DOPE: Doubly Optimistic and Pessimistic Exploration for Safe Reinforcement Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Safe reinforcement learning is extremely challenging. Not only must the agent explore an unknown environment, it must do so while ensuring no safety constraint violations. The problem is typically posed as a constrained Markov decision process (MDP) under an unknown model, often with the learning agent having access to a safe suboptimal baseline policy. Recent results obtain an $\tilde{O}(|\mathcal{S}|\sqrt{|\mathcal{S}|K})$ objective regret in K episodes for an MDP with $|\mathcal{S}|$ states, while being safe at all times. The main idea is to combine a reward bonus for exploration (optimism) with a conservative constraint (pessimism). However, the approach is so pessimistic that empirical results indicate an inordinately long learning process that keeps on applying sequences of the safe baseline policy. Our key insight is that such excessive pessimism hinders exploration, and needs to be combated by optimism with respect to the model. This insight yields DOPE, which has a double dose of optimism with respect to model and reward, while being pessimistic with respect to the constraints. We show that DOPE reduces the objective regret to $\tilde{O}(|\mathcal{S}|\sqrt{K})$, with no constraint violation. Furthermore, we show in empirical studies that DOPE has a dramatic performance improvement as compared to earlier approaches.

1 Introduction

Constrained Markov Decision Processes (CMDPs) impose restrictions that pertain to resource or safety constraints of the system. For example, the average radiated power in a wireless communication system must be restricted due to user health and battery lifetime considerations or the frequency of braking or accelerating in an autonomous vehicle must be kept bounded to ensure passenger comfort. Since these systems have complex dynamics, a constrained reinforcement learning (CRL) approach is attractive for determining an optimal control policy. But how do we ensure that safety or resource availability constraints are not violated while learning such an optimal control policy?

Our goal is to develop a framework for safe exploration without constraint violation (with high probability) 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, ensuring *safe exploration* in the CRL setting has received less attention. The problem is challenging, since we do not allow constraint violation either during learning or deployment, while ensuring a low regret in terms of the optimal objective. Our aim is to explore 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.

There has been much recent interest in model-based RL approaches to solving a constrained MDP, the most relevant of which we have summarized in Table 1. The setup is of a finite horizon episodic CMDP, with a state space of size $|\mathcal{S}|$, an action space of size $|\mathcal{A}|$ and a horizon of length H. Regret is measured over the first K episodes of the algorithm. Both the attained objective and constraint satisfaction are computed in an expected sense for a given policy. Allowing constraint violations

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Table 1: Comparison of safe RL algorithms for CMDPs. The algorithms are: OptCMDP, OptCMDP-Bonus [17], AlwaysSafe [34] and OptPessLP [26]. This table is presented for $K \geq \text{poly}(|S|, |A|, H)$, with polynomial terms independent of K omitted. Expected regret results are in high probability.

ALGORITHM	MODEL OPTIMISM	REWARD OPTIMISM	CONSTRAINT PESSIMISM	OBJECTIVE REGRET	CONSTRAINT REGRET	EMPIRICAL PERF.
OPTCMDP	✓	✗	✗	$\tilde{O}(\lvert\mathcal{S}\rvert\sqrt{\lvert\mathcal{A}\rvert H^4 K})$	$\tilde{O}(\lvert\mathcal{S}\rvert\sqrt{\lvert\mathcal{A}\rvert H^4 K})$	-
OPTCMDP-B	✗	✓	✗	$\tilde{O}(\lvert\mathcal{S}\rvert\sqrt{\lvert\mathcal{A}\rvert H^4 K})$	$\tilde{O}(\lvert\mathcal{S}\rvert\sqrt{\lvert\mathcal{A}\rvert H^4 K})$	-
ALWAYSSAFE	✓	✗	HEURISTIC	UNKNOWN	0	✗
OPTPESS-LP	✗	✓	✓	$\tilde{O}(\frac{H^3}{\overline{C}-\overline{C}_b}\sqrt{\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert K})$	0	✗
DOPE	✓	✓	✓	$\tilde{O}(\frac{H^3}{\overline{C}-\overline{C}_b}\lvert\mathcal{S}\rvert\sqrt{\lvert\mathcal{A}\rvert K})$	0	✓

during learning means that the algorithm suffers both an objective regret and a constraint regret. Examples of algorithms following this method are OptCMDP, OptCMDP-Bonus [17]. The algorithms use different ways of incentivizing exploration to obtain samples to build up an empirical system model. OptCMDP uses the idea of optimism in the face of uncertainty from the model perspective and solves an extended linear program to find the best model (highest reward) under the samples gathered thus far. OptCMDP-Bonus uses a different approach of adding a bonus to the reward term to incentivize exploration of state-action pairs that have fewer samples, which we consider as a form of optimism from the reward perspective. While the objective regret of both these algorithms is $\tilde{O}(\sqrt{K})$, the constraint regret too is of the same order, since constraints may be violated during learning.

An algorithm that begins with no knowledge of the model will have exploration steps during learning that might violate constraints in expectation. Hence, safe RL approaches assume the availability of an inexpert baseline policy that does not violate the constraints, but is insufficient to explore the entire state action space. So it cannot be simply applied until a high-accuracy model is obtained. A heuristic approach entitled AlwaysSafe [34] assumes a factored CMDP that allows the easy generation of a safe baseline policy. It then combines the optimism with respect to the model of OptCMDP with a heuristically chosen hardening of constraints. This guarantees no constraint violations, but does not have regret guarantee with respect to the objective. Its empirical performance is variable and its use is limited to factored CMDP problems.

The OptPessLP algorithm [26] formalizes the idea of coupling optimism and pessimism by starting with OptCMDPBonus that has an optimistic reward, and systematically applying decreasing levels of pessimism with respect to the constraint violations. The approach is successful in ensuring the twin goals of a $\tilde{O}(\sqrt{K})$ objective regret, while ensuring no constraint violations. However, the authors do not present any empirical performance evaluation results. When we implemented OptPessLP, we found that the performance is singularly bad in that linear regret persists for a large number of samples, and the tapering off to $\tilde{O}(\sqrt{K})$ regret behavior does not appear to happen quickly. The problem with this algorithm is that it is so pessimistic with regard to constraints that it heavily disincentivizes exploration and ends up choosing the base policy for long sequences.

