Safe Exploration for Constrained Reinforcement Learning with Provable Guarantees

Archana Bura, Aria HasanzadeZonuzy, Dileep Kalathil, Srinivas Shakkottai, and Jean-Francois Chamberland *

Abstract

We consider the problem of learning an episodic safe control policy that minimizes an objective function, while satisfying necessary safety constraints – both during learning and deployment. We formulate this safety constrained reinforcement learning (RL) problem using the framework of a finite-horizon Constrained Markov Decision Process (CMDP) with an unknown transition probability function. Here, we model the safety requirements as constraints on the expected cumulative costs that must be satisfied during all episodes of learning. We propose a model-based safe RL algorithm that we call the Optimistic-Pessimistic Safe Reinforcement Learning (OPSRL) algorithm, and show that it achieves an $\tilde{O}(S^2\sqrt{AH^7K/(\bar{C} - \bar{C}_b)})$ cumulative regret without violating the safety constraints during learning, where $S$ is the number of states, $A$ is the number of actions, $H$ is the horizon length, $K$ is the number of learning episodes, and $(\bar{C} - \bar{C}_b)$ is the safety gap, i.e., the difference between the constraint value and the cost of a known safe baseline policy. The scaling as $\tilde{O}(\sqrt{K})$ is the same as the traditional approach where constraints may be violated during learning, which means that our algorithm suffers no additional regret in spite of providing a safety guarantee. Our key idea is to use an optimistic exploration approach with pessimistic constraint enforcement for learning the policy. This approach simultaneously incentivizes the exploration of unknown states while imposing a penalty for visiting states that are likely to cause violation of safety constraints. We validate our algorithm by evaluating its performance on benchmark problems against conventional approaches.

1 Introduction

Constrained Markov Decision Processes (CMDPs) arise in the context of many cyber-physical systems, which are physical systems that are controlled algorithmically. Here, there are often restrictions on either the frequency with which actions may be used, or on the state visitation frequency, both of which pertain to resource or safety constraints of the system. For example, the average radiated power in wireless communication systems must be restricted due to user health and battery lifetime considerations, the average frequency in a power system must be held within limits to ensure optimal appliance performance, and the frequency of braking or accelerating in an autonomous vehicle must be kept bounded to ensure passenger comfort. Oftentimes, there are existing policies, developed empirically over time that can ensure constraint satisfaction, but might not be optimal in optimizing system performance. Could such safe baseline policies be used in a constrained reinforcement learning (CRL) setting as the basis to learn optimal policies while ensuring exploration remains safe? Would the regret incurred due to such safe learning be higher than the traditional approach of allowing constraint violations while learning?

In this paper, our goal is to develop a framework for safe exploration for solving CMDP problems where the model is unknown. While there has been much work on RL for both the MDP and
the CMDP setting earlier, ensuring safe exploration in the CRL setting has received less attention. However, safe exploration with a known baseline policy setting is perhaps the most relevant for real-world cyber-physical systems. This is because such systems are central to the functioning of society with known safe baseline policies that already exist, and so allowing unsafe exploration with constraint violation—even during learning—is typically unacceptable. Crucial to our study is the development of algorithms that expand the envelope of exploration over such safe baselines while ensuring that constraints are never violated, and characterizing their regret over time in maximizing the objective.

**Main Contributions:** Our goal is to develop an algorithm that will ensure no violation of the constraints both during learning and deployment, while having a low regret in terms of the optimal objective cost. We follow a model-based approach in an episodic setting under which the model (the transition kernel of the CMDP) is empirically determined as samples from the system are gathered. Our solution procedure and contributions are as follows:

(i) We develop a model-based algorithm entitled Optimistic-Pessimistic Safe Reinforcement Learning (OPSRL). The algorithm proceeds in an online fashion with model updates made as samples are gathered from the environment. As with the traditional approach of optimism in the face of uncertainty (OFU), the algorithm utilizes a confidence-ball around the empirically estimated model to expand it to a class of possible models.

(ii) Since the OFU-style approach may lead to selecting policies that are unsafe (violate the safety constraint), we add a penalty term to the original constraint cost to obtain a pessimistic constraint cost function, whose value depends on the uncertainty about the model. We call this approach pessimism in the face of uncertainty (PFU). Further, we use the same uncertainty term to reduce the objective cost to promote exploration. Thus, we have optimism with respect to the model and objective cost to promote exploration, and pessimism with respect to the constraint cost to ensure safety, leading to the phrase “optimistic-pessimistic” to describe our method.

(iii) Since the problem under the optimistic model might be infeasible initially, the algorithm uses the safe baseline policy for several steps until feasibility is attained, after which it switches to the OPSRL approach.

(iv) We derive regret bounds on the performance of our algorithm. Since we explicitly ensure that the episodic constraints are never violated (in expectation), there is no regret on that front. However, one might expect that the focus on safety might imply a worse regret than an unconstrained problem. Interestingly, it turns out that the regret scales as $O(\sqrt{K})$, where $K$ is the number of episodes, which is consistent with the unconstrained case, i.e., there is no increase in regret (in an order sense) due to safe exploration.

As discussed above, safety is crucial in cyber-physical systems, and yet we desire learning algorithms that can improve upon well-known safe baseline policies. Our result indicates that little is lost under safe exploration in solving unknown CMDPs via RL, indicating the feasibility of utilizing the methodology in real-world cyber-physical systems.

**1.1 Related Work**

**Constrained RL:** Constrained Markov Decision Processes (CMDP) has been an active area of research [Altman, 1999], with a number of applications in domains such as power systems [Singh et al., 2018] [Li et al., 2019], communication networks [Altman, 2002] [Singh and Kumar, 2018], and robotics [Ding et al., 2013] [Chow et al., 2015]. CMDP planning problem (with a known model) is typically solved using a linear programming approach or using a Lagrangian approach [Altman, 1999]. In Borkar, 2005, the author proposed an actor-critic reinforcement learning algorithm for learning the optimal policy of an infinite horizon average cost CMDP when the model is unknown. This work uses multi-timescale stochastic approximation theory to show asymptotic global convergence.
This approach is also utilized in function approximation settings with asymptotic local convergence guarantees [Bhatnagar and Lakshmanan, 2012, Chow et al., 2017, Tessler et al., 2018]. Policy gradient algorithms for CMDPs have also been developed [Achiam et al., 2017, Yang et al., 2019, Zhang et al., 2020], though they lack rigorous convergence guarantees. Policy gradient algorithms with provable guarantees have also been developed recently [Paternain et al., 2019, Ding et al., 2020]. However, none of the above mentioned works address the problem of safe exploration and provide any guarantees on the constraint violations during learning.

**Safe Multi-Armed Bandits:** In [Wu et al., 2016, Kazerouni et al., 2017], the authors study conservative exploration problem in linear bandits, where the cumulative reward of the selected actions is greater than the cumulative reward of a baseline policy. A more challenging problem of safe exploration with stage-wise safety constraint is studied in [Amani et al., 2019, Khezeli and Bitar, 2020, Moradipari et al., 2020, Pacchiano et al., 2021]. Linear bandits with more general constraints have also been studied [Parulekar et al., 2020, Liu et al., 2021]. Though these works address different aspects of safe exploration, they do not consider the more challenging RL setting which involves an underlying dynamical system with unknown model.

**Safe Online Convex Optimization:** Online convex optimization [Hazan, 2016] with stochastic constraints [Yu et al., 2017] and adversarial constraints [Neely and Yu, 2017, Yuan and Lamperski, 2018, Liakopoulos et al., 2019] have been addressed in the literature. These works, however, allow constraint violations during learning, and focus on characterizing the cumulative constraint violation along with the standard regret. A safe Frank-Wolf algorithm for convex optimization with unknown linear stochastic constraint has been studied in [Usmanova et al., 2019]. Safe online convex optimization works also do not consider the RL setting with an unknown model.

**Safe Exploration in RL:** The exploration-exploitation problem in finite horizon CMDPs, with constraints violations during learning, is addressed in [Efroni et al., 2020]. They propose four different algorithms and characterize the optimality regret and constraint regret. Optimality regret is similar to the standard regret and the constraint regret is a measure of the cumulative constraint violations. They show that one of the proposed algorithms achieves $O(\sqrt{SH^3AK})$ optimality regret and constraint regret. In [Singh et al., 2020], the authors address the exploration problem in infinite horizon average cost CMDPs with constraints violations during learning. In [Ding et al., 2021], the authors consider a finite horizon CMDP and propose a linear function approximation based algorithm for efficient exploration. They show that the proposed algorithm achieves $O(dH^3\sqrt{K})$ optimality regret and constraint regret, where $d$ is the dimension of the feature map. Algorithms with probably approximately correct (PAC) guarantees on the sample complexity for a finite horizon CMDP were developed in [HasanzadeZouzou et al., 2021]. They show that the number of samples needed to learn an $\epsilon$-optimal policy (i.e., the resulting objective value is at most $\epsilon$ larger than the optimal value and the constraint values are at most $\epsilon$ larger than the allowed), with a probability greater than $(1-\delta)$, is $\tilde{O}(\frac{SAH^3}{\epsilon^2} \log \frac{1}{\delta})$. This work also allows constraint violations during learning. [Kalagarla et al., 2021] also propose a similar algorithm with similar sample complexity result.

We emphasize that all the above mentioned works that address the safe exploration problem in RL allow constraint violations during learning. In particular, these algorithms incur $\tilde{O}(\sqrt{K})$ constraint regret in addition to $\tilde{O}(\sqrt{K})$ optimality regret. In sharp contrast to this, our work addresses a much more difficult problem of safe exploration without violating the safety constraints during learning. We propose an algorithm that achieves $\tilde{O}(\sqrt{S^3AH^7K})$ optimality regret without incurring any constraint regret.

We note that the problem of learning the optimal policy of a CMDP without violating the constraints was studied in [Zheng and Ratliff, 2020]. However, unlike the standard RL setting, they assumed that the model is known and only the cost functions are unknown. In this work, we are addressing a much more difficult RL problem with unknown model.

**Other Related Works:** In [Turchetta et al., 2016, Berkenkamp et al., 2017, Wachi et al., 2018, Wachi and Seul, 2020], the authors address the safe RL problem by utilizing Gaussian processes to model the (deterministic) transition probabilities and value functions. However, they lack rigorous
regret guarantees. Safe learning using control barrier function has also been studied recently [Cheng et al., 2019] [Taylor et al., 2020]. These works, in terms of the algorithms and analysis, differ significantly from our approach.

Notations: For any integer $M$, $[M]$ denotes the set $\{1, \ldots, M\}$. For any two real numbers $a, b$, $a \lor b := \max\{a, b\}$.

2 Preliminaries and Problem Formulation

2.1 Constrained Markov Decision Process

We address the safe exploration problem using the framework of episodic Constrained Markov Decision Process (CMDP) [Altman, 1999]. We consider a CMDP, denoted as $M = (\mathcal{S}, \mathcal{A}, r, c, H, P, \bar{C})$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the objective cost function, $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the constraint cost function, and $H$ is the episode length. The transition probability function $P(s'|s, a)$ represents the probability of transitioning to state $s'$ when action $a$ is taken at state $s$. In the RL context, the transition matrix $P$ is also called the model of CMDP. Finally, $\bar{C}$ is a scalar which specifies the safety constraint in terms of the maximum permissible value for the expected cumulative constraint cost. We consider a finite MDP setting where the cardinality of the state space, denoted by $S = |\mathcal{S}|$, and the cardinality of the action space, denoted by $A = |\mathcal{A}|$, are finite. Also, without loss of generality, we assume that the costs $r$ and $c$ are bounded in $[0, 1]$.

