Learning with Safety Constraints: Sample Complexity of Reinforcement Learning for Constrained MDPs

Aria HasanzadeZonuzy, Archana Bura, Dileep Kalathil, Srinivas Shakkottai
Texas A & M University
{azonuzy, archanabura, dileep.kalathil, sshakkot} @tamu.edu

Abstract

Many physical systems have underlying safety considerations that require that the policy employed ensures the satisfaction of a set of constraints. The analytical formulation usually takes the form of a Constrained Markov Decision Process (CMDP).We focus on the case where the CMDP is unknown, and RL algorithms obtain samples to discover the model and compute an optimal constrained policy. Our goal is to characterize the relationship between safety constraints and the number of samples needed to ensure a desired level of accuracy—both objective maximization and constraint satisfaction—in a PAC sense. We explore two classes of RL algorithms, namely, (i) a generative model based approach, wherein samples are taken initially to estimate a model, and (ii) an online approach, wherein the model is updated as samples are obtained. Our main finding is that compared to the best known bounds of the unconstrained regime, the sample complexity of constrained RL algorithms are increased by a factor that is logarithmic in the number of constraints, which suggests that the approach may be easily utilized in real systems.

Introduction

Markov Decision Processes (MDPs) are used to model a variety of systems for which stationary control policies are appropriate. In many cyber-physical systems (algorithmically controlled physical systems) restrictions may be placed on functions of the probability with which states may be visited. For example, in power systems, the frequency must be kept within tolerable limits, and allowing it to go outside these tolerances often might be unsafe. Similarly, in communication systems the number of transmissions that may be made in a time interval is limited by an average radiated power constraint due to interference and human safety considerations. The number of constraints can be large, since they can represent physical limitations (e.g., communication or transmission link capacities), performance requirements (per-flow packet delays, tolerable frequencies) and so on. The Constrained-MDP (CMDP) framework is used to model such circumstances (Altman 1999).

In this paper, our objective is to design simple algorithms to solve CMDP problems under an unknown model. Whereas the goal of a typical model-based RL approach would take as few samples as possible to quickly determine the optimal policy, minimizing the number of samples taken is even more important in the CMDP setting. This because constraints are violated during the learning process, and it might be critical to keep the number of such violations as low as possible due to safety considerations mentioned earlier, and yet ensure that the system objectives are maximized. Hence, determining how the joint metrics of objective maximization and safety violation evolve over time as the model becomes more and more accurate is crucial to understand the efficacy of a proposed RL algorithm for CMDPs.

Main Contributions: Our goal is to analyze the sample complexity of solving CMDPs to a desired accuracy with a high probability in both objective and constraints in the context of finite horizon (episodic) problems. We focus on two figures of merit pertaining to objective maximization and constraint satisfaction in a probably-approximately-correct (PAC) sense. Our main contributions are as follows:

(i) We develop two model-based algorithms, namely, (i) a generative approach that obtains samples initially then creates a model, and (ii) an online approach in which the model is updated as time proceeds. In both cases, the estimated model might have no solution, and we utilize a confidence-ball around the estimate to ensure that a solution may be found with high probability (assuming that the real model has a solution).

(ii) The algorithms follow the general pattern of model construction or update, followed by a solution using linear programming (LP) of the CMDP generated in this manner, with the addendum that the LP is extended to account for the fact that a search is made over the entire ball of models given the current samples. This procedure not only contributes to optimism as (Efroni, Mannor, and Pirotta 2020), but also guarantees feasibility of the solution.

(iii) We develop PAC-type sample complexity bounds for both algorithms, accounting for both objective maximization and constraint satisfaction. The general intuition is that the model accuracy should be higher than in the unconstrained case and, our main finding agrees with this intuition. Furthermore, comparing our main results with lower bounds on sample complexity of MDPs (Azar, Munos, and Kappen 2013; Dann and Brunskill 2015), we discover that the in-

crease in the sample complexity is by a logarithmic factor in the number of constraints and a size of state space. However, there are no lower bound results for CMDPs to the best of our knowledge.

As mentioned above, the number of constraints in cyber-physical systems can be large. Our result indicating logarithmic scaling with the number of constraints indicates that the number of constraints is not a major concern in solving unknown CMDPs via RL, hence indicating that the practicality of applying the constrained RL approach to cyber-physical systems applications.

Related Work: Much work in the space of CMDP has been driven by problems of control, and many of the algorithmic approaches and applications have taken a control-theoretic view (Altman 1999, 2002; Borkar 2005; Borkar and Jain 2014; Singh and Kumar 2018; Singh, Hou, and Kumar 2014). The approach taken is to study the problem under a known model, and showing asymptotic convergence of the solution method proposed. There are also studies on constrained partially observable MDPs such as (Isom, Meyn, and Braatz 2008; Kim et al. 2011). Both of these works propose algorithms based on value iteration requiring solving linear program or constrained quadratic program.