The issue upon which algorithm performance depends is the choice of how to combine optimism and pessimism to obtain both order optimal and empirically good objective regret performance, while ensuring no constraint violations happen. Our insight is that optimism with respect to the model is a key enabler of exploration, and can be coupled with the addition of optimism with respect to the reward. This *double dose of optimism*—both with respect to model and reward—could ensure that pessimistic hardening of constraints does not excessively retard exploration. Following this insight, we develop DOPE, a doubly optimistic and pessimistic exploration approach. We are able to show that DOPE not only attains $\tilde{O}(\sqrt{K})$ objective regret behavior with zero constraint regret with high probability, it also reduces the objective regret bound over OptPessLP by a factor of $\sqrt{|\mathcal{S}|}$. We conduct performance analysis simulations under representative CMDP problems and show that DOPE easily outperforms all the earlier approaches. Thus, the idea of double optimism is not only valuable from the order optimal algorithm design perspective, it also shows good empirical regret performance, indicating the feasibility of utilizing the methodology in real-world systems.

Related Work:

Constrained RL: Constrained Markov Decision Processes (CMDP) has been an active area of research [2], with applications in domains such as power systems [37, 24], communication networks [3, 36], and robotics [16, 10]. CMDP planing problem (with a known model) is typically solved

using a linear programming approach or a Lagrangian approach [2]. In [9], the author proposed an actor-critic RL algorithm for learning the asymptotically optimal policy for an infinite horizon average cost CMDP when the model is unknown. This approach is also utilized in function approximation settings with asymptotic local convergence guarantees [8, 11, 38]. Policy gradient algorithms for CMDPs have also been developed [1, 41, 44, 33, 15]. However, none these works address the problem of safe exploration to provide guarantees on the constraint violations during learning.

Safe Multi-Armed Bandits: The problem of safe exploration in linear bandits with stage-wise safety constraint is studied in [4, 23, 29, 31]. Linear bandits with more general constraints have also been studied [32, 27]. These works do not consider the more challenging RL setting which involves an underlying dynamical system with unknown model.

Safe Online Convex Optimization: Online convex optimization [19] has been studied with stochastic constraints [42] and adversarial constraints [30, 43, 25]. These allow constraint violation during learning and characterize the cumulative amount of violation. A safe Frank-Wolf algorithm for convex optimization with unknown linear stochastic constraints has been studied in [39]. However, these too do not consider the RL setting with an unknown model.

Exploration in Constrained RL: There has been much work in this area *with* constraint violations during learning, including the work discussed in the introduction [17]. These include [35, 18, 22], which derive bounds either on the objective and constraint regret or on the sample complexity of learning an ϵ-optimal policy. Other works on safe RL include [40], where a model-free approach is considered, and [5], that peratins to offline RL. These are complementary to our model-based approach. The problem of learning the optimal policy of a CMDP without violating the constraints was also studied in [45]. However, they assume that the model is known and only the cost functions are unknown, whereas we address more difficult problem with unknown model and cost functions.

Notations: For any integer M, $[M]$ denotes the set $\{1, \ldots, M\}$. For any two real numbers a, b, $a \vee b := \max\{a, b\}$. For any given set $\mathcal{X}$, $\Delta(\mathcal{X})$ denotes the probability simplex over the set $\mathcal{X}$, and $|\mathcal{X}|$ denotes the cardinality of the set $\mathcal{X}$.

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) [2]. We consider a CMDP, denoted as $M = \langle \mathcal{S}, \mathcal{A}, r, c, P, H, \bar{C} \rangle$ with $r = (r_h)_{h=1}^H, c = (c_h)_{h=1}^H, P = (P_h)_{h=1}^H$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, H is the episode length, $r_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the objective cost function at time step $h \in [H]$, $c_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the constraint cost function at time step $h \in [H]$, P_h is the transition probability function with $P_h(s'|s, a)$ representing the probability of transitioning to state s' when action a is taken at state s at time h. In the RL context, the transition matrix P is also called the model of the CMDP. Finally, $\bar{C}$ is a scalar that specifies the safety constraint in terms of the maximum permissible value for the expected cumulative constraint cost. We consider the setting where $|\mathcal{S}|$ and $|\mathcal{A}|$ are finite. Also, without loss of generality, we assume that costs r and c are bounded in $[0, 1]$.

A non-stationary randomized policy $\pi = (\pi_h)_{h=1}^H, \pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ specifies the control action to be taken at each time step $h \in [H]$. 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 $l : [H] \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, the value function of a policy π corresponding to time step $h \in [H]$ given a state $s \in \mathcal{S}$ is defined as

$$V_{l,h}^{\pi}(s; P) = \mathbb{E}[\sum_{\tau=h}^{H} l_\tau(s_\tau, a_\tau)|s_h = s], \tag{1}$$

where $a_\tau \sim \pi_\tau(s_\tau, \cdot),\ s_{\tau+1} \sim P_\tau(\cdot|s_\tau, a_\tau)$. Since we are mainly interested in the value of a policy starting from $h = 1$, we simply denote $V_{l,1}^{\pi}(s; P)$ as $V_l^{\pi}(s; P)$. For the rest of the paper, we will assume that the initial state is s_1 is fixed. So, we will simply denote $V_l^{\pi}(s_1; P)$ as $V_l^{\pi}(P)$, when it is clear from the context. This standard assumption [17, 14] can be made without loss of generality.

