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

Aria HasanzadeZonuzy, Archana Bura, Dileep Kalathil and Srinivas Shakkottai [*]

**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.

## 1 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 [1].

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.

[*]Authors are with the Department of Electrical and Computer Engineering, Texas A& M University, Texas, USA. Email:{azonuzy, dileep.kalathil, sshakkot}@tamu.edu

**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 [2], 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 [3,4], we discover that the increase 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 [1,5–9]. 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 [10,11]. 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 [12–15] under Lagrangian methods to show zero eventual duality gap. [16] 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 [17].

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 [18–20] 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, [21] and [22] present results in the context of unknown reward functions, with either a known stochastic or deterministic transition kernel. Other work [23] focuses on asymptotic convergence, and so does not provide an estimate on the learning rate. Finally, [2] 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 [24], 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.

## 2 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 $\pi$ 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 $\pi$ while being at state $s$ at time-step $h$.

When policy $\pi$ is fixed, the underlying Markov Decision Process turns into a Markov chain. The transition kernel of this Markov chain is $P_\pi$, 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 $P_\pi^{-1}$ and $P_\pi^0$ 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 $\pi$ as

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

where action $a_h$ is chosen according to policy $\pi$ 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 $\pi$ is

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

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

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

Again, the local variance of $i^{th}$ constraint function at time-step $h$ under policy $\pi$, i.e. $\sigma_{i,h}^{\pi^2}$ is defined similar to local variance of value function (2).

Finally, the general finite-horizon CMDP problem is

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

**Assumption 2.** *We assume that there exists some policy $\pi$ that satisfies the constraints in* (4)*. Hence, this CMDP problem is feasible with optimal policy $\pi^*$ 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*

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

*for a given $\epsilon$ and $\delta$.*

Note that this definition includes both objective maximization and constraint violations, as opposed to a traditional definition that only considers the objective [25].

# 3 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

$$\begin{aligned}&\mathcal{M}_{\delta_P} := \{M' : r'(s,a) = r(s,a), &(5)\\ &c'(i,s,a) = c(i,s,a), H' = H, s'_0 = s_0\\ &|P'(s'|s,a) - \widehat{P}(s'|s,a)| \leq &(6)\\ &\min\Big(\sqrt{\frac{2\widehat{P}(s'|s,a)(1-\widehat{P}(s'|s,a))}{n}\log\frac{4}{\delta_P}} + \frac{2}{3n}\log\frac{4}{\delta_P},\\ &\sqrt{\frac{\log 4/\delta_P}{2n}}\Big)\forall s, a, s', i\},\end{aligned}$$

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

[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.

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

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

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 Appendix 8 . Algorithm 1 describes Optimistic-GMBL.

**Algorithm 1** Optimistic-GMBL

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

### 3.1 PAC Analysis of Optimistic-GMBL

Here, we present the sample complexity result of Optimistic-GMBL. Time complexity result and analysis will be provided in Appendix 8.

**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*

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

*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 [3] in the following manner. First, is 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 [26]. However, there is no such guarantee for any general CMDP problem formulation [1]. 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 [1]. 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 Appendix 8.

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 [27] and empirical Bernstein's inequality [28], 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 $\pi$*

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

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

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

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

# 4 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 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 $k$ :

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

where $\delta_1$ 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:

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

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 $C'^{\pi}_{i,0}(s_0)$ 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$-\gamma$ [29] and its finite-horizon counterpart UCFH [4] with several differences. Unlike UCRL-$\gamma$ and UCFH, Algorithm 2 updates the model at the beginning of each episode, which allows for faster model construction. Also,

**Algorithm 2** Online-CRL

```
Input: accuracy $\epsilon$ and failure tolerance $\delta$.
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
        $a_t \sim \tilde{\pi}_k(s_t), s_{t+1} \sim P(\cdot|s_t, a_t), n(s_t, a_t) + +, n(s_{t+1}, s_t, a_t) + +$.
    $k + +$
```

since we desire a policy that pertains to a CMDP using an linear programming approach [1], 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.

### 4.1 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*

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

*episodes.*

To prove Theorem 2, we follow an approach motivated by [29] and its finite-horizon version [4]. 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 [2] 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 [24], and following the [2] 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 Appendix 8.

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

$$\begin{aligned} w_k(s,a) &:= \sum_{h=0}^{H-1} \mathbb{P}(s_h = s, a \sim \tilde{\pi}_k(s_h,\cdot,h)) \\ &= \sum_{h=0}^{H-1} P_{\tilde{\pi}_k}^{h-1} \mathbb{I}\{s=\cdot, a \sim \tilde{\pi}_k(s,\cdot,h)\}(s_0). \end{aligned}$$

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

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

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

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

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 $\epsilon$ 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 $\kappa$ and $\iota$ 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 $(\kappa, \iota)$. 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 $(\kappa, \iota)$ 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 $\epsilon$ by applying Lemma 5. Eventually, taking union bound yields the result. □

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

In the second scenario class, we consider a $3 \times 3$ 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

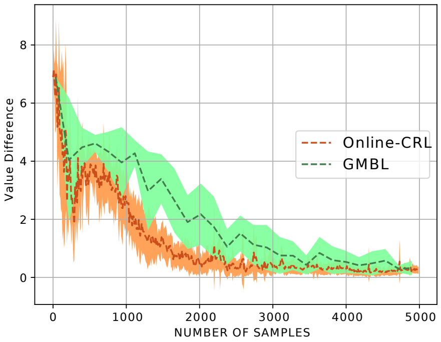

Figure 1: Value Difference for Scenario 1a

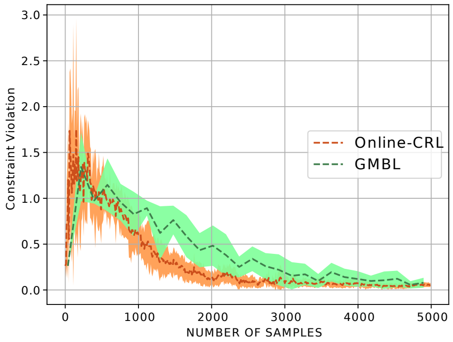

Figure 2: Constraint Violation for Scenario 1a

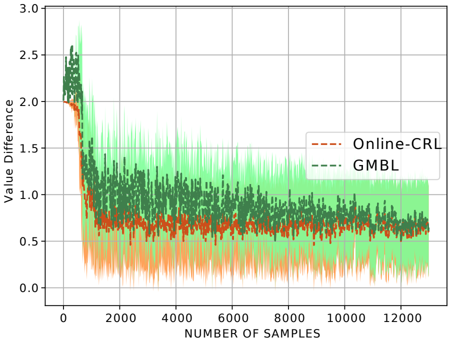

Figure 3: Value Difference for Scenario 1b

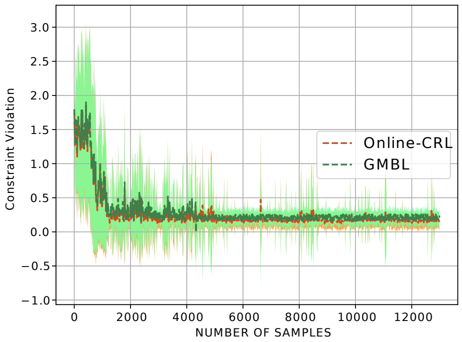

Figure 4: Constraint Violation for Scenario 1b

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

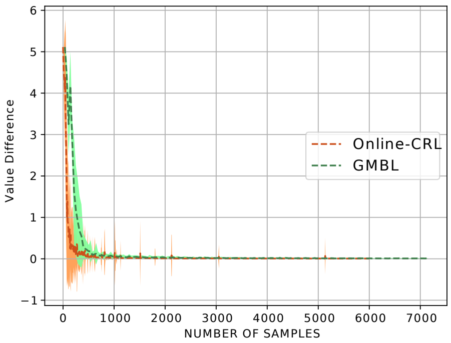

Figure 5: Value Difference for Scenario 2

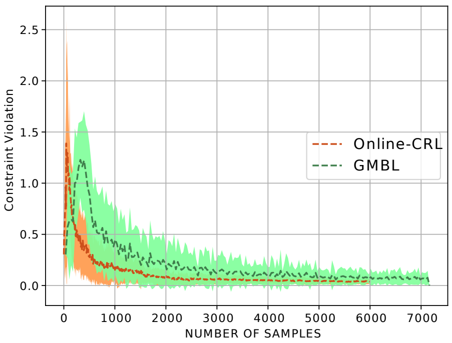

Figure 6: Constraint Violation for Scenario 2

# 6 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.

# 7 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.

# References

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

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

[3] 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.

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

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

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

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

[8] 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.

[9] 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.

[10] Joshua D Isom, Sean P Meyn, and Richard D Braatz. Piecewise linear dynamic programming for constrained pomdps. In *AAAI*, volume 1, pages 291–296, 2008.

[11] Dongho Kim, Jaesong Lee, Kee-Eung Kim, and Pascal Poupart. Point-based value iteration for constrained pomdps. In *IJCAI*, pages 1968–1974, 2011.

[12] 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.

[13] 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.

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

[15] 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.

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

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

[18] 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.

[19] 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.

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

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

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

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

[24] 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.

[25] 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.

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

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

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

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

[30] Fan Chung and Linyuan Lu. Concentration inequalities and martingale inequalities: a survey. *Internet Mathematics*, 3(1):79–127, 2006.