Extending CMDP approaches to the context on an unknown model has also mostly focused on asymptotic convergence (Bhatnagar and Lakshmanan 2012; Chow et al. 2018; Tessler, Mankowitz, and Mannor 2018; Paternain et al. 2019) under Lagrangian methods to show zero eventual duality gap. (Liu, Ding, and Liu 2019) also proposes an algorithm based on Lagrangian method, but proves that this algorithm achieves a small eventual gap. On the other hand empirical works built on Lagrangian method has also been proposed (Liang, Que, and Modiano 2018).

A parallel theme has been related to the constrained bandit case, wherein the the underlying problem, while not directly being an MDP, bears a strong relation to it. Work such as (Badanidiyuru, Kleinberg, and Slivkins 2013; Wu et al. 2015; Amani, Alizadeh, and Thrampoulidis 2019) consider such constraints, either in a knapsack sense, or on the type of controls that may be applied in a linear bandit context.

Closest to our theme are parallel works on CMDPs. For instance, (Zheng and Ratliff 2020) and (Wachi and Sui 2020) present results in the context of unknown reward functions, with either a known stochastic or deterministic transition kernel. Other work (Satija, Amortila, and Pineau 2020) focuses on asymptotic convergence, and so does not provide an estimate on the learning rate. Finally, (Efroni, Mannor, and Pirotta 2020) explores algorithms and themes similar to ours, but focuses on characterizing objective and constrained regret under different flavors of online algorithms, which can be seen as complementary to or work. Since there is no direct relation between regret and sample complexity (Dann, Lattimore, and Brunskill 2017), applying their regret approach to our setting gives relatively weak sample complexity bounds. Our discovery of a general principle of logarithmic increase in sample complexity with the number of constraints also distinguishes our work.

Notation and Problem Formulation

Notation and Setup: We consider a general finite-horizon CMDP formulation. There are a set of states S and set of actions A. The reward matrix is denoted by r, under which $r(s,a)$ is the reward for any 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 $i \in \{1,\dots,N\}$. Also, the vector $\bar{C}$ is used to denote the value of the constraints (i.e., the bound that must be satisfied). The probability of reaching another state s' while being at state s and taking action a is determined by transition kernel $P(s'|s,a)$. At the beginning of each horizon, we begin from a fixed initial state s_0. As the CMDP has a finite horizon, the length of each horizon, or episode, is considered to be a fixed value H. Hence, the CMDP is defined by the tuple $M = \langle S, A, P, r, c, \bar{C}, s_0, H\rangle$.

Assumption 1. *We assume S and A are finite sets with cardinalities $|S|$ and $|A|$. Further, we assume that the immediate reward $r(s,a)$ is taken from the interval $[0,1]$ and immediate cost lies in $[0,1]$. We also make an assumption that there are N constraints which for each $i \in \{1,\dots,N\}, \bar{C}_i \in [0, \bar{C}_{\max}]$.*

Next, to choose an action from A at time-step h, we define a policy π as a mapping from state-action space $S \times A$ to set of probability vectors defined over action space, i.e. $\pi : S \times A \to [0,1]^{|A|}$. So $\pi(s,\cdot,h)$ is a probability vector over A at time-step h. Also, $a \sim \pi(s,\cdot,h)$ means that action a is chosen according to policy π while being at state s at time-step h.

When policy π is fixed, the underlying Markov Decision Process turns into a Markov chain. The transition kernel of this Markov chain is P_π, which can be viewed as an operator. The operator $P_\pi f(s) = \mathbb{E}[f(s_{h+1})|s_h = s] = \sum_{s' \in S} P_\pi(s'|s) f(s')$ takes any function $f : S \to \mathbb{R}$ and returns the expected value of f in the next time step. For convenience, we define the multi-step version $P_\pi^h f(s) = P_\pi P_\pi \dots P_\pi f$, which is repeated h times. Further, we define P_π^{-1} and P_π^0 as the identity operator.

We consider cumulative finite horizon criteria for both the objective function and the constraint functions with identical horizon H. We define the value function of state s at time-step t under policy π as

$$V_t^\pi(s) = \mathbb{E}[\sum_{h=t}^{H-1} r(s_h, a_h); a_h \sim \pi(s_h,\cdot,h), s_t = s], \quad (1)$$

where action a_h is chosen according to policy π and expectation $\mathbb{E}[.]$ is taken w.r.t transition kernel P. Then, the local variance of the value function at time step h under policy π is

$$\sigma_h^{\pi^2}(s) = \mathbb{E}[(V_{h+1}^\pi(s_{h+1}) - P_\pi V_{h+1}^\pi(s))^2]. \quad (2)$$

Similar to the definition of the value function (1), the i^{th} constraint function at time t under policy π is formulated as

$$C_{i,t}^\pi(s) = \mathbb{E}[\sum_{h=t}^{H-1} c(i, s_h, a_h); a_t \sim \pi(s_h,\cdot,h), s_t = s]. \quad (3)$$

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

Finally, the general finite-horizon CMDP problem is

$$\max_{\pi} V_0^{\pi}(s_0) \text{ s.t. } C^{\pi}_{i,0}(s_0) \leq \bar{C}_i, \quad \forall i \in \{1,\dots,N\}. \quad (4)$$

Assumption 2. *We assume that there exists some policy π that satisfies the constraints in* (4)*. Hence, this CMDP problem is feasible with optimal policy π^* and optimal solution $V_0^*(s_0) = V_0^{\pi^*}(s_0)$.*

Note that we only consider learning feasible CMDPs, since otherwise no algorithm would be able to discover an optimal policy satisfying constraints.

