Model-Based Reinforcement Learning for Infinite-Horizon Discounted Constrained Markov Decision Processes

Aria HasanzadeZonuzy[1] , **Dileep Kalathil**[1] , **Srinivas Shakkottai**[1]
[1]Texas A & M University
{azonuzy, dileep.kalathil, sshakkot@tamu.edu}@tamu.edu,

Abstract

In many real-world reinforcement learning (RL) problems, in addition to maximizing the objective, the learning agent has to maintain some necessary safety constraints. We formulate the problem of learning a safe policy as an infinite-horizon discounted Constrained Markov Decision Process (CMDP) with an unknown transition probability matrix, where the safety requirements are modeled as constraints on expected cumulative costs. We propose two model-based constrained reinforcement learning (CRL) algorithms for learning a safe policy, namely, (i) GM-CRL algorithm, where the algorithm has access to a generative model, and (ii) UC-CRL algorithm, where the algorithm learns the model using an upper confidence style online exploration method. We characterize the sample complexity of these algorithms, i.e., the the number of samples needed to ensure a desired level of accuracy with high probability, both with respect to objective maximization and constraint satisfaction.

1 Introduction

Markov Decision Processes (MDPs) are a powerful approach to model stochastic systems where stationary control policies are appropriate. Many cyber-physical (i.e. physical systems controlled algorithmically) systems bear intrinsic limitations on the nature of control that may be applied. Hence, Constrained-MDP (CMDP) are an appropriate framework for the modeling and analysis of such systems [Altman, 1999].

In this paper, we aim to develop simple algorithms to learn near-optimal policies for a CMDP without knowing the system parameters. Although, a regular model-based RL algorithm attempts to collect as few samples as possible to quickly solve for the optimal policy, minimizing the number of samples taken is even more essential in the CMDP setting. This requirement is due to the existence of constraints in the CMDP setting, and it might be important to violate them as few times as possible while maximizing the objective of the system. Therefore, the behavior of a system with respect to (w.r.t) both objective maximization and safety violation over time is a crucial performance metric for a proposed RL algorithm for CMDPs.

Main Contributions: Our goal is to upper bound the number of samples required to learn a near-optimal policy while nearly satisfying the constraints with high probability (w.h.p.) in the context of the discounted infinite-horizon setting. Our contributions are mainly threefold:

(i) We design and analyze two model-based RL algorithms for CMDPs. One of them pursues a generative model based approach that obtains samples initially and creates a model. The other one is based on an online approach in which the model is updated over time-steps. With both algorithms, the estimated model might lead to infeasible situation. Thus, we utilize the idea of a confidence-ball around the estimated model such that the true model would belong to that ball w.h.p. This ensures that a solution may be found w.h.p. under the assumption that the real model has a solution.

(ii) Both algorithms follow a two-stage pattern of model construction and a CMDP solution. The algorithms use linear programming (LP) to solve the CMDP problem with additional linear constraints to incorporate the confidence-ball.

(iii) We characterize PAC-type sample complexity bounds for both algorithms, accounting for both objective maximization and constraint satisfaction. Intuitively, the model constructed by these algorithms must be more accurate than models created by unconstrained counterparts, which conjecture our main results are consistent with. Furthermore, a comparison of our main findings with lower bounds on sample complexity of MDPs [Azar *et al.*, 2013; Dann and Brunskill, 2015] shows an increase in our results by a logarithmic factor in the number of constraints and the size of the state space. However, there is no earlier work on lower bound of sample complexity of learning CMDPs to our best knowledge.

As mentioned above, cyber-physical systems might have a large number of constraints. However, our results indicate that the number of constraints should not be a major concern in implementation, since our bounds scale logarithmically with number of constraints. Hence, the results suggest that the constrained RL approach is likely applicable in a straightforward manner to cyber-physical systems.

Related Work: There are many articles studying the problem of controlling CMDPs with an algorithmic approach and control-theoretic view [Altman, 1999; Altman, 2002; Borkar, 2005; Borkar and Jain, 2014; Singh and Kumar, 2018; Singh *et al.*, 2014]. The results take the form of proving asymptotic convergence of their proposed methods under the assumption of the known model. There are also extensions of this approach to the context of an unknown model, where the focus is still on asymptotic behavior [Bhatnagar and Lakshmanan, 2012; Chow *et al.*, 2018; Tessler *et al.*, 2018; Paternain *et al.*, 2019]. These studies use Lagrangian method to show zero duality gap asymptotically. Further, [Liu *et al.*, 2019] also develops an algorithm based on the Lagrangian method, but with small eventual duality gap. Finally, empirical studies based on the Lagrangian method have also been presented [Liang *et al.*, 2018].

There are also studies on the constrained bandit case. Although bandits are not MDPs per se, they are strongly related to them. Articles such as [Badanidiyuru *et al.*, 2013; Wu *et al.*, 2015; Amani *et al.*, 2019] consider such constraints, either in a knapsack sense, or on the type of controls that may be applied in a linear bandit context.