# 8 Technical Appendix

## 8.1 Extended-Linear Programming

ELP is a Linear Programming, LP, formulation indeed. So, we first present generic LP which is used to solve CMDP problem of (4) [1], then build the idea of ELP based on that. To solve CMDP problem (4) via LP approach, we convert this problem to a linear programming problem formulated using new variables occupation measures. Now, consider $\mu$ as the finite-horizon state-action occupation measure under policy $\pi$ defined as

$$\mu(s,a,\pi,h) := \mathbb{P}(s_h = s, a_h = a | s_{h=0} = s_0), \tag{11}$$

where the probability is calculated w.r.t. underlying transition kernel under policy $\pi, P_\pi$. It is shown that objective function and constraint functions could be restated as functions of occupation measures. Then, the problem would become to find the optimal occupation measures.

Now, if we let $\mu$ be any generic occupation measure defined as (11), then the equivalent LP to CMDP problem (4) is

$$\begin{aligned}
\max_{\mu} & \sum_{s,a,h} \mu(s,a,h) r(s,a) \\
& \text{s.t.} \\
& \sum_{s,a,h} \mu(s,a,h) c(i,s,a) \le \bar{C}_i \quad \forall i, \\
& \sum_{a} \mu(s,a,h) = \sum_{s',a'} P(s|s',a') \mu(s',a',h-1) \ \ \forall h \in \{1,\dots,H-1\}, \\
& \sum_{a} \mu(s_0,a,0) = 1, \quad \sum_{a} \mu(s,a,0) = 0 \ \ \forall s \in S \backslash \{s_0\}, \\
& \mu(s,a,h) \ge 0 \ \ \forall s,a,h
\end{aligned} \tag{12}$$

It is proved that the LP (12) is equivalent to CMDP problem of (4), and the optimal policy computed by this LP is also the solution to CMDP problem in [1]. Eventually, the optimal policy $\pi^*$ is calculated as follows

$$\pi^*(s,a,h) = \frac{\mu(s,a,h)}{\sum_b \mu(s,b,h)}.$$

Now, given the estimated model $\widehat{P}$, we get the ELP formulation if we define new occupancy measure $q(s,a,s',h) = P(s'|s,a)\mu(s,a,h)$. Eventually, the ELP formulation is

$$\begin{aligned}
\max_{q} & \sum_{s,a,s',h} q(s,a,s',h) r(s,a) \\
& \text{s.t.} \\
& \sum_{s,a,s',h} q(s,a,s',h) c(i,s,a) \le \bar{C}_i \quad \forall i \in \{1,\dots,N\}, \\
& \sum_{a,s'} q(s,a,s',h) = \sum_{s',a'} q(s',a',s,h-1) \ \ \forall h \in \{1,\dots,H-1\}, \\
& \sum_{a,s'} q(s_0,a,s',0) = 1, \quad \sum_{a,s'} q(s,a,s',0) = 0 \ \ \forall s \in S \backslash \{s_0\}, \\
& q(s,a,s',h) \ge 0 \ \ \forall s,s' \in S, a \in A, h \in \{0,1,\dots,H-1\},
\end{aligned}$$

$$q(s,a,s',h) - (\widehat{P}(s'|s,a) + \beta(s,a,s')) \sum_{y} q(s,a,y,h) \leq 0 \;\; \forall s,a,s',h,$$

$$-\, q(s,a,s',h) + (\widehat{P}(s'|s,a) - \beta(s,a,s')) \sum_{y} q(s,a,y,h) \leq 0 \;\; \forall s,a,s',h,$$

where $\beta(s,a,s')$ is the radius of the confidence interval around $\widehat{P}(s'|s,a)$ which depends on the algorithm. The last two conditions in the above formulation include the confidence interval around $\widehat{P}$ and distinguish ELP from generic LP formulation. At the end, ELP outputs the optimistic policy, $\tilde{\pi}$ for Optimistic-GMBL and $\tilde{\pi}_k$ for Online-CRL, using the solution of above LP. Also, we can calculate an optimistic transition kernel denoted by $\tilde{P}$ by means of optimal $q(s,a,s',h)$. In brief, the optimistic transition kernel and optimistic policy are computed as follows

$$\tilde{P}(s'|s,a) = \frac{q(s,a,s',h,s_0)}{\sum_b q(s,a,b,h,s_0)}, \quad \tilde{\pi}^*(s,a,h) = \frac{\sum_{s'} q(s,a,s',h,s_0)}{\sum_{b,s'} q(s,b,s',h,s_0)}.$$

The details of ELP about the time and space complexity is briefed in [2], so we do not present them here.

## 8.2 Detailed Proofs for Upper PAC Bounds in Offline Mode

In this section, we assume that we have $n$ samples from each $(s,a)$ in every lemma presented.

***Proof of Lemma 1*:** Fix a state, action and next state, i.e. $s,a,s'$. Then, according to Hoeffding's inequality [27]

$$\mathbb{P}(|P(s'|s,a) - \widehat{P}(s'|s,a)| \leq \sqrt{\frac{\log 4/\delta_P}{2n}}) \geq 1 - \delta_P/2.$$

Now, we apply empirical Bernstein's inequality [28] and get

$$\mathbb{P}(|P(s'|s,a) - \widehat{P}(s'|s,a)| \leq \sqrt{\frac{2\widehat{P}(s'|s,a)(1-\widehat{P}(s'|s,a))}{n}\log\frac{4}{\delta_P}} + \frac{2}{3n}\log\frac{4}{\delta_P}) \geq 1 - \delta_P/2.$$

By combining these two inequalities and applying union bound, we get

$$\mathbb{P}(|P(s'|s,a) - \widehat{P}(s'|s,a)| \leq \min\{\sqrt{\frac{2\widehat{P}(s'|s,a)(1-\widehat{P}(s'|s,a))}{n}\log\frac{4}{\delta_P}} + \frac{2}{3n}\log\frac{4}{\delta_P}, \sqrt{\frac{\log 4/\delta_P}{2n}}\}) \geq 1 - \delta_P.$$

Finally, we get the result by applying union bound over all state, action and next states. □

**Lemma 6.** *Let $\delta_P \in (0,1)$. Assume $p, \widehat{p}, \tilde{p} \in [0,1]$ satisfy $\mathbb{P}(p \in \mathcal{P}_{\delta_P}) \geq 1 - \delta_P$ and $\tilde{p} \in \mathcal{P}_{\delta_P}$ where*

$$\mathcal{P}_{\delta_P} := \{p' \in [0,1] : |p' - \widehat{p}| \leq \min\Big(\sqrt{\frac{2\widehat{p}(1-\widehat{p})}{n}}\log 4/\delta_P + \frac{2}{3n}\log 4/\delta_P, \sqrt{\frac{\log 4/\delta_P}{2n}}\Big)\}.$$

*Then,*

$$|p - \tilde{p}| \leq \sqrt{\frac{8\tilde{p}(1-\tilde{p})}{n}\log 4/\delta_P} + 2\sqrt{2}\Big(\frac{\log 4/\delta_P}{n}\Big)^{\frac{3}{4}} + 3\sqrt{2}\frac{\log 4/\delta_P}{n}$$

*w.p. at least $1 - \delta_P$.*

*Proof.*

$$\begin{aligned}|p-\tilde{p}| \leq |p-\widehat{p}| + |\widehat{p}-\tilde{p}| &\leq 2\sqrt{\frac{2\widehat{p}(1-\widehat{p})}{n}\log 4/\delta_P} + \frac{4}{3n}\log 4/\delta_P \\ &\leq 2\sqrt{\frac{2\log 4/\delta_P}{n}(\tilde{p}+\sqrt{\frac{\log 4/\delta_P}{2n}})(1-\tilde{p}+\sqrt{\frac{\log 4/\delta_P}{2n}})} + \frac{4}{3n}\log 4/\delta_P \\ &= 2\sqrt{\frac{2\log 4/\delta_P}{n}\Big(\tilde{p}(1-\tilde{p})+\sqrt{\frac{\log 4/\delta_P}{2n}}+\frac{\log 4/\delta_P}{2n}\Big)} + \frac{4}{3n}\log 4/\delta_P \\ &\leq \sqrt{\frac{8\tilde{p}(1-\tilde{p})}{n}\log 4/\delta_P} + 2\sqrt{2}\Big(\frac{\log 4/\delta_P}{n}\Big)^{\frac{3}{4}} + 3\sqrt{2}\frac{\log 4/\delta_P}{n}.\end{aligned}$$

The first term in the first line is true w.p. at least $1-\delta_P$, hence the proof is complete. □

**Lemma 7.** *Suppose there are two CMDPs $M = \langle S, A, P, r, c, \bar{C}, s_0, H\rangle$ and $M' = \langle S, A, P', r, c, \bar{C}, s_0, H\rangle$ satisfying assumption 1. Then, under any policy $\pi$*

$$V_0^{\pi} - V_0'^{\pi} = \sum_{h=0}^{H-2} P_{\pi}'^{h-1}(P_{\pi} - P'_{\pi})V_{h+1}^{\pi} \textit{ and } V_0^{\pi} - V_0'^{\pi} = \sum_{h=0}^{H-2} P_{\pi}^{h-1}(P_{\pi} - P'_{\pi})V_{h+1}'^{\pi},$$

