s}}^*} > T_{s\to\bar{s}}^{\pi}$. Consider the following policy π': follow π starting from s and wait until $\bar{s}$ is reached (suppose that this happens in τ steps), then follow the optimal policy $\pi_{\bar{s}}^*$. Note that τ is a random variable and, by definition,

$$\mathbb{E}[\tau] = T_{s\to\bar{s}}^{\pi'}.$$

Let $(J^{\pi'}, v^{\pi'})$ be the average cost and the bias function of policy π'. First, note that $J^* = J^{\pi'}$, since π' is constructed the way that some policy π is utilized for a finite number of steps and the same policy $\pi_{\bar{s}}^*$ is used in the long term. Next, if v is a bias vector, v plus any constant is also a bias vector. Therefore, without loss of generality, we can apply the following transformation to v^* and $v^{\pi'}$:

$$\begin{aligned}
v^* &= v^* - \min_{s\in\mathcal{S}} v^*(s), \\
v^{\pi'} &= v^{\pi'} - \min_{s\in\mathcal{S}} v^{\pi'}(s).
\end{aligned} \tag{13}$$

Observe that by definition of the cost function $c_{\bar{s}}$, after transformation (13), $v^*(s) = T_{s\to\bar{s}}^{\pi_{\bar{s}}^*}$ and $v^{\pi'}(s) = T_{s\to\bar{s}}^{\pi'}$. Thus, for state s, we obtain

$$J^* + v^*(s) > J^{\pi'} + v^{\pi'}(s),$$

which contradicts the optimality of (J^*, v^*).

Now, if the MDP $M_{\bar{s}}$ is periodic, we apply the aperiodicity transformation from Puterman (1994) to get a new MDP $\tilde{M}_{\bar{s}}$: choose θ satisfying $0 < \theta < 1$ and define $\tilde{\mathcal{S}} = \mathcal{S}$, $\tilde{\mathcal{A}} = \mathcal{A}$, and

$$\begin{aligned}
\tilde{c}_{\bar{s}} &= \theta c_{\bar{s}}, \\
\tilde{p}(\cdot|s,a) &= (1-\theta)\mathbf{e}_s + \theta p(\cdot|s,a).
\end{aligned}$$

Note that $\tilde{M}_{\bar{s}}$ is communicating and aperiodic, and the previous reasoning applies to $\tilde{M}_{\bar{s}}$. Let $\tilde{J}^\pi, \tilde{v}^\pi, \tilde{T}^\pi_{s\to\bar{s}}$ denote the quantities associated with $\tilde{M}_{\bar{s}}$ for some policy π. Then, by Puterman (1994)[Proposition 8.5.8], these are related to the corresponding quantities for $M_{\bar{s}}$ as follows:

$$\tilde{J}^\pi = J^\pi,$$
$$\tilde{v}^\pi = v^\pi,$$
$$\tilde{T}^\pi_{s\to\bar{s}} = \frac{T^\pi_{s\to\bar{s}}}{\theta}.$$

Using these relations and the fact that we proved the result for $\tilde{M}_{\bar{s}}$ gives us the result for periodic MDPs. Since $\min_\pi T^\pi_{s\to\bar{s}} \le \max_{s,s'} \min_\pi T^\pi_{s\to s'}$, it immediately follows that $T^{\pi^*_{\bar{s}}}_{s\to\bar{s}} \le D$. □

Proof of Lemma 4.5. Let T_e be the first time when every (s,a)-pair is visited at least $\sqrt{T/A}$ times given $\{p_k \notin \mathcal{G}\}$, $T_e = \min\{t : N_t(s,a) \ge \sqrt{T/A} \quad \forall (s,a) \,|\, p_k \notin \mathcal{G}\}$.