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

$$\min_{\pi} \quad V_r^{\pi}(P) \quad \text{s.t.} \quad V_c^{\pi}(P) \leq \bar{C}. \tag{2}$$

We say that a policy π is a **safe policy** if $V_c^\pi(P) \leq \bar{C}$, i.e., if the expected cumulative constraint cost corresponding to the policy π is less than the maximum permissible value $\bar{C}$. The **set of safe policies**, denoted as Π_{safe}, is defined as $\Pi_{\text{safe}} = \{\pi : V_c^\pi(P) \leq \bar{C}\}$. Without loss of generality, we assume that the CMDP problem (2) is feasible, i.e., Π_{safe} is non-empty. Let π^* 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 [2]. 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 [2] to find the optimal policy. In this work, we will use the LP approach, consistent with model optimism. Details of solving (2) using the LP approach are in Appendix A.

2.2 Reinforcement Learning with Safe Exploration

The goal of the reinforcement learning with safe exploration is to solve (2), 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 Π_{safe}. Since Π_{safe} itself is defined based on the unknown P, the learning algorithm will not know Π_{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. 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 $a_{h,k} \sim \pi_{h,k}(s_{h,k}, \cdot)$, and incurs the costs $r_h(s_{h,k}, a_{h,k})$ and $c_h(s_{h,k}, a_{h,k})$. 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., $s_{1,k} = s_1$. We also assume that the maximum permissible cost $\bar{C}$ for any exploration policy is known and it is specified as part of the learning problem.

The performance of the RL algorithm is measured using the metric of *safe objective regret*. The safe objective regret is defined exactly as the standard regret of an RL algorithm for exploration in MDPs [20, 13, 6], but with an additional constraint that the exploration polices should belong to the safe set Π_{safe}. Formally, the safe objective regret $R(K)$ after K learning episodes is defined as

$$R(K) = \sum_{k=1}^{K} (V_r^{\pi_k}(P) - V_r^{\pi^*}(P)), \ \pi_k \in \Pi_{\text{safe}}, \forall k \in [K]. \tag{3}$$

Since Π_{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** π_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 π_b such that $V_c^{\pi_b}(P) = \bar{C}_b$, where $\bar{C}_b < \bar{C}$. The value $\bar{C}_b$ is also known to the algorithm.

Remark 2. Knowing a safe policy π_b is necessary for solving the safe RL problem because we require the constraint to always be satisfied. A similar assumption has been used in the case of safe exploration in linear bandits [4, 23, 31], as well as in earlier work on safe RL [34, 26].

3 Algorithm and Performance Guarantee

DOPE builds on the *optimism in the face of uncertainty (OFU)* style exploration algorithms for RL [20, 13], using such optimism, both in terms of the model, as well as to provide a reward bonus for under-explored state-actions. However, a naive OFU-style algorithm may lead to selecting exploration policies that are not in the safe set Π_{safe}. So we modify the selection of exploratory policy by incorporating *pessimism in the face of uncertainty (PFU)* on the constraints, making DOPE doubly optimistic and pessimistic in exploration.

DOPE 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 $n_{h,k}(s,a) = \sum_{k'=1}^{k-1} \mathbb{1}\{s_{h,k'} = s, a_{h,k'} = 1\}$ be the number of times the pair (s,a) was observed until the beginning of episode k. Similarly, define $n_{h,k}(s,a,s') = \sum_{k'=1}^{k-1} \mathbb{1}\{s_{h,k'} = s, a_{h,k'} = 1, s_{h+1,k'} = s'\}$. At the beginning of each episode k, DOPE estimates the model as $\widehat{P}_{h,k}(s'|s,a) = \frac{n_{h,k}(s,a,s')}{n_{h,k}(s,a)\vee 1}$. Similar to OFU-style algorithms, we construct a confidence set $\mathcal{P}_k$ around $\widehat{P}_k$ as $\mathcal{P}_k = \bigcap_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathcal{P}_k(s,a)$, where

$$\mathcal{P}_k(s,a) = \{P' : |P'_h(s'|s,a) - \widehat{P}_{h,k}(s'|s,a)| \le \beta^p_{h,k}(s,a,s'), \forall h \in [H], s' \in \mathcal{S}\}, \tag{4}$$

$$\beta^p_{h,k}(s,a,s') = \sqrt{\frac{4\text{Var}(\widehat{P}_{h,k}(s'|s,a))L}{n_{h,k}(s,a)\vee 1}} + \frac{14L}{3(n_{h,k}(s,a)\vee 1)}, \tag{5}$$

where $L = \log(\frac{2SAHK}{\delta})$, and $\text{Var}(\widehat{P}_{h,k}(s'|s,a)) = \widehat{P}_{h,k}(s'|s,a)(1 - \widehat{P}_{h,k}(s'|s,a))$. Using the empirical Bernstein inequality, we can show that the true model P is an element of $\mathcal{P}_k$ for any $k \in [K]$ with probability at least $1 - 2\delta$ (see Appendix C).

Similarly, at the beginning of each episode k, DOPE estimates the unknown objective and constraint costs as $\hat{r}_{h,k}(s,a) = \frac{\sum_{k'=1}^{K} r_h(s,a)\mathbb{1}\{s_{h,k'}=s,a_{h,k'}=a\}}{n_{h,k}(s,a)\vee 1}$, $\hat{c}_{h,k}(s,a) = \frac{\sum_{k'=1}^{K} c_h(s,a)\mathbb{1}\{s_{h,k'}=s,a_{h,k'}=a\}}{n_{h,k}(s,a)\vee 1}$. In keeping with OFU, we construct confidence sets $\mathcal{R}_k$ and $\mathcal{C}_k$ around $\hat{r}_k$ and $\hat{c}_k$ respectively, as

$$\begin{aligned}
\mathcal{R}_k &= \{\tilde{r} : |\tilde{r}_h(s,a) - \hat{r}_{h,k}(s,a)| \le \beta^l_{h,k}(s,a), \forall h, s, a \in [H]\times\mathcal{S}\times\mathcal{A}\},\\
\mathcal{C}_k &= \{\tilde{c} : |\tilde{c}_h(s,a) - \hat{c}_{h,k}(s,a)| \le \beta^l_{h,k}(s,a), \forall h, s, a \in [H]\times\mathcal{S}\times\mathcal{A}\},\\
\beta^l_{h,k}(s,a) &= \sqrt{\frac{L'}{n^k_h(s,a)\vee 1}},
\end{aligned} \tag{6}$$

where $L' = 2\log(\frac{6SAHK}{\delta})$. Using Hoeffding inequality, we can show that the true costs belong to $\mathcal{R}_k$ and $\mathcal{C}_k$ for any $k \in [K]$ with probability at least $1-\delta$ (see Appendix C). We define $\mathcal{M}_k = \mathcal{P}_k \cap \mathcal{R}_k \cap \mathcal{C}_k$ to be the total confidence ball around the empirical transitions and costs.