*and for any $i \in \{1, \dots, N\}$,*

$$C_{i,0}^{\pi} - C_{i,0}'^{\pi} = \sum_{h=0}^{H-2} P_{\pi}'^{h-1}(P_{\pi} - P'_{\pi})C_{i,h+1}^{\pi} \textit{ and } C_{i,0}^{\pi} - C_{i,0}'^{\pi} = \sum_{h=0}^{H-2} P_{\pi}^{h-1}(P_{\pi} - P'_{\pi})C_{i,h+1}'^{\pi}.$$

*Proof.* We only prove the first statement of value function since the proof procedure for cost is identical. For a fixed $h$ and $s$

$$\begin{aligned}&V_h^{\pi}(s) - V_h'^{\pi}(s) = r_{\pi}(s) + \sum_{s'} P_{\pi}(s'|s)V_{h+1}^{\pi}(s') - (r_{\pi}(s) + \sum_{s'} P'_{\pi}(s'|s)V_{h+1}'^{\pi}(s')) \\ &= \sum_{s'} P_{\pi}(s'|s)V_{h+1}^{\pi}(s') - \sum_{s'} P'_{\pi}(s'|s)V_{h+1}^{\pi}(s') + \sum_{s'} P'_{\pi}(s'|s)V_{h+1}^{\pi}(s') - \sum_{s'} P'_{\pi}(s'|s)V_{h+1}'^{\pi}(s') \\ &= \sum_{s'}(P_{\pi}(s'|s) - P'_{\pi}(s'|s))V_{h+1}^{\pi}(s') + \sum_{s'} P'_{\pi}(s'|s)(V_{h+1}^{\pi}(s') - V_{h+1}'^{\pi}(s')).\end{aligned}$$

Because $V_{H-1}^{\pi}(s) = V_{H-1}'^{\pi}(s) = r_{\pi}(s)$, if we expand the second term until $h = H-1$, we get the result. □

**Lemma 8.** *Let $\delta_P \in (0,1)$. Suppose there are two CMDPs $M = \langle S, A, P, r, c, \bar{C}, s_0, H\rangle$ and $M' = \langle S, A, P', r, c, \bar{C}, s_0, H\rangle$ satisfying assumption 1. Further assume*

$$|P(s'|s,a) - P'(s'|s,a)| \leq c_1 + c_2\sqrt{P'(s'|s,a) - (1 - P'(s'|s,a))}$$

*w.p. at least $1-\delta_P$ for each $s, s' \in S, a \in A$. Then, under any policy $\pi$*

$$|\sum_{s'}(P_{\pi}(s'|s) - P'_{\pi}(s'|s))V_{h+1}'^{\pi}(s')| \leq |S|c_1\|V_{h+1}'^{\pi}\|_{\infty} + c_2\sqrt{|S|}\sigma_h'^{\pi}(s)$$

*for any $(s,a) \in S \times A$ and $h \in [0, H-2]$ w.p. at least $1 - |S|\delta_P$, and*

$$|\sum_{s'}(P_{\pi}(s'|s) - P'_{\pi}(s'|s))C_{i,h+1}'^{\pi}(s')| \leq |S|c_1\|C_{i,h+1}'^{\pi}\|_{\infty} + c_2\sqrt{|S|}\sigma_{i,h}'^{\pi}(s)$$

*for any $(s,a) \in S \times A, i \in \{1, \dots, N\}$ and $h \in [0, H-2]$ w.p. at least $1 - |S|\delta_P$.*

*Proof.* We only prove the statement of value function since the proof procedure for cost is identical. Fix state $s$ and define for this fixed state $s$ the constant function $\bar{V}^\pi(s') = \sum_{s''} P'_\pi(s''|s)V'^\pi_{h+1}(s'')$ as the expected value function of the successor states of $s$. Note that $\bar{V}^\pi(s')$ is a constant function and so $\bar{V}^\pi(s') = \sum_{s''} P'_\pi(s''|s)\bar{V}^\pi(s'') = \sum_{s''} P_\pi(s''|s)\bar{V}^\pi(s'')$.

$$\begin{aligned}
&|\sum_{s'}(P_\pi(s'|s) - P'_\pi(s'|s))V'^\pi_{h+1}(s')| = |\sum_{s'}(P_\pi(s'|s) - P'_\pi(s'|s))V'^\pi_{h+1}(s') + \bar{V}^\pi(s) - \bar{V}^\pi(s)| \\
&= |\sum_{s'}(P_\pi(s'|s) - P'_\pi(s'|s))(V'^\pi_{h+1}(s') - \bar{V}^\pi(s'))| \\
&\leq \sum_{s'} |P_\pi(s'|s) - P'_\pi(s'|s)||V^\pi_{h+1}(s') - \bar{V}^\pi(s')| \qquad (13) \\
&\leq \sum_{s'}(c_1 + c_2\sqrt{P'_\pi(s'|s) - (1 - P'_\pi(s'|s))})|V^\pi_{h+1}(s') - \bar{V}^\pi(s')| \\
&\leq |S|c_1\|V'^\pi_{h+1}\|_\infty + c_2\sum_{s'}\sqrt{P'_\pi(s'|s)(1 - P'_\pi(s'|s))(V^\pi_{h+1}(s') - \bar{V}^\pi(s'))^2} \\
&\leq |S|c_1\|V'^\pi_{h+1}\|_\infty + c_2\sqrt{|S|\sum_{s'}P'_\pi(s'|s)(1 - P'_\pi(s'|s))(V^\pi_{h+1}(s') - \bar{V}^\pi(s'))^2} \qquad (14) \\
&\leq |S|c_1\|V'^\pi_{h+1}\|_\infty + c_2\sqrt{|S|\sum_{s'}P'_\pi(s'|s)(V^\pi_{h+1}(s') - \bar{V}^\pi(s'))^2} \\
&= |S|c_1\|V'^\pi_{h+1}\|_\infty + c_2\sqrt{|S|}\sigma'^\pi_h.
\end{aligned}$$

Inequality (13) holds w.p. at least $1 - |S|\delta_P$, since we used the assumption and applied the triangle inequality and union bound. We then applied the assumed bound on $|V'^\pi_{h+1}(s') - \bar{V}^\pi(s')|$ and bounded it by $\|V'^\pi_{h+1}\|_\infty$ as all value functions are non-negative. In inequality (14), we applied the Cauchy-Schwarz inequality and subsequently used the fact that each term is the sum is non-negative and that $(1 - P'_\pi(s'|s)) \leq 1$. The final equality follows from the definition of $\sigma'^\pi_h(s)$. □

**Lemma 9.** *Let $\delta_P \in (0, 1)$. Suppose there are two CMDPs $M = \langle S, A, P, r, c, \bar{C}, s_0, H\rangle$ and $M' = \langle S, A, P', r, c, \bar{C}, s_0, H\rangle$ satisfying assumption 1. Further assume*

$$|P(s'|s, a) - P'(s'|s, a)| \leq \frac{a}{\sqrt{n}}$$

*for all $s, s' \in S, a \in A$ w.p. at least $1 - \delta_P$. Then, under any policy $\pi$*

$$\|V^\pi_{H-1} - V'^\pi_{H-1}\|_\infty \leq \cdots \leq \|V^\pi_0 - V'^\pi_0\|_\infty \leq |S|H^2a\frac{1}{\sqrt{n}},$$

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

$$\|C^\pi_{i,H-1} - C'^\pi_{i,H-1}\|_\infty \leq \cdots \leq \|C^\pi_{i,0} - C'^\pi_{i,0}\|_\infty \leq |S|H^2a\frac{1}{\sqrt{n}}$$

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

*Proof.* We prove the statement of value function since the proof procedure for cost is identical. Let $\Delta_h = \max_s |V^\pi_h(s) - V'^\pi_h(s)|$. Then

$$\Delta_h = |V^\pi_h(s) - V'^\pi_h(s)| = |r_\pi(s) + \sum_{s'} P_\pi(s'|s)V^\pi_{h+1}(s') - (r_\pi(s) + \sum_{s'} P'_\pi(s'|s)V'^\pi_{h+1}(s'))|$$

$$
\begin{aligned}
&= |\sum_{s'} P_\pi(s'|s)V^\pi_{h+1}(s') - \sum_{s'} P'_\pi(s'|s)V^\pi_{h+1}(s') + \sum_{s'} P'_\pi(s'|s)V^\pi_{h+1}(s') - \sum_{s'} P'_\pi(s'|s)V'^\pi_{h+1}(s')| \\
&\leq \sum_{s'} |(P_\pi(s'|s) - P'_\pi(s'|s)|H + \Delta_{h+1} \\
&\leq |S|Ha\frac{1}{\sqrt{n}} + \Delta_{h+1}.
\end{aligned}
$$

Thus,

$$\Delta_h \leq |S|Ha\frac{1}{\sqrt{n}} + \Delta_{h+1}$$

w.p. at least $1 - |S|^2|A|\delta_P$ by applying union bound over all current state, action and next state. If we expand this recursively, we get

$$\Delta_{H-1} = 0 \leq \cdots \leq \Delta_0 \leq |S|H^2a\frac{1}{\sqrt{n}}$$

since $\Delta_{H-1} = \max_s |r_\pi(s) - r_\pi(s)| = 0$. By taking union bound over time-steps, we get the result holds w.p. at least $1 - |S|^2|A|H\delta_P$. Hence the proof is complete. □