A non-stationary randomized policy $\pi = (\pi^h)_{h=1}^H$ specifies the control action to be taken at each time step $h \in [H]$ of the episode, where $\pi^h : \mathcal{S} \to \Delta(\mathcal{A})$ and $\Delta(\mathcal{A})$ is the probability simplex over the action space $\mathcal{A}$. In particular, $\pi^h(s, a)$ denotes the probability of taking action $a$ when the state is $s$ at time step $h$. For an arbitrary cost function $l : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, the value function of a policy $\pi$ corresponding to time step $h \in [H]$ given a state $s \in \mathcal{S}$ is defined as

$$V^h_\pi(s; P) = \mathbb{E}_\pi \left[ \sum_{\tau = h}^H l(s^\tau, a^\tau) | s^h = s \right],$$

where $a^\tau \sim \pi(s^\tau, \cdot)$, $s^{\tau + 1} \sim P(\cdot | s^\tau, a^\tau)$.

Since we are mainly interested in the value of a policy starting from $h = 1$, we simply denote $V^1_\pi(s; P)$ as $V^1_\pi(s; P)$. For the rest of the paper, we will assume that the initial state is $s^1$ is fixed. So, we will simply denote $V^1_\pi(s^1; P)$ as $V^1_\pi(P)$. Note that this is a standard assumption [Efroni et al., 2020] [Ding et al., 2021] that can be taken without loss of generality.

The CMDP (planning) problem with a known model $P$ can then be stated as follows:

$$\min_{\pi} \; V^\pi_\pi(P),$$

subject to

$$V^\pi_\pi(P) \leq \bar{C}. \tag{2}$$

We say that a policy $\pi$ is a safe policy if $V^\pi_\pi(P) \leq \bar{C}$, i.e., if the expected cumulative constraint cost corresponding to the policy $\pi$ is less than the maximum permissible value $\bar{C}$. The set of safe policies, denoted as $\Pi_{\text{safe}}$, is defined as $\Pi_{\text{safe}} = \{ \pi : V^\pi_\pi(P) \leq \bar{C} \}$. Without loss of generality, we assume that the CMDP problem (2) is feasible, i.e., $\Pi_{\text{safe}}$ is non-empty. Let $\pi^*$ be the optimal safe policy, which is the solution of (2).

The CMDP (planning) problem is significantly different from the standard Markov Decision Process (MDP) (planning) problem [Altman, 1999]. Firstly, there may not exist an optimal deterministic policy for a CMDP, whereas the existence of a deterministic optimal policy is well known for a standard MDP. Secondly, there does not exist a Bellman optimality principle or Bellman equation
for CMDP. So, the standard dynamic programming solution approaches which rely on the Bellman equation cannot be directly applied to solve the CMDP problem.

There are two standard approaches for solving the CMDP problem, namely the Lagrangian approach and the linear programming (LP) approach. Both approaches exploit the zero duality gap property of the CMDP problem \cite{altman1999} to find the optimal policy. In this work, we will use the LP approach. The details for solving \ref{eq:cmdp} using the LP approach is given in Appendix A.

2.2 Reinforcement Learning with Safe Exploration

The goal of the reinforcement learning with safe exploration problem is to solve \ref{eq:cmdp}, but without the knowledge of the model $P$ a priori. Hence, the learning algorithm has to perform exploration by employing different policies to learn $P$. However, we also want the exploration for learning to have a safety guarantee, i.e., the policies employed during learning should belong to the set of safe policies $\Pi_{\text{safe}}$. Since $\Pi_{\text{safe}}$ itself is defined based on the unknown $P$, the learning algorithm will not know $\Pi_{\text{safe}}$ a priori. This makes the safe exploration problem extremely challenging.

We consider a model-based RL algorithm that repeatedly interacts with the environment in an episodic manner. We assume that the objective cost function $r(\cdot, \cdot)$ and the constraint cost function $c(\cdot, \cdot)$ are known to the algorithm. This will help us to focus on the more challenging problem of learning the model $P$. We note that the assumption of known cost function is standard in the works that address the exploration problem in RL \cite{dann2019}. Unknown cost functions can be addressed at the expense of additional notations. We also assume that the knowledge of the model is relaxed. This can be done by getting a conservative estimate of the value $V^\pi_s(P)$ by employing the policy for a sufficient number of episodes $K_o$.

Let $\pi_k = (\pi^h_k)_{h=1}^H$ be the policy employed by the algorithm in episode $k$. At each time step $h \in [H]$ in an episode $k$, the algorithm observes state $s^h_k$, selects action $a^h_k \sim \pi^h_k$, and incurs the costs $r(s^h_k, a^h_k)$ and $c(s^h_k, a^h_k)$. The next state $s^{h+1}_k$ is realized according to the probability vector $P(\cdot | s^h_k, a^h_k)$. As stated before, for simplicity, we assume that the initial state is fixed for each episode $k \in [K] := \{1, \ldots, K\}$, i.e., $s^1_k = s^i$.

The performance of the RL algorithm is measured using the metric of safe regret. The safe regret is defined exactly as the standard regret of an RL algorithm for exploration in MDPs \cite{jaksch2010,dann2017,azar2017}, but with an additional constraint that the exploration policies should belong to the safe set $\Pi_{\text{safe}}$. Formally, the safe regret $R(K)$ after $K$ learning episodes is defined as

$$R(K) = \sum_{k=1}^K (V^\pi_{s_k}(P) - V^\pi_{s_k}^*(P)),$$

subject to $\pi_k \in \Pi_{\text{safe}}$, $\forall k \in [K]$.

Since $\Pi_{\text{safe}}$ is unknown, clearly it is not possible to employ a safe policy without making any additional assumptions. We overcome this obvious limitation by assuming that the algorithm has access to a safe baseline policy $\pi_b$ such that $\pi_b \in \Pi_{\text{safe}}$. We formalize this assumption as follows.

**Assumption 1** (Safe baseline policy). The algorithm knows a safe baseline policy $\pi_b$ such that $V^\pi_{s_k}(P) = \bar{C}_b$, where $\bar{C}_b < \bar{C}$. The value $\bar{C}_b$ is also known to the algorithm.

**Remark 1.** Knowing a safe policy $\pi_b$ is necessary for solving the safe RL problem because we want the constraint to be satisfied from the very first learning episode. A similar assumption has been used in the case of safe exploration in linear bandits \cite{amani2019,khezeli2020,paichano2021}. However, the assumption of knowing the value $\bar{C}_b$ of the safe baseline policy can be relaxed. This can be done by getting a conservative estimate of the value $V^\pi_{s_k}(P)$ by employing the policy for a sufficient number of episodes $K_o$. 

5
3 Algorithm and Provable Performance Guarantee

In this section, we describe our algorithm for reinforcement learning with episodic safety guarantees. Our algorithm builds on the optimism in the face of uncertainty (OFU) style exploration algorithms for RL [Jaksch et al., 2010; Dann et al., 2017]. However, a naive OFU style algorithm may lead to selecting exploration policies that are not in the safe set \( \Pi_{\text{safe}} \). So we modify the selection of exploratory policy by incorporating a pessimism in the face of uncertainty (PFU) criteria. We call our algorithm as the Optimistic-Pessimistic Safe Reinforcement Learning (OPSRL) algorithm.

OPSRL algorithm operates in episodes, each of length \( H \). Define the filtration \( \mathcal{F}_k \) as the sigma algebra generated by the observations until the end of episode \( k \in [K] \), i.e., \( \mathcal{F}_k = (s^h_k, a^h_k, h \in [H], k' \in [k]) \). Let \( n_k(s, a) = \sum_{k'=1}^{k-1} \sum_{h=1}^{H} \mathbb{I}\{s^h_k = s, a^h_k = a\} \) be the number of times the pair \((s, a)\) was observed until the beginning of episode \( k \). Similarly, define \( n_k(s, a, s') = \sum_{k'=1}^{k-1} \sum_{h=1}^{H} \mathbb{1}\{s^{h'}_{k'} = s, a^{h'}_{k'} = a, s'\} \). At the beginning of each episode \( k \), OPSRL algorithm estimates the model as \( \hat{P}_k(s'|s, a) = \frac{n_k(s, a, s')}{n_k(s, a)\vee 1} \), where \( a \vee b := \max\{a, b\} \). Similar to the OFU style algorithms, we construct a confidence set \( \mathcal{P}_k \) around \( \hat{P}_k \) as

\[
\mathcal{P}_k = \bigcap_{(s, a) \in S \times A} \mathcal{P}_k(s, a), \\
\mathcal{P}_k(s, a) = \{P' : |P'(s'|s, a) - \hat{P}_k(s'|s, a)| \leq \beta_k(s, a, s'), \forall s' \in S\}, \\
\beta_k(s, a, s') = \sqrt{\frac{4 \text{Var}(\hat{P}_k(s'|s, a))L}{n_k(s, a)\vee 1}} + \frac{14L}{3(n_k(s, a)\vee 1)},
\]

where \( L = \log\left(\frac{2SAHK}{\delta}\right) \), and \( \text{Var}(\hat{P}_k(s'|s, a)) = \hat{P}_k(s'|s, a)(1 - \hat{P}_k(s'|s, a)) \). Using the empirical Bernstein inequality, we can show that the true model \( P \) is an element of \( \mathcal{P}_k \) for any \( k \) with probability at least \( 1 - \delta \) (see Appendix A).

It is tempting to use the standard OFU approach for selecting the exploration policies since this approach is known to provide sharp regret guarantees for exploration problems in RL. The standard OFU approach will find the optimistic model \( \mathcal{P}_k \) and optimistic policy \( \bar{\pi}_k \), where

\[
[\text{OFU Problem}] \quad \left(\bar{\pi}_k, \mathcal{P}_k\right) = \arg \min_{\pi', P'} V_{\pi'}^{\pi'}(P'), \\
\text{subject to } V_{\pi'}^P(P') \leq \bar{C}.
\]

The OFU problem (7) is feasible since the true model \( P \) is an element of \( \mathcal{P}_k \) (with high probability). In particular, \((\bar{\pi}_k, P) \) and \((\pi^*, P) \) are feasible solutions of (7). Moreover, (7) can be solved efficiently using an extended linear programming approach, as described in Appendix B. The policy \( \bar{\pi}_k \) ensures exploration while satisfying the constraint \( V_{\bar{\pi}}(\mathcal{P}_k) \leq \bar{C} \). However, this naive OFU approach overlooks the important issue that \( \bar{\pi}_k \) may not be a safe policy with respect to the true model \( P \). More precisely, it is possible to have \( V_{\bar{\pi}}(P) > \bar{C} \) even though \( V_{\bar{\pi}}(\mathcal{P}_k) \leq \bar{C} \). So, the standard OFU approach alone will not give a safe exploration strategy.