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

$$(\underline{\pi}_k, \underline{P}_k) = \underset{\pi', (P',r',c')\in\mathcal{M}^k}{\arg\min} \; V^{\pi'}_{r'}(P'), \;\; \text{s.t.} \;\; V^{\pi'}_{c'}(P') \le \bar{C}. \tag{7}$$

The OFU problem (7) is feasible since the true model M is an element of $\mathcal{M}_k$ (with high probability). In particular, (π_b, P) and (π^*, 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 $\underline{\pi}_k$ ensures exploration while satisfying the constraint $V^{\underline{\pi}_k}_c(\underline{P}_k) \le \bar{C}$. However, this naive OFU approach overlooks the important issue that $\underline{\pi}_k$ may not be a safe policy with respect to the true model P. More precisely, it is possible to have $V^{\underline{\pi}_k}_c(P) > \bar{C}$ even though $V^{\underline{\pi}_k}_c(\underline{P}_k) \le \bar{C}$. 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 add a pessimistic penalty to the empirical constraint cost to get the pessimistic constraint cost function $\bar{c}_k$ as

$$\bar{c}_{h,k}(s,a) = \hat{c}_{h,k}(s,a) + \beta^l_{h,k}(s,a) + H\bar{\beta}^p_{h,k}(s,a), \tag{8}$$

where $\bar{\beta}^p_{h,k}(s,a) = \sum_{s'\in\mathcal{S}} \beta^p_{h,k}(s,a,s')$. Since $\bar{\beta}^p_{h,k}(s,a)$ is $\tilde{O}(1/\sqrt{n_{h,k}(s,a)})$, (s,a) pairs that are less observed have a higher penalty, disincentivizing their exploration. However. such a pessimistic penalty may prevent the exploration that is necessary to learn the optimal policy. To overcome this issue, we also modify the empirical objective cost function by subtracting a term to incentivize exploration, to obtain an optimistic objective cost function

$$\bar{r}_{h,k}(s,a) = \hat{r}_{h,k}(s,a) - \frac{3H}{\bar{C} - \bar{C}_b}\beta^l_{h,k}(s,a) - \frac{H^2}{\bar{C} - \bar{C}_b}\bar{\beta}^p_{h,k}(s,a). \tag{9}$$

Since $\bar{\beta}^p_{h,k}(s,a)$ is $\tilde{O}(1/\sqrt{n_{h,k}(s,a)})$, (s,a) pairs that are less observed less will have a lowered cost to incentivize their exploration.

We select the policy π_k for episode k by solving the Doubly Optimistic-Pessimistic (DOP) problem:

$$(\pi_k, P_k) = \underset{\pi', P'\in\mathcal{P}_k}{\arg\min} \;\; V^{\pi'}_{\bar{r}_k}(P') \;\; \text{s.t.} \;\; V^{\pi'}_{\bar{c}_k}(P') \le \bar{C}. \tag{10}$$

Notice that DOPE is doubly optimistic by considering both the optimistic objective cost function in (9) and the optimistic model P_k from the confidence set $\mathcal{P}_k$ in (10), while being pessimistic on the constraint in (10). Later, in Lemma 18 in the appendix, we prove that (π_k, P_k) is indeed an optimistic solution. This is in contrast with [26], where the optimism is solely limited to the objective cost. We will show that our approach carefully balances double optimism and pessimism, yielding a regret minimizing learning algorithm with episodic safe exploration guarantees.

We note that the DOP problem (10) may not be feasible, especially in the first few episodes of learning. This is because, $\bar{\beta}^p_{h,k}(s,a)$ and $\beta^l_{h,k}(s,a)$ may be large during the initial phase of learning so that there may not be a policy π' and a model $P' \in \mathcal{P}_k$ that can satisfy the constraint $V^{\pi'}_{\bar{c}_k}(P') \leq \bar{C}$. We overcome this issue by employing a safe baseline policy π_b (as defined in Assumption 1) in the first K_o episodes, a value provided by Proposition 4. Since π_b is safe by definition, DOPE ensures safety during the first K_o episodes. We will later show that the DOP problem (10) will have a feasible solution after the first K_o episodes (see Proposition 4). For any episode $k \geq K_o$, DOPE employs policy π_k, which is the solution of (10). We will also show that π_k from (10) (once it becomes feasible) will indeed be a safe policy (see Proposition 5). We present DOPE formally in Algorithm 1.

Algorithm 1 Doubly Optimistic and Pessimistic Exploration (DOPE)

Input: $\delta \in (0,1)\,, r, c, \pi_b, \bar{C}_b, \bar{C}, K_o$
Initialization: $n_{h,k}(s,a) = n_{h,k}(s,a,s') = 0 \;\; \forall s, s' \in S, a \in A, h \in [H]$.
for episodes $k = 1, \ldots, K$ **do**
  Compute the estimates $\widehat{P}_k$, $\hat{r}_k$, $\hat{c}_k$ and the confidence set $\mathcal{M}_k$ according to (4) - (5), and (6).
  if $k \leq K_o$ **then**
    Select the exploration policy $\pi_k = \pi_b$
  else
    Select the exploration policy π_k according to (10)
  end if
  for $h = 1, 2, \ldots, H$ **do**
    Observe state $s_{h,k}$, select action $a_{h,k} \sim \pi_{h,k}(s_{h,k}, \cdot)$, incur the cost $r_h(s_{h,k}, a_{h,k})$ and $c_h(s_{h,k}, a_{h,k})$, and observe next state $s_{h+1,k} \sim P_h(\cdot|s_{h,k}, a_{h,k})$
    Update the counts: $n_{h,k}(s_{h,k}, a_{h,k}) \leftarrow n_{h,k}(s_{h,k}, a_{h,k}) + 1$, $n_{h,k}(s_{h,k}, a_{h,k}, s_{h+1,k}) \leftarrow n_{h,k}(s_{h,k}, a_{h,k}, s_{h+1,k}) + 1$
  end for
end for

We now present our main result, which shows that the DOPE algorithm achieves $\tilde{O}(\sqrt{K})$ regret *without* violating the safety constraints during learning, with high probability.