**Lemma 10.** *Let $\delta_P \in (0,1)$. Suppose there are two CMDPs $M = \langle S, A, P, r, c, \bar{C}, s_0, H\rangle$ and $M' = \langle S, A, P', r, c, \bar{C}, s_0, H\rangle$ satisfying assumption 1. Further assume*

$$|P(s'|s,a) - P'(s'|s,a)| \leq \frac{a}{\sqrt{n}}$$

*w.p. at least $1 - \delta_P$ for all $s, s' \in S, a \in A$. Then if $n \geq a|S|H^2$, at any time-step $h \in [0, H-1]$ and under any policy $\pi$*

$$\|\sigma^\pi_h - \sigma'^\pi_h\|_\infty \leq \frac{2\sqrt{2|S|H^2a}}{n^{1/4}},$$

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

$$\|\sigma^\pi_{i,h} - \sigma'^\pi_{i,h}\|_\infty \leq \frac{2\sqrt{2|S|H^2a}}{n^{1/4}}$$

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

*Proof.* We prove the statement of value function since the proof procedure for cost is identical. Fix a state $s$. Then,

$$
\begin{aligned}
\sigma^{\pi^2}_h(s) &= \sigma^{\pi^2}_h(s) - \mathbb{E}'[(V^\pi_{h+1}(s_{h+1}) - P'_\pi V^\pi_{h+1}(s))^2] + \mathbb{E}'[(V^\pi_{h+1}(s_{h+1}) - P'_\pi V^\pi_{h+1}(s))^2] \\
&\leq \sum_{s'}(P_\pi(s'|s) - P'_\pi(s'|s))V^{\pi^2}_{h+1}(s') - [(\sum_{s'} P_\pi(s'|s)V^\pi_{h+1}(s'))^2 - (\sum_{s'} P'_\pi(s'|s)V^\pi_{h+1}(s'))^2] \\
&\quad + [\sqrt{\mathbb{E}'[(V^\pi_{h+1}(s_{h+1}) - V'^\pi_1(s_1) - P'_\pi(V^\pi_{h+1} - V'^\pi_{h+1})(s))^2]} + \sqrt{\mathbb{E}'[(V'^\pi_{h+1}(s_{h+1}) - P'_\pi(V'^\pi_{h+1})(s))^2]}]^2,
\end{aligned}
$$

where we applied triangular inequality in the last line. And, please note that $\mathbb{E}'$ means expectation w.r.t. transition kernel $P'_\pi$. It is straightforward to show that $Var_{s'\sim P'_\pi(\cdot|s)}(V^\pi_h(s') - V'^\pi_h(s')) \leq \|V^\pi_h - V'^\pi_h\|^2_\infty$ implying

$$\sigma^{\pi^2}_h(s) \leq \sum_{s'}(P_\pi(s'|s) - P'_\pi(s'|s))V^{\pi^2}_{h+1}(s')$$

$$
\begin{aligned}
&- [\sum_{s'} (P_\pi(s'|s) - P'_\pi(s'|s)) V^\pi_{h+1}(s')][\sum_{s'} (P_\pi(s'|s) + P'_\pi(s'|s)) V^\pi_{h+1}(s')] \\
&+ (\|V^\pi_h - V'^\pi_h\|_\infty + \sigma'^\pi_h(s))^2
\end{aligned}
$$

w.p. at least $1 - |S|\delta_P$ Now, if we use Lemma 9, we get

$$
\begin{aligned}
\sigma^{\pi^2}_h(s) &\leq [\sigma^{'\pi}_h(s) + \frac{|S|H^2 a}{\sqrt{n}}]^2 + \frac{2|S|aH^2}{\sqrt{n}} \leq [\sigma^{'\pi}_h(s) + \frac{|S|H^2 a}{\sqrt{n}} + \frac{\sqrt{2|S|H^2 a}}{n^{1/4}}]^2 \\
&\leq [\sigma^{'\pi}_h(s) + \frac{2\sqrt{2|S|H^2 a}}{n^{1/4}}]^2,
\end{aligned}
$$

w.p. at least $1 - |S|^2|A|H\delta_P$.[2] In the last line, we used the fact that for any $x, y > 0$ we have $x^2 + y^2 \leq (x+y)^2$. And, the assumption on $n$, dominates the term with $\frac{1}{n^{1/4}}$ over $\sqrt{n}$. Eventually, the result follows by taking square root from both sides and union bound on both directions, i.e. $\sigma'^\pi_h(s) \leq \sigma^\pi_h(s) + \frac{2\sqrt{2|S|H^2 a}}{n^{1/4}}$.[3] □

**Lemma 11.** *[4] The variance of the value function defined as $\Sigma^\pi_t(s) = \mathbb{E}[(\sum_{h=t}^{H-1} r(s_h) - V^\pi_0(s))^2]$ satisfies a Bellman equation $\Sigma^\pi_t(s) = \sigma^{\pi^2}_t(s) + \sum_{s' \in S} P_\pi(s'|s) V^\pi_{t+1}(s')$ which gives $\Sigma^\pi_t(s) = \sum_{h=t}^{H} (P^{h-1}_\pi \sigma^{\pi^2}_h)(s)$. Since $0 \leq \Sigma^\pi_0(s) \leq H^2$, it follows that $0 \leq \sum_{h=0}^{H-1} (P^{h-1}_\pi \sigma^{\pi^2}_h)(s) \leq H^2$ for all $s \in S$.*

**Corollary 2.** *The result of Lemma 11 also holds for variance of cost functions.*

***Proof of Lemma 2:*** We only prove the statement of value function since the proof procedure for cost is identical. First, we apply Lemma 6 and get

$$
|P(s'|s,a) - \tilde{P}(s'|s,a)| \leq \sqrt{\frac{8P(\tilde{s'|}s,a)(1 - \tilde{P}(s'|s,a))}{n} \log 4/\delta_P} + 2\sqrt{2}\Big(\frac{\log 4/\delta_P}{n}\Big)^{\frac{3}{4}} + 3\sqrt{2}\frac{\log 4/\delta_P}{n}
$$

w.p. at least $1 - \delta_P$. So, let

$$
c_1 = 2\sqrt{2}\Big(\frac{\log 4/\delta_P}{n}\Big)^{\frac{3}{4}} + \frac{3\sqrt{2}\log 4/\delta_P}{n} \quad \text{and} \quad c_2 = \sqrt{\frac{8\log 4/\delta_P}{n}} \tag{15}
$$

Now, let fix state $s$ :

$$
\begin{aligned}
&|V^\pi_0(s) - \tilde{V}^\pi_0(s)| = |\sum_{h=0}^{H-2} \tilde{P}^{h-1}_\pi (P_\pi - \tilde{P}_\pi) V^\pi_{h+1}|(s) && (16) \\
&\leq \sum_{h=0}^{H-2} \tilde{P}^{h-1}_\pi |(P_\pi - \tilde{P}_\pi) V^\pi_{h+1}|(s) \leq \sum_{h=0}^{H-2} \tilde{P}^{h-1}_\pi (|S|c_1 \|V^\pi_{h+1}\|_\infty + c_2\sqrt{|S|}\sigma^\pi_h)(s) && (17) \\
&\leq |S|H^2 c_1 + c_2\sqrt{|S|} \sum_{h=0}^{H-1} (\tilde{P}^{h-1}_\pi \sigma^\pi_h)(s) && (18) \\
&\leq |S|H^2 c_1 + c_2\sqrt{|S|} \sum_{h=0}^{H-1} (\tilde{P}^{h-1}(\tilde{\sigma}^\pi_h + \frac{2^{1.25}|S|^{0.5}H(\log 4/\delta_P)^{0.25}}{n^{1/4}})(s) && (19) \\
&\leq |S|H^2 c_1 + c_2\sqrt{|S|H}\sqrt{\sum_{h=0}^{H-1} (\tilde{P}^{h-1}\tilde{\sigma}^{\pi^2}_h)(s)} + c_2 H\sqrt{|S|}\frac{2^{1.25}|S|^{0.5}H(\log 4/\delta_P)^{0.25}}{n^{1/4}} && (20)
\end{aligned}
$$

[2] Please note that when the assumption on transition kernel holds, then $\sum_{s'} (P_\pi(s'|s) - P'_\pi(s'|s)) V^{\pi^2}_{h+1}(s')$ and $\|V^\pi_h - V'^\pi_h\|_\infty$ are dependent. And, we can consider the one with lower probability.

[3] Here, we also know that the high probability bound on $|\sigma^\pi_h(s) - \sigma'^\pi_h(s)|$ is dependent over all $(s, a)$.

$$= \frac{3\sqrt{2}|S|H^2 \log 4/\delta_P}{n} + \frac{2\sqrt{2}|S|H^2(\log 4/\delta_P)^{\frac{3}{4}}}{n^{\frac{3}{4}}} + \sqrt{\frac{8|S|H^3 \log 4/\delta_P}{n}} + \frac{2^{2.75}|S|H^2(\log 4/\delta_P)^{\frac{3}{4}}}{n^{\frac{3}{4}}} \tag{21}$$

$$\le \sqrt{32\frac{|S|H^3}{n}}. \tag{22}$$

In equation (16), we used Lemma 7. Then, we applied Lemma 8 to obtain inequality (17). Next, we bound $\|V^{\pi}_{h+1}\|_\infty$ by $H$ in inequality (18). To get inequality (19), we use Lemma 10, since we can bound $P(\cdot|s,a) - \tilde{P}(\cdot|s,a)$ by $c_2$. And, we applied Cauchy-Scharwz inequality to get inequality (20). To get inequality (21), we applied Lemma 11 and substituting $c_1$ and $c_2$ according to equations (15). Finally, inequality (22) follows from the fact that $n \geq 2592|S|^2H^2 \log 4/\delta_P$. Since the result is true for every $s \in S$, hence the proof is complete. □

**Proof of Theorem 1:** Let $\delta_P \in (0,1)$. First, we know that optimistic planning problem (7) is feasible w.p. at least $1 - |S|^2|A|\delta_P$. The following events are dependent on this event. Thus, we consider the lowest probability of feasibility and following events.