Constrained-RL Problem: The Constrained RL problem formulation is identical to the CMDP optimization problem of (4), but without being aware of values of transition kernel P.[1] Our goal is 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*

$$\mathbb{P}\Big(V_0^{\mathcal{A}}(s_0) \geq V_0^{\pi^*}(s_0) - \epsilon \text{ and } C^{\mathcal{A}}_{i,0}(s_0) \leq \bar{C}_i + \epsilon \; \forall i \in \{1,\dots,N\}\Big) \geq 1-\delta$$

for a given ϵ and δ.

Note that this definition includes both objective maximization and constraint violations, as opposed to a traditional definition that only considers the objective (Strehl and Littman 2008).

Sample Complexity Result of Generative Model Based Learning

In this section, we introduce a generative model based CMDP learning algorithm called Optimistic Generative Model Based Learning, or Optimistic-GMBL. According to Optimistic-GMBL, we sample each state-action pair n number of times uniformly across all state-action pairs, count the number of times each transition occurs $n(s',s,a)$ for each next state s', and construct an empirical model of transition kernel denoted by $\widehat{P}(s'|s,a) = \frac{n(s',s,a)}{n} \; \forall (s',s,a)$. Then Optimistic-GMBL 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 horizon of the true CMDP, but with transition kernels close to true model. This class of CMDPs is defined as

$$\mathcal{M}_{\delta_P} := \{M' : r'(s,a) = r(s,a), \quad (5)$$
$$c'(i,s,a) = c(i,s,a), H' = H, s'_0 = s_0$$
$$|P'(s'|s,a) - \widehat{P}(s'|s,a)| \leq \quad (6)$$

$$\min\Big(\sqrt{\frac{2\widehat{P}(s'|s,a)(1-\widehat{P}(s'|s,a))}{n}\log\frac{4}{\delta_P}} + \frac{2}{3n}\log\frac{4}{\delta_P}, \sqrt{\frac{\log 4/\delta_P}{2n}}\Big) \forall s,a,s',i\},$$

where δ_P is defined in Algorithm 1. For any $M' \in \mathcal{M}$, objective function $V_0^{'\pi}(s_0)$ and cost functions $C^{'\pi}_{i,0}(s_0)$ are computed w.r.t. the corresponding transition kernel P' according to equations (1) and (3) respectively.

Finally, Optimistic-GMBL maximizes the objective function among all possible transition kernels, while satisfying constraints (if feasible). More specifically, it solves the optimistic planning problem below

$$\max_{\pi, M' \in \mathcal{M}_{\delta_P}} V_0^{'\pi}(s_0) \quad \text{s.t.} \quad C^{'\pi}_{i,0}(s_0) \leq \bar{C}_i \;\; \forall i. \quad (7)$$

Optimistic-GMBL uses Extended Linear Programming, or **ELP**, to solve the problem of (7). This method inputs $\mathcal{M}_{\delta_P}$ and outputs $\tilde{\pi}$ for the optimal solution. The description of ELP is provided in (HasanzadeZonuzy, Kalathil, and Shakkottai 2020). Algorithm 1 describes Optimistic-GMBL.

Algorithm 1 Optimistic-GMBL

1: Input: accuracy ϵ and failure tolerance δ.
2: Set $\delta_P = \frac{\delta}{12(N+2)|S|^2|A|H}$.
3: Set $n(s',s,a) = 0 \; \forall (s,a,s')$.
4: **for** each $(s,a) \in S \times A$ **do**
5: Sample (s,a), $n = \frac{256}{\epsilon^2}|S|H^3 \log\frac{12(N+2)|S||A|H}{\delta}$ and update $n(s',s,a)$.
6: $\widehat{P}(s'|s,a) = \frac{n(s',s,a)}{n} \; \forall s'$.
7: Construct $\mathcal{M}_{\delta_P}$ according to (5).
8: Output $\tilde{\pi} = \text{ELP}(\mathcal{M}_{\delta_P})$.

PAC Analysis of Optimistic-GMBL

Here, we present the sample complexity result of Optimistic-GMBL. Time complexity result and analysis will be provided in (HasanzadeZonuzy, Kalathil, and Shakkottai 2020).