Theorem 3. *Fix any $\delta \in (0,1)$. Consider the DOPE algorithm with K_o as specified in Proposition 4. Let $\{\pi_k, k \in [K]\}$ be the sequence of policies generated by the DOPE algorithm. Then, with probability at least $1 - 5\delta$, $\pi_k \in \Pi_{\text{safe}}$ for all $k \in [K]$. Moreover, with probability at least $1 - 5\delta$, the regret of the DOPE algorithm satisfies*

$$R(K) \leq \tilde{\mathcal{O}}\left(\frac{SH^3}{(\bar{C} - \bar{C}_b)}\sqrt{AK}\right).$$

4 Analysis

We now provide the technical analysis of DOPE, concluding with the proof outline of Theorem 3.

4.1 Preliminaries

For an arbitrary policy π' and transition probability function P', define $\epsilon^{\pi'}_k(P')$ and $\eta^{\pi'}_k(P')$ as

$$\epsilon^{\pi'}_k(P') = H\mathbb{E}[\sum_{h=1}^{H} \bar{\beta}^p_{h,k}(s_{h,k}, a_{h,k})|\pi', P', \mathcal{F}_{k-1}], \eta^{\pi'}_k(P') = \mathbb{E}[\sum_{h=1}^{H} \beta^l_{h,k}(s_{h,k}, a_{h,k})|\pi', P', \mathcal{F}_{k-1}]. \tag{11}$$

Then, it is straightforward to show that (see (23) - (24) in the Appendix) $V_{\bar{c}_k}^{\pi'}(P') = V_{\hat{c}_k}^{\pi'}(P') + \eta_k^{\pi'}(P) + \epsilon_k^{\pi'}(P'), V_{\bar{r}_k}^{\pi'}(P') = V_{\hat{r}_k}^{\pi'}(P') - \frac{3H}{\bar{C}-\bar{C}_b}\eta_k^{\pi'}(P') - \frac{H}{\bar{C}-\bar{C}_b}\epsilon_k^{\pi'}(P')$. The analysis utilizes this decomposition of $V_{\bar{c}_k}^{\pi'}(P')$ and $V_{\bar{r}_k}^{\pi'}(P')$, and the properties of $\epsilon_k^{\pi'}(P')$ and $\eta_k^{\pi'}(P')$. Under the good event G defined as in Lemma 12, we can show, $V_{\hat{c}_k}^{\pi'}(P') - \eta_k^{\pi'}(P') \leq V_c^{\pi'}(P') \leq V_{\hat{c}_k}^{\pi'}(P') + \eta_k^{\pi'}(P')$.

4.2 Feasibility of the OP Problem

Even though (π_b, P) is a feasible solution to the original CMDP problem (2), it may not be a feasible for the DOP problem (10) in the initial phase of learning. To see this, note that $V_{\bar{c}_k}^{\pi_b}(P) = V_{\hat{c}_k}^{\pi_b}(P) + \eta_k^{\pi_b}(P) + \epsilon_k^{\pi_b}(P)$, and since $V_c^{\pi_b}(P) \geq V_{\hat{c}_k}^{\pi_b}(P) - \eta_k^{\pi_b}(P)$ under the good event, and $V_c^{\pi_b}(P) = \bar{C}_b$, we will have $V_{\bar{c}_k}^{\pi_b}(P) \leq \bar{C}$ if $2\eta_k^{\pi_b}(P) + \epsilon_k^{\pi_b}(P) \leq (\bar{C} - \bar{C}_b)$. So, (π_b, P) is a feasible solution for (10) if $2\eta_k^{\pi_b}(P) + \epsilon_k^{\pi_b}(P) \leq (\bar{C} - \bar{C}_b)$. This sufficient condition may not be satisfied for initial episodes. However, since $\epsilon_k^{\pi_b}(P)$ and $\eta_k^{\pi_b}(P)$ are decreasing in k, if (π_b, P) becomes a feasible solution for (10) at episode k', then it will remain feasible for all episodes $k \geq k'$. Also, since $\bar{\beta}_{h,k}^p$ and $\beta_{h,k}^l$ decrease 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 4. *Under the DOPE algorithm, with a probability greater that $1 - 5\delta$, (π_b, P) is a feasible solution for the OP problem (10) for all $k \geq K_o$, where $K_o = \tilde{\mathcal{O}}(\frac{S^2AH^4}{(\bar{C}-\bar{C}_b)^2})$.*

4.3 Safety Exploration Guarantee

We show that the DOPE algorithm provides a safe exploration guarantee, i.e., $\pi_k \in \Pi_{\text{safe}}$ for all $k \in [K]$ with high probability, where π_k is the exploration policy employed by DOPE in episode k. This is achieved by the carefully designed pessimistic constraint of the DOP problem (10).