Since $T_k \le T_{k-1} + 1$ and $\mathbb{P}\left(\exists(s,a) : N_t(s,a) < \sqrt{T/A} \,|\, p_k \notin \mathcal{G}\right)$ is non-increasing in t, i.e., $\mathbb{P}\left(\exists(s,a) : N_t(s,a) < \sqrt{T/A} \,|\, p_k \notin \mathcal{G}\right) \le \mathbb{P}\left(\exists(s,a) : N_{t-1}(s,a) < \sqrt{T/A} \,|\, p_k \notin \mathcal{G}\right)$, for $k > 1$, we observe

$$\sum_{t=t_k}^{t_{k+1}-1} \mathbb{P}\left(\exists(s,a) : N_{t_k}(s,a) < \sqrt{T/A} \,|\, p_k \notin \mathcal{G}\right) \le \mathbb{P}\left(\exists(s,a) : N_{t_k}(s,a) < \sqrt{T/A} \,|\, p_k \notin \mathcal{G}\right) + \sum_{t=t_{k-1}}^{t_k-1} \mathbb{P}\left(\exists(s,a) : N_t(s,a) < \sqrt{T/A} \,|\, p_k \notin \mathcal{G}\right).$$

Next, by noting that $\mathbb{P}\left(\exists(s,a) : N_t(s,a) < \sqrt{T/A} \,|\, p_k \notin \mathcal{G}\right) = \mathbb{P}\left(T_e > t\right)$, we have

$$\begin{aligned}
&\sum_{k:t_k\le T} \mathbb{E}\left[\sum_{t=t_k}^{t_{k+1}-1} \mathbb{I}\left\{\exists(s,a) : N_{t_k}(s,a) < \sqrt{T/A}\right\} \,|\, p_k \notin \mathcal{G}\right] = \sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} \mathbb{P}\left(\exists(s,a) : N_{t_k}(s,a) < \sqrt{T/A} \,|\, p_k \notin \mathcal{G}\right) \\
&\le 1 + \sum_{k=2}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} \mathbb{P}\left(\exists(s,a) : N_{t_k}(s,a) < \sqrt{T/A} \,|\, p_k \notin \mathcal{G}\right) \\
&\le 1 + \sum_{k=2}^{K_T}\left[\mathbb{P}\left(\exists(s,a) : N_{t_k}(s,a) < \sqrt{T/A} \,|\, p_k \notin \mathcal{G}\right) + \sum_{t=t_{k-1}}^{t_k-1} \mathbb{P}\left(\exists(s,a) : N_t(s,a) < \sqrt{T/A} \,|\, p_k \notin \mathcal{G}\right)\right] \\
&= 1 + \sum_{k=2}^{K_T}\left[\mathbb{P}\left(T_e > t_k\right) + \sum_{t=t_{k-1}}^{t_k-1} \mathbb{P}\left(T_e > t\right)\right] = 1 + \sum_{k=2}^{K_T} \mathbb{P}\left(T_e > t_k\right) + \sum_{t=1}^{T} \mathbb{P}\left(T_e > t\right) \le 1 + 2\mathbb{E}[T_e],
\end{aligned}$$

where the last inequality follows from the tail sum formula $\mathbb{E}[T_e] = \sum_{t=0}^{\infty} \mathbb{P}(T_e > t)$. Finally, by Lemma A.1, we have $\mathbb{E}[T_e] \le DS\sqrt{AT}$, which gives

$$\sum_{k:t_k\le T} \mathbb{E}\left[\sum_{t=t_k}^{t_{k+1}-1} \mathbb{I}\left\{\exists(s,a) : N_{t_k}(s,a) < \sqrt{T/A}\right\} \,|\, p_k \notin \mathcal{G}\right] \le 2DS\sqrt{AT} + 1.$$

□

A.4. Auxiliary lemmas

Lemma A.2 (Regret of the main cost on the good event)**.** *Under Assumption 2.1, conditioned on the good event* $\{p_k \in \mathcal{G}\}$,

$$\sum_{k=1}^{K_T} \mathbb{E}\left[R_{0,k} | p_k \in \mathcal{G}\right] \mathbb{P}\left(p_k \in \mathcal{G}\right) \le (D+1)\sqrt{2SAT\log(T)} + 49DS\sqrt{AT\log(AT)}.$$

Most of the analysis here recovers the analysis of Ouyang et al. (2017). Nonetheless, for the sake of clarity, we provide the complete proof of Lemma A.2.