Now, we have

$$V_0^{\pi^*}(s_0) - \sqrt{32\frac{|S|H^3 \log 4/\delta_P}{n}} \le \tilde{V}_0^{\pi^*}(s_0) \le V_0^{\pi^*}(s_0) + \sqrt{\frac{32|S|H^3 \log 4/\delta_P}{n}}$$

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

$$V_0^{\tilde{\pi}}(s_0) - \sqrt{\frac{32|S|H^3 \log 4/\delta_P}{n}} \le \tilde{V}_0^{\tilde{\pi}}(s_0) \le V_0^{\tilde{\pi}}(s_0) + \sqrt{\frac{32|S|H^3 \log 4/\delta_P}{n}}$$

w.p. at least $1 - 3|S|^2|A|H\delta_P$ according to Lemma 2. On the other hand, we know that $\tilde{V}_0^{\pi^*}(s_0) \le \tilde{V}_0^{\tilde{\pi}}(s_0)$. Thus, by combining these results we get

$$V_0^{\pi^*}(s_0) - \sqrt{\frac{32|S|H^3 \log 4/\delta_P}{n}} \le \tilde{V}_0^{\pi^*}(s) \le \tilde{V}_0^{\tilde{\pi}}(s_0) \le V_0^{\tilde{\pi}}(s) + \sqrt{\frac{32|S|H^3 \log 4/\delta_P}{n}}.$$

It yields that $V_0^{\tilde{\pi}}(s_0) \geq V_0^{\pi^*}(s_0) - 2\sqrt{\frac{32|S|H^3 \log 4/\delta_P}{n}}$ w.p. at least $1 - 6|S|^2|A|H\delta_P$ by union bound.

On the other hand, for any $i \in \{1, \ldots, N\}$ we have

$$C_{i,0}^{\tilde{\pi}}(s_0) \le \tilde{C}_{i,0}^{\tilde{\pi}}(s_0) + \sqrt{\frac{32|S|H^3 \log 4/\delta_P}{n}} \le \bar{C}_i + \sqrt{\frac{32|S|H^3 \log 4/\delta_P}{n}}$$

w.p. at least $1 - 3|S|^2|A|H\delta_P$ according to Lemma 2. By taking union bound, we get that all statements for value and cost functions hold w.p. at least $1 - (3N+6)|S|^2|A|H\delta_P$. Hence, putting $\epsilon = 2\sqrt{\frac{32|S|H^3 \log 4/\delta_P}{n}}$ and $\delta = 12(N+2)|S|^2|A|H\delta_P$ concludes the proof. Please note that $\epsilon < \frac{2}{9}\sqrt{\frac{H}{|S|}}$ would satisfy the assumption in Lemma 2. □

## 8.3 Detailed Proof for Theorem 2

First, we bound total number of model updates in Algorithm 2.

**Lemma 12.** *The total number of updates under algorithm 2 is bounded by* $U_{\max} = |S|^2|A|m$.

*Proof.* Let fix a $(s,a)-$pair. Note that $n(s,a)$ is not decreasing and also it increases up to $|S|mH$. And, since update of model happens at the beginning of each episode, then maximum number of updates due to a single $(s,a)$ happens at most $|S|m$ number of times. Thus, maximum number of updates due to all $(s,a)-$pairs is no larger than $|S|^2|A|m$ □

***Proof of Lemma 3*****:** At each episode with model update $k$ and for each $(s, a)$, by Hoeffding's inequality [27] we have

$$|P(s'|s,a) - \hat{P}(s'|s,a)| \leq \sqrt{\frac{\log{(4/\delta_1)}}{2n(s,a)}}$$

holds w.p. at least $1 - \delta_1/2$.

By empirical Brenstein's inequality [28] we have

$$|P(s'|s,a) - \widehat{P}(s'|s,a)| \leq \sqrt{\frac{2\widehat{P}(s'|s,a)(1 - \widehat{P}(s'|s,a))}{n(s,a)}\log\frac{4}{\delta_1}} + \frac{2}{3n(s,a)}\log\frac{4}{\delta_1}$$

w.p. at least $1 - \delta_1/2$.

Combining above two inequalities and applying union bound, we get

$$\mathbb{P}(|P(s'|s,a) - \widehat{P}(s'|s,a)| \leq \min\{\sqrt{\frac{2\widehat{P}(s'|s,a)(1 - \widehat{P}(s'|s,a))}{n(s,a)}\log\frac{4}{\delta_1}} + \frac{2}{3n(s,a)}\log\frac{4}{\delta_1}, \sqrt{\frac{\log 4/\delta_1}{2n(s,a)}}\}) \geq 1 - \delta_1.$$

Finally, we get the result by applying union bound over all model updates and next states. □

Now, we start proving Lemma 4. But, first we provide some useful lemmas.

**Lemma 13.** *Total number of observations of* $(s, a) \in X_{k,\kappa,\iota}$ *with* $\kappa \in [1, |S| - 1]$ *and* $\iota > 0$ *over all phases* $k$ *is at most* $3|S \times A|mw_\iota\kappa$. $w_\iota = \min\{w_k(s,a) : \iota_k(s,a) = \iota\}$.

*Proof.* Note that $w_{\iota+1} = 2w_\iota$ for $\iota > 0$. Consider a phase $k$ and a fixed $(s, a) \in X_{k,\kappa,\iota}$. Since we assumed $\iota_k(s, a) = \iota$, then $w_\iota \leq w_k(s, a) \leq 2w_\iota$. Similarly, from $\kappa_k(s, a) = \kappa$ we have $\frac{n_k(s,a)}{2mw_k(s,a)} \leq \kappa \leq \frac{n_k(s,a)}{mw_k(s,a)}$ which implies

$$mw_\iota\kappa \leq mw_k(s,a)\kappa \leq n_k(s,a) \leq 2mw_k(s,a)\kappa \leq 4mw_\iota\kappa. \tag{23}$$

Therefore, each $(s, a)$ in $\{(s, a) \in X_{k,\kappa,\iota} : k \in \mathbb{N}\}$ can only be observed $3mw_\iota\kappa$. Then, the total observations is at most $3|S \times A|mw_\iota\kappa$. □

**Lemma 14.** *Number of episodes* $E_{\kappa,\iota}$ *in phases with* $|X_{k,\kappa,\iota}| > \kappa$ *is bounded for* $\alpha \geq 3$ *w.h.p.*

$$P(E_{\kappa,\iota} > \alpha N) \leq \exp{(-\frac{\beta w_\iota(\kappa+1)N}{H})},$$

*where* $N = |S \times A|m$ *and* $\beta = \frac{\alpha(3/\alpha-1)^2}{7/3-1/\alpha}$.

*Proof.* Let $\nu_k := \sum_{h=0}^{H-1}\mathbb{I}\{(s_h, a_h) \in X_{k,\kappa,\iota}\}$ be number of observations of $(s, a)$ with $|X_{k,\kappa,\iota}| > \kappa$. We have $k \in \{1, ..., E_{\kappa,\iota}\}$.

In these episodes $|X_{k,\kappa,\iota}| \geq \kappa + 1$ and all $(s, a)$ in partition $(\kappa, \iota)$ have $w_k(s, a) \geq w_\iota$, then

$$\mathbb{E}[\nu_k|\nu_1, ..., \nu_{k-1}] \geq (\kappa + 1)w_\iota.$$

Also $\mathbb{V}[\nu_k|\nu_1, ..., \nu_{k-1}] \leq \mathbb{E}[\nu_k|\nu_1, ..., \nu_{k-1}]H$ since $\nu_k \in [0, H]$.