For any $k \leq K_o$, since $\pi_k = \pi_b$, and it is safe by Assumption 1. For $k \geq K_o$, (10) is feasible according to Proposition 4. Since (π_k, P_k) is the solution of (10), we have $V_{\bar{c}_k}^{\pi_k}(P_k) = V_{\hat{c}_k}^{\pi_k}(P^k) + \eta_k^{\pi_k}(P_k) + \epsilon_k^{\pi_k}(P_k) \leq \bar{C}$. This implies that $V_{\hat{c}_k}^{\pi_k}(P_k) + \eta_k^{\pi_k}(P_k) \leq \bar{C} - \epsilon_k^{\pi_k}(P_k)$. We have that $V_c^{\pi_k}(P_k) \leq V_{\hat{c}^k}^{\pi_k}(P_k) + \eta_k^{\pi_k}(P_k)$ under the good event, and hence, the above equation implies that $V_c^{\pi_k}(P_k) \leq \bar{C} - \epsilon_k^{\pi_k}(P_k)$, i.e., π_k satisfies a tighter constraint with respect to the model P_k. However, it is not obvious that the policy π_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 $\epsilon_k^{\pi_k}(P_k)$, i.e., $V_c^{\pi_k}(P) - V_c^{\pi_k}(P_k) \leq \epsilon_k^{\pi_k}(P_k)$. This will then immediately yield that $V_c^{\pi_k}(P) \leq V_c^{\pi_k}(P_k) + \epsilon_k^{\pi_k}(P_k) \leq \bar{C}$, which is the true safety constraint. The key idea is in the design of the pessimistic cost function $\bar{c}_k(\cdot,\cdot)$ 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 DOPE below.

Proposition 5. *Let $\{\pi_k, k \in [K]\}$ be the sequence of policies generated by the DOPE algorithm. Then π_k is safe for all $k \in [K]$, i.e., $V_c^{\pi_k}(P) \leq \bar{C}$, for all $k \in [K]$, with a probability greater than $1 - 5\delta$.*

4.4 Regret Analysis

The regret analysis for most OFU style RL algorithms follows the standard approach of decomposing the regret into two terms as $R_k = V_r^{\pi_k}(P) - V_r^{\pi^*}(P) = (V_r^{\pi_k}(P) - V_r^{\pi_k}(P_k)) + (V_r^{\pi_k}(P_k) - V_r^{\pi^*}(P))$, where R_k denotes the regret in episode k. The first term is the difference between value functions of the selected policy π_k with respect to the true model P and optimistic model P_k. Bounding this term is the key technical analysis part of most of the OFU algorithms for the unconstrained MDP [20, 12] and also the CMDP [17]. In the standard OFU style analysis for the unconstrained problem, since $P \in \mathcal{P}_k$ for all k, it can be easily observed that (π^*, P) is a feasible solution for the OFU problem (7) for all $k \in [K]$. Moreover, since (π_k, P_k) is the optimal solution in k^{th} episode, we get $V_r^{\pi_k}(P_k) \leq V_r^{\pi^*}(P)$. So, the second term will be non-positive, and hence can be dropped from the regret analysis. However, in our setting, the second term can be positive since (π^*, P) may not be a feasible solution of the DOP problem (10) due to the pessimistic constraint. This

necessitates a different approach for bounding the regret. Existing work [26] only considers optimism in the objective cost, and hence their proof closely follows that of OptCMDP-Bonus algorithm in [17] with pessimistic constraints. In analyzing the regret of DOPE, we need to handle the optimism in objective cost as well as the model in regret terms, along with the pessimistic constraints. This make the analysis particularly challenging. The full proof is detailed in the appendix.

5 Experiments

We now evaluate DOPE via experiments. We have two relevant metrics, namely, (i) objective regret, defined in (3) that measures the optimality gap of the algorithm, and (ii) constraint regret, defined as $\sum_{k=1}^{K} \max\{0, V_c^{\pi_k}(P) - \bar{C}\}$, where π_k is the output of the algorithm in question at episode k. This measures the safety gap of the algorithm. Our candidate algorithms are (i) OptCMDP, (ii) AlwaysSafe, (iii) OptPessLP and (iv) DOPE, all described in the introduction. OptCMDP is expected to show constraint violations, while the other three should show zero constraint regret. We consider two environments here, with a third environment presented in the appendix. AlwaysSafe can directly be used only with a factored CMDP, and only applies to the first environment presented. We use the π_T variant of this algorithm [34].

Factored CMDP: We first consider a CMDP where the safety relavant features of the model can be separated, as shown in [34]. This CMDP contains states $\{1, 2, 3\}$ arranged in a circle, and allows 2 actions $\{1, 2\}$ in each state, to move right or stay put, respectively. The transitions move the agent to its right state with probability 1, if action 1 is taken. If action 2 is taken, it remains in the same state with probability 1. Action 1 does not incur any objective cost or constraint cost. Taking action 2 gives a objective cost equals to the state number, and a constraint cost of 1. We choose episode length $H = 6$, and constraint as $\bar{C} = 3$. The structure of this CMDP is such that AlwaysSafe is able to extract a safe baseline policy from it.

Media Streaming CMDP: Our second 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 μ_1 corresponds to the fast service. Packets received at the device are stored in a media buffer and played out according to a Bernoulli process with parameter γ. 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

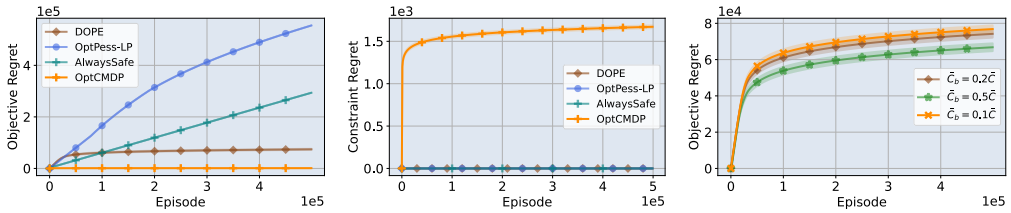
$$s_{h+1} = \min\{\max(0, s_h + A_h - B_h), N\},$$

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 $r(s, a) = \mathbb{1}\{s = 0\}$, while the constraint cost is $c(s, a) = \mathbb{1}\{a = 1\}$, i.e., we desire to minimize the outage cost, while limiting the usage of fast service. We consider a horizon $H = 10$, and $\bar{C} = \frac{H}{2}$.