Theorem 1. *Consider any finite-horizon CMDP $M = \langle S, A, P, r, c, \bar{C}, s_0, H\rangle$ satisfying assumptions 1 and 2, and CMDP problem formulation of* (4)*. Then, for any $\epsilon \in (0, \frac{2}{9}\sqrt{\frac{H}{|S|}})$ and $\delta \in (0,1)$, algorithm 1 creates a model CMDP $\tilde{M} = \langle S, A, \tilde{P}, r, c, \bar{C}, s_0, H\rangle$ and outputs policy $\tilde{\pi}$ such that*

$$\mathbb{P}(V_0^{\tilde{\pi}}(s_0) \geq V_0^{\pi^*}(s_0) - \epsilon \text{ and } C^{\tilde{\pi}}_{i,0}(s_0) \leq \bar{C}_i + \epsilon \; \forall i \in \{1,2,\dots,N\}) \geq 1-\delta,$$

with at least total sampling budget of

$$\frac{256}{\epsilon^2}|S|^2|A|H^3 \log\frac{12(N+2)|S||A|H}{\delta}.$$

The proof of Theorem 1 differs from the traditional analysis framework of unconstrained RL (Azar, Munos, and Kappen 2013) in the following manner. First, is the role played

[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.

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 any 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, Optimistic-GMBL 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 $O(|S|^2 \log |S|)$ sample complexity in the size of state space.

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 (HasanzadeZonuzy, Kalathil, and Shakkottai 2020).

First, we show that true CMDP lies inside the $\mathcal{M}_{\delta_P}$ with high probability, w.h.p. So, the problem (7) would be 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 get 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 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 $n \geq 2592|S|^2H^2 \log 4/\delta_P$, under any policy π*

$$\|V_0^\pi - \tilde{V}_0^\pi\|_\infty \leq \sqrt{\frac{32|S|H^3}{n}}$$

w.p. at least $1 - 3|S|^2|A|H\delta_P$, and for any $i \in \{1, \dots, N\}$,

$$\|C_{i,0}^\pi - \tilde{C}_{i,0}^\pi\|_\infty \leq \sqrt{\frac{32|S|H^3}{n}}$$

w.p. at least $1 - 3|S|^2|A|H\delta_P$.

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 at each time-step h, $(P_\pi - \tilde{P}_\pi)V_h^\pi(s) \leq O(\sqrt{\frac{|S|}{n}}\sigma_h^\pi(s))$. Applying this bound to $|\tilde{V}_0^\pi(s_0) - V_0^\pi(s_0)|$ and from the fact that $\sigma_h^\pi(s)$ is close to $\tilde{\sigma}_h^\pi(s)$ by $\frac{\sqrt{|S|H^2}}{n^{1/4}}$, 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 $|V_0^{\pi^*}(s_0) - \tilde{V}_0^{\pi^*}(s_0)| \leq O(\sqrt{\frac{|S|H^3}{n}})$ w.h.p. by means of Lemma 2. Similarly, $|V_0^{\tilde{\pi}}(s_0) - \tilde{V}_0^{\tilde{\pi}}(s_0)| \leq O(\sqrt{\frac{|S|H^3}{n}})$ w.h.p. Next, we use the fact that $\tilde{V}_0^{\pi^*}(s_0) \leq \tilde{V}_0^{\tilde{\pi}}(s_0)$ and obtain

$$V_0^{\tilde{\pi}}(s_0) \geq V_0^{\pi^*}(s_0) - O(\sqrt{\frac{|S|H^3}{n}}).$$

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 $i \in \{1, \dots, N\}$ we have $|C_{i,0}^{\tilde{\pi}}(s_0) - \tilde{C}_{i,0}^{\tilde{\pi}}(s_0)| \leq O(\sqrt{\frac{|S|H^3}{n}})$ w.h.p. Also, we know that $\tilde{C}_{i,0}^{\tilde{\pi}}(s_0) \leq \bar{C}_i$, since $\tilde{\pi}$ is solution of the ELP. Hence, we obtain

$$C_{i,0}^{\tilde{\pi}}(s_0) \leq \bar{C}_i + O(\sqrt{\frac{|S|H^3}{n}})$$

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

Sample Complexity Result of Online Learning

The Optimistic-GMBL approach requires that every state-action pair in the system be sampled a certain number of times before a policy is computed. However, many applications may not be able to utilize this approach since it may not be possible to reach those states without the application of some policy, or they might be unsafe and so should not be sampled often. Hence, we need an approach that can collect samples from the environment by means of an online algorithm.

