LEARNING AND CONTROL WITH DELAY AND SAFETY CONSTRAINTS IN NETWORKED SYSTEMS

A Dissertation

by

ARIA HASANZADEZONUZY

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Srinivas Shakkottai
Committee Members,	Dileep Kalathil
	P.R. Kumar
	Suman Chakravorty
Head of Department,	Mirosalv M. Begovic

December 2021

Major Subject: Electrical Engineering

ABSTRACT

Many physical systems have underlying safety or capacity considerations that require that the control policy employed ensures the satisfaction of a set of constraints. For instance, in a data network, in addition to maximizing the utility of the users, the controller has to maintain necessary link capacities constraints. Such systems can often be modeled as a constrained Markov Decision Process (CMDP), but the model itself might have unknown or rapidly changing system parameters, which calls for a learning-based solution approach.

Our goal in this thesis is to develop Reinforcement Learning (RL) algorithms to learn a generic CMDP problem, and explore applicatiosn to communication networks. Here, our goal is to characterize the relationship between constraints and the number of samples needed to ensure a desired level of accuracy. We explore two classes of algorithms, (i) algorithms based on Linear Programming (LP), (ii) algorithms based on a Lagrangian approach. Each of these classes is divided into two sub-classes according to sample collection process. On the one hand, we may collect samples uniformly across state-action pairs, and then develop a control policy based on these samples—called the generative model based approach. On the other hand, we may collect samples in an online manner by applying a policy on the system, and then continually refining the policy as more samples become available—called the online learning approach. We characterize the sample complexity of the algorithms following both these approaches to obtain near-optimal policies.

We then consider the question of CMDPs in the context of data networks. We desire to solve the problem of serving real-time flows over a multi-hop wireless network. Each flow is composed of packets that have strict deadlines, and the goal is to maximize the weighted timely throughput of the system. Consistent with recent developments using mm-wave communications, we assume that the links are directional, but are lossy, and have unknown probabilities of successful packet transmission. An average link utilization budget constrains the system. The problem thus takes the form of a CMDP with an unknown transition kernel, and we develop new algorithms well suited for data network problems using the insights of RL algorithms for generic CMDPs.

ACKNOWLEDGMENTS

I would like to thank my family, my advisor and my friends who helped me in their own ways.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professors Srinivas Shakkottai, Dileep Kalathil and P.R. Kumar of the Department of Electrical and Computer Engineering and Professor Suman Chakravorty of the Department of Aerospace Engineering.

The simulations of Chapter 2 were conducted by Archana Bura of the Department of Electrical Engineering.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by grants from the National Science Foundation ECCS 1839816, CNS 1719384, EFMA 1440969, CNS 1149458 and the Army Research Office grant W911NF-19-1-0367. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsoring agencies.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

1. INTRODUCTION

Many physical systems have underlying safety considerations that require that the control policy employed ensures the satisfaction of a set of constraints. These constraints might take range from per-packet deadline guarantees in a communication system, to the need to ensure that cars do not hit each other in an autonomous platoon. Often, the system of interest can be modeled as a constrained Markov Decision Process (CMDP), but the model itself might have unknown or rapidly changing system parameters. Thus, conventional control approaches are not attractive and reinforcement learning (RL) is called for. In other cases, the model might be well defined, but many interacting systems, such as autonomous vehicles might render the system complex. Our focus in this thesis is to develop control algorithms for networked systems that require the satisfaction of constraints for their safe and efficient operation.

In chapters $2, 3$ and 4, 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. In chapter 2 and 3 we solely focus on reducing the sample complexity for infinite-horizon and finite-horizon CMDPs respectively. However, in chapter 4 we concentrate on computationally efficient algorithms with cost of sample complexity. Our main finding in these chapters 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.

We next take up the problem of constrained decision making in data networks. In chapter 5, we study the problem of broadcasting real-time flows in multi-hop wireless networks. We consider that each packet has a stringent deadline, and each node in the network obtains some utility based on the number of packets delivered to it on time for each flow. We propose a distributed protocol, the delegated-set routing (DSR) protocol, that incurs virtually no overhead of coordination among nodes. We also develop distributed algorithms that aim to maximize the total system utility under DSR. The utility of our DSR protocol and distributed algorithms are demonstrated by both theoretical analysis and simulation results, where we show that our algorithms achieve better performance even when compared against centralized throughput optimal policies.

Finally, in chapter 6, we bring together the constrained network optimization and constrained RL approaches in the context of serving real-time flows over a multi-hop wireless network. Each flow is composed of packets that have strict deadlines, and the goal is to maximize the weighted timely throughput of the system. Consistent with recent developments using mm-wave communications, we assume that the links are directional, but are lossy, and have unknown probabilities of successful packet transmission. An average link utilization budget (similar to a power constraint) constrains the system. We pose the problem in the form of a CMDP with an unknown transition kernel. We use a duality approach to decompose the problem into an inner unconstrained MDP with link usage costs, and an outer link-cost update step. For the inner MDP, we develop model-based reinforcement learning algorithms that sample links by sending packets to learn the link statistics. While the first algorithm type samples links at will at the beginning and constructs the model, the second type is an online approach that can only use packets from flows to sample links that they traverse. The approach to the outer problem follows gradient descent. We characterize the sample complexity (number of packets transmitted) to obtain near-optimal policies, to show that a basic online approach has a poorer sample complexity bound, it can be modified to obtain an online algorithm that has excellent empirical performance.

2. LEARNING WITH SAFETY CONSTRAINTS: SAMPLE COMPLEXITY OF REINFORCEMENT LEARNING FOR CONSTRAINED MDPs*

2.1 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 [4].

In this chapter, 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 maximiza-

*Reprinted with permission from [1].

tion 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 [5], 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 [6, 7], 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 [4, 8, 9, 10, 11, 12]. 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 [13, 14]. 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 [15, 16, 17, 18] under Lagrangian methods to show zero eventual duality gap. [19] 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 [20].

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 [21, 22, 23] 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, [24] and [25] present results in the context of unknown reward functions, with either a known stochastic or deterministic transition kernel. Other work [26] focuses on asymptotic convergence, and so does not provide an estimate on the learning rate. Finally, [5] 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 [27], 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.

2.2 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 $i \in \{1,\ldots,N\}$. 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 $i \in \{1,\ldots,N\}, \bar{C}_i \in [0, \bar{C}_{\max}]$.*

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 $P_\pi^h f(s) = P_\pi P_\pi \ldots P_\pi f$, 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

$$V_t^\pi(s) = \mathbb{E}[\sum_{h=t}^{H-1} r(s_h, a_h); a_h \sim \pi(s_h, \cdot, h), s_t = s], \tag{2.1}$$

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

$$\sigma_h^{\pi^2}(s) = \mathbb{E}[(V_{h+1}^\pi(s_{h+1}) - P_\pi V_{h+1}^\pi(s))^2]. \tag{2.2}$$

Similar to the definition of the value function (3.1), the i^{th} constraint function at time t under policy π is formulated as

$$C^{\pi}_{i,t}(s) = \mathbb{E}[\sum_{h=t}^{H-1} c(i, s_h, a_h); a_t \sim \pi(s_h, \cdot, h), s_t = s]. \tag{2.3}$$

Again, the local variance of i^{th} constraint function at time-step h under policy π, i.e. $\sigma^{\pi^2}_{i,h}$ is defined similar to local variance of value function (4.2).

Finally, the general finite-horizon CMDP problem is

$$\max_{\pi} V^{\pi}_0(s_0) \text{ s.t. } C^{\pi}_{i,0}(s_0) \leq \bar{C}_i, \quad \forall i \in \{1, \ldots, N\}. \tag{2.4}$$

Assumption 2. *We assume that there exists some policy π that satisfies the constraints in* (3.4). *Hence, this CMDP problem is feasible with optimal policy π^* and optimal solution $V^*_0(s_0) = V^{\pi^*}_0(s_0)$.*

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 (3.4), but without being aware of values of transition kernel P.[†] 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*

$$\mathbb{P}\Big(V^{\mathcal{A}}_0(s_0) \geq V^{\pi^*}_0(s_0) - \epsilon \text{ and } C^{\mathcal{A}}_{i,0}(s_0) \leq \bar{C}_i + \epsilon \, \forall i \in \{1, \ldots, N\}\Big) \geq 1 - \delta$$

[†] 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.

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 [28].

2.3 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

$$\begin{aligned}
&\mathcal{M}_{\delta_P} := \{M' : r'(s, a) = r(s, a), && (2.5)\\
&c'(i, s, a) = c(i, s, a), H' = H, s'_0 = s_0 \\
&|P'(s'|s, a) - \widehat{P}(s'|s, a)| \leq && (2.6)\\
&\min\Big(\sqrt{\frac{2\widehat{P}(s'|s, a)(1 - \widehat{P}(s'|s, a))}{n} \log\frac{4}{\delta_P}} + \frac{2}{3n}\log\frac{4}{\delta_P}, \\
&\sqrt{\frac{\log 4/\delta_P}{2n}}\Big) \forall s, a, s', i\},
\end{aligned}$$

where δ_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 (3.1) and (3.3) 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

$$\max_{\pi, M' \in \mathcal{M}_{\delta_P}} V_0^{'\pi}(s_0) \quad \text{s.t.} \quad C_{i,0}^{'\pi}(s_0) \leq \bar{C}_i \;\; \forall i. \tag{2.7}$$

Optimistic-GMBL uses Extended Linear Programming, or **ELP**, to solve the problem of (6.33). This method inputs $\mathcal{M}_{\delta_P}$ and outputs $\tilde{\pi}$ for the optimal solution. The description of ELP is provided in supplementary materials. 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 (4.21).
Output $\tilde{\pi} = \text{ELP}(\mathcal{M}_{\delta_P})$.
```

2.3.1 PAC Analysis of Optimistic-GMBL

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

Theorem 1. *Consider any finite-horizon CMDP $M = \langle S, A, P, r, c, \bar{C}, s_0, H \rangle$ satisfying assumptions 3 and 4, and CMDP problem formulation of* (3.4). *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*

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

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 12 differs from the traditional analysis framework of unconstrained RL [6] 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 [29]. However, there is no such guarantee for any general CMDP problem formulation [4]. 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 (6.33) 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 [4]. 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 12. Then we sketch the proof of this theorem. The detailed proofs are provided in supplementary materials.

First, we show that true CMDP lies inside the $\mathcal{M}_{\delta_P}$ with high probability, w.h.p. So, the problem (6.33) would be feasible w.h.p., since the original CMDP problem is assumed to be feasible according to Assumption 4.

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 [30] and empirical Bernstein's inequality [31], we get that each $P(s'|s, a)$ is inside the confidence set defined by (6.32) with probability at least $1 - \delta_P$. Applying the union bound yields the result. □

Now, we present the core lemma required for proving Theorem 12 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*

$$\|V_0^\pi - \tilde{V}_0^\pi\|_\infty \leq \sqrt{\frac{128|S|H^3 \log 4/\delta_P}{n}}$$

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

$$\|C_{i,0}^\pi - \tilde{C}_{i,0}^\pi\|_\infty \leq \sqrt{\frac{128|S|H^3 \log 4/\delta_P}{n}}$$

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 12*:** From Lemma 6, we know that the optimistic planning problem (6.33) 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 28. 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 28 to bound constraint violation. Thus, for each $i \in \{1, \ldots, N\}$ 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 $\epsilon = O(\sqrt{\frac{|S|H^3}{n}})$. □

2.4 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 8, 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 (6.32). 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 :

$$\begin{aligned}
&\mathcal{M}_k := \{M' : r'(s,a) = r(s,a), \\
&c'(i,s,a) = c(i,s,a), H' = H, s'_0 = s_0 \\
&|P'(s'|s,a) - \widehat{P}(s'|s,a)| \leq \\
&\min\Big(\sqrt{\frac{2\widehat{P}(s'|s,a)(1-\widehat{P}(s'|s,a))}{n(s,a)}\log\frac{4}{\delta_1}} \\
&+ \frac{2}{3n(s,a)}\log\frac{4}{\delta_1}, \sqrt{\frac{\log 4/\delta_1}{2n(s,a)}}\Big)\ \forall s, s', a, i\},
\end{aligned} \tag{2.8}$$

where δ_1 is defined in Algorithm 8.

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 (6.33), 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 (3.1) and (3.3) w.r.t. underlying transition kernel P', respectively.

This algorithm draws inspiration from the infinite-horizon algorithm UCRL$-\gamma$ [32] and its finite-horizon counterpart UCFH [7] with several differences. Unlike UCRL-γ and UCFH, Algorithm 8 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 [4], we must ensure that all constraints are linear. Hence, unlike UCFH, Algorithm 8

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 (6.35) and (6.36).
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 (6.31).
    $\tilde{\pi}_k = \text{ELP}(\mathcal{M}_k)$.
    for $t = 1, \ldots, H$ do
        $a_t \sim \tilde{\pi}_k(s_t), s_{t+1} \sim P(\cdot|s_t, a_t), n(s_t, a_t) ++, n(s_{t+1}, s_t, a_t) ++$.
    $k ++$
```

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.

2.4.1 PAC Analysis of Online-CRL

We now present the PAC bound of Algorithm 8.

Theorem 2. *Consider CMDP $M = \langle S, A, r, c, \bar{C}, s_0, H \rangle$ satisfying assumptions 3 and 4. 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 13, we follow an approach motivated by [32] and its finite-horizon version [7]. 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 [5] while using an algorithm reminiscent of Algorithm 8, 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 [27], and following the [5] 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 13. Finally, we sketch the proof of Theorem 13. The detailed analysis is provided in supplementary materials.

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* 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

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

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

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

which indicates how often (s, a) has been observed relative to its importance. Value of m is defined in Algorithm 8. 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*

$$m \geq \frac{6H^2}{\epsilon} \log \frac{2(N+1)E_{\max}}{\delta}, \tag{2.9}$$

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 28 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*

$$m = 1280\frac{|S|H^2}{\epsilon^2}(\log_2 \log_2 H)^2 \log_2^2\Big(\frac{8|S|^2H^2}{\epsilon}\Big) \log \frac{4}{\delta_1}, \tag{2.10}$$

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 13*:** First, we apply Lemma 30 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 33. 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 35. Eventually, taking union bound yields the result. □

2.5 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 3×3 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

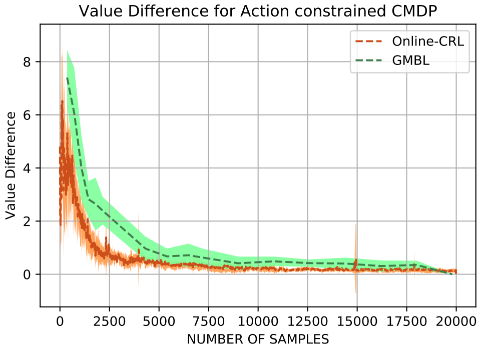

Figure 2.1: Value Difference for Scenario 1a. Reprinted with permission from [1]

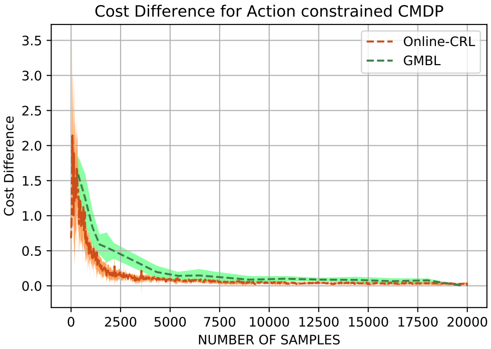

Figure 2.2: Constraint Violation for Scenario 1a. Reprinted with permission from [1]

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 2.1, 2.3 and 2.5, 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 2.2, 2.4, and 2.6, which illustrates constraint violation.

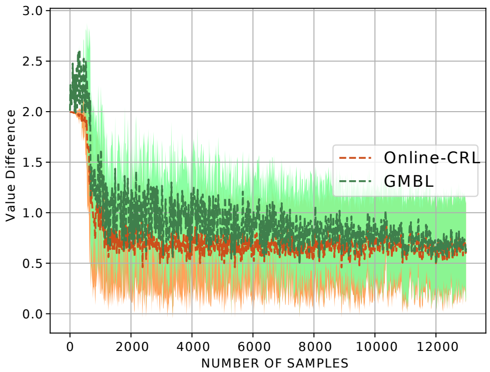

Figure 2.3: Value Difference for Scenario 1b. Reprinted with permission from [1]

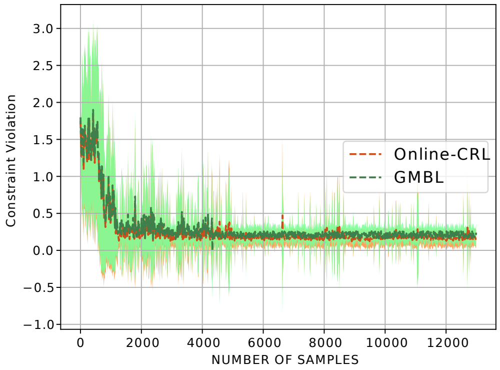

Figure 2.4: Constraint Violation for Scenario 1b. Reprinted with permission from [1]

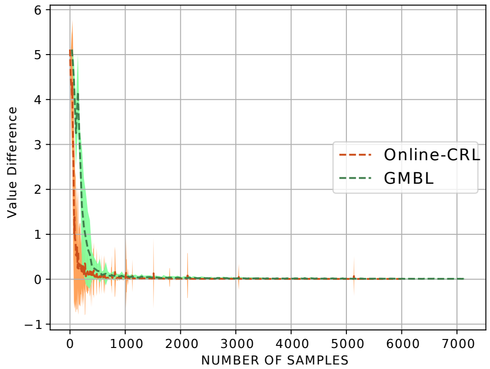

Figure 2.5: Value Difference for Scenario 2. Reprinted with permission from [1]

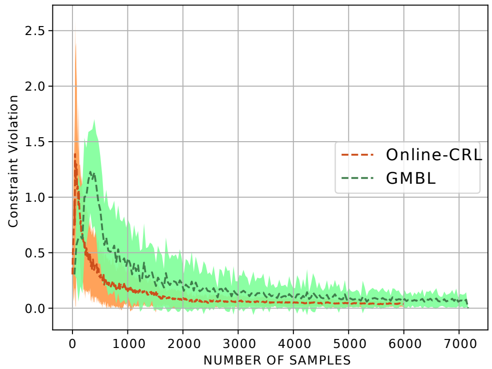

Figure 2.6: Constraint Violation for Scenario 2. Reprinted with permission from [1]

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

2.6 Conclusion

In this chapter, we studied the problem of learning stationary policies for finite-horizon CMDPs using Linear programming. We developed two types of algorithms and analyzed their sample complexity results—Optimistic-GMBL and online-CRL. Our most prominent result states a logarithmic increase in sample complexity compared to unconstrained regime. In the next chapter, we will study the case of infinite horizon CMDPs.

3. MODEL-BASED REINFORCEMENT LEARNING FOR INFINITE-HORIZON DISCOUNTED CONSTRAINED MARKOV DECISION PROCESSES[*]

3.1 Introduction

The previous chapter focuses on learning finite horizon (episodic) Constrained Markov Decision Processes (CMDPs). However, there are many physical systems that are consistent with a discounted infinite horizon reward. Therefore, in this chapter, we aim to develop simple algorithms to learn near-optimal policies for an infinite horizon CMDP without knowing the system parameters. Although, a regular model-based RL algorithm attempts to collect as few samples as possible to quickly solve for the optimal policy, minimizing the number of samples taken is even more essential in the CMDP setting. This requirement is due to the existence of constraints in the CMDP setting, and it might be important to violate them as few times as possible while maximizing the objective of the system. Therefore, the behavior of a system with respect to (w.r.t) both objective maximization and safety violation over time is a crucial performance metric for a proposed RL algorithm for CMDPs.

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

Our contributions are mainly threefold:

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

[*] Reprinted with permission from [2]

has a solution.

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

(iii) We characterize PAC-type sample complexity bounds for both algorithms, accounting for both objective maximization and constraint satisfaction.

Intuitively, the model constructed by these algorithms must be more accurate than models created by unconstrained counterparts, which conjecture our main results are consistent with. Furthermore, a comparison of our main findings with lower bounds on sample complexity of MDPs [6, 7] shows an increase in our results by a logarithmic factor in the number of constraints and the size of the state space. However, there is no earlier work on lower bound of sample complexity of learning CMDPs to our best knowledge.

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

Related Work: There are many articles studying the problem of controlling CMDPs with an algorithmic approach and control-theoretic view [4, 8, 9, 10, 11, 12]. The results take the form of proving asymptotic convergence of their proposed methods under the assumption of the known model. There are also extensions of this approach to the context of an unknown model, where the focus is still on asymptotic behavior [15, 16, 17, 18]. These studies use Lagrangian method to show zero duality gap asymptotically. Further, [19] also develops an algorithm based on the Lagrangian method, but with small eventual duality gap. Finally, empirical studies based on the Lagrangian method have also been presented [20].

There are also studies on the constrained bandit case. Although bandits are not MDPs per se, they are strongly related to them. Articles such as [21, 22, 23] consider such constraints, either in a knapsack sense, or on the type of controls that may be applied in a linear bandit context.

More related to our work theme are parallel studies on CMDPs. For example, [24] and [25] provide results with the assumption of unknown reward functions, with either a known or deterministic transition kernel. There are other works [26] focusing on proving asymptotic convergence without providing a bound on learning rate. Finally, closest related work to this article is [5] which explores algorithms similar to ours in finite-horizon setting, but concentrating on characterizing objective and constrained regret bounds. Now, regret and sample complexity bounds are not directly translatable [27], and converting their regret bounds to our setting gives relatively weak sample complexity bounds. Specifically, our main results with logarithmic increase in sample complexity with the number of constraints differentiates our work.

3.2 Notation and Problem Formulation

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

Assumption 3. *State and action sets S and A are assumed to be finite with cardinalities $|S|$ and $|A|$. In addition, the immediate cost and immediate reward $r(s, a)$ are assumed to be taken from the interval $[0, 1]$. Number of constraints is also assumed to be N which for each $i \in \{1, \ldots, N\}, \bar{C}_i \in [0, \bar{C}_{\max}]$.*

Now, we define a stationary policy $\pi : S \times A \to [0, 1]^{|A|}$ as a mapping from state-action space

$S \times A$ to set of probability vectors defined over action space in order to choose an action at any time-step t. Henceforth, $\pi(s, a)$ represents the probability of choosing the action a when the system is at state s. Also, $a \sim \pi(s, \cdot)$ means that action a is chosen according to stationary policy π while being at state s.

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

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

$$V^\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t); a_t \sim \pi(s_t, \cdot), s_{t=0} = s_0], \tag{3.1}$$

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

$$\begin{aligned} \sigma^2_{V_\pi}(s) &= \gamma^2 \mathbb{E}[(V^\pi(s_{t+1}) - P_\pi V^\pi(s))^2] \\ &= \gamma^2 P_\pi[(V^\pi - P_\pi V^\pi)^2](s). \end{aligned} \tag{3.2}$$

Analogous to the definition of the value function (3.1), the i^{th} constraint function under policy π is defined as

$$C_i^\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t c(i, s_t, a_t); a_t \sim \pi(s_t, \cdot), s_{t=0} = s_0]. \tag{3.3}$$

Again, the local variance of i^{th} constraint function under policy π, i.e. $\sigma^2_{C_i^\pi}$ is defined similar to local variance of value function (4.2).

Eventually, the general infinite-horizon CMDP problem is

$$\max_{\pi} V^{\pi}(s_0) \text{ s.t. } C_i^{\pi}(s_0) \leq \bar{C}_i, \quad \forall i \in \{1, \ldots, N\}. \tag{3.4}$$

Assumption 4. *We assume that the CMDP problem of* (3.4) *is feasible with optimal policy π^* and optimal solution $V^*(s_0) = V^{\pi^*}(s_0)$.*

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

Constrained-RL Problem: The Constrained RL problem formulation is identical to the CMDP optimization problem of (3.4) with one difference. Here, we are not aware of the values of the transition kernel P.[†] We desire to provide model-based algorithms and determine the sample complexity results in a PAC sense, which is defined as follows:

Definition 2. *For an algorithm $\mathcal{A}$, sample complexity is the number of samples that $\mathcal{A}$ requires to achieve*

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

for a given ϵ and δ.

Note that with this definition, we include both objective maximization and constraint violations as opposed to the traditional definition that only considers the objective [28].

3.3 Sample Complexity Result of Generative Model Based Learning

Generative model based learning is a well known approach to learn an optimal policy for an MDP. However, naive application of this approach to CMDPs may not end with a feasible solution. Hence, we explore the generative model based approach for CMDPs, and propose a generative model based CMDP learning algorithm called Generative Model-Constrained RL (GM-CRL).

[†] 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.

According to GM-CRL, each state-action pair is sampled n number of times uniformly across all state-action pairs, the number of times each transition occurs $n(s', s, a)$ for each next state s' is counted, and an empirical model of transition kernel denoted by $\widehat{P}(s'|s,a) = \frac{n(s',s,a)}{n}\ \forall(s', s, a)$ is constructed.

Unlike MDP problem formulation, there is no guarantee such that CMDP problem formulation w.r.t. $\widehat{P}$ is feasible. In order to resolve the feasibility concern, we expand the space of transition kernels to include the true transition kernel P, noting that the CMDP problem w.r.t. P is feasible from Assumption 4. The algorithmic layout of this approach is as follows. GM-CRL creates a class of CMDPs using the empirical model. This class is denoted by $\mathcal{M}_{\delta_P}$ and contains CMDPs with identical reward, cost matrices, $\bar{C}$, initial state s_0 and discount factor of the true CMDP, but with transition kernels close to true model. This class of CMDPs is defined as

$$\mathcal{M}_{\delta_P} := \{M' : r'(s,a) = r(s,a), c'(i,s,a) = c(i,s,a), \gamma' = \gamma, \tag{3.5}$$

$$|P'(s'|s,a) - \widehat{P}(s'|s,a)| \leq$$

$$\min\Big(\sqrt{\frac{2\widehat{P}(s'|s,a)(1-\widehat{P}(s'|s,a))}{n}\log\frac{4}{\delta_P}} + \frac{2}{3n}\log\frac{4}{\delta_P}, \sqrt{\frac{\log 4/\delta_P}{2n}}\Big)\forall s, a, s', i\}, \tag{3.6}$$

where δ_P is defined in Algorithm 3. Note that for any $M' \in \mathcal{M}$, objective function $V'^{\pi}(s_0)$ and cost functions $C_i'^{\pi}(s_0)$ are computed w.r.t. the corresponding transition kernel P' according to equations (3.1) and (3.3) respectively.

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

$$\max_{\pi, M' \in \mathcal{M}_{\delta_P}} V'^{\pi}(s_0) \quad \text{s.t.} \quad C_i'^{\pi}(s_0) \leq \bar{C}_i \ \ \forall i. \tag{3.7}$$

To solve the problem of (6.33), GM-CRL uses Extended Linear Programming, or **ELP**. This method takes $\mathcal{M}_{\delta_P}$ as input and gives $\tilde{\pi}$ for the optimal solution. The description of ELP is provided in supplementary materials. Algorithm 3 describes GM-CRL.

Algorithm 3 GM-CRL

```
Input: accuracy $\epsilon$ and failure tolerance $\delta$.
Set $\delta_P = \frac{\delta}{5(N+2)|S|^3|A|}$.
Set $n(s', s, a) = 0 \; \forall (s, a, s')$.
for each $(s, a) \in S \times A$ do
    Sample $(s, a), n = \frac{1152(\log 2)^2\gamma^2}{\epsilon^2(1-\gamma)^3}|S|^2|A| \log \frac{4}{\delta_P}$ and update $n(s', s, a)$.
    $\widehat{P}(s'|s, a) = \frac{n(s',s,a)}{n} \; \forall s'$.
Construct $\mathcal{M}_{\delta_P}$ according to (3.5).
Output $\tilde{\pi} = \text{ELP}(\mathcal{M}_{\delta_P})$.
```

3.3.1 PAC Analysis of GM-CRL

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

Theorem 3. *Consider any infinite-horizon CMDP $M = \langle S, A, P, r, c, \bar{C}, s_0, \gamma \rangle$ satisfying assumptions 3 and 4, and CMDP problem formulation of* (3.4). *Then, for any $\epsilon \in (0, \frac{0.22\gamma}{\sqrt{|S|(1-\gamma)}})$ and $\delta \in (0, 1)$, algorithm 3 creates a model CMDP $\tilde{M} = \langle S, A, \tilde{P}, r, c, \bar{C}, s_0, \gamma \rangle$ and outputs policy $\tilde{\pi}$ such that*

$$\mathbb{P}(V^{\tilde{\pi}}(s_0) \geq V^{\pi^*}(s_0) - \epsilon \;\; \textit{and}$$

$$C_i^{\tilde{\pi}}(s_0) \leq \bar{C}_i + \epsilon \;\; \forall i \in \{1, 2, \ldots, N\}) \geq 1 - \delta,$$

with at least total sampling budget of

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

The proof of Theorem 12 is different from the traditional analysis framework of unconstrained RL [6] in the following manner. First, consider the role played by optimism in model construction. The notion of optimism is not required for learning unconstrained MDPs with generative models, because any estimated model is always feasible [29]. However, there is no such guarantee for a general CMDP problem formulation [4]. Specifically, simply substituting the true kernel P by the estimated one $\widehat{P}$ is not appropriate, since there is no assurance of feasibility of that problem. Hence, GM-CRL converts the CMDP problem under the estimated transition kernel to an optimistic planning problem (6.33) 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 [4]. We require a uniform PAC result over set of all policies and set of value and constraint functions, which in turn leads to quadratic sample complexity in the size of state space; i.e., a scaling of $O(|S|^2)$.

Corollary 2. *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 12. Then we sketch the proof of this theorem. The detailed proofs are provided in supplementary materials.

First, we show that true CMDP lies inside the $\mathcal{M}_{\delta_P}$ with high probability, w.h.p. Hence, the problem (6.33) is feasible w.h.p., since the original CMDP problem is assumed to be feasible according to Assumption 4.

Lemma 6.

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

Proof Sketch: Fix a state-action pair (s, a) and next state s'. Then, according to combination of Hoefding's inequality [30] and empirical Bernstein's inequality [31], we obtain that each $P(s'|s, a)$ is inside the confidence set defined by (6.32) with probability at least $1 - \delta_P$. Applying the union bound yields the result. □

Now, we present the core lemma required for proving Theorem 12 and its proof sketch. Using this lemma, we bound the mismatch in objective and constraint functions when we have n number of samples from each (s, a). This bound applies uniformly over the set of policies and set of value and constraint functions. The result also enables us to bound the objective and constraint functions individually. Then we apply the union bound on all objective and constraint functions. This process is the reason why the number of constraints appear logarithmically in the sample complexity result.

Lemma 7. *Let* $\delta_P \in (0, 1)$. *Then, if* $n \geq 11819 \frac{|S|^2 \log 4/\delta_P}{(1-\gamma)^2}$, *under any policy* π

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

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

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

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

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

***Proof Sketch of Theorem 12*:** From Lemma 6, we know that the optimistic planning problem (6.33) is feasible w.h.p. Hence, we can obtain an optimistic policy $\tilde{\pi}$. The rest of this proof consists of two major parts.

First, we prove $\epsilon-$optimality of objective function w.h.p. Considering policy π^* we obtain $|V^{\pi^*}(s_0) - \tilde{V}^{\pi^*}(s_0)| \leq O(\sqrt{\frac{|S|}{(1-\gamma)^3 n}})$ w.h.p. by means of Lemma 28. Similarly, $|V^{\tilde{\pi}}(s_0) -$

$\tilde{V}^{\tilde{\pi}}(s_0)| \leq O(\sqrt{\frac{|S|}{(1-\gamma)^3 n}})$ w.h.p. Next, we use the fact that $\tilde{V}^{\pi^*}(s_0) \leq \tilde{V}^{\tilde{\pi}}(s_0)$ and obtain

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

Next, we show that each constraint is violated at most by ϵ w.h.p. Here, we use the second part of Lemma 28 to bound constraint violation. Thus, for each $i \in \{1, \ldots, N\}$ we have $|C_i^{\tilde{\pi}}(s_0) - \tilde{C}_i^{\tilde{\pi}}(s_0)| \leq O(\sqrt{\frac{|S|}{(1-\gamma)^3 n}})$ w.h.p. Also, we know that $\tilde{C}_i^{\tilde{\pi}}(s_0) \leq \bar{C}_i$, since $\tilde{\pi}$ is solution of the ELP. Hence, we obtain

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

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

3.4 Sample Complexity Result of Online Learning

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

Upper Confidence Constrained-RL, or UC-CRL described in Algorithm 4, is an online method proceeding over time-steps. At each time-step t, UC-CRL constructs an empirical model $\widehat{P}$ using state-action visitation frequencies, i.e., $\widehat{P}(s'|s,a) = \frac{n(s',s,a)}{n(s,a)}$, where $n(s',s,a)$ and $n(s,a)$ are visitation frequencies. Then, we use $\widehat{P}$ to create a confidence interval around each element $\widehat{P}(s'|s,a)$ using same concentration inequalities of GM-CRL defined by (6.32). Next, UC-CRL constructs set of infinite-horizon CMDPs $\mathcal{M}_t$ which any CMDP $M' \in \mathcal{M}_t$ has identical discount factor and reward and cost matrices to the true CMDP M, but different transition kernels from the concentration inequalities. $\mathcal{M}_t$ is identical to $\mathcal{M}_{\delta_P}$ except for the use of $n(s,a)$ instead of n. Thus the class of CMPDs is defined as below at each time-step t :

$$\begin{aligned}
&\mathcal{M}_t := \{M' : r'(s,a) = r(s,a), c'(i,s,a) = c(i,s,a), \gamma' = \gamma,\\
&|P'(s'|s,a) - \widehat{P}(s'|s,a)| \leq\\
&\min\Big(\sqrt{\frac{2\widehat{P}(s'|s,a)(1-\widehat{P}(s'|s,a))}{n(s,a)}}\log\frac{4}{\delta_1} + \frac{2}{3n(s,a)}\log\frac{4}{\delta_1}, \sqrt{\frac{\log 4/\delta_1}{2n(s,a)}}\Big)\ \ \forall s,s',a,i\},
\end{aligned} \tag{3.8}$$

where δ_1 is defined in Algorithm 4.

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

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

This problem is identical to the problem of (6.33), except for substituting $\mathcal{M}_{\delta_P}$ with $\mathcal{M}_t$. Here, for any $M' \in \mathcal{M}_t$, $V'^{\pi}(s_0)$ and $C_i'^{\pi}(s_0)$ are computed according to (3.1) and (3.3) w.r.t. underlying transition kernel P', respectively.

Algorithm 4 UC-CRL

```
Input: accuracy $\epsilon$ and failure tolerance $\delta$.
Set $m$ according to (6.35) and (6.36).
Set $t = 1, w_{\min} = \frac{\epsilon(1-\gamma)}{4|S|}, U_{\max} = |S|^2|A|m, \delta_1 = \frac{\delta}{4(N+1)|S|U_{\max}}$.
Set $n(s,a) = n(s',s,a) = 0 \ \forall s,s' \in S, a \in A$.
while there is $(s,a)$ with $n(s,a) < \frac{|S|m}{1-\gamma}$ do
    $\widehat{P}(s'|s,a) = \frac{n(s',s,a)}{n(s,a)} \ \forall (s,a)$ with $n(s,a) > 0$ and $s' \in S$.
    Construct $\mathcal{M}_t$ according to (3.8).
    $\tilde{\pi}_t = \text{ELP}(\mathcal{M}_t)$.
    $a_t \sim \tilde{\pi}_t(s_t), s_{t+1} \sim P(\cdot|s_t,a_t)$
    if $n(s_t,a_t) < \frac{|S|m}{1-\gamma}$ then
        $n(s_t,a_t)++, n(s_{t+1},s_t,a_t)++$.
    $t++$
```

UC-CRL is inspired by the infinite-horizon algorithm UCRL$-\gamma$ [32] and its finite-horizon equivalent UCFH [7] with differences. Similar to UCRL$-\gamma$, Algorithm 4 uses a combination of the empirical Bernstein's and Hoeffding's inequalities. These concentration inequalities allow us to ensure linearity of constraints (i.e., we can indeed use an extended linear program to solve for the constrained optimistic policy). However, the constraints of UCFH contain non-linear expressions preventing us from employing ELP. Furthermore, unlike UCRL$-\gamma$ and UCFH, Algorithm 4 updates the model at each time-step rather than at the beginning of long phases. This procedure allows for faster model construction. Finally, since we are solving a CMDP, this algorithm utilizes ELP instead of Extended Value Iteration which is used by UCRL$-\gamma$.

3.4.1 PAC Analysis of UC-CRL

We now present the PAC bound of Algorithm 4.

Theorem 4. *Consider CMDP $M = \langle S, A, P, r, c, \bar{C}, s_0, \gamma \rangle$ satisfying assumptions 3 and 4. For any $0 < \epsilon, \delta < 1$, under UC-CRL we have:*

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

for all but at most

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

time-steps.

We follow an approach motivated by [32] and its finite-horizon version [7] to prove Theorem 4. However, there are several differences in our technique, and we need to accommodate the frequent model update in our proof. We will show that, unlike existing approaches, we can update the model at each time-step, without increasing the sample complexity. Thus, we are able to obtain PAC bounds that match the unconstrained case, and only increase by logarithmic factor with the number of constraints.

There are also recent works on characterizing the regret of constrained-RL in a finite-horizon setting [5] with an algorithm similar to Algorithm 4. An important emerging question is whether one can immediately convert these regret results into sample complexity bounds? A naive translation of the regret bounds of [5] would give us a PAC result $\tilde{O}(\frac{|S|^2|A|H^4}{\epsilon^2})$. For comparing finite-horizon setting with infinite-horizon one, we can replace H with $\frac{1}{1-\gamma}$ to obtain a PAC result for the equivalent infinite-horizon algorithm. Considering this, the approach followed by [5] gives a PAC bound which is looser than our result by a factor of $\frac{1}{(1-\gamma)^2}$. Therefore, this alternative option does not lead to the strong bounds that we are able to obtain, and matches existing PAC results of the unconstrained case.

Now, we present the notions of *knownness* and *importance* for state-action pairs and base our proof on these notions. Then we present the key lemmas needed for proving Theorem 4. Finally, we provide a proof sketch for Theorem 4. The detailed analysis is provided in supplementary materials.

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

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

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

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

Then, the *importance* ι_t of (s, a) at time-step t is defined as its relative weight compared to $w_{\min} :=$

$\frac{\epsilon(1-\gamma)}{4|S|}$ on a log-scale

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

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

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

which indicates how often (s,a) has been observed relative to its importance. Value of m is defined in Algorithm 4. Now, we can categorize $(s,a)-$pairs into subsets

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

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

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

Lemma 8. *$M \in \mathcal{M}_t$ for all time-steps t with probability at least $1 - \frac{\delta}{2(N+1)}$.*

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

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

Lemma 9. *Suppose E is the number of time-steps t for which there are κ and ι with $|X_{t,\kappa,\iota}| > \kappa$, i.e. $E = \sum_{t=1}^{\infty} \mathbb{I}\{\exists(\kappa,\iota) : |X_{t,\kappa,\iota}| > \kappa\}$ and let*

$$m \ge \frac{4}{\epsilon(1-\gamma)^3} \log \frac{2(N+1)E_{\max}}{\delta}, \tag{3.9}$$

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

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

Finally, the next lemma bounds the mismatch between objective and constraint functions of the optimistic model and true model. This lemma functions similarly to Lemma 28 for GM-CRL. It provides a uniform PAC result over value and constraint functions. Hence, it enables us to have individual PAC results for any objective and constraint functions. As discussed in GM-CRL section, this process is responsible for a $\log N$ increase in the PAC result.

Lemma 10. *Assume $M \in \mathcal{M}_t$. If $|X_{t,\kappa,\iota}| \le \kappa$ for all (κ,ι) and $0 < \epsilon \le 1$ and*

$$\begin{aligned} m \ge & 1280 \frac{|S|}{\epsilon^2(1-\gamma)^2} (\log_2 \log_2 (\frac{1}{1-\gamma}))^2 \log_2^2 \Big(\frac{8|S|^2}{\epsilon(1-\gamma)^2} \Big) \\ & \times \log \frac{4}{\delta_1}, \end{aligned} \tag{3.10}$$

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

Proof Sketch: First, we show $|\tilde{P}(s'|s,a) - P(s'|s,a)| \le O(\sqrt{\frac{P(s'|s,a)(1-P(s'|s,a))}{n}})$ for each s', s, a. Then we prove that at each time-step $t, (P_\pi - \tilde{P}_\pi)V^\pi(s) \le O(\sqrt{\frac{|S|}{n}}\sigma_{V^\pi}(s))$. Next we partition the state-action based on knownness, i.e., whether they belong to X_t or not. By using all

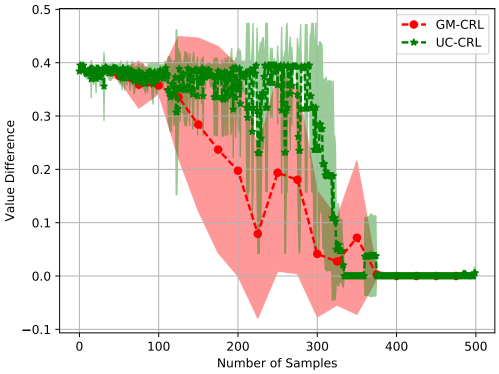

Figure 3.1: Value Difference. Reprinted with permission from [2]

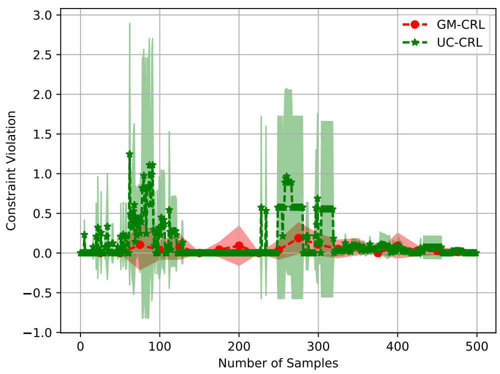

Figure 3.2: Constraint Violation. Reprinted with permission from [2]

bounds and sequence of CMDPs, we obtain a bound on $|\tilde{V}^\pi(s_0) - V^\pi(s_0)|$. Eventually, we use the definition of weights to get the final result. This procedure is also applicable to each constraint function i. □

***Proof Sketch of Theorem 4*:** We first use Lemma 30 and show that true CMDP is admissible ,i.e. $M \in \mathcal{M}_t$ for every time-step, w.p. at least $1 - \frac{\delta}{2(N+1)}$. Hence, the optimistic planning problem becomes feasible and an optimistic policy $\tilde{\pi}_t$ exists w.h.p. Further, we use Lemma 33 to bound the number of time-steps where $|X_{t,\kappa,\iota}| > \kappa$ w.h.p. Thus, for other time-steps where $|X_{t,\kappa,\iota}| \leq \kappa$, we apply Lemma 35 we show that objective function is $\epsilon-$optimal and all constraint functions are violated by ϵ. Eventually, we obtain the result by means of union bound. □

3.5 Experimental Results

We conduct experiments on CMDPs similar to a grid world MDP, wherein each square is the location of the agent. Here, we consider a 5×5 with 25 states and 4 actions. The goal of the agent is to reach a final state starting from a fixed initial state in least number of steps possible. When the agent arrives at the final state, it receives a reward of 1. It is then transitioned to initial state at the next time-step. 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. In this experiment, we consider restricting the number of moves to the right.

We simulate GM-CRL and UC-CRL. In order to compare them, we consider two performance metrics. The first is the difference in value function calculated by

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

where π' is whether GM-CRL or UC-CRL. The second performance metric is of constraint violation, which is calculated by

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

since we have one constraint in each scenario. Further, we average each data point on every figure over 10 runs.

As seen in the Figure 3.1 both GM-CRL and UC-CRL reach the optimal value simultaneously, although UC-CRL incurs higher error with smaller number of samples. We discover that both algorithms performance is almost indistinguishable w.r.t. value difference metric. Similar behavior appears in Figure 3.2, which illustrates constraint violation. These results match our theoretical findings where the PAC results of both algorithms are much the same.

3.6 Conclusion

This chapter covers the learning problem of infinite horizon CMDPs. Similar to previous chapter, we designed two algorithms based on LP approach to learn an infinite horizon CMDP. Further, we showed that there is also a logarithmic increase in sample complexity result compared to unconstrained MDPs.

However, the RL algorithms presented in previous and current chapters are computationally expensive due to the need for solving an LP at each policy update. In the next chapter, we focus on developing more efficient RL algorithms for CMDPs in terms of computational complexity.

4. MODEL-BASED COMPUTATIONALLY EFFICIENT REINFORCEMENT LEARNING FOR FINITE-HORIZON CONSTRAINED MARKOV DECISION PROCESSES

4.1 Introduction

Prior to this chapter, we concentrated on analyzing the sample complexity of learning algorithms for CMDPs using LP approach. However, solving an LP requires high amount of computation power. Hence in this chapter, our objective is to design computationally efficient algorithms to solve CMDP problems where the dynamics are not known. RL algorithms with provable sample complexity results have been proposed to solve this class of problems in prior work [33]. However, those algorithms are based on a Linear programming (LP) approach, which is computationally expensive, particularly in the CMDP scenario. Here, we design different algorithms based on Lagrangian approach [4] that much reduces the cost of computation. Further, we analyze the sample complexity of the proposed algorithms in a probably-approximately-correct (PAC) sense.

Main Contributions: Our main contributions are as follows:

(i) We present two model-based RL algorithms with two different settings, (i) a generative setting where we initially collect samples from the environment and create a model, and (ii) an online setting where the samples are collected by interacting with the environment.

(ii) Both of the algorithms follow the Lagrangian approach which transforms the CMDP problem to a min-max problem. This process contributes to reduction in computational complexity of the algorithms compared to their LP-based counterparts [33]. Here, we guarantee the transformation of the CMDP problem to min-max problem yields an identical solution.

(iii) We provide PAC-type sample complexity results for both algorithms accounting for the differences in the corresponding Lagrangian functions. Comparing the presented sample complexity result with the LP-based algorithms [33] shows that Lagrangian-based algorithms suffer from higher sample complexity, while being computationally efficient. This is because the Lagrangian approach introduces new variables, causing enlargement of the reward space.

Related Work: There is extensive work in the space of CMDP focusing on problems of control, and many of the algorithmic approaches and applications have taken a control-theoretic view [4, 8, 9, 10, 11, 12]. These articles assume that the dynamics of the environment is known, and show the asymptotic convergence of their proposed solutions

Asymptotic convergence of algorithms for solving CMDPs under unknown model is also studied by [15, 16, 17, 18] under Lagrangian methods to show zero eventual duality gap. Another work [19] also proposes an algorithm based on Lagrangian method with small eventual duality-gap.