Experiment Setup: OptCMDP and OptPessLP algorithms have not been implemented earlier. For accurate comparison, we simulate all the algorithms true to their original formulations of cost functions and confidence intervals. Our experiments are carried out for 20 random runs, and averaged to obtain the regret plots. For DOPE, we choose K_0 to be as specified in Proposition 4. Full details on the algorithm parameters and experiment settings are provided in Appendix E.

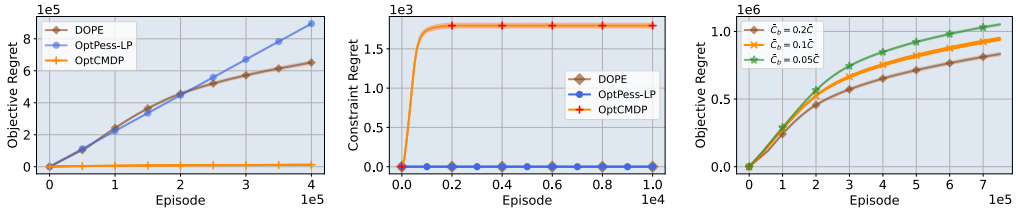
Baseline Policies: Both OptPessLP and DOPE require baseline policies. We select the baseline policies as the optimal solutions of the given CMDP with a constraint $\bar{C}_b = 0.2\bar{C}$. This choice is to showcase the efficacy of DOPE, despite starting with a conservative baseline policy.

Results for Factored CMDP: Fig. 1(a) shows the objective regret of the algorithms in this environment. OptCMDP has a good objective regret performance as expected, but shows constraint violations. AlwasySafe fails to achieve $\tilde{O}(\sqrt{K})$ regret in the episodes shown. OptPessLP takes a long time to attain $\sqrt{K}$ behavior, which means that is chooses the baseline policy for an extended period, and shows high empirical regret. This suggests that reward optimism of OptPessLP is insufficient to balance the pessimism in the constraint. DOPE not only achieves the desired $\tilde{O}(\sqrt{K})$ regret, but also does so in fewer episodes compared to the other two algorithms, despite using the same baseline policy as OptPessLP. Furthermore, it has low empirical regret. Fig. 1(b) shows that the constraint violation regret is zero for all the episodes of learning for all the safe algorithms, while OptCMDP shows a large constraint violation regret.



(a) Objective Regret (b) Constraint Regret (c) DOPE Objective Regret, varying $\bar{C}_b$

Figure 1: Illustrating the Objective Regret and Constraint Regret for a Factored CMDP environment.



(a) Objective Regret (b) Constraint Regret (c) DOPE Objective Regret, varying $\bar{C}_b$

Figure 2: Illustrating the Objective Regret and Constraint Regret for the Media Streaming Environment.

Results for Media Streaming CMDP: Fig 2(a) compares the objective regret across algorithms. The value of DOPE over OptPessLP is more apparent here, since, after a linear growth phase, the objective regret of DOPE changes to a square-root scaling. OptPessLP has not explored sufficiently at this point, and hence suffers high linear regret. Finally, OptCMDP has the same square-root regret scaling, but is fastest, since it is not constrained by safe exploration. Fig 2(b) compares the regret in constraint violation for these algorithms. As expected, DOPE and OptPessLP do not violate the constraint, while the OptCMDP algorithm significantly violates the constraints during learning.

Effect of Baseline Policy: We compare the objective regret of DOPE under different baseline policies in Fig 1(c) and Fig 2(c). We see that less conservative (but safe) baselines result in lower regret, but the difference is not excessive, implying that the exact baseline policy chosen is not crucial.

Summary: DOPE has two valuable properties: (i) Faster rate of shift to $\sqrt{K}$ behavior, since the linear regret phase where it applies the baseline policy is relatively short, and (ii) $\tilde{O}(\sqrt{K})$ regret with respect to optimal, which together mean that the empirical regret is lower than other approaches.

Limitations: Our goal is to find an RL algorithm with no constraint violation with high probability in the tabular setting. Since safe exploration is a basic problem, we follow the usual approach in the literature of first establishing the fundamental theory results for the tabular setting [20, 13, 6]. We also note that most of the existing work on exploration in safe RL is in the tabular setting [17, 35, 26]. In our future work, we plan to employ the DOPE approach in a function approximation setting.

6 Conclusion

We considered the safe exploration problem in reinforcement learning, wherein a safety constraint must be satisfied during learning and evaluation. Earlier approaches to constrained RL have proposed optimism on the model, optimism on reward, and pessimism on constraints as means of modulating exploration, but none have shown order optimal regret, no safety violation, and good empirical performance simultaneously. We started with the conjecture that double optimism combined with pessimism is the key to attaining the ideal balance for fast and safe exploration, and design DOPE that carefully combines these elements. We showed that DOPE not only attains order-optimal $\tilde{O}(\sqrt{K})$ regret without violating safety constraints, but also reduces the best known regret bound by a factor of $\sqrt{|\mathcal{S}|}$. Furthermore, it has significantly better empirical performance than existing approaches. We thus make a case for adoption of the approach for real world use cases and extension to large scale RL problems using function approximation.

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained Policy Optimization. In *International Conference on Machine Learning*, pages 22–31. PMLR, 2017.

[2] Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

[3] Eitan Altman. Applications of Markov decision processes in communication networks. In *Handbook of Markov decision processes*, pages 489–536. Springer, 2002.

[4] Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems*, 32:9256–9266, 2019.

[5] Sanae Amani and Lin F. Yang. Doubly pessimistic algorithms for strictly safe off-policy optimization. In *2022 56th Annual Conference on Information Sciences and Systems (CISS)*, pages 113–118, 2022.

[6] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

[7] Dimitri P Bertsekas et al. *Dynamic programming and optimal control: Vol. 1*. Athena scientific Belmont, 2000.

[8] Shalabh 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.

[9] Vivek S Borkar. An actor-critic algorithm for constrained Markov decision processes. *Systems & control letters*, 54(3):207–213, 2005.

[10] Yin-Lam Chow, Marco Pavone, Brian M Sadler, and Stefano 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.