Online Constrained-RL, or Online-CRL described in Algorithm 2, is an online method proceeding in episodes with length H. At the beginning of each episode k, Online-CRL constructs an empirical model $\widehat{P}$ according to 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. This empirical model $\widehat{P}$ induces a set of finite-horizon CMDPs $\mathcal{M}_k$ which any CMDP $M' \in \mathcal{M}_k$ has identical horizon and reward and cost matrices. However, for any $(s, a) \in S \times A$ and $s' \in S$, $P'(s'|s, a)$ lies inside a confidence interval induced

Algorithm 2 Online-CRL

Input: accuracy ϵ and failure tolerance δ.
Set $k = 1, w_{\min} = \frac{\epsilon}{4H|S|}, U_{\max} = |S|^2|A|m, \delta_1 = \frac{\delta}{4(N+1)|S|U_{\max}}$.
Set m according to (9) and (10).
Set $n(s,a) = n(s',s,a) = 0 \ \forall s, s' \in S, a \in A$.
while there is (s,a) with $n(s,a) < |S|mH$ **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 $\mathcal{M}_k$ according to (8).
  $\tilde{\pi}_k = \text{ELP}(\mathcal{M}_k)$.
  for $t = 1, \ldots, H$ **do**
    $a_t \sim \tilde{\pi}_k(s_t), s_{t+1} \sim P(\cdot|s_t,a_t), n(s_t,a_t) ++, n(s_{t+1},s_t,a_t) ++$.
  $k++$

by $\widehat{P}$. To construct a confidence interval for any element of $P'(s'|s,a)$, we use identical concentration inequalities to Optimistic GMBL as defined by (6). The only difference is the use of $n(s,a)$ instead of n. Thus the class of CMPDs is defined as below at each episode k :

$$\begin{aligned}
&\mathcal{M}_k := \{M' : r'(s,a) = r(s,a), \\
&c'(i,s,a) = c(i,s,a), H' = H, s'_0 = s_0 \\
&|P'(s'|s,a) - \widehat{P}(s'|s,a)| \leq \\
&\min\Big(\sqrt{\frac{2\widehat{P}(s'|s,a)(1-\widehat{P}(s'|s,a))}{n(s,a)}\log\frac{4}{\delta_1}} \\
&+ \frac{2}{3n(s,a)}\log\frac{4}{\delta_1}, \sqrt{\frac{\log 4/\delta_1}{2n(s,a)}}\Big) \ \forall s, s', a, i\},
\end{aligned} \tag{8}$$

where δ_1 is defined in Algorithm 2.

Next, we use ELP to obtain an optimistic policy $\tilde{\pi}_k$, which is the solution of optimistic CMDP problem below:

$$\max_{\pi, M' \in \mathcal{M}_k} V_0^{'\pi}(s_0) \ \text{ s.t. } \ C_{i,0}^{'\pi}(s_0) \leq \bar{C}_i \ \forall \, i.$$

This problem is exactly the same as problem of (7), except for substituting $\mathcal{M}_{\delta_P}$ with $\mathcal{M}_k$. Here, for any $M' \in \mathcal{M}_k$, $V_0'^{\pi}(s_0)$ and $C_{i,0}'^{\pi}(s_0)$ are computed according to (1) and (3) w.r.t. underlying transition kernel P', respectively.

This algorithm draws inspiration from the infinite-horizon algorithm UCRL$-\gamma$ (Lattimore and Hutter 2014) and its finite-horizon counterpart UCFH (Dann and Brunskill 2015) with several differences. Unlike UCRL-γ and UCFH, Algorithm 2 updates the model at the beginning of each episode, which allows for faster model construction. Also, since we desire a policy that pertains to a CMDP using an linear programming approach (Altman 1999), we must ensure that all constraints are linear. Hence, unlike UCFH, Algorithm 2 utilizes a combination of the empirical Bernstein's and Hoeffding's inequalities, which allows 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 are non-linear and require the use of extended value iteration coupled with a complex sub-routine, which cannot be utilized in the constrained RL case. Thus, we are able to obtain strong bounds on sample complexity similar to UCFH, but yet ensure that the solution approach only uses a linear program.

PAC Analysis of Online-CRL

We now present the PAC bound of Algorithm 2.

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

$$\begin{aligned}
&\mathbb{P}(V_0^{\tilde{\pi}_k}(s_0) \geq V_0^{\pi^*}(s_0) - \epsilon \ \textit{and} \\
&C_{i,0}^{\tilde{\pi}_k}(s_0) \leq \bar{C}_i + \epsilon \ \forall i \in \{1,2,\ldots,N\}) \geq 1-\delta,
\end{aligned}$$

for all but at most

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

episodes.

To prove Theorem 2, we follow an approach motivated by (Lattimore and Hutter 2014) and its finite-horizon version (Dann and Brunskill 2015). However, there are several differences in our technique. As mentioned above, one of the differences is with regard to restricting ourselves to only linear concentration inequalities. We will show that excluding non-linear concentration inequalities pertaining to variance does not increase the sample complexity, and utilizing the fact that the number of successor states is less that $|S|$ leads to matching sample complexity in terms of $|S|$ with the UCFH algorithm. Furthermore, we are able to show that, unlike existing approaches, we can update the model at each episode, again 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 results on characterizing the regret of constrained-RL (Efroni, Mannor, and Pirotta 2020) while using an algorithm reminiscent of Algorithm 2, and the question arises as to whether one can immediately translate these regret results into sample complexity bounds? However, regret and sample complexity results do not directly follow from one another (Dann, Lattimore, and Brunskill 2017), and following the (Efroni, Mannor, and Pirotta 2020) approach gives a PAC result $\tilde{O}(\frac{|S|^2|A|H^4}{\epsilon^2})$, which is looser than our result by a factor of H^2. Thus, this alternative option does not provide the strong bounds that we are able to obtain to match existing PAC results of the unconstrained case.