More related to our work theme are parallel studies on CMDPs. For example, [Zheng and Ratliff, 2020] and [Wachi and Sui, 2020] provide results with the assumption of unknown reward functions, with either a known or deterministic transition kernel. There are other works [Satija *et al.*, 2020] focusing on proving asymptotic convergence without providing a bound on learning rate. Finally, closest related work to this article is [Efroni *et al.*, 2020] which explores algorithms similar to ours in finite-horizon setting, but concentrating on characterizing objective and constrained regret bounds. Now, regret and sample complexity bounds are not directly translatable [Dann *et al.*, 2017], and converting their regret bounds to our setting gives relatively weak sample complexity bounds. Specifically, our main results with logarithmic increase in sample complexity with the number of constraints differentiates our work.

2 Notation and Problem Formulation

Notation and Setup: Our focus is on an infinite-horizon CMDP defined by a tuple $M = \langle S, A, P, r, c, \bar{C}, s_0, \gamma\rangle$. S and A represent the sets of states and actions respectively. Additionally, $P(s'|s,a)$ is used to indicate the probability of reaching state s' by taking action a while being at state s. We define $r(s,a)$ as the reward for each state-action pair (s,a). We assume that there are N constraints. We use c to denote the cost matrix, where $c(i,s,a)$ is the immediate cost incurred by the i^{th} constraint in (s,a) where $i \in \{1,\ldots,N\}$. Further, the value of the constraints (i.e. the bound that must be satisfied) are determined by the vector $\bar{C}$. Also, initial state is specified by s_0. Finally, we use γ for discount factor. In this study, the discount factor is unique for both objective function and constraint functions where they shall be defined later.

Assumption 1. *State and action sets S and A are assumed to be finite with cardinalities $|S|$ and $|A|$. In addition, the immediate cost and immediate reward $r(s,a)$ are assumed to be taken from the interval* $[0,1]$. *Number of constraints is also assumed to be N which for each $i \in \{1,\ldots,N\}, \bar{C}_i \in$* $[0, \bar{C}_{\max}]$.

Now, we define a stationary policy $\pi : S \times A \rightarrow [0,1]^{|A|}$ as a mapping from state-action space $S \times A$ to set of probability vectors defined over action space in order to choose an action at any time-step t. Henceforth, $\pi(s,a)$ represents the probability of choosing the action a when the system is at state s. Also, $a \sim \pi(s,\cdot)$ means that action a is chosen according to stationary policy π while being at state s.

Fixing a policy π transforms the underlying MDP to a Markov chain. The transition kernel of this Markov chain is P_π, which can be viewed as an operator. The operator $P_\pi f(s) = \mathbb{E}[f(s_{t+1})|s_t = s] = \sum_{s'\in S} P_\pi(s'|s) f(s')$ takes any function $f : S \rightarrow \mathbb{R}$ and returns the expected value of f in the next time-step. For convenience, we define the multistep version $P_\pi^t f(s) = P_\pi P_\pi \ldots P_\pi f$, which is repeated t times. Further, we define P_π^0 as the identity operator.

For the objective and constraint functions, we consider discounted infinite-horizon criteria with identical discount factor γ. We define the value function of state s under policy π as

$$V^\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t,a_t); a_t \sim \pi(s_t,\cdot), s_{t=0} = s_0], \quad (1)$$

where expectation $\mathbb{E}[\cdot]$ is taken w.r.t transition kernel P. Next, the local variance of the value function at time step t under policy π is

$$\begin{aligned}\sigma^2_{V_\pi}(s) &= \gamma^2 \mathbb{E}[(V^\pi(s_{t+1}) - P_\pi V^\pi(s))^2] \quad (2)\\ &= \gamma^2 P_\pi[(V^\pi - P_\pi V^\pi)^2](s).\end{aligned}$$

Analogous to the definition of the value function (1), the i^{th} constraint function under policy π is defined as

$$C_i^\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t c(i,s_t,a_t); a_t \sim \pi(s_t,\cdot), s_{t=0} = s_0]. \quad (3)$$

Again, the local variance of i^{th} constraint function under policy π, i.e. $\sigma^2_{C_i^\pi}$ is defined similar to local variance of value function (2).

Eventually, the general infinite-horizon CMDP problem is

$$\max_\pi V^\pi(s_0) \text{ s.t. } C_i^\pi(s_0) \le \bar{C}_i, \quad \forall i \in \{1,\ldots,N\}. \quad (4)$$

Assumption 2. *We assume that the CMDP problem of* (4) *is feasible with optimal policy π^* and optimal solution $V^*(s_0) = V^{\pi^*}(s_0)$.*

Note that we only consider learning feasible CMDPs by this assumption.

Constrained-RL Problem: The Constrained RL problem formulation is identical to the CMDP optimization problem of (4) with one difference. Here, we are not aware of the values of the transition kernel P.[1] We desire to provide model-based algorithms and determine the sample complexity results in a PAC sense, which is defined as follows:

Definition 1. *For an algorithm $\mathcal{A}$, sample complexity is the number of samples that $\mathcal{A}$ requires to achieve*

$$\mathbb{P}\Big(V^{\mathcal{A}}(s_0) \ge V^{\pi^*}(s_0) - \epsilon \text{ and } C_i^{\mathcal{A}}(s_0) \le \bar{C}_i + \epsilon\ \forall i \in \{1,\ldots,N\}\Big) \ge 1 - \delta$$

for a given ϵ and δ.