In order to ensure that the exploration policy employed at any episode is safe, we propose to solve a slightly different problem with a pessimistic constraint cost function. In particular, we add a pessimistic penalty to the original constraint cost to get the pessimistic constraint cost function \( c_k \) as

\[
c_k(s, a) = c(s, a) + H\tilde{\beta}_k(s, a),
\]

where \( \tilde{\beta}_k(s, a) = \sum_{s' \in S} \tilde{\beta}_k(s, a, s') \). Since \( \tilde{\beta}_k(s, a) \) is \( O(1/\sqrt{n_k(s, a)}) \), the \((s, a)\) pairs that are observed less will have a higher penalty than the ones observed more.
The pessimistic penalty may however prevent the exploration that is necessary to learn the optimal policy. In order to overcome this issue, we also modify the original objective cost function by subtracting a term that will incentivize exploration, to get the optimistic objective cost function \( r_k(s, a) = r(s, a) - \frac{H^2}{C - C_b} \bar{\beta}_k(s, a) \).

By the definition of \( \bar{\beta}_k \), the \((s, a)\) pairs that are observed less will have a less optimistic objective cost which will incentivize the algorithm to explore those states.

We then select the policy \( \pi_k \) for episode \( k \) by solving the following Optimistic-Pessimistic (OP) problem:

\[
[\text{OP Problem}] \quad (\pi_k, P_k) = \arg \min_{\pi', P' \in P_k} V_{r_k}'(P') \\
\text{subject to } V_{c_k}'(P') \leq \bar{C}.
\]

We show that our approach carefully balances the optimism and pessimism, yielding a regret minimizing learning algorithm with episodic safe exploration guarantees. The name OPSRL indicates this clever combination of optimism and pessimism.

We note that the OP problem (10) may not be feasible, especially in the first few episodes of learning. This is because, \( \bar{\beta}_k(s, a) \) may be large during the initial phase of learning so that there may not be a policy \( \pi' \) and a model \( P' \in P_k \) that can satisfy the constraint \( V_{c_k}'(P') \leq \bar{C} \). We overcome this issue by employing a safe baseline policy \( \pi_b \) (as defined in Assumption 1) in the first \( K_o \) episodes. Since \( \pi_b \) is safe by definition, OPSRL algorithm ensures safety during the first \( K_o \) episodes. We will later show that the OP problem (10) will have a feasible solution after the first \( K_o \) episodes (see Proposition 1).

For any episode \( k \geq K_o \), OPSRL algorithm will employ the policy \( \pi_k \), which is the solution of (10). We will also show that \( \pi_k \) from (10) (once it becomes feasible) will indeed be a safe policy (see Proposition 2). We present the OPSRL algorithm formally in Algorithm 1.

---

**Algorithm 1 Optimistic-Pessimistic Safe Reinforcement Learning (OPSRL)**

1. **Input:** \( \delta \in (0, 1), r, c, \pi_b, C_b, \bar{C}, K_o \)
2. **Initialization:** \( n_k(s, a) = n_k(s, a, s') = 0 \) \( \forall s, s' \in S, a \in A \).
3. **for** episodes \( k = 1, \ldots, K_o \) **do**
   4. **if** \( k \leq K_o \) **then**
   5. **else**
   6. **end if**
   7. **for** \( h = 1, 2, \ldots, H \) **do**
   8. Observe state \( s_h^k \), select action \( a_h^k \sim \pi_h^k(s_h^k, \cdot) \),
   9. incur the cost \( r(s_h^k, a_h^k) \) and \( c(s_h^k, a_h^k) \), and observe next state \( s_{h+1}^k \sim P(\cdot|s_h^k, a_h^k) \)
   10. Update the counts: \( n_k(s_h^k, a_h^k) \leftarrow n_k(s_h^k, a_h^k) + 1 \), \( n_k(s_h^k, a_h^k, s_{h+1}^k) \leftarrow n_k(s_h^k, a_h^k, s_{h+1}^k) + 1 \)
   11. **end for**
12. **end for**

---

We now present our main result.

**Theorem 1.** Fix any \( \delta \in (0, 1) \). Consider the OPSRL algorithm with \( K_o \) as specified in Proposition 1. Let \( \{\pi_k, k \in [K]\} \) be the sequence of policies generated by the OPSRL algorithm. Then, with
probability at least $1 - 3\delta$, $\pi_k \in \Pi_{\text{safe}}$ for all $k \in [K]$. Moreover, with probability at least $1 - 3\delta$, the regret of the OPSRL algorithm satisfies

$$R(K) \leq \tilde{O}\left(\frac{S^2AH^4}{(C - C_b)^2} + \frac{S^2}{(C - C_b)} \sqrt{AH^2K}\right).$$

Remark 2. The above theorem shows that the OPSRL algorithm achieves $\tilde{O}(\sqrt{K})$ regret without violating the safety constraints during learning, with high probability.

4 Analysis

We now provide the technical analysis of our OPSRL algorithm. We start with some preliminary notions that are necessary for the analysis. We then show that OP problem (10) is feasible after some finite number of episodes and the policies generated by the OPSRL algorithm are indeed safe. Finally, we give the proof outline of Theorem 1 along with intuitive explanation.

4.1 Preliminaries

For an arbitrary policy $\pi'$ and transition probability function $P'$, define $\epsilon_k'(P')$ as

$$\epsilon_k'(P') = H\mathbb{E}\left[\sum_{h=1}^{H} \beta_k(s', a^h) | \pi', P', \mathcal{F}_{k-1} \right].$$

Then, it is straightforward to show that (see [23] - [24] in the Appendix)

$$V_c^\pi_k(P') = V_c^\pi'(P') + \epsilon_k'(P'),$$
$$V_r^\pi_k(P') = V_r^\pi'(P') - \frac{H}{C - C_b} \epsilon_k'(P').$$

(12)

Our analysis will make use of the above decompositions of $V_c^\pi_k(P')$ and $V_r^\pi_k(P')$ and the properties of $\epsilon_k'(P')$.

4.2 Feasibility of the OP Problem

Even though $(\pi_b, P)$ is a feasible solution for the original CMDP problem (2), it may not be a feasible solution for the OP problem (10) in the initial phase of learning. To see this, note that since $V^\pi_b(P) = V_c^\pi_b(P) + \epsilon_k^\pi_b(P)$ and $V_c^\pi_b(P) = C_b$, we will have $V^\pi_b(P) \leq C$ if and only if $\epsilon_k^\pi_b(P) \leq (C - C_b)$. So, $(\pi_b, P)$ is a feasible solution for (10) if and only if $\epsilon_k^\pi_b(P) \leq (C - C_b)$. This condition may not be satisfied for initial episodes. However, since $\epsilon_k^\pi_b(P)$ is decreasing in $k$, if $(\pi_b, P)$ becomes a feasible solution for (10) at episode $k'$, then it remains to be feasible solution for all episodes $k \geq k'$. Also, since $\beta_k$ decreases with $k$, one can expect that (10) becomes feasible after some number of episodes. We use these intuitions, along with some technical lemmas to show the following result.

Proposition 1. Under the OPSRL algorithm, with a probability greater that $1 - 3\delta$, $(\pi_b, P)$ is a feasible solution for the OP problem (10) for all $k \geq K_o$, where $K_o = \tilde{O}\left(\frac{S^2AH^3}{(C - C_b)^2}\right)$. The above result, however, only shows that $\pi_b$ becomes a feasible policy after some finite number of episodes. A natural question is, will $\pi_b$ be the only feasible policy? In such a case, the OPSRL algorithm may not provide enough exploration to learn the optimal policy. We alleviate these concerns by showing that there exist feasible policies that will visit all state-action pairs that are visited by the optimal policy. We defer this result to Appendix (D.1).
4.3 Safety Exploration Guarantee

We now show that the OPSRL Algorithm provides safe exploration guarantee. In particular, \( \pi_k \in \Pi_{sa} \) for all \( k \in [K] \) with high probability, where \( \pi_k \) is the exploration policy employed by OPSRL algorithm in episode \( k \). This is achieved by the cleverly designed pessimistic constraint of the OP problem \([10]\).

For any episode \( k \leq K_o \), we have \( \pi_k = \pi_b \), and it is safe by Assumption \([1]\). For \( k \geq K_o \), \([10]\) is feasible according to Proposition \([1]\). Since \((\pi_c, P_k)\) is the solution of \([10]\), we have \( V_c^{\pi_k}(P_k) = V_{\pi_k}(P_k) = c_{P_k}^\pi(P_k) \leq C \). This also implies that \( V_c^{\pi_k}(P_k) \leq c_{\pi_k}(P_k) \), i.e., \( \pi_k \) satisfies a tighter constraint with respect to the model \( P_k \). However, it is not obvious that the policy \( \pi_k \) will be safe with respect to the true model \( P \) because \( V_c^{\pi_k}(P) \) may be larger than \( V_c^{\pi_k}(P_k) \) due to the change from \( P_k \) to \( P \).

We, however, show that \( V_c^{\pi_k}(P) \) cannot be larger than \( V_c^{\pi_k}(P_k) \) by more than \( c_{\pi_k}(P_k) \), i.e., \( V_c^{\pi_k}(P) - V_c^{\pi_k}(P_k) \leq c_{\pi_k}(P_k) \). This will then immediately yield that \( V_c^{\pi_k}(P) \leq V_c^{\pi_k}(P_k) + c_{\pi_k}(P_k) \leq C \), which is the true safety constraint. The key idea is in the design of the pessimistic cost function \( c_{\pi_k}(\cdot, \cdot) \) such that its pessimistic effect will balance the change in the value function (from \( V_c^{\pi_k}(P_k) \) to \( V_c^{\pi_k}(P) \)) due to the optimistic selection of the model \( P_k \).

We formally state the safety guarantee of OPSRL algorithm below.

**Proposition 2.** Let \( \{\pi_k, k \in [K]\} \) be the sequence of policies generated by the OPSRL algorithm. Then \( \pi_k \) is safe for all \( k \in [K] \), i.e., \( V_c^{\pi_k}(P) \leq C \), for all \( k \in [K] \), with a probability greater than \( 1 - 3\delta \).

4.4 Regret Analysis

To see the key challenge arising in the regret analysis of OPSRL, consider the regret \( R_k \) in episode \( k \), decomposed into two terms as

\[
R_k = V_c^{\pi_k}(P) - V_c^{\pi_\ast}(P) = V_c^{\pi_k}(P) - V_c^{\pi_k}(P_k) + V_c^{\pi_k}(P_k) - V_c^{\pi_\ast}(P) + \underbrace{V_c^{\pi_k}(P_k) - V_c^{\pi_k}(P)}_{\text{Term I}} + \underbrace{V_c^{\pi_k}(P) - V_c^{\pi_k}(P_k)}_{\text{Term II}}.
\]

Term I is the difference between value functions of the selected policy \( \pi_k \) with respect to the true model \( P \) and optimistic model \( P_k \). This term is standard and closely follows the analysis for OFU style algorithms for the unconstrained MDP [Jaksch et al., 2010, Dann and Brunskill, 2015]. We formally state the regret due to Term I below.