Proof. First, we rewrite equation (3) in terms of the state-action pair (Chen et al., 2022):

$$J^{\pi}(s;c,p) + q^{\pi}(s,a;p) = c(s,a) + \sum_{s'} p(s'|s,a) v^{\pi}(s';p), \tag{14}$$

where $v^{\pi}(s;p)$ and $q^{\pi}(s,a;p)$ are connected by $v^{\pi}(s;p) = \sum_a \pi(a|s) q^{\pi}(s,a;p)$.

Conditioned on the good event $\{p_k \in \mathcal{G}\}$, every policy π_k is the solution of LP (4)-(7), and we can apply the Bellman equation (14) to $c_0(s_t,a_t)$, and decompose $R_{0,k}$ into the following terms.

$$\begin{aligned}
&\sum_{k=1}^{K_T} \mathbb{E}\left[R_{0,k} | p_k \in \mathcal{G}\right] \mathbb{P}\left(p_k \in \mathcal{G}\right) \leq \sum_{k=1}^{K_T} \mathbb{E}\left[R_{0,k} | p_k \in \mathcal{G}\right] = \sum_{k=1}^{K_T} \mathbb{E}\left[\sum_{t=t_k}^{t_{k+1}-1} \left(c_0(s_t,a_t) - J^{\pi_*}(c_0;p_*)\right)\right] \\
&= \sum_{k=1}^{K_T} \mathbb{E}\left[\sum_{t=t_k}^{t_{k+1}-1} \left(J^{\pi_k}(c_0;p_k) - J^{\pi_*}(c_0;p_*) + q^{\pi_k}(s_t,a_t;p_k) - \sum_{s'\in\mathcal{S}} p_k(s'|s_t,a_t) v^{\pi_k}(s',p_k)\right)\right] \\
&= \underbrace{\sum_k \mathbb{E}\left[\sum_t \left(J^{\pi_k}(c_0;p_k) - J^{\pi_*}(c_0;p_*)\right)\right]}_{R_0} + \underbrace{\sum_k \mathbb{E}\left[\sum_t \left(q^{\pi_k}(s_t,a_t;p_k) - v^{\pi_k}(s_t;p_k)\right)\right]}_{R_1} \\
&+ \underbrace{\sum_k \mathbb{E}\left[\sum_t \left[v^{\pi_k}(s_t;p_k) - v^{\pi_k}(s_{t+1};p_k)\right]\right]}_{R_2} + \underbrace{\sum_k \mathbb{E}\left[\sum_t \left[v^{\pi_k}(s_{t+1};p_k) - \sum_{s'} p_k(s'|s_t,a_t) v^{\pi_k}(s';p_k)\right]\right]}_{R_3}.
\end{aligned}$$

Now, we note that $R_1 = 0$ as

$$\mathbb{E}[q^{\pi_k}(s_t,a_t;p_k) - v^{\pi_k}(s_t;p_k)] = \mathbb{E}[q^{\pi_k}(s_t,a_t;p_k) - \sum_a \pi_k(a|s_t) q^{\pi}(s_t,a;p_k)] = 0. \tag{15}$$

Next, applying lemmas A.4, A.5, A.6 to R_0, R_2, R_3, correspondingly, gives us the result. □

Lemma A.3 (Regret of the auxiliary costs on the good event)**.** *Under Assumption 2.1, conditioned on the good event* $\{p_k \in \mathcal{G}\}$,

$$\sum_{k=1}^{K_T} \mathbb{E}\left[R_{i,k} | p_k \in \mathcal{G}\right] \mathbb{P}\left(p_k \in \mathcal{G}\right) \leq (D+1)\sqrt{2SAT\log(T)} + 49DS\sqrt{AT\log(AT)}.$$