Now, we define the continuation:

$$\nu_k^+ := \begin{cases} \nu_k & i \leq E_{\kappa,\iota} \\ w_\iota(\kappa+1) & \text{O.W.} \end{cases}$$

and centralized auxiliary sequence

$$\bar{\nu}_k := \frac{\nu_k^+ w_\iota(\kappa+1)}{\mathbb{E}[\nu_k^+|\nu_1^+, ..., \nu_{k-1}^+]}.$$

By construction

$$\mathbb{E}[\bar{\nu}_k|\bar{\nu}_1, ..., \bar{\nu}_{k-1}] = w_\iota(\kappa+1).$$

According to lemma 13, we have $E_{\kappa,\iota} > \alpha N$ if

$$\sum_{k=1}^{\alpha N} \bar{\nu}_k \le 3N w_\iota \kappa \le 3N w_\iota(\kappa+1).$$

Now, we define martingale below

$$B_k := \mathbb{E}\left[\sum_{j=1}^{\alpha N} \bar{\nu}_j | \bar{\nu}_1, ..., \bar{\nu}_k\right] = \sum_{j=1}^{k} \bar{\nu}_j + \sum_{j=k+1}^{\alpha N} \mathbb{E}[\bar{\nu}_j|\bar{\nu}_1, ..., \bar{\nu}_i],$$

which gives $B_0 = \alpha N w_\iota(\kappa+1)$ and $B_{\alpha N} = \sum_{k=1}^{\alpha N} \bar{\nu}_k$. Now, since $\nu_k^+ \in [0, H]$

$$|B_{k+1} - B_k| = |\bar{\nu}_k - \mathbb{E}[\bar{\nu}_k|\bar{\nu}_1, ..., \bar{\nu}_{k-1}]| = \left|\frac{w_\iota(\kappa+1)(\nu_k^+ - \mathbb{E}[\nu_k^+|\bar{\nu}_1, ..., \bar{\nu}_{k-1}])}{\mathbb{E}[\nu_k^+|\nu_1^+, ..., \nu_{k-1}^+]}\right| \le |\nu_k^+ - \mathbb{E}[\nu_k^+|\bar{\nu}_1, ..., \bar{\nu}_{k-1}]| \le H.$$

Using

$$\sigma^2 := \sum_{k=1}^{\alpha N} \mathbb{V}[B_k - B_{k-1}|B_1 - B_0, ..., B_{k-1} - B_{k-2}] = \sum_{k=1}^{\alpha N} \mathbb{V}[\bar{\nu}_k|\bar{\nu}_1, ..., \bar{\nu}_{k-1}] \le \alpha N H w_\iota(\kappa+1) = HB_0$$

we can apply Theorem 22 of [30] and obtain

$$\mathbb{P}(E_{\kappa,\iota} > \alpha N) \le \mathbb{P}\left(\sum_{k=1}^{\alpha N} \bar{\nu}_k \le 3N w_\iota(\kappa+1)\right) = \mathbb{P}(B_{\alpha N} - B_0 \le 3B_0/\alpha - B_0) =\le \exp\left(-\frac{(3/\alpha - 1)^2 B_0^2}{2\sigma^2 + H(1/3 - 1/\alpha)B_0}\right)$$

for $\alpha \ge 3$. By simplifying it we get

$$\mathbb{P}(E_{\kappa,\iota} > \alpha N) \le \exp -\frac{\alpha(3/\alpha)^2}{7/3 - 1/\alpha}\frac{N w_\iota(\kappa+1)}{H}.$$

□

***Proof of Lemma 4*:** Since $w_k(s,a) \le H$, we have that $\frac{w_k(s,a)}{w_{min}} < \frac{H}{w_{min}}$ and so $\iota_k(s,a) \le H/w_{min} = 4H^2|S|/\epsilon$. In addition, $|X_{k,\kappa,\iota}| \le |S \times A|$ for all $k, \kappa, \iota$ and so $|X_{k,\kappa,\iota}| > \kappa$ can only be true for $\kappa \le |S|$. Hence, only $E_{max} = \log_2 \frac{H}{w_{min}} \log_2 |S|$ possible values for $(\kappa, \iota)$ exists that can have $|X_{k,\kappa,\iota}| > \kappa$. By union bound over all $(\kappa, \iota)$ and lemma 14, we get

$$\begin{aligned}\mathbb{P}(E \le \alpha N E_{max}) &\ge \mathbb{P}(\max_{(\kappa,\iota)} \le \alpha N) \ge 1 - E_{max} \exp\left(-\frac{\beta w_\iota(\kappa+1)N}{H}\right)\\ &\ge 1 - E_{max} \exp\left(-\frac{\beta w_{min} N}{H}\right) = 1 - E_{max} \exp\left(-\frac{\beta w_{min} m|S \times A|}{H}\right)\\ &= 1 - E_{max} \exp\left(-\frac{\beta \epsilon m|S \times A|}{4H^2|S|}\right).\end{aligned}$$

Bounding the right hand-side by $1 - \delta/2$ and solving for $m$ gives

$$1 - E_{max} \exp\left(-\frac{\beta \epsilon m |S \times A|}{4H^2|S|}\right) \geq 1 - \delta/2 \Leftrightarrow m \geq \frac{4H^2|S|}{|S \times A|\beta\epsilon} \ln \frac{2E_{max}}{\delta}.$$

Hence, the condition

$$m \geq \frac{4H^2}{\beta\epsilon} \ln \frac{2E_{max}}{\delta}$$

is sufficient for desired result to hold. Plugging in $\alpha = 6$ and $\beta = \frac{\alpha(3/\alpha-1)^2}{7/3-1/\alpha}$ would obtain the statement to show. □

Next, we need the following corollaries to prove Lemma 5.

**Corollary 3.** *If we substitute the $\delta_P$ with $\delta_1$ in Lemma 6, the result will pertain.*

**Corollary 4.** *If we substitute the $\delta_P$ with $\delta_1$ in Lemma 8, the result will pertain.*

***Proof of Lemma 5:*** We only prove the statement of value function since the proof procedure for cost is identical.

Before proceeding, in this lemma we reason about a sequence of CMDPs $M_d$ which have the same transition probabilities but different reward matrix $r^{(d)}$ and cost matrices $c^{(d)}$. Here, we only present the definition of $r^{(d)}$, as definition of $c^{(d)}$ is identical to $r^{(d)}$. For $d = 0$, the reward matrix is the original reward function $r$ of $M$ ($r^{(0)} = r$.) The following reward matrices are then defined recursively as $r^{(2d+2)} = \max_h \sigma^{(d),2}_{h:H-1}$, where $\sigma^{(d),2}_{h:H-1}$ is local variance of the value function w.r.t. the rewards $r^{(d)}$. Note that for every $d$ and $h = 0, ..., H-1$ and $s \in S$, we have $r^{(d)}(s) \in [0, H^d]$.

In addition, we will drop the notations $k$ and policy $\tilde{\pi}_k$ in the following lemmas, since the statements are for a fixed episode $k$ and all value functions, reward matrices and transition kernels are defined under policy $\tilde{\pi}_k$.

Now,

$$\begin{aligned}
\Delta_d := |V_0^{(d)}(s_0) - \tilde{V}_0^{(d)}(s_0)| &= |\sum_{h=0}^{H-2} P^{h-1}(P - \tilde{P})\tilde{V}_{h+1}^{(d)}(s_0)| \\
&\leq \sum_{h=0}^{H-1} P^{h-1}|P - \tilde{P}\tilde{V}_{h+1}^{(d)}|(s_0) \\
&= \sum_{h=0}^{H-1} P^{h-1}\left(\sum_{s,a\in S\times A} \mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s,\cdot,h)\}|(P - \tilde{P})\tilde{V}_{h+1}^{(d)}|\right)(s_0) \\
&= \sum_{s,a\in S\times A}\sum_{h=0}^{H-1} P^{h-1}\left(\mathbb{I}\{s = \cdot, a = \tilde{\pi}(s,\cdot,h)\}|(P - \tilde{P})\tilde{V}_{h+1}^{(d)}|\right)(s_0) \\
&= \sum_{s,a\in S\times A}\sum_{h=0}^{H-1} P^{h-1}\left(\mathbb{I}\{s = \cdot, a = \tilde{\pi}(s,\cdot,h)\}|(P - \tilde{P})\tilde{V}_{h+1}^{(d)}(s)|\right)(s_0)
\end{aligned}$$

The first equality follows from Lemma 7, the second step from the fact that $V_{h+1} \geq 0$ and $P^{h-1}$ being non-expansive. In the third, we introduce an indicator function which does not change the value as we sum over all $(s, a)$ pairs. The fourth step relies on the linearity of $P$ operators. In the fifth step, we realize that $\mathbb{I}\{s = ., a \sim \tilde{\pi}(s,\cdot,h)\}|(P - \tilde{P})\tilde{V}_{h+1}^{(d)}(\cdot)|$ is a function that takes nonzero values for input $s$. We can therefore replace the argument of the second term with $s$ without changing the value. The term becomes constant and by linearity of $P$, we can write

$$
\begin{aligned}
&|V_0^{(d)}(s_0) - \tilde{V}_0^{(d)}(s_0)| = \Delta_d \leq \sum_{s,a \in S \times A} \sum_{h=0}^{H-1} P^{h-1} \left( \mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\} |(P - \tilde{P}) \tilde{V}_{h+1}^{(d)}(s)| \right) (s_0) \\
&\leq \sum_{s,a \notin X} \sum_{h=0}^{H-1} \|\tilde{V}_{h+1}^{(d)}\|_\infty (P^{h-1} \mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0) \\
&+ \sum_{s,a \in X} \sum_{h=0}^{H-1} |(P - \tilde{P}) \tilde{V}_{h+1}^{(d)}(s)| (P^{h-1} \mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0) \\
&\leq \sum_{s,a \notin X} \sum_{h=0}^{H-1} H^{d+1} (P^{h-1} \mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0) \\
&+ \sum_{s,a \in X} \sum_{h=0}^{H-1} |(P - \tilde{P}) \tilde{V}_{h+1}^{(d)}(s)| (P^{h-1} \mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0) \\
&\leq \sum_{s,a \notin X} \sum_{h=0}^{H-1} H^{d+1} (P^{h-1} \mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0) \\
&+ \sum_{s,a \in X} \sum_{h=0}^{H-1} ||S| c_1(s,a) H^{d+1} + c_2(s,a) \sqrt{|S|} \tilde{\sigma}_h^{(d)}(s,a)| (P^{h-1} \mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0) \\
&\leq \sum_{s,a \notin X} \sum_{h=0}^{H} H^{d+1} (P^{h-1} \mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0) \\
&+ \sum_{s,a \in X} \sum_{h=0}^{H} ||S| c_1(s,a) H^{d+1}| (P^{h-1} \mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0) \\
&+ \sum_{s,a \in X} \sum_{h=0}^{H-1} |\sqrt{|S|} c_2(s,a) \tilde{\sigma}_h^{(d)}(s,a)| (P^{h-1} \mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0) \\
&\leq \sum_{s,a \notin X} H^{d+1} w(s,a) + \sum_{s,a \in X} |S| c_1(s,a) H^{d+1} w(s,a) \\
&+ \sum_{s,a \in X} \sqrt{|S|} c_2(s,a) \sum_{h=0}^{H-1} \tilde{\sigma}_h^{(d)}(s,a) (P^{h-1} \mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0) \\
&\leq w_{min} |S| H^{d+1} + \sum_{s,a \in X} |S| c_1(s,a) H^{d+1} w(s,a) + \sum_{s,a \in X} \sqrt{|S|} c_2(s,a) \sum_{h=0}^{H-1} \tilde{\sigma}_h^{(d)}(s,a) (P^{h-1} \mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0) \\
&= \frac{\epsilon}{4} H^d + \sum_{s,a \in X} |S| c_1(s,a) H^{d+1} w(s,a) + \sum_{s,a \in X} \sqrt{|S|} c_2(s,a) \sum_{h=0}^{H-1} \tilde{\sigma}_h^{(d)}(s,a) (P^{h-1} \mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0)
\end{aligned}
$$