Now, we introduce the notions of *knownness* and *importance* for state-action pairs and base our proof on these notions. Then we present the key lemmas required to prove Theorem 2. Finally, we sketch the proof of Theorem 2. The detailed analysis is provided in (HasanzadeZonuzy, Kalathil, and Shakkottai 2020).

Let the *weight* of $(s,a)-$pair in an episode k under policy $\tilde{\pi}_k$ be its expected frequency in that episode

$$w_k(s,a) := \sum_{h=0}^{H-1} \mathbb{P}(s_h = s, a \sim \tilde{\pi}_k(s_h, \cdot, h))$$

$$= \sum_{h=0}^{H-1} P_{\tilde{\pi}_k}^{h-1} \mathbb{I}\{s = \cdot, a \sim \tilde{\pi}_k(s, \cdot, h)\}(s_0).$$

Then, the *importance* ι_k of (s, a) at episode k is defined as its relative weight compared to $w_{\min} := \frac{\epsilon}{4H|S|}$ on a log-scale

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

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

$$\kappa_k(s, a) := \max\{z_i : z_i \leq \frac{n_k(s, a)}{m w_k(s, a)}\} \in \{0, 1, 2, 4, \dots\},$$

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

$$X_{k,\kappa,\iota} := \{(s, a) \in X_k : \kappa_k(s, a) = \kappa, \iota_k(s, a) = \iota\}$$
$$\text{and } \; \bar{X}_k = S \times A \setminus X_k,$$

where $X_k = \{(s, a) : \iota_k(s, a) > 0\}$ is the active set and $\bar{X}_k$ is the set of $(s, a)-$pairs that are very unlikely under policy $\tilde{\pi}_k$. We will show that if $|X_{k,\kappa,\iota}| \leq \kappa$ is satisfied, then the model of Online-CRL would achieve near-optimality while violating constraints at most by ϵ w.h.p. This condition indicates that important state-action pairs under policy $\tilde{\pi}_k$ are visited a sufficiently large number of times. Hence, the model of Online-CRL will be accurate enough to obtain PAC bounds.

Now, first we show that true model belongs to $\mathcal{M}_k$ for every episode k w.h.p.

Lemma 3. *$M \in \mathcal{M}_k$ for all episodes k with probability at least $1 - \frac{\delta}{2(N+1)}$.*

Proof Sketch: Fix a (s, a), next state s' and an episode k. Then, $P(s'|s, a)$ lies inside the confidence set constructed by the combined Bernstein's and Hoeffding's inequalities. Taking the union bound over maximum number of model updates, $U_{\max}$, and next states would yield the result. □

Next, we bound the number of episodes that the condition $|X_{k,\kappa,\iota}| \leq \kappa$ is violated w.h.p.

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

$$m \geq \frac{6H^2}{\epsilon} \log \frac{2(N+1)E_{\max}}{\delta}, \tag{9}$$

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

Proof sketch: The proof of this lemma is divided into two stages. First, we provide a bound on the total number of times a fixed (s, a) could be observed in a particular $X_{k,\kappa,\iota}$ in all episodes. Then, we present a high probability bound on the number of episodes that $|X_{k,\kappa,\iota}| > \kappa$ for a fixed (κ, ι). Finally, we obtain the result by means of martingale concentration and union bound. □

Finally, the next lemma provides a bound on the mismatch between objective and constraint functions of the optimistic model and true model. The role of this lemma is similar to Lemma 2 for Optimistic-GMBL. It provides a PAC result, which is uniform over value and constraint functions. Hence, it is possible to have individual PAC results for any objective and constraint functions. As discussed in the context of Optimistic-GMBL, this process is responsible for a $\log N$ increase in the sample complexity result.

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

$$m = 1280 \frac{|S|H^2}{\epsilon^2} (\log_2 \log_2 H)^2 \log_2^2\Big(\frac{8|S|^2H^2}{\epsilon}\Big) \log \frac{4}{\delta_1}, \tag{10}$$

then $|\tilde{V}_0^{\tilde{\pi}_k}(s_0) - V_0^{\tilde{\pi}_k}(s_0)| \leq \epsilon$ and for any i, $|\tilde{C}_{i,0}^{\tilde{\pi}_k}(s_0) - C_{i,0}^{\tilde{\pi}_k}(s_0)| \leq \epsilon$.