[11] Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

[12] Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, 28:2818–2826, 2015.

[13] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *arXiv preprint arXiv:1703.07710*, 2017.

[14] Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3304–3312. PMLR, 2021.

[15] Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo R Jovanovic. Natural policy gradient primal-dual method for constrained Markov decision processes. In *NeurIPS*, 2020.

[16] Xu Chu Ding, Alessandro Pinto, and Amit Surana. Strategic planning under uncertainties via constrained Markov decision processes. In *2013 IEEE International Conference on Robotics and Automation*, pages 4568–4575. IEEE, 2013.

[17] Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained MDPs. *arXiv preprint arXiv:2003.02189*, 2020.

[18] Aria HasanzadeZonuzy, Archana Bura, Dileep Kalathil, and Srinivas 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.

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

[20] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

[21] Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial Markov decision processes with bandit feedback and unknown transition. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 4860–4869. PMLR, 13–18 Jul 2020.

[22] Krishna C Kalagarla, Rahul Jain, and Pierluigi 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.

[23] Kia Khezeli and Eilyan Bitar. Safe linear stochastic bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10202–10209, 2020.

[24] Hepeng Li, Zhiqiang Wan, and Haibo He. Constrained EV charging scheduling based on safe deep reinforcement learning. *IEEE Transactions on Smart Grid*, 11(3):2427–2439, 2019.

[25] Nikolaos Liakopoulos, Apostolos Destounis, Georgios Paschos, Thrasyvoulos Spyropoulos, and Panayotis Mertikopoulos. Cautious regret minimization: Online optimization with long-term budget constraints. In *International Conference on Machine Learning*, pages 3944–3952. PMLR, 2019.

[26] Tao Liu, Ruida Zhou, Dileep Kalathil, PR Kumar, and Chao Tian. Learning Policies with Zero or Bounded Constraint Violation for Constrained MDPs. In *Thirty-fifth Conference on Neural Information Processing Systems*, 2021.

[27] Xin Liu, Bin Li, Pengyi Shi, and Lei Ying. An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. *arXiv preprint arXiv:2102.05295*, 2021.

[28] Andreas Maurer and Massimiliano 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.

[29] Ahmadreza Moradipari, Christos Thrampoulidis, and Mahnoosh Alizadeh. Stage-wise conservative linear bandits. *Advances in Neural Information Processing Systems*, 33, 2020.

[30] Michael J Neely and Hao Yu. Online convex optimization with time-varying constraints. *arXiv preprint arXiv:1702.04783*, 2017.

[31] Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 2827–2835. PMLR, 2021.

[32] Advait Parulekar, Soumya Basu, Aditya Gopalan, Karthikeyan Shanmugam, and Sanjay Shakkottai. Stochastic linear bandits with protected subspace. *arXiv preprint arXiv:2011.01016*, 2020.

[33] Santiago Paternain, Luiz Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. *Advances in Neural Information Processing Systems*, 32:7555–7565, 2019.

[34] Thiago D Simão, Nils Jansen, and Matthijs TJ Spaan. AlwaysSafe: Reinforcement learning without safety constraint violations during training. 2021.

[35] Rahul Singh, Abhishek Gupta, and Ness B Shroff. Learning in Markov decision processes under constraints. *arXiv preprint arXiv:2002.12435*, 2020.

[36] Rahul Singh and PR 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.

[37] Rahul Singh, PR Kumar, and Le Xie. Decentralized control via dynamic stochastic prices: The independent system operator problem. *IEEE Transactions on Automatic Control*, 63(10):3206–3220, 2018.

[38] Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2018.

[39] Ilnura Usmanova, Andreas Krause, and Maryam Kamgarpour. Safe convex learning under uncertain constraints. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2106–2114. PMLR, 2019.

[40] Honghao Wei, Xin Liu, and Lei Ying. Triple-q: A model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation. In *International Conference on Artificial Intelligence and Statistics*, pages 3274–3307. PMLR, 2022.

[41] Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J Ramadge. Projection-based constrained policy optimization. In *International Conference on Learning Representations*, 2019.

[42] Hao Yu, Michael J Neely, and Xiaohan Wei. Online convex optimization with stochastic constraints. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1427–1437, 2017.

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

[44] Yiming Zhang, Quan Vuong, and Keith Ross. First order constrained optimization in policy space. *Advances in Neural Information Processing Systems*, 33, 2020.

[45] Liyuan Zheng and Lillian Ratliff. Constrained upper confidence reinforcement learning. In *Learning for Dynamics and Control*, pages 620–629. PMLR, 2020.

[46] Alexander Zimin and Gergely Neu. Online learning in episodic Markovian decision processes by relative entropy policy search. In *Neural Information Processing Systems 26*, 2013.

Societal Impact and Ethics Statement

Reinforcement learning has much potential for application to a variety of cyber-physical systems, such as the power grid, robotics and other systems where guarantees on the operating region of the system must be met. Our work provides a theoretical basis for the design of controllers that can be applied in such scenarios. The approaches presented in the paper were tested on simulated environments, and did not involve any human interaction. We do not see any ethical concerns with our research approach.

A note of caution with our approach is that the policy generated is only as good as the training environment, and many examples exist wherein the policy generated is optimal according to its training, but violates basic truths known to human operators and could fail quite badly. Indeed, our approach does not provide sample-path guarantees, and the system could well move into deleterious states for a small fraction of the time, which might be completely unacceptable and trigger hard fail safes, such as breakers in a power system. Understanding the right application environments with excellent domain knowledge is hence needed before any practical success can be claimed.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] In Theorem 3 and Section 5
   (b) Did you describe the limitations of your work? [Yes] In Section 5
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Societal Impact and Ethics Statement.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] Described in Societal Impact and Ethics Statement.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Assumption 1, and assumptions in statement of Theorem 3
   (b) Did you include complete proofs of all theoretical results? [Yes] Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results? [Yes] Supplemental material
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Section 5, and Appendix E
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Figures contain error bars.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Appendix E
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Code is released as supplemtental material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects... [N/A]