Proof. Similarly to Lemma A.2, conditioned on the good event $\{p_k \in \mathcal{G}\}$, we can decompose $R_{i,k}$ as follows:

$$\begin{aligned}
&\sum_{k=1}^{K_T} \mathbb{E}\left[R_{i,k} | p_k \in \mathcal{G}\right] \mathbb{P}\left(p_k \in \mathcal{G}\right) \leq \sum_{k=1}^{K_T} \mathbb{E}\left[R_{i,k} | p_k \in \mathcal{G}\right] = \sum_{k=1}^{K_T} \mathbb{E}\left[\sum_{t=t_k}^{t_{k+1}-1} \left(c_i(s_t,a_t) - \tau_i\right)\right] \\
&= \sum_{k=1}^{K_T} \mathbb{E}\left[\sum_{t=t_k}^{t_{k+1}-1} \left(J^{\pi_k}(c_i;p_k) - \tau_i + q^{\pi_k}(s_t,a_t;p_k) - \sum_{s'\in\mathcal{S}} p_k(s'|s_t,a_t) v^{\pi_k}(s',p_k)\right)\right] \\
&= \underbrace{\sum_k \mathbb{E}\left[\sum_t \left(J^{\pi_k}(c_i;p_k) - \tau_i\right]\right.}_{R_0} + \underbrace{\sum_k \mathbb{E}\left[\sum_t \left(q^{\pi_k}(s_t,a_t;p_k) - v^{\pi_k}(s_t;p_k)\right)\right]}_{R_1} \\
&+ \underbrace{\sum_k \mathbb{E}\left[\sum_t \left[v^{\pi_k}(s_t;p_k) - v^{\pi_k}(s_{t+1};p_k)\right]\right]}_{R_2} + \underbrace{\sum_k \mathbb{E}\left[\sum_t \left[v^{\pi_k}(s_{t+1};p_k) - \sum_{s'} p_k(s'|s_t,a_t) v^{\pi_k}(s';p_k)\right]\right]}_{R_3}.
\end{aligned}$$

Now, we note that $(J^{\pi_k}(c_i; p_k) - \tau_i)$ is negative on the good event $\{p_k \in \mathcal{G}\}$ for all k, and term R_0 can be dismissed. $R_1 = 0$ because of (15), and R_2 and R_3 regret terms can be bounded by Lemmas A.5 and A.6 correspondingly. □

Lemma A.4 (Lemma 3 from (Ouyang et al., 2017)). *For any cost function $c : \mathcal{S} \times \mathcal{A} \to [0, 1]$,*

$$\mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\left(J^{\pi_k}(c; p_k) - J^{\pi_*}(c; p_*)\right)\right] \leq K_T \leq \sqrt{2SAT\log(T)}.$$

Lemma A.5 (Lemma 4 from (Ouyang et al., 2017)).

$$\mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\left(v^{\pi_k}(s_t; p_k) - v^{\pi_k}(s_{t+1}; p_k)\right)\right] \leq DK_T \leq D\sqrt{2SAT\log(T)}.$$

Lemma A.6 (Lemma 5 from (Ouyang et al., 2017)).

$$\mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\left(v^{\pi_k}(s_{t+1}; p_k) - \sum_{s'\in\mathcal{S}} p_k(s'|s_t, a_t)v^{\pi_k}(s'; p_k)\right)\right] \leq 49DS\sqrt{AT\log(AT)}.$$