Proof Sketch: We first use algebraic operations to obtain $|\tilde{P}(s'|s, a) - P(s'|s, a)| \leq O(\sqrt{\frac{P(s'|s,a)(1-P(s'|s,a))}{n}})$ for each s', s, a. Then we show that at each time-step h, $(P_\pi - \tilde{P}_\pi)V_h^\pi(s) \leq O(\sqrt{\frac{|S|}{n}}\sigma_h^\pi(s))$. Then we divide the state-action based on knownness, i.e., whether they belong to X_k or not. By applying all bounds and using the fact that $\sigma_h^\pi(s)$ is close to $\tilde{\sigma}_h^\pi(s)$ by $\frac{\sqrt{|S|H^2}}{n^{1/4}}$, we obtain a bound on $|\tilde{V}_0^\pi(s_0) - V_0^\pi(s_0)|$. 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*:** First, we apply Lemma 3 and show that $M \in \mathcal{M}_k$ for every k w.p. at least $1 - \frac{\delta}{2(N+1)}$. Therefore, the optimistic planning problem would be feasible and an optimistic policy $\tilde{\pi}_k$ exists w.h.p. Furthermore, we bound the number of episodes where $|X_{k,\kappa,\iota}| > \kappa$ w.h.p. by means of Lemma 4. Thus, for other episodes where $|X_{k,\kappa,\iota}| \leq \kappa$, we show that objective function is $\epsilon-$optimal and all constraint functions are violated by ϵ by applying Lemma 5. Eventually, taking union bound yields the result. □

Experimental Results

We conduct experiments on CMDPs akin to a grid world MDP, wherein each square indicates the location of the agent. The goal of the is to start at the fixed start state and reach the final state in H steps. The agent obtains a reward of 1 when reaching the goal. Transitions are stochastic, and given any action, there is probability of self and other transitions, as well as transitioning to other state as intended by the action. We consider two classes of CMDPs under this setting, namely, (i) state occupancy constraints, and (ii) action frequency constraints, which represent the types of constraints that might appear in real systems.

For the first scenario class, we augment the unconstrained MDP by an action budget constraint. We restrict the number

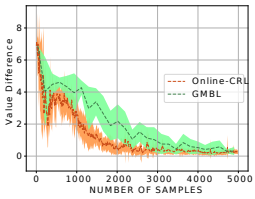

Figure 1: Value Difference-Scenario 1a

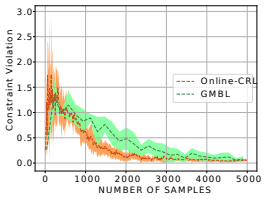

Figure 2: Constraint Violation-Scenario 1a

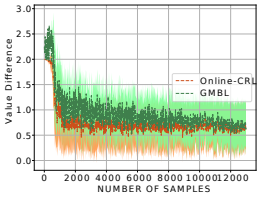

Figure 3: Value Difference Scenario 1b

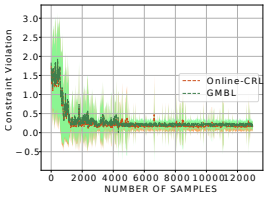

Figure 4: Constraint Violation-Scenario 1b

of moves to the right, while ensuring that a feasible path to the goal exists. Here, we consider a 3×3 and 5×5 grid as examples, with 9 state states and 25 states respectively, and with 4 actions. The 3×3 and 5×5 examples are labeled as scenario 1a and scenario 1b.

In the second scenario class, we consider a 3×3 grid world with a particular state is "bad" for the CMDP, so the agent must avoid entering it frequently or at all. The bad state has higher probability of transitioning out of itself compared to the rest of the states. But, if the agent enters this state, a cost is levied. Thus, the constraint is to limit the probability of entering the bad state, and to set the constraint threshold to 0. This means that the optimal policy for CMDP is to avoid the bad state altogether. This process is equivalent to incurring an immediate cost of 1 when the agent finds itself in the bad state.

We simulate Optimistic-GMBL and Online-CRL for these scenarios. Here, we consider two performance metrics. One, difference in value function calculated by

$$V_0^{\pi^*}(s_0) - V_0^{\pi'}(s_0).$$

where π' is whether Optimistic-GMBL or Online-CRL. The second performance metric is constraint violation which is calculated by

$$\max(C_0^{\pi'}(s_0) - \bar{C}, 0).$$

since we have one constraint in each scenario. Further, we average each data point on every figure over 25 runs.

As seen in the Figures 1, 3 and 5, both Optimistic-GMBL and Online-CRL reach the optimal values in both scenarios. We observe that the Online-CRL algorithm, despite having fewer number of samples, does consistently better than the Optimistic-GMBL algorithm in both the scenarios. Similar behavior appears in figures 2, 4, and 6, which illustrates constraint violation. Intuitively, Online-CRL outperforms

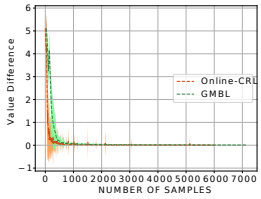

Figure 5: Value Difference-Scenario 2

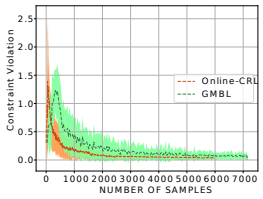

Figure 6: Constraint Violation-Scenario 2

Optimistic-GMBL empirically because it samples the important state-action pairs often, and hence resolves uncertainty quickly.