**Lemma 1.** Let \( K_o \) be as defined in Proposition \([7]\). Under the OPSRL algorithm, with a probability greater than \((1 - 3\delta)\), we have

\[
\sum_{k=K_o}^{K} V_c^{\pi_k}(P) - V_c^{\pi_k}(P_k) \leq \tilde{O}(S\sqrt{AH^3K}).
\]

However, Term II is more complicated. In the standard OFU style analysis for the unconstrained problem, since \( P \in P_k \) for all \( k \), it can be easily observed that \((\pi_\ast, P)\) is a feasible solution for the OFU problem \([7]\) for all \( k \in [K] \). Moreover, since \((\pi_k, P_k)\) is the optimal solution in \( k \)th episode, we get \( V_c^{\pi_k}(P_k) \leq V_c^{\pi_\ast}(P) \). So, Term II will be non-positive, and hence it can be dropped from the regret analysis. However, in our setting, the second term can be positive since \((\pi_\ast, P)\) may not be a feasible solution of the OP problem \([10]\) due to the pessimistic constraint. This necessitates a different approach for bounding term II. We formally state the regret due to Term II below.
Lemma 2. Let $K_o$ as defined in Proposition 1. Under the OPSRL algorithm, with a probability greater than $(1 - 3\delta)$, we have

$$\sum_{k=K_o}^{K} V_{\pi_k}(P_k) - V_{\pi^*}(P) \leq \tilde{O}\left(\frac{S^2}{(C - C_b)}\sqrt{AH^2K}\right).$$

We can now get the regret bound given in Theorem 1 by using the results of Lemma 1, Lemma 2 and Proposition 1. The formal proof is given in Appendix (D.4).

5 Experiments

We now validate the OPSRL algorithm via simulations. We have two relevant metrics, namely (i) optimality regret, defined in (3) that measures the optimality gap of the algorithm, and (ii) constraint regret, defined as $\sum_{k=1}^{K} \max\{0, V_{\pi_k}(P) - C\}$, where $\pi_k$ is the output of the algorithm in question at episode $k$. This measures the safety gap of the algorithm. Note that OPSRL will have a zero constraint violation regret due to its safety guarantee, whereas other algorithms are likely to do worse.

Media Control: Our first environment represents media streaming to a device from a wireless base station, which provides high and low service rates at different costs. These service rates have independent Bernoulli distributions, with parameters $\mu_1 = 0.9$, and $\mu_2 = 0.1$, where $\mu_1$ corresponds to the fast service. Packets received at the device are stored in a media buffer and played out according to a process that is Bernoulli with parameter $\gamma$. We denote the number of incoming packets into the buffer is denoted $A_h$, and the number of packets leaving the buffer $B_h$. The media buffer length is the state and evolves as

$$s_{h+1} = \min\{\max(0, s_h + A_h - B_h), N\},$$

where $N = 20$ is the maximum buffer length. The action space is $\{1, 2\}$, where action 1 corresponds to using the fast service. The objective cost is $r(s, a) = \mathbb{I}\{s = 0\}$, while the constraint cost is $c(s, a) = \mathbb{I}\{a = 1\}$, i.e., we want to minimize the outage cost, while limiting the usage of fast service. We consider a horizon $H = 10$.

Experiment setup and Results: We compare our OPSRL algorithm with two other approaches, namely (i) OptCMDP Algorithm [Efroni et al., 2020], which solves the CMDP optimally but will violate constraints during learning, and (ii) UCRL2 [Jaksch et al., 2010], which ignores constraints, but provides a baseline on the $\tilde{O}(\sqrt{K})$ scaling in the unconstrained case. The baseline policies for
OPSRL are selected as the optimal solutions of problems with stricter constraints (i.e., they will be quite conservative). We use $\tilde{C}_b = 0.2\tilde{C}$. We run each experiment 5 times.

Fig.1a compares the optimality regret for different algorithms. The optimality regret of OPSRL grows linearly for the initial $K_0$ episodes, and then changes to a square-root scaling with the number of episodes. Both UCRL2 and OptCMDP have the same square-root regret scaling, but the constant is lower than OPSRL due to the initial safe learning phase of OPSRL. Fig.1b compares the regret in constraint violation for OPSRL and OptCMDP algorithms for the Media control environment. We see that OPSRL does not violate the constraint, while the OptCMDP algorithm violates the constraints during learning. UCRL2 does not account for constraints, and has a linear regret scaling that we do not illustrate.

Finally, Fig.1c shows the optimality regret of OPSRL algorithm with different baseline policies. We see that less conservative (but safe) baselines allow for lower initial regret indicating the value of selecting good baseline policies.

Additional experiments: Due to page limitations, we have included additional experiments in Appendix E.

6 Conclusion

In this work, we considered the safe exploration problem in reinforcement learning, where the exploration policy used in each episode of the learning must satisfy some safety requirements. We addressed this problem using the framework of an episodic CMDP, where the safety requirements are modeled as constraints on the expected cumulative costs that must be satisfied during all episodes of learning. We proposed a model-based online reinforcement learning algorithm, called OPSRL algorithm, to solve this problem. The OPSRL algorithm follows an ‘optimistic-pessimistic’ approach that balances the optimistic exploration and pessimistic constraint satisfaction. We showed that the OPSRL algorithm achieves $\tilde{O}(\sqrt{K})$ regret without violating the safety constraints during learning. This matches the regret guarantees of online RL algorithms in unconstrained setting, which shows that there is no increase in regret (in an order sense) due to safe exploration.

In the future, we plan to extend our safe learning approach to large scale reinforcement learning problems using function approximations.
References

J. Achiam, D. Held, A. Tamar, and P. Abbeel. Constrained policy optimization. In International Conference on Machine Learning, pages 22–31. PMLR, 2017.

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

E. Altman. Applications of markov decision processes in communication networks. In Handbook of Markov decision processes, pages 489–536. Springer, 2002.

S. Amani, M. Alizadeh, and C. Thrampoulidis. Linear stochastic bandits under safety constraints. Advances in Neural Information Processing Systems, 32:9256–9266, 2019.

M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In International Conference on Machine Learning, pages 263–272. PMLR, 2017.

F. Berkenkamp, M. Turchetta, A. P. Schoellig, and A. Krause. Safe model-based reinforcement learning with stability guarantees. In Proceedings of the 31st International Conference on Neural Information Processing Systems, pages 908–919, 2017.

D. P. Bertsekas et al. Dynamic programming and optimal control: Vol. 1. Athena scientific Belmont, 2000.

S. Bhatnagar and K. Lakshmanan. An online actor–critic algorithm with function approximation for constrained markov decision processes. Journal of Optimization Theory and Applications, 153(3):688–708, 2012.

V. S. Borkar. An actor-critic algorithm for constrained markov decision processes. Systems & control letters, 54(3):207–213, 2005.

R. Cheng, G. Orosz, R. M. Murray, and J. W. Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 3387–3395, 2019.

Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone. Risk-constrained reinforcement learning with percentile risk criteria. The Journal of Machine Learning Research, 18(1):6070–6120, 2017.

Y.-L. Chow, M. Pavone, B. M. Sadler, and S. Carpin. Trading safety versus performance: Rapid deployment of robotic swarms with robust performance constraints. Journal of Dynamic Systems, Measurement, and Control, 137(3):031005, 2015.

A. Cohen, T. Koren, and Y. Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In International Conference on Machine Learning, pages 1300–1309. PMLR, 2019.

C. Dann and E. Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. Advances in Neural Information Processing Systems, 28:2818–2826, 2015.

C. Dann, T. Lattimore, and E. Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. arXiv preprint arXiv:1703.07710, 2017.

D. Ding, K. Zhang, T. Basar, and M. R. Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. In NeurIPS, 2020.

D. Ding, X. Wei, Z. Yang, Z. Wang, and M. Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In International Conference on Artificial Intelligence and Statistics, pages 3304–3312. PMLR, 2021.
X. C. Ding, A. Pinto, and A. Surana. Strategic planning under uncertainties via constrained markov decision processes. In 2013 IEEE International Conference on Robotics and Automation, pages 4568–4575. IEEE, 2013.

Y. Efroni, N. Merlis, M. Ghavamzadeh, and S. Mannor. Tight regret bounds for model-based reinforcement learning with greedy policies. arXiv preprint arXiv:1905.11527, 2019.

Y. Efroni, S. Mannor, and M. Pirotta. Exploration-exploitation in constrained mdps. arXiv preprint arXiv:2003.02189, 2020.

A. HasanzadeZonuzy, A. Bura, D. Kalathil, and S. Shakkottai. Learning with safety constraints: Sample complexity of reinforcement learning for constrained mdps. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 7667–7674, 2021.

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

T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. Journal of Machine Learning Research, 11(4), 2010.

C. Jin, T. Jin, H. Luo, S. Sra, and T. Yu. Learning adversarial Markov decision processes with bandit feedback and unknown transition. In H. D. III and A. Singh, editors, Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 4860–4869. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/jin20c.html.

K. C. Kalagarla, R. Jain, and P. Nuzzo. A sample-efficient algorithm for episodic finite-horizon mdp with constraints. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 8030–8037, 2021.

A. Kazerouni, M. Ghavamzadeh, Y. Abbasi-Yadkori, and B. Van Roy. Conservative contextual linear bandits. In Proceedings of the 31st International Conference on Neural Information Processing Systems, pages 3913–3922, 2017.

K. Khezeli and E. Bitar. Safe linear stochastic bandits. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 10202–10209, 2020.

H. Li, Z. Wan, and H. He. Constrained ev charging scheduling based on safe deep reinforcement learning. IEEE Transactions on Smart Grid, 11(3):2427–2439, 2019.

N. Liakopoulos, A. Destounis, G. Paschos, T. Spyropoulos, and P. Mertikopoulos. Cautious regret minimization: Online optimization with long-term budget constraints. In International Conference on Machine Learning, pages 3944–3952. PMLR, 2019.

X. Liu, B. Li, P. Shi, and L. Ying. An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. arXiv preprint arXiv:2102.05295, 2021.

A. Maurer and M. Pontil. Empirical bernstein bounds and sample-variance penalization. In COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009, 2009.

A. Moradipari, C. Thrampoulidis, and M. Alizadeh. Stage-wise conservative linear bandits. Advances in Neural Information Processing Systems, 33, 2020.

M. J. Neely and H. Yu. Online convex optimization with time-varying constraints. arXiv preprint arXiv:1702.04783, 2017.
A. Pacchiano, M. Ghavamzadeh, P. Bartlett, and H. Jiang. Stochastic bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 2827–2835. PMLR, 2021.

A. Parulekar, S. Basu, A. Gopalan, K. Shanmugam, and S. Shakkottai. Stochastic linear bandits with protected subspace. *arXiv preprint arXiv:2011.01016*, 2020.

S. Paternain, L. Chamon, M. Calvo-Fullana, and A. Ribeiro. Constrained reinforcement learning has zero duality gap. *Advances in Neural Information Processing Systems*, 32:7555–7565, 2019.

R. Singh and P. Kumar. Throughput optimal decentralized scheduling of multihop networks with end-to-end deadline constraints: Unreliable links. *IEEE Transactions on Automatic Control*, 64(1): 127–142, 2018.

R. Singh, P. Kumar, and L. Xie. Decentralized control via dynamic stochastic prices: The independent system operator problem. *IEEE Transactions on Automatic Control*, 63(10):3206–3220, 2018.

R. Singh, A. Gupta, and N. B. Shroff. Learning in markov decision processes under constraints. *arXiv preprint arXiv:2002.12435*, 2020.

A. Taylor, A. Singletary, Y. Yue, and A. Ames. Learning for safety-critical control with control barrier functions. In *Learning for Dynamics and Control*, pages 708–717. PMLR, 2020.