[1] We only assume that transition kernel is unknown and the extension to unknown reward and cost matrices is straightforward, and does not require additional methodology.

Note that with this definition, we include both objective maximization and constraint violations as opposed to the traditional definition that only considers the objective [Strehl and Littman, 2008].

3 Sample Complexity Result of Generative Model Based Learning

Generative model based learning is a well known approach to learn an optimal policy for an MDP. However, naive application of this approach to CMDPs may not end with a feasible solution. Hence, we explore the generative model based approach for CMDPs, and propose a generative model based CMDP learning algorithm called Generative Model-Constrained RL (GM-CRL). According to GM-CRL, each state-action pair is sampled n number of times uniformly across all state-action pairs, the number of times each transition occurs $n(s', s, a)$ for each next state s' is counted, and an empirical model of transition kernel denoted by $\widehat{P}(s'|s,a) = \frac{n(s',s,a)}{n}\ \forall(s', s, a)$ is constructed.

Unlike MDP problem formulation, there is no guarantee such that CMDP problem formulation w.r.t. $\widehat{P}$ is feasible. In order to resolve the feasibility concern, we expand the space of transition kernels to include the true transition kernel P, noting that the CMDP problem w.r.t. P is feasible from Assumption 2. The algorithmic layout of this approach is as follows. GM-CRL creates a class of CMDPs using the empirical model. This class is denoted by $\mathcal{M}_{\delta_P}$ and contains CMDPs with identical reward, cost matrices, $\bar{C}$, initial state s_0 and discount factor of the true CMDP, but with transition kernels close to true model. This class of CMDPs is defined as

$$\begin{aligned}&\mathcal{M}_{\delta_P} := &(5)\\ &\{M' : r'(s,a) = r(s,a), c'(i,s,a) = c(i,s,a), \gamma' = \gamma, &\\ &|P'(s'|s,a) - \widehat{P}(s'|s,a)| \le &\\ &\min\Big(\sqrt{\frac{2\widehat{P}(s'|s,a)(1-\widehat{P}(s'|s,a))}{n}\log\frac{4}{\delta_P}} + \frac{2}{3n}\log\frac{4}{\delta_P}, &\\ & &(6)\\ &\sqrt{\frac{\log 4/\delta_P}{2n}}\Big)\forall s, a, s', i\},&\end{aligned}$$

where δ_P is defined in Algorithm 1. Note that for any $M' \in \mathcal{M}$, objective function $V'^{\pi}(s_0)$ and cost functions $C'^{\pi}_i(s_0)$ are computed w.r.t. the corresponding transition kernel P' according to equations (1) and (3) respectively.

At the end, GM-CRL maximizes the objective function among all possible transition kernels, while satisfying constraints (if feasible). More specifically, it solves the optimistic planning problem below

$$\max_{\pi, M' \in \mathcal{M}_{\delta_P}} V'^{\pi}(s_0) \quad \text{s.t.} \quad C'^{\pi}_i(s_0) \le \bar{C}_i \ \ \forall i. \tag{7}$$

To solve the problem of (7), GM-CRL uses Extended Linear Programming, or **ELP**. This method takes $\mathcal{M}_{\delta_P}$ as input and gives $\tilde{\pi}$ for the optimal solution. The description of ELP is provided in supplementary materials. Algorithm 1 describes GM-CRL.

Algorithm 1 GM-CRL

```
Input: accuracy $\epsilon$ and failure tolerance $\delta$.
Set $\delta_P = \frac{\delta}{5(N+2)|S|^3|A|}$.
Set $n(s', s, a) = 0\ \forall(s, a, s')$.
for each $(s, a) \in S \times A$ do
    Sample $(s, a), n = \frac{1152(\log 2)^2\gamma^2}{\epsilon^2(1-\gamma)^3}|S|^2|A| \log\frac{4}{\delta_P}$ and update $n(s', s, a)$.
    $\widehat{P}(s'|s, a) = \frac{n(s',s,a)}{n}\ \forall s'$.
end for
Construct $\mathcal{M}_{\delta_P}$ according to (5).
Output $\tilde{\pi} = \text{ELP}(\mathcal{M}_{\delta_P})$.
```

3.1 PAC Analysis of GM-CRL

Here, we present the sample complexity result of GM-CRL.