Closest to our approach are parallel works on CMDPs. For example, [24] and [25] present results for CMDP problems with unknown reward functions. Other work [5] presents similar methods to ours but with a focus on bounding the regret, instead of sample complexity that is our focus of our work. Because the nature of the metric [5] considers for regret and our metric for sample complexity are different, we cannot directly translate their regret bound to a sample complexity bound that we desire. Further, [27] shows that there is only a loose relation between regret and sample complexity, which means that a separate analysis on sample complexity is needed. Finally, [33] designs and analyzes similar algorithms to learn CMDP problems. However, their approach is basically different since they are using LP-based algorithms that have high computational complexity.

4.2 Problem Formulation and Notation

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 $P(s'|s, a)$. 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 $M = \langle S, A, P, r, c, s_0, \bar{C}, H \rangle$.

Assumption 5. *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 $i \in \{1, \ldots, N\}, \bar{C}_i \in [\bar{C}_{\min}, \bar{C}_{\max}]$.*

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 stationary[*] 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 $P_\pi^h f(s) = P_\pi P_\pi \ldots P_\pi f$, 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

$$V_t^\pi(s) = \mathbb{E}[\sum_{h=t}^{H-1} r(s_h, a_h); a_h \sim \pi(s_h, \cdot, h), s_t = s], \tag{4.1}$$

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

$$\sigma_h^{\pi^2}(s) = \mathbb{E}[(V_{h+1}^\pi(s_{h+1}) - P_\pi V_{h+1}^\pi(s))^2]. \tag{4.2}$$

Similar to the definition of the value function (4.1), the i^{th} constraint function at time t under

[*] Here, stationary means that the policy does not change over episodes. However, it can be a function of the time-step within the episode.

policy π is formulated as

$$C^{\pi}_{i,t}(s) = \mathbb{E}[\sum_{h=0}^{H-1} c(i, s_h, a_h); a_t \sim \pi(s_h, \cdot, h), s_t = s]. \tag{4.3}$$

Again, the local variance of i^{th} constraint function at time-step h under policy π, i.e. $\sigma^{\pi^2}_{i,h}$ is defined similar to local variance of value function (4.2).

Finally, the general finite-horizon CMDP problem is

$$\max_{\pi} V^{\pi}_0(s_0) \text{ s.t. } C^{\pi}_{i,0}(s_0) \leq \bar{C}_i, \quad \forall i \in \{1, \ldots, N\}. \tag{4.4}$$

Assumption 6. *We assume that there exists some policy π that satisfies the constraints in* (4.4)*. Hence, this CMDP problem is feasible with optimal policy π^* and optimal solution $V^*_0(s_0) = V^{\pi^*}_0(s_0)$.*

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

Assumption 7. *Let assume there exists a policy π_0 incurring zero return and cost, i.e.*

$$V^{\pi_0}_0(s_0) = 0 \textit{ and } C^{\pi_0}_{i,0}(s_0) = 0 \ \forall i$$

for every transition kernel.

One of the ways to solve the optimization problem (4.4) is via **Dual Decomposition** technique as stated in [4]. Thus, we introduce an $N-$dimensional Lagrange multiplier vector λ where the i^{th} element λ_i corresponds to the i^{th} constraint and it is non-negative, $\lambda_i \geq 0$. So, the Lagrangian function is

$$L(\pi, \lambda) = V^{\pi}_0(s_0) - \sum_{i=1}^{N} \lambda_i (C^{\pi}_{i,0}(s_0) - \bar{C}_i). \tag{4.5}$$

Rearranging the above equation yields

$$L(\pi, \lambda) = V_0^{\pi(\lambda)}(\lambda) + \sum_{i=1}^{N} \lambda_i \bar{C}_i, \tag{4.6}$$

where $V_0^{\pi(\lambda)}(\lambda)$ is the value function defined as

$$V_0^{\pi}(\lambda) = V_0^{\pi}(s_0) - \sum_{i=1}^{N} \lambda_i C_{i,0}^{\pi}(s_0) = \mathbb{E}[\sum_{h=0}^{H-1}(r(s_h, a_h) - \sum_{i=1}^{N} \lambda_i c(i, s_h, a_h)); a_h \sim \pi(s_h, \cdot, h)]. \tag{4.7}$$

So, we can create a new reward matrix $r_c(\lambda, s, a)$ as

$$r_c(\lambda, s, a) = r(s, a) - \sum_{i=1}^{N} \lambda_i c(i, s, a). \tag{4.8}$$

Then, the dual function is

$$D(\lambda) = \max_{\pi} L(\pi, \lambda), \tag{4.9}$$

and dual problem would be

$$\min_{\lambda \geq 0} D(\lambda), \tag{4.10}$$

where $\lambda \geq 0$ means every element being non-negative.

We denote the optimal solution of (4.10) by λ^*. Here, we consider CMDP problems where there is no **duality-gap** i.e. $D(\lambda^*) = V_0^{\pi^*}(s_0)$. Assumption 6 guarantees that we obtain 0 duality-gap according to Slater's condition [4].

To solve the dual problem of (4.10), [4] applies gradient descent method. In this method, the vector λ is initialized by an arbitrary value. The CMDP problem would turn to an MDP problem when the λ is fixed. Hence, a policy w.r.t fixed λ, $\pi^*(\lambda)$, is computed. Next, the gradient w.r.t. $\pi^*(\lambda)$ is calculated. Finally, the new λ is calculated and this procedure is carried on until $D(\lambda)$ converges to $D(\lambda^*)$.

Constrained-RL Problem: The Constrained RL problem formulation is identical to the CMDP optimization problem of (4.4), but without being aware of values of transition kernel P.[†] Our goal is to provide model-based algorithms and determine the sample complexity results in a PAC sense, which is defined as follows:

Definition 3. *For an algorithm $\mathcal{A}$, sample complexity is the number of samples that $\mathcal{A}$ requires to achieve*

$$\mathbb{P}\Big(L(\pi(\mathcal{A}), \lambda(\mathcal{A})) \leq L(\pi^*, \lambda^*) + \epsilon\Big) \geq 1 - \delta$$

for a given ϵ and δ where $\pi(\mathcal{A})$ and $\lambda(\mathcal{A})$ are outcomes of the algorithm.

4.3 GMBL-Dual

In this section, we introduce a generative model based CMDP learning algorithm called Generative Model Based Learning-Dual, or GMBL-Dual. According to GMBL-Dual, 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 $\widehat{P}(s'|s, a) = \frac{n(s',s,a)}{n} \; \forall (s', s, a)$.

We now consider a different constrained MDP that is identical to the CMDP defined in Section 4.2 except that its transition kernel is $\widehat{P}$ instead of P. The expectation w.r.t. $\widehat{P}$ is denoted by $\widehat{\mathbb{E}}[\cdot]$. We define the quantities $\widehat{V}_0^\pi(s_0, \lambda)$ and $\widehat{C}_{i,0}^\pi(s_0)$ in the same way as in (4.7) and (4.3) but by replacing $\mathbb{E}$ by $\widehat{\mathbb{E}}$. The quantities $\widehat{L}(\pi, \lambda), \widehat{D}(\lambda)$ can also now be defined in a similar way as in (4.6) and (4.9) by replacing $V_0^\pi(s_0, \lambda)$ with $\widehat{V}_0^\pi(s_0, \lambda)$. The optimal dual variable $\widehat{\lambda}^*$ is defined as $\widehat{\lambda}^* = \arg\min_\lambda \widehat{D}(\lambda)$. We also define

$$\widehat{\pi}(\lambda) = \arg\max_\pi \widehat{V}_0^\pi(\lambda), \;\; \widehat{V}_0^*(\lambda) = \widehat{V}_0^{\widehat{\pi}(\lambda)}(\lambda). \tag{4.11}$$

Note that $\widehat{\pi}(\lambda)$ and $\widehat{V}_0^*(\lambda)$ can be computed by standard finite horizon dynamic programming [34],

[†] 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.

and we omit the details.

The GMBL-Dual algorithm is summarized in Algorithm

Algorithm 5 GMBL-Dual

```
Input: accuracy $\epsilon$ and failure tolerance $\delta$.
Set $\delta_P = \frac{\delta}{18|S|^2|A|H}$.
Initialize $\lambda_i(0) = \widehat{\lambda}_i = 0 \forall i \in [1, N]$.
Set $n(s', s, a) = 0\ \forall (s, a, s'), K$and$\alpha$ according to (4.13)
for each $(s, a) \in S \times A$ do
    Sample $(s, a), n = \frac{128}{\epsilon^2}|S|H^3(1 + NB_\lambda)^2 \log \frac{72|S|^2|A|H}{\delta}$ and update $n(s', s, a)$.
    $\widehat{P}(s'|s, a) = \frac{n(s',s,a)}{n}\ \forall s'$.
for $k = 0, 1, \ldots, K$ do
    $\widehat{\pi}_k$ = BackwardInduction($M(\lambda)$).
    $\lambda_i(k + 1) = \Pi_\Lambda(\lambda_i(k) - \alpha(\bar{C}_i - \widehat{C}^{\widehat{\pi}_k}_{i,0}(s_0)))$ ‡
    $\widehat{\lambda} += \lambda(k)$
$\widehat{\lambda} /= K$
Output $\widehat{\pi}$ = BackwardInduction($M(\widehat{\lambda})$)
```

We next present the sample complexity of GMBL-Dual.

Theorem 5. *For any $\delta \in (0, 1)$ and $\epsilon \in (0, \frac{2}{9}(1 + NB_\lambda)\sqrt{\frac{H}{|S|}})$, GMBL-Dual algorithm with*

$$n(\epsilon, \delta) \geq \frac{128|S|H^3(1 + NB_\lambda)^2 \log (72|S|^2|A|H/\delta)}{\epsilon^2} \tag{4.12}$$

and parameters

$$K = \Big(\frac{3\sqrt{N}H(H + \bar{C}_{\max})}{\epsilon\bar{C}_{\min}}\Big)^2, \quad \alpha = \frac{\epsilon}{3N(H + \bar{C}_{\max})}, \tag{4.13}$$

achieves a $\widehat{\lambda}$ and $\widehat{\pi}$ such that

$$\mathbb{P}\left(|L(\widehat{\pi}, \widehat{\lambda}) - L(\pi^*, \lambda^*)| \leq \epsilon\right) \geq 1 - \delta.$$

The proof requires the use of multiple smaller results that we first present below, followed by their integration to yield the proof of the main theorem.

Proposition 1. *Let $\lambda_{\max} = \max_i \lambda_i^*$. Then under Assumption 7, $\lambda_{\max} < B_\lambda = \frac{H}{C_{\min}}$.*

Proof. The proof is provided in Appendix C. □

Lemma 11. *With the parameters K and α given by* (4.13)*, we obtain $|\widehat{D}(\widehat{\lambda}) - \widehat{D}(\hat{\lambda}^*)| \leq \epsilon/3$*

This follows from the standard rate of convergence analysis of projected subgradient descent algorithm for convex functions and proposition 1. For completeness we first reproduce that result. We use the following result.

Theorem 6. *[35] Let $g : \mathcal{X} \to \mathbb{R}^d$ be a convex function with $\|\nabla g\| \leq B_1$ where ∇g is the subgradient of g. Also, assume that the domain of $g(\cdot)$ is bounded, i.e. $\|x\| \leq B_2, \forall x \in \mathcal{X}$. Consider the projected gradient descent algorithm $x_{k+1} = \Pi_{\mathcal{X}}[x_k - \alpha \nabla g(x_k)]$ where $\Pi_{\mathcal{X}}$ is the projection operator. Then, with $\alpha = B_2/(B_1\sqrt{K})$,*

$$g(\frac{1}{K}\sum_{k=1}^{K} x_k) - g(x^*) \leq \frac{B_1 B_2}{\sqrt{K}}.$$

Proof of Lemma 11: We first show that the subgradient of $\widehat{D}(\cdot)$ at λ, denoted by $\nabla \widehat{D}(\lambda)$ is given by

$$\nabla \widehat{D}(\lambda) = [\bar{C}_i - \widehat{C}_{i,0}^{\widehat{\pi}(\lambda)}(s_0)]_i.$$

Indeed, for any given λ', λ,

$$\begin{aligned}
\widehat{D}(\lambda') = \max_{\pi} \widehat{L}(\pi, \lambda') \geq \widehat{L}(\widehat{\pi}(\lambda), \lambda') &= \widehat{V}_0^{\widehat{\pi}(s_0,\lambda)} + \sum_i \lambda_i'(\bar{C}_i - \widehat{C}_{i,0}^{\widehat{\pi}(\lambda)}(s_0)) \\
&= \widehat{V}_0^{\widehat{\pi}(s_0,\lambda)} + \sum_i \lambda_i(\bar{C}_i - \widehat{C}_{i,0}^{\widehat{\pi}(\lambda)}(s_0)) + \sum_i (\lambda_i' - \lambda_i)(\bar{C}_i - \widehat{C}_{i,0}^{\widehat{\pi}(\lambda)}(s_0)) \\
&= \widehat{D}(\lambda) + \sum_i (\lambda_i' - \lambda_i)(\bar{C}_i - \widehat{C}_{i,0}^{\widehat{\pi}(\lambda)}(s_0))
\end{aligned}$$

and hence the claim follows by the definition of subgradient.

In order to bound $\|\nabla \widehat{D}(\lambda)\|$, first note that $\widehat{C}_{i,0}^{\widehat{\pi}_f(\lambda)}(s_0) \leq H$. Also, $\bar{C}_i \leq \bar{C}_{\max}$. Hence, $\|\nabla \widehat{D}(\lambda)\| = \sqrt{N}(H + \bar{C}_{\max})$.

Now, considering Proposition 1, we project $\lambda(k)$ to set $[0, B_\lambda]$. Using Theorem 15, we get the desired result. □

Lemma 12. *Let* $\delta_P \in (0,1)$. *Then, if* $n \geq 2592|S|^2H^2 \log 4/\delta_P$, *for a given* $\lambda \in [0, B_\lambda]$ *under any policy* π

$$\|V_0^\pi(\lambda) - \widehat{V}_0^\pi(\lambda)\|_\infty \leq \sqrt{\frac{128|S|H^3(1+NB_\lambda)^2 \log 4/\delta_P}{n}}$$

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

Proof. The proof procedure is identical to proof of Lemma 2 of [33] with adjustment of $\|V_0^\pi(\lambda) - \widehat{V}_0^\pi(\lambda)\|_\infty \leq H(1 + NB_\lambda)$. □

Lemma 13. *Let* $\delta_P \in (0,1)$. *Then, if* $n \geq 2592|S|^2H^2 \log 4/\delta_P$, *for a given* $\lambda \in [0, B_\lambda]$

$$|\widehat{D}(\lambda) - D(\lambda)| \leq \sqrt{\frac{128|S|H^3(1+NB_\lambda)^2 \log 4/\delta_P}{n}}$$

w.p. at least $1 - 6|S|^2|A|H\delta_P$.

Proof. For a given λ, consider two policies $\pi(\lambda)$ and $\widehat{\pi}(\lambda)$. Then, according to Lemma 38 we have

$$V_0^{\pi(\lambda)}(s_0,\lambda) \leq \widehat{V}_0^{\pi(\lambda)}(s_0,\lambda) + \epsilon' \leq \widehat{V}_0^{\widehat{\pi}(\lambda)}(s_0,\lambda) + \epsilon' \tag{4.14}$$

w.p. at least $1 - 3|S|^2|A|H\delta_P$ where $\epsilon' = \sqrt{\frac{128|S|H^3(1+NB_\lambda)^2 \log 4/\delta_P}{n}}$. Please notice that the second inequality is due to the fact $\widehat{\pi}(\lambda) = \arg\max_\pi \widehat{V}_0^\pi(\lambda)$. Next, we have

$$\widehat{V}_0^{\widehat{\pi}(\lambda)}(s_0,\lambda) \leq V_0^{\widehat{\pi}(\lambda)}(s_0,\lambda) + \epsilon' \leq V_0^{\pi(\lambda)}(s_0,\lambda) + \epsilon' \tag{4.15}$$

w.p. at least $1 - 3|S|^2|A|H\delta_P$. Now, combining the two inequalities (6.49) and (6.50), we get

$$|\widehat{V}_0^{\widehat{\pi}(\lambda)}(s_0, \lambda) - V_0^{\pi(\lambda)}(s_0, \lambda)| \leq \epsilon'$$

w.p. at least $1 - 6|S|^2|A|H\delta_P$. Using the above inequality, we get

$$|\widehat{D}(\lambda) - D(\lambda)| = |\widehat{V}_0^{\widehat{\pi}(\lambda)}(s_0, \lambda) - V_0^{\pi(\lambda)}(s_0, \lambda))| \leq \epsilon'$$

w.p. at least $1 - 6|S|^2|A|H\delta_P$. Hence the proof is complete. □

Lemma 14. *Let $\delta_P \in (0, 1)$. Then, if $n \geq 2592|S|^2H^2 \log 4/\delta_P$, for a given $\lambda \in [0, B_\lambda]$*

$$|\widehat{D}(\widehat{\lambda}^*) - D(\lambda^*)| \leq \sqrt{\frac{128|S|H^3(1 + NB_\lambda)^2 \log 4/\delta_P}{n}}$$

w.p. at least $1 - 12|S|^2|A|H\delta_P$.

Proof. Since $\widehat{\lambda}^* = \arg\min_\lambda \widehat{D}(\lambda)$, then $\widehat{D}(\widehat{\lambda}^*) \leq \widehat{D}(\lambda^*)$. Next, we have

$$\widehat{D}(\lambda^*) \leq D(\lambda^*) + \epsilon' \tag{4.16}$$

w.p. at least $1 - 6|S|^2|A|H\delta_P$ where $\epsilon' = \sqrt{\frac{128|S|H^3(1+NB_\lambda)^2 \log 4/\delta_P}{n}}$ according to Lemma 39. Therefore, we have

$$\widehat{D}(\widehat{\lambda}^*) \leq D(\lambda^*) + \epsilon' \tag{4.17}$$

w.p. at least $1 - 6|S|^2|A|H\delta_P$. Taking identical steps, we get

$$D(\lambda^*) \leq \widehat{D}(\widehat{\lambda}^*) + \epsilon' \tag{4.18}$$

w.p. at least $1 - 6|S|^2|A|H\delta_P$. Finally, combining the inequalities (6.52) and (6.53) yields the result. □

Now, we are ready to prove Theorem 14.

Proof of Theorem 14: We expand the result of theorem 14

$$\begin{aligned}|L(\widehat{\pi}, \widehat{\lambda}) - L(\pi^*, \lambda^*)| &= |\sum_i \widehat{\lambda}_i \bar{C}_i + V_0^{\widehat{\pi}}(s_0, \widehat{\lambda}) - D(\lambda^*)| \\ &= |\sum_i \widehat{\lambda}_i \bar{C}_i + \widehat{V}_0^{\widehat{\pi}}(s_0, \widehat{\lambda}) - D(\lambda^*) + V_0^{\widehat{\pi}}(s_0, \widehat{\lambda}) - \widehat{V}_0^{\widehat{\pi}}(s_0, \widehat{\lambda})| \\ &\leq |\widehat{D}(\widehat{\lambda}) - D(\lambda^*)| + |V_0^{\widehat{\pi}}(s_0, \widehat{\lambda}) - \widehat{V}_0^{\widehat{\pi}}(s_0, \widehat{\lambda})|.\end{aligned}$$

First, we bound $|\widehat{D}(\widehat{\lambda}) - D(\lambda^*)|$ by and expanding it further

$$\begin{aligned}&|\widehat{D}(\widehat{\lambda}) - D(\lambda^*)| = |\widehat{D}(\widehat{\lambda}) - \widehat{D}(\widehat{\lambda}^*) + \widehat{D}(\widehat{\lambda}^*) - D(\lambda^*)| \\ &\leq |\widehat{D}(\widehat{\lambda}) - \widehat{D}(\widehat{\lambda}^*)| + |\widehat{D}(\widehat{\lambda}^*) - D(\lambda^*)| \leq \frac{\epsilon}{3} + \epsilon' \qquad (4.19)\end{aligned}$$

w.p. at least $1 - 12|S|^2|A|H\delta_P$ where $\epsilon' = \sqrt{\frac{128|S|H^3(1+NB_\lambda)^2 \log 4/\delta_P}{n}}$ according to Lemmas 11 and 40.

Next,

$$|V_0^{\widehat{\pi}}(s_0, \widehat{\lambda}) - \widehat{V}_0^{\widehat{\pi}}(s_0, \widehat{\lambda})| \leq \epsilon' \qquad (4.20)$$

w.p. at least $1 - 6|S|^2|A|H\delta_P$ according to Lemma 38.

Eventually, we combine two inequalities (6.54) and (6.55) and get

$$\begin{aligned}&|L(\widehat{\pi}, \widehat{\lambda}) - L(\pi^*, \lambda^*)| \\ &\leq \frac{\epsilon}{3} + 2\sqrt{\frac{128|S|H^3(1+NB_\lambda)^2 \log 4/\delta_P}{n}}\end{aligned}$$

w.p. at least $1-18|S|^2|A|H\delta_P$. Hence, putting $\epsilon = 3\sqrt{\frac{128|S|H^3(1+NB_\lambda)^2 \log 4/\delta_P}{n}}$ and $\delta = 18|S|^2|A|H\delta_P$ completes the proof. □

4.4 Online-CRL-Dual

Online Constrained-RL Dual, or Online-CRL-Dual described in Algorithm 6, is an online method proceeding in episodes with length H. At the beginning of each episode k, Online-CRL-Dual 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. Furthermore, the reward matrix r_c is created by equation (4.8) using updated Lagrange multipliers.

For any λ, the empirical model $\widehat{P}$ and reward matrix r_c induce a set of finite-horizon MDPs $\mathcal{M}_\lambda$ which any MDP $M' \in \mathcal{M}_\lambda$ has identical horizon and reward matrix. However, for any $(s,a) \in S \times A$ and $s' \in S$, $P'(s'|s,a)$ lies inside a confidence interval induced by $\widehat{P}$. M_λ is defined as:

$$\mathcal{M}_\lambda := \{M' : r'_c(\lambda, s, a) = r_c(\lambda, s, a), H' = H, \tag{4.21}$$

$$|P'(s'|s,a) - \widehat{P}(s'|s,a)| \leq \min\Big(\sqrt{\frac{2\widehat{P}(s'|s,a)(1-\widehat{P}(s'|s,a))}{n(s,a)}}\log\frac{4}{\delta_P} + \frac{2}{3n(s,a)}\log\frac{4}{\delta_P}, \sqrt{\frac{\log 4/\delta_P}{2n(s,a)}}\Big) \tag{4.22}$$

$$\forall s, a, s'\},$$

where δ_P is defined in Algorithm 6. For any $M' \in \mathcal{M}_\lambda$, 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 (4.1) and (4.3) respectively. Accordingly, $L'(\pi,\lambda)$ and $D'(\lambda)$ are computed via equations (4.6) and (4.9). The optimistic model among all models in $\mathcal{M}_\lambda$ is denoted using $\sim$ such as $\tilde{V}_0^{\tilde{\pi}(\lambda)}(s_0)$. Here,

$$\tilde{\pi}(\lambda) = \underset{\pi, M' \in \mathcal{M}_\lambda}{\arg\max}\, V_0'^{\pi}(s_0, \lambda).$$

Next, at the end of the episode k Lagrange multipliers are updated according to **Stochastic**

Subgradient Descent method as follows:

$$\lambda_i^{(k+1)} = \Pi_\lambda(\lambda_i^{(k)} + \alpha(\tilde{C}_{i,0}^{\tilde{\pi}_k}(s_0) - \bar{C}_i)) \tag{4.23}$$

where $\Pi_\lambda(\cdot)$ is projection to $[0, B_\lambda]$. B_λ is the upper bound on each λ_i^*.

Next, Online-CRL-Dual computes $\tilde{\lambda}^{(k)} = \frac{\sum_{j=0}^{k} \lambda^{(j)}}{k}$ and constructs $M_{\tilde{\lambda}^{(k)}}$. Finally, it output $\tilde{\pi}_k$ as solution to $M_{\tilde{\lambda}^{(k)}}$. Algorithm 6 briefs Online-CRL-Dual.

Algorithm 6 Online-CRL-Dual

```
Input: accuracy $\epsilon$ and failure tolerance $\delta$.
Set $m$ by means of (6.35) and (6.36)
Set $w_{\min} = \frac{\epsilon}{20HNB_\lambda|S|}, U_{\max} = |S|^2|A|m, \delta_P = \frac{\delta}{8|S|U_{\max}}$
Set $K$ and $\alpha$ as in (4.24)
$\tilde{\lambda}_i = \tilde{\lambda}_i^{(0)} = \lambda_i^{(0)} = 0 \ \forall i.$
Set $n(s,a) = n(s,a,s') = 0 \ \forall s,a,s'$
while there is $(s,a)$ with $n(s,a) < |S|mH$ and $k < K$ do
    $\widehat{P}(s'|s,a) = \frac{n(s',s,a)}{\max\{n(s,a),1,\}} \ \forall(s,a)$ with $n(s,a) > 0$ and $s' \in S$.
    Construct $M_{\tilde{\lambda}^{(k)}}$ using (4.21)
    Solve $M_{\tilde{\lambda}^{(k)}}$ by EVI and get $\tilde{\pi}_k(\tilde{\lambda}^{(k)})$
    for $t = 1, \ldots, H$ do
        $a_t \sim \tilde{\pi}_k(\tilde{\lambda}^{(k)}, s_t), s_{t+1} \sim P(\cdot|s_t, a_t)$
        if $n(s_t, a_t) < |S|mH$ : then
            $n(s_t, a_t) + +, n(s_t, a_t, s_{t+1}) + +$
    Construct $M_{\lambda^{(k)}}$ using (4.21)
    Solve $M_{\lambda^{(k)}}$ by EVI and get $\tilde{\pi}_k(\lambda^{(k)})$
    Evaluate $\tilde{C}_{i,0}^{\tilde{\pi}_k(\lambda^{(k)})}(s_0) \ \forall i.$
    Update $\lambda^{(k+1)}$ using (4.23)
    $\tilde{\lambda} \ = + \ \lambda^{(k+1)}$
    $k + +$
    $\tilde{\lambda}^{(k)} = \frac{\tilde{\lambda}}{k}$
```

This algorithm draws inspiration from the infinite-horizon algorithm UCRL$-\gamma$ [32] and its finite-horizon counterpart UCFH [7] with several differences. Unlike UCRL-γ and UCFH, Algorithm 6 updates the model at the beginning of each episode, which allows for faster model construction.

To provide PAC result for sample complexity of Algorithm 6, we make the following assumption as well.

4.4.1 PAC Analysis of Online-CRL-Dual

Here, we present the PAC result of Algorithm 6

Theorem 7. *Consider CMDP $M = \langle S, A, P, r, c, \bar{C}, H\rangle$ satisfying assumptions 5, 6 and 7. For any $0 < \epsilon, \delta < 1$ under Algorithm 6 we have*

$$L(\tilde{\pi}_k(\tilde{\lambda}^{(k)}), \tilde{\lambda}^{(k)}) \leq L(\pi^*, \lambda^*) + \epsilon$$

w.p. at least $1 - \delta$ after

$$\max\{K, O(\frac{|S|^2|A|N^2 B_\lambda^2 H^2}{\epsilon^2} \log \frac{1}{\delta})\}$$

number of episodes with K and α determined as

$$K = \Big(\frac{5\sqrt{N}H(\bar{C}_{\max} + H)}{\epsilon \bar{C}_{\min}}\Big)^2, \quad \alpha = \frac{\epsilon}{5N(\bar{C}_{\max} + H)^2}. \tag{4.24}$$

The proof pf Theorem 7 consists of two stages. First, we prove that $V_0^{\tilde{\pi}_k(\tilde{\lambda}^{(k)})}(s_0, \tilde{\lambda}^{(k)})$ and $V_0^{\pi^*(\tilde{\lambda}^{(k)})}(s_0, \tilde{\lambda}^{(k)})$ are close to each other w.h.p. for all but except some number of episodes. The proof of this section is carried out by means of **unconstrained MDP analysis** techniques, mainly from [7]. Next, we prove that $\tilde{D}_k(\tilde{\lambda}^{(k)})$ is also close to $D(\lambda^*)$ w.h.p. Here, we focus on employing **stochastic subgradient method**. Before explaining each stage, we present the following proposition which is useful for both stages.

4.4.2 First Stage: Unconstrained MDP Analysis

To prove this stage, we follow an approach motivated by [32] and its finite-horizon version [7]. 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.

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 7. Finally, we sketch the proof of Theorem 7. The detailed analysis is provided in supplementary materials.

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 $w_{\min} := \frac{\epsilon}{20H|S|}$ on a log-scale

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

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

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

which indicates how often (s, a) has been observed relative to its importance. Value of m is defined in Algorithm 6. 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 sub-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 15. *[7] $M \in \mathcal{M}_k$ for all episodes k with probability at least $1 - \frac{\delta}{8}$.*

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

Lemma 16. *[7] 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*

$$m \geq \frac{30H^2}{\epsilon} \log \frac{8E_{\max}}{\delta}, \tag{4.25}$$

where $E_{\max} = \log_2 \frac{H}{w_{\min}} \log_2 |S|$. Then, $\mathbb{P}(E \leq 6|S||A|mE_{\max}) \geq 1 - \frac{\delta}{8}$.

Finally, the next lemma provides a bound on the mismatch between objective function of the optimistic model and true model. It provides a PAC result for objective functions with any reward matrix $r_c(\lambda)$ for any $\lambda \in [0, B_\lambda]$.

Lemma 17. *Assume $M \in \mathcal{M}_k$. If $|X_{k,\kappa,\iota}| \leq \kappa$ for all (κ,ι) and $0 < \epsilon \leq 1$ and*

$$m \geq 12800 \frac{|S|N^2 B_\lambda^2 H^2}{\epsilon^2} (\log_2 \log_2 H)^2 \log_2^2\Big(\frac{40|S|^2H^2}{\epsilon}\Big) \log \frac{6}{\delta_P}, \tag{4.26}$$

then $|\tilde{V}_0^{\tilde{\pi}_k(\bar{\lambda}^{(k)})}(s_0, \bar{\lambda}^{(k)}) - V_0^{\tilde{\pi}_k(\bar{\lambda}^{(k)})}(s_0, \bar{\lambda}^{(k)})| \leq \frac{\epsilon}{5}$.

Proof. The proof is provided in Appendix C. □

Lemma 18. *For any* $0 < \epsilon, \delta \le 1$, *we have*

$$\mathbb{P}(|\tilde{V}_0^{\tilde{\pi}_k(\bar{\lambda}^{(k)})}(s_0, \bar{\lambda}^{(k)}) - V_0^{\tilde{\pi}_k(\bar{\lambda}^{(k)})}(s_0, \bar{\lambda}^{(k)})| \le \frac{\epsilon}{5}) \ge 1 - \frac{\delta}{4}, \quad \mathbb{P}(\tilde{D}_k(\bar{\lambda}^{(k)}) \le D(\bar{\lambda}^{(k)}) + \frac{\epsilon}{5}) \ge 1 - \frac{\delta}{4} \tag{4.27}$$

for all episodes except at most

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

and eventually for any $\lambda \in [0, B_\lambda]$

$$\mathbb{P}(\tilde{D}(\lambda) \le D(\lambda) + \frac{\epsilon}{5}) \ge 1 - \frac{\delta}{4}. \tag{4.28}$$

Proof. The proof is provided in Appendix C. □

4.4.3 Second Stage: Stochastic Subgradient Descent

First, we define good events. For the given accuracy $\epsilon \in (0,1)$, Algorithm 6 would collect samples $(s_t, a_t)_t$ where eventually each state-action pair (s,a) would be visited equal number of times i.e. $|S|mH$. Please note that value of m depends on ϵ. Each time the algorithm runs, it pursues different trajectory. Among all the trajectories with equal number of samples from each (s,a), we collect the ones in a set denoted by $F_{\epsilon/5}$ that satisfy the following

$$\|\tilde{V}_0^{\tilde{\pi}(\lambda)}(\lambda) - V_0^{\pi^*(\lambda)}(\lambda)\|_\infty \le \frac{\epsilon}{5} \text{ and } |\tilde{D}(\lambda) - D(\lambda)| \le \frac{\epsilon}{5} \;\; \forall\lambda \in [0, B_\lambda] \tag{4.29}$$

and call it set of “good events". Later in Lemma 19, we will show that

$$\mathbb{P}(F_{\epsilon/5}) \ge 1 - \frac{\delta}{4}.$$

Now, suppose we are given a partial trajectory denoted by $\tau_{\epsilon/5,T}$ which consists of observations up to T time-steps of a complete trajectory belonging to $F_{\epsilon/5}$. Thus for a fixed episode k, we get $\tilde{D}_k(\lambda) = \mathbb{E}[\tilde{D}(\lambda)|F_{\epsilon/5}]$.

Lemma 19. *For any given $\lambda \in [0, B_\lambda]$ under algorithm 6 finally we obtain*

$$\mathbb{P}(F_{\epsilon/5}) \geq 1 - \frac{\delta}{4}.$$

Proof. The proof is provided in Appendix C. □

Lemma 20. *Let $g : \mathcal{X} \to \mathbb{R}^d$ be a convex function with bounded domain, i.e. $\|x\| \leq B_1, \forall x \in \mathcal{X}$. Also assume that there is $\tilde{g}$ where $\mathbb{E}[\tilde{g}] \in \partial g$ and it is bounded, i.e. $\|\tilde{g}\| \leq B_2$. Consider the projected stochastic subgradient descent algorithm $x_{k+1} = \Pi_{\mathcal{X}}(x_k - \alpha\tilde{g}(x_k))$ where $\Pi_{\mathcal{X}}$ is the projection operator, for K steps. Then after K iterations with $\alpha = \frac{B_1}{B_2\sqrt{K}}$ we have*

$$\mathbb{E}[g(\frac{1}{K}\sum_{k=1}^{K} x_k)] - g(x^*) \leq \frac{B_1 B_2}{\sqrt{K}},$$

Proof. The proof is provided in Appendix C. □

Proposition 2. *With parameters K and α determined by* (4.24)*, after $k \geq \max\{K,$ $\tilde{O}\Big(\frac{|S|^2|A|N^2B_\lambda^2H^2}{\epsilon^2}\log\frac{1}{\delta}\Big)\}$ we get*

$$\mathbb{P}(\tilde{D}_k(\tilde{\lambda}^{(k)}) \leq \tilde{D}(\tilde{\lambda}^*) + \frac{3\epsilon}{5}) \geq 1 - \frac{\delta}{2},$$

where $\tilde{\lambda}^* := \arg\min_\lambda \tilde{D}(\lambda)$.

Proof. The proof is provided in Appendix C. □

Lemma 21. *For any $0 < \epsilon, \delta < 1$, under Algorithm 6 we get*

$$\mathbb{P}(\tilde{D}_k(\tilde{\lambda}^{(k)}) \leq D(\lambda^*) + \frac{4\epsilon}{5}) \geq 1 - \frac{3\delta}{4}.$$

Proof. The proof is provided in Appendix C. □

4.4.4 Proof of Theorem 7

Consider the following

$$\begin{aligned}
L(\tilde{\pi}_k(\tilde{\lambda}^{(k)}), \tilde{\lambda}^{(k)}) - D(\lambda^*) &= V_0^{\tilde{\pi}_k(\tilde{\lambda}^k)}(s_0, \tilde{\lambda}^{(k)}) + \sum_i \tilde{\lambda}_i^{(k)} \bar{C}_i - D(\lambda^*) \\
&= V_0^{\tilde{\pi}_k(\tilde{\lambda}^k)}(s_0, \tilde{\lambda}^{(k)}) - \tilde{V}_0^{\tilde{\pi}_k(\tilde{\lambda}^k)}(s_0, \tilde{\lambda}^{(k)}) + \tilde{V}_0^{\tilde{\pi}_k(\tilde{\lambda}^k)}(s_0, \tilde{\lambda}^{(k)}) + \sum_i \tilde{\lambda}_i^{(k)} \bar{C}_i - D(\lambda^*) \\
&= V_0^{\tilde{\pi}_k(\bar{\lambda}^k)}(s_0, \tilde{\lambda}^{(k)}) - \tilde{V}_0^{\tilde{\pi}_k(\tilde{\lambda}^k)}(s_0, \tilde{\lambda}) + \tilde{D}_k(\tilde{\lambda}^{(k)}) - D(\lambda^*)
\end{aligned}$$

Now, we consider the two following sections:

- **Section** 1**:** We prove $|V_0^{\tilde{\pi}_k(\tilde{\lambda}^{(k)})}(s_0, \tilde{\lambda}^{(k)}) - \tilde{V}_0^{\tilde{\pi}_k(\tilde{\lambda}^{(k)})}(s_0, \tilde{\lambda}^{(k)})|$ is small w.h.p. after certain number of episodes.
- **Section** 2**:** We prove $|\tilde{D}_k(\tilde{\lambda}^{(k)}) - D(\lambda^*)|$ is small w.h.p. after certain number of episodes.

Section 1**:** To prove this part, we apply Lemma 18 and get

$$\mathbb{P}(\tilde{V}_0^{\tilde{\pi}_k(\tilde{\lambda}^{(k)})}(s_0, \tilde{\lambda}^{(k)}) - V_0^{\tilde{\pi}_k(\tilde{\lambda}^{(k)})}(s_0, \tilde{\lambda}^{(k)}) \leq \frac{\epsilon}{5}) \geq 1 - \frac{\delta}{4}. \tag{4.30}$$

Section 2**:** Now, we prove the second part. By applying Lemma 21, we get

$$\mathbb{P}(\tilde{D}_k(\tilde{\lambda}^{(k)}) - D(\lambda^*) \leq \frac{4\epsilon}{5}) \geq 1 - \frac{3\delta}{4} \tag{4.31}$$

with parameters K and α specified by (4.24). Therefore, both inequalities (4.30) and (4.31) are satisfied after $\max\{K, 6E_{\max}|S||A|m\}$ which is

$$\begin{aligned}
&61440 \frac{|S|^2|A|N^2 B_\lambda^2 H^2 (\bar{C}_{\max} + H)^2}{\bar{C}_{\min}^2 \epsilon^2} (\log_2 \log_2 H)^2 \log_2^2\Big(\frac{4|S|H^2}{\epsilon}\Big) \log_2^2\Big(\frac{8H^2|S|^2}{\epsilon}\Big) \\
&\times \log\Big(\frac{2048|S|^4|A|H^2}{\epsilon^2\delta} (\log_2 \log_2 H)^2 \log_2^2\Big(\frac{8H^2|S|^2}{\epsilon}\Big)\Big).
\end{aligned}$$

Finally, combining inequalities (4.30) and (4.31) and applying union bound would yield the result.
□

4.5 Conclusion

In this chapter, we developed two algorithms based on Lagrangian approach in order to achieve computationally efficient learning algorithms. We showed, however, that this efficiency yields in an increase in sample complexity due to expansion of space of reward matrices.

Our next goal is to tailor the algorithms for specific application scenarios. Specifically, we desire to study routing and scheduling in multihop wireless networks. First, we study the problem of routing and scheduling for broadcasting with known parameters in the next chapter. Then, we present learning algorithms for data networks in the final chapter.

5. BROADCASTING REAL-TIME FLOWS IN INTEGRATED ACCESS AND BACKHAUL 5G NETWORKS*

5.1 Introduction

Mutli-hop broadcasting in wireless networks, which entails disseminating information to every device in the system via retransmissions at multiple nodes, is an important mechanism to coordinate devices in networked systems. Furthermore, many applications of broadcast communications are safety-critical, and timely deliveries of information is crucial to maintain the robustness and safety of the system. For example, multi-hop broadcasting is needed to disseminate timely safety information among connected vehicles in vehicular ad hoc networks (VANETs), to announce control decisions in networked control systems and Internet of Things (IoT), and to exchange locations and flight paths among unmanned aerial vehicles (UAVs) for Unmanned Aircraft System Traffic Management (UTM).

The cellular infrastructure that will enable these time-critical broadcast wireless applications will be 5G networks that are currently being designed to support ultra-low latency, ultra-high throughput communications. These networks will utilize the highly directional and high bandwidth mm-wave band, which suffers from high attenuation and sensitively to fading. This requires the relatively dense deployment of small base stations at spacings of about 250 m. However, providing fiber backhaul to all of these base stations is prohibitively expensive. An important development in this context is *Integrated Access and Backhaul* (IAB) [36, 37], under which there are a few base stations with fiber backhaul that act as gateways to many others that are connected via a mm-wave wireless mesh backhaul. This mm-wave backhaul creates a directional wireless network between the nodes, but routing across these is highly dynamic and subject to the vagaries of the wireless channel. The same mm-wave spectrum also is used to provide access to end-users, i.e., both access and backhaul are integrated over mm-wave.

Motivated by the above features of emerging networks, this chapter studies the problem of

*Reprinted with permission from [3]

designing algorithms for broadcasting real-time flows with strict per-packet end-to-end deadlines in directional wireless mesh networks. Here, real-time flow imposes a strict deadline for each of its packets, and packets that cannot be delivered before their respective deadlines are dropped from the system. From the IAB perspective, our goal is to ensure that each broadcast packet is delivered to an appropriate IAB base station before its deadline, at which point it is immediately transmitted to its respective end user. Each IAB node in the network then obtains some utility based on the time-average number of on-time packets that it receive from each flow. The goal of this chapter is to maximize the total timely-utility of the whole network.

There are several important challenges that need to be addressed for broadcasting real-time flows in such multi-hop mmWave networks. First, since it is difficult to coordinate a large network in real-time, centralized algorithms that require the instant knowledge of the state of each node and packet are usually infeasible to implement. Hence, we need distributed algorithms, where each node makes decisions using its local information. Second, as mentioned above, transmissions in the mmWave band can be unreliable. Finally, broadcasting algorithms need to explicitly address the deadline requirement of each flow.

Main Results and Organization

In this chapter, we propose a new protocol for broadcasting in multi-hop mmWave networks, namely, the *delegated-set routing* (DSR) protocol. DSR has two important features: First, it is a distributed protocol where all the required coordination among nodes can be conveyed in the headers of packets once the topology of the network is known. Hence, there is virtually no overhead of coordination after topology creation process. Second, DSR allows each node to dynamically change its transmission strategies based on the deadlines of its packets and random events, such as transmission failures, it experiences.

Relaxing the link utilization constraint (number of transmissions allowed per time slot) to an average one, and using dual decomposition techniques, we also propose a distributed algorithm that aims to maximize the total system-wide utility under DSR. This algorithm only requires minimal and infrequent information exchange among nodes. We analytically prove that our algorithm achieves the optimal total utility under an average link capacity constraint. The key novelty lies in a natural decomposition into packet-by-packet and link-by-link updates that need minimal coordination. These lead to a steepest-ascent-type control associated with each packet, and a sub-gradient type of update at links. This algorithm also gives rise to a simple index policy when link utilization constraints of all links need to be satisfied at every instant.

We evaluate our algorithms through simulations on representative network graphs. We compare our algorithms against recent studies on throughput optimal algorithms, including one that is designed specifically for broadcast, and one that is universal in terms of being able to support unicast, multicast and broadcast. We show that despite some of these algorithms being centralized and complex, our algorithm, which is designed specifically for simplicity and delay optimality, achieves better performance.

The chapter is organized as follows. Section 5.2 reviews existing studies on broadcasting and multi-hop networks. Section 5.3 describes our system model for multi-hop networks with real-time broadcast flows. Section 5.4 describes the additional structure imposed by the DSR protocol, as well as an epoch-wise approach to policy selection. Section 5.5 applies dual-decomposition, which turns out to be the basis of our distributed algorithm. Section 5.6 proposes distributed algorithms that optimize DSR, as well as the index policy that can ensure hard capacity constraints are met. Section 5.7 presents our simulation results.

5.2 Related Work

Broadcasting/multicasting is a fundamental functionality of networks, and has been studied in a substantial body of literature. One of the earliest policies for broadcasting/multicasting in ad hoc networks is via flooding [38, 39]. However, such policies can lead to severe packet collision frequency, and excessive redundant retransmissions, as shown by Ni et al. [40]. Gandhi et al. [41] and Huang et al. [42] have shown that the problem of minimizing delay in wireless ad hoc networks is NP-hard, and have proposed approximation algorithms aiming to reduce delay. These studies rely on centralized algorithms.

There has been much interest in throughput optimal broadcasting/multicasting. For instance, Sarkar and Tassiulas [43] proposed a scheduling and routing policy that relies on pre-computed spanning trees, which might be difficult to maintain and compute in scalable sized networks. Ho and Viswanathan [44] and Yuan et al. [45] propose network coding based policies in the context, which, however, leads to additional computation complexity. Zhang et al. [46] and Sinha et al. [47] consider multi-hop broadcasting problems in Directed Acyclic Graphs (DAG), which are not applicable to networks with arbitrary topology. Sinha et al. [48] also propose a centralized throughput optimal broadcasting policy for networks with arbitrary topology, which might be difficult to deploy in a large scale system. Furthermore, the throughput maximization focus of all the above does not directly allow for meeting stringent deadline guarantees.

Given the rising application of wireless networks to safety-critical and realtime applications, there has been much recent interest in deadline constrained multi-hop communication. Xiong et al. [49] proposed a delay-aware throughput optimal policy for multi-hop networks. Their policy, however, can not provide stringent delay guarantees. Mao et al. [50] propose a hard deadline guaranteed policy, under the assumption that all routes in the network are fixed. Li and Eryilmaz [51] consider serving flows with stringent deadlines in a multi-hop system, and their proposed framework can be extended to incorporate routing decisions. However, their policies are heuristic, and optimality cannot be shown. Singh and Kumar [52] relax the deadline constrained optimization problem in the manner of the Whittle's relaxation for multi-armed bandits, and proposed decentralized optimal solutions. However, both it and the above body of work on deadline constrained communication only considers unicast traffic, and it is not clear how it applies to broadcasting/multicasting networks.

5.3 System Model

We consider a multi-hop network that consists of N wireless nodes operating in the mmWave band motivated by the IAB system. Here, the nodes correspond to fixed IAB base stations, and the network topology is known to all nodes. The available spectrum is divided into multiple half-duplex channels, and nodes can use these channels to send and receive packets from multiple nodes simultaneously. Furthermore, these channels are *directional* in that transmissions on different links do not interfere with each other, consistent with empirical observations in IAB test deployments [37]. These links can have different constraints on the supportable number of transmissions in each time slot, as well as their reliabilities.