C. Tessler, D. J. Mankowitz, and S. Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2018.

M. Turchetta, F. Berkenkamp, and A. Krause. Safe exploration in finite markov decision processes with gaussian processes. *Advances in Neural Information Processing Systems*, 29:4312–4320, 2016.

I. Usmanova, A. Krause, and M. Kamgarpour. Safe convex learning under uncertain constraints. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2106–2114. PMLR, 2019.

A. Wachi and Y. Sui. Safe reinforcement learning in constrained markov decision processes. In *International Conference on Machine Learning*, pages 9797–9806. PMLR, 2020.

A. Wachi, Y. Sui, Y. Yue, and M. Ono. Safe exploration and optimization of constrained mdps using gaussian processes. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Y. Wu, R. Shariff, T. Lattimore, and C. Szepesvári. Conservative bandits. In *International Conference on Machine Learning*, pages 1254–1262. PMLR, 2016.

T.-Y. Yang, J. Rosca, K. Narasimhan, and P. J. Ramadge. Projection-based constrained policy optimization. In *International Conference on Learning Representations*, 2019.

H. Yu, M. J. Neely, and X. Wei. Online convex optimization with stochastic constraints. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1427–1437, 2017.

J. Yuan and A. Lamperski. Online convex optimization for cumulative constraints. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6140–6149, 2018.

Y. Zhang, Q. Vuong, and K. Ross. First order constrained optimization in policy space. *Advances in Neural Information Processing Systems*, 33, 2020.

L. Zheng and L. Ratliff. Constrained upper confidence reinforcement learning. In *Learning for Dynamics and Control*, pages 620–629. PMLR, 2020.

A. Zimin and G. Neu. Online learning in episodic markovian decision processes by relative entropy policy search. In *Neural Information Processing Systems 26*, 2013.
A Linear Programming Method for Solving the CMDP Problem

Here we give a brief description on solving the CMDP problem (2) using the linear programming method when the model $P$ is known. The details can be found in [Efroni et al., 2020, Section 2].

The first step is to reformulate (2) using occupancy measure [Altman, 1999, Zimin and Neu, 2013]. For a given policy $\pi$ and an initial state $s_1$, the state-action occupation measure $w_\pi$ for the MDP with model $P$ is defined as

$$w_\pi^h(s, a; P) = \mathbb{E}[\{s^h = s, a^h = a\} | s^1, P, \pi] = P(s^h = s, a^h = a | s^1, P, \pi).$$  \hspace{1cm} (13)

Given the occupancy measure, the policy that generated it can easily be computed as

$$\pi^h(s, a) = \frac{w_\pi^h(s, a; P)}{\sum_b w_\pi^h(s, b; P)}. \hspace{1cm} (14)$$

The occupancy measure of any policy $\pi$ for an MDP with model $P$ should satisfy the following conditions. We omit the explicit dependence on $\pi$ and $P$ from the notation of $w$ for simplicity.

$$\sum_a w^h(s, a) = \sum_{s', a'} P(s | s', a') w^{h-1}(s', a'), \ \forall h \in [H] \setminus \{1\} \hspace{1cm} (15)$$

$$\sum_a w^1(s, a) = 1 \{s = s^1\}, \ \forall s \in S, w^h(s, a) \geq 0, \ \forall (s, a, h) \in S \times A \times [H] \hspace{1cm} (16)$$

From the above conditions, it is easy to show that $\sum_{s, a} w^h(s, a) = 1$. So, occupancy measures are indeed probability measures. Since the set of occupancy measures for a model $P$, denoted as $W(P)$, is defined by a set of affine constraints, it is straightforward to show that $W(P)$ is convex. We state this fact formally below.

**Proposition 3.** The set of occupancy measures for an MDP with model $P$, denoted as $W(P)$, is convex.

Recall that the value of a policy $\pi$ for an arbitrary cost function $l : S \times A \rightarrow \mathbb{R}$ with a given initial state $s^1$ is defined as $V_\pi^l(P) = \mathbb{E}[\sum_{h=1}^H l(s^h, a^h) | s^1 = s, \pi, P]$. It can then be expressed using the occupancy measure as

$$V_\pi^l(P) = \sum_{h, s, a} w_\pi^h(s, a; P) l(s, a) = l^T w_\pi(P),$$

where $w_\pi(P) \in \mathbb{R}^{SAH}$ with $(s, a, h)$ element is given by $w_\pi^h(s, a; P)$ and $l \in \mathbb{R}^{SAH}$ with $(s, a, h)$ element is given by $l(s, a), \forall h \in [H]$. The CMDP problem (2) can then be written as

$$\min_{\pi} \ r^T w_\pi(P) \text{ s.t. } c^T w_\pi(P) \leq \bar{C}. \hspace{1cm} (17)$$

Using the properties of the occupancy measures, the reformulated CMDP problem (17) can be rewritten as an LP, where the optimization variables are occupancy measures [Zimin and Neu, 2013, Efroni et al., 2020]. More precisely, the CMDP problem (2) and its equivalent (17) can be written...
as

$$\min_{w} \sum_{h,s,a} w^h(s,a)r(s,a)$$ \hspace{1cm} (18a)$$

subject to $$\sum_{h,s,a} w^h(s,a)c(s,a) \leq \bar{C}$$ \hspace{1cm} (18b)$$

$$\sum_{a} w^h(s,a) = \sum_{s',a'} P(s'|s,a')w^{h-1}(s',a'), \forall h \in [H] \setminus \{1\}$$ \hspace{1cm} (18c)$$

$$z^h(s,a) = 1\{s = s^1\}, \forall s \in \mathcal{S}$$ \hspace{1cm} (18d)$$

$$w^h(s,a) \geq 0, \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$$ \hspace{1cm} (18e)$$

From the optimal solution $$w^*$$. of (18), the optimal policy $$\pi^*$$ for the CMDP problem (2) can be computed using (14).

B Extended Linear Programming Method for Solving (7) and (10)

The OFU problem (7) and the pessimistic optimistic problem (10) may appear much more challenging than the CMDP problem (2) because they involve a minimization over all models in $$\mathcal{P}_k$$, which is non-trivial. However, finding the optimistic model (and the corresponding optimistic policy) from a given confidence set is a standard step in OFU style algorithms for exploration in RL [Jaksch et al., 2010, Efroni et al., 2020]. In the case of standard (unconstrained) MDP, this problem is solved using an approach called extended value iteration [Jaksch et al., 2010]. In the case of constrained MDP, (7) (and similarly (10)) can be solved by an approach called extended linear programming. The details are given in [Efroni et al., 2020]. We give a brief description below for completeness. Note that the description below mainly focus on solving (7). Solving (10) is identical, just by replacing the constraint cost function $$c(\cdot, \cdot)$$ with pessimist constraint cost function $$c_k(\cdot, \cdot)$$, and is mentioned at the end of this subsection.

Define the state-action-state occupancy measure $$z_\pi$$ as $$z^h_{\pi}(s,a,s';P) = P(s'|s,a)w^h_{\pi}(s,a;P)$$. The extended LP formulation corresponding to (7) is then given as follows:

$$\max_{z} \sum_{s,a,s',h} z^h(s,a,s')r(s,a)$$ \hspace{1cm} (19a)$$

s.t. $$\sum_{s,a,s',h} z^h(s,a,s')c(s,a) \leq \bar{C}$$ \hspace{1cm} (19b)$$

$$\sum_{a,s'} z^h(s,a,s') = \sum_{s',a'} z^{h-1}(s',a'), \forall h \in [H] \setminus \{1\}, s \in \mathcal{S}$$ \hspace{1cm} (19c)$$

$$\sum_{a,s'} z^1(s,a,s') = 1\{s = s^1\}, \forall s \in \mathcal{S}$$ \hspace{1cm} (19d)$$

$$z^h(s,a,s') \geq 0, \forall (s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$$, \hspace{1cm} (19e)$$

$$\sum_{y} z^h(s,a,y) \leq 0, \forall (s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$$ \hspace{1cm} (19f)$$

$$-z^h(s,a,s') + (\hat{P}_k(s'|s,a) - \beta_k(s,a,s')) \sum_{y} z^h(s,a,y) \leq 0, \forall (s,a,s',h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$$ \hspace{1cm} (19g)$$
The last two conditions (19f) and (19g) distinguish the extended LP formulation from the LP formulation for CMDP. These constraints are based on the Bernstein confidence sets around the empirical model \( \hat{P}_k \).

From the solutions \( \hat{z}^* \) of the extended LP, we can obtain the solution of (7) as
\[
\hat{P}_k(s'|s,a) = \frac{\hat{z}^{h,*}(s,a,s')}{\sum_y \hat{z}^{h,*}(s,a,y)}, \quad \hat{r}_k^h(s,a) = \frac{\sum_{b,s'} \hat{z}^{h,*}(s,b,s')}{\sum_{b,s'} \hat{z}^{h,*}(s,b,s')}.
\] (20)

C Useful Technical Results

Here we reproduce the supporting technical results that are required for analyzing our OPSRL algorithm. We begin by stating the following concentration inequality, known as empirical Bernstein inequality [Maurer and Pontil 2009, Theorem 4].

**Lemma 3** (Empirical Bernstein Inequality). Let \( Z = (Z_1, \ldots, Z_n) \) be i.i.d random vector with values in \([0,1]^n\), and let \( \delta \in (0,1) \). Then, with probability at least \( 1 - \delta \), it holds that
\[
\mathbb{E}[Z] - \frac{1}{n} \sum_{i=1}^{n} Z_i \leq \sqrt{\frac{2\text{Var}(Z) \log(\frac{2}{\delta})}{n}} + \frac{7 \log(\frac{2}{\delta})}{3(n-1)},
\]
where \( \text{Var}(Z) \) is the sample variance.

We can get the following result using empirical Bernstein inequality and union bound. This result is widely used in the literature now, for example see [Jin et al. 2020, Proof of Lemma 2].

**Lemma 4.** With probability at least \( 1 - 2\delta \), for all \((s,a,s') \in S \times A \times S, k \in [K]\), we have
\[
|P(s'|s,a) - \hat{P}_k(s'|s,a)| \leq \sqrt{\frac{4\text{Var}(\hat{P}_k(s'|s,a)) \log \left(\frac{2SAKH}{\delta}\right)}{n_k(s,a) \lor 1}} + \frac{14 \log \left(\frac{2SAKH}{\delta}\right)}{3(n_k(s,a) \lor 1)}.
\]

Recall (from (4) - (6)) that
\[
\beta_k(s,a,s') = \sqrt{\frac{4\text{Var}(\hat{P}_k(s'|s,a)) \log \left(\frac{2SAKH}{\delta}\right)}{n_k(s,a) \lor 1}} + \frac{14 \log \left(\frac{2SAKH}{\delta}\right)}{3(n_k(s,a) \lor 1)},
\]
\[
\mathcal{P}_k(s,a) = \left\{ P' : |P'(s'|s,a) - \hat{P}_k(s'|s,a)| \leq \beta_k(s,a,s'), \forall s' \in S \right\}, \mathcal{P}_k = \bigcap_{(s,a) \in S \times A} \mathcal{P}_k(s,a).
\]

Define the event
\[
F_p = \{ P \in \mathcal{P}_k, \forall k \in [K] \}.
\] (21)

Then, using Lemma 4 we can get the following result immediately.