In the second inequality, we split the sum over all $(s, a)$ pairs and used the fact that $P$ and $\tilde{P}$ are non-expansive. The next step follows from $\|V_{h+1}^{(d)}\|_\infty \leq \|V_0^{(d)}\|_\infty \leq H^{d+1}$. We then apply Lemma 8 and subsequently use that all terms are nonnegative and the definition of $w(s, a)$. Eventually, the last two lines come from the fact that $w(s, a) \leq w_{min}$ for all $(s, a)$ not in the active set. Besides, please note that we are analyzing under the given policy $\tilde{\pi}$, which implies that there are only $|S|$ nonzero $w$ in non-active set.

Using the assumption that $M \in \mathcal{M}$ and $\tilde{M} \in \mathcal{M}$ from the fact that ELP chooses the optimistic CMDP in $\mathcal{M}$, we can apply Corollary 3 and get that

$$c_1(s,a) = 2\sqrt{2}\Big(\frac{\log 4/\delta_1}{n(s,a)}\Big)^{3/4} + 3\sqrt{2}\frac{\log 4/\delta_1}{n(s,a)} \quad \text{and} \quad c_2(s,a) = \sqrt{\frac{8}{n(s,a)}\log 4/\delta_1}.$$

Plugging definitions above we have

$$\Delta_d \leq \frac{\epsilon}{4}H^d + 2\sqrt{2}|S|H^{d+1}\log 4/{\delta_1}^{3/4} \sum_{s,a \in X} \frac{w(s,a)}{n(s,a)^{3/4}} + 3\sqrt{2}|S|H^{d+1}\log 4/\delta_1 \sum_{s,a \in X} \frac{w(s,a)}{n(s,a)}$$
$$+ \sqrt{8|S|\log 4/\delta_1} \sum_{s,a \in X} \frac{1}{\sqrt{n(s,a)}} \sum_{h=0}^{H-1} \tilde{\sigma}_h^{(d)}(s,a)(P^{h-1}\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s,\cdot,h)\})(s_0)$$

Hence, we bound

$$\Delta_d \leq \frac{\epsilon}{4}H^d + U_d(s_0) + Y_d(s_0) + Z_d(s_0)$$

as a sum of three terms which we will consider individually in the following. The first term is

$$\begin{aligned}
U_d(s_0) &= 2\sqrt{2}|S|H^{d+1}\log 4/{\delta_1}^{3/4} \sum_{s,a \in X} \frac{w(s,a)}{n(s,a)^{3/4}} \\
&\leq 2\sqrt{2}|S|H^{d+5/4}\log 4/{\delta_1}^{3/4} \sum_{\kappa,\iota \in \mathcal{K}\times\mathcal{I}} \sum_{s,a \in X_{\kappa,\iota}} \Big(\frac{w(s,a)}{n(s,a)}\Big)^{3/4} \\
&\leq 2\sqrt{2}|S|H^{d+5/4}\log 4/{\delta_1}^{3/4} \sum_{\kappa,\iota \in \mathcal{K}\times\mathcal{I}} \Big(\frac{|X_{\kappa,\iota}|}{m\kappa}\Big)^{3/4} \\
&\leq 2\sqrt{2}|S|H^{d+5/4}\log 4/{\delta_1}^{3/4} \sum_{\kappa,\iota \in \mathcal{K}\times\mathcal{I}} \Big(\frac{1}{m}\Big)^{3/4} \\
&\leq 2\sqrt{2}|S|H^{d+5/4}\log 4/{\delta_1}^{3/4} \mathcal{K}\times\mathcal{I}\Big(\frac{1}{m}\Big)^{3/4}.
\end{aligned}$$

In the second line, we used Cauchy-Scharwz. Next, we used the fact that for $s, a \in X_{\kappa,\iota}$, we have $n(s,a) \geq mw(s,a)\kappa$, refer to equation (23). Finally, we applied the assumption of $|X_{\kappa,\iota}| \leq \kappa$. Please note that $\mathcal{K}\times\mathcal{I}$ is the set of all possible $(\kappa,\iota)$ pairs.

The next term is

$$Y_d(s_0) = 3\sqrt{2}|S|H^{d+1}\log 4/\delta_1 \sum_{s,a \in X} \frac{w(s,a)}{n(s,a)} \leq 3\sqrt{2}|S|H^{d+1}\log 4/\delta_1 \sum_{\kappa,\iota} \frac{|X_{\kappa,\iota}|}{m\kappa} \leq \frac{3\sqrt{2}|S|H^{d+1}\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m}$$

which we used $n(s,a) \geq mw(s,a)\kappa$ again.

The last term is

$$Z_d(s_0) = \sqrt{8|S|\log 4/\delta_1} \sum_{s,a \in X} \frac{1}{\sqrt{n(s,a)}} \sum_{h=0}^{H-1} \tilde{\sigma}_h^{(d)}(s,a)(P^{h-1}\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s,\cdot,h)\})(s_0) \leq \sqrt{8|S|\log 4/\delta_1}$$
$$\times \sum_{s,a \in X} \frac{1}{\sqrt{n(s,a)}} \sqrt{\sum_{h=0}^{H-1} P^{h-1}\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s,\cdot,h)\}(s_0)} \sqrt{\sum_{h=0}^{H-1} \tilde{\sigma}_h^{(d)^2}(s,a)P^{h-1}\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s,\cdot,h)\}(s_0)}$$
$$= \sqrt{8|S|\log 4/\delta_1} \sum_{s,a \in X} \sqrt{\frac{w(s,a)}{n(s,a)} \sum_{h=0}^{H-1} \tilde{\sigma}_h^{(d)^2}(s,a)P^{h-1}\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s,\cdot,h)\}(s_0)}$$

$$
\begin{aligned}
&= \sqrt{8|S|\log 4/\delta_1} \sum_{\kappa,\iota} \sum_{s,a\in X_{\kappa,\iota}} \sqrt{\frac{w(s,a)}{n(s,a)} \sum_{h=0}^{H-1} \tilde{\sigma}_h^{(d)^2}(s,a) P^{h-1}\mathbb{I}\{s=\cdot, a\sim\tilde{\pi}(s,\cdot,h)\}(s_0)} \\
&\le \sqrt{8|S|\log 4/\delta_1} \sum_{\kappa,\iota} \sqrt{|X_{\kappa,\iota}| \sum_{s,a\in X_{\kappa,\iota}} \frac{w(s,a)}{n(s,a)} \sum_{h=0}^{H-1} \tilde{\sigma}_h^{(d)^2}(s,a) P^{h-1}\mathbb{I}\{s=\cdot, a\sim\tilde{\pi}(s,\cdot,h)\}(s_0)} \\
&\le \sqrt{8|S|\log 4/\delta_1} \sum_{\kappa,\iota} \sqrt{\frac{1}{m} \sum_{s,a\in X_{\kappa,\iota}} \sum_{h=0}^{H-1} \tilde{\sigma}_h^{(d)^2}(s,a) P^{h-1}\mathbb{I}\{s=\cdot, a\sim\tilde{\pi}(s,\cdot,h)\}(s_0)} \\
&\le \sqrt{\frac{8|S|\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m} \sum_{s,a\in X} \sum_{h=0}^{H-1} \tilde{\sigma}_h^{(d)^2}(s,a) P^{h-1}\mathbb{I}\{s=\cdot, a\sim\tilde{\pi}(s,\cdot,h)\}(s_0)} \\
&\le \sqrt{\frac{8|S|\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m} \sum_{s,a\in S\times A} \sum_{h=0}^{H-1} \tilde{\sigma}_h^{(d)^2}(s,a) P^{h-1}\mathbb{I}\{s=\cdot, a\sim\tilde{\pi}(s,\cdot,h)\}(s_0)} \\
&= \sqrt{\frac{8|S|\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m} \sum_{h=0}^{H-1} P^{h-1}\tilde{\sigma}_h^{(d)^2}(s_0)} \\
&\le \sqrt{\frac{8|S|H^{2d+3}\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m}}.
\end{aligned}
$$