Time is slotted and numbered as $t = 1, 2, \ldots$. We assume that link l can transmit T_l packets in each time slot, and that each transmission will be successfully received by the receiver with probability P_l. At the end of each time slot, the receiver sends an aggregated ACK indicating which packets it has successfully received in the time slot to the transmitter. Where we need to indicate the transmitter and the receiver of a link, we use $l = n \to m$ to indicate that link l has transmitter n and receiver m.

We consider F real-time broadcast flows, using s_f to indicate the source node of flow f. At the beginning of each time slot t, $a_f(t)$ packets of flow f arrive at node s_f. We assume that $[a_f(1), a_f(2), \ldots]$ is a sequence of i.i.d. random variables with mean A_f. Moreover, each flow f specifies a per-packet end-to-end deadline of D_f time slots. Packets from flow f are only useful for D_f time slots from their respective arrival times at their source nodes, and are dropped from the network when they expire. Due to communication constraints, it is likely that some nodes cannot receive all packets from each flow. We therefore measure the performance of node n on flow f by its *timely-throughput*, defined as the long-term average number of packets from flow f that are successfully delivered to node n within the deadline.

Let Ω be a set of stationary packet scheduling policies. Hence, given the state of the system consisting of the locations and expiry times of all existing packets, a policy $\omega \in \Omega$ is a rule that decides which packet to transmit on what link, subject to communication constraints. For each stationary policy $\omega \in \Omega$, let $x^{\omega}_{n,f}(t)$ be the number of packets from f that are delivered to n at time t under ω, i.e., these are the packets that survived the deadline constraint. Also, let $\epsilon^{\omega}_{l,f}(t)$ be the number of packets from flow f transmitted over link l at time t under ω. Since ω is a stationary policy, and all packets that expire are immediately dropped, we can define

$$\mu^{\omega}_{n,f} := \liminf_{T\to\infty} \frac{\sum_{t=1}^{T} x^{\omega}_{n,f}(t)}{T}$$

as the timely-throughput of node n on flow f under ω, and

$$\bar{\epsilon}^{\omega}_{l,f} := \limsup_{T \to \infty} \frac{\sum_{t=1}^{T} \epsilon^{\omega}_{l,f}(t)}{T}$$

as the average number of transmissions for flow f over link l under ω.

Now, finding the optimal total utility with respect to timely-throughputs over all the N nodes under DSR is equivalent to finding the stationary policy that maximizes the total timely-utility under link utilization constraints, which can be written as

Relaxed Timely-Utility Maximization (R-TUM)

$$\text{Max} \sum_{n=1}^{N} \sum_{f=1}^{F} U_{n,f}(\mu^{\omega}_{n,f}) \tag{5.1}$$

$$\text{s.t. } \omega \in \Omega, \tag{5.2}$$

$$\sum_{f=1}^{F} \bar{\epsilon}^{\omega}_{l,f} \leq T_l, \forall l. \tag{5.3}$$

Notice that whereas the R-TUM problem above requires each delivered packet to satisfy its deadline constraint, it only requires that the *long-term average* number of transmissions over link l, $\sum_{f=1}^{F} \bar{\epsilon}^{\omega}_{l,f}$ be no larger than T_l. This link utilization constraint relaxation is in the same manner as [52]. In a practical system, such a relaxation might be akin to imposing an average transmit power constraint rather than a hard one. We will first design policies that pertain to this relaxed link-utilization constraint. Using the insights gained, we will also develop a policy that enforces a *hard link-utilization constraint,* i.e.,

$$\sum_{f=1}^{F} \epsilon^{\omega}_{l,f}(t) \leq T_l, \forall l, t. \tag{5.4}$$

Solving the R-TUM problem could be posed as a Markov Decision Process (MDP), where the state of the system at any given point of time consists of the locations and expiry times of all existing packets. However, such a solution is infeasible to implement in practice. First, it is straightforward to show that the number of different system states is at least doubly exponential in N, and hence standard algorithms for finding the optimal MDP-based solution will result in prohibitive complexity. Second, even after one finds the optimal MDP-based solution, it may be impossible to implement it in a distributed fashion, since the complete state needs to be known at each node. In what follows, we impose additional structure on the policy space to render it tractable.

5.4 A Structured Approach to Real-Time Broadcasting

We now introduce two elements of structure to the policy space to enable its solution as a distributed convex optimization problem.

5.4.1 Delegated-Set Routing (DSR)

Ensuring a per-packet deadline guarantee requires that we retain flexibility in routing to dynamically choose the next hop node for a packet based on current state. Thus, source routing on a per-packet basis is not satisfactory. However, for distributed implementation, we also need to ensure that there is no ambiguity as to which neighboring node is responsible for transmitting a packet to a given node. We resolve these seemingly opposite requirements via a protocol that we term *delegated-set routing (DSR).*

For each node n that possesses a packet i at time $t,$ we define the delegated-set of node n as the subset of nodes that n is responsible for forwarding packets, possibly through multi-hop transmissions. First, to ensure routing flexibility, whenever a node n decides to forward a packet to another node m, node n delegates a subset of its own delegated-set to $m,$ and specifies this subset in the packet header. If the transmission is successful, this subset is removed from the delegated-set of $n,$ since it is now the responsibility of m to forward the packet to this subset. Second, in order to avoid duplicate transmissions (ambiguity on which node should transmit a given packet), the DSR protocol requires that the delegated-sets of different nodes for the same packet are chosen to be disjoint.

To illustrate how DSR works, consider the network as shown in Fig. 5.1. When a packet

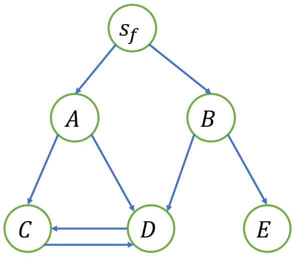

Figure 5.1: An example illustrating DSR. Reprinted with permission from [3]

arrives at the source node s_f, the delegated set of s_f is every node in the network, since it is the responsibility of s_f to broadcast the packet to the entire network. Suppose in the first time slot, s_f transmits the packet to A, and delegates the subset $\{A, C, D\}$ to A. If the transmission is successful, the delegated-set of s_f becomes $\{s_f, B, E\}$, while the delegated-set of A is $\{A, C, D\}$. In the next time slot, when s_f transmits the packet to B, it needs to delegate the subset $\{B, E\}$ to B. In particular, s_f cannot include node D in the delegated-set for B, since D is already in the delegated-set of A.

We note that the ability to dynamically adjust routing decisions is an important feature that distinguishes DSR from many existing studies on multi-hop broadcasting, such as [48] and [43]. These studies adopt source-routing, where the source node determines the routing decision of each packet, and intermediate nodes cannot change the decision. As s_f cannot foresee whether the transmissions from A to C will be successful, it cannot take an optimal routing decision.

5.4.2 Epoch-wise Stationary Policies

Our second aspect of adding structure to the policy space is to expand it from Ω, the set of all stationary policies, to the set of all *epoch-wise stationary policies*. In an epoch-wise stationary policy, time is divided into epochs of equal length. The epoch-wise stationary policy adopts a stationary policy $\omega^+[i]$ in each epoch i. The duration of an epoch is chosen to be large enough so that the average performance of $\omega^+[i]$ in epoch i is not influenced by the system state at the beginning of the epoch. Specifically, an epoch-wise stationary policy is defined as follows:

Definition 4. *An epoch-wise stationary policy is a sequence of stationary policies* $\omega^+ = (\omega[i])_{i=1}^{\infty}, \omega[i] \in$

$\Omega\}$, *where* $\omega[i]$ *is used in epoch* i*. The length of an epoch is chosen so that, under* ω^+*,*

$$\mu_{n,f}^{\omega^+} := \liminf_{T\to\infty} \frac{\sum_{t=1}^{T} x_{n,f}^{\omega^+}(t)}{T} = \liminf_{I\to\infty} \frac{\sum_{i}^{I} \mu_{n,f}^{\omega[i]}}{I},$$

and

$$\bar{\epsilon}_{l,f}^{\omega^+} := \limsup_{T\to\infty} \frac{\sum_{t=1}^{T} \epsilon_{l,f}^{\omega^+}(t)}{T} = \limsup_{I\to\infty} \frac{\sum_{i}^{I} \bar{\epsilon}_{l,f}^{\omega[i]}}{I}.$$

We can now define Ω^+ as the set of all epoch-wise stationary policies. For each epoch-wise stationary policy ω^+, let $\gamma^{\omega^+} := [[\mu_{n,f}^{\omega^+}, 1 \leq n \leq N, 1 \leq f \leq F], [\bar{\epsilon}_{l,f}^{\omega^+}, 1 \leq l \leq L, 1 \leq f \leq F]]$ be the vector of timely-throughputs and average link uses under ω^+. Also, let $\Gamma := \{\gamma^{\omega^+} | \omega^+ \in \Omega^+\}$ be the set of attainable vectors of timely-throughputs and average link uses under all epoch-wise stationary policies. An important advantage of considering the policy space Ω^+ is that Γ is a convex set.

Lemma 22. Γ *is convex.*

Proof. Consider two epoch-wise stationary policies ω_1^+, ω_2^+, and a number $0 < a < 1$, we will show that there exists an epoch-wise stationary policy ω_a^+ such that $\gamma^{\omega_a^+} = a\gamma^{\omega_1^+} + (1-a)\gamma^{\omega_2^+}$.

We construct ω_a^+ as follows: In epoch i, if $\lfloor ai \rfloor > \lfloor a(i-1) \rfloor$, then ω_a^+ uses $\omega_1[\lfloor ai \rfloor]$ in epoch i. Otherwise, ω_a^+ uses $\omega_2[i - \lfloor ai \rfloor]$ in epoch i.

It is straightforward to check that, in the first I epochs, ω_a^+ consists of the first $\lfloor aI \rfloor$ stationary policies from ω_1^+ and the first $\lceil (1-a)I \rceil$ stationary policies from ω_2^+. Therefore, $\gamma^{\omega_a^+} = a\gamma^{\omega_1^+} + (1-a)\gamma^{\omega_2^+}$. □

Since Γ is a convex set, it is straightforward to verify that the optimization problem (5.1)–(5.3) subject the policy space Γ is a convex optimization problem.

5.5 Solution Overview

Although the problem **R-TUM**, (5.1) – (5.3), is convex, solving it directly remains challenging because there is no simple characterization of Γ. In this section, we present a general framework of solving **R-TUM** through dual decomposition. The exact distributed algorithm will be presented in the next section.

5.5.1 Dual Problem Formulation

Let λ_l be the Lagrange multiplier with respect to the constraint $\sum_{f=1}^{F} \bar{\epsilon}_{l,f}^{\omega+} \leq T_l$ in (5.3), and λ be the vector of all λ_l, $l = 1, 2, \dots, L$. The Lagrangian of **R-TUM** is then

$$\mathcal{L}(\gamma^{\omega^+}, \lambda) = \sum_{n=1}^{N} \sum_{f=1}^{F} U_{n,f}(\mu_{n,f}^{\omega+}) - \sum_{l=1}^{L} \lambda_l \left(\sum_{f=1}^{F} \bar{\epsilon}_{l,f}^{\omega+} - T_l \right), \tag{5.5}$$

and the dual objective function is

$$\mathcal{D}(\lambda) = \max_{\gamma \in \Gamma} \mathcal{L}(\gamma, \lambda). \tag{5.6}$$

The dual problem of **R-TUM** is to find a non-negative vector λ that minimizes $\mathcal{D}(\lambda)$.

We first show that strong duality holds for **R-TUM**.

Theorem 8. *Let $\mathcal{P}^*$ be the optimal solution to **R-TUM**, and $\mathcal{D}^* := \min_{\lambda:\lambda_l \geq 0, \forall l} \mathcal{D}(\lambda)$, then $\mathcal{P}^* = \mathcal{D}^*$.*

Proof. Since **Relaxed Utility** is convex, we only need to check the Slater's condition:

1. $\Gamma \neq \emptyset$.

2. Constraint (5.3) is a linear inequality.

3. Consider the policy that never transmits any packets. Under this policy, the number of transmissions over link l is 0, which is strictly less than T_l, for all l.

Therefore, Slater's condition holds, and the proof is complete. □

Hence, solving **R-TUM** is equivalent to solving the dual problem, which consists of two steps: First, given a vector λ, we need to find the dual objective function $\mathcal{D}(\lambda)$. Second, we need to find the vector λ that minimizes $\mathcal{D}(\lambda)$.

5.5.2 Packet-By-Packet Decomposition for the Dual Objective

We first present an iterative algorithm that finds $\mathcal{D}(\lambda) = \max_{\gamma \in \Gamma} \mathcal{L}(\gamma, \lambda)$ for a given λ using the steepest ascent algorithm. For each stationary policy ω, let γ^{ω} be defined to be the vector of timely-throughputs and link usages under ω. Then the steepest ascent algorithm constructs a sequence of epoch-wise stationary policies that ultimately converges to the optimal epoch-wise stationary policy. The algorithm proceeds as follows:

1. Set $k \leftarrow 1$
2. Let ω_k^+ be the round-robin epoch-wise stationary policy that follows the sequence $\{\omega_1, \omega_2, \ldots, \omega_k, \omega_1, \omega_2, \ldots$
3. Let ω_{k+1} be the stationary policy that maximizes the directional derivative, $\nabla \mathcal{L}(\gamma^{\omega_k^+}, \lambda) \cdot \gamma^{\omega_{k+1}}$.
4. Set $k \leftarrow k+1$ and repeat step 2.

Based on our construction of ω_k^+, we have $\gamma^{\omega_k^+} = \frac{\sum_{j=1}^{k} \gamma^{\omega_j}}{k}$. Therefore $\gamma^{\omega_{k+1}^+} - \gamma^{\omega_k^+} = \frac{\gamma^{\omega_{k+1}} - \gamma^{\omega_k^+}}{k+1}$. Effectively, for each k, our steepest ascent algorithm finds ω_{k+1}^+ that maximizes the directional derivative $\nabla \mathcal{L}(\gamma^{\omega_k^+}, \lambda) \cdot (\gamma^{\omega_{k+1}^+} - \gamma^{\omega_k^+})$ among all epoch-wise stationary policies with step size $\frac{1}{k+1}$. Following the analysis presented in Boyd et al. [53] Section 9.4.3, it is straightforward to show the following:

Theorem 9. *Under our steepest ascent algorithm, $\mathcal{L}(\gamma^{\omega_k^+}, \lambda)$ converges to $\mathcal{D}(\lambda)$, as $k \to \infty$.*

Notice that the critical step in our steepest ascent policy is to find ω_{k+1} that maximizes $\nabla L(\gamma^{\omega_k^+}, \lambda) \cdot \gamma^{\omega_{k+1}}$. We have

$$
\begin{aligned}
&\nabla \mathcal{L}(\gamma^{\omega_k^+}, \lambda) \cdot \gamma^{\omega_{k+1}} \\
=&\sum_{n,f} \frac{\partial}{\partial \mu_{n,f}} \mathcal{L}(\gamma^{\omega_k^+}, \lambda) \mu_{n,f}^{\omega_{k+1}} + \sum_{l,f} \frac{\partial}{\partial \bar{\epsilon}_{l,f}} \mathcal{L}(\gamma^{\omega_k^+}, \lambda) \bar{\epsilon}_{l,f}^{\omega_{k+1}} \\
=&\sum_{f=1}^{F} \left\{ \sum_{n=1}^{N} U'_{n,f}(\mu_{n,f}^{\omega_k^+}) \mu_{n,f}^{\omega_{k+1}} - \sum_{l=1}^{L} \lambda_l \bar{\epsilon}_{l,f}^{\omega_{k+1}} \right\}.
\end{aligned}
$$

This naturally gives us a flow-by-flow decomposition in the sense that $\nabla\mathcal{L}(\gamma^{\omega_k^+}, \lambda) \cdot \gamma^{\omega_{k+1}}$ can be maximized by maximizing

$$\sum_{n=1}^{N} U'_{n,f}(\mu_{n,f}^{\omega_k^+})\mu_{n,f}^{\omega_{k+1}} - \sum_{l=1}^{L} \lambda_l \bar{\epsilon}_{l,f}^{\omega_{k+1}} \tag{5.7}$$

for each flow f individually. Moreover, note that, after normalizing with the average packet arrival rate of flow f, $\mu_{n,f}^{\omega_{k+1}}$ is the average delivery per-packet from flow f to node n, and $\bar{\epsilon}_{l,f}^{\omega_{k+1}}$ is the average number of transmissions per packet over link l for flow f.

For each packet i from flow f, let $y_{n,f,i}$ be a random variable representing the event that packet i is successfully delivered to node n within its deadline of d_f. Also, let $z_{l,f,i}$ be the random variable indicating the number times that link l transmits i. Then $\mathbb{E}[y_{n,f,i}]$ is the success probability that packet i is delivered to node n, while $\mathbb{E}[z_{l,f,i}]$ is the expected number of times that link l transmits i. Therefore, from (5.7), maximizing $\nabla\mathcal{L}(\gamma^{\omega_k^+}, \lambda) \cdot \gamma^{\omega_{k+1}}$ can be achieved by maximizing

$$\sum_{n=1}^{N} U'_{n,f}(\mu_{n,f}^{\omega_k^+})\, \mathbb{E}[y_{n,f,i}] - \sum_{l=1}^{L} \lambda_l\, \mathbb{E}[z_{l,f,i}] \tag{5.8}$$

for each packet i.

We note that such packet-by-packet decomposition allows distributed algorithms for finding the optimal solution since, instead of considering the system state as a whole, each packet only needs to maximize (5.8) on its own, without considering the states of other packets.

5.5.3 Link-by-Link Update for the Dual Problem

After finding $\mathcal{D}(\lambda)$, we now proceed to find the solution to the dual problem, $\min_{\lambda:\lambda_l \geq 0, \forall l} \mathcal{D}(\lambda)$. Our solution is based on the subgradient method. We first find the subgradient of $\mathcal{D}(\lambda)$.

Theorem 10. *Let* $\gamma(\lambda) = [[\mu_{n,f}(\lambda)], [\bar{\epsilon}_{l,f}(\lambda)]] := \arg\max_{\gamma\in\Gamma} \mathcal{L}(\gamma, \lambda)$*, then the L-dimensional vector* $[T_l - \sum_{f=1}^{F} \bar{\epsilon}_{l,f}(\lambda)]$ *is a subgradient for* $\mathcal{D}(\lambda)$*.*

Proof. For any arbitrary λ':

$$\begin{aligned}
&\mathcal{D}(\lambda') = \max_{\gamma} \mathcal{L}(\gamma, \lambda') \geq \mathcal{L}(\gamma(\lambda), \lambda') \\
=&\sum_{n=1}^{N}\sum_{f=1}^{F} U_{n,f}(\mu_{n,f}(\lambda)) - \sum_{l=1}^{L} \lambda'_l (\sum_{f=1}^{F} \bar{\epsilon}_{l,f}(\lambda) - T_l) \\
=&\sum_{n=1}^{N}\sum_{f=1}^{F} U_{n,f}(\mu_{n,f}(\lambda)) - \sum_{l=1}^{L} \lambda_l (\sum_{f=1}^{F} \bar{\epsilon}_{l,f}(\lambda) - T_l) \\
&+ \sum_{l=1}^{L} (\lambda_l - \lambda'_l)(\sum_{f=1}^{F} \bar{\epsilon}_{l,f}(\lambda) - T_l) \\
=&\mathcal{D}(\lambda) + (\lambda' - \lambda) \cdot [T_l - \sum_{f=1}^{F} \bar{\epsilon}_{l,f}(\lambda)].
\end{aligned}$$

Thus, $[T_l - \sum_{f=1}^{F} \bar{\epsilon}_{l,f}(\lambda)]$ is a subgradient of $\mathcal{D}(\lambda)$. □

The subgradient method finds the optimal λ that minimizes $\mathcal{D}(\lambda)$ iteratively. Starting with an arbitrary vector $\lambda(1)$, the subgradient method finds $\lambda(k+1) = [\lambda_l(k+1)]$ by setting

$$\lambda_l(k+1) = \left[\lambda_l(k) - \beta_k \left(T_l - \sum_{f=1}^{F} \bar{\epsilon}_{l,f}(\lambda(k)) \right) \right]^+, \tag{5.9}$$

where $x^+ := \max\{0, x\}$.

Theorem 11. *If the sequence β_k is chosen so that $\beta_k \geq 0, \forall k$, $\sum_{k=1}^{\infty} \beta_k = \infty$, and $\lim_{k\to\infty} \beta_k = 0$, then $\mathcal{D}(\lambda(k)) \to \min_{\lambda:\lambda_l \geq 0, \forall l} \mathcal{D}(\lambda)$, as $k \to \infty$.*

Proof. This is the direct result of Theorem 8.9.2 in [54]. □

Recall that $\sum_{f=1}^{F} \bar{\epsilon}_{l,f}(\lambda(k)))$ is the average number of transmissions that link l makes. Therefore, for link l to update λ_l by (5.9), link l only needs to know its own link constraint and the number of transmissions it makes. Hence, this subgradient method allows for a distributed update of λ_l.

5.6 Optimization of DSR

Under DSR, the transmission strategy for a node having a packet i consists of two parts: determining which node to transmit the packet i to, and determining what delegated-set to assign to the receiver. In this section, we discuss the optimal transmission strategy that maximizes (5.8) under the design of DSR.

Fix a packet i from flow f. For each subset of nodes π, let $\mathcal{L}_\pi$ be the set of links whose transmitter and receiver are both in π. Also, for each node n, subset of nodes π, and integer $\tau \in [0, d_f]$, define

$$\begin{aligned} W_f(n, \pi, \tau) = & \\ & \max\left(\sum_{k\in\pi} U'_{k,f}(\mu_{k,f}^{\omega_k^+})\,\mathbb{E}[y_{k,f,i}] - \sum_{l\in\mathcal{L}_\pi} \lambda_l\,\mathbb{E}[z_{l,f,i}]\right) \end{aligned} \tag{5.10}$$

if node n receives the packet i and delegated-set π, and the packet i has τ time slots before meeting its deadline.

By the definition of $W_f(n, \pi, \tau)$, finding the optimal transmission strategy that maximizes (5.8) is equivalent to finding the value of $W_f(s_f, \{1, 2, \ldots, N\}, d_f)$, as well as the transmission strategy that achieves it.

We use dynamic programming to find $W_f(n, \pi, \tau)$. Suppose node n receives the packet i and delegated-set π, and packet i has τ time slots before meeting its deadline. Also suppose that node n decides to transmit the packet to m and designates the delegated-set π^m to m. If the transmission is successful, then, in the next time slot, node n has a delegated-set of $\pi - \pi^m$, node m has a delegated-set of π^m, and packet i has $\tau - 1$ time slots before its deadline. By the definition of $W_f(\cdot)$, we have, given that the transmission is successful,

$$\begin{aligned} & \max\left(\sum_{k\in\pi} U'_{k,f}(\mu_{k,f}^{\omega_k^+})\,\mathbb{E}[y_{k,f,i}] - \sum_{l\in\mathcal{L}_\pi} \lambda_l\,\mathbb{E}[z_{l,f,i}]\right) \\ & = W_f(n, \pi - \pi^m, \tau - 1) + W_f(m, \pi^m, \tau - 1) - \lambda_{n\to m}. \end{aligned} \tag{5.11}$$

On the other hand, if the transmission fails, then, in the next time slot, node n still has the delegated-set π and packet i has $\tau - 1$ time slots before its deadline. Given that the transmission fails, we have

$$\max\left(\sum_{k\in\pi} U'_{k,f}(\mu_{k,f}^{\omega_k^+})\,\mathbb{E}[y_{k,f,i}] - \sum_{l\in\mathcal{L}_\pi} \lambda_l\,\mathbb{E}[z_{l,f,i}]\right)$$

$$= W_f(n,\pi,\tau-1) - \lambda_{n\to m}. \tag{5.12}$$

Since each transmission from n to m succeeds with probability $P_{n\to m}$, we have, given that n transmits packet i and assigns delegated-set π^m to m,

$$\max\left(\sum_{k\in\pi} U'_{k,f}(\mu_{k,f}^{\omega_k^+})\,\mathbb{E}[y_{k,f,i}] - \sum_{l\in\mathcal{L}_\pi} \lambda_l\,\mathbb{E}[z_{l,f,i}]\right)$$

$$= P_{n\to m} \times (5.11) + (1 - P_{n\to m}) \times (5.12). \tag{5.13}$$

Based on the above analysis, we can write down the following iterative equation:

$$\begin{aligned}
&W_f(n,\pi,\tau) = \max\{W_f(n,\pi,\tau-1),\\
&\max_{m,\pi^m: m\in\pi^m, \pi^m\subset\pi}[P_{n\to m}(W_f(n,\pi-\pi^m,\tau-1)\\
&+ W_f(m,\pi^m,\tau-1)) + (1-P_{n\to m})W_f(n,\pi,\tau-1)\\
&- \lambda_{n\to m}]\},
\end{aligned} \tag{5.14}$$

with boundary condition

$$W_f(n,\pi,0) = r_{i,n} = U'_{n,f}(\mu_{n,f}), \tag{5.15}$$

where the term $W_f(n,\pi,\tau-1)$ in (5.14) represents the case when n does not transmit the packet at all. Eq. (5.14) and (5.15) allows a dynamic programming algorithm to find $W_f(n,\pi,\tau)$ for all f, n, π, and τ. As we will show in Section 5.7, our algorithm can be easily carried out in medium-sized networks.

5.6.1 Index-DSR for Per-Time-Slot Link Constraint

The Dynamic Program in (5.14) can be directly combined with the dual decomposition in Section 5.5 to achieve the optimal solution of **R-TUM** problem under DSR. In this section, we further propose an index policy that satisfies the per-time-slot link utilization constraint $\sum_{i,v} \epsilon_{i,v,l}(t) \leq T_l$, for all t, of the original **TUM** problem. The index-DSR policy would be to transmit the maximum number of packets among all possible packets to be transmitted so that the per-time-slot link constraint is not violated.

We make several changes to the dynamic program and the dual decomposition technique. First, we change the iterative equation (5.14) to

$$\begin{aligned}
&W_f(n,\pi,\tau) = \\
&\max_{m,\pi^m: m\in\pi^m, \pi^m\subset\pi} [P_{n\to m}(W_f(n,\pi-\pi^m,\tau-1) \\
&+ W_f(m,\pi^m,\tau-1)) + (1-P_{n\to m})W_f(n,\pi,\tau-1) \\
&- \lambda_{n\to m}]\}, \qquad (5.16)
\end{aligned}$$

as long as there is a link from n to another node in π, and

$$W_f(n,\pi,\tau) = W_f(n,\pi,\tau-1), \qquad (5.17)$$

otherwise. In other words, we force each node n to find a link to transmit each packet. We also define $m^*(n,\pi,\tau)$ and $\pi^{m*}(n,\pi,\tau)$ as the optimal m and π^m that achieves $W_f(n,\pi,\tau)$. We note that, since we now force each node n to find a link to transmit each packet, it is possible that $W_f(n,\pi,\tau)$ is negative for some (n,π,τ).

Second, in each time slot t and for each link $n \to m$, we find all packets possessed by n with delegated-set π, τ slots until their respective deadlines, and $m^*(n,\pi,\tau) = m$. We sort these packets in descending order of $W_f(n,\pi,\tau)$, and let $\epsilon'_{n\to m}(t)$ be the number of these packets with

$W_f(n, \pi, \tau) > 0$. In other words, $\epsilon'_{n \to m}(t)$ is the number of packets whose optimal strategy yields a positive return by transmitting over the link $n \to m$. After sorting these packets, link $n \to m$ simply transmit the first $T_{n \to m}$ packets. Finally, the price of each link is updated by (5.9).

5.7 Simulation Results

In this section, we present simulation results that compare the performance of our policy against a policy proposed in [55] called Universal Max-Weight (UMW), and a policy proposed by Sinha, Paschos, and Modiano in [48] that we call SPM. We first provide a brief description of these two policies, and then present our simulation settings and results.

5.7.1 Overview of UMW and SPM

The UMW policy solves the problem of throughput-optimal packet dissemination in a network with arbitrary topology with different types of traffic, e.g., unicast, multicast and broadcast. In both the centralized and distributed versions of UMW, the route of each packet is decided at the origin. This route is a weighted tree that is constructed using the edge weights at time of decision at the origin. Hence, if the route of the packet turns to be inappropriate during packet dissemination, it cannot be modified. Although this policy also has a heuristic version that can be implemented in a distributed fashion, we consider our comparison against the centralized version, which has better performance than the distributed one.

The SPM policy is designed specifically for throughput optimal broadcast. SPM is a virtual-queue based algorithm, where virtual-queues are defined for subsets of nodes. These virtual queues keep track of a kind of backpressure, while accounting for the fact that packets are duplicated in the broadcast regime. A feature of this work is that each slot is sub-divided into L minislots, where L is the number of links in the network, and a random link is activated in each mini-slot. Here, a packet may be retransmitted multiple times over the mini-slots comprising a slot (i.e., it could potentially reach all nodes in just one slot). To ensure consistency with the slot model, we modify this algorithm to only allow packet state updates at each slot, rather than at each mini-slot.

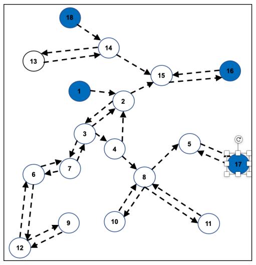

Figure 5.2: Scenario 1: 11−node network. Reprinted with permission from [3]

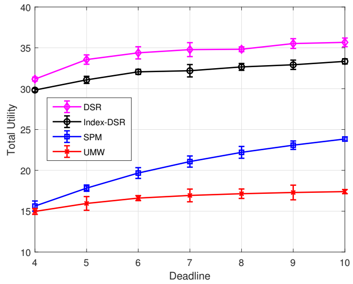

Figure 5.3: Scenario 1: Linear Utility Function. Reprinted with permission from [3]

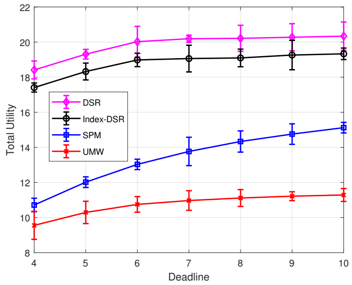

Figure 5.4: Scenario 1: Logarithmic Utility Function. Reprinted with permission from [3]

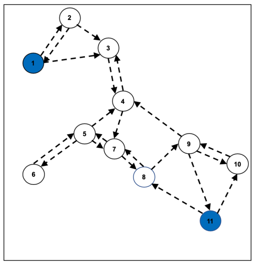

Figure 5.5: Scenario 2: 18−node network. Reprinted with permission from [3]

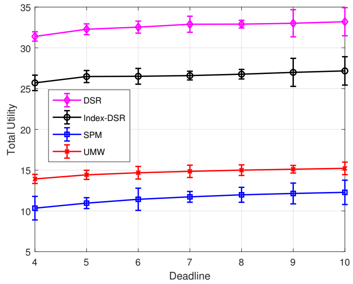

Figure 5.6: Scenario 2: Linear Utility Function. Reprinted with permission from [3]

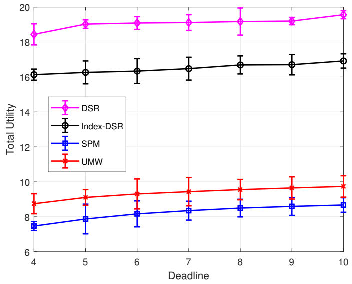

Figure 5.7: Scenario 2: Logarithmic Utility Function. Reprinted with permission from [3]

5.7.2 Simulation Settings and Results

In this study, we consider two different simulation scenarios motivated by designs for IAB network deployments [36, 37]. Here, we have two kinds of nodes, namely, (i) gateway nodes with fiber drops (shown in red), and (ii) wireless-only nodes with mm-wave backhaul (shown as blue nodes). We assume that gateways communicate reliably between each other with zero latency, since they are connected to the same backend switch (consistent with IAB architecture). The two scenarios represent different levels of gateway availability. The first scenario is a an $11-$node network with 2 fiber drops as in figure (5.2), while the second scenario is an $18-$ node IAB network with 9 fiber drops as shown in figure (5.5). Hence, Scenario 1 is illustrative performance in a network with multiple wireless hops, whereas Scenario 2 illustrates performance in a more densely connected network.

In both scenarios, there are two broadcast flows. One of the flows originates at a fiber-connected gateway node, and the other one from a wireless-only node. For each link l, P_l is randomly chosen from $[0.5, 1.0]$, and T_l is randomly chosen from $[1, 5]$. Each flow generates packets according to a Poisson random process, where source node of flow 1 has a mean arrival rate of 1.5 packets per time slot, and source node of second flow has a mean arrival rate of 2 packets per time slot. Since UWM and SPM only aim to maximize throughput, we first consider a linear utility function $U_{n,f}(\mu_{n,f}) = \mu_{n,f}$ to make a fair comparison. In this case, the total utility of the system is the same as the total timely-throughputs. In a second case, we also consider a logarithmic utility function $U_{n,f}(\mu_{n,f}) = \log(\mu_{n,f} + 1)$, which models the idea that the utility of the end user might be a non-negative, concave and increasing function of timely throughput. We assume that the two flows have the same deadline of D time slots, and vary D from 4 to 10. We test four the optimal DSR protocol (for the relaxed problem), the Index-DSR protocol, the UWM policy, and the SPM policy.

The simulation results for the linear utility function and the logarithmic utility function for scenario in figure (5.2) are shown in figures (5.3) and (5.4), respectively. The performance of DSR is an upper bound, since it is the optimal solution under a relaxed constraint. The Index-DSR protocol outperforms UWM, possibly because of more dynamic routing of each packet under

Index-DSR. This also shows that UMW might be providing bursty service to nodes, since deadlines are often violated and packets are dropped, leading to poor throughput. The Index-DSR policy outperforms SPM in all cases despite the assumption taht SPM can compute the reachable subgraph for each packet instantly.

The results for second IAB scenario, depicted in (5.5) shows similar results in terms performance of DSR-based algorithms for much the same reasons specified above. However, results of figures (5.6) and (5.7) shows that UMW has better performance than SPM, unlike the results obtained in (5.3) and (5.7). This result appears to be due to the density of the network. The UMW policy manages to deliver more unexpired packets to the destination since it has to traverse fewer hops. SPM is also handicapped by the fact that we force it to obey a slot-by-slot state update model like all the other protocols (although we do allow it to utilize its minislot-based transmission model). Ultimately, these results demonstrate the efficiency and flexibility of the DSR protocol.

5.8 Conclusion

In this chapter, We studied the problem of broadcasting real-time flows with hard per-packet deadlines in a multi-hop wireless network. We considered the IAB node deployment and proposed an optimal algorithm–DSR and near-optimal algorithm–index-DSR. These algorithms assume that the link reliabilities are known. In the next chapter, we disregard this assumption and present learning algorithms.

6. REINFORCEMENT LEARNING FOR SCHEDULING AND ROUTING REAL-TIME FLOWS IN INTEGRATED ACCESS AND BACKHAUL 5G NETWORKS

6.1 Introduction

The previous chapter studied the problem of routing and scheduling when the system parameters (link reliabilities) are known. In this chapter, we do not consider this assumption moving into learning direction. The problem of maximizing timely throughput can be posed in the manner of reinforcement learning (RL) over a Constrained Markov Decision Process (CMDP). Here, the state of the system is the tuple of location and remaining lifetime of each packet, and a unit reward is obtained each time that an unexpired packet is delivered successfully to end-user. The available actions are the choices of links that can be used for forwarding the packet at each node, and the randomness of the MDP kernel stems from the randomness of the links. The constraints of this problem are on the number of transmissions permissible per link at each time, while the fact that the probabilities of success or failure at each link is unknown implies the need for a learning approach.

Multiple challenges must be addressed to successfully solve the CMDP problem of deadline constrained flows. First, reinforcement learning must be employed to estimate the link reliabilities using as few packets as possible. Second, we must ensure that per-packet deadline guarantees are met. Finally, it is untenable to solve a global MDP that requires state information about every packet and node in the system, and a simple distributed implementation of the policy is desired.

From RL point of view, 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, and yet ensure that the weighted timely throughput is maximized. Hence, determining how the joint metrics of timely throughput maximization and capacity constraint violation evolve over time as the model becomes more and more accurate is crucial to understand the efficacy of a proposed RL algorithms for CMDPs.

Main Results

Our main results are based on two general frameworks presented in [4]. First, we formulate an LP according to CMDP problem [4] and analyze the sample complexity of solving that 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 with the LP framework 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.

(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.

(iii) We develop PAC-type sample complexity bounds for both algorithms, accounting for both objective maximization and constraint satisfaction.

Next, we build on a framework of a general solution methodology for CMDPs using a dual decomposition approach of Altman [4]. Here, the CMDP problem is solved via a two step procedure of (i) maximizing the objective (solving an MDP) under fixed Lagrange multipliers corresponding to the constraints, and (ii) a gradient descent step over the Lagrange multipliers. This algorithm follows a procedure under which each link is sampled a given number of times to determine its statistics to a desired level of accuracy, and the resulting (noisy) model of the system is used as an input to the CMDP framework of Altman [4].

Our work is perhaps the first to consider a learning approach towards solving CMDPs in the context of optimal wireless scheduling design. The main contribution of our work is to design algorithms that explicitly account for the overhead of learning link reliabilities while computing the optimal packet and link scheduling policy. We follow the general theme of *model-based reinforcement learning,* under which the intent is to efficiently determine the transition kernel of the MDP under study, and explicitly solve it to obtain the optimal policy. This approach is particularly suited to our problem, as it has a well defined structure under which the unknown sources of randomness in the system are parametrized by the success probabilities of the links. Our performance analysis goal is to characterize the so-called *sample complexity* of our algorithms, *i.e.,* we wish to determine the number of packet transmissions needed to ensure that the value of the packet transmission policy differs from that of the optimal policy at most by a parameter ϵ with a high probability.

Our numerical evaluation is over topologies similar to those proposed for IAB trails [37]. We compare our RL-based algorithms with the optimal solution value assuming that the model (link success probabilities) are known to show how the accuracy improves with increasing sample complexity.

6.2 Related Work

There has been much work in the past several years on provably throughput optimal scheduling policies, starting with seminal work of Tassiulas et al. [56], and follow up works [57, 58] leading to the so-called backpressure type scheduling policies. Recent work in this space has focused on throughput optimal broadcast under networks with different topologies [59, 55]. With the rise of real-time streaming applications that require hard delay guarantees, a different approach is needed as backpressure cannot provide delay optimality. Work in this space focuses on scheduling such real-time flows, wherein an MDP formulation is avoided due to the emphasis on a single (typically downlink) wireless hop [60, 61].

The design of scheduling algorithms that can support hard deadline constrains in the multi-hop context has been the topic of recent study. For instance, Xiong et al. [49] introduce delay-awareness into the protocol, without, however, enabling hard deadline guarantees. Other work, such as that by Mao et al. [50] provide such guarantees under fixed routing, while that by Li et al. [51] is only able to do so in a heuristic manner without optimality guarantees. The fundamental issue here is the need to solve a global MDP for taking scheduling/routing decisions, and the work of Singh et al. [62] is the first to use an average link utilization constraint to enable a simple and distributed solution. The approach has been further generalized to the broadcast setting by HasanzadeZonuzy et al. [63].

The use of AI methods in communication networks has recently been the subject of much interest, with most work focusing on bandit-style approaches to learning the sources of randomness in the system. For example, Krishnasamy et al. [64] use posterior sampling with some additional learning effort in order to small queuing regret in a system with a single queue and many wireless channel. Combes et al. [65] and Gupta et al. [66] both use a marginal posterior sampling approach in the context of power allocation in the context of a system in which channel statistics are unknown. Talebi et al. [67] also consider a bandit approach to routing over links whose statistics are unknown.

Unlike the above body of work on learning in wireless networks, our problem of delay constrained unicast flows does not admit a bandit-type of solution due to the hard delay constraint that implies that the state of each packet in the system consists of both a location and a time to live. Hence, while the source of randomness in our problem lies in unreliable links (like earlier work), our formulation is very different and takes the form of a constrained MDP that explicitly accounts for state, rather than the bandit formulation considered earlier.

In addition, 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 [4, 8, 9, 10, 11, 12] for solving a general CMDP problem. 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 [13, 14]. 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 [15, 16, 17, 18] under Lagrangian methods to show zero eventual duality gap. [19] 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 [20].

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 [21, 22, 23] 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, [24] and [25] present results in the context of unknown reward functions, with either a known stochastic or deterministic transition kernel. Other work [26] focuses on asymptotic convergence, and so does not provide an estimate on the learning rate. Finally, [5] 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 [27], 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.

6.3 Problem Formulation

In this section, we formally describe our model and the constrained MDP (CMDP) formulation for maximizing the weighted timely throughput of the system. The setup of both approaches are similar to Singh et al. [62], and employs the relaxed transmission constraint and Lagrangian decomposition technique proposed in that work to obtain simple per-packet MDPs that are conducive to attain decentralized optimal policy. To achieve this goal, we explain briefly two different approaches to solve a CMDP problem presented in [4] in detail. Finally, we state the problem formally.

6.3.1 System Model

We consider a communication network described by a directed graph $G = (\mathcal{S}, \mathcal{L})$, where $\mathcal{S}$ is the set of nodes and $\mathcal{L}$ is the set of links. The cardinality of $\mathcal{S}, \mathcal{L}$ are denoted by $S, |\mathcal{L}|$ respectively.

Let $\mathcal{L}_j$ be set of the outgoing links from node $j \in \mathcal{S}$. A directed link $l = (j, k)$ indicates that node j can transmit data packet to node k. We use self loops to indicate the decision not to transmit at a node, *i.e.,* $(j, j) \in \mathcal{L}_j$ for all $j \in \mathcal{S}$. We model unreliability of network links by assuming that a transmission over link l is successful with a probability p_l. We also assume that the time is slotted, and one time slot is the time needed to transmit one packet over any link in the network.

We consider a set of finite number of flows F with size $|F|$ indexed by $f \in \{1, \ldots, |F|\}$. s_f and d_f indicate the source node and destination node of flow $f \in F$, respectively. Let $A_f(t)$ denotes the set of packets arriving at node s_f at time t that are in flow f. The average arrival rate of flow f is then defined as $\rho_f = \lim_{T\to\infty} \sum_{t=1}^{T} |A_f(t)|/T$. We denote $\rho_{\text{tot}} = \sum_f \rho_f$. Each packet of flow f has a maximum end-to-end delay τ_f associated with it. A packet of flow f that has arrived at s_f at time t needs to be delivered to d_f before time $t + \tau_f$, or else it will be discarded. We assume that $max_f \tau_f = \tau_{\max} < \infty$.

The *timely throughput* for flow f under a scheduling policy π, R_f^π, is the expected value of the number of packets delivered prior to deadline expiry per unit time,

$$R_f^\pi = \liminf_{T\to\infty} \frac{1}{T} \mathbb{E} \sum_{t=1}^{T} x_f^\pi(t), \tag{6.1}$$

where $x_f^\pi(t)$ is the number of packets of flow f successfully delivered to d_f under policy π at time t.

The *average link utilization* for link l under policy π, denoted by C_l^π is defined as

$$C_l^\pi := \limsup_{T\to\infty} \frac{1}{T} \mathbb{E} \sum_{f\in F} \sum_{t=1}^{T} c_{l,f}^\pi(t), \tag{6.2}$$

where $c_{l,f}^\pi(t)$ is the number of packet transmissions for flow f on link l under policy π at time t. This is the relaxation proposed in [62]. In practice, such a relaxation might correspond to an average transmit power constraint. It is pointed out in [62] that the gap between this approach and the hard constraint becomes small in the heavy traffic regime. We will also consider such a hard constraint in the numerical simulations.

The optimal scheduling problem is to find a policy π^* that solves the following optimization problem

$$[\textbf{OSP}] \ \ \max_{\pi} \ \sum_{f \in F} \beta_f R_f^{\pi}, \ \ \text{s.t } C_l^{\pi} \leq C_l, \forall l \in \mathcal{L} \tag{6.3}$$

6.3.2 Constrained MDP Formulation

We now formulate OSP using the framework of constrained Markov Decision Processes (CMDP). We first define the per-packet finite-horizon MDP corresponding to each flow as if there is no link capacity by specifying the states, actions, rewards and transition kernel. Next, we show how the per-packet MDPs weld together by imposing link capacity constraints. Then, we formulate the network CMDP using the per-packet MDP definitions, cost matrices and constraints.

State. Let $s_{i,f}(t)$ denote the state of the packet i from flow f at time t, defined as the node at which that packet from flow f is located at time t. If the packet has been delivered to its destination, or if it has been discarded from the network by time t, then $s_{i,f}(t)$ is defined as the terminal state s_{term}. The state of the network at time t, $s(t)$, is then defined as $s(t) = (s_{i,f}(t), i \in \cup_{\tau=0}^{\tau_f} A_f(t - \tau), f \in F)$.

Action. The scheduling action $a_{i,f}(t)$ for packet i in flow f at time t is defined the link on which that packet is transmitted at time t. Hence, $a_{i,f}(t) \in \mathcal{L}_{s_{i,f}(t)}$. The scheduling action for the network at time t, $a(t)$, is then defined as $a(t) = (a_{i,f}(t), i \in \cup_{\tau=0}^{\tau_f} A_f(t - \tau), f \in \mathcal{F})$. A scheduling policy π maps the state of the system $s(t)$ to the scheduling action $a(t)$, *i.e.,* $a(t) = \pi(s(t))$.

Transition Kernel. We denote the transition kernel of the MDP as $P(k|j, l)$, which is the probability that the $s_{i,f}(t + 1) = k$ given that $s_{i,f}(t) = j$ and $a_{i,f}(t) = l$. Clearly,

$$P(k|j, l) = \begin{cases} p_l & \text{if } l = (j, k) \\ 1 - p_l & \text{if } j = k \\ 0 & \text{O.W.} \end{cases} \tag{6.4}$$

We assume that $p_l = 1$ for $l = (j, j)$ for all $j \in \mathcal{S}$. Note that the transition kernel is the same for all packets in all the flows.