**Lemma 5.** Let \( F_p \) be the event defined as in (21). Then, \( \mathbb{P}(F_p) \geq 1 - 2\delta \)

We now define the event \( F_w \) as follows
\[
F_w = \left\{ n_k(s,a) \geq \frac{1}{2} \sum_{j<k} w_j(s,a) - H \log \frac{SAH}{\delta}, \forall(s,a,s') \in S \times A \times S, k \in [K] \right\},
\] (22)
where \( w_j(s,a) = \sum_{h=1}^{H} w_{nj}^h(s,a) \). We have the following result from [Dann et al. 2017, Corollary 4.4].
Lemma 6 (Corollary E.4., [Dann et al., 2017]). Let \( F_w \) be the event defined as in (22). Then, \( \mathbb{P}(F_w) \geq 1 - \delta \).

We now define the **good event** \( G = F_p \cap F_w \). Using union bound, we can show that \( \mathbb{P}(G) \geq 1 - 3\delta \). Since our analysis is based on this good event, we formally state it as a lemma.

Lemma 7. Define the good event \( G = F_p \cap F_w \), where \( F_p \) is defined as in (21) and \( F_w \) is defined as in (22). Then, \( \mathbb{P}(G) \geq 1 - 3\delta \).

We will also use the following results for analyzing the performance of our OPSRL algorithm.

Lemma 8 (Lemma 38, [Efroni et al., 2019]). Under the event \( F_w \),
\[
\sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E} \left[ \frac{1}{\sqrt{n_k(s_h^k, a_h^k)} \lor 1} \right] \mathbb{I}_{F_k-1} \leq \tilde{O}(\sqrt{SAHK + SAH}).
\]

Lemma 9 (Lemma 39, [Efroni et al., 2019]). Under the event \( F_w \),
\[
\sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E} \left[ \frac{1}{n_k(s_h^k, a_h^k) \lor 1} \right] \mathbb{I}_{F_k-1} \leq \tilde{O}(SAH).
\]

The following result, called value difference lemma [Dann et al., 2017, Lemma E.15], helps us to analyze the difference between the value functions of two MDPs that differ in their transition probability functions.

Lemma 10 (Value difference lemma). Consider two MDPs \( M = (\mathcal{S}, \mathcal{A}, l, P) \) and \( M' = (\mathcal{S}, \mathcal{A}, l, P') \) that differ only in their transition probability functions. For any policy \( \pi \), state \( s \in \mathcal{S} \), and time step \( h \in [H] \), the following relation holds.
\[
V^{h,\pi}_1(s; P) - V^{h,\pi}_1(s; P') = \mathbb{E} \left[ \sum_{\tau=0}^{H} ((P - P')(\cdot|s, a)\tau) \mathbb{I}_{V^{\tau+1,\pi}_1(\cdot; P')} \mathbb{I}_{s^{\tau+1} = s, \pi, P'} \right]
\]

\[
= \mathbb{E} \left[ \sum_{\tau=0}^{H} ((P' - P)(\cdot|s, a)\tau) \mathbb{I}_{V^{\tau+1,\pi}_1(\cdot; P')} \mathbb{I}_{s^{\tau+1} = s, \pi, P'} \right].
\]

D Proof of the Main Results
All the results we prove in this section are conditioned on the good event \( G \) defined in Section C. So, the results hold true with a probability greater than \( 1 - 3\delta \) according to Lemma 7. We will omit stating this conditioning under \( G \) in each statement to avoid repetition.

D.1 Proofs of Proposition 1 and Corollary 1
First note that
\[
\mathbb{E} \left[ \sum_{h=1}^{H} c_k(s_h^h, a_h^h)|\pi', P', F_{k-1} \right] = \mathbb{E} \left[ \sum_{h=1}^{H} c(s_h^h, a_h^h)|\pi', P' \right] + H \mathbb{E} \left[ \sum_{h=1}^{H} \bar{\beta}_k(s_h^h, a_h^h)|\pi', P', F_{k-1} \right]
\]
\[
= V^{\pi'}_c(P') + \epsilon^c_k(P'),
\]
\[
\mathbb{E} \left[ \sum_{h=1}^{H} r_k(s_h^h, a_h^h)|\pi', P', F_{k-1} \right] = \mathbb{E} \left[ \sum_{h=1}^{H} r(s_h^h, a_h^h)|\pi', P' \right] - \frac{H^2}{C - C_b} \mathbb{E} \left[ \sum_{h=1}^{H} \bar{\beta}_k(s_h^h, a_h^h)|\pi', P', F_{k-1} \right]
\]
\[
= V^{\pi'}_r(P') - \frac{H}{C - C_b} \epsilon^r_k(P').
\]

We now prove the following result.
Lemma 11. Let $\epsilon_k^\pi(P')$ be as defined in [11]. Also, let $\{\pi_k\}$ be the sequence of policies generated by OPSRL algorithm. Then, for any $K' \leq K$, with a probability greater than $1 - 3\delta$,

$$\sum_{k=1}^{K'} \epsilon_k^\pi(P) \leq \tilde{O}(S\sqrt{AH^3K'})$$

Proof.

$$\sum_{k=1}^{K'} \epsilon_k^\pi(P) = H \sum_{k=1}^{K'} \mathbb{E}[\sum_{h=1}^{H} \beta_k(s^k_h, a^k_h) | \pi_k, P, F_{k-1}] = H \sum_{k=1}^{K'} \mathbb{E}[\sum_{h=1}^{H} \beta_k(s^k_h, a^k_h) | \pi_k, P, F_{k-1}]
\leq \left(\frac{4L}{n_k(s^k_h, a^k_h) \vee 1} \right) \sum_{k=1}^{K'} \mathbb{E} \left[ \sum_{h=1}^{H} \sqrt{\hat{P}_k(s^k_h, a^k_h) | \pi_k, P, F_{k-1}} \right]
\quad + H \sum_{k=1}^{K'} \mathbb{E} \left[ \sum_{h=1}^{H} \frac{(14/3)L}{n_k(s^k_h, a^k_h) \vee 1} | \pi_k, P, F_{k-1} \right]
\leq 2H\sqrt{S\sqrt{L}} \sum_{k=1}^{K'} \mathbb{E} \left[ \sum_{h=1}^{H} \frac{1}{n_k(s^k_h, a^k_h) \vee 1} | \pi_k, P, F_{k-1} \right]
\quad + \left(\frac{14}{3}\right)HSL \sum_{k=1}^{K'} \mathbb{E} \left[ \sum_{h=1}^{H} \frac{1}{n_k(s^k_h, a^k_h) \vee 1} | \pi_k, P, F_{k-1} \right]
\leq H\sqrt{S\sqrt{L}} \tilde{O}(\sqrt{SAH^3K'}) + HSL \tilde{O}(SAH) \leq \tilde{O}(S\sqrt{AH^3K'}) \quad (25)$$

Here, we get inequality (a) by the definition of $\beta_k$ (c.f. [6]). To get (b), note that $\sum_{s' \in S} \sqrt{\hat{P}_k(s'|s^k_h, a^k_h)} \leq \sqrt{\sum_{s'} \hat{P}_k(s'|s^k_h, a^k_h) \vee S}$ by Cauchy-Schwarz inequality and $\sum_{s'} \hat{P}_k(s'|s^k_h, a^k_h) = 1$. We get (c) using Lemma 8 and Lemma 9.

We now give the proof of Proposition 1.

Proof of Proposition 1. First note that even though $(\pi_b, P)$ is a feasible solution for the original CMDP problem [2], it may not be feasible for the OP problem [10]. To see this, note that since $V^*_c(P) = V^*_c(P') + \epsilon_k^\pi(P')$ and $V^*_c(P) = \hat{C}_b$, we will have $V^*_c(P) \leq \hat{C}$ if only if $\epsilon_k^\pi(P) \leq (C - \hat{C}_b)$. So, $(\pi_b, P)$ is a feasible solution for [10] if only if $\epsilon_k^\pi(P) \leq (C - \hat{C}_b)$. This condition may not be satisfied in the initial episodes.

However, since $\epsilon_k^\pi(P)$ is decreasing in $k$, if $(\pi_b, P)$ becomes a feasible solution for [10] at episode $k'$, then it will remain to be a feasible solution for all episodes $k \geq k'$.

Suppose $\pi_k = \pi_b$ for all $k \leq K'$. Also, suppose $(\pi_b, P)$ is not a feasible solution for [10] until episode $K' + 1$. Then, $\epsilon_k^\pi(P) > C - \hat{C}_b$ for all $k \leq K'$. So, we should get

$$K'(C - \hat{C}_b) \leq \sum_{k=1}^{K'} \epsilon_k^\pi(P) = \sum_{k=1}^{K'} \epsilon_k^\pi(P) \leq \tilde{O}(S\sqrt{AH^3K'})$$

where the last inequality is from Lemma 11. However, this inequality is violated for $K' \geq \tilde{O}(\frac{S^2AH^3}{(C - \hat{C}_b)^2})$.

So, $(\pi_b, P)$ is a feasible solution for [10] for any episode $k \geq K_o = \tilde{O}(\frac{S^2AH^3}{(C - \hat{C}_b)^2})$ provided that $\pi_k = \pi_b$ for all $k \leq K_o$.

\[19\]
The above result, however, only shows that $\pi_b$ becomes a feasible policy after some finite number of episodes. A natural question is, is $\pi_b$ the only feasible policy? In such a case, the OPSRL algorithm may not provide enough exploration to learn the optimal policy.

We alleviate the concerns about the above possible issue by showing that for all $k \geq K_\alpha$, there exists a feasible solution $(\pi', P)$ for the OP problem (10) such that $w^\pi_h(s,a;P) > 0$ for every $(s,a) \in S \times A$ with $w^\pi_h(s,a;P) > 0$. Informally, this implies that $\pi'$ will visit all state-action pairs that will be visited by the optimal policy $\pi^*$. This result can be derived as a corollary for Proposition 1.

**Corollary 1.** For any $k \geq K_\alpha$, there exists a policy $\pi'$ such that $(\pi', P)$ is a feasible solution for (10) and $w^\pi_h(s,a;P) > 0$ if $w^\pi_h(s,a;P) > 0$, $\forall (s,a,h) \in S \times A \times [H]$.

**Proof of Corollary 1.** Following the same steps as in the proof of Proposition 1, we can show that for $k \geq K_\alpha = \tilde{O}(\frac{\sum_{c\in HSL}(C(c_\epsilon - C_\epsilon))}{\epsilon^2})$, $\epsilon^\pi_k(P) \leq (\tilde{C} - C_\epsilon)/2$. We omit the constants for notational convenience.

So, for $k \geq K_\alpha$, we get $V^{\pi_h}_c(P) = V^{\pi_h}_c(P) + \epsilon^\pi_k(P) \leq (\tilde{C} + C_\epsilon)/2$.

Using the occupancy measures $w^\pi_h$ and $w^\pi_h'$, define a new occupancy measure $w^h(s,a;P) = (1 - \alpha)w^\pi_h(s,a;P) + \alpha w^\pi_h'(s,a;P)$ for an $\alpha > 0$. Note that $w'$ is a valid occupancy measure since the set of occupancy measure is convex (c.f. Proposition 3). Let $\pi'$ be the policy corresponding to the occupancy measure $w'$, which can be obtained according to (14) so that $w' = w_{\pi'}$. Note that $w^h(s,a;P) > 0$ if $w^\pi_h(s,a;P) > 0$, $\forall (s,a,h) \in S \times A \times [H]$, by construction.