Conclusion

This paper introduced the notion of sample complexity in objective maximization and constraint satisfaction for understanding the performance of RL algorithms for safety-constrained applications. We developed two types of algorithms—Optimistic-GMBL and online-CRL. The main finding of a logarithmic factor increase in sample complexity over the unconstrained regime suggests value of the approach to real systems.

Broader Impact

Reinforcement learning has shown great success in domains that are action constrained, such as robotics, but less so on systems that are safety constrained in terms of the occupancy measure generated by the policy employed. These include a variety of cyber-physical systems (CPS) such as the power grid, and other utilities, where guarantees on the operating region of the system must be met—ideally deterministically, but within some bounds with high probability in practice.

It is in the space of control of such CPS that our work is applicable, and could potentially have an impact on a wide variety of supervisory control and data acquisition (SCADA) systems. Many of them already employ empirically determined policies validated through large scale simulations, and it is not hard to visualize them as being driven by RL-based policies. Sample complexity bounds reveal how much information is needed to obtain what level of guarantee of safe operability, and hence are a way of determining if a policy has been well enough trained to be actually used.

However, a note of caution with this 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 violate basic truths known to human operators and could fail quite badly. Indeed, our approach does not provide sample-path constraints, 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.

Acknowledgments

This work was supported in part by the National Science Foundation grants CRII-CPS-1850206, NECCS-EPCN-1839616, CNS-1955696, ECCS-1839816, CNS1719384, CPS-2038963, and Army Research Office grant W911NF-19-1-0367.

References

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

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

Amani, S.; Alizadeh, M.; and Thrampoulidis, C. 2019. Linear stochastic bandits under safety constraints. In *Advances in Neural Information Processing Systems*, 9252–9262.

Azar, M. G.; Munos, R.; and Kappen, H. J. 2013. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning* 91(3): 325–349.

Badanidiyuru, A.; Kleinberg, R.; and Slivkins, A. 2013. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, 207–216. IEEE.

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

Borkar, V.; and Jain, R. 2014. Risk-constrained Markov decision processes. *IEEE Transactions on Automatic Control* 59(9): 2574–2579.

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

Chow, Y.; Nachum, O.; Duenez-Guzman, E.; and Ghavamzadeh, M. 2018. A lyapunov-based approach to safe reinforcement learning. In *Advances in Neural Information Processing Systems*, 8092–8101.

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

Dann, C.; Lattimore, T.; and Brunskill, E. 2017. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, 5713–5723.

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

HasanzadeZonuzy, A.; Kalathil, D.; and Shakkottai, S. 2020. Learning with Safety Constraints: Sample Complexity of Reinforcement Learning for Constrained MDPs. *arXiv preprint arXiv:2008.00311* .

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

Isom, J. D.; Meyn, S. P.; and Braatz, R. D. 2008. Piecewise Linear Dynamic Programming for Constrained POMDPs. In *AAAI*, volume 1, 291–296.

Kim, D.; Lee, J.; Kim, K.-E.; and Poupart, P. 2011. Point-based value iteration for constrained POMDPs. In *IJCAI*, 1968–1974.

Lattimore, T.; and Hutter, M. 2014. Near-optimal PAC bounds for discounted MDPs. *Theoretical Computer Science* 558: 125–143.

Liang, Q.; Que, F.; and Modiano, E. 2018. Accelerated primal-dual policy optimization for safe reinforcement learning. *arXiv preprint arXiv:1802.06480* .

Liu, Y.; Ding, J.; and Liu, X. 2019. IPO: Interior-point policy optimization under constraints. *arXiv preprint arXiv:1910.09615* .

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

Paternain, S.; Chamon, L.; Calvo-Fullana, M.; and Ribeiro, A. 2019. Constrained Reinforcement Learning Has Zero Duality Gap. In *Advances in Neural Information Processing Systems*, 7553–7563.

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

Satija, H.; Amortila, P.; and Pineau, J. 2020. Constrained Markov Decision Processes via Backward Value Functions. *arXiv preprint arXiv:2008.11811* .

Singh, R.; Hou, I.-H.; and Kumar, P. 2014. Fluctuation analysis of debt based policies for wireless networks with hard delay constraints. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, 2400–2408. IEEE.

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

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

Tessler, C.; Mankowitz, D. J.; and Mannor, S. 2018. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074* .

Wachi, A.; and Sui, Y. 2020. Safe Reinforcement Learning in Constrained Markov Decision Processes. *arXiv preprint arXiv:2008.06626* .

Wu, H.; Srikant, R.; Liu, X.; and Jiang, C. 2015. Algorithms with logarithmic or sublinear regret for constrained contextual bandits. In *Advances in Neural Information Processing Systems*, 433–441.

Zheng, L.; and Ratliff, L. J. 2020. Constrained upper confidence reinforcement learning. *arXiv preprint arXiv:2001.09377* .