Theorem 1. *Consider any infinite-horizon CMDP $M = \langle S, A, P, r, c, \bar{C}, s_0, \gamma\rangle$ satisfying assumptions 1 and 2, and CMDP problem formulation of* (4). *Then, for any $\epsilon \in (0, \frac{0.22\gamma}{\sqrt{|S|(1-\gamma)}})$ and $\delta \in (0, 1)$, algorithm 1 creates a model CMDP $\tilde{M} = \langle S, A, \tilde{P}, r, c, \bar{C}, s_0, \gamma\rangle$ and outputs policy $\tilde{\pi}$ such that*

$$\begin{aligned}&\mathbb{P}(V^{\tilde{\pi}}(s_0) \ge V^{\pi^*}(s_0) - \epsilon \ \textit{and}\\ &C_i^{\tilde{\pi}}(s_0) \le \bar{C}_i + \epsilon \ \ \forall i \in \{1, 2, \ldots, N\}) \ge 1 - \delta,\end{aligned}$$

with at least total sampling budget of

$$\frac{1152(\log 2)^2\gamma^2}{\epsilon^2(1-\gamma)^3}|S|^2|A| \log\frac{20(N+2)|S|^3|A|}{\delta}.$$

The proof of Theorem 1 is different from the traditional analysis framework of unconstrained RL [Azar *et al.*, 2013] in the following manner. First, consider the role played by optimism in model construction. The notion of optimism is not required for learning unconstrained MDPs with generative models, because any estimated model is always feasible [Puterman, 2014]. However, there is no such guarantee for a general CMDP problem formulation [Altman, 1999]. Specifically, simply substituting the true kernel P by the estimated one $\widehat{P}$ is not appropriate, since there is no assurance of feasibility of that problem. Hence, GM-CRL converts the CMDP problem under the estimated transition kernel to an optimistic planning problem (7) and an ELP-based solution.

Second, the core of the analysis of every unconstrained MDP is based on being able to characterize the optimal policy via the Bellman operator. This technique enables one to obtain a sample complexity that scales with the size of the state space as $O(|S|)$. However, we cannot use this approach to characterize the optimal policy in a CMDP [Altman, 1999]. We require a uniform PAC result over set of all policies and set of value and constraint functions, which in turn leads to quadratic sample complexity in the size of state space; i.e., a scaling of $O(|S|^2)$.

Corollary 1. *In case of $N = 0$, the problem would become regular unconstrained MDP. And, the sample complexity result with $N = 0$ would also hold for unconstrained case.*

Now, we present some of the lemmas that are essential to prove Theorem 1. Then we sketch the proof of this theorem. The detailed proofs are provided in supplementary materials.

First, we show that true CMDP lies inside the $\mathcal{M}_{\delta_P}$ with high probability, w.h.p. Hence, the problem (7) is feasible w.h.p., since the original CMDP problem is assumed to be feasible according to Assumption 2.

Lemma 1.

$$\mathbb{P}(M \in \mathcal{M}_{\delta_P}) \geq 1 - |S|^2|A|\delta_P.$$

Proof Sketch: Fix a state-action pair (s, a) and next state s'. Then, according to combination of Hoefding's inequality [Hoeffding, 1994] and empirical Bernstein's inequality [Maurer and Pontil, 2009], we obtain that each $P(s'|s, a)$ is inside the confidence set defined by (6) with probability at least $1 - \delta_P$. Applying the union bound yields the result. □

Now, we present the core lemma required for proving Theorem 1 and its proof sketch. Using this lemma, we bound the mismatch in objective and constraint functions when we have n number of samples from each (s, a). This bound applies uniformly over the set of policies and set of value and constraint functions. The result also enables us to bound the objective and constraint functions individually. Then we apply the union bound on all objective and constraint functions. This process is the reason why the number of constraints appear logarithmically in the sample complexity result.

Lemma 2. *Let* $\delta_P \in (0, 1)$. *Then, if* $n \geq 11819\frac{|S|^2 \log 4/\delta_P}{(1-\gamma)^2}$, *under any policy* π

$$\|V^\pi - \tilde{V}^\pi\|_\infty \leq 3\gamma \log 2\sqrt{\frac{32|S| \log 4/\delta_P}{(1-\gamma)^3 n}}$$

w.p. at least $1 - 5|S|^3|A|\delta_P$, *and for any* $i \in \{1, \ldots, N\}$,

$$\|C_i^\pi - \tilde{C}_i^\pi\|_\infty \leq 3\gamma \log 2\sqrt{\frac{32|S| \log 4/\delta_P}{(1-\gamma)^3 n}}$$

w.p. at least $1 - 5|S|^3|A|\delta_P$.

Proof Sketch: We first show that $|\tilde{P}(s'|s, a) - P(s'|s, a)| \leq O(\sqrt{\frac{P(s'|s,a)(1-P(s'|s,a))}{n}})$ for each s', s, a. Then, we show that $(P_\pi - \tilde{P}_\pi)V^\pi(s) \leq O(\sqrt{\frac{|S|}{n}}\sigma_{V^\pi}(s))$. Applying this bound to $|\tilde{V}^\pi(s_0) - V^\pi(s_0)|$ and from the fact that $\sigma_{V^\pi}(s)$ is close to $\tilde{\sigma}_{V^\pi}(s)$ by $O(\frac{\sqrt{|S|}}{(1-\gamma)n^{1/4}})$, we obtain the result. This procedure is also applicable to each constraint function i. □

***Proof Sketch of Theorem 1*:** From Lemma 1, we know that the optimistic planning problem (7) is feasible w.h.p. Hence, we can obtain an optimistic policy $\tilde{\pi}$. The rest of this proof consists of two major parts.