Now, we will show that $(\pi', P)$ is a feasible solution for the OP problem (10) for all $k \geq K_\alpha$.

First, note that $\epsilon^\pi_k(P) \leq 7HSL$ for any policy $\pi$. So, $V^{\pi_h}_c(P) = V^{\pi_h}_c(P) + \epsilon^\pi_k(P) \leq 8HSL$. We also know that value function is linear in the occupancy measure (c.f. Section A). Due to this linearity, we get $V^{\pi_h}_c(P) = (1 - \alpha)V^{\pi_h}_c(P) + \alpha V^{\pi_h}_c(P) \leq (1 - \alpha)(\tilde{C} + C_\epsilon)/2 + \alpha 8HSL$. Selecting $\alpha < \frac{\tilde{C} - C_\epsilon}{16HSL}$ will ensure that $V^{\pi_h}_c(P) < \tilde{C}$, which in turn guarantees that $(\pi', P)$ is feasible.

**D.2 Proof of Proposition 2.**

**Proof.** For any episode $k \leq K_\alpha$, we have $\pi_k = \pi_b$, and it is safe by Assumption 1. For $k \geq K_\alpha$, (10) is feasible according to Proposition 1. Since $(\pi_k, P_k)$ is the solution of (10), we have $V^{\pi_k}_c(P_k) \leq \tilde{C}$.

We will now show that $V^{\pi_k}_c(P) \leq \tilde{C}$, conditioned on the good event $G$.

By the value difference lemma (Lemma 10), we have

\[
V^{\pi_k}_c(P) - V^{\pi_k}_c(P_k) = \mathbb{E}\sum_{h=1}^{H}((P - P_k)(\cdot|s_h^k, a_h^k))^T V^{\pi_h+1,\pi_k}(\cdot; P)|\pi_k, P_k, F_{k-1})
\]

\[\leq \mathbb{E}\sum_{h=1}^{H} \|((P - P_k)(\cdot|s_h^k, a_h^k))\|_1 \|V^{\pi_h+1,\pi_k}(\cdot; P)\|_\infty|\pi_k, P_k, F_{k-1})
\]

\[\leq H \mathbb{E}\sum_{h=1}^{H} \beta_k(s_h^k, a_h^k)|\pi_k, P_k, F_{k-1}) = \bar{\beta}_k(P_k).
\]

(26)

Here, we get (a) by H"older’s inequality inequality. To get (b), we make use of two observations. First, note that $\|V^{\pi_h+1,\pi_k}(\cdot; P)\|_\infty \leq H$ because the expected cumulative cost cannot be greater than $H$ since $|c(\cdot, \cdot)| \leq 1$ by assumption. Second, under the good event $G$, $\sum_{s'} |P(s'|s, a) - P_k(s'|s, a)| \leq \sum_{s'} |\beta_k(s, a, s') = \bar{\beta}_k(s, a)$.

From (26), we get

\[V^{\pi_k}_c(P) \leq V^{\pi_k}_c(P_k) + \epsilon^\pi_k(P_k) \leq V^{\pi_k}_c(P_k) \leq \tilde{C},
\]

where (c) is by definition and (d) is from the fact that $(\pi_k, P_k)$ is the solution of (10). So, $V^{\pi_k}_c(P) \leq \tilde{C}$, and hence $\pi_k$ is safe, under the good event $G$. So, this is statement holds with a probability greater than $1 - 3\delta$, according to Lemma 7. 

\[\square\]
D.3 Proof of Lemma 1

Proof.

\[
\sum_{k=K_o}^{K} V_r^{{\pi}_k}(P) - V_r^{{\pi}^*(P)} \overset{(a)}{=} \sum_{k=K_o}^{K} \mathbb{E} \left[ \sum_{h=1}^{H} ((P_k - P)(\cdot | s_h^k, a_h^k))^T V_r^{h+1, {\pi}_k}(\cdot; P_k) | {\pi}_k, P, F_{k-1} \right]
\]

\[
\overset{(b)}{\leq} \sum_{k=K_o}^{K} \mathbb{E} \left[ \sum_{h=1}^{H} \| (P_k - P)(\cdot | s_h^k, a_h^k) \|_1 \| V_r^{h+1, {\pi}_k}(\cdot; P_k) \|_{\infty} | {\pi}_k, P \right]
\]

\[
\overset{(c)}{\leq} H \sum_{k=K_o}^{K} \mathbb{E} \left[ \sum_{h=1}^{H} \sum_{s' \in S} \beta_k(s_h^k, a_h^k, s') | {\pi}_k, P, F_{k-1} \right]
\]

\[
\overset{(d)}{\leq} H \sum_{k=K_o}^{K} \mathbb{E} \left[ \sum_{h=1}^{H} \tilde{\beta}_k(s_h^k, a_h^k) | {\pi}_k, P, F_{k-1} \right] \overset{(e)}{=} \sum_{k=K_o}^{K} \epsilon_k^\pi(P) \overset{(f)}{\leq} \tilde{O}(S\sqrt{AH^3K}).
\]

Here (a) is obtained by the value difference lemma (Lemma 10), (b) by by Holder’s inequality, (c) by conditioning on \(G\) and using the fact value function is upper bounded by \(H\), (d) by definition of \(\tilde{\beta}\), and (e) by definition of \(\epsilon_k^\pi(P)\). Final inequality (f) is obtained from Lemma 11.

\[\square\]

D.4 Proofs of Lemma 2 and Theorem 1

We first prove two supporting lemmas.

**Lemma 12.** Let \((\pi_k, P_k)\) be the optimal solution corresponding to the OP problem (10). Then,

\[V_r^{\pi_k}(P_k) - V_r^{\pi^*(P)} \leq \frac{H}{(C - \bar{C}_b)} \epsilon_k^\pi(P_k).
\]

**Proof.** We will first consider a more general version of the OP problem (10) as

\[(\tilde{\pi}_k, \tilde{P}_k) = \arg \min_{\pi', P' \in P_k} V_{r_k}(P') \text{ subject to } V_{r_k}(P') \leq \bar{C}, \quad (27)
\]

where we change \(r_k\) in (10) to \(\bar{r}_k\) above, with \(\bar{r}_k(s, a) = r(s, a) - bH\tilde{\beta}_k(s, a)\). Note that (27) reduces to (10) for \(b = \frac{H}{C - \bar{C}_b}\) and hence it is indeed a general version.

Using the occupancy measures \(w_{\bar{h}}^k\) and \(w_{\bar{s}}^k\), define a new occupancy measure \(\tilde{w}^h(s, a) = (1 - \alpha_k)w_{\bar{h}}^k(s, a; P) + \alpha_kw_{\bar{s}}^k(s, a; P)\) for an \(\alpha_k > 0\). Note that \(\tilde{w}\) is a valid occupancy measure since the set of occupancy measure is convex (c.f. Proposition 3). Let \(\tilde{\pi}\) be the policy corresponding to the occupancy measure \(\tilde{w}\), which can be obtained according to (14) so that \(\tilde{w} = w_{\tilde{\pi}}\).

**Claim 1:** \((\tilde{\pi}, P)\) is a feasible solution for (27) when \(\alpha_k\) satisfies the sufficient condition

\[\alpha_k \leq \frac{\bar{C} - \bar{C}_b - \epsilon_k^\pi(P)}{\bar{C} - \bar{C}_b + \epsilon_k^\pi(P) + \epsilon_k^\pi(P)}.
\]

**Proof of Claim 1:** Since value function is a linear function of the occupancy measure, we have \(V_{\tilde{\pi}}(\tilde{P}) = (1 - \alpha_k)(V_{\pi} + \epsilon_k^\pi(P)) + \alpha_k(V_{\pi} + \epsilon_k^\pi(P))\). For \((\tilde{\pi}, P)\) to be a feasible solution for (27), it must be true that \(V_{\tilde{\pi}}(\tilde{P}) \leq \bar{C}\). Using the fact that \(V_{\pi} = \bar{C}_b\) and \(V_{\pi}^* \leq \bar{C}\), it is sufficient to get an \(\alpha_k\) such that

\[\bar{C} - \bar{C}_b + \epsilon_k^\pi(P) + \alpha_k(\bar{C} + \epsilon_k^\pi(P)) \leq \bar{C}.
\]

This yields a sufficient condition (28). Note that \(\alpha_k\) is non-negative because \(\epsilon_k^\pi(P) \leq \bar{C} - \bar{C}_b\) for \(k \geq K_o\), as shown in the proof of Proposition 1. This concludes the proof of Claim 1.
Claim 2: \( V_{\bar{r}_k}^{\pi_k}(\bar{P}_k) \leq V_r^{\pi^*}(P) \) if \( b \) satisfies the sufficient condition

\[
b \geq \frac{H}{C - C_b}. \tag{29}
\]

Proof of Claim 2: Selecting an \( \alpha_k \) that satisfies the condition \( \bar{C} \). \((\bar{\pi}, P)\) is a feasible solution of \((27)\). Since \((\bar{\pi}_k, \bar{P}_k)\) is the optimal solution of \((27)\), we have \(V_{\bar{r}_k}^{\pi_k}(\bar{P}_k) \leq V_{\bar{r}_k}^{\bar{\pi}_k}(P)\). So, it is sufficient to find a \( b \) such that \( V_{\bar{r}_k}^{\pi_k}(P) \leq V_{\bar{r}_k}^{\bar{\pi}_k}(P) \). Using the linearity of the value function w.r.t. occupancy measure, this is equivalent to \((1 - \alpha_k)(V_{\bar{r}_k}^{\pi_k} - b_{\bar{r}_k}^{\pi_k}(P)) + \alpha_k(V_{\bar{r}_k}^{\bar{\pi}_k} - b_{\bar{r}_k}^{\bar{\pi}_k}(P)) \leq V_{\bar{r}_k}^{\bar{\pi}_k}(P)\). This will yield the condition \( b \geq \frac{V_{\bar{r}_k}^{\pi_k}(P) - V_{\bar{r}_k}^{\bar{\pi}_k}(P)}{C - C_b} \). Now, choosing \( \alpha_k \) that satisfies the condition \( \bar{C} \) will ensure that \( \frac{\alpha_k}{1 - \alpha_k} \leq \frac{C - C_b - \bar{C}}{C - C_b} \). Using this in the previous inequality for \( b \), we get the sufficient condition \( b \geq \frac{C - C_b - \bar{C}}{C - C_b} \). Since \( V_{\bar{r}_k}^{\pi_k}(P) \leq H \) and \( V_{\bar{r}_k}^{\bar{\pi}_k}(P) \geq 0 \), we get the sufficient condition \( \bar{C} \). This concludes the proof of Claim 2.

Now, let \( b = \frac{H}{C - C_b} \). So, \( \bar{r}_k = r_k \) and \((\bar{\pi}_k, \bar{P}_k) = (\pi_k, P_k)\). By Claim 2, we have \(V_{\bar{r}_k}^{\pi_k}(P_k) \leq V_{\bar{r}_k}^{\bar{\pi}_k}(P)\). This implies,

\[
V_{\bar{r}_k}^{\pi_k}(P_k) - \frac{H}{C - C_b} \epsilon_k^{\pi_k}(P_k) \leq V_{\bar{r}_k}^{\bar{\pi}_k}(P).
\]

After rearranging, we get the desired result.