Furthermore, transition kernel under policy π is

$$P_\pi(k|j) = \sum_l P(k|j,l)\pi(l|j).$$

Reward. Let $r_f(j)$ denote the reward for a packet in flow f for being in state j. We define

$$r_f(j) = \begin{cases} \beta_f & \text{if } j = d_f \\ 0 & \text{O.W.} \end{cases} \tag{6.5}$$

Per-Packet MDP. For each flow f, we can denote the per-packet finite-horizon MDP by a tuple $M_f = \langle \mathcal{S}, \mathcal{L}, P, r_f, \tau_f \rangle$. Here, the horizon is τ_f. Like any other finite-horizon MDP, we define value function of state j at time $t \in [0, \tau_f)$ under any policy π as below:

$$V^\pi_{f,t}(j) = \mathbb{E}[\sum_{h=t}^{\tau_f} r_f(s(h))|a(h) \sim \pi(s(h)), s(t) = j]. \tag{6.6}$$

Now, we need to specify the cost matrices and constraint vector to formulate the CMDP. These definitions are independent from the per-packet MDPs.

Cost Matrices and Constraint Vector. A capacity constraint on each link l implies a cost matrix and a constraint to be satisfied. The cost matrix is defined as

$$c_l(j,l) = \begin{cases} 1 & \text{if } l \in \mathcal{L}_j \\ 0 & \text{O.W.} \end{cases} \tag{6.7}$$

and the constraint is denoted by C_l.$[C_l]_l$ is used to denote the constraint vector.

Then, for each flow f the cost function of state j under policy π at time-step $t \in [0, \tau_f)$ is determined as below

$$C^\pi_{l,f,t}(j) = \mathbb{E}[\sum_{h=t}^{\tau_f} c_l(s(h), a(h))|a(h) \sim \pi(s(h)), s(t) = j]. \tag{6.8}$$

Now, we rewrite term $\beta_f R_f^\pi$ using definition (6.1)

$$\begin{aligned}
\beta_f R_f^\pi &= \beta_f \liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{\infty} x_f^\pi(t) = \liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{\infty} \beta_f x_f^\pi(t) \\
&= \liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{\infty} |A_f(t)| V_{f,0}^\pi(s_f) \\
&= V_{f,0}^\pi(s_f) \liminf_{T\to\infty} \frac{\sum_{t=1}^{\infty} |A_f(t)|}{T} = \rho_f V_{f,0}^\pi(s_f).
\end{aligned}$$

Analogously, we use definition (6.2) and get

$$C_l^\pi = \sum_f \rho_f C_{l,f,0}^\pi(s_f).$$

Now, we use tuple $\langle \mathcal{S}, \mathcal{L}, P, \{\rho_f\}_f, \{s_f\}_f \{r_f\}_f, \{c_l\}_l, [C_l]_l, \{\tau_f\}_f \rangle$ to denote the Network-CMDP and formulate the OSP equivalently as,

$$\begin{aligned}
&[\textbf{Network-CMDP}] \\
&\max_\pi \sum_f \rho_f V_{f,0}^\pi(s_f) \\
&\text{s.t} \sum_f \rho_f C_{l,f,0}^\pi(s_f) \leq C_l \;\; \forall l \in \mathcal{L}.
\end{aligned} \tag{6.9}$$

The network-CMDP problem of (6.9) is quite different from generic CMDP formulations presented in [4]. Unlike generic CMDPs, network-CMDP consists of multiple decision processes. One solution could be approaching network-CMDP as one generic CMDP and apply existing methods. This approach would lead to a solution but in expense of high computation power. Besides, these decision processes do not have identical horizon length in general which does not align with existing methods for CMDPs.

Further, network-CMDPs integrate multiple costs caused by different decision processes and impose them as a constraint. This integration prohibits us from decomposing these decision processes. However, we show how we tackle these issues by using existing methods for CMDPs in a different way.

Network-CMDP problem formulated by (6.9) could be solved in two ways. First solution approach is converting the problem to **Linear Programming (LP)** using *occupancy measures*. The other way is solving via **Lagrange multipliers**. Both of the solution techniques are extensively discussed in [4]. In this chapter, we utilize both approaches to design and analyze Reinforcement Learning algorithms.

6.3.3 LP Representation of CMDP

LP is one technique to solve CMDP problem (6.9) [4]. To convert CMDP problem to a linear programming problem, we introduce occupation measures. The finite-horizon state-action occupation measure at time-step τ under policy π is defined as

$$\mu(j, l, \pi, \tau, f) := \mathbb{P}(j_{f,\tau} = j, l_{f,\tau} = l), \tag{6.10}$$

where the probability is calculated w.r.t. underlying transition kernel under policy $\pi; P_\pi$. It is shown that both objective function and cost functions could be restated as functions of occupation measures. Then, the problem would become to find the optimal occupation measures. This procedure may be accomplished by creating a Linear Program that is equivalent to network-CMDP problem [4]. Here, we present the equivalent LP formulation of network-CMDP problem (6.9) by

[5]. Let μ be any generic occupation measure. Then, the equivalent LP would be

$$\begin{aligned}
&\max_{\mu} \sum_{j,f,\tau} \mu(j,l,\tau,f)\rho_f r_f(j) \\
&\text{s.t.} \\
&\sum_{l,\tau} \sum_{f} \rho_f \mu(j,l,\tau,f) c_l(j,l) \leq C_l, \\
&\sum_{l} \mu(j,l,\tau,f) = \sum_{j',l'} P(j|j',l')\mu(j',l',\tau-1,f), \\
&\sum_{l} \mu(s_f,l,0,f) = 1, \;\; \sum_{l} \mu(s_f,l,0,f) = 0 \;\; \forall f, \\
&\mu(j,j,\tau,f) \geq 0.
\end{aligned} \tag{6.11}$$

We can prove that LP (B.3) is equivalent to network-CMDP problem (6.9) by means of traditional CMDP methods in [4]. However, [68] shows it directly. Finally, the optimal policy π^* would be

$$\pi^*(j,l,\tau,f) = \frac{\mu(j,l,\tau,f)}{\sum_{l'} \mu(j,l',\tau,f)}.$$

6.3.4 Packet-by-Packet Decomposition

We next describe the decomposition approach that reduces the complexity of the problem by turning it into a per-packet MDP, rather than having to consider a global problem that accounts for the states of all packets in the system at each transmission decision.

The Lagrange Dual is a usual approach towards the solution of a CMDP [4]. The Lagrangian can be written as,

$$
\begin{aligned}
L(\pi, \lambda) = \sum_{l \in \mathcal{L}} \lambda_l C_l + \lim_{T \to \infty} \frac{1}{T} \, \mathbb{E} \sum_{t=1}^{T} \sum_{f \in F} \sum_{i \in A_f(t)} \sum_{\tau=0}^{\tau_f} \\
(r_f(s^{\pi}_{i,f}(t+\tau)) - \sum_{l} \lambda_l \mathbb{I}\{a^{\pi}_{i,f}(t+\tau) = l\}), \qquad (6.12)
\end{aligned}
$$

considering equivalent formulation of **OSP**, presented by (6.9). Noting that the rewards and transition probabilities are the same for each packet i in a given flow f, we define

$$
\begin{aligned}
V^{\pi}_{f,0}(s_f, \lambda) = \mathbb{E}[\sum_{\tau=0}^{\tau_f} (r_f(s^{\pi}_{i,f}(t+\tau)) - \sum_{l} \lambda_l \mathbb{I}\{a^{\pi}_{i,f}(t+\tau) = l\}) \\
|i, f, s^{\pi}_{i,f}(t) = s_f], \qquad (6.13)
\end{aligned}
$$

where $\mathbb{E}$ is the expectation w.r.t. to the underlying transition kernel under the policy π. Then the Lagrangian (6.12) can be written as

$$
\begin{aligned}
L(\pi, \lambda) &= \sum_{l \in \mathcal{L}} \lambda_l C_l + \sum_{f \in F} \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{i \in A_f(t)} V^{\pi}_{f,0}(s_f, \lambda) \\
&= \sum_{l \in \mathcal{L}} \lambda_l C_l + \sum_{f \in F} \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} |A_f(t)| V^{\pi}_{f,0}(s_f, \lambda) \\
&= \sum_{l \in \mathcal{L}} \lambda_l C_l + \sum_{f \in F} \rho_f V^{\pi}_{f,0}(s_f, \lambda). \qquad (6.14)
\end{aligned}
$$

The dual function $D(\lambda)$ and ‘dual policy’ $\pi(\lambda)$, and the optimal dual variable are defined as

$$
\begin{aligned}
&D(\lambda) = \max_{\pi} \; L(\pi, \lambda), \; \pi(\lambda) = \arg\max_{\pi} \; L(\pi, \lambda), \qquad (6.15) \\
&\lambda^* = \arg\min_{\lambda \geq 0} \; D(\lambda).
\end{aligned}
$$

Since there is no duality gap [4], the optimal policy π^* for the **[CMDP]** is the same as $\pi(\lambda^*)$.

Note that given a λ, $V^{\pi}_{f,0}(s_f, \lambda)$ for a given flow f does not depend on other flows. Hence, rather than finding an optimal joint policy $\pi(\lambda)$ for all flows, we can instead find an optimal policy $\pi_f(\lambda)$

for each flow separately. More precisely,

$$
\begin{aligned}
D(\lambda) = \max_{\pi} L(\pi, \lambda) &= \sum_{l \in \mathcal{L}} \lambda_l C_l + \max_{\pi} \sum_{f \in F} \rho_f V^{\pi}_{f,0}(s_f, \lambda) \\
&= \sum_{l \in \mathcal{L}} \lambda_l C_l + \sum_{f \in F} \rho_f \max_{\pi_f} V^{\pi_f}_{f,0}(s_f, \lambda) \\
&= \sum_{l \in \mathcal{L}} \lambda_l C_l + \sum_{f \in F} \rho_f V^{*}_{f,0}(s_f, \lambda),
\end{aligned}
$$

where,

$$
V^{*}_{f,0}(\lambda) = \max_{\pi_f} V^{\pi_f}_{f,0}(\lambda), \text{ and, } \pi_f(\lambda) = \arg\max_{\pi_f} V^{\pi_f}_{f,0}(\lambda) \tag{6.16}
$$

Now, $\pi_f(\lambda)$ and $V^{*}_{f,0}(\lambda)$ can be computed by standard finite horizon dynamic programming if we know the transition kernel P (equivalently, the link probabilities p_l).

However, as discussed in the Introduction, p_ls are unknown a priori. We thus propose a reinforcement learning approach for learning p_ls and at the same time solving for the optimal policy.

6.3.5 Constrained RL Formulation

The Constrained RL problem formulation is identical to the CMDP optimization problem of (6.9), but without being aware of values of transition kernel P.[*] Our goal is to provide model-based algorithms and determine the sample complexity results in a PAC sense. As we have two solution approaches for solving a CMDP, we present two definitions of sample complexity that extends the notion of sample complexity for unconstrained regime [28] to Constrained-RL. The way these definitions include the objective maximization and constraint violations differentiates them.

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

6.3.5.1 Sample Complexity of Algorithms based on LP

Definition 5. *Let $\mathcal{A}$ be an algorithm and $\pi(\mathcal{A})$ be the output of this algorithm. Then, sample complexity for $\mathcal{A}$ is the number of packets that $\mathcal{A}$ requires to achieve*

$$\mathbb{P}\Big(\sum_f \rho_f V_{f,0}^{\pi(\mathcal{A})}(s_f) \geq \sum_f \rho_f V_{f,0}^{\pi^*}(s_f) - \epsilon \\ \text{and } \sum_f \rho_f C_{l,f,0}^{\pi(\mathcal{A})} \leq C_l + \epsilon \ \forall l \Big) \geq 1 - \delta$$

for a given ϵ and δ.

Definition 5 includes satisfaction of both objective maximization and constraint violations individually. This definition mostly suits the algorithms outputting a policy which have analytical bounds on sub-optimality and constraint violations. Such algorihtms are usually based on LP approach.

6.3.5.2 Sample Complexity of Algorithms based on per-packet Decomposition

Definition 6. *Let $\mathcal{A}$ be an algorithm based on per-packet decomposition and $\pi(\mathcal{A})$ and $\lambda(\mathcal{A})$ be the output of this algorithm. Then, sample complexity for $\mathcal{A}$ is the number of packets that $\mathcal{A}$ requires to achieve*

$$\mathbb{P}\Big(L(\pi(\mathcal{A}), \lambda(\mathcal{A})) \leq L(\pi^*, \lambda^*) + \epsilon \Big) \geq 1 - \delta$$

for a given ϵ and δ.

Definition 6 regards the notion of sample complexity for Constrained-RL in a different manner. This definition integrates the objective maximization and constraint violation whereas Definition 5 which considers these individually. This integration is made by including Lagrange multipliers. Therefore, Definition 6 is appropriate for algorithms based on dual decomposition. Please notice that although we are not able to characterize the objective sub-optimality and constraint violations individually by Definition 6, but we obtain the luxury of less computational complexity.

6.4 Reinforcement Learning Solutions

This section consists of two categories of (RL) algorithms. Algorithms based on (i) LP approach , and (ii) dual decomposition. In each category, we propose two identical subcategories of model-based RL algorithms for solving the CMDP corresponding to timely throughput maximization. In the following we elaborate on each category.

6.4.1 Algorithms based on LP

In this section, we present two algorithms based on LP. These two algorithms employ LP as an alternate and more precise tool compared to Dual Decomposition. First, we propose and analyze a generative model-based algorithm. Then, we conclude this section by an online algorithm.

6.4.1.1 GMBL-LP

Here, we introduce a generative model based network learning algorithm called Generative Model Based Learning-LP, or GMBL-LP. According to GMBL-LP, we sample each link n number of times uniformly across all links, count the number of times each transition occurs $n(k, j, l)$ for each next node k, and construct an empirical model of transition kernel denoted by $\widehat{P}(k|j, l) = \frac{n(k,j,l)}{n}$ $\forall(k, j, l)$. Then GMBL-LP substitutes the empirical model with true model and solves the following optimization problem by means LP

$$\max_{\pi} \sum_{f} \rho_f \widehat{V}^{\pi}_{f,0}(s_f) \;\; \text{s.t} \;\; \sum_{f} \rho_f \widehat{C}^{\pi}_{l,f,0}(s_f) \leq C_l \;\; \forall l \in \mathcal{L} \tag{6.17}$$

where the value functions $\widehat{V}^{\pi}_{f,0}(s_f)$ and cost functions $\widehat{C}^{\pi}_{f,l,0}(s_f)$ are calculated w.r.t. to transition kernel $\widehat{P}$ using equations (6.6) and (6.8) respectively.

Algorithm 7 describes GMBL-LP.

Now, we present the sample complexity result of GMBL-LP.

Theorem 12. *GMBL-LP algorithm with*

$$n(\epsilon, \delta) \geq \frac{72 S \rho_{tot}^2 \tau_{\max}^3 \beta_{\max}^2 \log 4/\delta_P}{\epsilon^2}$$

Algorithm 7 GMBL-LP

```
Input: accuracy $\epsilon$ and failure tolerance $\delta$.
Set $\delta_P = \frac{\delta}{3(|\mathcal{L}|+2)S^2|F||\mathcal{L}|\tau_{\max}}$
Set $n(k,j,l) = 0 \; \forall (j,l,k)$.
**for** each $j \in \mathcal{S}$ **do**
    Sample $l \in \mathcal{L}_j, n = \frac{72 S \rho_{\text{tot}} \tau_{\max}^3 \beta_{\max}^2 \log 4/\delta_P}{\epsilon^2}$ and update $n(k,j,l)$.
    $\widehat{P}(k|j,l) = \frac{n(k,j,l)}{n} \; \forall k$.
Output $\widehat{\pi} = \text{LP}(\widehat{M})$.
```

for $\epsilon < \frac{\beta_{\max}\rho_{tot}}{18 \log 4/\delta_P}\sqrt{\frac{2\tau_{\max}}{S}}$ *achieves a* $\widehat{\pi}$ *such that*

$$\begin{aligned}\mathbb{P}\Big(\sum_f \rho_f V_{f,0}^{\widehat{\pi}_f} \geq \sum_f \rho_f V_{f,0}^{\pi_f^*} - \epsilon \textit{ and}& \\ \sum_f \rho_f C_{l,f,0}^{\widehat{\pi}} \leq C_l + \epsilon\Big) &\geq 1 - \delta\end{aligned}$$

where δ_P *is defined in Algorithm 7.*

The proof of Theorem 12 resembles both traditional analysis frameworks of unconstrained RL [6] and constrained RL [33]. First, unlike [33], we are not required to apply the notion of optimism. Because, we are equipped with "do-not-transmit" action at every state (location). This makes the network-CMDP problem feasible under any transition kernel (6.9). Thus, the problem of (6.17) is feasible, which allows us to avoid applying optimism. Another benefit is that we are able design an algorithm for data network purposes with less computational complexity compared to its counterpart algorithm for general CMDPs [33].

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 [4]. 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.

Now, we present some of the propositions that are essential to prove Theorem 12. Then we sketch the proof of this theorem.

Now, we present the lemmas required for proving Theorem 12 and its proof. Using these propositions, 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 23. *Suppose there is a network-CMDPs $M = \langle \mathcal{S}, \mathcal{L}, P, \{r_f\}_f, \{c_l\}_l, [C_l]_l, \{s_f\}_f, \{\tau_f\}_f \rangle$.Then, for any flow f, under any policy π*

$$V^{\pi}_{f,0} - \widehat{V}^{\pi}_{f,0} = \sum_{h=0}^{\tau_f - 2} \widehat{P}^{h-1}_{\pi} (P_{\pi} - \widehat{P}_{\pi}) V^{\pi}_{f,h+1} \text{ and}$$
$$V^{\pi}_{f,0} - \widehat{V}^{\pi}_{f,0} = \sum_{h=0}^{\tau_f - 2} P^{h-1}_{\pi} (P_{\pi} - \widehat{P}_{\pi}) \widehat{V}^{\pi}_{f,h+1},$$

and for any l,

$$C^{\pi}_{f,l,0} - \widehat{C}^{\pi}_{f,l,0} = \sum_{h=0}^{\tau_f - 2} \widehat{P}^{h-1}_{\pi} (P_{\pi} - \widehat{P}_{\pi}) C^{\pi}_{f,l,h+1} \text{ and}$$
$$C^{\pi}_{f,l,0} - \widehat{C}^{\pi}_{f,l,0} = \sum_{h=0}^{\tau_f - 2} P^{h-1}_{\pi} (P_{\pi} - \widehat{P}_{\pi}) \widehat{C}^{\pi}_{f,l,h+1}.$$

Proof. We only prove the first statement of value function since the proof procedure for cost is

identical. For a fixed h and j

$$
\begin{aligned}
&V^{\pi}_{f,h}(j) - \widehat{V}^{\pi}_{f,h}(j) = r_f(j) + \sum_k P_\pi(k|j) V^{\pi}_{f,h+1}(k) \\
&- (r_f(j) + \sum_k \widehat{P}_\pi(k|j) \widehat{V}^{\pi}_{f,h+1}(k)) \\
&= \sum_k P_\pi(k|j) V^{\pi}_{f,h+1}(k) - \sum_k \widehat{P}_\pi(k|j) V^{\pi}_{f,h+1}(k) \\
&+ \sum_k \widehat{P}_\pi(k|j) V^{\pi}_{f,h+1}(k) - \sum_k \widehat{P}_\pi(k|j) \widehat{V}^{\pi}_{f,h+1}(k) \\
&= \sum_k (P_\pi(k|j) - \widehat{P}_\pi(k|j)) V^{\pi}_{f,h+1}(k) \\
&+ \sum_k \widehat{P}_\pi(k|j)(V^{\pi}_{f,h+1}(k) - \widehat{V}^{\pi}_{f,h+1}(k)).
\end{aligned}
$$

Because $V^{\pi}_{f,\tau_f-1}(j) = \widehat{V}^{\pi}_{f,\tau_f-1}(j) = r_f(j)$, if we expand the second term until $h = \tau_f - 1$, we get the result. □

Lemma 24. *Let $\delta_P \in (0,1)$ and*

$$
|p_l - \widehat{p}_l| \le c_1 + c_2\sqrt{\widehat{p}_l(1-\widehat{p}_l)}
$$

w.p. at least $1-\delta_P$ for each $l \in \mathcal{L}$. Then, for any flow f under any policy π

$$
\begin{aligned}
&|\sum_k (P_\pi(k|j) - \widehat{P}_\pi(k|j)) \widehat{V}^{\pi}_{f,h+1}(k)| \\
&\qquad \le 2c_1 \|\widehat{V}^{\pi}_{f,h+1}\|_\infty + c_2\sqrt{2}\widehat{\sigma}^{\pi}_{f,h}(j)
\end{aligned}
$$

for any $(j,l) \in \mathcal{S} \times \mathcal{L}$ and $h \in [0, \tau_f - 2]$ w.p. at least $1 - 2\delta_P$, and

$$
\begin{aligned}
&|\sum_k (P_\pi(k|j) - \widehat{P}_\pi(k|j)) \widehat{C}^{\pi}_{f,l,h+1}(k)| \\
&\qquad \le 2c_1 \|\widehat{C}^{\pi}_{f,l,h+1}\|_\infty + c_2\sqrt{2}\widehat{\sigma}^{\pi}_{f,l,h}(j)
\end{aligned}
$$

for any $(j,l) \in \mathcal{S} \times \mathcal{L}, l \in \mathcal{L}$ *and* $h \in [0, \tau_f - 2]$ *w.p. at least* $1 - 2\delta_P$.

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

$$\begin{aligned}
&|\sum_k (P_{\pi}(k|j) - \widehat{P}_{\pi}(k|j))\widehat{V}_{f,h+1}^{\pi}(k)| \\
&= |\sum_k (P_{\pi}(k|j) - \widehat{P}_{\pi}(k|j))\widehat{V}_{f,h+1}^{\pi}(k) + \bar{V}_f^{\pi}(j) - \bar{V}_j^{\pi}(j)|
\end{aligned}$$

$$= |\sum_k (P_{\pi}(k|j) - \widehat{P}_{\pi}(k|j))(\widehat{V}_{f,h+1}^{\pi}(k) - \bar{V}_f^{\pi}(k))| \tag{6.18}$$

$$\leq \sum_k |P_{\pi}(k|j) - \widehat{P}_{\pi}(k|j)||\widehat{V}_{f,h+1}^{\pi}(k) - \bar{V}_f^{\pi}(k)| \tag{6.19}$$

$$\begin{aligned}
&\leq \sum_k (c_1 + c_2\sqrt{\widehat{P}_{\pi}(k|j) - (1 - \widehat{P}_{\pi}(k|j))})|\widehat{V}_{f,h+1}^{\pi}(k) - \bar{V}_f^{\pi}(k)| \\
&\leq Sc_1\|\widehat{V}_{f,h+1}^{\pi}\|_{\infty} \\
&+ c_2 \sum_k \sqrt{\widehat{P}_{\pi}(k|j)(1 - \widehat{P}_{\pi}(k|j))(\widehat{V}_{f,h+1}^{\pi}(k) - \bar{V}_f^{\pi}(k))^2} \\
&\leq Sc_1\|\widehat{V}_{f,h+1}^{\pi}\|_{\infty}
\end{aligned}$$

$$+ c_2\sqrt{S\sum_k \widehat{P}_{\pi}(k|j)(1 - \widehat{P}_{\pi}(k|j))(\widehat{V}_{f,h+1}^{\pi}(k) - \bar{V}_f^{\pi}(k))^2} \tag{6.20}$$

$$\begin{aligned}
&\leq Sc_1\|\widehat{V}_{f,h+1}^{\pi}\|_{\infty} + c_2\sqrt{S\sum_k \widehat{P}_{\pi}(k|j)(\widehat{V}_{f,h+1}^{\pi}(k) - \bar{V}_f^{\pi}(k))^2} \\
&= Sc_1\|\widehat{V}_{f,h+1}^{\pi}\|_{\infty} + c_2\sqrt{S}\widehat{\sigma}_{f,h}^{\pi}(j).
\end{aligned}$$

Inequality (B.7) holds w.p. at least $1 - S\delta_P$, since we used the assumption and applied the triangle inequality and union bound. Please note that the it is straightforward to show that assumption leads to the form of inequality (6.18). We then applied the assumed bound on $|\widehat{V}_{f,h+1}^{\pi}(k) - \bar{V}_f^{\pi}(k)|$ and bounded it by $\|\widehat{V}_{f,h+1}^{\pi}\|_{\infty}$ as all value functions are non-negative. In inequality (B.8), we applied the Cauchy-Schwarz inequality and subsequently used the fact that each term is the sum is non-negative and that $(1 - \widehat{P}_\pi(k|j)) \leq 1$. The final equality follows from the definition of $\widehat{\sigma}^\pi_{f,h}(j)$. $\square$

Lemma 25. *Let $\delta_P \in (0,1)$ and*

$$|p_l - \widehat{p}_l| \leq \frac{c_3}{\sqrt{n}}$$

for all $l \in \mathcal{L}$ w.p. at least $1 - \delta_P$. Then, for any flow f under any policy π

$$\|V^\pi_{f,\tau_f-1} - \widehat{V}^\pi_{f,\tau_f-1}\|_\infty \leq \cdots \leq \|V^\pi_{f,0} - \widehat{V}^\pi_{f,0}\|_\infty \leq \frac{c_3 \tau_f \beta_f}{\sqrt{n}},$$

w.p. at least $1 - 2S|\mathcal{L}|\tau_f \delta_P$, and for any l

$$\begin{aligned} &\|C^\pi_{f,l,\tau_f-1} - \widehat{C}^\pi_{f,l,\tau_f-1}\|_\infty \leq \cdots \\ &\cdots \leq \|C^\pi_{f,l,0} - \widehat{C}^\pi_{f,l,0}\|_\infty \leq \frac{c_3 \tau_f \beta_f}{\sqrt{n}} \end{aligned}$$

w.p. at least $1 - 2S|\mathcal{L}|\tau_f \delta_P$.

Proof. We prove the statement of value function since the proof procedure for cost is identical. Let

$\Delta_h = \max_j |V^\pi_{f,h}(j) - \widehat{V}^\pi_{f,h}(j)|$. Then

$$\begin{aligned}
&\Delta_h = |V^\pi_{f,h}(j) - \widehat{V}^\pi_{f,h}(j)| = |r_f(j) + \sum_k P_\pi(k|j)V^\pi_{f,h+1}(k) \\
&- (r_f(j) + \sum_k \widehat{P}_\pi(k|j)\widehat{V}^\pi_{f,h+1}(k))| \\
&= |\sum_k P_\pi(k|j)V^\pi_{f,h+1}(k) - \sum_k \widehat{P}_\pi(k|j)V^\pi_{f,h+1}(k) \\
&+ \sum_k \widehat{P}_\pi(k|j)V^\pi_{f,h+1}(k) - \sum_k \widehat{P}_\pi(k|s)\widehat{V}^\pi_{f,h+1}(k)| \\
&= |\sum_{l\in\mathcal{L}_j} \pi(j,f,h,l)(p_l V^\pi_{f,h+1}(j') + (1-p_l)V^\pi_{f,h+1}(j) \\
&- \widehat{p}_l\widehat{V}^\pi_{f,h+1}(j') - (1-\widehat{p}_l)\widehat{V}^\pi_{f,h+1}(j) \pm \widehat{p}_l V^\pi_{f,h+1}(j') \\
&\pm (1-\widehat{p}_l)V^\pi_{f,h+1}(j))| \\
&= |\sum_{l\in\mathcal{L}_j} \pi(j,f,h,l)((p_l - \widehat{p}_l)V^\pi_{f,h+1}(j') + (\widehat{p}_l - p_l)V^\pi_{f,h+1}(j) \\
&+ \widehat{p}_l(V^\pi_{f,h+1}(j') - \widehat{V}^\pi_{f,h+1}(j')) \\
&+ (1-\widehat{p}_l)(V^\pi_{f,h+1}(j) - \widehat{V}^\pi_{f,h+1}(j)))| \\
&\leq \sum_{l\in\mathcal{L}_j} \pi(j,f,h,l)(|p_l - \widehat{p}_l|\beta_f + \Delta_{h+1}) \\
&\leq \frac{c_3\beta_f}{\sqrt{n}} + \Delta_{h+1}
\end{aligned}$$

Here, $l = (j, j')$. Thus,

$$\Delta_h \leq \frac{c_3\beta_f}{\sqrt{n}} + \Delta_{h+1}$$

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

$$\Delta_{\tau_f - 1} = 0 \leq \cdots \leq \Delta_0 \leq \frac{c_3\tau_f\beta_f}{\sqrt{n}}$$

since $\Delta_{\tau_f-1} = \max_j |r_f(j) - r_f(j)| = 0$. By taking union bound over time-steps, we get the result holds w.p. at least $1 - 2S|\mathcal{L}|\tau_f\delta_P$. Hence the proof is complete. □

Lemma 26. *Let $\delta_P \in (0,1)$ and*

$$|p_l - \widehat{p}_l| \leq \frac{c_3}{\sqrt{n}}$$

w.p. at least $1 - \delta_P$ for all $l \in \mathcal{L}$. Then if $n \geq \frac{c_3^2}{36S^2}$, for any flow f, at any time-step $h \in [0, \tau_f - 1]$ and under any policy π

$$\|\sigma^\pi_{f,h} - \widehat{\sigma}^\pi_{f,h}\|_\infty \leq \frac{2\sqrt{6c_3S}\tau_f\beta_f}{n^{1/4}},$$

w.p. at least $1 - 2S^2\mathcal{L}\tau_f\delta_P$, and similarly for any $l \in \mathcal{L}$

$$\|\sigma^\pi_{f,l,h} - \widehat{\sigma}^\pi_{f,l,h}\|_\infty \leq \frac{2\sqrt{6c_3S}\tau_f\beta_f}{n^{1/4}}$$

w.p. at least $1 - 2S^2|\mathcal{L}|\tau_f\delta_P$.

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

$$\begin{aligned}
&\sigma^{\pi^2}_{f,h}(j) = \sigma^{\pi^2}_{f,h}(j) - \widehat{\mathbb{E}}[(V^\pi_{f,h+1}(j_{h+1}) - \widehat{P}_\pi V^\pi_{f,h+1}(j))^2] \\
&+ \widehat{\mathbb{E}}[(V^\pi_{f,h+1}(j_{h+1}) - \widehat{P}_\pi V^\pi_{f,h+1}(j))^2] \\
&\leq \sum_k (P_\pi(k|j) - \widehat{P}_\pi(k|j)) V^{\pi^2}_{f,h+1}(k) \\
&- [(\sum_k P_\pi(k|j) V^\pi_{f,h+1}(k))^2 - (\sum_k \widehat{P}_\pi(k|j) V^\pi_{f,h+1}(k))^2] \\
&+ \sqrt{\widehat{\mathbb{E}}[(V^\pi_{f,h+1}(j_{h+1}) - \widehat{V}^\pi_1(j_1) - \widehat{P}_\pi(V^\pi_{f,h+1} - \widehat{V}^\pi_{f,h+1})(j))^2]} \\
&+ \sqrt{\widehat{\mathbb{E}}[(\widehat{V}^\pi_{f,h+1}(j_{h+1}) - \widehat{P}_\pi(\widehat{V}^\pi_{f,h+1})(j))^2]}^2,
\end{aligned}$$

where we applied triangular inequality in the last line. And, please note that $\widehat{\mathbb{E}}$ means expectation w.r.t. transition kernel $\widehat{P}_\pi$. It is straightforward to show that $Var_{k\sim\widehat{P}_\pi(\cdot|j)}(V^\pi_{f,h}(k) - \widehat{V}^\pi_{f,h}(k)) \leq \|V^\pi_{f,h} - \widehat{V}^\pi_{f,h}\|^2_\infty$ implying

$$\begin{aligned}\sigma^{\pi^2}_{f,h}(j) \leq & \sum_k (P_\pi(k|j) - \widehat{P}_\pi(k|j))V^{\pi^2}_{f,h+1}(k) \\ & - [\sum_k (P_\pi(k|j) - \widehat{P}_\pi(k|j))V^\pi_{f,h+1}(k)] \\ & \times [\sum_k (P_\pi(k|j) + \widehat{P}_\pi(k|j))V^\pi_{f,h+1}(k)] \\ & + (\|V^\pi_{f,h} - \widehat{V}^\pi_{f,h}\|_\infty + \widehat{\sigma}^\pi_{f,h}(j))^2.\end{aligned}$$

Now, if we use Lemma 53, we get

$$\begin{aligned}\sigma^{\pi^2}_{f,h}(j) &\leq [\widehat{\sigma}^\pi_{f,h}(j) + \frac{c_3\tau_f\beta_f}{\sqrt{n}}]^2 + \frac{6c_3S\beta_f^2}{\sqrt{n}} \\ &\leq [\widehat{\sigma}^\pi_{f,h}(j) + \frac{c_3\tau_f\beta_f}{\sqrt{n}}]^2 + \frac{6c_3S\tau_f^2\beta_f^2}{\sqrt{n}} \\ &\leq [\widehat{\sigma}^\pi_{f,h}(j) + \frac{c_3\tau_f\beta_f}{\sqrt{n}} + \frac{\sqrt{6c_3S}\tau_f\beta_f}{n^{1/4}}]^2 \\ &\leq [\widehat{\sigma}^\pi_{f,h}(j) + \frac{2\sqrt{6c_3S}\tau_f\beta_f}{n^{1/4}}]^2\end{aligned}$$

w.p. at least $1 - 2S^2|\mathcal{L}|\tau_f\delta_P$.[†] Next, we multiplied the $\frac{6c_3S\beta_f^2}{\sqrt{n}}$ by τ_f^2 since $\tau_f \geq 1$. Then, 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. $\widehat{\sigma}^\pi_{f,h}(j) \leq \sigma^\pi_{f,h}(j) + \frac{2\sqrt{6c_3S}\tau_f\beta_f}{n^{1/4}}$. □

Lemma 27. *[7] For any flow f, the variance of the value function defined as $\Sigma^\pi_{f,t}(j) = \mathbb{E}[(\sum_{h=t}^{\tau_f-1} r_f(j_h) - V^\pi_{f,0}(j))^2]$ satisfies a Bellman equation $\Sigma^\pi_{f,t}(j) = \sigma^{\pi^2}_{f,t}(j) + \sum_k P_\pi(k|j)V^\pi_{f,t+1}(k)$ which gives $\Sigma^\pi_{f,t}(j) = \sum_{h=t}^{\tau_f}(P_\pi^{h-1}\sigma^{\pi^2}_{f,h})(j)$. Since $0 \leq \Sigma^\pi_{f,0}(j) \leq (\tau_f\beta_f)^2$, it follows that $0 \leq \sum_{h=0}^{\tau_f-1}(P_\pi^{h-1}\sigma^{\pi^2}_{f,h})(j) \leq (\tau_f\beta_f)^2$ for all $j \in \mathcal{S}$.*

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

Corollary 3. *The result of Lemma 63 also holds for variance of cost functions.*

Lemma 28. *Let* $\delta_P \in (0,1)$. *Then, if* $n \geq 11664 S^2 \tau_f^2 \log 4/\delta_P{}^3$, *for any flow* f *under any policy* π

$$\|V_{f,0}^{\pi} - \widehat{V}_{f,0}^{\pi}\|_{\infty} \leq \sqrt{18 \frac{S \tau_f^3 \beta_f^2 \log 4/\delta_P}{n}}$$

w.p. at least ..., and for any $l \in \mathcal{L}$,

$$\|C_{f,l,0}^{\pi} - \tilde{C}_{f,l,0}^{\pi}\|_{\infty} \leq \sqrt{18 \frac{S \tau_f^3 \beta_f^2 \log 4/\delta_P}{n}}.$$

w.p. at least $1 - 3S^2 |\mathcal{L}| \tau_f \delta_P$.

Proof. We only prove the statement of value function since the proof procedure for cost is identical. First, let

$$c_1 = \frac{2}{3n} \log \frac{4}{\delta_P} \text{ and } c_2 = \sqrt{\frac{2 \log 4/\delta_P}{n}} \tag{6.21}$$

Now, let fix state j :

$$|V^{\pi}_{f,0}(j) - \widehat{V}^{\pi}_{f,0}(j)| = |\sum_{h=0}^{\tau_f - 2} \widehat{P}^{h-1}_{\pi}(P_{\pi} - \widehat{P}_{\pi})V^{\pi}_{f,h+1}|(j) \tag{6.22}$$

$$\leq \sum_{h=0}^{\tau_f - 2} \widehat{P}^{h-1}_{\pi}|(P_{\pi} - \widehat{P}_{\pi})V^{\pi}_{f,h+1}|(j)$$

$$\leq \sum_{h=0}^{\tau_f - 2} \widehat{P}^{h-1}_{\pi}(Sc_1\|V^{\pi}_{f,h+1}\|_{\infty} + c_2\sqrt{S}\sigma^{\pi}_{f,h})(j) \tag{6.23}$$

$$\leq S\tau_f^2\beta_f c_1 + c_2\sqrt{S}\sum_{h=0}^{\tau_f - 1}(\widehat{P}^{h-1}_{\pi}\sigma^{\pi}_{f,h})(j) \tag{6.24}$$

$$\leq S\tau_f^2\beta_f c_1$$

$$+ c_2\sqrt{S}\sum_{h=0}^{\tau_f - 1}(\widehat{P}^{h-1}(\widehat{\sigma}^{\pi}_{f,h} + \frac{2\sqrt{6}(\log 4/\delta_P)^{0.25}S^{0.5}\tau_f\beta_f}{n^{1/4}})(j) \tag{6.25}$$

$$\leq S\tau_f^2\beta_f c_1 + c_2\sqrt{S\tau_f}\sqrt{\sum_{h=0}^{\tau_f - 1}(\widehat{P}^{h-1}\widehat{\sigma}^{\pi^2}_{f,h})(j)} \tag{6.26}$$

$$+ c_2\tau_f\sqrt{S}\frac{2\sqrt{6}(\log 4/\delta_P)^{0.25}S^{0.5}\tau_f\beta_f}{n^{1/4}} \tag{6.27}$$

$$= \frac{2S\tau_f^2\beta_f}{3n} \tag{6.28}$$

$$+ \sqrt{\frac{2S\tau_f^3\beta_f^2\log 4/\delta_P}{n}} + \frac{4 \times 3^{0.5}S\tau_f^2\beta_f\log 4/\delta_P{}^{0.75}}{n^{0.75}} \tag{6.29}$$

$$\leq \sqrt{18\frac{S\tau_f^3\beta_f^2\log 4/\delta_P}{n}}. \tag{6.30}$$

In equation (B.16), we used Lemma 51. Then, we applied Lemma 52 to obtain inequality (B.17). Next, we bound $\|V^{\pi}_{f,h+1}\|_{\infty}$ by $\tau_f\beta_f$ in inequality (B.18). To get inequality (B.19), we use Lemma 54, since we can bound $p_l - \widehat{p}_l$ by c_2. And, we applied Cauchy-Scharwz inequality to get inequality (B.20). To get inequality (B.21) and (6.28), we applied Lemma 63 and substituting c_1 and c_2 according to equations (B.15). Finally, inequality (A.12) follows from the fact that $n \geq 11664S^2\tau_f^2\log 4/\delta_P{}^3$. Since the result is true for every $j \in S$, hence the proof is complete. □

Now, we are ready to prove the Theorem 12

Proof of Theorem 12: Let $\delta_P \in (0,1)$. First, we know that optimization problem (6.17) is feasible. Now, for each flow f, we have

$$V_{f,0}^{\pi^*}(s_f) - \sqrt{18\frac{S\tau_f^3\beta_f^2 \log 4/\delta_P}{n}} \leq \widehat{V}_{f,0}^{\pi^*}(s_f)$$
$$\leq V_{f,0}^{\pi^*}(s_f) + \sqrt{18\frac{S\tau_f^3\beta_f^2 \log 4/\delta_P}{n}}$$

w.p. at least $1 - 3S^2|\mathcal{L}|\tau_f\delta_P$ and

$$V_{f,0}^{\widehat{\pi}}(s_f) - \sqrt{18\frac{S\tau_f^3\beta_f^2 \log 4/\delta_P}{n}} \leq \widehat{V}_{f,0}^{\widehat{\pi}}(s_f)$$
$$\leq V_{f,0}^{\widehat{\pi}}(s_f) + \sqrt{18\frac{S\tau_f^3\beta_f^2 \log 4/\delta_P}{n}}$$

w.p. at least $1 - 3S^2|\mathcal{L}|\tau_f\delta_P$ according to Lemma 28. On the other hand, we know that $\widehat{V}_{f,0}^{\pi^*}(s_f) \leq \widehat{V}_{f,0}^{\widehat{\pi}}(s_f)$. Thus, by combining these results we get

$$V_{f,0}^{\pi^*}(s_f) - \sqrt{18\frac{S\tau_f^3\beta_f^2 \log 4/\delta_P}{n}} \leq \widehat{V}_{f,0}^{\pi^*}(s_f)$$
$$\leq \widehat{V}_{f,0}^{\widehat{\pi}}(s_f) \leq V_{f,0}^{\widehat{\pi}}(s_f) + \sqrt{18\frac{S\tau_f^3\beta_f^2 \log 4/\delta_P}{n}}.$$

It yields that $V_{f,0}^{\widehat{\pi}}(s_f) \geq V_{f,0}^{\pi^*}(s_f) - 2\sqrt{18\frac{S\tau_f^3\beta_f^2 \log 4/\delta_P}{n}}$ w.p. at least $1 - 6S^2|\mathcal{L}|\tau_f\delta_P$ by union bound. Therefore,

$$\sum_f \rho_f V_{f,0}^{\widehat{\pi}}(s_f) \geq \sum_f V_{f,0}^{\pi^*}(s_f) - 2\rho_{\text{tot}}\sqrt{18\frac{S\tau_{\max}^3\beta_{\max}^2 \log 4/\delta_P}{n}}$$

w.p. at least $1 - 6S^2|F||\mathcal{L}|\tau_{\max}\delta_P$.

Now, for any flow f and any $l \in \mathcal{L}$ we have

$$C^{\widehat{\pi}}_{f,l,0}(s_f) \leq \widehat{C}^{\widehat{\pi}}_{f,l,0}(s_f) + \sqrt{18\frac{S\tau_f^3\beta_f^2 \log 4/\delta_P}{n}}$$

w.p. at least $1 - 3S^2|\mathcal{L}|\tau_f\delta_P$ according to Lemma 28. Finally,

$$\begin{aligned}
&\sum_f \rho_f C^{\widehat{\pi}}_{f,l,0}(s_f) \\
&\leq \sum_f \rho_f \widehat{C}^{\widehat{\pi}}_{f,l,0}(s_f) + \rho_{\text{tot}}\sqrt{18\frac{S\tau_{\max}^3\beta_{\max}^2 \log 4/\delta_P}{n}} \\
&\leq C_l + \rho_{\text{tot}}\sqrt{18\frac{S\tau_{\max}^3\beta_{\max}^2 \log 4/\delta_P}{n}}
\end{aligned}$$

w.p. at least $1 - 3S^2|F||\mathcal{L}|\tau_{\max}\delta_P$.

By taking union bound, we get that all statements for value and cost functions hold w.p. at least $1 - 3(|\mathcal{L}|+2)S^2|F||\mathcal{L}|\tau_{\max}\delta_P$. Hence, putting $\epsilon = 2\rho_{\text{tot}}\sqrt{18\frac{S\tau_{\max}^3\beta_{\max}^2 \log 4/\delta_P}{n}}$ and $\delta = 3(|\mathcal{L}| + 2)S^2|F||\mathcal{L}|\tau_{\max}\delta_P$ concludes the proof. Please note that $\epsilon < \frac{\beta_{\max}\rho_{\text{tot}}}{18 \log 4/\delta_P}\sqrt{\frac{2\tau_{\max}}{S}}$ would satisfy the assumption in Lemma 28. □

6.4.1.2 Online-CRL-LP

Online Constrained-RL-LP, or Online-CRL-LP described in Algorithm 8, is an online method proceeding in episodes with length of $\tau_{\max}$. At the beginning of each episode e, Online-CRL-LP constructs an empirical model $\widehat{P}$ according to link visitation frequencies, i.e., $\widehat{P}(k|j,l) = \frac{n(k,j,l)}{n(j,l)}$, where $n(k,j,l)$ and $n(j,l)$ are visitation frequencies. This empirical model $\widehat{P}$ induces a set of finite-horizon MDPs for each flow. Considering the constraints on link capacities, we get a set of network-CMDPs $\mathcal{M}_e$ which any network-CMDP $M' \in \mathcal{M}_e$ has identical horizon and reward and cost matrices. However, for any $(j,l) \in \mathcal{S} \times \mathcal{L}$ and $k \in \mathcal{S}$, $P'(k|j,l)$ lies inside a confidence interval induced by $\widehat{P}$. To construct a confidence interval for any element of $P'(k|j,l)$, we use concentration inequalities as defined by (6.32). Thus the class of network-CMDPs is defined as below at each episode e :

$$\begin{aligned}
&\mathcal{M}_e := \{M' : \{r'_f(j)\}_f = \{r_f(j)\}_f, \{\rho'_f\}_f = \{\rho_f\}_f && (6.31)\\
&c'_l(j,l) = c_l(j,l), C'_l = C_l, \{\tau'_f\}_f = \{\tau_f\}_f, \\
&|P'(k|j,l) - \widehat{P}(k|j,l)| \leq && (6.32)\\
&\min\Big(\sqrt{\frac{2\widehat{P}(k|j,l)(1-\widehat{P}(k|j,l))}{n(j,l)}}\log\frac{4}{\delta_1} + \frac{2}{3n(j,l)}\log\frac{4}{\delta_1}, \\
&\sqrt{\frac{\log 4/\delta_P}{2n(j,l)}}\Big)\forall j,l,k,f\},
\end{aligned}$$

where δ_1 is defined in Algorithm 8. Here, for any $M' \in \mathcal{M}_e$, and flow f, $V'^{\pi}_{f,0}(s_f)$ and $C'^{\pi}_{f,l,0}(s_f)$ are computed according to (6.6) and (6.8) w.r.t. underlying transition kernel P', respectively.

Finally, Online-CRL-LP maximizes the objective functions among all possible transition kernels, while satisfying constraints (if feasible). More specifically, it solves the optimistic planning problem below

$$\max_{\pi, M' \in \mathcal{M}_e} \sum_f \rho_f V'^{\pi}_{f,0}(s_f) \text{ s.t. } \sum_f \rho_f C'^{\pi}_{f,l,0}(s_f) \leq C_l \ \forall\, l. \tag{6.33}$$

Online-CRL-LP uses Extended Linear Programming, or **ELP**, to solve the problem of (6.33). This method inputs $\mathcal{M}_{\delta_P}$ and outputs $\tilde{\pi}$ for the optimal solution. The description of ELP is provided in supplementary materials.

Algorithm 8 describes Online-CRL-LP.