First, we prove $\epsilon-$optimality of objective function w.h.p. Considering policy π^* we obtain $|V^{\pi^*}(s_0) - \tilde{V}^{\pi^*}(s_0)| \leq O(\sqrt{\frac{|S|}{(1-\gamma)^3 n}})$ w.h.p. by means of Lemma 2. Similarly, $|V^{\tilde{\pi}}(s_0) - \tilde{V}^{\tilde{\pi}}(s_0)| \leq O(\sqrt{\frac{|S|}{(1-\gamma)^3 n}})$ w.h.p. Next, we use the fact that $\tilde{V}^{\pi^*}(s_0) \leq \tilde{V}^{\tilde{\pi}}(s_0)$ and obtain

$$V^{\tilde{\pi}}(s_0) \geq V^{\pi^*}(s_0) - O(\sqrt{\frac{|S|}{(1-\gamma)^3 n}}).$$

Next, we show that each constraint is violated at most by ϵ w.h.p. Here, we use the second part of Lemma 2 to bound constraint violation. Thus, for each $i \in \{1, \ldots, N\}$ we have $|C_i^{\tilde{\pi}}(s_0) - \tilde{C}_i^{\tilde{\pi}}(s_0)| \leq O(\sqrt{\frac{|S|}{(1-\gamma)^3 n}})$ w.h.p. Also, we know that $\tilde{C}_i^{\tilde{\pi}}(s_0) \leq \bar{C}_i$, since $\tilde{\pi}$ is solution of the ELP. Hence, we obtain

$$C_i^{\tilde{\pi}}(s_0) \leq \bar{C}_i + O(\sqrt{\frac{|S|}{(1-\gamma)^3 n}})$$

w.h.p. Finally, we obtain the end result by applying the union bound, and obtaining n by solving $\epsilon = O(\sqrt{\frac{|S|}{(1-\gamma)^3 n}})$. □

4 Sample Complexity Result of Online Learning

The GM-CRL approach operates in a way that every state-action pair in the system is sampled a certain number of times before a policy is computed. However, there are applications that are not capable of utilizing this approach, since it may not be possible to reach those states without the employment of some policy, or they might be unsafe, and so should not be sampled often. Hence, we have to find an approach that can collect samples from the environment by means of an online algorithm.

Upper Confidence Constrained-RL, or UC-CRL described in Algorithm 2, is an online method proceeding over time-steps. At each time-step t, UC-CRL constructs an empirical model $\widehat{P}$ using state-action visitation frequencies, i.e., $\widehat{P}(s'|s, a) = \frac{n(s',s,a)}{n(s,a)}$, where $n(s', s, a)$ and $n(s, a)$ are visitation frequencies. Then, we use $\widehat{P}$ to create a confidence interval around each element $\widehat{P}(s'|s, a)$ using same concentration inequalities of GM-CRL defined by (6). Next, UC-CRL constructs set of infinite-horizon CMDPs $\mathcal{M}_t$ which any CMDP $M' \in \mathcal{M}_t$ has identical discount factor and reward and cost matrices to the true CMDP M, but different transition kernels from the concentration inequalities. $\mathcal{M}_t$ is identical to $\mathcal{M}_{\delta_P}$ except for the use of $n(s, a)$ instead of n. Thus the class of CMPDs is defined as below at each time-step t :

$$\begin{aligned}
&\mathcal{M}_t := \\
&\{M' : r'(s, a) = r(s, a), c'(i, s, a) = c(i, s, a), \gamma' = \gamma, \\
&|P'(s'|s, a) - \widehat{P}(s'|s, a)| \leq \\
&\min\Big(\sqrt{\frac{2\widehat{P}(s'|s, a)(1 - \widehat{P}(s'|s, a))}{n(s, a)}} \log\frac{4}{\delta_1} + \\
&\frac{2}{3n(s, a)} \log\frac{4}{\delta_1}, \sqrt{\frac{\log 4/\delta_1}{2n(s, a)}}\Big) \;\; \forall s, s', a, i\},
\end{aligned} \tag{8}$$

where δ_1 is defined in Algorithm 2.

Subsequently, UC-CRL uses ELP to solve the optimistic CMDP problem below and get the optimistic policy $\tilde{\pi}_t$:

$$\max_{\pi, M' \in \mathcal{M}_t} V'^{\pi}(s_0) \text{ s.t. } C_i'^{\pi}(s_0) \leq \bar{C}_i \ \forall\, i.$$

This problem is identical to the problem of (7), except for substituting $\mathcal{M}_{\delta_P}$ with $\mathcal{M}_t$. Here, for any $M' \in \mathcal{M}_t$, $V'^{\pi}(s_0)$ and $C_i'^{\pi}(s_0)$ are computed according to (1) and (3) w.r.t. underlying transition kernel P', respectively.