Lemma A.7 (Lemma 17 from (Jaksch et al., 2010)). *For any $t \geq 1$, the probability that the true MDP M is not contained in the set of plausible MDPs $\mathcal{M}(t) = \left\{(\mathcal{S}, \mathcal{A}, p', \boldsymbol{c}, \tau, \rho) : \|p'(\cdot|s, a) - p_k(\cdot|s, a)\|_1 \leq \sqrt{\frac{14S\log(2At_k/\delta)}{\max\{1, N_{t_k}(s,a)\}}}\right\}$ at time t is at most $\frac{\delta}{15t}$, that is*

$$\mathbb{P}\{M \notin \mathcal{M}(t)\} < \frac{\delta}{15t^6}.$$

B. Experimental details

B.1. Baselines: OFU-based algorithms

We use three OFU-based algorithms from the existing literature for comparison: C-UCRL (Zheng & Ratliff, 2020), UCRL-CMDP (Singh et al., 2023), and FHA (Alg. 3) (Chen et al., 2022). These algorithms rely on the knowledge of different CMDP components, e.g., UCRL-CMDP relies on knowledge of rewards r, whereas C-UCRL uses the knowledge of transitions p. To enable fair comparison, all algorithms were extended to the unknown reward/costs and unknown probability transitions setting. Specifically, we assume that each algorithm knows only the states space $\mathcal{S}$ and the action space $\mathcal{A}$, substituting the unknown elements with their empirical estimates:

$$\bar{r}_t(s, a) = \frac{\sum_{j=1}^{t-1}\mathbb{I}\{s_t = s, a_t = a\}r_t}{N_t(s, a) \vee 1}, \quad \forall s \in \mathcal{S}, a \in \mathcal{A}, \tag{16}$$

$$\bar{c}_{i,t}(s, a) = \frac{\sum_{j=1}^{t-1}\mathbb{I}\{s_t = s, a_t = a\}c_{i,t}}{N_t(s, a) \vee 1}, \quad \forall s \in \mathcal{S}, a \in \mathcal{A}, \quad i = 1 \ldots, m, \tag{17}$$

$$\bar{p}_t(s, a, s') = \frac{N_t(s, a, s')}{N_t(s, a) \vee 1}, \quad \forall s, s' \in \mathcal{S}, a \in \mathcal{A}. \tag{18}$$

where r is the reward function (inverse main cost c_0) and $N_t(s, a)$ and $N_t(s, a, s')$ denote the number of visits to (s, a) and (s, a, s') respectively.

Further, we provide algorithmic-specific details separately for each baseline:

1. C-UCRL follows a principle of "optimism in the face of reward uncertainty; pessimism in the face of cost uncertainty." This algorithm, which was developed in (Zheng & Ratliff, 2020), considers conservative (safe) exploration by overestimating both rewards and costs:

$$\hat{r}_t(s,a) = \bar{r}_t(s,a) + b_t(s,a) \quad \text{and} \quad \hat{c}_t(s,a) = \bar{r}_t(s,a) + b_t(s,a).$$

C-UCRL proceeds in episodes of linearly increasing number of rounds kh, where k is the episode index and h is the fixed duration given as an input. In each epoch, the random policy [2] is executed for h steps for additional exploration, and then policy π_k is applied for $(k-1)h$ number of steps, making kh the total duration of episode k.

2. Unlike the previous algorithm, where uncertainty was taken into account by enhancing rewards and costs, UCRL-CMDP (Singh et al., 2023) constructs confidence set $\mathcal{C}_t$ over $\bar{p}_t$:

$$\mathcal{C}_t = \{p : |p(s,a,s') - \bar{p}_t(s,a,s')| \leq b_t(s,a) \quad \forall (s,a)\}.$$

UCRL-CMDP algorithm proceeds in episodes of fixed duration of $\lceil T^\alpha \rceil$, where α is an input of the algorithm. At the beginning of each round, the agent solves the following constrained optimization problem in which the decision variables are (i) Occupation measure $\mu(s,a)$, and (ii) "Candidate" transition p':

$$\max_{\mu, p' \in \mathcal{C}_t} \sum_{s,a} \mu(s,a) r(s,a), \tag{19}$$

$$\text{s.t.} \quad \sum_{s,a} \mu(s,a) c_i(s,a) \leq \tau_i, \quad i = 1, \ldots, m, \tag{20}$$

$$\sum_a \mu(s,a) = \sum_{s',a} \mu(s',a) p'(s',a,s), \quad \forall s \in \mathcal{S}, \tag{21}$$

$$\mu(s,a) \geq 0, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \quad \sum_{s,a} \mu(s,a) = 1, \tag{22}$$

Note that program (19)-(22) is not linear anymore as $\mu(s',a)$ is being multiplied by $p'(s',a,s)$ in equation (21). This is a serious drawback of UCRL-CMDP algorithm because, as we show in the experiments, program (19)-(22) becomes computationally inefficient for even moderate problems.