Lemma 13. With a probability greater than \((1 - 3\delta)\), \( \sum_{k=K_0}^{K} \epsilon_k^{\pi_k}(P_k) \leq \tilde{O}(S^2 \sqrt{AH^{3K}}) \).

Proof. We can write,

\[
\sum_{k=K_0}^{K} \epsilon_k^{\pi_k}(P_k) = \sum_{k=K_0}^{K} (\epsilon_k^{\pi_k}(P_k) - \epsilon_k^{\pi_k}(P)) + \sum_{k=K_0}^{K} \epsilon_k^{\pi_k}(P). \tag{30}
\]

We will consider the first summation:

\[
\sum_{k=K_0}^{K} \epsilon_k^{\pi_k}(P_k) - \epsilon_k^{\pi_k}(P) = H \sum_{k=K_0}^{K} \mathbb{E} \left[ \sum_{h=1}^{H} \bar{\beta}_k(s_k^h, a_k^h)|\pi_k, P_k, F_{k-1} \right] - \mathbb{E} \left[ \sum_{h=1}^{H} \bar{\beta}_k(s_k^h, a_k^h)|\pi_k, P, F_{k-1} \right]
\]

\[
\leq H \sum_{k=K_0}^{K} \mathbb{E} \left[ \sum_{h=1}^{H} (P_k - P)(s_k^h, a_k^h) \bar{\pi}_k, P_k, F_{k-1} \right]
\]

\[
\leq H \sum_{k=K_0}^{K} \mathbb{E} \left[ \sum_{h=1}^{H} ||P_k - P||_{1}||V_{\bar{\beta}_k}^{h+1, \pi_k}||_{\infty, P, F_{k-1}} \right]
\]

\[
\leq 7H^2 SL \sum_{k=K_0}^{K} \mathbb{E} \left[ \sum_{h=1}^{H} \bar{\beta}_k(s_k^h, a_k^h)|\pi_k, P, F_{k-1} \right]
\]

\[
\leq 7HSL \sum_{k=K_0}^{K} \epsilon_k^{\pi_k}(P). \tag{31}
\]

We get (a) by the definition of value function with \( \bar{\beta}_k \) as the cost function, and then applying the value difference lemma (Lemma 10). Since \( \delta_k(s, a) \leq 7SL \), we get \(||V_{\bar{\beta}_k}^{h+1, \pi_k}||_{\infty} \leq 7HSL \). So, the
We denote by $\mathcal{G}$. Using (31) in (30), we get

$$
\sum_{k=K_0}^{K} \epsilon^*_k(P_k) \leq 8HSL \sum_{k=K_0}^{K} \epsilon^*_k(P) \overset{(c)}{\leq} 8HSL \hat{O}(S^2AH^3K),
$$

where (c) follows from Lemma [11] \hfill \Box

We now give the proof of Lemma [2]

Proof of Lemma 2

$$
\sum_{k=K_0}^{K} V^r_{\pi_k}(P_k) - V^*_{\pi}(P) \overset{(a)}{\leq} \sum_{k=K_0}^{K} \frac{H}{(C - C_b)} \epsilon^*_k(P_k) \overset{(b)}{\leq} \hat{O}(\frac{S^2}{(C - C_b)} \sqrt{AH^7K}).
$$

Here, (a) follows from Lemma [12] and (b) follows from Lemma [13] \hfill \Box

We now present the proof of Theorem 1.

Proof of Theorem 1. The regret for OPSRL algorithm after $K$ episodes is decomposed as,

$$
R(K) = \sum_{k=1}^{K_0} (V^r_{\pi_k}(P) - V^*_{\pi}(P)) + \sum_{k=K_0+1}^{K} V^*_{\pi_k}(P) - V^*_{\pi_k}(P_k) + \sum_{k=K_0+1}^{K} V^*_{\pi_k}(P_k) - V^*_{\pi}(P).
$$

Since $(V^r_{\pi_k}(P) - V^*_{\pi}(P)) \leq H$, we can bound the first term using the bound on $K_0$ from Proposition 1. We bound the second term using Lemma 1 and the third term using Lemma 2. This will give, with probability $1 - 3\delta$,

$$
R(K) \leq \hat{O}(\frac{S^2AH^4}{(C - C_b)^2}) + \hat{O}(S\sqrt{AH^3K}) + \hat{O}(\frac{S^2}{(C - C_b)} \sqrt{AH^7K}).
$$

Moreover, from proposition 2, we have that $\pi_k \in \Pi_{safe}$ for all $k \in [K]$, with probability $1 - 3\delta$. \hfill \Box

E Experiments

We now validate our theoretical findings with additional simulations. We once again fix $K$, the maximum number of episodes in which we employ the algorithms. As before, we measure the performance of the algorithms using two forms of regret metrics, namely, (i) Optimality Regret, defined as in [3], and (ii) Constraint Regret, defined as $\sum_{k=1}^{K} \max \{0, V^*_c(P) - \hat{C}\}$, where $\pi_k$ is the output of the algorithm in question at episode $k$. We chose two constrained MDPs for the validation purpose. In the main paper, we presented a Media control Environment. We now present the details of Inventory control Environment, and provide regret results. For each of these environments, we average the performance of each algorithm over 5 simulation runs.

E.1 Inventory Control Environment

We consider a single product inventory control problem, similar to the one in Bertsekas et al. [2000]. Our environment evolves according to a finite horizon CMDP, with horizon length $H = 7$, where each time step $h \in [H]$ represents a day of the week. In this problem, our goal is to maximize the expected total revenue over a week, while keeping the expected total costs in that week below a certain level. We do not backlog the demands.

The storage has a maximum capacity $N = 6$, which means it can store a maximum of 6 items. We denote by $s_h$, the state of the environment, as the amount of inventory available at $h^{th}$ day. The
action \( a_h \) is the amount of inventory the agent purchases such that the inventory does not overflow. Thus, the action space \( \mathcal{A}_s \in \{0, \ldots, N - s\} \), for the state \( s \). The exogenous demand is represented by \( d_h \), which is a random variable representing the stochastic demand for the inventory on the \( h^{th} \) day. We assume \( d_h \) to be in \( \{0, \cdots N\} \) with distribution \([0.3, 0.2, 0.2, 0.05, 0.05]\). If the demand is higher than the inventory and supply, the excess demand will not be met. The state evolution then follows,

\[
s_{h+1} = \max\{0, s_h + a_h - d_h\}.
\]

We define the rewards and costs as follows. The revenue is generated as, \( f(s, a, s') = 8(s + a - s') \), when \( s' > 0 \), and is 0 otherwise. The reward obtained in state \((s, a)\) is then the expected revenue over all next states \( s' \), \( r(s, a) = \mathbb{E}[f(s, a, s')] \). The cost associated with the inventory has two components. Firstly, there is a purchase cost when the inventory is brought in, which is a fixed cost of 4 units, plus a variable cost of \( 2a \), which increases with the amount of purchase. Secondly, we also have a non-decreasing holding cost \( s \), for storing the inventory. Hence, the cost in \((s, a)\) is \( c(s, a) = 4 + 2a + s \). We normalize the rewards and costs to be in the range \([0, 1]\). Our goal is to maximize the expected total revenue over a week \((H = 7)\), while keeping the expected total costs in that week below a threshold \( \bar{C} \).

### E.2 Experiment setup and Results

We now validate our OPSRL algorithm with this environment. As before, we compare it against the OptCMDP Algorithm 1 in [Efroni et al., 2020], and UCRL2 algorithm from [Jaksch et al., 2010]. We use UCRL2 algorithm as a baseline for the \( \bar{O}(\sqrt{K}) \) optimality regret. The OptCMDP algorithm characterizes the objective regret and the constraint regret of theCMDP. We expect to see that the OPSRL algorithm achieves the same optimality regret, as these algorithms in order sense, while being safe in all episodes of learning. We emphasize that the performance metric for UCRL2 is computed with respect to the optimal objective value of the unconstrained problem, instead of the constrained problem. As before, we choose the optimal policy from a conservative constrained problem (with a much stricter constraint) as the baseline policy. We use \( \bar{C}_b = 0.1\bar{C} \).

Fig. (2a) compares the optimality regret for inventory control environment incurred by each algorithm with respect to number of episodes. As we see in this figure, in the initial episodes, the objective regret of OPSRL grows linearly with number of episodes. Later, the growth rate of regret changes to square root of number of episodes. We see that this change of behavior happens after \( K_0 \) episodes specified by Proposition 1. Hence, the linear growth rate indeed corresponds to the time when the base policy is employed. In conclusion, the regret for OPSRL algorithm depicted in Figures 2a matches the result of Theorem 1. Further, we also observe that the objective regret of OptCMDP is lower than OPSRL. This behaviour can be attributed to the fact that in order to perform safe exploration, OPSRL includes a pessimistic penalty in the constraint \( (8) \). This argument also explains UCRL2 outperforming OPSRL.

Fig. (2b) compares the regret in constraint violation for OPSRL and OptCMDP algorithms, for the inventory control setting. Here, we see that OPSRL does not violate the constraint, while OptCMDP incurs a regret that grows sublinearly. This figure shows that OPSRL does indeed perform safe exploration as we presented in the theory, while OptCMDP violates the constraints during learning. UCRL2 has an enormous linearly growing constraint violation regret compared to other two algorithms, which hinders us to show the distinction between the rest of the algorithms clearly in the figure. Hence, we omitted that detail. Finally, Fig. (2c) compares the optimality regret for various baseline policies.

### E.3 Media Control Environment

We now provide a detailed description of our environment in the main paper. Here, we choose to control Media streaming to a device from a wireless base station. The base station provides two
types of service to the device, a fast service and a slow service. The packets received are stored in a media buffer at the device. Here, we want to minimize the cost of having an empty buffer, while keeping the utilization of fast service below certain level.

We denote by $A_h$, the number of incoming packets into the buffer, and by $B_h$, the number of packets leaving the buffer. The state of the environment is denoted as $s_h$ in $h^{th}$ step, which is the media buffer length. It evolves as follows:

$$s_{h+1} = \min\{\max(0, s_h + A_h - B_h), N\},$$

where we consider $N = 20$, as the maximum buffer length in our experiment. The action space is $\mathcal{A} = \{1, 2\}$, i.e., the action is to use either fast server 1 or slow server 2. We assume that the service rates of the servers have independent Bernoulli distributions, with parameters $\mu_1 = 0.9$, and $\mu_2 = 0.1$, where $\mu_1$ corresponds to the fast service. The media playback at the device is also Bernoulli with parameter $\gamma$. Hence, $A_h$ is a random variable with mean either $\mu_1$ or $\mu_2$ depending on the action taken, and $B_h$ is a random variable with mean $\gamma$. These components constitute the unknown transition dynamics of our environment.

The objective cost is $r(s, a) = 1_{\{s = 0\}}$, i.e., it has a value of 1, when the buffer hits zero, and is zero everywhere else. Our constraint cost is $c(s, a) = 1_{\{a = 1\}}$, i.e., there is a constraint cost of 1 when the fast service is used, and is zero otherwise. We then constrain the expected number of times the fast service is used to $H$ in a horizon of length $H = 10$. 

Figure 2: Illustrating the Optimality Regret and Constraint Regret for the Inventory Control Environment.