Algorithm 2 UC-CRL

```
Input: accuracy $\epsilon$ and failure tolerance $\delta$.
Set $m$ according to (9) and (10).
Set $t = 1, w_{\min} = \frac{\epsilon(1-\gamma)}{4|S|}, U_{\max} = |S|^2|A|m, \delta_1 = \frac{\delta}{4(N+1)|S|U_{\max}}$.
Set $n(s,a) = n(s',s,a) = 0 \ \forall s, s' \in S, a \in A$.
while there is $(s,a)$ with $n(s,a) < \frac{|S|m}{1-\gamma}$ do
    $\widehat{P}(s'|s,a) = \frac{n(s',s,a)}{n(s,a)} \ \ \forall (s,a)$ with $n(s,a) > 0$ and $s' \in S$.
    Construct $\mathcal{M}_t$ according to (8).
    $\tilde{\pi}_t = \text{ELP}(\mathcal{M}_t)$.
    $a_t \sim \tilde{\pi}_t(s_t), s_{t+1} \sim P(\cdot|s_t, a_t)$
    if $n(s_t, a_t) < \frac{|S|m}{1-\gamma}$ then
        $n(s_t, a_t) ++, n(s_{t+1}, s_t, a_t) ++$.
    end if
    $t ++$
end while
```

UC-CRL is inspired by the infinite-horizon algorithm UCRL$-\gamma$ [Lattimore and Hutter, 2014] and its finite-horizon equivalent UCFH [Dann and Brunskill, 2015] with differences. Similar to UCRL$-\gamma$, Algorithm 2 uses a combination of the empirical Bernstein's and Hoeffding's inequalities. These concentration inequalities allow us to ensure linearity of constraints (i.e., we can indeed use an extended linear program to solve for the constrained optimistic policy). However, the constraints of UCFH contain non-linear expressions preventing us from employing ELP. Furthermore, unlike UCRL$-\gamma$ and UCFH, Algorithm 2 updates the model at each time-step rather than at the beginning of long phases. This procedure allows for faster model construction. Finally, since we are solving a CMDP, this algorithm utilizes ELP instead of Extended Value Iteration which is used by UCRL$-\gamma$.

4.1 PAC Analysis of UC-CRL

We now present the PAC bound of Algorithm 2.

Theorem 2. *Consider CMDP* $M = \langle S, A, P, r, c, \bar{C}, s_0, \gamma \rangle$ *satisfying assumptions 1 and 2. For any* $0 < \epsilon, \delta < 1$, *under UC-CRL we have:*

$$\mathbb{P}(V^{\tilde{\pi}_t}(s_0) \geq V^{\pi^*}(s_0) - \epsilon \ \textit{ and }$$
$$C_i^{\tilde{\pi}_t}(s_0) \leq \bar{C}_i + \epsilon \ \forall i \in \{1, 2, \ldots, N\}) \geq 1 - \delta,$$

for all but at most

$$\tilde{O}\Big(\frac{|S|^2|A|}{\epsilon^2(1-\gamma)^3} \log \frac{(N+1)}{\delta}\Big)$$

time-steps.

We follow an approach motivated by [Lattimore and Hutter, 2014] and its finite-horizon version [Dann and Brunskill, 2015] to prove Theorem 2. However, there are several differences in our technique, and we need to accommodate the frequent model update in our proof. We will show that, unlike existing approaches, we can update the model at each time-step, without increasing the sample complexity. Thus, we are able to obtain PAC bounds that match the unconstrained case, and only increase by logarithmic factor with the number of constraints.

There are also recent works on characterizing the regret of constrained-RL in a finite-horizon setting [Efroni *et al.*, 2020] with an algorithm similar to Algorithm 2. An important emerging question is whether one can immediately convert these regret results into sample complexity bounds? A naive translation of the regret bounds of [Efroni *et al.*, 2020] would give us a PAC result $\tilde{O}(\frac{|S|^2|A|H^4}{\epsilon^2})$. For comparing finite-horizon setting with infinite-horizon one, we can replace H with $\frac{1}{1-\gamma}$ to obtain a PAC result for the equivalent infinite-horizon algorithm. Considering this, the approach followed by [Efroni *et al.*, 2020] gives a PAC bound which is looser than our result by a factor of $\frac{1}{(1-\gamma)^2}$. Therefore, this alternative option does not lead to the strong bounds that we are able to obtain, and matches existing PAC results of the unconstrained case.

Now, we present the notions of *knownness* and *importance* for state-action pairs and base our proof on these notions. Then we present the key lemmas needed for proving Theorem 2. Finally, we provide a proof sketch for Theorem 2. The detailed analysis is provided in supplementary materials.