This algorithm draws inspiration from the constrained-RL algorithm Online-CRL [33] with several differences. The formulation of network-CMDP differs from generic CMDP definition in two ways. First, the objective of generic CMDP is defined for one decision process with one horizon, while the network-CMDP contains multiple decision processes with different horizon lengths. Second, the constraints imposed on a network-CMDP are integration of multiple decision processes. This situation does not happen to general CMDPs. In spite of these differences, we

Algorithm 8 Online-CRL-LP

```
Input: accuracy $\epsilon$ and failure tolerance $\delta$.
Set $e = 1, w_{\min} = \frac{|F|\epsilon}{4\tau_{\max} S \rho_{\text{tot}} \beta_{\max}}, U_{\max} = S|\mathcal{L}|m, \delta_1 = \frac{\delta}{4|F|SU_{\max}}$.
Set $m$ according to (6.35) and (6.36).
Set $n(j,l) = n(k,j,l) = 0 \ \forall j, d_l \in \mathcal{S}, l \in \mathcal{L}$.
while there is $(j,l)$ with $n(j,l) < Sm\tau_{\max}$ do
    $\widehat{P}(k|j,l) = \frac{n(k,j,l)}{n(j,l)} \ \forall (j,l)$ with $n(j,l) > 0$ and $d_l \in \mathcal{S}$.
    Construct $\mathcal{M}_e$ according to (6.31).
    $\tilde{\pi}_e = \text{ELP}(\mathcal{M}_e)$.
    for Each flow $f$ do
        for $t = 1, \ldots, \tau_f$ do
            $l \sim \tilde{\pi}_e(j,f,l,h), j_{t+1} \sim P(\cdot|j_t,l_t), n(j_t,l_t)++, n(j_{t+1},j_t,l_t)++$.
    $e++$
```

show that we are able to use existing RL tools for problems defined as network-CMDPs.

We now present the PAC bound of Algorithm 8.

Theorem 13. *For any* $0 < \epsilon, \delta < 1$, *under Online-CRL-LP we have:*

$$\mathbb{P}\Big(\sum_f \rho_f V^{\tilde{\pi}_e}_{f,0}(s_f) \geq \sum_f \rho_f V^{\pi^*}_{f,0}(s_f) - \epsilon \ \textit{and}$$

$$\sum_f \rho_f C^{\tilde{\pi}_e}_{f,l,0}(s_f) \leq C_l + \epsilon \, \forall l \in \mathcal{L}\Big) \geq 1 - \delta,$$

for all but at most

$$\tilde{O}\Big(\frac{S|\mathcal{L}|\rho_{tot}^2 \tau_{\max}^2 \beta_{\max}^2}{\epsilon^2}\Big)$$

episodes.

To prove Theorem 13, we follow an approach motivated by [33]. However, there are several differences in our technique. As mentioned above, one of the differences is with regard to multiple decision processes and integration of these processes within constraints. We will show how to approach this problem and obtain matching sample complexity results.

There are also recent results on characterizing the regret of constrained-RL [5] while using an algorithm reminiscent of Algorithm 8, 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 [27], and following the [5] 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 links and base our proof on these notions. Then we present the key lemmas required to prove Theorem 13. Finally, we sketch the proof of Theorem 13. The detailed analysis is provided in supplementary materials.

Let the weight of $(j, l)-$pair in an episode e for flow f under policy $\tilde{\pi}_e$ be its expected frequency in that episode

$$\begin{aligned} w_{f,e}(j,l) &:= \sum_{h=0}^{\tau_f-1} \mathbb{P}(j_h = j, l \sim \tilde{\pi}_e(j_h, f, \cdot, h)) \\ &= \sum_{h=0}^{\tau_f-1} P_{\tilde{\pi}_e}^{h-1} \mathbb{I}\{j = \cdot, l \sim \tilde{\pi}_e(j, f, \cdot, h)\}(s_f). \end{aligned}$$

Further, we define the *weight* of $(j, l)-$pair in an episode e under policy $\tilde{\pi}_e$ be the cumulative weights of all flows:

$$w_e(j,l) = \sum_f w_{f,e}(j,l).$$

Then, the *importance* $\iota_{e,f}$ of (j, l) at episode e is defined as its relative weight compared to $w_{\min} := \frac{\epsilon}{4\tau_{\max}S}$ on a log-scale

$$\begin{gathered} \iota_e(j,l) := \min\{z_i : z_i \geq \frac{w_e(j,l)}{w_{\min}}\} \\ \text{where } \; z_1 = 0 \text{ and } z_i = 2^{i-2} \;\; \forall i = 2, 3, \ldots . \end{gathered}$$

[‡] Here, $|S|$ and $|A|$ are number state and action spaces respectively and H represents length of horizon.

Note that $\iota_e(j,l) \in \{0,1,2,4,8,16,\dots\}$ is an integer indicating the influence of the link on the value function of $\tilde{\pi}_e$. Similarly, we define *knownness* as

$$\kappa_e(j,l) := \max\{z_i : z_i \leq \frac{n_e(j,l)}{mw_e(j,l)}\} \in \{0,1,2,4,\dots\},$$

which indicates how often (j,l) has been observed relative to its importance. Value of m is defined in Algorithm 8. Now, we can categorize $(j,l)-$pairs into subsets

$$X_{e,\kappa,\iota} := \{(j,l) \in X_e : \kappa_e(j,l) = \kappa, \iota_e(j,l) = \iota\}$$
$$\text{and } \bar{X}_e = \mathcal{S} \times \mathcal{A} \setminus X_e,$$

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

Here, we present the lemmas required for proving the Theorem 13

Lemma 29. *The total number of updates under algorithm 8 is bounded by* $U_{\max} = S|\mathcal{L}|m$.

Proof. Let fix a link l. Note that $n(j,l)$ is not decreasing and also it increases up to $Sm\tau_{\max}$. And, since update of model happens at the beginning of each episode, then maximum number of updates due to a single l happens at most Sm number of times. Thus, maximum number of updates due to all l is no larger than $S|\mathcal{L}|m$. □

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

Lemma 30. $M \in \mathcal{M}_e$ *for all episodes* e *with probability at least* $1 - \frac{\delta}{2|F|(|\mathcal{L}|+1)}$.

Proof. At each episode with model update e and for each l, by Hoeffding's inequality [30] we have

$$|P(k|j,l) - \widehat{P}(k|j,l)| \leq \sqrt{\frac{\log{(4/\delta_1)}}{2n(j,l)}}$$

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

By empirical Brenstein's inequality [31] we have

$$|P(k|j,l) - \widehat{P}(k|j,l)| \leq \sqrt{\frac{2\widehat{P}(k|j,l)(1-\widehat{P}(k|j,l))}{n(j,l)}\log\frac{4}{\delta_1}} \\ + \frac{2}{3n(j,l)}\log\frac{4}{\delta_1}$$

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

Combining above two inequalities and applying union bound, we get

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

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

Lemma 31. *Total number of observations of* $(j,l) \in X_{e,\kappa,\iota}$ *with* $\kappa \in [1, S-1]$ *and* $\iota > 0$ *over all episodes* e *is at most* $3|\mathcal{L}|mw_\iota\kappa$. $w_\iota = \min\{w_e(j,l) : \iota_{e,f}(j,l) = \iota\}$.

Proof. Note that $w_{\iota+1} = 2w_\iota$ for $\iota > 0$. Consider an episode e and a fixed $(j,l) \in X_{e,\kappa,\iota}$. Since we assumed $\iota_e(j,l) = \iota$, then $w_\iota \leq w_e(j,l) \leq 2w_\iota$. Similarly, from $\kappa_e(j,l) = \kappa$ we have $\frac{n_k(j,l)}{2mw_k(j,l)} \leq \kappa \leq \frac{n_k(j,l)}{mw_e(j,l)}$ which implies

$$mw_\iota\kappa \leq mw_e(j,l)\kappa \leq n_e(j,l) \leq 2mw_e(j,l)\kappa \leq 4mw_\iota\kappa. \tag{6.34}$$

Therefore, each (j,l) in $\{(j,l) \in X_{e,\kappa,\iota} : e \in \mathbb{N}\}$ can only be observed $3mw_\iota\kappa$. Then, the total observations is at most $3|\mathcal{L}|mw_\iota\kappa$. □

Lemma 32. *Number of episodes $E_{\kappa,\iota}$ in episodes with $|X_{e,\kappa,\iota}| > \kappa$ is bounded for $\alpha \geq 3$ w.h.p.*

$$\mathbb{P}(E_{\kappa,\iota} > \alpha N) \leq \exp\left(-\frac{\eta w_\iota(\kappa+1)N}{|F|\tau_{\max}}\right),$$

where $N = |\mathcal{L}|m$ and $\eta = \frac{\alpha(3/\alpha-1)^2}{7/3-1/\alpha}$.

Proof. Let $\nu_e := \sum_{h=0}^{\tau_{\max}-1} \mathbb{I}\{(j_h, l_h) \in X_{e,\kappa,\iota}\}$ be number of observations of (j,l) with $|X_{e,\kappa,\iota}| > \kappa$. We have $e \in \{1, ..., E_{\kappa,\iota}\}$.

In these episodes $|X_{e,\kappa,\iota}| \geq \kappa + 1$ and all (j,l) in partition (κ,ι) have $w_e(j,l) \geq w_\iota$, then

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

Also $\mathbb{V}[\nu_e|\nu_1, ..., \nu_{e-1}] \leq \mathbb{E}[\nu_e|\nu_1, ..., \nu_{e-1}]\tau_{\max}$ since $\nu_e \in [0, |F|\tau_{\max}]$.

Now, we define the continuation:

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

and centralized auxiliary sequence

$$\bar{\nu}_e := \frac{\nu_e^+ w_\iota(\kappa+1)}{\mathbb{E}[\nu_e^+|\nu_1^+, \dots, \nu_{e-1}^+]}.$$

By construction

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

According to lemma 57, we have $E_{\kappa,\iota} > \alpha N$ if

$$\sum_{e=1}^{\alpha N} \bar{\nu}_e \leq 3Nw_\iota \kappa \leq 3Nw_\iota(\kappa+1).$$

Now, we define martingale below

$$B_e := \mathbb{E}\left[\sum_{i=1}^{\alpha N} \bar{\nu}_i | \bar{\nu}_1, \ldots, \bar{\nu}_e\right] = \sum_{i=1}^{e} \bar{\nu}_i + \sum_{i=e+1}^{\alpha N} \mathbb{E}[\bar{\nu}_i | \bar{\nu}_1, \ldots, \bar{\nu}_i],$$

which gives $B_0 = \alpha N w_\iota(\kappa+1)$ and $B_{\alpha N} = \sum_{e=1}^{\alpha N} \bar{\nu}_e$. Now, since $\nu_e^+ \in [0, |F|\tau_{\max}]$

$$\begin{aligned}
|B_{e+1} - B_e| &= |\bar{\nu}_e - \mathbb{E}[\bar{\nu}_e | \bar{\nu}_1, \ldots, \bar{\nu}_{e-1}]| \\
&= \left| \frac{w_\iota(\kappa+1)(\nu_e^+ - \mathbb{E}[\nu_e^+ | \bar{\nu}_1, \ldots, \bar{\nu}_{e-1}])}{\mathbb{E}[\nu_e^+ | \nu_1^+, \ldots, \nu_{e-1}^+]} \right| \\
&\leq |\nu_e^+ - \mathbb{E}[\nu_e^+ | \bar{\nu}_1, \ldots, \bar{\nu}_{e-1}]| \leq |F|\tau_{\max}.
\end{aligned}$$

Using

$$\begin{aligned}
\sigma^2 &:= \sum_{e=1}^{\alpha N} \mathbb{V}[B_e - B_{e-1} | B_1 - B_0, \ldots, B_{e-1} - B_{e-2}] \\
&= \sum_{e=1}^{\alpha N} \mathbb{V}[\bar{\nu}_e | \bar{\nu}_1, \ldots, \bar{\nu}_{e-1}] \leq \alpha N |F| \tau_{\max} w_\iota(\kappa+1) \\
&= B_0 |F| \tau_{\max}
\end{aligned}$$

We can apply Theorem 22 of [69] and obtain

$$\begin{aligned}
\mathbb{P}(E_{\kappa,\iota} > \alpha N) &\leq \mathbb{P}\left(\sum_{e=1}^{\alpha N} \bar{\nu}_e \leq 3Nw_\iota(\kappa+1)\right) \\
&= \mathbb{P}(B_{\alpha N} - B_0 \leq 3B_0/\alpha - B_0) \\
&\leq \exp\left(-\frac{(3/\alpha - 1)^2 B_0^2}{2\sigma^2 + |F|\tau_{\max}(1/3 - 1/\alpha)B_0}\right)
\end{aligned}$$

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

$$\mathbb{P}(E_{\kappa,\iota} > \alpha N) \leq \exp - \frac{\alpha(3/\alpha)^2}{7/3 - 1/\alpha} \frac{N w_\iota(\kappa+1)}{|F|\tau_{\max}}.$$

□

Lemma 33. *Suppose E is the number of episodes e for which there are κ and ι with $|X_{e,\kappa,\iota}| > \kappa$, i.e. $E = \sum_{e=1}^{\infty} \mathbb{I}\{\exists(\kappa,\iota) : |X_{e,\kappa,\iota}| > \kappa\}$ and let*

$$m \geq \frac{6\tau_{\max}^2}{\epsilon} \log \frac{2|F|(|\mathcal{L}|+1)E_{\max}}{\delta}, \tag{6.35}$$

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

Proof. Since $w_e(j,l) \leq |F|\tau_{\max}$, we have that $\frac{w_e(j,l)}{w_{min}} < \frac{|F|\tau_{\max}}{w_{min}}$ and so $\iota_e(j,l) \leq |F|\tau_{\max}/w_{min} = 4\tau_{\max}^2 S/\epsilon$. In addition, $|X_{e,\kappa,\iota}| \leq |\mathcal{L}|$ for all e, κ, ι and so $|X_{e,\kappa,\iota}| > \kappa$ can only be true for $\kappa \leq S$. Hence, only $E_{max} = \log_2 \frac{|F|\tau_{\max}}{w_{min}} \log_2 S$ possible values for (κ,ι) exists that can have $|X_{e,\kappa,\iota}| > \kappa$. By union bound over all (κ,ι) and lemma 58, we get

$$\begin{aligned}
\mathbb{P}(E \leq \alpha N E_{\max}) &\geq \mathbb{P}(\max_{(\kappa,\iota)} E_{\kappa,\iota} \leq \alpha N) \\
&\geq 1 - E_{\max} \exp\left(-\frac{\eta w_\iota(\kappa+1)N}{|F|\tau_{\max}}\right) \\
&\geq 1 - E_{\max} \exp\left(-\frac{\eta w_{min} N}{|F|\tau_{\max}}\right) \\
&= 1 - E_{\max} \exp\left(-\frac{\eta w_{\min} m|\mathcal{L}|}{|F|\tau_{\max}}\right) \\
&= 1 - E_{\max} \exp\left(-\frac{\eta \epsilon m|\mathcal{L}|}{4\tau_{\max}^2 S}\right).
\end{aligned}$$

Bounding the right hand-side by $1 - \frac{\delta}{2|F|(|\mathcal{L}|+1)}$ and solving for m gives

$$\begin{aligned}
1 - E_{\max} \exp\left(-\frac{\eta\epsilon m|\mathcal{L}|}{4\tau_{\max}^2|S|}\right) &\geq 1 - \delta/2 \\
\Leftrightarrow m &\geq \frac{4\tau_{\max}^2 S}{|\mathcal{L}|\eta\epsilon} \ln \frac{2E_{\max}}{\delta}.
\end{aligned}$$

Hence, the condition

$$m \geq \frac{4\tau_{\max}^2}{\eta\epsilon} \ln \frac{2E_{\max}}{\delta}$$

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

Lemma 34. *Let* $\delta_1 \in (0,1)$. *Assume* $p, \widehat{p}, \tilde{p} \in [0,1]$ *satisfy* $\mathbb{P}(p \in \mathcal{P}_{\delta_1}) \geq 1 - \delta_1$ *and* $\tilde{p} \in \mathcal{P}_{\delta_1}$ *where*

$$\begin{aligned}\mathcal{P}_{\delta_1} := \{p' \in [0,1] : |p' - \widehat{p}| \leq \min\Big(&\sqrt{\frac{2\widehat{p}(1-\widehat{p})}{n}\log 4/\delta_1} \\ &+ \frac{2}{3n}\log 4/\delta_1, \sqrt{\frac{\log 4/\delta_P}{2n}}\Big)\}.\end{aligned}$$

Then,

$$\begin{aligned}|p - \tilde{p}| \leq &\sqrt{\frac{8\tilde{p}(1-\tilde{p})}{n}\log 4/\delta_1} + 2\sqrt{2}\Big(\frac{\log 4/\delta_1}{n}\Big)^{\frac{3}{4}} \\ &+ 3\sqrt{2}\frac{\log 4/\delta_1}{n}\end{aligned}$$

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

Proof.

$$\begin{aligned}|p - \tilde{p}| &\leq |p - \widehat{p}| + |\widehat{p} - \tilde{p}| \leq 2\sqrt{\frac{2\widehat{p}(1-\widehat{p})}{n}\log 4/\delta_1} + \frac{4}{3n}\log 4/\delta_1 \\ &\leq 2\sqrt{\frac{2\log 4/\delta_1}{n}(\tilde{p} + \sqrt{\frac{\log 4/\delta_1}{2n}})(1 - \tilde{p} + \sqrt{\frac{\log 4/\delta_1}{2n}})} + \frac{4}{3n}\log 4/\delta_1 \\ &= 2\sqrt{\frac{2\log 4/\delta_1}{n}\Big(\tilde{p}(1-\tilde{p}) + \sqrt{\frac{\log 4/\delta_1}{2n}} + \frac{\log 4/\delta_1}{2n}\Big)} + \frac{4}{3n}\log 4/\delta_1 \\ &\leq \sqrt{\frac{8\tilde{p}(1-\tilde{p})}{n}\log 4/\delta_1} + 2\sqrt{2}\Big(\frac{\log 4/\delta_1}{n}\Big)^{\frac{3}{4}} + 3\sqrt{2}\frac{\log 4/\delta_1}{n}.\end{aligned}$$

The first term in the first line is true w.p. at least $1 - \delta_1$, hence the proof is complete. □

Corollary 4. *If we substitute the* δ_P *with* δ_1 *in Lemma 52, the result will pertain.*

Lemma 35. *Assume $M \in \mathcal{M}_e$. If $|X_{e,\kappa,\iota}| \leq \kappa$ for all (κ, ι) and $0 < \epsilon \leq 1$ and*

$$\begin{aligned} m = 1280\frac{S\tau_{\max}^2\rho_{tot}^2\beta_{\max}^2}{\epsilon^2}(\log_2 \log_2 \tau_{\max})^2 \\ \log_2^2\Big(\frac{8S^2\rho_{tot}^2\beta_{\max}^2\tau_{\max}^2}{\epsilon}\Big)\log\frac{4}{\delta_1}, \end{aligned} \tag{6.36}$$

then for any flow f, $|\tilde{V}_{f,0}^{\tilde{\pi}_e}(s_f) - V_{f,0}^{\tilde{\pi}_e}(s_f)| \leq \frac{\epsilon}{\rho_{tot}}$ and for any l, $|\tilde{C}_{f,l,0}^{\tilde{\pi}_e}(s_f) - C_{f,l,0}^{\tilde{\pi}_e}(s_f)| \leq \frac{\epsilon}{\rho_{tot}}$.

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

Consider an individual flow f. Before proceeding, in this lemma we reason about a sequence of CMDPs $M_{f,d}$ which have the same transition probabilities but different reward matrix $r_f^{(d)}$ and cost matrices $c_l^{(d)}$. Here, we only present the definition of $r_f^{(d)}$, as definition of $c_l^{(d)}$ is identical to $r_f^{(d)}$. For $d = 0$, the reward matrix is the original reward function r_f of M_f ($r_f^{(0)} = r_f$.) The following reward matrices are then defined recursively as $r_f^{(2d+2)} = \max_h \sigma_{f,h:\tau_f-1}^{(d),2}$, where $\sigma_{f,h:\tau_f-1}^{(d),2}$ is local variance of the value function w.r.t. the rewards $r_f^{(d)}$. Note that for every d and $h = 0, \ldots, \tau_f - 1$ and $j \in \mathcal{S}$, we have $r_f^{(d)}(j) \in [0, \beta_f^d \tau_f^d]$.

In addition, we will drop the notations k, f and policy $\tilde{\pi}_e$ in the following lemmas, since the statements are for a fixed episode k and flow f and all value functions, reward matrices and transition kernels are defined under policy $\tilde{\pi}_e$. Please note that s_0 is the source node of that particular flow.

Now,

$$\Delta_d := |V_0^{(d)}(s_0) - \tilde{V}_0^{(d)}(s_0)| = |\sum_{h=0}^{\tau-2} P^{h-1}(P - \tilde{P})\tilde{V}_{h+1}^{(d)}(s_0)|$$

$$\leq \sum_{h=0}^{\tau-1} P^{h-1}|P - \tilde{P}\tilde{V}_{h+1}^{(d)}|(s_0)$$

$$= \sum_{h=0}^{\tau-1} P^{h-1}$$

$$\left(\sum_{j\in\mathcal{S},l\in\mathcal{L}_j} \mathbb{I}\{j = \cdot, l \sim \tilde{\pi}(j, \cdot, h)\}|(P - \tilde{P})\tilde{V}_{h+1}^{(d)}|\right)(s_0)$$

$$= \sum_{j\in\mathcal{S},l\in\mathcal{L}_j} \sum_{h=0}^{\tau-1} P^{h-1}$$

$$\left(\mathbb{I}\{j = \cdot, l = \tilde{\pi}(j, \cdot, h)\}|(P - \tilde{P})\tilde{V}_{h+1}^{(d)}|\right)(s_0)$$

$$= \sum_{j\in\mathcal{S},l\in\mathcal{L}_j} \sum_{h=0}^{\tau-1} P^{h-1}$$

$$\left(\mathbb{I}\{j = \cdot, l = \tilde{\pi}(j, \cdot, h)\}|(P - \tilde{P})\tilde{V}_{h+1}^{(d)}(j)|\right)(s_0)$$

The first equality follows from Lemma 51, 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 (j, l) pairs. The fourth step relies on the linearity of P operators. In the fifth step, we realize that $\mathbb{I}\{j = \cdot, l \sim \tilde{\pi}(j, \cdot, h)\}|(P - \tilde{P})\tilde{V}_{h+1}^{(d)}(\cdot)|$ is a function that takes nonzero values for input j. We can therefore replace the argument of the second term with j without changing the value. The term becomes constant and by linearity of P, we can write

$$|V_0^{(d)}(s_0) - \tilde{V}_0^{(d)}(s_0)| = \Delta_d \leq \sum_{j \in \mathcal{S}, l \in \mathcal{L}_j} \sum_{h=0}^{\tau-1} P^{h-1} \left(\mathbb{I}\{j = \cdot, l \sim \tilde{\pi}(j, \cdot, h)\} |(P - \tilde{P})\tilde{V}_{h+1}^{(d)}(s)| \right)(s_0)$$

$$\leq \sum_{j,l \notin X} \sum_{h=0}^{\tau-1} \|\tilde{V}_{h+1}^{(d)}\|_\infty (P^{h-1}\mathbb{I}\{j = \cdot, l \sim \tilde{\pi}(j, \cdot, h)\})(s_0)$$

$$+ \sum_{j,l \in X} \sum_{h=0}^{\tau-1} |(P - \tilde{P})\tilde{V}_{h+1}^{(d)}(j)| \times (P^{h-1}\mathbb{I}\{j = \cdot, l \sim \tilde{\pi}(j, \cdot, h)\})(s_0)$$

$$\leq \sum_{j,l \notin X} \sum_{h=0}^{\tau-1} \beta^{d+1}\tau^{d+1} (P^{h-1}\mathbb{I}\{j = \cdot, l \sim \tilde{\pi}(j, \cdot, h)\})(s_0)$$

$$+ \sum_{j,l \in X} \sum_{h=0}^{\tau-1} |(P - \tilde{P})\tilde{V}_{h+1}^{(d)}(j)| \times (P^{h-1}\mathbb{I}\{j = \cdot, l \sim \tilde{\pi}(j, \cdot, h)\})(s_0)$$

$$\leq \sum_{j,l \notin X} \sum_{h=0}^{\tau-1} \beta^{d+1}\tau^{d+1} (P^{h-1}\mathbb{I}\{j = \cdot, l \sim \tilde{\pi}(j, \cdot, h)\})(s_0)$$

$$+ \sum_{j,l \in X} \sum_{h=0}^{\tau-1} |Sc_1(j,l)\beta^{d+1}\tau^{d+1} + c_2(j,l)\sqrt{S}\tilde{\sigma}_h^{(d)}(j,l)| \times (P^{h-1}\mathbb{I}\{j = \cdot, l \sim \tilde{\pi}(j, \cdot, h)\})(s_0)$$

$$\leq \sum_{j,l \notin X} \sum_{h=0}^{\tau} \beta^{d+1}\tau^{d+1} (P^{h-1}\mathbb{I}\{j = \cdot, l \sim \tilde{\pi}(j, \cdot, h)\})(s_0)$$

$$+ \sum_{j,l \in X} \sum_{h=0}^{\tau} |Sc_1(j,l)\beta^{d+1}\tau^{d+1}| (P^{h-1}\mathbb{I}\{j = \cdot, l \sim \tilde{\pi}(j, \cdot, h)\})(s_0)$$

$$+ \sum_{j,l \in X} \sum_{h=0}^{\tau-1} |\sqrt{S}c_2(j,l)\tilde{\sigma}_h^{(d)}(j,l)| \times (P^{h-1}\mathbb{I}\{j = \cdot, l \sim \tilde{\pi}(j, \cdot, h)\})(s_0)$$

$$\leq \sum_{j,l \notin X} \beta^{d+1}\tau^{d+1}w(j,l) + \sum_{j,l \in X} Sc_1(j,l)\tau^{d+1}w(j,l)$$

$$+ \sum_{j,l \in X} \sqrt{S}c_2(j,l) \times \sum_{h=0}^{\tau-1} \tilde{\sigma}_h^{(d)}(j,l)(P^{h-1}\mathbb{I}\{j = \cdot, l \sim \tilde{\pi}(j, \cdot, h)\})(s_0)$$

$$\leq w_{\min}S\beta^{d+1}\tau^{d+1} + \sum_{j,l \in X} Sc_1(j,l)\beta^{d+1}\tau^{d+1}w(j,l)$$

$$+ \sum_{j,l \in X} \sqrt{S}c_2(j,l) \times \sum_{h=0}^{\tau-1} \tilde{\sigma}_h^{(d)}(j,l)(P^{h-1}\mathbb{I}\{j = \cdot, l \sim \tilde{\pi}(j, \cdot, h)\})(s_0)$$

In the second inequality, we split the sum over all (j, l) pairs and used the fact that P and $\tilde{P}$ are non-expansive. The next step follows from $\|V_{h+1}^{(d)}\|_\infty \leq \|V_0^{(d)}\|_\infty \leq \beta^{d+1}\tau^{d+1}$. We then apply Lemma 52 and subsequently use that all terms are nonnegative and the definition of $w(j, l)$. Therefore,

$$
\begin{aligned}
&\Delta_d \leq \\
&\frac{|F|\beta^d\tau^d\epsilon}{4\rho_{\text{tot}}} + \sum_{j,l\in X} Sc_1(j,l)\beta^{d+1}\tau^{d+1}w(j,l) + \sum_{j,l\in X}\sqrt{S}c_2(j,l) \\
&\times \sum_{h=0}^{\tau-1}\tilde{\sigma}_h^{(d)}(j,l)(P^{h-1}\mathbb{I}\{j=\cdot, l\sim\tilde{\pi}(j,\cdot,h)\})(s_0)
\end{aligned}
$$

Eventually, this step comes from the fact that $w(j, l) \leq w_{\min}$ for all (j, l) 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 Lemma 60 and get that

$$
\begin{aligned}
c_1(j,l) &= 2\sqrt{2}\Big(\frac{\log 4/\delta_1}{n(j,l)}\Big)^{3/4} + 3\sqrt{2}\frac{\log 4/\delta_1}{n(j,l)} \\
\text{and}\quad c_2(j,l) &= \sqrt{\frac{8}{n(j,l)}\log 4/\delta_1}.
\end{aligned}
$$

Plugging definitions above we have

$$
\begin{aligned}
\Delta_d \leq\ &\frac{|F|\beta^d\tau^d\epsilon}{4\rho_{\text{tot}}} + 2\sqrt{2}S\beta^{d+1}\tau^{d+1}\log 4/\delta_1{}^{3/4}\sum_{j,l\in X}\frac{w(j,l)}{n(j,l)^{3/4}} \\
&+ 3\sqrt{2}S\beta^{d+1}\tau^{d+1}\log 4/\delta_1\sum_{j,l\in X}\frac{w(j,l)}{n(j,l)} \\
&+ \sqrt{8S\log 4/\delta_1}\sum_{j,l\in X}\frac{1}{\sqrt{n(j,l)}}\sum_{h=0}^{\tau-1}\tilde{\sigma}_h^{(d)}(j,l) \\
&\times (P^{h-1}\mathbb{I}\{j=\cdot, l\sim\tilde{\pi}(j,\cdot,h)\})(s_0)
\end{aligned}
$$

Hence, we bound

$$\Delta_d \leq \frac{|F|\beta^d \tau^d \epsilon}{4\rho_{\text{tot}}} + U_d(s_0) + Y_d(s_0) + Z_d(s_0)$$

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

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

In the second line, we used Cauchy-Scharwz. Next, we used the fact that for $s, a \in X_{\kappa,\iota}$, we have $n(j,l) \geq mw(j,l)\kappa$, refer to equation (B.22). 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 (κ, ι) pairs.

The next term is

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

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

$$Y_d(s_0) \leq \sqrt{\frac{8S\beta^{2d+3}\tau^{2d+3}\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m}}. \tag{6.37}$$

However, we can improve this bound as follows

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

In the third step, we used Lemma 63 and definition of $r^{(2d+2)}$.

The last term is

$$
\begin{aligned}
Z_d(s_0) &= 3\sqrt{2}S\beta^{d+1}\tau^{d+1}\log 4/\delta_1 \sum_{j,l\in X} \frac{w(j,l)}{n(j,l)} \\
&\leq 3\sqrt{2}S\beta^{d+1}\tau^{d+1}\log 4/\delta_1 \sum_{\kappa,\iota} \frac{|X_{\kappa,\iota}|}{m\kappa} \\
&\leq \frac{3\sqrt{2}S\beta^{d+1}\tau^{d+1}\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m}
\end{aligned}
$$

which we used $n(j,l) \geq mw(j,l)\kappa$ again.

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

$$
\begin{aligned}
\Delta_d \leq & \frac{|F|\beta^d\tau^d\epsilon}{4\rho_{\text{tot}}} + 2\sqrt{2}S\beta^{d+5/4}\tau^{d+5/4}\log 4/\delta_1{}^{3/4} \\
& \times |\mathcal{K}\times\mathcal{I}|\Big(\frac{1}{m}\Big)^{3/4} + \frac{3\sqrt{2}S\beta^{d+1}\tau^{d+1}\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m} \\
& + \sqrt{\frac{8S\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m}\beta^{2d+2}\tau^{2d+2}} \\
& + \sqrt{\frac{8S\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m}\Delta_{2d+2}}.
\end{aligned}
$$

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 $\gamma = \lceil \frac{\log\tau}{2\log 2} \rceil$ with $\mathcal{D} = \{0, 2, 6, 14, \ldots, \gamma\}$, we get

$$
\begin{aligned}
\Delta_0 \leq & \sum_{d\in\mathcal{D}\setminus\gamma} \Big(\frac{8S\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m}\Big)^{\frac{d}{d+2}} \\
& \times \Big[\frac{|F|\beta^d\tau^2\epsilon}{4\rho_{\text{tot}}} + 3\sqrt{\frac{8S\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|\tau^{2d+2}}{m}}\Big]^{\frac{2}{d+2}} \\
& + \Big(\frac{8S\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m}\Big)^{\frac{\gamma}{\gamma+2}} \\
& \times \Big[\frac{|F|\beta^\gamma\tau^\gamma\epsilon}{4\rho_{\text{tot}}} + 3\sqrt{\frac{8S\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|\tau^{2\gamma+2}}{m}}\Big]^{\frac{2}{\gamma+2}}.
\end{aligned}
$$

Here, we used inequality (B.23) to bound $Z_\gamma(s_0)$. Finally, the proof completes if we let

$$m = 1280\frac{S\beta^2\rho_{\text{tot}}^2\tau^2}{\epsilon^2}(\log_2\log_2\tau)^2\log_2^2\Big(\frac{8S^2\beta^2\rho_{\text{tot}}^2\tau^2}{\epsilon}\Big)\log\frac{6}{\delta_1}.$$

□

***Proof of Theorem 13*:** By Lemma 33, we know that number of episodes where $|X_{\kappa,\iota}| > \kappa$ for some κ, ι is bounded by $6E_{\max}|\mathcal{L}|m$ with probability at least $1 - \frac{\delta}{2|F|(|\mathcal{L}|+1)}$. For all other episodes, we have by Lemma 35 that for any flow f and any link l

$$|\tilde{V}_{f,0}^{\tilde{\pi}_e}(s_f) - V_{f,0}^{\tilde{\pi}_e}(s_f)| \leq \frac{\epsilon}{\rho_{\text{tot}}},\ |\tilde{C}_{f,l,0}^{\tilde{\pi}_e}(s_f) - C_{f,l,0}^{\tilde{\pi}_e}(s_f)| \leq \frac{\epsilon}{\rho_{\text{tot}}}. \tag{6.38}$$

Using Lemma 30, we get that $M \in \mathcal{M}_e$ for any episode e w.p. at least $1 - \frac{\delta}{2|F|(|\mathcal{L}|+1)}$. Further, we know that ELP outputs the policy $\tilde{\pi}_e$ such that

$$\tilde{V}_{f,0}^{\tilde{\pi}_e}(s_f) \geq V_{f,0}^{\pi^*}(s_f), \quad \tilde{C}_{f,l,0}^{\tilde{\pi}_e}(s_f) \leq C_l \ \forall l \tag{6.39}$$

w.p. at least $1 - \frac{\delta}{2|F|(|\mathcal{L}|+1)}$. Combining the inequalities (C.2) with inequalities (B.25), we get that for all episodes with $|X_{\kappa,\iota}| \leq \kappa$ for all κ, ι

$$V_{f,0}^{\tilde{\pi}_e}(s_f) \geq V_{f,0}^{\pi^*}(s_f) - \frac{\epsilon}{\rho_{\text{tot}}}$$

w.p. at least $1 - \frac{\delta}{2|F|(|\mathcal{L}|+1)}$. Thus we have

$$\sum_f \rho_f V_{f,0}^{\tilde{\pi}_e}(s_f) \geq \sum_f \rho_f V_{f,0}^{\pi^*}(s_f) - \epsilon$$

w.p. at least $1 - \frac{\delta}{2(|\mathcal{L}|+1)}$.

Next for any $l, C^{\tilde{\pi}_e}_{f,l,0}(s_f) \leq \tilde{C}^{\tilde{\pi}_e}_{f,l,0}(s_f) + \frac{\epsilon}{\rho_{\text{tot}}}$. So,

$$\sum_f \rho_f C^{\tilde{\pi}_e}_{f,l,0}(s_f) \leq \sum_f \rho_f \tilde{C}^{\tilde{\pi}_e}_{f,l,0}(s_f) + \epsilon \leq C_l + \epsilon$$

w.p. 1, since the optimistic problem of (6.33) is feasible in this context. Applying the union bound we get the desired result, if m satisfies

$$\begin{aligned} m &\geq 1280 \frac{S\beta^2_{\max}\rho^2_{\text{tot}}\tau^2_{\max}}{\epsilon^2}(\log_2\log_2\tau_{\max})^2 \\ &\times \log^2_2\Big(\frac{8\beta^2_{\max}\rho^2_{\text{tot}}\tau^2_{\max}S^2}{\epsilon}\Big)\log\frac{4}{\delta_1} \text{ and} \\ m &\geq \frac{6\tau^2_{\max}}{\epsilon}\log\frac{2E_{\max}}{\delta}. \end{aligned}$$

From the definitions, we get

$$\log\frac{4}{\delta_1} = \log\frac{4|F|SU_{\max}}{\delta} = \log\frac{4|F|S|\mathcal{L}|m}{\delta}.$$

Thus,

$$\begin{aligned} m \geq\ & 1280 \frac{S\beta^2_{\max}\rho^2_{\text{tot}}\tau^2_{\max}}{\epsilon^2}(\log_2\log_2\tau_{\max})^2 \\ &\times \log^2_2\Big(\frac{8\beta^2_{\max}\rho^2_{\text{tot}}\tau^2_{\max}S^2}{\epsilon}\Big)\log\frac{4|F|S|\mathcal{L}|m}{\delta}. \end{aligned}$$

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

$$\begin{aligned} m \geq\ & 2560 \frac{S\beta^2_{\max}\rho^2_{\text{tot}}\tau^2_{\max}}{\epsilon^2}(\log_2\log_2\tau_{\max})^2 \\ &\times \log^2_2\Big(\frac{8\beta^2_{\max}\rho^2_{\text{tot}}\tau^2_{\max}S^2}{\epsilon}\Big) \\ &\times \Big[\log\Big(\frac{2048|F|S|\mathcal{L}|\tau^2_{\max}(\log_2\log_2\tau_{\max})^2}{\epsilon^2\delta} \\ &+ \log\log^2_2\Big(\frac{8\tau^2_{\max}|S|^2}{\epsilon}\Big)\Big)\Big]. \end{aligned}$$

Also,

$$E_{\max} = \log_2 S \log_2 \frac{4S\beta_{\max}^2 \rho_{\text{tot}}^2 \tau_{\max}^2}{\epsilon} \le \log_2^2 \frac{4S\beta_{\max}^2 \rho_{\text{tot}}^2 \tau_{\max}^2}{\epsilon}$$

and

$$\begin{aligned}
&\log \frac{2|F|E_{\max}}{\delta} \\
&= \log \frac{2|F| \log_2 S \log_2(4S\beta_{\max}^2 \rho_{\text{tot}}^2 \tau_{\max}^2/\epsilon)}{\delta} \\
&\le \log \frac{2|F| \log_2^2(4S\beta_{\max}^2 \rho_{\text{tot}}^2 \tau_{\max}^2/\epsilon)}{\delta} \\
&\le \log \frac{16|F|S^3|\mathcal{L}|\tau_{\max}^2}{\epsilon\delta}.
\end{aligned}$$

Setting

$$\begin{aligned}
m = 2560 &\frac{S\beta_{\max}^2 \rho_{\text{tot}}^2 \tau_{\max}^2}{\epsilon^2} (\log_2 \log_2 \tau_{\max})^2 && (6.40) \\
\times \log_2^2 &\Big(\frac{8\beta_{\max}^2 \rho_{\text{tot}}^2 \tau_{\max}^2 S^2}{\epsilon} \Big) && (6.41) \\
\times \Big[&\log \Big(\frac{2048|F|S^3|\mathcal{L}|\tau_{\max}^2}{\epsilon^2\delta} (\log_2 \log_2 \tau_{\max})^2 \\
+ &\log \log_2^2 \Big(\frac{8\tau_{\max}^2 S^2}{\epsilon} \Big) \Big) \Big].
\end{aligned}$$

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

$$\begin{aligned}
6E_{\max}|\mathcal{L}|m = 15360 &\frac{S|\mathcal{L}|\beta_{\max}^2 \rho_{\text{tot}}^2 \tau_{\max}^2}{\epsilon^2} (\log_2 \log_2 \tau_{\max})^2 \\
&\times \log_2^2 \Big(\frac{4S\beta_{\max}^2 \rho_{\text{tot}}^2 \tau_{\max}^2}{\epsilon} \Big) \\
&\times \Big[\log \Big(\frac{2048|F|S^3|\mathcal{L}|\tau_{\max}^2}{\epsilon^2\delta} (\log_2 \log_2 \tau_{\max})^2 \\
&+ \log \log_2^2 \Big(\frac{8\tau_{\max}^2 S^2}{\epsilon} \Big) \Big) \Big]
\end{aligned}$$

sub-optimal episodes. □

6.4.2 Algorithms based on Dual Decomposition

In this section, we propose two model-based RL algorithms. Both algorithms operate in a loop consisting of two steps. First, they solve the per-packet MDP under the current model to obtain sub-optimal solution of $V_f^*(\lambda)$ with high probability. Next, the Lagrange multipliers, λ are updated according to estimated model. The difference between the two algorithms is in how they sample the system in order to construct the model and update it, using an offline or online approach. We will show how both algorithms would result in an $\epsilon-$optimal policy with high probability. We also characterize the sample complexity according to Definition 6.

6.4.2.1 Generative Model-Based Learning-Dual

According to the GMBL-Dual algorithm, we use a traditional channel sounding approach, and simply send n packets over each link for estimating the reliability of each link p_l. For each link $l = (j, k)$, the transmission of a packet is 'successful' if the packet transmitted from node j in the link l reaches node j in one time slot. We define the empirical link reliability, $\hat{p}_l$, as the ratio of the successful transmission to the total number of transmission. Given $\hat{p}_l$, we can define the approximate transmission kernel $\hat{P}$ as (6.4) by replacing p_l with $\hat{p}_l$. It is straight forward to see that $\hat{p}_l$ is an unbiased estimator of p_l and $\hat{P}$ is an unbiased estimator of P. The expectation w.r.t. to this approximate transition kernel $\widehat{P}$ is denoted by $\hat{\mathbb{E}}[\cdot]$.

We now consider a different constrained MDP that is identical to the CMDP defined in Section 6.3 except that its transition kernel is $\hat{P}$ instead of P. The expectation w.r.t. $\hat{P}$ is denoted by $\hat{\mathbb{E}}[\cdot]$. We define the quantities $\hat{V}_f^\pi(\lambda)$ in the same way as in (6.13) but by replacing $\mathbb{E}$ by $\hat{\mathbb{E}}$. The quantities $\hat{L}(\pi, \lambda), \hat{D}(\lambda)$ can also now be defined in a similar way as in (6.14) and (6.15) by replacing $V_f^\pi(\lambda)$ with $\hat{V}_f^\pi(\lambda)$. The optimal dual variable $\hat{\lambda}^*$ is defined as $\hat{\lambda}^* = \arg\min_\lambda \hat{D}(\lambda)$. We also define

$$\hat{\pi}_f(\lambda) = \arg\max_{\pi_f} \hat{V}_f^{\pi_f}(\lambda), \quad \hat{V}_f^*(\lambda) = \hat{V}_f^{\hat{\pi}_f(\lambda)}(\lambda). \tag{6.42}$$

Note that $\hat{\pi}_f(\lambda)$ and $\hat{V}_f^*(\lambda)$ can be computed by standard finite horizon dynamic programming [34], and we omit the details.

We also define the following quantities for describing the GMBL algorithm succinctly.

$$\hat{C}^{\pi_f}_{l,f} = \hat{\mathbb{E}}[\sum_{\tau=0}^{\tau_f} \mathbf{1}\{a^{\pi_f}_{i,f}(t+\tau) = l\} | i, f, s^{\pi_f}_{i,f}(t) = s_f], \tag{6.43}$$

$$\hat{R}^{\pi_f}_f = \mathbb{E}[\sum_{\tau=0}^{\tau_f} r_f(s^{\pi_f}_{i,f}(t+\tau) | i, f, s^{\pi_f}_{i,f}(t) = s_f], \tag{6.44}$$

$$\hat{C}^{\pi}_l = \sum_f \rho_f \hat{C}^{\pi_f}_{l,f}, \;\; \hat{R}^{\pi} = \sum_f \rho_f \hat{R}^{\pi_f}_f, \text{ for } \pi = (\pi_f)_{f \in \mathcal{F}}, \tag{6.45}$$

Here π is the joint policy given by the collection of each individual policy π_f. Note that from (6.13), (6.43), (6.44), we can write

$$\hat{V}^{\pi_f}_f(\lambda) = \hat{R}^{\pi_f}_f - \sum_l \lambda_l \hat{C}^{\pi_f}_{l,f}, \tag{6.46}$$

$$\hat{L}(\pi, \lambda) = \hat{R}^{\pi} + \sum_l \lambda_l (C_l - \hat{C}^{\pi}_l).$$

The GMBL-Dual algorithm is summarized in Algorithm 9.

Algorithm 9 Generative Model-Based Learning-Dual (GMBL-Dual)

```
Input: accuracy $\epsilon, \delta$. Initialize $\lambda_l(0) = 0, \forall l \in \mathcal{L}$
Send $n = n(\epsilon, \delta)$ packets in each link $l \in \mathcal{L}$
Estimate the link probability $\hat{p}_l$ by transmitting $n$ packets across all links uniformly
Construct the approximate transition kernel $\hat{P}$
for $m$ from 1 to $M$ do
    For each flow $f$, compute $\pi_f(m) = \hat{\pi}_f(\lambda(m))$ according to (6.42). Define $\pi(m) = (\pi_f(m))_{f \in \mathcal{F}}$
    Compute $\hat{C}^{\pi(m)}_l$ according to (6.43) and (6.45)
    Compute $\lambda_l(m+1)$ for each link $l$ as
        $$\lambda_l(m+1) = \Pi_{\Lambda}(\lambda_l(m) - \alpha(C_l - \hat{C}^{\pi(m)}_l))^{\S}$$
Compute $\hat{\lambda}(M) = \frac{1}{M} \sum_{m=1}^{M} \lambda(m)$.
Compute $\pi_f(M+1) = \hat{\pi}_f(\hat{\lambda}(M))$
Output: $\hat{\pi} = (\pi_f(M+1))_{f \in F}, \;\; \hat{\lambda} = \hat{\lambda}(M)$
```

We next present the sample complexity of GMBL-Dual.

Theorem 14. *GMBL-Dual algorithm with*

$$n(\epsilon, \delta) \geq \frac{162 S \rho_{tot}^2 \tau_{\max}^3 (\beta_{\max} + |\mathcal{L}|\lambda_{\max})^2 \log \frac{72|F|S^2|\mathcal{L}|\tau_{\max}}{\delta}}{\epsilon^2} \tag{6.47}$$

and parameters

$$\begin{aligned} M &= \frac{36|\mathcal{L}|(\tau_{\max}\rho_{tot} + C_{\max})^2 \lambda_{\max}^2}{\epsilon^2}, \\ \alpha &= \frac{\epsilon}{3|\mathcal{L}|(\tau_{\max}\rho_{tot} + C_{\max})^2}, \end{aligned} \tag{6.48}$$

where $\lambda_{\max} = \frac{\rho_{tot}\beta_{\max}}{C_{\min}}$ *and* $C_{\min} = \min_l C_l$*, achieves a* $\hat{\lambda}$ *and* $\hat{\pi}$ *such that*

$$\mathbb{P}\left(|L(\hat{\pi}, \hat{\lambda}) - L(\pi^*, \lambda^*)| \leq \epsilon\right) \geq (1 - \delta).$$

The proof requires the use of multiple smaller results that we first present below, followed by their integration to yield the proof of the main theorem.