3. FHA (Finite Horizon Approximation for CMDP) divides the T timesteps into K rounds and treats each episode as an episodic finite-horizon CMDP. Fix some episode k. Through the lens of occupancy measure that is defined on $S \times A \times H \times S$ space (where H is the length of the episode), FHA optimizes the following linear program:

$$\max_{\mu} \sum_h \sum_{s,a} r(s,a) \sum_{s'} \mu(s,a,h,s'), \tag{23}$$

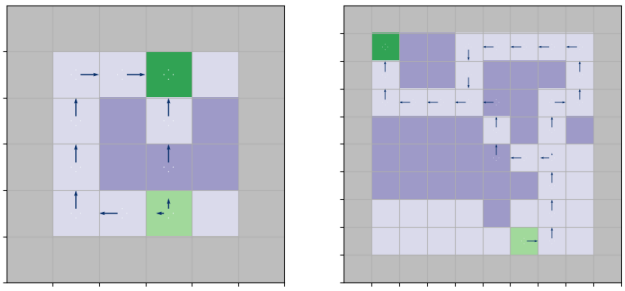
$$s.t. \sum_h \sum_{s,a} c_i(s,a) \sum_{s'} \mu(s,a,h,s') \leq H\tau_i + sp(p_*), \tag{24}$$

$$P_\mu \in \{p : |p(s,a,s') - \bar{p}_k(s,a,s')| \leq b_k(s,a) \quad \forall (s,a)\},, \tag{25}$$

where $P_\mu(s,a,s') = \frac{\mu(s,a,h,s')}{\sum_{s'} \mu(s,a,h,s')} \quad \forall h = 1, \ldots H.$

Although program (23)-(25) is linear, we emphasize that this algorithm requires finding an optimal occupancy measure for each H and each K, resulting in $O(S^2AT)$ decision variables. As we mentioned in the experiments, this is prohibitive even for moderate-sized CMDPs.

[2] Original algorithm utilizes a safe baseline during the first h rounds in each epoch, which is assumed to be known. However, to make the comparison as fair as possible, we assume that a random policy is applied instead.



(a) Marsrover 4x4 (b) Marsrover 8x8

Figure 3. Marsrover gridworlds. The initial position is light green, the goal is dark green, the walls are gray, and risky states are purple. Figure 3(a) illustrates 4x4 Marsrover environment. Figure 3(b) illustrates 8x8 Marsrover environment. In both cases, the agent's task is to get from the initial state to the goal state, and the optimal policy combines with some probabilities fast and safe ways, which are indicated by arrows on the pictures.