Let the *weight* of $(s,a)-$pair under any policy π be its discounted expected frequency

$$\begin{aligned} & w^{\pi}(s,a|s') \\ & := \mathbb{I}\{(s', \pi(s')) = (s,a)\} + \gamma \sum_{s''} P_{\pi}(s''|s') w^{\pi}(s,a|s''). \end{aligned}$$

Using this general definition, we define the weight of (s,a) under policy $\tilde{\pi}_t$ as

$$w_t(s,a) = w^{\tilde{\pi}_t}(s,a|s_t).$$

Then, the *importance* ι_t of (s,a) at time-step t is defined as its relative weight compared to $w_{\min} := \frac{\epsilon(1-\gamma)}{4|S|}$ on a log-scale

$$\iota_t(s,a) := \min\{z_j : z_j \geq \frac{w_t(s,a)}{w_{\min}}\}$$
$$\text{where } \ z_1 = 0 \text{ and } z_j = 2^{j-2} \ \forall j = 2, 3, \ldots.$$

Note that $\iota_t(s,a) \in \{0, 1, 2, 4, 8, 16, \ldots\}$ is an integer indicating the influence of the state-action pair on the value function of $\tilde{\pi}_t$. Similarly, we define *knownness* as

$$\kappa_t(s,a) := \max\{z_i : z_i \leq \frac{n_t(s,a)}{m w_t(s,a)}\} \in \{0, 1, 2, 4, \ldots\},$$

which indicates how often (s,a) has been observed relative to its importance. Value of m is defined in Algorithm 2. Now, we can categorize $(s, a)-$pairs into subsets

$$X_{t,\kappa,\iota} := \{(s, a) \in X_t : \kappa_t(s, a) = \kappa, \iota_t(s, a) = \iota\}$$
$$\text{and } \bar{X}_t = S \times A \setminus X_t,$$

where $X_t = \{(s, a) : \iota_t(s, a) > 0\}$ is the active set and $\bar{X}_t$ is the set of $(s, a)-$pairs that are very unlikely under policy $\tilde{\pi}_t$. We will show that if the criteria $|X_{t,\kappa,\iota}| \leq \kappa$ is met, then the model of UC-CRL would achieve near-optimal policies where these policies would violate constraints at most by ϵ w.h.p. This condition specifies that important state-action pairs under policy $\tilde{\pi}_t$ are visited a sufficiently large number of times. Thus, the model of UC-CRL will be accurate enough to obtain PAC bounds.

Now, we first show that the true model lies in $\mathcal{M}_t$ for every time-step t w.h.p.

Lemma 3. *$M \in \mathcal{M}_t$ for all time-steps t with probability at least $1 - \frac{\delta}{2(N+1)}$.*

Proof Sketch: Let consider a fixed (s, a), next state s' and a time-step t. Then, $P(s'|s, a)$ belongs to the confidence set constructed by the combined Bernstein's and Hoeffding's inequalities. By taking the union bound over maximum number of model updates, $U_{\max}$, and next states we obtain the result. □

Next, we bound the number of time-steps in which the condition $|X_{t,\kappa,\iota}| \leq \kappa$ is violated w.h.p.

Lemma 4. *Suppose E is the number of time-steps t for which there are κ and ι with $|X_{t,\kappa,\iota}| > \kappa$, i.e. $E = \sum_{t=1}^{\infty} \mathbb{I}\{\exists(\kappa, \iota) : |X_{t,\kappa,\iota}| > \kappa\}$ and let*

$$m \geq \frac{4}{\epsilon(1-\gamma)^3} \log \frac{2(N+1)E_{\max}}{\delta}, \tag{9}$$

where $E_{\max} = \log_2 \frac{1}{w_{\min}(1-\gamma)} \log_2 |S|$. Then, $\mathbb{P}(E \leq 6|S||A|mE_{\max}) \geq 1 - \frac{\delta}{2(N+1)}$.

Proof sketch: This lemma is proven in two stages. First, we bound the total number of times a fixed (s, a) could be observed in a particular $X_{t,\kappa,\iota}$ over all time-steps. Then, we provide a high probability bound on the number of time-steps that $|X_{t,\kappa,\iota}| > \kappa$ for a fixed (κ, ι). Finally, we get the result using of martingale concentration and union bound. □

Finally, the next lemma bounds the mismatch between objective and constraint functions of the optimistic model and true model. This lemma functions similarly to Lemma 2 for GM-CRL. It provides a uniform PAC result over value and constraint functions. Hence, it enables us to have individual PAC results for any objective and constraint functions. As discussed in GM-CRL section, this process is responsible for a $\log N$ increase in the PAC result.

Lemma 5. *Assume $M \in \mathcal{M}_t$. If $|X_{t,\kappa,\iota}| \leq \kappa$ for all (κ, ι) and $0 < \epsilon \leq 1$ and*

$$m \geq 1280 \frac{|S|}{\epsilon^2(1-\gamma)^2} (\log_2 \log_2(\frac{1}{1-\gamma}))^2 \log_2^2\Big(\frac{8|S|^2}{\epsilon(1-\gamma)^2}\Big) \times \log \frac{4}{\delta_1}, \tag{10}$$

then $|\tilde{V}^{\tilde{\pi}_t}(s_0) - V^{\tilde{\pi}_t}(s_0)| \leq \epsilon$ and for any i, $|\tilde{C}_i^{\tilde{\pi}_t}(s_0) - C_i^{\tilde{\pi}_t}(s_0)| \leq \epsilon$.