Lemma 36. *Let* $\lambda_{\max} = \max_l \lambda_l^*$*. Then,* $\lambda_{\max} < \frac{\rho_{tot}\beta_{\max}}{C_{\min}}$*.*

Proof. Considering **OSP** of (6.3) and reward matrix of each flow f (6.5), it is obvious that optimal policy for **OSP** would be transmitting packets rather than no-transmission, so that **OSP** would yield positive result. Therefore, it means that each policy $\pi_f(\lambda^*)$ is consisted of transmitting packets. Comparing value of transmission policy with no-transmission and knowing the fact that no-transmission policy yield 0 reward for each flow, we conclude that $V_f^*(\lambda^*)$ must be positive for each flow f, otherwise no-transmission policy must have been chosen.

On the other hand, optimal policy of the network might utilize some of the links, not all the links. Due to Complementary Slackness, [70], if link l is not used, then λ_l^* would be 0. And, chosen links would be utilized such that average utilization would be equal to their constraint, i.e. $\sum_f \rho_f C_{f,l,0}^{\pi^*}(s_f) = C_l$. Hence, we continue with those links that their λ_l^* would be positive.

For any flow f, we know that

$$0 < V_{f,0}^{\pi^*}(\lambda^*) \leq \beta_f - \sum_{l \in q_f} \lambda_l^* \sigma_{l,f},$$

where q_f is the path from s_f to d_f and $\sigma_{l,f}$ is expected number of times that the packet of flow f is being transmitted over link l. Therefore,

$$0 < \beta_f - \sum_{l \in q_f} \lambda_l^* \sigma_{l,f}.$$

Multiplying above term by ρ_f and summing both sides over f, we get

$$0 < \sum_f \rho_f \beta_f - \sum_f \sum_{l \in q_f} \lambda_l^* \rho_f \sigma_{l,f}.$$

On the other hand, by rearranging summations we get

$$0 < \sum_f \beta_f - \sum_f \sum_{l \in q_f} \lambda_l^* \sigma_{l,f} \leq \rho_{\text{tot}} \beta_{\max} - \sum_{l \in G^*} \lambda_l^* (\sum_f \rho_f \sigma_{l,f}),$$

where G^* is subgraph of the network with all active links.

Further, we can write $\rho_f \sigma_{l,f}$ using $C_{f,l,0}^{\pi^*}$, where

$$C_{f,l,0}^{\pi^*}(s_f) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} c_{f,l}^{\pi^*}(t).$$

So

$$\begin{aligned} 0 < \sum_f \beta_f - & \sum_f \sum_{l \in q_f} \lambda_l^* \rho_f \sigma_{l,f} \\ & \leq \rho_{\text{tot}} \beta_{\max} - \sum_{l \in G^*} \lambda_l^* (\sum_f \rho_f C_{f,l,0}^{\pi^*}(s_f)). \end{aligned}$$

Next, ignoring the middle term we get

$$\sum_{l \in G^*} \lambda_l^* (\sum_f \rho_f C_{f,l,0}^{\pi^*}(s_f)) < \rho_{\text{tot}} \beta_{\max}.$$

Since all λ_l^* and $C_{f,l,0}^{\pi^*}(s_f)$ are positive, we pick the $\lambda_{\max}$ and the link corresponding to it and get

$$\lambda_{\max} C_{l(\max)}^{\pi^*} < \rho_{\text{tot}} \beta_{\max}.$$

Now, we find the lower bound for the left-hand-side. We know that $C_{l(\max)}^{\pi^*} \geq C_{\min}$ according to Complementary Slackness. Hence,

$$C_{\min} \lambda_{\max} < \rho_{\text{tot}} \beta_{\max}.$$

Therefor, proof completes. □

Lemma 37. *With the parameters M and α given by* (6.48)*, we obtain $|\hat{D}(\hat{\lambda}) - \hat{D}(\hat{\lambda}^*)| \leq \epsilon/3$*

This follows from the standard rate of convergence analysis of projected subgradient descent algorithm for convex functions and proposition 36. For completeness we first reproduce that result. We use the following result.

Theorem 15. *[35] Let $g : \mathcal{X} \to \mathbb{R}^d$ be a convex function with $||\nabla g|| \leq B_1$. Also, assume that the domain of $g(\cdot)$ is bounded, i.e. $||x|| \leq B_2, \forall x \in \mathcal{X}$. Consider the projected gradient descent algorithm $x_{m+1} = \Pi_{\mathcal{X}}[x_m - \alpha \nabla g(x_m)]$ where $\Pi_{\mathcal{X}}$ is the projection operator. Then, with $\alpha = B_2/(B_1\sqrt{M})$,*

$$g(\frac{1}{M}\sum_{m=1}^{M} x_m) - g(x^*) \leq \frac{B_1 B_2}{\sqrt{M}}$$

Proof of Lemma 37: We first show that the subgradient of $\widehat{D}(\cdot)$ at λ, denoted by $\nabla \widehat{D}(\lambda)$ is given by

$$\nabla \widehat{D}(\lambda) = (C_l - \sum_f \rho_f \widehat{C}^{\widehat{\pi}(\lambda)}_{f,l,0}(s_f))_{l \in \mathcal{L}}.$$

Indeed, for any given λ', λ,

$$\begin{aligned}
\widehat{D}(\lambda') &= \max_{\pi} \hat{L}(\pi, \lambda') \ge \widehat{L}(\widehat{\pi}(\lambda), \lambda') \\
&= \widehat{R}^{\widehat{\pi}(\lambda)} + \sum_l \lambda'_l (C_l - \sum_f \rho_f \widehat{C}^{\widehat{\pi}(\lambda)}_{f,l,0}(s_f)) \\
&= \widehat{R}^{\widehat{\pi}(\lambda)} + \sum_l \lambda_l (C_l - \sum_f \rho_f \widehat{C}^{\widehat{\pi}(\lambda)}_{f,l,0}(s_f)) \\
&+ \sum_l (\lambda'_l - \lambda_l)(C_l - \sum_f \rho_f \widehat{C}^{\widehat{\pi}(\lambda)}_{f,l,0}(s_f)) \\
&= \widehat{D}(\lambda) + \sum_l (\lambda'_l - \lambda_l)(C_l - \sum_f \rho_f \widehat{C}^{\widehat{\pi}(\lambda)}_{f,l,0}(s_f))
\end{aligned}$$

and hence the claim follows by the definition of subgradient.

In order to bound $\|\nabla \hat{D}(\lambda)\|$, first note that from (6.43) $\widehat{C}^{\widehat{\pi}_f(\lambda)}_{f,l,0}(s_f) \le \tau_{max}$. So, $\sum_f \rho_f \widehat{C}^{\pi(\lambda)}_{f,l,0}(s_f) \le \tau_{\max}\rho_{\text{tot}}$. Also, $C_l \le C_{\max}$. Hence, $\|\nabla D(\lambda)\| = \sqrt{\mathcal{L}}(\tau_{\max}\rho_{\text{tot}} + C_{\max})$.

Now, considering Lemma 36, we project $\lambda(m)$ to set $[0, 2\lambda_{\max}]$. Using Theorem 15, we get the desired result. □

Lemma 38. *Let* $\delta_P \in (0,1)$. *Then, if* $n \ge 11664 S^2 \tau_f^2 \log 4/{\delta_P}^3$, *for any flow* f *and for a given* $\lambda \in [0, \lambda_{\max}]$ *under any policy* π

$$\|V^{\pi}_{f,0}(\lambda) - \widehat{V}^{\pi}_{f,0}(\lambda)\|_\infty \le \sqrt{18 \frac{S\tau_f^3(\beta_f + |\mathcal{L}|\lambda_{\max})^2 \log 4/\delta_P}{n}}$$

w.p. at least $1 - 3S^2|\mathcal{L}|\tau_f\delta_P$.

Proof. The proof procedure is identical to proof of Lemma 28 with adjustment of $\|V^{\pi}_{f,0}(\lambda) - \widehat{V}^{\pi}_{f,0}(\lambda)\|_\infty \le \beta_f + |\mathcal{L}|\lambda_{\max}$. □

Lemma 39. *Let* $\delta_P \in (0,1)$. *Then, if* $n \geq 11664 S^2 \tau_f^2 \log 4/\delta_P{}^3$, *for a given* $\lambda \in [0, \lambda_{\max}]$

$$|\widehat{D}(\lambda) - D(\lambda)| \leq \rho_{tot}\sqrt{18\frac{S\tau_{\max}^3(\beta_{\max} + |\mathcal{L}|\lambda_{\max})^2 \log 4/\delta_P}{n}}$$

w.p. at least $1 - 6|F|S^2|\mathcal{L}|\tau_{\max}\delta_P$.

Proof. For a given λ, consider two policies $\pi(\lambda)$ and $\widehat{\pi}(\lambda)$. Then for any flow f, according to Lemma 38 we have

$$V_{f,0}^{\pi_f(\lambda)}(s_f, \lambda) \leq \widehat{V}_{f,0}^{\pi_f(\lambda)}(s_f, \lambda) + \epsilon' \leq \widehat{V}_{f,0}^{\widehat{\pi}_f(\lambda)}(s_f, \lambda) + \epsilon' \tag{6.49}$$

w.p. at least $1 - 3S^2|\mathcal{L}|\tau_f\delta_P$ where $\epsilon' = \sqrt{18\frac{S\tau_f^3(\beta_f + |\mathcal{L}|\lambda_{\max})^2 \log 4/\delta_P}{n}}$. Please notice that the second inequality is due to the fact $\widehat{\pi}_f(\lambda) = \arg\max_\pi \widehat{V}_{f,0}^{\pi}(\lambda)$. Next, we have

$$\widehat{V}_{f,0}^{\widehat{\pi}_f(\lambda)}(s_f, \lambda) \leq V_{f,0}^{\widehat{\pi}_f(\lambda)}(s_f, \lambda) + \epsilon' \leq V_{f,0}^{\pi_f(\lambda)}(s_f, \lambda) + \epsilon' \tag{6.50}$$

w.p. at least $1 - 3S^2|\mathcal{L}|\tau_f\delta_P$. Now, combining the two inequalities (6.49) and (6.50), we get

$$|\widehat{V}_{f,0}^{\widehat{\pi}_f(\lambda)}(s_f, \lambda) - V_{f,0}^{\pi_f(\lambda)}(s_f, \lambda)| \leq \epsilon'$$

w.p. at least $1 - 6S^2|\mathcal{L}|\tau_f\delta_P$. Using the above inequality, we get

$$\begin{aligned}|\widehat{D}(\lambda) - D(\lambda)| &= |\sum_f \rho_f(\widehat{V}_{f,0}^{\widehat{\pi}_f(\lambda)}(s_f) - V_{f,0}^{\pi_f(\lambda)}(s_f))| \\ &\leq \sum_f \rho_f|\widehat{V}_{f,0}^{\widehat{\pi}_f(\lambda)}(s_f) - V_{f,0}^{\pi_f(\lambda)}(s_f)| \leq \rho_{\text{tot}}\epsilon'\end{aligned}$$

w.p. at least $1 - 6|F|S^2|\mathcal{L}|\tau_{\max}\delta_P$. Hence the proof is complete. □

Lemma 40. *Let* $\delta_P \in (0,1)$*. Then, if* $n \geq 11664 S^2 \tau_f^2 \log 4/{\delta_P}^3$*, for a given* $\lambda \in [0, \lambda_{\max}]$

$$|\widehat{D}(\widehat{\lambda}^*) - D(\lambda^*)| \leq \rho_{tot}\sqrt{18\frac{S\tau_f^3(\beta_f + |\mathcal{L}|\lambda_{\max})^2 \log 4/\delta_P}{n}}$$

w.p. at least $1 - 12|F|S^2|\mathcal{L}|\tau_{\max}\delta_P$.

Proof. Since $\widehat{\lambda}^* = \arg\min_\lambda \widehat{D}(\lambda)$, then $\widehat{D}(\widehat{\lambda}^*) \leq \widehat{D}(\lambda^*)$. Next, we have

$$\widehat{D}(\lambda^*) \leq D(\lambda^*) + \epsilon' \tag{6.51}$$

w.p. at least $1 - 6|F|S^2|\mathcal{L}|\tau_{\max}\delta_P$ where $\epsilon' = \rho_{\text{tot}}\sqrt{18\frac{S\tau_{\max}^3(\beta_{\max}+|\mathcal{L}|\lambda_{\max})^2 \log 4/\delta_P}{n}}$ according to Lemma 39. Therefore, we have

$$\widehat{D}(\widehat{\lambda}^*) \leq D(\lambda^*) + \epsilon' \tag{6.52}$$

w.p. at least $1 - 6|F|S^2|\mathcal{L}|\tau_{\max}\delta_P$. Taking identical steps, we get

$$D(\lambda^*) \leq \widehat{D}(\widehat{\lambda}^*) + \epsilon' \tag{6.53}$$

w.p. at least $1 - 6|F|S^2|\mathcal{L}|\tau_{\max}\delta_P$. Finally, combining the inequalities (6.52) and (6.53) yields the result. □

Now, we are ready to prove Theorem 14.

Proof of Theorem 14: We expand the result of theorem 14

$$
\begin{aligned}
&|L(\widehat{\pi},\widehat{\lambda}) - L(\pi^*,\lambda^*)| \\
&= |\sum_l \widehat{\lambda}_l C_l + \sum_f \rho_f V_{f,0}^{\widehat{\pi}_f}(s_f,\widehat{\lambda}) - D(\lambda^*)| \\
&= |\sum_l \widehat{\lambda}_l C_l + \sum_f \rho_f \widehat{V}_{f,0}^{\widehat{\pi}_f}(s_f,\widehat{\lambda}) - D(\lambda^*) \\
&+ \sum_f \rho_f (V_{f,0}^{\widehat{\pi}_f}(s_f,\widehat{\lambda}) - \widehat{V}_{f,0}^{\widehat{\pi}_f}(s_f,\widehat{\lambda}))| \\
&\le |\widehat{D}(\widehat{\lambda}) - D(\lambda^*)| + \sum_f \rho_f |V_{f,0}^{\widehat{\pi}_f}(s_f,\widehat{\lambda}) - \widehat{V}_{f,0}^{\widehat{\pi}_f}(s_f,\widehat{\lambda})|.
\end{aligned}
$$

First, we bound $|\widehat{D}(\widehat{\lambda}) - D(\lambda^*)|$ by and expanding it further

$$
\begin{aligned}
&|\widehat{D}(\widehat{\lambda}) - D(\lambda^*)| = |\widehat{D}(\widehat{\lambda}) - \widehat{D}(\widehat{\lambda}^*) + \widehat{D}(\widehat{\lambda}^*) - D(\lambda^*)| \\
&\le |\widehat{D}(\widehat{\lambda}) - \widehat{D}(\widehat{\lambda}^*)| + |\widehat{D}(\widehat{\lambda}^*) - D(\lambda^*)| \le \frac{\epsilon}{3} + \epsilon' \qquad (6.54)
\end{aligned}
$$

w.p. at least $1 - 12|F|S^2|\mathcal{L}|\tau_{\max}\delta_P$ where $\epsilon' = \rho_{\text{tot}}\sqrt{18\frac{S\tau_{\max}^3(\beta_{\max}+|\mathcal{L}|\lambda_{\max})^2\log 4/\delta_P}{n}}$ according to Lemmas 37 and 40.

Next,

$$
\sum_f \rho_f |V_{f,0}^{\widehat{\pi}_f}(s_f,\widehat{\lambda}) - \widehat{V}_{f,0}^{\widehat{\pi}_f}(s_f,\widehat{\lambda})| \le \epsilon' \qquad (6.55)
$$

w.p. at least $1 - 6|F|S^2|\mathcal{L}|\tau_{\max}\delta_P$ according to Lemma 38.

Eventually, we combine two inequalities (6.54) and (6.55) and get

$$
\begin{aligned}
&|L(\widehat{\pi},\widehat{\lambda}) - L(\pi^*,\lambda^*)| \\
&\le \frac{\epsilon}{3} + 2\rho_{\text{tot}}\sqrt{18\frac{S\tau_{\max}^3(\beta_{\max}+|\mathcal{L}|\lambda_{\max})^2\log 4/\delta_P}{n}}
\end{aligned}
$$

w.p. at least $1 - 18|F|S^2|\mathcal{L}|\tau_{\max}\delta_P$. Hence, putting $\epsilon = 3\rho_{\text{tot}}\sqrt{18\frac{S\tau_{\max}^3(\beta_{\max}+|\mathcal{L}|\lambda_{\max})^2\log 4/\delta_P}{n}}$ and

$\delta = 18|F||S^2||\mathcal{L}|\tau_{\max}\delta_P$ completes the proof. □

6.5 Simulation Results

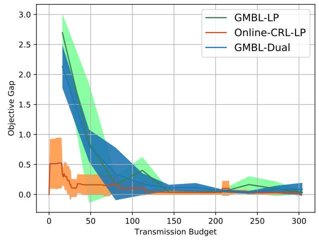

Figure 6.1: Objective Error-10-Node Network

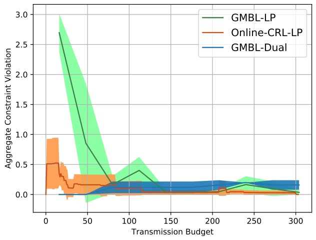

Figure 6.2: Constraint Violation-10-Node Network

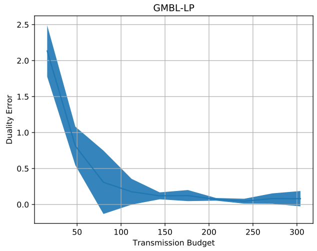

Figure 6.3: Duality Gap Error-10-Node Network

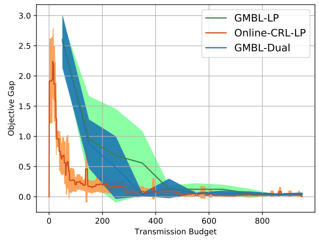

Figure 6.4: Objective Error-20-Node Network

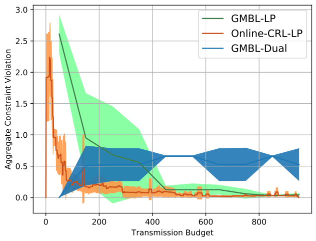

Figure 6.5: Constraint Violation-20-Node Network

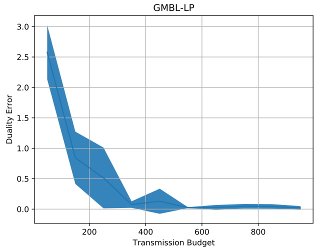

Figure 6.6: Duality Gap Error-20-Node Network

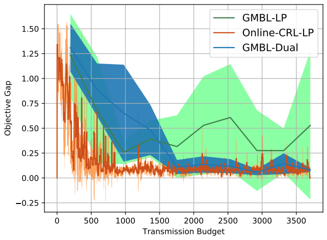

Figure 6.7: Objective Error-40-Node Network

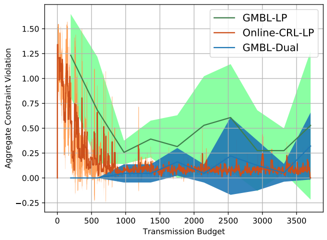

Figure 6.8: Constraint Violation-40-Node Network

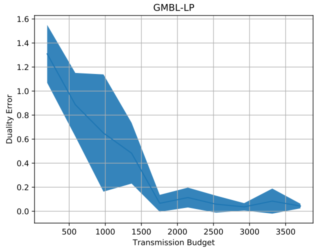

Figure 6.9: Duality Gap Error-40-Node Network

In this section, we present simulation results to compare the performance of the GMBL-LP, Online-CRL-LP and GMBL-Dual algorithms with respect to the optimal policy in the context of attaining high weighted timely throughput in an IAB network. We develop three simulation scenarios that are motivated by IAB node deployment features such as density of nodes and mm-wave communication [36]. The three scenarios represent different node densities. We consider networks with $10, 20$ and 40 nodes with $20, 80$ and 320 number of links respectively. In every scenario, for each link l, p_l is uniformly randomly chosen from $[0.5, 1.0]$, while C_l is chosen from $[1, 5]$. We have two unicast flows in all scenarios. Packet arrivals to the system follow a Poisson number of arrivals to each source node in each time slot. In all scenarios, both flows have same weight of 2.

The performance metrics of interest are the error in the objective function which is the weighted throughput engendered by the policy that is the outcome of the algorithm and aggregation of capacity constraint violations. We also consider the error corresponding to duality gap only for GMBL-Dual algorithm. We define this error as

$$\left| \sum_{l} \lambda_l^M C_l + \sum_{f} V_f^{\pi_f(\lambda_M)}(\lambda_M) - D(\lambda^*) \right|, \tag{6.56}$$

where λ_M and $\pi_f(\lambda_M)$ are the Lagrange multipliers and policy that result from the execution of GMBL-Dual algorithm. The error depends on the number of sub-gradient updates, M, which we empirically set as 100 for good error performance.

Our graphs relate to sample complexity, reward, constraint violation and duality gap. We set a packet budget for learning the model, and identify the error for each of our candidate algorithms. Figures 6.1, 6.4 and 6.7 depict the relation between the error and transmission budget empirically in the three scenarios. The graphs show that increasing the transmit budget reduces error for all the algorithms which is inconsistent with Theorems 12, 13 and 14. However, Online-CRL-LP outperforms both GMBL-LP and GMBL-Dual algorithms. This observation implicates that Online-CRL-LP mostly concentrates on effective links, while the other two algorithms do not distinguish between links. Here, we also observe that GMBL-LP and GMBL-Dual algorithms perform similarly in terms of obtaining true objective. However, the figures show that when the density of the network increases, GMBL-LP algorithm experiences more deviations compared to other two algorithms.

Next, we compare the total amount of capacity constraint violations of output of three algorithms in figures 6.2, 6.5 and 6.8. These figures illustrate Online-CRL-LP also outperforms the other two algorithms in terms of constraint violation. We also observe that while Online-CRL-LP and GMBL-LP algorithms show a decrease in constraint violation, GMBL-Dual algorithm incur an increase in constraint violation. This result happens from implementing deterministic policy instead of stochastic policies. Because, we do not have access to obtain a stochastic policy for a given MDP. Resolving this issue needs developing a tool to construct a stochastic policy for a given MDP, or implementing two time-scale approximation which results in a totally different approach.

Further, we illustrate duality-gap error only for GMBL-Dual algorithm in figures 6.3, 6.6 and 6.9 because this result is only available for this algorithm. All the figures show that the duality gap error characterized according to equation (6.56) decreases as the transmission budget increases. These figures also implicate that even though the constraint violation does not decrease, the duality gap error decreases which is consistent with Theorem 14.

Finally, comparing the graphs of each performance metric under the three network scenarios indicate that a larger number of sources of randomness causes higher error. For example, the figures 6.1, 6.4 and 6.7 show that the error is higher for scenario 2 (which has more links) than scenario 1 for every transmit budget. This phenomenon is also seen in figures 6.2, 6.5 and 6.8, and 6.3, 6.6 and 6.9, which is consistent with Theorems 12, 13 and 14.

6.6 Conclusion

In this chapter, we considered the problem of maximizing the throughput of unicast flows with strict per-packet deadlines over a multi-hop wireless network, motivated by 5G IAB mm-wave networks. The problem formulation took the form on a CMDP, and, assuming that the link statistics are unknown, can be solved using LP and dual-decomposition approaches. We proposed a model-based RL approaches, and developed three of algorithms, based on offline channel sounding.

7. SUMMARY AND DIRECTIONS OF FUTURE RESEARCH

In this thesis, we study learning in the face of constraints as applied to networked systems. In the first three chapters, we introduced two notions of sample complexity for understanding the performance of RL algorithms for safety-constrained applications. We developed two categories of algorithms—offline GMBL and online algorithms in first three chapters. The main findings of chapters 2 and 3 points to a logarithmic factor increase in sample complexity over the unconstrained regime suggesting the value of the approach to real systems. However, these algorithms are based on LP approach which are computationally expensive. Therefore, in chapter 4, we concentrate on reducing the computational complexity of the constrained-RL algorithms by considering dual approach.

In the next chapter, we studied the problem of broadcasting real-time flows with hard per-packet deadlines in a multi-hop wireless network as a prospective application for our learning algorithms. This problem is computationally complex due to the need to solve an MDP over the network graph. We relax the problem using average link utilization constraints, and come up with a novel decomposition approach that enables its solution in a distributed fashion. We propose the DSR algorithm that maximizes the total timely-utility. The algorithm has a low complexity, and has a really low coordination overhead. We also develop a simple *index* policy based on DSR that is able to meet hard link utilization constraints. We simulate the variants of the algorithm, comparing against several recent throughout optimal algorithms . In all cases, DSR and the index policy have a better performance in terms of total timely-utility. We conclude that throughput and delay optimality are fundamentally different, but simple near-optimal solutions are possible in the delay-constrained case.

In final chapter, we considered the problem of maximizing the throughput of unicast flows with strict per-packet deadlines over a multi-hop wireless network, motivated by 5G IAB mm-wave networks. The problem formulation took the form on a CMDP, and, assuming that the link statistics are unknown, can be solved using LP and dual-decomposition approaches. We proposed a model-based RL approaches, and developed three of algorithms, based on offline channel sounding. We introduced two notions of sample complexity. First notion is defined for objective maximization and constraint satisfaction individually for understanding the performance of RL algorithms for networks. With this definition, we developed two types of algorithms—GMBL-LP and online-CRL-LP. Second notion of sample complexity integrates objective maximization and constraint violations together. We designed algorithm GMBL-Dual based on this definition. Finally, we compared the performance of the three algorithms and showed that they have almost similar performance.

REFERENCES

[1] A. HasanzadeZonuzy, A. Bura, D. Kalathil, and S. Shakkottai, “Learning with safety constraints: Sample complexity of reinforcement learning for constrained mdps,” *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[2] A. HasanzadeZonuzy, D. Kalathil, and S. Shakkottai, “Model-based reinforcement learning for infinite-horizon discounted constrained markov decision processes,” *30th International Joint Conference on Artificial Intelligence*, 2021.

[3] A. HasanzadeZonuzy, I.-H. Hou, and S. Shakkottai, “Broadcasting real-time flows in integrated backhaul and access 5g networks,” in *2019 International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT)*, pp. 1–8, 2019.

[4] E. Altman, *Constrained Markov decision processes*, vol. 7. CRC Press, 1999.

[5] Y. Efroni, S. Mannor, and M. Pirotta, “Exploration-exploitation in constrained mdps,” *arXiv preprint arXiv:2003.02189*, 2020.

[6] M. G. Azar, R. Munos, and H. J. Kappen, “Minimax pac bounds on the sample complexity of reinforcement learning with a generative model,” *Machine learning*, vol. 91, no. 3, pp. 325–349, 2013.

[7] C. Dann and E. Brunskill, “Sample complexity of episodic fixed-horizon reinforcement learning,” in *Advances in Neural Information Processing Systems*, pp. 2818–2826, 2015.

[8] E. Altman, “Applications of Markov decision processes in communication networks,” in *Handbook of Markov decision processes*, pp. 489–536, Springer, 2002.

[9] V. S. Borkar, “An actor-critic algorithm for constrained Markov decision processes,” *Systems & control letters*, vol. 54, no. 3, pp. 207–213, 2005.

[10] V. Borkar and R. Jain, “Risk-constrained Markov decision processes,” *IEEE Transactions on Automatic Control*, vol. 59, no. 9, pp. 2574–2579, 2014.

[11] R. Singh and P. Kumar, "Throughput optimal decentralized scheduling of multihop networks with end-to-end deadline constraints: Unreliable links," *IEEE Transactions on Automatic Control*, vol. 64, no. 1, pp. 127–142, 2018.

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

[13] J. D. Isom, S. P. Meyn, and R. D. Braatz, "Piecewise linear dynamic programming for constrained pomdps.," in *AAAI*, vol. 1, pp. 291–296, 2008.

[14] D. Kim, J. Lee, K.-E. Kim, and P. Poupart, "Point-based value iteration for constrained pomdps," in *IJCAI*, pp. 1968–1974, 2011.

[15] S. Bhatnagar and K. Lakshmanan, "An online actor–critic algorithm with function approximation for constrained Markov decision processes," *Journal of Optimization Theory and Applications*, vol. 153, no. 3, pp. 688–708, 2012.

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

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

[18] S. Paternain, L. Chamon, M. Calvo-Fullana, and A. Ribeiro, "Constrained reinforcement learning has zero duality gap," in *Advances in Neural Information Processing Systems*, pp. 7553–7563, 2019.

[19] Y. Liu, J. Ding, and X. Liu, "Ipo: Interior-point policy optimization under constraints," *arXiv preprint arXiv:1910.09615*, 2019.

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

[21] A. Badanidiyuru, R. Kleinberg, and A. Slivkins, “Bandits with knapsacks,” in *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 207–216, IEEE, 2013.

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

[23] S. Amani, M. Alizadeh, and C. Thrampoulidis, “Linear stochastic bandits under safety constraints,” in *Advances in Neural Information Processing Systems*, pp. 9252–9262, 2019.

[24] L. Zheng and L. J. Ratliff, “Constrained upper confidence reinforcement learning,” *arXiv preprint arXiv:2001.09377*, 2020.

[25] A. Wachi and Y. Sui, “Safe reinforcement learning in constrained markov decision processes,” *arXiv preprint arXiv:2008.06626*, 2020.

[26] H. Satija, P. Amortila, and J. Pineau, “Constrained markov decision processes via backward value functions,” *arXiv preprint arXiv:2008.11811*, 2020.

[27] C. Dann, T. Lattimore, and E. Brunskill, “Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning,” in *Advances in Neural Information Processing Systems*, pp. 5713–5723, 2017.

[28] A. L. Strehl and M. L. Littman, “An analysis of model-based interval estimation for markov decision processes,” *Journal of Computer and System Sciences*, vol. 74, no. 8, pp. 1309–1331, 2008.

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

[30] W. Hoeffding, “Probability inequalities for sums of bounded random variables,” in *The Collected Works of Wassily Hoeffding*, pp. 409–426, Springer, 1994.

[31] A. Maurer and M. Pontil, “Empirical bernstein bounds and sample variance penalization,” *arXiv preprint arXiv:0907.3740*, 2009.

[32] T. Lattimore and M. Hutter, "Near-optimal pac bounds for discounted mdps," *Theoretical Computer Science*, vol. 558, pp. 125–143, 2014.

[33] A. HasanzadeZonuzy, A. Bura, D. Kalathil, and S. Shakkottai, "Learning with safety constraints: Sample complexity of reinforcement learning for constrained mdps," *arXiv preprint arXiv:2008.00311*, 2020.

[34] D. P. Bertsekas, D. P. Bertsekas, D. P. Bertsekas, and D. P. Bertsekas, *Dynamic programming and optimal control*, vol. 1. Athena scientific Belmont, MA, 1995.

[35] S. Bubeck *et al.*, "Convex optimization: Algorithms and complexity," *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.

[36] M. N. Islam, S. Subramanian, and A. Sampath, "Integrated access backhaul in millimeter wave networks," in *Wireless Communications and Networking Conference (WCNC), 2017 IEEE*, pp. 1–6, IEEE, 2017.

[37] M. N. Islam, N. Abedini, G. Hampel, S. Subramanian, and J. Li, "Investigation of performance in integrated access and backhaul networks," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 597–602, IEEE, 2018.

[38] C. Ho, K. Obraczka, G. Tsudik, and K. Viswanath, "Flooding for reliable multicast in multi-hop ad hoc networks," in *In Proceedings of the International Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications (DIALM)*, 1999.

[39] J. Jetcheva, Y. Hu, D. Maltz, and D. Johnson, "A simple protocol for multicast and broadcast in mobile ad hoc networks," in *In Internet Draft, draft-ietf-nanet-simple-mbcast-00.txt*, June 2001.

[40] S. Ni, Y. Tseng, Y. Chen, and J. Sheu, "The broadcast storm problem in a mobile ad hoc network," in *In Proceedings of the fifth annual ACM/IEEE international conference on Mobile computing and networking*, pp. 151 – 162, 1999.

[41] R. Gandhi, S. Parthasarathy, and A. Mishra, "Minimizing broadcast latency and redundancy in ad hoc networks," in *in Proceedings of 4th ACM international symposium on Mobile ad hoc networking and computing*, pp. 222 – 232, 2003.

[42] S. C. Huang, P.-J. Wan, X. Jia, H. Du, and W. Shang, "Minimum-latency broadcast scheduling in wireless ad hoc networks," in *Proc. of IEEE INFOCOM*, 2007.

[43] S. Sarkar and L. Tassiulas, "A framework for routing and congestion control for multicast information flows," *IEEE Transactions on Information Theory*, vol. 48, no. 10, pp. 2690 – 2708, 2002.

[44] T. Ho and H. Viswanathan, "Dynamic algorithms for multicast with intra-session network coding," in *In Proc. 43rd Annual Allerton Conference on Communication Control, and Computing*, 2005.

[45] J. Yuan, Z. Li, W. Yu, and B. Li, "A cross-layer optimization framework for multihop multicast in wireless mesh networks," *Selected Areas in Communications, IEEE Journal on*, vol. 24, no. 11, 2006.

[46] S. Zhang, M. Chen, Z. Li, and L. Huang, "Optimal distributed broadcasting with per-neighbor queues in acyclic overlay networks with arbitrary underlay capacity constraints," in *In Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pp. 814 – 818, 2013.

[47] A. Sinha, L. Tassiulas, and E. Modiano, "Throughput-optimal broadcast in wireless networks with dynamic topology," in *Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pp. 21 – 30, July 2016.

[48] A. Sinha, G. Paschos, and E. Modiano, "Throughput-optimal multi-hop broadcast algorithms," in *Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pp. 51 – 60, July 2016.

[49] H. Xiong, R. Li, A. Eryilmaz, and E. Ekici, “Delay-aware cross-layer design for network utility maximization in multi-hop networks,” *Selected Areas in Communications, IEEE Journal on*, vol. 29, pp. 951 – 959, 2011.

[50] Z. Mao, C. E. Koksal, and N. B. Shroff, “Online packet scheduling with hard deadlines in multihop communication networks,” in *Proc. of IEEE INFOCOM*, pp. 2463 – 2471, 2013.

[51] R. Li and A. Eryilmaz, “Scheduling end-to-end deadline-constrained traffic with reliability requirements in multi-hop networks,” *Selected Areas in Communications, IEEE Journal on*, vol. 20, pp. 1649 – 1662, Otc 2012.

[52] R. Singh and P. R. Kumar, “Optimizing quality of experience of dynamic video streaming over fading wireless networks,” in *In Proceedings of 54th IEEE Conference on Decision and Control*, Dec 2015.

[53] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[54] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear programming: theory and algorithms*. John Wiley & Sons, 2013.

[55] A. Sinha and E. Modiano, “Optimal control for generalized network-flow problems,” *IEEE/ACM Transactions on Networking (TON)*, vol. 26, no. 1, pp. 506–519, 2018.

[56] L. Tassiulas and A. Ephremides, “Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks,” in *29th IEEE Conference on Decision and Control*, pp. 2130–2132, IEEE, 1990.

[57] X. Lin and N. B. Shroff, “Joint rate control and scheduling in multihop wireless networks,” in *2004 43rd IEEE Conference on Decision and Control (CDC)(IEEE Cat. No. 04CH37601)*, vol. 2, pp. 1484–1489, IEEE, 2004.

[58] A. Eryilmaz and R. Srikant, “Joint congestion control, routing, and mac for stability and fairness in wireless networks,” *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1514–1524, 2006.

[59] A. Sinha, G. Paschos, and E. Modiano, "Throughput-optimal multi-hop broadcast algorithms," *IEEE/ACM Transactions on Networking*, 2017.

[60] I.-H. Hou, V. Borkar, and P. R. Kumar, "A theory of QoS for wireless," in *Proc. IEEE International Conference on Computer Communications (INFOCOM)*, (Rio de Janeiro, Brazil), April 2009.

[61] R. Li, A. Eryilmaz, and B. Li, "Throughput-optimal wireless scheduling with regulated inter-service times," in *2013 Proceedings IEEE INFOCOM*, pp. 2616–2624, IEEE, 2013.

[62] R. Singh and P. Kumar, "Throughput optimal decentralized scheduling of multihop networks with end-to-end deadline constraints: Unreliable links," *IEEE Transactions on Automatic Control*, vol. 64, no. 1, pp. 127–142, 2019.

[63] A. HasanzadeZonuzy, I.-H. Hou, and S. Shakkottai, "Broadcasting real-time flows in integrated backhaul and access 5G networks," in *WiOpt 2019*, 2019.

[64] S. Krishnasamy, R. Sen, R. Johari, and S. Shakkottai, "Regret of queueing bandits," in *Advances in Neural Information Processing Systems*, pp. 1669–1677, 2016.

[65] R. Combes, A. Proutiere, D. Yun, J. Ok, and Y. Yi, "Optimal rate sampling in 802.11 systems," in *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, pp. 2760–2767, IEEE, 2014.

[66] H. Gupta, A. Eryilmaz, and R. Srikant, "Low-complexity, low-regret link rate selection in rapidly-varying wireless channels," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pp. 540–548, IEEE, 2018.

[67] M. S. Talebi, Z. Zou, R. Combes, A. Proutiere, and M. Johansson, "Stochastic online shortest path routing: The value of feedback," *IEEE Transactions on Automatic Control*, vol. 63, no. 4, pp. 915–930, 2017.

[68] R. Singh and P. Kumar, "Optimizing quality of experience of dynamic video streaming over fading wireless networks," in *2015 54th IEEE Conference on Decision and Control (CDC)*, pp. 7195–7200, IEEE, 2015.

[69] F. Chung and L. Lu, “Concentration inequalities and martingale inequalities: a survey,” *Internet Mathematics*, vol. 3, no. 1, pp. 79–127, 2006.

[70] D. Bertsekas and A. Nedic, “Convex analysis and optimization (conservative),” 2003.

APPENDIX A

APPENDIX OF CHAPTER 2

A.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 (3.4) [4], then build the idea of ELP based on that. To solve CMDP problem (3.4) via LP approach, we convert this problem to a linear programming problem formulated using new variables occupation measures. Now, consider μ as the finite-horizon state-action occupation measure under policy π defined as

$$\mu(s,a,\pi,h) := \mathbb{P}(s_h = s, a_h = a | s_{h=0} = s_0), \tag{A.1}$$

where the probability is calculated w.r.t. underlying transition kernel under policy π, P_π. 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 μ be any generic occupation measure defined as (B.1), then the equivalent LP to CMDP problem (3.4) is

$$\begin{aligned}
&\max_{\mu} \sum_{s,a,h} \mu(s,a,h) r(s,a) \\
&\qquad \text{s.t.} \\
&\qquad \sum_{s,a,h} \mu(s,a,h) c(i,s,a) \leq \bar{C}_i \quad \forall i, \\
&\qquad \sum_{a} \mu(s,a,h) = \sum_{s',a'} P(s|s',a') \mu(s',a',h-1) \ \ \forall h \in \{1,\dots,H-1\}, \\
&\qquad \sum_{a} \mu(s_0,a,0) = 1, \ \ \sum_{a} \mu(s,a,0) = 0 \ \ \forall s \in S \backslash \{s_0\}, \\
&\qquad \mu(s,a,h) \geq 0 \ \ \forall s,a,h
\end{aligned} \tag{A.2}$$

It is proved that the LP (B.3) is equivalent to CMDP problem of (3.4), and the optimal policy computed by this LP is also the solution to CMDP problem in [4]. Eventually, the optimal policy π^* 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

$$\begin{aligned}
&\max_q \sum_{s,a,s',h} q(s,a,s',h) r(s,a)\\
&\quad \text{s.t.}\\
&\quad \sum_{s,a,s',h} q(s,a,s',h) c(i,s,a) \le \bar{C}_i \quad \forall i \in \{1,\ldots,N\},\\
&\quad \sum_{a,s'} q(s,a,s',h) = \sum_{s',a'} q(s',a',s,h-1) \;\; \forall h \in \{1,\ldots,H-1\},\\
&\quad \sum_{a,s'} q(s_0,a,s',0) = 1, \;\; \sum_{a,s'} q(s,a,s',0) = 0 \;\; \forall s \in S \backslash \{s_0\},\\
&\quad q(s,a,s',h) \ge 0 \;\; \forall s,s' \in S, a \in A, h \in \{0,1,\ldots,H-1\},\\
&\quad q(s,a,s',h) - (\widehat{P}(s'|s,a) + \beta(s,a,s')) \sum_y q(s,a,y,h) \le 0 \;\; \forall s,a,s',h,\\
&\quad -q(s,a,s',h) + (\widehat{P}(s'|s,a) - \beta(s,a,s')) \sum_y q(s,a,y,h) \le 0 \;\; \forall s,a,s',h,
\end{aligned}$$

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 [5], so we do not present them here.

A.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 6*:** Fix a state, action and next state, i.e. s, a, s'. Then, according to Hoeffding's inequality [30]

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

Now, we apply empirical Bernstein's inequality [31] 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 41. *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 42. *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 3. Then, under any policy* π

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

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

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

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 43. *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 3. 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 π

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

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 $(s, a) \in S \times A, i \in \{1, \ldots, N\}$ 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 $\bar{V}^\pi(s') = \sum_{s''} P'_\pi(s''|s)V_{h+1}'^\pi(s'')$ as the expected value function of the successor states of s. Note that $\bar{V}^\pi(s')$ is a constant function and so $\bar{V}^\pi(s') = \sum_{s''} P'_\pi(s''|s)\bar{V}^\pi(s'') = \sum_{s''} P_\pi(s''|s)\bar{V}^\pi(s'')$.

$$
\begin{aligned}
&|\sum_{s'}(P_\pi(s'|s) - P'_\pi(s'|s))V'^\pi_{h+1}(s')| = |\sum_{s'}(P_\pi(s'|s) - P'_\pi(s'|s))V'^\pi_{h+1}(s') + \bar{V}^\pi(s) - \bar{V}^\pi(s)| \\
&= |\sum_{s'}(P_\pi(s'|s) - P'_\pi(s'|s))(V'^\pi_{h+1}(s') - \bar{V}^\pi(s'))| \\
&\leq \sum_{s'} |P_\pi(s'|s) - P'_\pi(s'|s)||V^\pi_{h+1}(s') - \bar{V}^\pi(s')| && \text{(A.3)} \\
&\leq \sum_{s'}(c_1 + c_2\sqrt{P'_\pi(s'|s) - (1 - P'_\pi(s'|s))})|V^\pi_{h+1}(s') - \bar{V}^\pi(s')| \\
&\leq |S|c_1\|V'^\pi_{h+1}\|_\infty + c_2\sum_{s'}\sqrt{P'_\pi(s'|s)(1 - P'_\pi(s'|s))(V^\pi_{h+1}(s') - \bar{V}^\pi(s'))^2} \\
&\leq |S|c_1\|V'^\pi_{h+1}\|_\infty + c_2\sqrt{|S|\sum_{s'} P'_\pi(s'|s)(1 - P'_\pi(s'|s))(V^\pi_{h+1}(s') - \bar{V}^\pi(s'))^2} && \text{(A.4)} \\
&\leq |S|c_1\|V'^\pi_{h+1}\|_\infty + c_2\sqrt{|S|\sum_{s'} P'_\pi(s'|s)(V^\pi_{h+1}(s') - \bar{V}^\pi(s'))^2} \\
&= |S|c_1\|V'^\pi_{h+1}\|_\infty + c_2\sqrt{|S|}\sigma'^\pi_h.
\end{aligned}
$$

Inequality (B.7) 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 (B.8), 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 44. *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 3. 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* π

$$\|V_{H-1}^{\pi} - V_{H-1}'^{\pi}\|_\infty \leq \cdots \leq \|V_0^{\pi} - V_0'^{\pi}\|_\infty \leq |S|H^2 a\frac{1}{\sqrt{n}},$$

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

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

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 $\Delta_h = \max_s |V_h^\pi(s) - V_h'^\pi(s)|$. Then

$$\begin{aligned}
&\Delta_h = |V_h^\pi(s) - V_h'^\pi(s)| = |r_\pi(s) + \sum_{s'} P_\pi(s'|s)V_{h+1}^\pi(s') - (r_\pi(s) + \sum_{s'} P_\pi'(s'|s)V_{h+1}'^\pi(s'))| \\
&= |\sum_{s'} P_\pi(s'|s)V_{h+1}^\pi(s') - \sum_{s'} P_\pi'(s'|s)V_{h+1}^\pi(s') + \sum_{s'} P_\pi'(s'|s)V_{h+1}^\pi(s') - \sum_{s'} P_\pi'(s'|s)V_{h+1}'^\pi(s')| \\
&\leq \sum_{s'} |(P_\pi(s'|s) - P_\pi'(s'|s)|H + \Delta_{h+1} \\
&\leq |S|Ha\frac{1}{\sqrt{n}} + \Delta_{h+1}.
\end{aligned}$$

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

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

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 45. *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 3. 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 π

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

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

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

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,

$$\begin{aligned}
\sigma_h^{\pi^2}(s) &= \sigma_h^{\pi^2}(s) - \mathbb{E}'[(V_{h+1}^\pi(s_{h+1}) - P_\pi' V_{h+1}^\pi(s))^2] + \mathbb{E}'[(V_{h+1}^\pi(s_{h+1}) - P_\pi' V_{h+1}^\pi(s))^2] \\
&\leq \sum_{s'}(P_\pi(s'|s) - P_\pi'(s'|s))V_{h+1}^{\pi^2}(s') - [(\sum_{s'} P_\pi(s'|s)V_{h+1}^\pi(s'))^2 - (\sum_{s'} P_\pi'(s'|s)V_{h+1}^\pi(s'))^2] \\
&\quad + [\sqrt{\mathbb{E}'[(V_{h+1}^\pi(s_{h+1}) - V_1'^\pi(s_1) - P_\pi'(V_{h+1}^\pi - V_{h+1}'^\pi)(s))^2]} + \sqrt{\mathbb{E}'[(V_{h+1}'^\pi(s_{h+1}) - P_\pi'(V_{h+1}'^\pi)(s))^2]}]^2,
\end{aligned}$$

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

$\|V_h^\pi - V_h^{\prime\pi}\|_\infty^2$ implying

$$\begin{aligned}\sigma_h^{\pi^2}(s) \leq & \sum_{s'}(P_\pi(s'|s) - P'_\pi(s'|s))V_{h+1}^{\pi^2}(s') \\ & - [\sum_{s'}(P_\pi(s'|s) - P'_\pi(s'|s))V_{h+1}^\pi(s')][\sum_{s'}(P_\pi(s'|s) + P'_\pi(s'|s))V_{h+1}^\pi(s')] \\ & + (\|V_h^\pi - V_h^{\prime\pi}\|_\infty + \sigma_h^{\prime\pi}(s))^2\end{aligned}$$

w.p. at least $1 - |S|\delta_P$ Now, if we use Lemma 53, we get

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

w.p. at least $1 - |S|^2|A|H\delta_P$.[*] 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. $\sigma_h^{\prime\pi}(s) \leq \sigma_h^\pi(s) + \frac{2\sqrt{2|S|H^2a}}{n^{1/4}}$.[†] □

Lemma 46. *[7] The variance of the value function defined as* $\Sigma_t^\pi(s) = \mathbb{E}[(\sum_{h=t}^{H-1} r(s_h) - V_0^\pi(s))^2]$ *satisfies a Bellman equation* $\Sigma_t^\pi(s) = \sigma_t^{\pi^2}(s) + \sum_{s' \in S} P_\pi(s'|s)V_{t+1}^\pi(s')$ *which gives* $\Sigma_t^\pi(s) = \sum_{h=t}^{H}(P_\pi^{h-1}\sigma_h^{\pi^2})(s)$*. Since* $0 \leq \Sigma_0^\pi(s) \leq H^2$*, it follows that* $0 \leq \sum_{h=0}^{H-1}(P_\pi^{h-1}\sigma_h^{\pi^2})(s) \leq H^2$ *for all* $s \in S$*.*

Corollary 5. *The result of Lemma 63 also holds for variance of cost functions.*

Proof of Lemma 28: We only prove the statement of value function since the proof procedure for cost is identical. First, we apply Lemma 60 and get

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

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

[†] Here, we also know that the high probability bound on $|\sigma_h^\pi(s) - \sigma_h^{\prime\pi}(s)|$ is dependent over all (s, a).