B.2. Environments

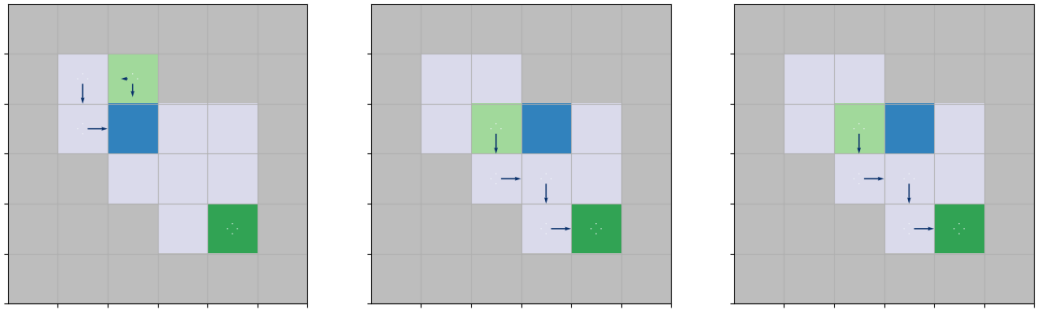
To demonstrate the performance of the algorithms, we consider three gridworld environments in our analysis. There are four actions possible in each state, $\mathcal{A} = \{up, down, right, left\}$, which cause the corresponding state transitions, except that actions that would take the agent to the wall leave the state unchanged. Due to the stochastic environment, transitions are stochastic (i.e., even if the agent's action is to go *up*, the environment can send the agent with a small probability *left*). Typically, the gridworld is an episodic task where the agent receives cost 1 (equivalently reward -1) on all transitions until the terminal state is reached. We reduce the episodic setting to the infinite-horizon setting by connecting terminal states to the initial state. Since there is no terminal state in the infinite-horizon setting, we call it the goal state instead. Thus, every time the agent reaches the goal, it receives a cost of 0 (or reward of 0), and every action from the goal state sends the agent to the initial state. We introduce constraints by considering the following specifications of a gridworld environment: Marsrover and Box environments.

Marsrover. This environment was used in (Tessler et al., 2019; Zheng & Ratliff, 2020; Brantley et al., 2020). The agent must move from the initial position to the goal avoiding risky states. Figure (3) illustrates the CMDP structure: the initial position is light green, the goal is dark green, the walls are gray, and risky states are purple. "In the Mars exploration problem, those darker states are the states with a large slope that the agents want to avoid. The constraint we enforce is the upper bound of the per-step probability of stepping into those states with large slope – i.e., the more risky or potentially unsafe states to explore" (Zheng & Ratliff, 2020). Each time the agent appears in a purple state incurs an auxiliary cost of 1. Other states incur no auxiliary costs.

Without constraints, the optimal policy is obviously to always go *up* from the initial state. However, with constraints, the optimal policy is a randomized policy that goes *left* and *up* with some probabilities, as illustrated in Figure 3(a). In experiments, we consider two marsrover gridworlds: 4x4, as shown in Figure 3(a), and 8x8, depicted in Figure 3(b).

Box. Another conceptually different specification of a gridworld is Box environment from (Leike et al., 2017). Unlike the Marsrover example, there are no static risky states; instead, there is an obstacle, a box, which is only "pushable" (see Figure 4(a)). Moving onto the blue tile (the box) pushes the box one tile into the same direction if that tile is empty; otherwise, the move fails as if the tile were a wall. The main idea of Box environment is "to minimize effects unrelated to their main objectives, especially those that are irreversible or difficult to reverse" (Leike et al., 2017). If the agent takes the fast way (i.e., goes down from its initial state; see Figure 4(c)) and pushes the box into the corner, the agent will never be able to get it back, and the initial configuration would be irreversible. In contrast, if the agent chooses the safe way (i.e., approaches the box from the left side), it pushes the box to the reversible state (see Figure 4(b)). This example illustrates situations of performing the task without breaking a vase in its path, scratching the furniture, bumping into humans, etc.

Each action incurs an auxiliary cost of 1 if the box is in a corner (cells adjacent to at least two walls) and no auxiliary costs otherwise. Similarly to the Marsrover example, without safety constraints, the optimal policy is to take a fast way (go down from the initial state). However, with constraints, the optimal policy is a randomized policy that goes down and left from the initial state.



(a) Box (Main) (b) Box (Safe) (c) Box (Fast)

Figure 4. Box gridworld. The initial position is light green, the goal is dark green, the walls are gray, and risky states are purple. Figure 4(a) illustrates the initial configuration. The agent's task is to get from the initial state to the goal state, and the optimal policy combines with some probabilities fast and safe ways, which are indicated by arrows on the pictures. Figure 4(b)-4(c) illustrates safe and fast ways.