Proof Sketch: First, we show $|\tilde{P}(s'|s, a) - P(s'|s, a)| \leq O(\sqrt{\frac{P(s'|s,a)(1-P(s'|s,a))}{n}})$ for each s', s, a. Then we prove that at each time-step t, $(P_\pi - \tilde{P}_\pi)V^\pi(s) \leq O(\sqrt{\frac{|S|}{n}}\sigma_{V^\pi}(s))$. Next we partition the state-action based on knownness, i.e., whether they belong to X_t or not. By using all bounds and sequence of CMDPs, we obtain a bound on $|\tilde{V}^\pi(s_0) - V^\pi(s_0)|$. Eventually, we use the definition of weights to get the final result. This procedure is also applicable to each constraint function i. □

***Proof Sketch of Theorem 2*:** We first use Lemma 3 and show that true CMDP is admissible ,i.e. $M \in \mathcal{M}_t$ for every time-step, w.p. at least $1 - \frac{\delta}{2(N+1)}$. Hence, the optimistic planning problem becomes feasible and an optimistic policy $\tilde{\pi}_t$ exists w.h.p. Further, we use Lemma 4 to bound the number of time-steps where $|X_{t,\kappa,\iota}| > \kappa$ w.h.p. Thus, for other time-steps where $|X_{t,\kappa,\iota}| \leq \kappa$, we apply Lemma 5 we show that objective function is $\epsilon-$optimal and all constraint functions are violated by ϵ. Eventually, we obtain the result by means of union bound. □

5 Conclusion

In this paper, we presented the notion of sample complexity in objective maximization and constraint satisfaction in order to understand the performance of RL algorithms for safety-constrained applications. We proposed and analyzed two types of algorithms—GM-CRL and UC-CRL. The main finding of a logarithmic factor increase in sample complexity compared to unconstrained regime indicates the value of the algorithms in applying them to real systems.

Acknowledgments

This work was funded in part by grants NSF-Intel CNS 1719384, NSF CPS-2038963, NSF CNS 1955696, ARO W911NF-19-1-0367, NSF CRII- CPS-1850206, NSF CAREER-EPCN- 2045783.

References

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

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

[Amani *et al.*, 2019] Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. In *Advances in Neural Information Processing Systems*, pages 9252–9262, 2019.

[Azar *et al.*, 2013] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349, 2013.

[Badanidiyuru *et al.*, 2013] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on*

Foundations of Computer Science, pages 207–216. IEEE, 2013.

[Bhatnagar and Lakshmanan, 2012] 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.

[Borkar and Jain, 2014] Vivek Borkar and Rahul Jain. Risk-constrained Markov decision processes. *IEEE Transactions on Automatic Control*, 59(9):2574–2579, 2014.

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

[Chow *et al.*, 2018] Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 8092–8101, 2018.

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

[Dann *et al.*, 2017] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5713–5723, 2017.

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

[Hoeffding, 1994] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

[Lattimore and Hutter, 2014] Tor Lattimore and Marcus Hutter. Near-optimal pac bounds for discounted mdps. *Theoretical Computer Science*, 558:125–143, 2014.

[Liang *et al.*, 2018] Qingkai Liang, Fanyu Que, and Eytan Modiano. Accelerated primal-dual policy optimization for safe reinforcement learning. *arXiv preprint arXiv:1802.06480*, 2018.

[Liu *et al.*, 2019] Yongshuai Liu, Jiaxin Ding, and Xin Liu. Ipo: Interior-point policy optimization under constraints. *arXiv preprint arXiv:1910.09615*, 2019.

[Maurer and Pontil, 2009] Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

[Paternain *et al.*, 2019] Santiago Paternain, Luiz Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. In *Advances in Neural Information Processing Systems*, pages 7553–7563, 2019.

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

[Satija *et al.*, 2020] Harsh Satija, Philip Amortila, and Joelle Pineau. Constrained markov decision processes via backward value functions. *arXiv preprint arXiv:2008.11811*, 2020.

[Singh and Kumar, 2018] 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.

[Singh *et al.*, 2014] Rahul Singh, I-Hong Hou, and PR Kumar. Fluctuation analysis of debt based policies for wireless networks with hard delay constraints. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, pages 2400–2408. IEEE, 2014.

[Strehl and Littman, 2008] Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

[Tessler *et al.*, 2018] Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074*, 2018.

[Wachi and Sui, 2020] Akifumi Wachi and Yanan Sui. Safe reinforcement learning in constrained markov decision processes. *arXiv preprint arXiv:2008.06626*, 2020.

[Wu *et al.*, 2015] Huasen Wu, Rayadurgam Srikant, Xin Liu, and Chong Jiang. Algorithms with logarithmic or sublinear regret for constrained contextual bandits. In *Advances in Neural Information Processing Systems*, pages 433–441, 2015.

[Zheng and Ratliff, 2020] Liyuan Zheng and Lillian J Ratliff. Constrained upper confidence reinforcement learning. *arXiv preprint arXiv:2001.09377*, 2020.