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

$$c_1 = 2\sqrt{2}\Big(\frac{\log 4/\delta_P}{n}\Big)^{\frac{3}{4}} + \frac{3\sqrt{2}\log 4/\delta_P}{n} \text{ and } c_2 = \sqrt{\frac{8\log 4/\delta_P}{n}} \tag{A.5}$$

Now, let fix state s :

$$|V_0^\pi(s) - \tilde{V}_0^\pi(s)| = |\sum_{h=0}^{H-2}\tilde{P}_\pi^{h-1}(P_\pi - \tilde{P}_\pi)V_{h+1}^\pi|(s) \tag{A.6}$$

$$\le \sum_{h=0}^{H-2}\tilde{P}_\pi^{h-1}|(P_\pi - \tilde{P}_\pi)V_{h+1}^\pi|(s) \le \sum_{h=0}^{H-2}\tilde{P}_\pi^{h-1}(|S|c_1\|V_{h+1}^\pi\|_\infty + c_2\sqrt{|S|}\sigma_h^\pi)(s) \tag{A.7}$$

$$\le |S|H^2c_1 + c_2\sqrt{|S|}\sum_{h=0}^{H-1}(\tilde{P}_\pi^{h-1}\sigma_h^\pi)(s) \tag{A.8}$$

$$\le |S|H^2c_1 + c_2\sqrt{|S|}\sum_{h=0}^{H-1}(\tilde{P}^{h-1}(\tilde{\sigma}_h^\pi + \frac{2^{1.25}|S|^{0.5}H(\log 4/\delta_P)^{0.25}}{n^{1/4}})(s) \tag{A.9}$$

$$\le |S|H^2c_1 + c_2\sqrt{|S|H}\sqrt{\sum_{h=0}^{H-1}(\tilde{P}^{h-1}\tilde{\sigma}_h^{\pi^2})(s)} + c_2H\sqrt{|S|}\frac{2^{1.25}|S|^{0.5}H(\log 4/\delta_P)^{0.25}}{n^{1/4}} \tag{A.10}$$

$$= \frac{3\sqrt{2}|S|H^2\log 4/\delta_P}{n} + \frac{2\sqrt{2}|S|H^2(\log 4/\delta_P)^{\frac{3}{4}}}{n^{\frac{3}{4}}} + \sqrt{\frac{8|S|H^3\log 4/\delta_P}{n}} + \frac{2^{2.75}|S|H^2(\log 4/\delta_P)^{\frac{3}{4}}}{n^{\frac{3}{4}}} \tag{A.11}$$

$$\le \sqrt{128\frac{|S|H^3\log 4/\delta_P}{n}}. \tag{A.12}$$

In equation (B.16), we used Lemma 51. Then, we applied Lemma 52 to obtain inequality (B.17). Next, we bound $\|V_{h+1}^\pi\|_\infty$ by H in inequality (B.18). To get inequality (B.19), we use Lemma 54, 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 (B.20). To get inequality (B.21), we applied Lemma 63 and substituting c_1 and c_2 according to equations (B.15). Finally, inequality (A.12) follows from the fact that $n \ge 2592|S|^2H^2\log 4/\delta_P$. Since the result is true for every $s \in S$, hence the proof is complete. □

Proof of Theorem 12: Let $\delta_P \in (0,1)$. First, we know that optimistic planning problem (6.33) 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

$$V_0^{\pi^*}(s_0) - \sqrt{128\frac{|S|H^3 \log 4/\delta_P}{n}} \leq \tilde{V}_0^{\pi^*}(s_0) \leq V_0^{\pi^*}(s_0) + \sqrt{\frac{128|S|H^3 \log 4/\delta_P}{n}}$$

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

$$V_0^{\tilde{\pi}}(s_0) - \sqrt{\frac{128|S|H^3 \log 4/\delta_P}{n}} \leq \tilde{V}_0^{\tilde{\pi}}(s_0) \leq V_0^{\tilde{\pi}}(s_0) + \sqrt{\frac{128|S|H^3 \log 4/\delta_P}{n}}$$

w.p. at least $1-3|S|^2|A|H\delta_P$ according to Lemma 28. On the other hand, we know that $\tilde{V}_0^{\pi^*}(s_0) \leq \tilde{V}_0^{\tilde{\pi}}(s_0)$. Thus, by combining these results we get

$$V_0^{\pi^*}(s_0) - \sqrt{\frac{128|S|H^3 \log 4/\delta_P}{n}} \leq \tilde{V}_0^{\pi^*}(s) \leq \tilde{V}_0^{\tilde{\pi}}(s_0) \leq V_0^{\tilde{\pi}}(s) + \sqrt{\frac{128|S|H^3 \log 4/\delta_P}{n}}.$$

It yields that $V_0^{\tilde{\pi}}(s_0) \geq V_0^{\pi^*}(s_0) - 2\sqrt{\frac{128|S|H^3 \log 4/\delta_P}{n}}$ 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

$$C_{i,0}^{\tilde{\pi}}(s_0) \leq \tilde{C}_{i,0}^{\tilde{\pi}}(s_0) + \sqrt{\frac{128|S|H^3 \log 4/\delta_P}{n}} \leq \bar{C}_i + \sqrt{\frac{128|S|H^3 \log 4/\delta_P}{n}}$$

w.p. at least $1 - 3|S|^2|A|H\delta_P$ according to Lemma 28. 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 $\epsilon = 2\sqrt{\frac{128|S|H^3 \log 4/\delta_P}{n}}$ 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 28. □

A.3 Detailed Proof for Theorem 13

First, we bound total number of model updates in Algorithm 8.

Lemma 47. *The total number of updates under algorithm 8 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 30*:** At each episode with model update k and for each (s, a), by Hoeffding's inequality [30] 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 [31] 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 33. But, first we provide some useful lemmas.

Lemma 48. *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

$$mw_\iota\kappa \leq mw_k(s, a)\kappa \leq n_k(s, a) \leq 2mw_k(s, a)\kappa \leq 4mw_\iota\kappa. \tag{A.13}$$

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 49. *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 (κ, ι) 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 57, we have $E_{\kappa,\iota} > \alpha N$ if

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

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 $\nu_k^+ \in [0, H]$

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

Using

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

we can apply Theorem 22 of [69] and obtain

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

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

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

□

Proof of Lemma 33**:** Since $w_k(s,a) \leq H$, we have that $\frac{w_k(s,a)}{w_{min}} < \frac{H}{w_{min}}$ and so $\iota_k(s,a) \leq H/w_{min} = 4H^2|S|/\epsilon$. In addition, $|X_{k,\kappa,\iota}| \leq |S \times A|$ for all k, κ, ι and so $|X_{k,\kappa,\iota}| > \kappa$ can only be true for $\kappa \leq |S|$. Hence, only $E_{max} = \log_2 \frac{H}{w_{min}} \log_2 |S|$ possible values for (κ, ι) exists that can have $|X_{k,\kappa,\iota}| > \kappa$. By union bound over all (κ, ι) and lemma 58, we get

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

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

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

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 35.

Corollary 6. *If we substitute the δ_P with δ_1 in Lemma 60, the result will pertain.*

Corollary 7. *If we substitute the δ_P with δ_1 in Lemma 52, the result will pertain.*

Proof of Lemma 35: 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 $r^{(2d+2)} = \max_h \sigma_{h:H-1}^{(d),2}$, where $\sigma_{h:H-1}^{(d),2}$ 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 51, 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

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

$$\leq \sum_{s,a \notin X} \sum_{h=0}^{H-1} \|\tilde{V}_{h+1}^{(d)}\|_\infty (P^{h-1}\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0)$$

$$+ \sum_{s,a \in X} \sum_{h=0}^{H-1} |(P - \tilde{P})\tilde{V}_{h+1}^{(d)}(s)| (P^{h-1}\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0)$$

$$\leq \sum_{s,a \notin X} \sum_{h=0}^{H-1} H^{d+1} (P^{h-1}\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0)$$

$$+ \sum_{s,a \in X} \sum_{h=0}^{H-1} |(P - \tilde{P})\tilde{V}_{h+1}^{(d)}(s)| (P^{h-1}\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0)$$

$$\leq \sum_{s,a \notin X} \sum_{h=0}^{H-1} H^{d+1} (P^{h-1}\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0)$$

$$+ \sum_{s,a \in X} \sum_{h=0}^{H-1} ||S|c_1(s,a)H^{d+1} + c_2(s,a)\sqrt{|S|}\tilde{\sigma}_h^{(d)}(s,a)| (P^{h-1}\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0)$$

$$\leq \sum_{s,a \notin X} \sum_{h=0}^{H} H^{d+1} (P^{h-1}\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0)$$

$$+ \sum_{s,a \in X} \sum_{h=0}^{H} ||S|c_1(s,a)H^{d+1}| (P^{h-1}\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0)$$

$$+ \sum_{s,a \in X} \sum_{h=0}^{H-1} |\sqrt{|S|}c_2(s,a)\tilde{\sigma}_h^{(d)}(s,a)| (P^{h-1}\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0)$$

$$\leq \sum_{s,a \notin X} H^{d+1} w(s,a) + \sum_{s,a \in X} |S|c_1(s,a)H^{d+1}w(s,a)$$

$$+ \sum_{s,a \in X} \sqrt{|S|}c_2(s,a) \sum_{h=0}^{H-1} \tilde{\sigma}_h^{(d)}(s,a)(P^{h-1}\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0)$$

$$\leq w_{min}|S|H^{d+1} + \sum_{s,a \in X} |S|c_1(s,a)H^{d+1}w(s,a)$$

$$+ \sum_{s,a \in X} \sqrt{|S|}c_2(s,a) \sum_{h=0}^{H-1} \tilde{\sigma}_h^{(d)}(s,a)(P^{h-1}\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0)$$

$$= \frac{\epsilon}{4}H^d + \sum_{s,a \in X} |S|c_1(s,a)H^{d+1}w(s,a)$$

$$+ \sum_{s,a \in X} \sqrt{|S|}c_2(s,a) \sum_{h=0}^{H-1} \tilde{\sigma}_h^{(d)}(s,a)(P^{h-1}\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot, h)\})(s_0)$$

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 52 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 9 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

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

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 (B.22). 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 (κ,ι) pairs.

The next term is

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

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

The last term is

$$
\begin{aligned}
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} \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)} \\
&\times \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)} \\
&= \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)} \\
&\leq \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)} \\
&\leq \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)} \\
&\leq \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)} \\
&\leq \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)} \\
&\leq \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 $|X_{\kappa,\iota}| \leq \kappa$ 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 $\|\tilde{\sigma}_h^{(d)}\|_\infty \leq H^{2d+2}$. Thus, we have

$$Z_d(s_0) \leq \sqrt{\frac{8|S|H^{2d+3}\log 4/\delta_1|\mathcal{K}\times\mathcal{I}|}{m}}. \tag{A.14}$$

However, we can improve this bound as follows

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

In the third step, we used Lemma 63 and definition of $r^{(2d+2)}$.

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

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

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 (B.23) to bound $Z_\beta(s_0)$. Finally, the proof completes if we let

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

$\square$.

Proof of Theorem 13: By Lemma 33, we know that number of episodes where $|X_{\kappa,\iota}| > \kappa$ for some κ, ι 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 35 that for any $i \in \{1, \ldots, N\}$

$$|\tilde{V}_0^{\tilde{\pi}_k}(s_0) - V_0^{\tilde{\pi}_k}(s_0)| \leq \epsilon, \quad |\tilde{C}_{i,0}^{\tilde{\pi}_k}(s_0) - C_0^{\tilde{\pi}_k}(s_0)| \leq \epsilon. \tag{A.15}$$

Using Lemma 30, 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

$$\tilde{V}_0^{\tilde{\pi}_k}(s_0) \geq V_0^{\pi^*}(s_0), \quad \tilde{C}_{i,0}^{\tilde{\pi}_k}(s_0) \leq \bar{C}_i \;\; i \in \{1, \ldots, N\} \tag{A.16}$$

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

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

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

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

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

$$m = 2560\frac{|S|H^2}{\epsilon^2}(\log_2\log_2 H)^2\log_2^2\Big(\frac{8H^2|S|^2}{\epsilon}\Big) \qquad \text{(A.17)}$$
$$\times \log\Big(\frac{2048(N+1)|S|^4|A|H^2}{\epsilon^2\delta}(\log_2\log_2 H)^2\log_2^2\Big(\frac{8H^2|S|^2}{\epsilon}\Big)\Big).$$

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

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

sub-optimal episodes. □

APPENDIX B

APPENDIX OF CHAPTER 3

B.1 Solving CMDP and Optimistic CMDP

There does not exists a standard dynamic programming approach (value iteration or policy iteration) for solving CMDP. Instead, CMDPs are typically solved either using a Lagrangian approach or a linear programming approach (LP). In this chapter, we focus on the LP approach. Here, we first briefly discuss the LP based approach for solving CMDP (3.4). We then show that this approach can be extended to solve the optimistic CMDP problem 6.33.

As shown in [4], we can rewrite the CMDP problem as an LP using occupation measures. The occupancy measure μ under policy is defined as

$$\mu(s, a, \pi) := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_\pi(s_t = s, a_t = a | s_{t=0} = s_0), \tag{B.1}$$

where the probability $\mathbb{P}(\cdot)$ is calculated w.r.t. underlying transition kernel under policy π, P_π. It is easy to check that

$$(1 - \gamma)V^\pi = \sum_{s,a} \mu(s, a, \pi) r(s, a), \;\; (1 - \gamma)C_i^\pi = \sum_{s,a} \mu(s, a, \pi) c(i, s, a), \forall i. \tag{B.2}$$

Let μ be any generic occupation measure defined as (B.1). Now, the CMDP problem can be restated as an LP, to find the best occupation measure as follows [4].

$$\max_{\mu} \quad \sum_{s,a} \mu(s,a) r(s,a) \tag{B.3}$$

$$\text{s.t.} \quad \sum_{s,a} \mu(s,a) c(i,s,a) \leq \bar{C}_i/(1-\gamma), \ \ \forall i \in \{1,\ldots,N\}, \tag{B.4}$$

$$\sum_{s',a} \mu(s',a)(\mathbb{I}(s=s') - \gamma P(s'|s,a)) = (1-\gamma)\mathbb{I}(s=s_0) \ \ \forall s \in S \tag{B.5}$$

$$\mu(s,a) \geq 0 \ \ \forall s \in S, a \in A \tag{B.6}$$

It is proved that the LP (B.3) is equivalent to CMDP problem of (3.4), and the optimal policy computed by this LP is also the solution to CMDP problem in [4]. Eventually, the optimal policy π^* is calculated as follows

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

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

$$\begin{aligned}
&\max_{q} \sum_{s,a,s'} q(s,a,s') r(s,a) \\
&\quad \text{s.t.} \\
&\quad \sum_{s,a,s'} q(s,a,s') c(i,s,a) \leq \bar{C}_i/(1-\gamma) \ \ \forall i \in \{1,\ldots,N\}, \\
&\quad \sum_{s',a} [\mathbb{I}(s'=s)(\sum_{s''} q(s'',s',a)) - \gamma q(s',s,a)] = (1-\gamma)\mathbb{I}(s=s_0) \ \ \forall s \in S \\
&\quad q(s,a,s') \geq 0 \ \ \forall s,s' \in S, a \in A, \\
&\quad q(s,a,s') - (\widehat{P}(s'|s,a) + \beta(s,a,s')) \sum_{y} q(s,a,y) \leq 0 \ \ \forall s,a,s', \\
&\quad - q(s,a,s') + (\widehat{P}(s'|s,a) - \beta(s,a,s')) \sum_{y} q(s,a,y) \leq 0 \ \ \forall s,a,s'
\end{aligned}$$

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 GM-CRL and $\tilde{\pi}_t$ for UC-CRL, using the solution of above LP. Also, we can calculate an optimistic transition kernel denoted by $\tilde{P}$ by means of optimal $q(s,a,s')$. In brief, the optimistic transition kernel and optimistic policy are computed as follows

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

B.2 Detailed Proofs for Sample Complexity of Generative Model Based Learning

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

***Proof of Lemma 6*:** Fix a state, action and next state, i.e. s,a,s'. Then, according to Hoeffding's inequality [30]

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

Now, we apply empirical Bernstein's inequality [31] 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 50. *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 51. *Suppose there are two CMDPs* $M = \langle S, A, P, r, c, \bar{C}, s_0, \gamma\rangle$ *and* $M' = \langle S, A, P', r, c, \bar{C}, s_0\gamma\rangle$ *satisfying assumption 3. Then, under any policy* π

$$V^{\pi} - V'^{\pi} = \gamma(I - \gamma P'_{\pi})^{-1}(P_{\pi} - P'_{\pi})V^{\pi} \;\; \textit{and} \;\; V^{\pi} - V'^{\pi} = \gamma(I - \gamma P_{\pi})^{-1}(P_{\pi} - P'_{\pi})V'^{\pi},$$

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

$$C_i^{\pi} - C_i'^{\pi} = \gamma(I - \gamma P'_{\pi})^{-1}(P_{\pi} - P'_{\pi})C_i^{\pi} \;\; \textit{and} \;\; C_i^{\pi} - C_i'^{\pi} = \gamma(I - \gamma P_{\pi})^{-1}(P_{\pi} - P'_{\pi})C_i'^{\pi}.$$

Proof. We only prove the first statement of value function since the proof procedure for cost is

identical. For a fixed s

$$\begin{aligned}
V^\pi(s) - V'^\pi(s) &= r_\pi(s) + \gamma \sum_{s'} P_\pi(s'|s)V^\pi(s') - (r_\pi(s) + \gamma \sum_{s'} P'_\pi(s'|s)V'^\pi(s')) \\
&= \gamma\Big(\sum_{s'} P_\pi(s'|s)V^\pi(s') - \sum_{s'} P'_\pi(s'|s)V^\pi(s') + \sum_{s'} P'_\pi(s'|s)V^\pi(s') - \sum_{s'} P'_\pi(s'|s)V'^\pi(s')\Big) \\
&= \sum_{s'} \gamma(P_\pi(s'|s) - P'_\pi(s'|s))V^\pi(s') + \gamma \sum_{s'} P'_\pi(s'|s)(V^\pi(s') - V'^\pi(s')).
\end{aligned}$$

So, for vector $V^\pi - V'^\pi$ we have

$$V^\pi - V'^\pi = \gamma(P_\pi - P'_\pi)V^\pi + \gamma P'_\pi(V^\pi - V'^\pi).$$

Reorganizing the above equation would yield the result. □

Lemma 52. *Let $\delta_P \in (0,1)$. Suppose there are two CMDPs $M =< S, A, P, r, c, \bar{C}, \gamma >$ and $M' =< S, A, P', r, c, \bar{C}, \gamma >$ satisfying assumption 3. 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 π

$$|\sum_{s'}(P_\pi(s'|s) - P'_\pi(s'|s))V'^\pi(s')| \leq |S|c_1\|V'^\pi\|_\infty + c_2\sqrt{|S|}\sigma^2_{V'^\pi}(s)$$

for any $(s,a) \in S \times A$ w.p. at least $1 - |S||A|\delta_P$, and

$$|\sum_{s'}(P_\pi(s'|s) - P'_\pi(s'|s))C'^\pi_i(s')| \leq |S|c_1\|C'^\pi_i\|_\infty + c_2\sqrt{|S|}\sigma^2_{C'^\pi_i}(s)$$

for any $(s,a) \in S \times A, i \in \{1, \ldots, N\}$ w.p. at least $1 - |S||A|\delta_P$.

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 $\bar{V}^\pi(s') = \sum_{s''} P'_\pi(s''|s)V'^\pi(s'')$ as the expected value function of the successor states of s. Note that $\bar{V}^\pi(s')$ is a constant function and so $\bar{V}^\pi(s') = \sum_{s''} P'_\pi(s''|s)\bar{V}^\pi(s'') = \sum_{s''} P_\pi(s''|s)\bar{V}^\pi(s'')$.

$$|\sum_{s'}(P_\pi(s'|s) - P'_\pi(s'|s))V'^\pi(s')| = |\sum_{s'}(P_\pi(s'|s) - P'_\pi(s'|s))V'^\pi(s') + \bar{V}^\pi(s) - \bar{V}^\pi(s)|$$

$$= |\sum_{s'}(P_\pi(s'|s) - P'_\pi(s'|s))(V'^\pi(s') - \bar{V}^\pi(s'))|$$

$$\leq \sum_{s'} |P_\pi(s'|s) - P'_\pi(s'|s)||V^\pi(s') - \bar{V}^\pi(s')| \tag{B.7}$$

$$\leq \sum_{s'}(c_1 + c_2\sqrt{P'_\pi(s'|s)(1 - P'_\pi(s'|s))})|V^\pi(s') - \bar{V}^\pi(s')|$$

$$\leq |S|c_1\|V'^\pi\|_\infty + c_2\sum_{s'}\sqrt{P'_\pi(s'|s)(1 - P'_\pi(s'|s))(V^\pi(s') - \bar{V}^\pi(s'))^2}$$

$$\leq |S|c_1\|V'^\pi\|_\infty + c_2\sqrt{|S|\sum_{s'} P'_\pi(s'|s)(1 - P'_\pi(s'|s))(V^\pi(s') - \bar{V}^\pi(s'))^2} \tag{B.8}$$

$$\leq |S|c_1\|V'^\pi\|_\infty + c_2\sqrt{|S|\sum_{s'} P'_\pi(s'|s)(V^\pi(s') - \bar{V}^\pi(s'))^2}$$

$$= |S|c_1\|V'^\pi\|_\infty + c_2\sqrt{|S|}\sigma^2_{V'^\pi}.$$

Inequality (B.7) holds w.p. at least $1 - |S||A|\delta_P$, since we used the assumption and applied the triangle inequality and union bound. Please note that the premise of the Lemma is true w.p. at least $1 - \delta_P$ for every (s, a). Thus, the union bound gives $1 - |S||A|\delta_P$. Next, we then applied the assumed bound on $|V'^\pi(s') - \bar{V}^\pi(s')|$ and bounded it by $\|V'^\pi\|_\infty$ as all value functions are non-negative. In inequality (B.8), 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 $\sigma^2_{V'^\pi}$. □

Lemma 53. *Let $\delta_P \in (0, 1)$. Suppose there are two CMDPs $M = \langle S, A, P, r, c, \bar{C}, s_0, \gamma\rangle$ and $M' = \langle S, A, P', r, c, \bar{C}, s_0, \gamma\rangle$ satisfying assumption 3. Further assume*

$$|P(s'|s, a) - P'(s'|s, a)| \leq \frac{c_3}{\sqrt{n}}$$

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

$$\|V^\pi - V'^\pi\|_\infty \le \frac{c_3 \gamma |S|}{(1-\gamma)^2 \sqrt{n}},$$

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

$$\|C_i^\pi - C_i'^\pi\|_\infty \le \frac{c_3 \gamma |S|}{(1-\gamma)^2 \sqrt{n}}$$

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

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

$$\begin{aligned}
&\Delta = |V^\pi(s) - V'^\pi(s)| = |r_\pi(s) + \gamma \sum_{s'} P_\pi(s'|s)V^\pi(s') - (r_\pi(s) + \gamma \sum_{s'} P'_\pi(s'|s)V'^\pi(s'))| \\
&= \gamma |\sum_{s'} P_\pi(s'|s)V^\pi(s') - \sum_{s'} P'_\pi(s'|s)V^\pi(s') + \sum_{s'} P'_\pi(s'|s)V^\pi(s') - \sum_{s'} P'_\pi(s'|s)V'^\pi(s')| \\
&\le \gamma \sum_{s'} |(P_\pi(s'|s) - P'_\pi(s'|s)| \|V^\pi\|_\infty + \gamma\Delta \\
&\le \frac{c_3 \gamma |S|}{(1-\gamma)\sqrt{n}} + \gamma\Delta.
\end{aligned}$$

Because, $\|V^\pi\|_\infty \le \frac{1}{(1-\gamma)}$. Thus,

$$\Delta \le \frac{c_3 \gamma |S|}{(1-\gamma)^2 \sqrt{n}}$$

w.p. at least $1 - |S|^2|A|\delta_P$ by applying union bound over all current state, action and next state. Hence the proof is complete. □

Lemma 54. *Let $\delta_P \in (0,1)$. Suppose there are two CMDPs $M = \langle S, A, P, r, c, \bar{C}, s_0, \gamma \rangle$ and*

$M' = \langle S, A, P', r, c, \bar{C}, s_0\gamma\rangle$ *satisfying assumption 3. Further assume*

$$|P(s'|s,a) - P'(s'|s,a)| \le \frac{c_3}{\sqrt{n}}$$

w.p. at least $1 - \delta_P$ *for all* $s, s' \in S, a \in A$*. Then if* $n \ge \frac{\gamma^2 c_3 |S|}{12(1-\gamma)^2}$*, under any policy* π

$$\|\sigma_{V^\pi} - \sigma_{V'^\pi}\|_\infty \le \frac{2\gamma\sqrt{3c_3|S|}}{(1-\gamma)n^{1/4}},$$

w.p. at least $1 - 4|S|^3|A|\delta_P$*, and similarly for any* $i \in \{1, \dots, N\}$

$$\|\sigma_{C_i^\pi} - \sigma_{C_i'^\pi}\|_\infty \le \frac{2\gamma\sqrt{3c_3|S|}}{(1-\gamma)n^{1/4}}$$

w.p. at least $1 - 4|S|^3|A|\delta_P$*.*

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

$$\begin{aligned}\sigma^2_{V^\pi}(s) &= \sigma^2_{V^\pi}(s) - \gamma^2\mathbb{E}'[(V^\pi - P'_\pi V^\pi)^2(s)] + \gamma^2\mathbb{E}'[(V^\pi - P'_\pi V^\pi)^2(s)] \\ &\le \gamma^2 \sum_{s'}(P_\pi(s'|s) - P'_\pi(s'|s))V^{\pi^2}(s') - \gamma^2[(\sum_{s'} P_\pi(s'|s)V^\pi(s'))^2 - (\sum_{s'} P'_\pi(s'|s)V^\pi(s'))^2] \\ &+ [\gamma\sqrt{\mathbb{E}'[((V^\pi - V'^\pi) - P'_\pi(V^\pi - V'^\pi))^2(s)]} + \sqrt{\gamma^2\mathbb{E}'[(V'^\pi - P'_\pi V'^\pi)^2(s)]}]^2,\end{aligned}$$

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

$\|V^\pi - V'^\pi\|_\infty^2$ implying

$$\begin{aligned}
\sigma^2_{V^\pi}(s) \leq \gamma^2 &\sum_{s'}(P_\pi(s'|s) - P'_\pi(s'|s))V^{\pi^2}(s') && \text{(B.9)}\\
&- \gamma^2[\sum_{s'}(P_\pi(s'|s) - P'_\pi(s'|s))V^\pi(s')][\sum_{s'}(P_\pi(s'|s) + P'_\pi(s'|s))V^\pi(s')] && \text{(B.10)}\\
&+ (\gamma\|V^\pi - V'^\pi\|_\infty + \sigma_{V'^\pi}(s))^2 && \text{(B.11)}\\
&\leq [\sigma_{V'^\pi}(s) + \frac{\gamma^2 c_3|S|}{(1-\gamma)^2\sqrt{n}}]^2 + \frac{3\gamma^2 c_3|S|}{(1-\gamma)^2\sqrt{n}} && \text{(B.12)}\\
&\leq [\sigma_{V'^\pi}(s) + \frac{\gamma^2 c_3|S|}{(1-\gamma)^2\sqrt{n}} + \frac{\gamma\sqrt{3c_3|S|}}{(1-\gamma)n^{1/4}}]^2 && \text{(B.13)}\\
&\leq [\sigma_{V'^\pi}(s) + \frac{2\gamma\sqrt{3c_3|S|}}{(1-\gamma)n^{1/4}}]^2. && \text{(B.14)}
\end{aligned}$$

To obtain inequality (B.12), we did the following. First, we used the premise of the Lemma on transition kernel and the fact that $|V^\pi(s')| \leq \frac{1}{1-\gamma}$ on lines (B.9) and (B.10). Corresponding inequality holds w.p. at least $1 - |S||A|\delta_P$ by taking union bound over all (s, a), because it must be true for all (s, a). Next, we used Lemma 53 and get $\|V^\pi - V^{'\pi}\|_\infty \leq \frac{\gamma^2 c_3|S|}{(1-\gamma)^2\sqrt{n}}$ w.p. at least $1 - |S|^2|A|\delta_P$. Further, we used the fact that for any $x, y > 0$ we have $x^2 + y^2 \leq (x + y)^2$ to get inequality (B.13) which holds w.p. at least $1 - 2|S|^2|A|\delta_P$. The final inequality (B.14) is obtained by considering the assumption on n, which dominates the term with $\frac{1}{n^{1/4}}$ over $\sqrt{n}$. Eventually, the result follows by taking square root from both sides and taking union bound on all states and both directions, i.e. $\sigma_{V'^\pi}(s) \leq \sigma^\pi(s) + \frac{2\gamma\sqrt{3c_3|S|}}{(1-\gamma)n^{1/4}}$.

□

Lemma 55. *[6] The variance of the value function defined as* $\Sigma_{V^\pi}(s) = \mathbb{E}[(\sum_{t=0}^\infty \gamma^t r(s_t) - V^\pi(s))^2]$ *satisfies a Bellman equation* $\Sigma_{V^\pi}(s) = \sigma^2_{V^\pi}(s) + \gamma^2 \sum_{s' \in S} P_\pi(s'|s)\Sigma_{V^\pi}(s')$ *which gives* $\Sigma_{V^\pi} = (I - \gamma^2 P_\pi)^{-1}\sigma^2_{V^\pi}$*. Since* $0 \leq \Sigma_{V^\pi}(s) \leq \frac{1}{(1-\gamma)^2}$*, it follows that for every* $s \in S$

$$0 \leq (I - \gamma^2 P_\pi)^{-1}\sigma^2_{V^\pi}(s) \leq \frac{1}{(1-\gamma)^2}$$

and

$$0 \leq (I-\gamma^2 P_\pi)^{-1}\sigma_{V^\pi}(s) \leq \frac{2\log 2}{(1-\gamma)^{1.5}}$$

for all $s \in S$.

Corollary 8. *The result of Lemma 63 also holds for variance of cost functions.*

Proof of Lemma 28: We only prove the statement of value function since the proof procedure for cost is identical. First, we apply Lemma 60 and get

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

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

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

Now, let fix state s :

$$|V^\pi(s)-\tilde{V}^\pi(s)| = \gamma|(I-\gamma P_\pi)^{-1}(P_\pi-\tilde{P}_\pi)\tilde{V}^\pi|(s) \tag{B.16}$$

$$\leq \gamma(I-\gamma P_\pi)^{-1}|(P_\pi-\tilde{P}_\pi)\tilde{V}^\pi|(s) \leq \gamma(I-\gamma P_\pi)^{-1}(|S|c_1\|\tilde{V}^\pi\|_\infty\mathbf{1} + c_2\sqrt{|S|}\sigma_{\tilde{V}^\pi})(s) \tag{B.17}$$

$$\leq \frac{\gamma|S|c_1}{(1-\gamma)^2} + c_2\gamma\sqrt{|S|}((I-\gamma P_\pi)^{-1}\sigma_{\tilde{V}^\pi})(s) \tag{B.18}$$

$$\leq \frac{\gamma|S|c_1}{(1-\gamma)^2} + c_2\gamma\sqrt{|S|}((I-\gamma P_\pi)^{-1}(\sigma_{V^\pi} + \frac{2^{2.25}3^{0.5}\gamma|S|^{0.5}(\log 4/\delta_P)^{0.25}}{(1-\gamma)n^{1/4}}\mathbf{1}))(s) \tag{B.19}$$

$$\leq \frac{\gamma|S|c_1}{(1-\gamma)^2} + \frac{2\log 2c_2\gamma\sqrt{|S|}}{(1-\gamma)^{1.5}} + \frac{2^{2.25}3^{0.5}\gamma^2c_2|S|(\log 4/\delta_P)^{0.25}}{(1-\gamma)^2n^{1/4}} \tag{B.20}$$

$$= \frac{2^{2.5}\gamma\log 2|S|^{0.5}\log 4/{\delta_P}^{0.5}}{(1-\gamma)^{1.5}n^{0.5}} + \frac{2^{1.5}\gamma|S|\log 4/{\delta_P}^{0.75}(1+2^{2.25}3^{0.5}\gamma)}{(1-\gamma)^2n^{0.75}} + \frac{2^{0.5}3\gamma|S|\log 4/\delta_P}{(1-\gamma)^2n}$$

$$\leq 3\frac{2^{2.5}\gamma\log 2|S|^{0.5}\log 4/{\delta_P}^{0.5}}{(1-\gamma)^{1.5}n^{0.5}}. \tag{B.21}$$

In equation (B.16), we used Lemma 51. Then, we applied Lemma 52 to obtain inequality (B.17) which holds w.p. at least $1-|S||A|\delta_P$. Please note that $\mathbf{1}$ is a $|S|-$dimensional vector with all elements being 1. Next, we bound $\|\tilde{V}^{\pi}\|_{\infty}$ by $\frac{1}{1-\gamma}$ in inequality (B.18). Also note that $(I-\gamma P_{\pi})^{-1}\mathbf{1}(s) \leq \frac{1}{1-\gamma}$. To get inequality (B.19), we use Lemma 54, since we can bound $P(\cdot|s,a)-\tilde{P}(\cdot|s,a)$ by $2c_2$ due to the condition on n. This inequality holds w.p. at least $1-4|S|^3|A|\delta_P$. Inequality of line (B.20) is obtained by bounding $(I-\gamma P)^{-1}\sigma_{V^{\pi}}(s)$ by $\frac{2\log 2}{(1-\gamma)^{1.5}}$ using Lemma 63. Finally, inequality (B.21) is according to the condition on n. Since the result must be true for every $s \in S$, we take union bound over all s. Hence, the proof is complete. □

Proof of Theorem 12: Let $\delta_P \in (0,1)$. First, we know that optimistic planning problem (6.33) 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 between feasibility and following events.

Now, we have

$$V^{\pi^*}(s_0) - 3\gamma\log 2\sqrt{\frac{32|S|\log 4/\delta_P}{(1-\gamma)^3 n}} \leq \tilde{V}^{\pi^*}(s_0) \leq V^{\pi^*}(s_0) + 3\gamma\log 2\sqrt{\frac{32|S|\log 4/\delta_P}{(1-\gamma)^3 n}}$$

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

$$V^{\tilde{\pi}}(s_0) - 3\gamma\log 2\sqrt{\frac{32|S|\log 4/\delta_P}{(1-\gamma)^3 n}} \leq \tilde{V}^{\tilde{\pi}}(s_0) \leq V^{\tilde{\pi}}(s_0) + 3\gamma\log 2\sqrt{\frac{32|S|\log 4/\delta_P}{(1-\gamma)^3 n}}$$

w.p. at least $1-5|S|^3|A|\delta_P$ according to Lemma 28. On the other hand, we know that $\tilde{V}^{\pi^*}(s_0) \leq \tilde{V}^{\tilde{\pi}}(s_0)$. Thus, by combining these results we get

$$V^{\pi^*}(s_0) - 3\gamma\log 2\sqrt{\frac{32|S|\log 4/\delta_P}{(1-\gamma)^3 n}} \leq \tilde{V}^{\pi^*}(s) \leq \tilde{V}^{\tilde{\pi}}(s_0) \leq V^{\tilde{\pi}}(s) + 3\gamma\log 2\sqrt{\frac{32|S|\log 4/\delta_P}{(1-\gamma)^3 n}}.$$

It yields that $V^{\tilde{\pi}}(s_0) \geq V^{\pi^*}(s_0) - 6\gamma\log 2\sqrt{\frac{32|S|\log 4/\delta_P}{(1-\gamma)^3 n}}$ w.p. at least $1-10|S|^3|A|\delta_P$ by union bound.

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

$$C_i^{\tilde{\pi}}(s_0) \leq \tilde{C}_i^{\tilde{\pi}}(s_0) + 3\gamma \log 2\sqrt{\frac{32|S| \log 4/\delta_P}{(1-\gamma)^3 n}} \leq \bar{C}_i + 3\gamma \log 2\sqrt{\frac{32|S| \log 4/\delta_P}{(1-\gamma)^3 n}}$$

w.p. at least $1 - 5|S|^3|A|\delta_P$ according to Lemma 28. By taking union bound, we get that all statements for value and cost functions hold w.p. at least $1-(5N+10)|S|^3|A|\delta_P$. Hence, putting $\epsilon = 6\gamma \log 2\sqrt{\frac{32|S| \log 4/\delta_P}{(1-\gamma)^3 n}}$ and $\delta = (5N+10)|S|^3|A|\delta_P$ concludes the proof. Please note that $\epsilon < \frac{0.22\gamma}{\sqrt{|S|(1-\gamma)}}$ would satisfy the assumption in Lemma 28. □

B.3 Detailed Proof for Theorem 4

First, we bound total number of model updates in Algorithm 4.

Lemma 56. *The total number of updates under algorithm 4 is bounded by* $U_{\max} = \frac{|S|^2|A|m}{1-\gamma}$.

Proof. Let fix a $(s,a)-$pair. Note that $n(s,a)$ is not decreasing and also it increases up to $\frac{|S|m}{(1-\gamma)}$. And, since update of model happens at the beginning of each time-step, then maximum number of updates due to a single (s,a) happens at most $\frac{|S|m}{(1-\gamma)}$ number of times. Thus, maximum number of updates due to all $(s,a)-$pairs is no larger than $\frac{|S|^2|A|m}{1-\gamma}$ □

***Proof of Lemma 30*:** At each time-step with model update t and for each (s,a), by Hoeffding's inequality [30] 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 [31] 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 33. But, first we provide some useful lemmas.

Lemma 57. *Total number of observations of* $(s,a) \in X_{t,\kappa,\iota}$ *with* $\kappa \in [1, |S|-1]$ *and* $\iota > 0$ *over all time-steps* t *is at most* $3|S \times A|mw_\iota\kappa$*.* $w_\iota = \min\{w_t(s,a) : \iota_t(s,a) = \iota\}$*.*

Proof. Note that $w_{\iota+1} = 2w_\iota$ for $\iota > 0$. Consider a time-step t and a fixed $(s,a) \in X_{t,\kappa,\iota}$. Since we assumed $\iota_t(s,a) = \iota$, then $w_\iota \leq w_t(s,a) \leq 2w_\iota$. Similarly, from $\kappa_t(s,a) = \kappa$ we have $\frac{n_t(s,a)}{2mw_t(s,a)} \leq \kappa \leq \frac{n_t(s,a)}{mw_t(s,a)}$ which implies

$$mw_\iota\kappa \leq mw_t(s,a)\kappa \leq n_t(s,a) \leq 2mw_t(s,a)\kappa \leq 4mw_\iota\kappa. \tag{B.22}$$

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

Lemma 58. *Number of time-steps* $E_{\kappa,\iota}$ *with* $|X_{t,\kappa,\iota}| > \kappa$ *is bounded for* $\alpha \geq 3$ *w.h.p.*

$$P(E_{\kappa,\iota} > \alpha K) \leq \exp(-\beta w_\iota(\kappa+1)K(1-\gamma)^2),$$

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

Proof. Let $\nu_t := \sum_{k=0}^{t} \gamma^k \mathbb{I}\{(s_k,a_k) \in X_{k,\kappa,\iota}\}$ be discounted number of observations of (s,a) with $|X_{k,\kappa,\iota}| > \kappa$. We have $t \in \{1, ..., E_{\kappa,\iota}\}$.

In these time-steps $|X_{t,\kappa,\iota}| \geq \kappa+1$ and all (s,a) in partition (κ,ι) have $w_t(s,a) \geq w_\iota$, then

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

Also $\mathbb{V}[\nu_t|\nu_1,\ldots,\nu_{t-1}] \leq \mathbb{E}[\nu_t|\nu_1,\ldots,\nu_{t-1}]/(1-\gamma)$ since $\nu_t \in [0, \frac{1}{1-\gamma}]$.

Now, we define the continuation:

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

and centralized auxiliary sequence

$$\bar{\nu}_t := \frac{\nu_t^+ w_\iota(\kappa+1)}{\mathbb{E}[\nu_t^+|\nu_1^+,\ldots,\nu_{t-1}^+]}.$$

By construction

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

According to lemma 57, we have $E_{\kappa,\iota} > \alpha K$ if

$$\sum_{t=1}^{\alpha K} \bar{\nu}_t \leq 3Kw_\iota\kappa(1-\gamma) \leq 3Kw_\iota(\kappa+1)(1-\gamma).$$

The factor of $(1-\gamma)$ is due to fact that $\bar{\nu}_t$ is discounted number of visitation. While, $E_{\kappa,\iota}$ is total number of visitations. Hence, we normalize that by multiplying the right hand side by $(1-\gamma)$.

Now, we define martingale below

$$B_t := \mathbb{E}\left[\sum_{j=1}^{\alpha K} \bar{\nu}_j|\bar{\nu}_1,\ldots,\bar{\nu}_t\right] = \sum_{j=1}^{t} \bar{\nu}_j + \sum_{j=t+1}^{\alpha K} \mathbb{E}[\bar{\nu}_j|\bar{\nu}_1,...,\bar{\nu}_t],$$

which gives $B_0 = \alpha K w_\iota(\kappa+1)$ and $B_{\alpha K} = \sum_{t=1}^{\alpha K} \bar{\nu}_t$. Now, since $\nu_t^+ \in [0, \frac{1}{1-\gamma}]$

$$|B_{t+1} - B_t| = |\bar{\nu}_t - \mathbb{E}[\bar{\nu}_t|\bar{\nu}_1,\ldots,\bar{\nu}_{t-1}]| = \left|\frac{w_\iota(\kappa+1)(\nu_t^+ - \mathbb{E}[\nu_t^+|\bar{\nu}_1,\ldots,\bar{\nu}_{t-1}])}{\mathbb{E}[\nu_t^+|\nu_,^+\ldots,\nu_{t-1}^+]}\right|$$
$$\leq |\nu_t^+ - \mathbb{E}[\nu_t^+|\bar{\nu}_1,\ldots,\bar{\nu}_{t-1}]| \leq \frac{1}{1-\gamma}.$$

Using

$$\sigma^2 := \sum_{t=1}^{\alpha K} \mathbb{V}[B_t - B_{t-1} | B_1 - B_0, \dots, B_{t-1} - B_{t-2}] = \sum_{t=1}^{\alpha K} \mathbb{V}[\bar{\nu}_t | \bar{\nu}_1, \dots, \bar{\nu}_{t-1}] \le \alpha \frac{K w_\iota(\kappa+1)}{1-\gamma} = \frac{B_0}{1-\gamma}.$$

We can apply Theorem 22 of [69] and obtain

$$\begin{aligned}\mathbb{P}(E_{\kappa,\iota} > \alpha K) &\le \mathbb{P}\left(\sum_{t=1}^{\alpha K} \bar{\nu}_t \le 3Kw_\iota(\kappa+1)(1-\gamma)\right) = \mathbb{P}(B_{\alpha K} - B_0 \le (3B_0/\alpha - B_0)(1-\gamma)) \\ &\le \exp\left(-\frac{(3/\alpha-1)^2 B_0^2 (1-\gamma)}{2\sigma^2 + (1/3 - 1/\alpha)B_0/(1-\gamma)}\right)\end{aligned}$$

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

$$\mathbb{P}(E_{\kappa,\iota} > \alpha K) \le \exp\left(-Kw_\iota(\kappa+1)(1-\gamma)^2 \frac{\alpha(3/\alpha)^2}{7/3 - 1/\alpha}\right).$$

□

***Proof of Lemma 33*:** Since $w_t(s,a) \le \frac{1}{1-\gamma}$, we have that $\frac{w_t(s,a)}{w_{min}} < \frac{1}{w_{min}(1-\gamma)}$ and so $\iota_t(s,a) \le \frac{1}{(1-\gamma)w_{\min}} = \frac{4|S|}{\epsilon(1-\gamma)^2}$. In addition, $|X_{t,\kappa,\iota}| \le |S|$ for all t, κ, ι and so $|X_{t,\kappa,\iota}| > \kappa$ can only be true for $\kappa \le |S|$. Hence, only $E_{max} = \log_2 \frac{1}{w_{min}(1-\gamma)} \log_2 |S|$ possible values for (κ, ι) exists that can have $|X_{t,\kappa,\iota}| > \kappa$. By union bound over all (κ, ι) and lemma 58, we get

$$\begin{aligned}\mathbb{P}(E \le \alpha K E_{max}) &\ge \mathbb{P}(\max_{(\kappa,\iota)} \le \alpha K) \ge 1 - E_{max} \exp\left(-\beta w_\iota(\kappa+1)K(1-\gamma)^2\right) \\ &\ge 1 - E_{max}\exp\left(-\beta w_{min} K (1-\gamma)^2\right) = 1 - E_{max} \exp\left(-\beta w_{min} m |S \times A| (1-\gamma)^2\right) \\ &= 1 - E_{max} \exp\left(-\frac{\beta \epsilon m |S \times A| (1-\gamma)^3}{4|S|}\right).\end{aligned}$$

Bounding the right hand-side by $1 - \frac{\delta}{2(N+1)}$ and solving for m gives

$$1 - E_{max} \exp\left(-\frac{\beta \epsilon m |S \times A|(1-\gamma)^3}{4|S|}\right) \ge 1 - \frac{\delta}{2(N+1)} \Leftrightarrow m \ge \frac{4|S|}{|S \times A|(1-\gamma)^3 \beta\epsilon} \log \frac{2E_{max}}{\delta}.$$

Hence, the condition

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

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 35.