B.3. Simulation results

Figure 5 shows the reward (inverse main cost) and average consumption (auxiliary cost) behavior of PSConRL, C-UCRL, UCRL-CMDP, and FHA (Alg. 3) illustrating how the regret from Figure 2 is accumulated. The top row shows the reward performance. The bottom row presents the average consumption of the auxiliary cost.

Taking a closer look at Marsrover environments (left and middle columns), we see that all algorithms converge to the optimal solution (top row), and their average consumption (middle row) satisfies the constraints in the long run. In the Box example (right column), we see that C-UCRL is stuck with the suboptimal solution. The algorithm exploits safe policy once it is learned, which corresponds to the near-linear regret behavior in Figure 2. Alternatively, PSConRL converges to the optimal solution relatively quickly (middle and bottom graphs).

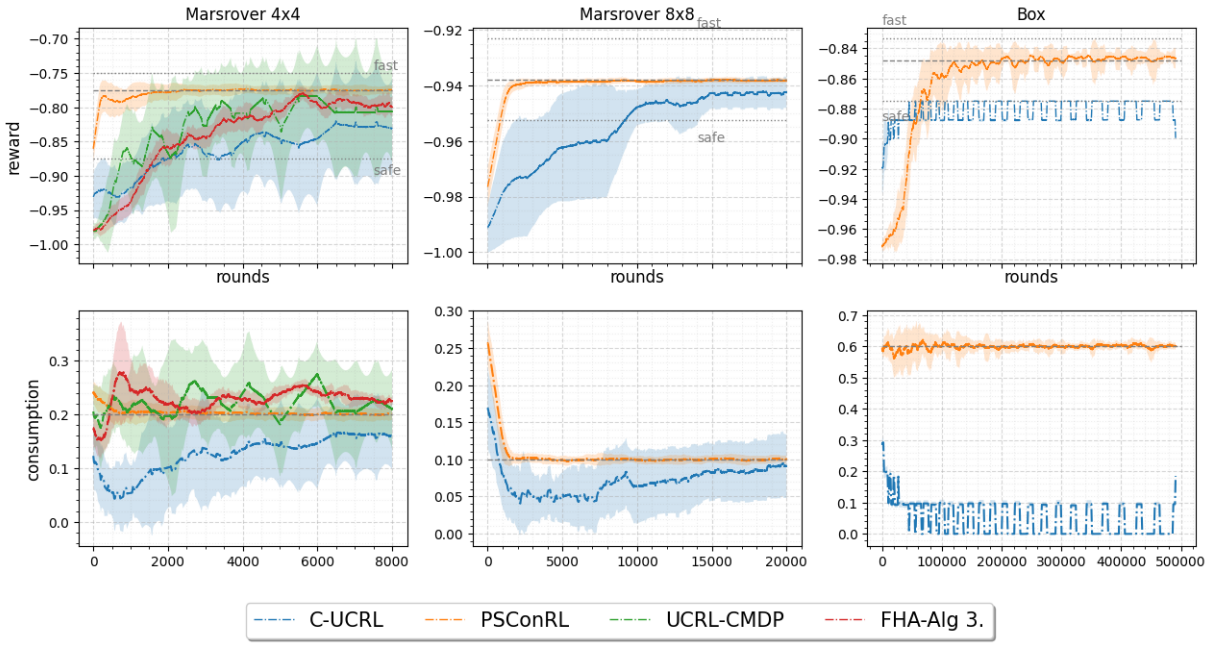


Figure 5. **(Top row)** shows the average reward (inverse average main cost); the dashed line shows the optimal behavior, and the dotted lines depict the reward level of safe and fast policies. **(Bottom row)** shows the average consumption of the auxiliary cost; the constraint thresholds are 0.2 for Marsrover 4x4, 0.1 for Marsrover 8x8, and 0.6 for Box. Results are averaged over 100 runs for Marsrover 4x4 and over 30 runs for Marsrover 8x8 and Box.