In the second line, we applied Cauchy-Scharwz inequality. Then, we used the definition of $w(s,a)$ to get to third step. Next, we split the sum and applied Cauchy-Scharwz again to obtain fifth step. Furthermore, we applied the assumption of $|X_{\kappa,\iota}| \le \kappa$ to get sixth step. Next, we applied Cauchy-Scharwz inequality to obtain seventh step. And, the final step follows from the facts that $P^{h-1}$ is non-expansive and $\|\tilde{\sigma}_h^{(d)}\|_\infty \le H^{2d+2}$. Thus, we have

$$
Z_d(s_0) \le \sqrt{\frac{8|S|H^{2d+3}\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m}}. \tag{24}
$$

However, we can improve this bound as follows

$$
\begin{aligned}
Z_d(s_0) &\le \sqrt{\frac{8|S|\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m} \sum_{h=0}^{H-1} P^{h-1}\tilde{\sigma}_h^{(d)^2}(s_0)} \\
&= \sqrt{\frac{8|S|\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m} \sum_{h=0}^{H-1} P^{h-1}\tilde{\sigma}_h^{(d)^2}(s_0) - \tilde{P}^{h-1}\tilde{\sigma}_h^{(d)^2}(s_0) + \tilde{P}^{h-1}\tilde{\sigma}_h^{(d)^2}(s_0)} \\
&\le \sqrt{\frac{8|S|\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m} \Big(H^{2d+2} + \sum_{h=0}^{H-1} P^{h-1}r^{(2d+2)}(s_0) - \tilde{P}^{h-1}r^{(2d+2)}(s_0)\Big)} \\
&= \sqrt{\frac{8|S|\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m} \Big(H^{2d+2} + V_0^{(2d+2)}(s_0) - \tilde{V}_0^{(2d+2)}(s_0)\Big)} \\
&= \sqrt{\frac{8|S|\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m} (H^{2d+2} + \Delta_{2d+2})} \\
&\le \sqrt{\frac{8|S|\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m} H^{2d+2}} + \sqrt{\frac{8|S|\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m} \Delta_{2d+2}}.
\end{aligned}
$$

In the third step, we used Lemma 11 and definition of $r^{(2d+2)}$.

Now, if we put all the pieces together, we have

$$\Delta_d \leq \frac{\epsilon}{4}H^d + 2\sqrt{2}|S|H^{d+5/4}\log 4/\delta_1{}^{3/4}\mathcal{K}\times\mathcal{I}\Big(\frac{1}{m}\Big)^{3/4} + \frac{3\sqrt{2}|S|H^{d+1}\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m}$$
$$+ \sqrt{\frac{8|S|\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m}}H^{2d+2} + \sqrt{\frac{8|S|\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m}}\Delta_{2d+2}.$$

If we choose $m$ sufficiently large which will be shown later, then it is straightforward to show that $U_d(s_0) \leq Z_d(s_0)$ and $Y_d(s_0) \leq Z_d(s_0)$. Hence, if we expand the above inequality up to depth $\beta = \lceil \frac{\log H}{2\log 2} \rceil$ with $\mathcal{D} = \{0, 2, 6, 14, \ldots, \beta\}$, we get

$$\Delta_0 \leq \sum_{d\in\mathcal{D}\setminus\beta} \Big(\frac{8|S|\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m}\Big)^{\frac{d}{d+2}} \Big[\frac{\epsilon}{4}H^d + 3\sqrt{\frac{8|S|\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|H^{2d+2}}{m}}\Big]^{\frac{2}{d+2}}$$
$$+ \Big(\frac{8|S|\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m}\Big)^{\frac{\beta}{\beta+2}} \Big[\frac{\epsilon}{4}H^\beta + 3\sqrt{\frac{8|S|\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|H^{2\beta+2}}{m}}\Big]^{\frac{2}{\beta+2}}.$$

Here, we used inequality (24) to bound $Z_\beta(s_0)$. Finally, the proof completes if we let

$$m = 1280\frac{|S|H^2}{\epsilon^2}(\log_2\log_2 H)^2\log_2^2\Big(\frac{8|S|^2H^2}{\epsilon}\Big)\log\frac{6}{\delta_1}.$$

□.

**Proof of Theorem 2:** By Lemma 4, we know that number of episodes where $|X_{\kappa,\iota}| > \kappa$ for some $\kappa, \iota$ is bounded by $6E_{\max}|S||A|m$ with probability at least $1 - \frac{\delta}{2(N+1)}$. For all other episodes, we have by Lemma 5 that for any $i \in \{1, \ldots, N\}$

$$|\tilde{V}_0^{\tilde{\pi}_k}(s_0) - V_0^{\tilde{\pi}_k}(s_0)| \leq \epsilon, \quad |\tilde{C}_{i,0}^{\tilde{\pi}_k}(s_0) - C_0^{\tilde{\pi}_k}(s_0)| \leq \epsilon. \tag{25}$$

Using Lemma 3, we get that $M \in \mathcal{M}_k$ for any episode $k$ w.p. at least $1 - \frac{\delta}{2(N+1)}$. Further, we know that ELP outputs the policy $\tilde{\pi}_k$ such that

$$\tilde{V}_0^{\tilde{\pi}_k}(s_0) \geq V_0^{\pi^*}(s_0), \quad \tilde{C}_{i,0}^{\tilde{\pi}_k}(s_0) \leq \bar{C}_i \;\; i \in \{1, \ldots, N\} \tag{26}$$

w.p. at least $1 - \frac{\delta}{2(N+1)}$. Combining the inequalities (25) with inequalities (26), we get that for all episodes with $|X_{\kappa,\iota}| \leq \kappa$ for all $\kappa, \iota$

$$V_0^{\tilde{\pi}_k}(s_0) \geq V_0^{\pi^*}(s_0) - \epsilon$$

w.p. at least $1 - \frac{\delta}{2(N+1)}$ and for any $i$, $C_{i,0}^{\tilde{\pi}_k}(s_0) \leq \bar{C}_i + \epsilon$ w.p. at least $1 - \frac{\delta}{2(N+1)}$. Applying the union bound we get the desired result, if $m$ satisfies

$$m \geq 1280\frac{|S|H^2}{\epsilon^2}(\log_2\log_2 H)^2\log_2^2\Big(\frac{8H^2|S|^2}{\epsilon}\Big)\log\frac{4}{\delta_1} \;\; \text{and}$$
$$m \geq \frac{6H^2}{\epsilon}\log\frac{2(N+1)E_{\max}}{\delta}.$$

From the definitions, we get

$$\log\frac{4}{\delta_1} = \log\frac{4(N+1)|S|U_{\max}}{\delta} = \log\frac{4(N+1)|S|^2|A|m}{\delta}.$$

Thus,

$$m \geq 1280\frac{|S|H^2}{\epsilon^2}(\log_2\log_2 H)^2\log_2^2\Big(\frac{8H^2|S|^2}{\epsilon}\Big)\log\frac{4(N+1)|S|^2|A|m}{\delta}.$$

It is well-known fact that for any constant $B > 0, \nu \geq 2B\ln B$ implies $\nu \geq B\ln\nu$. Using this, we can set

$$\begin{aligned} m \geq 2560&\frac{|S|H^2}{\epsilon^2}(\log_2\log_2 H)^2\log_2^2\Big(\frac{8H^2|S|^2}{\epsilon}\Big)\\ &\times\log\Big(\frac{2048(N+1)|S|^3|A|H^2}{\epsilon^2\delta}(\log_2\log_2 H)^2\log_2^2\Big(\frac{8H^2|S|^2}{\epsilon}\Big)\Big).\end{aligned}$$

On the other hand,

$$E_{\max} = \log_2|S|\log_2\frac{4|S|H^2}{\epsilon} \leq \log_2^2\frac{4|S|H^2}{\epsilon}$$

and

$$\begin{aligned}\log\frac{2(N+1)E_{\max}}{\delta} &= \log\frac{2(N+1)\log_2|S|\log_2(4|S|H^2/\epsilon)}{\delta} \leq \log\frac{2(N+1)\log_2^2(4|S|H^2/\epsilon)}{\delta}\\ &\leq \log\frac{16(N+1)|S|^4|A|H^2}{\epsilon\delta}.\end{aligned}$$

Setting

$$\begin{aligned} m = 2560&\frac{|S|H^2}{\epsilon^2}(\log_2\log_2 H)^2\log_2^2\Big(\frac{8H^2|S|^2}{\epsilon}\Big)\\ &\times\log\Big(\frac{2048(N+1)|S|^4|A|H^2}{\epsilon^2\delta}(\log_2\log_2 H)^2\log_2^2\Big(\frac{8H^2|S|^2}{\epsilon}\Big)\Big).\end{aligned} \tag{27}$$

is therefore a valid choice for $m$ to ensure that with probability at least $1-\frac{\delta}{(N+1)}$, there are at most

$$\begin{aligned}6E_{\max}|S||A|m =&15360\frac{|S|^2|A|H^2}{\epsilon^2}(\log_2\log_2 H)^2\log_2^2\Big(\frac{4|S|H^2}{\epsilon}\Big)\log_2^2\Big(\frac{8H^2|S|^2}{\epsilon}\Big)\\ &\times\log\Big(\frac{2048(N+1)|S|^4|A|H^2}{\epsilon^2\delta}(\log_2\log_2 H)^2\log_2^2\Big(\frac{8H^2|S|^2}{\epsilon}\Big)\Big)\end{aligned}$$

sub-optimal episodes. □