Corollary 9. *If we substitute the δ_P with δ_1 in Lemma 60, the result will pertain.*

Corollary 10. *If we substitute the δ_P with δ_1 in Lemma 52, the result will pertain.*

Proof of Lemma 35: 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 under policy π as $r_\pi^{(2d+2)} = \sigma_{V^{(d)},\pi}^{(d),2}$, where $\sigma_{V^{(d)},\pi}^{(d),2}$ is local variance of the value function w.r.t. the rewards $r^{(d)}$. Note that for every d and $s \in S$, we have $r^{(d)}(s) \in [0, \frac{1}{(1-\gamma)^d}]$.

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

Now,

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

The first equality follows from Lemma 51, the second step from the fact that all elements os V is non-negative and $(I-\gamma P)^{-1}$ being non-expansive. In the third line, 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 $\mathbb{I}\{s = ., a\sim\tilde{\pi}(s,\cdot)\}|(P-\tilde{P})\tilde{V}^{(d)}(\cdot)|$ 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

$$|V^{(d)}(s_0) - \tilde{V}^{(d)}(s_0)| = \Delta_d \le \sum_{s,a \in S \times A} \gamma (I - \gamma P)^{-1} \Big(\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot)\} |(P - \tilde{P}) \tilde{V}^{(d)}(s)| \Big) (s_0)$$

$$\le \gamma \sum_{s,a \notin X} \|\tilde{V}^{(d)}\|_\infty (I - \gamma P)^{-1} (\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot)\})(s_0)$$

$$+ \gamma \sum_{s,a \in X} |(P - \tilde{P}) \tilde{V}^{(d)}(s)| (I - \gamma P)^{-1} \mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot)\})(s_0)$$

$$\le \gamma \sum_{s,a \notin X} \frac{1}{(1-\gamma)^{d+1}} (I - \gamma P)^{-1} (\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot)\})(s_0)$$

$$+ \gamma \sum_{s,a \in X} |(P - \tilde{P}) \tilde{V}^{(d)}(s)| (I - \gamma P)^{-1} \mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot)\})(s_0)$$

$$\le \gamma \sum_{s,a \notin X} \frac{1}{(1-\gamma)^{d+1}} (I - \gamma P)^{-1} (\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot)\})(s_0)$$

$$+ \gamma \sum_{s,a \in X} ||S| c_1(s,a) \frac{1}{(1-\gamma)^{d+1}} + c_2(s,a) \sqrt{|S|} \sigma_{\tilde{V}}^{(d)}(s) | (I - \gamma P)^{-1} (\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot)\})(s_0)$$

$$\le \gamma \sum_{s,a \notin X} \frac{1}{(1-\gamma)^{d+1}} (I - \gamma P)^{-1} (\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot)\})(s_0)$$

$$+ \gamma \sum_{s,a \in X} ||S| c_1(s,a) \frac{1}{(1-\gamma)^{d+1}} | (I - \gamma P)^{-1} (\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot)\})(s_0)$$

$$+ \gamma \sum_{s,a \in X} | \sqrt{|S|} c_2(s,a) \sigma_{\tilde{V}}^{(d)}(s) | (I - \gamma P)^{-1} (\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot)\})(s_0)$$

$$\le \sum_{s,a \notin X} \frac{1}{(1-\gamma)^{d+1}} w(s,a) + \sum_{s,a \in X} |S| c_1(s,a) \frac{1}{(1-\gamma)^{d+1}} w(s,a)$$

$$+ \gamma \sum_{s,a \in X} \sqrt{|S|} c_2(s,a) \sigma_{\tilde{V}}^{(d)}(s) (I - \gamma)^{-1} (\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot)\})(s_0)$$

$$\le \frac{w_{\min} |S|}{(1-\gamma)^{d+1}} + \sum_{s,a \in X} \frac{|S| c_1(s,a) w(s,a)}{(1-\gamma)^{d+1}}$$

$$+ \gamma \sum_{s,a \in X} \sqrt{|S|} c_2(s,a) \sigma_{\tilde{V}}^{(d)}(s) (I - \gamma P)^{-1} (\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot)\})(s_0)$$

$$= \frac{\epsilon}{4(1-\gamma)^d} + \sum_{s,a \in X} \frac{|S| c_1(s,a) w(s,a)}{(1-\gamma)^{d+1}}$$

$$+ \gamma \sum_{s,a \in X} \sqrt{|S|} c_2(s,a) \sigma_{\tilde{V}}^{(d)}(s) (I - \gamma P)^{-1} (\mathbb{I}\{s = \cdot, a \sim \tilde{\pi}(s, \cdot)\})(s_0)$$

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 $\|V^{(d)}\|_\infty \leq \frac{1}{(1-\gamma)^{d+1}}$. We then apply Lemma 52 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 9 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

$$\begin{aligned}\Delta_d \leq \frac{\epsilon}{4(1-\gamma)^d} &+ 2\sqrt{2}\frac{|S|}{(1-\gamma)^{d+1}}\log 4/{\delta_1}^{3/4}\sum_{s,a\in X}\frac{w(s,a)}{n(s,a)^{3/4}} + 3\sqrt{2}\frac{|S|}{(1-\gamma)^{d+1}}\log 4/\delta_1\sum_{s,a\in X}\frac{w(s,a)}{n(s,a)} \\ &+ \sqrt{8|S|\log 4/\delta_1}\sum_{s,a\in X}\frac{\gamma}{\sqrt{n(s,a)}}\sigma_{\tilde{V}}^{(d)}(s)(I-\gamma P)^{-1}(\mathbb{I}\{s=\cdot, a\sim\tilde{\pi}(s,\cdot)\})(s_0)\end{aligned}$$

Hence, we bound

$$\Delta_d \leq \frac{\epsilon}{4(1-\gamma)^d} + U_d(s_0) + Y_d(s_0) + Z_d(s_0)$$

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

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

In the second line, we used Cauchy-Scharwz. Next, we used the fact that for $s,a\in X_{\kappa,\iota}$, we have $n(s,a)\ge mw(s,a)\kappa$, refer to equation (B.22). Finally, we applied the assumption of $|X_{\kappa,\iota}|\le\kappa$. Please note that $\mathcal{K}\times\mathcal{I}$ is the set of all possible (κ,ι) pairs.

The next term is

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

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

The last term is

$$
\begin{aligned}
& Z_d(s_0) = \sqrt{8|S|\log 4/\delta_1} \sum_{s,a\in X} \frac{\gamma}{\sqrt{n(s,a)}} (I-\gamma P)^{-1} \sigma^{(d)}_{\tilde{V}}(s) (\mathbb{I}\{s=\cdot, a\sim \tilde{\pi}(s,\cdot)\})(s_0) \\
& \leq \sqrt{8|S|\log 4/\delta_1} \sum_{s,a\in X} \frac{\gamma}{\sqrt{n(s,a)}} \sqrt{(I-\gamma P)^{-1} \mathbb{I}\{s=\cdot, a\sim \tilde{\pi}(s,\cdot)\}(s_0)} \\
& \times \sqrt{(I-\gamma P)^{-1} \sigma^{(d),2}_{\tilde{V}}(s) \mathbb{I}\{s=\cdot, a\sim \tilde{\pi}(s,\cdot)\}(s_0)} \\
& = \gamma\sqrt{8|S|\log 4/\delta_1} \sum_{s,a\in X} \sqrt{\frac{w(s,a)}{n(s,a)} (I-\gamma P)^{-1} \sigma^{(d),2}_{\tilde{V}}(s) \mathbb{I}\{s=\cdot, a\sim \tilde{\pi}(s,\cdot)\}(s_0)} \\
& = \gamma\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)} (I-\gamma P)^{-1} \sigma^{(d),2}_{\tilde{V}}(s) \mathbb{I}\{s=\cdot, a\sim \tilde{\pi}(s,\cdot)\}(s_0)} \\
& \leq \gamma\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)} (I-\gamma P)^{-1} \sigma^{(d),2}_{\tilde{V}}(s) \mathbb{I}\{s=\cdot, a\sim \tilde{\pi}(s,\cdot)\}(s_0)} \\
& \leq \gamma\sqrt{8|S|\log 4/\delta_1} \sum_{\kappa,\iota} \sqrt{\frac{1}{m} \sum_{s,a\in X_{\kappa,\iota}} (I-\gamma P)^{-1} \sigma^{(d),2}_{h}(s) \mathbb{I}\{s=\cdot, a\sim \tilde{\pi}(s,\cdot)\}(s_0)} \\
& \leq \gamma\sqrt{\frac{8|S|\log 4/\delta_1 |\mathcal{K}\times\mathcal{I}|}{m} \sum_{s,a\in X} (I-\gamma P)^{-1} \sigma^{(d),2}_{\tilde{V}}(s) \mathbb{I}\{s=\cdot, a\sim \tilde{\pi}(s,\cdot)\}(s_0)} \\
& \leq \gamma\sqrt{\frac{8|S|\log 4/\delta_1 |\mathcal{K}\times\mathcal{I}|}{m} \sum_{s,a\in S\times A} (I-\gamma P)^{-1} \sigma^{(d),2}_{\tilde{V}}(s) \mathbb{I}\{s=\cdot, a\sim \tilde{\pi}(s,\cdot)\}(s_0)} \\
& = \gamma\sqrt{\frac{8|S|\log 4/\delta_1 |\mathcal{K}\times\mathcal{I}|}{m} (I-\gamma P)^{-1} \sigma^{(d),2}_{\tilde{V}}(s_0)} \\
& \leq \sqrt{\frac{8|S|\log 4/\delta_1 |\mathcal{K}\times\mathcal{I}|}{m(1-\gamma)^{2d+3}}}.
\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 $|X_{\kappa,\iota}| \leq \kappa$ to get sixth step. Next, we applied Cauchy-Scharwz inequality to obtain seventh step. And, the final step follows from the facts that

$\gamma \leq 1, (I-\gamma P)^{-1}(s_0) \leq \frac{1}{1-\gamma}$ and $\|\sigma_{\tilde{V}}^{(d)}\|_\infty \leq \frac{1}{(1-\gamma)^{2d+2}}$. Thus, we have

$$Z_d(s_0) \leq \sqrt{\frac{8|S| \log 4/\delta_1 |\mathcal{K} \times \mathcal{I}|}{m(1-\gamma)^{2d+3}}}. \tag{B.23}$$

However, we can improve this bound as follows

$$\begin{aligned}
&Z_d(s_0) \leq \gamma \sqrt{\frac{8|S| \log 4/\delta_1 |\mathcal{K} \times \mathcal{I}|}{m} (I-\gamma P)^{-1} \sigma_{\tilde{V}}^{(d),2}(s_0)} \\
&= \gamma \sqrt{\frac{8|S| \log 4/\delta_1 |\mathcal{K} \times \mathcal{I}|}{m} (I-\gamma P)^{-1} \sigma_{\tilde{V}}^{(d),2}(s_0) - (I-\gamma \tilde{P}^{-1}) \sigma_{\tilde{V}}^{(d),2}(s_0) + (I-\gamma \tilde{P})^{-1} \sigma_{\tilde{V}}^{(d),2}(s_0)} \\
&\leq \gamma \sqrt{\frac{8|S| \log 4/\delta_1 |\mathcal{K} \times \mathcal{I}|}{m} \Big(\frac{1}{(1-\gamma)^{2d+2}} + (I-\gamma P)^{-1} r^{(2d+2)}(s_0) - (I-\gamma \tilde{P})^{-1} r^{(2d+2)}(s_0) \Big)} \\
&= \gamma \sqrt{\frac{8|S| \log 4/\delta_1 |\mathcal{K} \times \mathcal{I}|}{m} \Big(\frac{1}{(1-\gamma)^{2d+2}} + V^{(2d+2)}(s_0) - \tilde{V}^{(2d+2)}(s_0) \Big)} \\
&= \gamma \sqrt{\frac{8|S| \log 4/\delta_1 |\mathcal{K} \times \mathcal{I}|}{m} (\frac{1}{(1-\gamma)^{2d+2}} + \Delta_{2d+2})} \\
&\leq \sqrt{\frac{8|S| \log 4/\delta_1 |\mathcal{K} \times \mathcal{I}|}{m(1-\gamma)^{2d+2}}} + \sqrt{\frac{8|S| \log 4/\delta_1 |\mathcal{K} \times \mathcal{I}|}{m} \Delta_{2d+2}}.
\end{aligned}$$

In the third step, we used Lemma 63 and definition of $r^{(2d+2)}$. In the last line, we used the fact that $\gamma \leq 1$.

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

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

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

$\beta = \lceil \frac{1}{2\log 2} \log \frac{1}{1-\gamma} \rceil$ with $\mathcal{D} = \{0, 2, 6, 14, \ldots, \beta\}$, we get

$$\Delta_0 \leq \sum_{d \in \mathcal{D} \backslash \beta} \Big(\frac{8|S| \log 4/\delta_1 |\mathcal{K} \times \mathcal{I}|}{m} \Big)^{\frac{d}{d+2}} \Big[\frac{\epsilon(1-\gamma)^d}{4} + 3\sqrt{\frac{8|S| \log 4/\delta_1 |\mathcal{K} \times \mathcal{I}|}{m(1-\gamma)^{2d+2}}} \Big]^{\frac{2}{d+2}}$$
$$+ \Big(\frac{8|S| \log 4/\delta_1 |\mathcal{K} \times \mathcal{I}|}{m} \Big)^{\frac{\beta}{\beta+2}} \Big[\frac{\epsilon}{4(1-\gamma)^\beta} + 3\sqrt{\frac{8|S| \log 4/\delta_1 |\mathcal{K} \times \mathcal{I}|}{m(1-\gamma)^{2\beta+2}}} \Big]^{\frac{2}{\beta+2}}.$$

Here, we used inequality (B.23) to bound $Z_\beta(s_0)$. Finally, the proof completes if we let

$$m = 1280 \frac{|S|}{\epsilon^2 (1-\gamma)^2} (\log_2 \log_2 (\frac{1}{(1-\gamma)}))^2 \log_2^2 \Big(\frac{8|S|^2}{\epsilon(1-\gamma)^2} \Big) \log \frac{6}{\delta_1}.$$

□.

Proof of Theorem 4: By Lemma 33, we know that number of time-steps where $|X_{\kappa,\iota}| > \kappa$ for some κ, ι is bounded by $6E_{\max}|S||A|m$ with probability at least $1 - \frac{\delta}{2(N+1)}$. For all other time-steps, we have by Lemma 35 that for any $i \in \{1, \ldots, N\}$

$$|\tilde{V}^{\tilde{\pi}_t}(s_0) - V^{\tilde{\pi}_t}(s_0)| \leq \epsilon, \quad |\tilde{C}_i^{\tilde{\pi}_t}(s_0) - C_i^{\tilde{\pi}_t}(s_0)| \leq \epsilon. \tag{B.24}$$

Using Lemma 30, we get that $M \in \mathcal{M}_t$ for any time-step t w.p. at least $1 - \frac{\delta}{2(N+1)}$. Further, we know that ELP outputs the policy $\tilde{\pi}_t$ such that

$$\tilde{V}^{\tilde{\pi}_t}(s_0) \geq V^{\pi^*}(s_0), \quad \tilde{C}_i^{\tilde{\pi}_t}(s_0) \leq \bar{C}_i \;\; i \in \{1, \ldots, N\} \tag{B.25}$$

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

$$V^{\tilde{\pi}_t}(s_0) \geq V^{\pi^*}(s_0) - \epsilon$$

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

$$m \geq 1280\frac{|S|}{\epsilon^2(1-\gamma)^2}(\log_2\log_2(\frac{1}{1-\gamma}))^2\log_2^2\Big(\frac{8|S|^2}{\epsilon(1-\gamma)^2}\Big)\log\frac{4}{\delta_1} \text{ and}$$
$$m \geq \frac{4}{\epsilon(1-\gamma)^3}\log\frac{2(N+1)E_{\max}}{\delta}.$$

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,

$$m \geq 1280\frac{|S|}{\epsilon^2(1-\gamma)^3}(\log_2\log_2(\frac{1}{1-\gamma}))^2\log_2^2\Big(\frac{8|S|^2}{\epsilon(1-\gamma)^2}\Big)\log\frac{4(N+1)|S|^2|A|m}{\delta}. \tag{B.26}$$

It is well-known fact that for any constant $B > 0, \nu \geq 2B\log B$ implies $\nu \geq B\log\nu$. we can use this sufficiency condition to satisfy (B.26) by setting m as follows:

$$m \geq 2560\frac{|S|}{\epsilon^2(1-\gamma)^3}(\log_2\log_2(\frac{1}{1-\gamma}))^2\log_2^2\Big(\frac{8|S|^2}{\epsilon(1-\gamma)^2}\Big)$$
$$\times\log\Big(\frac{5120(N+1)|S|^3|A|}{\epsilon^2\delta(1-\gamma)^3}(\log_2\log_2(\frac{1}{1-\gamma}))^2\log_2^2\Big(\frac{8|S|^2}{\epsilon(1-\gamma)^2}\Big)\Big).$$

Also,

$$E_{\max} = \log_2|S|\log_2\frac{4|S|}{\epsilon(1-\gamma)^2} \leq \log_2^2\frac{4|S|}{\epsilon(1-\gamma)^2}$$

and

$$\log\frac{2(N+1)E_{\max}}{\delta} = \log\frac{2(N+1)\log_2|S|\log_2(\frac{4|S|}{\epsilon(1-\gamma)^2})}{\delta} \leq \log\frac{2(N+1)\log_2^2(\frac{4|S|}{\epsilon(1-\gamma)^2})}{\delta}$$
$$\leq \log\frac{16(N+1)|S|^4|A|}{\epsilon\delta(1-\gamma)^3}.$$

Setting

$$\begin{aligned} m = 2560 & \frac{|S|}{\epsilon^2(1-\gamma)^3}(\log_2\log_2(\frac{1}{1-\gamma}))^2\log_2^2\Big(\frac{8|S|^2}{\epsilon(1-\gamma)^2}\Big) \\ & \times \log\Big(\frac{5120(N+1)|S|^4|A|}{\epsilon^2\delta(1-\gamma)^3}(\log_2\log_2(\frac{1}{1-\gamma}))^2\log_2^2\Big(\frac{8|S|^2}{\epsilon(1-\gamma)^2}\Big)\Big). \end{aligned} \tag{B.27}$$

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

$$\begin{aligned} 6E_{\max}|S||A|m = & 15360\frac{|S|^2|A|}{\epsilon^2(1-\gamma)^3}(\log_2\log_2(\frac{1}{1-\gamma}))^2\log_2^2\Big(\frac{4|S|}{\epsilon(1-\gamma)^2}\Big)\log_2^2\Big(\frac{8|S|^2}{\epsilon(1-\gamma)^2}\Big) \\ & \times \log\Big(\frac{5120(N+1)|S|^4|A|}{\epsilon^2\delta(1-\gamma)^3}(\log_2\log_2(\frac{1}{1-\gamma}))^2\log_2^2\Big(\frac{8|S|^2}{\epsilon(1-\gamma)^2}\Big)\Big) \end{aligned}$$

time-steps which neither are optimal nor near-optimal. □

APPENDIX C

APPENDIX OF CHAPTER 4

C.1 GMBL-Dual

Proof of Proposition 1: Considering optimization problem of (4.4) and Assumptions 5 and 6, it is obvious that optimal policy of the problem (4.4) would yield positive result, i.e. $V_0^{\pi^*}(s_0) = V_0^{\pi^*(\lambda^*)}(s_0) > 0$. Comparing value of π^* with do-nothing policy and knowing the fact that do-nothing policy yield 0 reward, we conclude that $V_0^{\pi^*(\lambda^*)}(s_0, \lambda^*)$ must be positive, otherwise do-nothing policy must have been chosen.

On the other hand, optimal policy might satisfy some of constraints tightly, not all of them. Due to Complementary Slackness, if constraint i is not completely satisfied, then λ_i^* would be 0. First, consider the case where there is no tight satisfaction. Then, all the λ_i^* would be 0. So, we focus on the case where there is at least one positive λ_i^*. Thus, we have

$$0 < V_0^{\pi^*(\lambda^*)}(s_0, \lambda^*) = V_0^{\pi^*(\lambda^*)}(s_0) - \sum_i \lambda_i^* \bar{C}_i \leq H - \sum_i \lambda_i^* \bar{C}_i,$$

Therefore,

$$0 \leq H - \sum_i \lambda_i^* \bar{C}_i \leq H - B_\lambda \bar{C}_{\min}.$$

The second inequality comes from the fact that $\lambda_{\max}$ must appear, and the worst case is that it appears at least once with the least $\bar{C}_i$. Hence the proof completes. □

C.2 Online-CRL-Dual

Before proceeding with Lemma 17, for any $\lambda \in [0, B_\lambda]$ we reason about a sequence of MDPs $M_{d,\lambda}$ which have the same transition probabilities but different reward matrix $r_c^{(d)}(\lambda)$. For $d = 0$, the reward matrix is the original reward function $r_c(\lambda)$ of M_λ ($r_c^{(0)}(\lambda) = r_c(\lambda)$.) The following reward matrices are then defined recursively as $r_c^{(2d+2)}(\lambda) = \max_h \sigma_h^{(d),2}$, where $\sigma_{h:H-1}^{(d),2}$ is local variance of the value function w.r.t. the rewards $r_c^{(d)}(\lambda)$. Note that for every d and $h = 0, ..., H-1$ and $(s,a) \in S \times A$, we have $r_c^{(d)}(\lambda, s, a) \in [0, (HNB_\lambda)^d]$.

In addition, we will drop the notations k, $\tilde{\lambda}^{(k)}$ and policy $\tilde{\pi}_k(\tilde{\lambda}^{(k)})$ 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 $\tilde{\pi}_k(\tilde{\lambda}^{(k)})$.

Now, we present and prove some lemmas required for Lemma 17.

Lemma 59.

$$V_0 - \tilde{V}_0 = \sum_{h=0}^{H-2} P^{h-1}(P - \tilde{P})\tilde{V}_{h+1}.$$

Proof. For a fixed h and s :

$$\begin{aligned} V_h(s) - \tilde{V}_h(s) &= r_c(s) + \sum_{s'} P(s'|s)V_{h+1}(s') - r_c(s) - \sum_{s'} \tilde{P}(s'|s)\tilde{V}_{h+1}(s') \\ &+ \sum_{s'} P(s'|s)\tilde{V}_{h+1}(s') - \sum_{s'} P(s'|s)\tilde{V}_{h+1}(s') \\ &= \sum_{s'} P(s'|s)(V_{h+1}(s') - \tilde{V}_{h+1}(s')) + \sum_{s'}(P(s'|s) - \tilde{P}(s'|s))\tilde{V}_{h+1}(s'). \end{aligned}$$

Since we have $V_{H-1}(s) = r_c(s) = \tilde{V}_{H-1}(s)$, we can recursively expand the first difference until $h = 0$ and get the result. □

Lemma 60. *Assume $p, \widehat{p}, \tilde{p} \in [0,1]$ satisfy $\widehat{p} \in \mathcal{P}$ and $\tilde{p} \in \mathcal{P}$ where*

$$\begin{aligned} \mathcal{P} := \{p' \in [0,1] : &|p - p'| \leq \sqrt{\frac{\ln 6/\delta_P}{2n}}, \\ &|p - p'| \leq \sqrt{\frac{2p(1-p)}{n}\ln(6/\delta_P)} + \frac{2}{3n}\ln(6/\delta_P), \\ &|p'(1-p') - p(1-p)| \leq \frac{2\ln(6/\delta_P)}{n-1}\}. \end{aligned}$$

Then

$$|p-\tilde{p}| \leq \sqrt{\frac{8\tilde{p}(1-\tilde{p})}{n}\ln(6/\delta_P)} + \frac{16}{3(n-1)}\ln(6/\delta_P).$$

Proof.

$$\begin{aligned}|p-\tilde{p}| &\leq |p-\widehat{p}| + |\widehat{p}-\tilde{p}| \leq 2\sqrt{\frac{2\widehat{p}(1-\widehat{p})}{n}\ln(6/\delta_P)} + 2\frac{2}{3n}\ln(6/\delta_P) \\ &\leq 2\sqrt{\frac{2}{n}\Big(\tilde{p}(1-\tilde{p}) + \frac{2\ln(6/\delta_P)}{n-1}\Big)\ln(6/\delta_P)} + \frac{4}{3n}\ln(6/\delta_P) \\ &\leq 2\sqrt{\frac{2\tilde{p}(1-\tilde{p})}{n}\ln(6\delta_P)} + 2\frac{2\ln(6/\delta_P)}{n-1} + \frac{4}{3n}\ln(6/\delta_P) \\ &\leq 2\sqrt{\frac{2\tilde{p}(1-\tilde{p})}{n}\ln(6\delta_P)} + \frac{16}{3(n-1)}\ln(6/\delta_P).\end{aligned}$$

□

Lemma 61. *Assume*

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

for $a = \pi_h(s)$ and all $s', s \in S$. Then

$$|\sum_{s'}(P(s'|s) - \tilde{P}(s'|s))\tilde{V}_{h+1}(s)| \leq c_1(s,a)|S|\|\tilde{V}_{h+1}\|_\infty + c_2(s,a)\sqrt{|S|}\tilde{\sigma}_h(s)$$

for any $(s,a) \in S \times A$.

Proof. Let s and $a = \pi_h(s)$ be fixed and define for this fixed s the constant function $\bar{V}(s') = \sum_{s''}\tilde{P}(s''|s')\tilde{V}_{h+1}(s'')$ as the expected value function of successor states of s. Note that $\bar{V}(s')$ is a constant function and so $\bar{V}(s') = \sum_{s''}\tilde{P}(s''|s')\bar{V}(s'') = \sum_{s''}P(s''|s')\bar{V}(s'')$.

$$
\begin{aligned}
&|\sum_{s'}(P(s'|s) - \tilde{P}(s'|s))\tilde{V}_{h+1}(s')| = |\sum_{s'}(P(s'|s) - \tilde{P}(s'|s))\tilde{V}_{h+1}(s') + \bar{V}(s) - \bar{V}(s)| \\
&= |\sum_{s'}(P(s'|s) - \tilde{P}(s'|s))(\tilde{V}_{h+1} - \bar{V})(s')| \\
&\leq \sum_{s'} |P(s'|s,a) - \tilde{P}(s'|s,a)||\tilde{V}_{h+1}(s') - \bar{V}(s')| \\
&\leq \sum_{s'} \Big(c_1(s,a) + c_2(s,a)\sqrt{\tilde{P}(s'|s,a)(1-\tilde{P}(s'|s,a))}\Big)|\tilde{V}_{h+1}(s') - \bar{V}(s')| \\
&\leq |S|c_1(s,a)\|\tilde{V}_{h+1}\|_\infty + c_2(s,a)\sum_{s'}\sqrt{\tilde{P}(s'|s,a)(1-\tilde{P}(s'|s,a))(\tilde{V}_{h+1}(s') - \bar{V}(s'))^2} \\
&\leq |S|c_1(s,a)\|\tilde{V}_{h+1}\|_\infty + c_2(s,a)\sqrt{|S|\sum_{s'}\tilde{P}(s'|s,a)(1-\tilde{P}(s,a))(\tilde{V}_{h+1}(s') - \bar{V}(s'))^2} \\
&\leq |S|c_1(s,a)\|\tilde{V}_{h+1}\|_\infty + c_2(s,a)\sqrt{|S|\sum_{s'}\tilde{P}(s'|s,a)(\tilde{V}_{h+1}(s') - \bar{V}(s'))^2} \\
&\leq |S|c_1(s,a)\|\tilde{V}_{h+1}\|_\infty + c_2(s,a)\sqrt{S}\tilde{\sigma}_h(s).
\end{aligned}
$$

In the first inequality, we wrote the definition of P and $\tilde{P}$ and applied the triangle inequality. We then applied the assumed bound and bounded $|\tilde{V}_{h+1}(s') - \bar{V}(s')|$ by $\|V_{h+1}\|_\infty$ as all value functions are non-negative. In fourth inequality, we applied Cauchy-Schwartz inequality and subsequently used the fact that each term in the sum is non-negative and that $(1 - \tilde{P}(s'|s,a)) \leq 1$. The final inequality follows from the definition of $\tilde{\sigma}_h$. □

Lemma 62. *Assume $M \in \mathcal{M}_k$. If $|X_{\kappa,\iota}| \leq \kappa$ for all (κ, ι), then*

$$\Delta_d := |V_{1:H}^{(d)}(s_0) - \tilde{V}^{(d)}(s_0)| \leq \hat{A}_d + \hat{B}_d + \min\{\hat{C}_d, \hat{C}'_d + \hat{C}''\sqrt{\Delta_{2d+2}}\}$$

where

$$\hat{A}_d = \frac{\epsilon}{20}(HNB_\lambda)^d, \hat{B}_d = \frac{64(HNB_\lambda)^{d+1}|\mathcal{K}\times\mathcal{I}|}{3m}\ln(6/\delta_P),$$

and

$$\hat{C}_d' = \sqrt{|\mathcal{K} \times \mathcal{I}| \frac{16}{m} (HNB_\lambda)^{2d+2} \ln{(6/\delta_P)}}, \hat{C}_d = \hat{C}_d' \sqrt{HNB_\lambda}, \hat{C}'' = \sqrt{|\mathcal{K} \times \mathcal{I}| \frac{16}{m} \ln{(6/\delta_P)}}.$$

Proof. In this lemma we assume that we are in a fixed phase k, so, we drop the index of k.

$$\begin{aligned}
\Delta_d = |V_0^{(d)}(s_0) - \tilde{V}_0^{(d)}(s_0)| &= |\sum_{h=0}^{H-2} P^{h-1}(P - \tilde{P})\tilde{V}_{h+1}^{(d)}(s_0)| \\
&\leq \sum_{h=0}^{(H-2)} P^{h-1} |(P - \tilde{P})\tilde{V}_{h+1}^{(d)}|(s_0) \\
&= \sum_{h=0}^{H-2} P^{h-1} \Big(\sum_{s,a} \mathbb{I}\{s = \cdot, a = \pi_h(s)\} |(P - \tilde{P})\tilde{V}_{h+1}^{(d)}| \Big)(s_0) \\
&= \sum_{s,a} \sum_{h=0}^{H-2} P^{h-1} \Big(\mathbb{I}\{s = \cdot, a = \pi_h(s)\} |(P - \tilde{P})\tilde{V}_{h+1}^{(d)}| \Big)(s_0) \\
&= \sum_{s,a} \sum_{h=0}^{H-2} P^{h-1} \Big(\mathbb{I}\{s = \cdot, a = \pi_h(s)\} |(P - \tilde{P})\tilde{V}_{h+1}^{(d)}(s)| \Big)(s_0)
\end{aligned}$$

The first equality follows from Lemma 59, 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^{h-1} operators. In the fifth step, we realize that $\mathbb{I}\{s = \cdot, a = \pi_h(s)\}|(P - \tilde{P})\tilde{V}_{h+1}^{(d)}(\cdot)|$ 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^{h-1}, we can write

$$
\begin{aligned}
&|V_0^{(d)}(s_0) - \tilde{V}_0^{(d)}(s_0)| = \Delta_d \leq \sum_{s,a}\sum_{h=0}^{H-2} P^{h-1}\Big(\mathbb{I}\{s=\cdot, a=\pi_h(s)\}|(P-\tilde{P})\tilde{V}_{h+1}^{(d)}(s)|\Big)(s_0) \\
&\leq \sum_{s,a\notin X}\sum_{h=0}^{H-2}\|\tilde{V}_{h+1}^{(d)}\|_\infty (P^{h-1}\mathbb{I}\{s=\cdot, a=\pi_h(s)\})(s_0) \\
&+ \sum_{s,a\in X}\sum_{h=0}^{H-2}|(P-\tilde{P})\tilde{V}_{h+1}^{(d)}(s)|(P^{h-1}\mathbb{I}\{s=\cdot, a=\pi_h(s)\})(s_0) \\
&\leq \sum_{s,a\notin X}\sum_{h=0}^{H-2}(HNB_\lambda)^{d+1}(P^{h-1}\mathbb{I}\{s=\cdot, a=\pi_h(s)\})(s_0) \\
&+ \sum_{s,a\in X}\sum_{h=0}^{H-2}|(P-\tilde{P})\tilde{V}_{h+1}^{(d)}(s)|(P^{h-1}\mathbb{I}\{s=\cdot, a=\pi_h(s)\})(s_0) \\
&\leq \sum_{s,a\notin X}\sum_{h=0}^{H-2}(HNB_\lambda)^{d+1}(P^{h-1}\mathbb{I}\{s=\cdot, a=\pi_h(s)\})(s_0) \\
&+ \sum_{s,a\in X}\sum_{h=0}^{H-2}||S|c_1(s,a)(HNB_\lambda)^{d+1} + c_2(s,a)\sqrt{|S|}\tilde{\sigma}_h^{(d)}(s,a)|(P^{h-1}\mathbb{I}\{s=\cdot, a=\pi_h(s)\})(s_0) \\
&\leq \sum_{s,a\notin X}\sum_{h=0}^{H-1}(HNB_\lambda)^{d+1}(P^{h-1}\mathbb{I}\{s=\cdot, a=\pi_h(s)\})(s_0) \\
&+ \sum_{s,a\in X}\sum_{h=0}^{H-1}||S|c_1(s,a)(HNB_\lambda)^{d+1}|(P^{h-1}\mathbb{I}\{s=\cdot, a=\pi_h(s)\})(s_0) \\
&+ \sum_{s,a\in X}\sum_{h=0}^{H-2}|c_2(s,a)\sqrt{|S|}\tilde{\sigma}_h^{(d)}(s,a)|(P^{h-1}\mathbb{I}\{s=\cdot, a=\pi_h(s)\})(s_0) \\
&\leq \sum_{s,a\notin X}(HNB_\lambda)^{d+1}w(s,a) + \sum_{s,a\in X}|S|c_1(s,a)(HNB_\lambda)^{d+1}w(s,a) \\
&+ \sum_{s,a\in X}\sqrt{|S|}c_2(s,a)\sum_{h=0}^{H-2}\tilde{\sigma}_h^{(d)}(s,a)(P^{h-1}\mathbb{I}\{s=\cdot, a=\pi_h(s)\})(s_0) \\
&\leq \sum_{s,a\notin X}(HNB_\lambda)^{d+1}w(s,a) + \sum_{s,a\in X}|S|c_1(s,a)(HNB_\lambda)^{d+1}w(s,a) \\
&+ \sum_{s,a\in X}\sqrt{|S|}c_2(s,a)\sum_{h=0}^{H-2}\tilde{\sigma}_h^{(d)}(s,a)(P^{h-1}\mathbb{I}\{s=\cdot, a=\pi_h(s)\})(s_0)
\end{aligned}
$$

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 $\|V_{h+1}^{(d)}\|_\infty \leq \|V_0^{(d)}\|_\infty \leq (HNB_\lambda)^{d+1}$. We then apply Lemma 61 and subsequently use that all terms are non-negative and the definition of $w(s,a)$. Using the assumption that $M \in \mathcal{M}$ and $\tilde{M} \in \mathcal{M}$, we can apply Lemma 60 and get that

$$c_2(s,a) = \sqrt{\frac{8}{n(s,a)} \ln \frac{6}{\delta_P}} \text{ and } c_1(s,a) = \frac{16}{3(n(s,a)-1)} \ln \frac{6}{\delta_P}.$$

Hence, we bound

$$\Delta_d \leq A(s_0) + B(s_0) + C(s_0)$$

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

$$A(s_0) = \sum_{s,a \notin X} (HNB_\lambda)^{d+1} w(s,a) \leq w_{\min} |S| (HNB_\lambda)^{d+1} \leq \frac{\epsilon (HNB_\lambda)^{d+1} |S|}{20 HNB_\lambda |S|} = \frac{\epsilon}{20} (HNB_\lambda)^d = \hat{A}_d$$

as $w(s,a) \leq w_{\min}$ for all (s,a) not in the active set and that the policy is deterministic, which implies that there are only $|S|$ nonzero w. The next term is

$$\begin{aligned} B(s_0) &= |S| \sum_{s,a \in X} w(s,a) (HNB_\lambda)^{d+1} \frac{16}{3(n(s,a)-1)} \ln \frac{6}{\delta_P} \\ &= 2 (HNB_\lambda)^{d+1} \ln \frac{6}{\delta_P} \sum_{\kappa,\iota} \sum_{s,a \in X_{\kappa,\iota}} w(s,a) \frac{16}{3(n(s,a)-1)} \\ &= \frac{32}{3} (HNB_\lambda)^{d+1} \ln \frac{6}{\delta_P} \sum_{\kappa,\iota} \sum_{s,a \in X_{\kappa,\iota}} \frac{w(s,a)}{n(s,a)} \frac{n(s,a)}{n(s,a)-1}. \end{aligned}$$

For $s, a \in X_{\kappa,\iota}$, we have $n(s,a) \geq m w(s,a) \kappa$ and so

$$\frac{w(s,a)}{n(s,a)} \leq \frac{1}{\kappa m}.$$

Further, for all relevant (s, a) pairs, we have $n(s, a) > 1$ which implies

$$B(s_0) \leq \frac{64(HNB_\lambda)^{d+1}}{3} \ln \frac{6}{\delta_P} \sum_{\kappa,\iota} \frac{|X_{\kappa,\iota}|}{\kappa m}$$

and since, $|X_{\kappa,\iota}| \leq \kappa$

$$B(s_0) \leq \frac{64(HNB_\lambda)^{d+1}|\mathcal{K} \times \mathcal{I}|}{3m} \ln \frac{6}{\delta_P} = \hat{B}_d$$

where $\mathcal{K} \times \mathcal{I}$ is the set of all possible (κ, ι) pairs. The last term is

$$\begin{aligned}
C(s_0) &= \sqrt{|S|} \sum_{s,a \in X} c_2(s,a) \sum_{h=0}^{H-2} \tilde{\sigma}_h^{(d)}(s,a) P^{h-1} \mathbb{I}\{s = \cdot, a = \pi_h(s)\}(s_0) \\
&\leq \sqrt{|S|} \sum_{s,a \in X} c_2(s,a) \sqrt{\sum_{h=0}^{H-2} P^{h-1} \mathbb{I}\{s = \cdot, a = \pi_h(s)\}} \\
&\times \sqrt{\sum_{h=0}^{H-2} \tilde{\sigma}_h^{(d),2}(s,a) P^{h-1} \mathbb{I}\{s = \cdot, a = \pi_h(s)\}(s_0)} \\
&\leq \sqrt{|S|} \sum_{s,a \in X} \sqrt{\frac{8w(s,a)}{n(s,a)} \ln \frac{6}{\delta_P} \sum_{h=0}^{H-2} \tilde{\sigma}_h^{(d),2}(s,a) P^{h-1} \mathbb{I}\{s = \cdot, a = \pi_h(s)\}(s_0)}
\end{aligned}$$

where we first applied the Cauchy-Schwartz inequality and then used the definition of $c_2(s, a)$ and

$w(s,a)$.

$$\begin{aligned}
C(s_0) &\le \sqrt{|S|}\sum_{\kappa,\iota}\sum_{s,a\in X_{\kappa,\iota}}\sqrt{\frac{8w(s,a)}{n(s,a)}\ln\frac{6}{\delta_P}\sum_{h=0}^{H-2}\tilde{\sigma}_h^{(d),2}(s,a)P^{h-1}\mathbb{I}\{s=\cdot,a=\pi_h(s)\}(s_0)} \\
&\le \sqrt{|S|}\sum_{\kappa,\iota}\sqrt{|X_{\kappa,\iota}|\sum_{s,a\in X_{\kappa,\iota}}\frac{8w(s,a)}{n(s,a)}\ln\frac{6}{\delta_P}\sum_{h=0}^{H-2}\tilde{\sigma}_h^{(d),2}(s,a)P^{h-1}\mathbb{I}\{s=\cdot,a=\pi_h(s)\}(s_0)} \\
&\le \sqrt{|S|}\sum_{\kappa,\iota}\sqrt{\sum_{s,a\in X_{\kappa,\iota}}\frac{8}{m}\ln\frac{6}{\delta_P}\sum_{h=0}^{H-2}\tilde{\sigma}_h^{(d),2}(s,a)P^{h-1}\mathbb{I}\{s=\cdot,a=\pi_h(s)\}(s_0)} \\
&\le \sqrt{|S||\mathcal{K}\times\mathcal{I}|\frac{8}{m}\ln\frac{6}{\delta_P}\sum_{s,a\in X}\sum_{h=0}^{H-2}\tilde{\sigma}_h^{(d),2}(s,a)P^{h-1}\mathbb{I}\{s=\cdot,a=\pi_h(s)\}(s_0)} \\
&\le \sqrt{|S||\mathcal{K}\times\mathcal{I}|\frac{8}{m}\ln\frac{6}{\delta_P}\sum_{s,a\in S\times A}\sum_{h=0}^{H-2}\tilde{\sigma}_h^{(d),2}(s,a)P^{h-1}\mathbb{I}\{s=\cdot,a=\pi_h(s)\}(s_0)} \\
&= \sqrt{|S||\mathcal{K}\times\mathcal{I}|\frac{8}{m}\ln\frac{6}{\delta_P}\sum_{h=0}^{H-2}P^{h-1}\tilde{\sigma}_h^{(d),2}(s_0)} \\
&\le \sqrt{|S||\mathcal{K}\times\mathcal{I}|\frac{8(HNB_\lambda)^{2d+3}\ln(6/\delta_P)}{m}} = \hat{C}_d.
\end{aligned}$$

We first split the sum and applied the Cauchy-Schwartz inequality. Then, we used again the fact that $\frac{w(s,a)}{n(s,a)} \le \frac{1}{\kappa m}$ and $|X_{\kappa,\iota}| \le \kappa$. In the fourth step, we applied Cauchy-Schwartz and the final inequality follows from $\|\tilde{\sigma}_h^{(d),2}\| \le (HNB_\lambda)^{2d+2}$ and the fact that P^{h-1} is non-expansive. Alternatively, we can rewrite the bound as:

$$\begin{aligned}
C(s_0) &\le \sqrt{|S||\mathcal{K}\times\mathcal{I}|\frac{8}{m}\ln\frac{6}{\delta_P}\sum_{h=0}^{H-2}P^{h-1}\tilde{\sigma}_h^{(d),2}(s_0)} \\
&= \sqrt{|S||\mathcal{K}\times\mathcal{I}|\frac{8}{m}\ln\frac{6}{\delta_P}\sum_{h=0}^{H-2}P^{h-1}\tilde{\sigma}_h^{(d),2}(s_0)-\tilde{P}^{h-1}\tilde{\sigma}_h^{(d),2}(s_0)+\tilde{P}^{h-1}\tilde{\sigma}_h^{(d),2}(s_0)}.
\end{aligned}$$

Next lemma 63 shows that the variance $\tilde{\Sigma}_0^{(d)}$ also satisfies the Bellman equation with local variances $\tilde{\sigma}_h^{(d),2}$. This insight allows us to bound $\sum_{h=0}^{H-2}\tilde{P}^{h-1}\tilde{\sigma}_h^{(d),2}(s_0) = \tilde{\Sigma}_0^{(d)}(s_0) \le (HNB_\lambda)^{2d+2}$. Also, note that $\tilde{\sigma}_h^{(d),2} = r_c^{(2d+2)}$ which gives us

$$\begin{aligned}
C(s_0) &\le \sqrt{|S||\mathcal{K}\times\mathcal{I}|\frac{8}{m}\ln\frac{6}{\delta_P}\Big((HNB_\lambda)^{2d+2} + \sum_{h=0}^{H-2} P^{h-1} r_c^{(2d+2)}(s_0) - \tilde{P}^{h-1} r_c^{(2d+2)}(s_0)\Big)} \\
&= \sqrt{|S||\mathcal{K}\times\mathcal{I}|\frac{8}{m}\ln\frac{6}{\delta_P}\Big((HNB_\lambda)^{2d+2} + V_0^{(2d+2)}(s_0) - \tilde{V}_0^{(2d+2)}(s_0)\Big)} \\
&\le \sqrt{|S||\mathcal{K}\times\mathcal{I}|\frac{8}{m}\ln\frac{6}{\delta_P}((HNB_\lambda)^{2d+2} + \Delta_{2d+2})} \\
&\le \sqrt{|S||\mathcal{K}\times\mathcal{I}|\frac{8}{m}\ln\frac{6}{\delta_P}(HNB_\lambda)^{2d+2}} + \sqrt{|S||\mathcal{K}\times\mathcal{I}|\frac{8}{m}\Delta_{2d+2}\ln\frac{6}{\delta_P}} = \hat{C}'_d + \hat{C}''\sqrt{\Delta_{2d+2}}.
\end{aligned}$$

□

Lemma 63. *[7] The variance of the value function defined as $\Sigma_t^\pi(s) = \mathbb{E}[(\sum_{h=t}^{H-1} r(s_h) - V_0^\pi(s))^2]$ satisfies a Bellman equation $\Sigma_t^\pi(s) = \sigma_t^{\pi^2}(s) + \sum_{s'\in S} P_\pi(s'|s)V_{t+1}^\pi(s')$ which gives $\Sigma_t^\pi(s) = \sum_{h=t}^{H}(P_\pi^{h-1}\sigma_h^{\pi^2})(s)$. Since $0 \le \Sigma_0^\pi(s) \le H^2$, it follows that $0 \le \sum_{h=0}^{H-1}(P_\pi^{h-1}\sigma_h^{\pi^2})(s) \le H^2$ for all $s \in S$.*

Proof of Lemma 17: The recursive bound from lemma 62

$$\Delta_d \le \hat{A}_d + \hat{B}_d + \hat{C}'_d + \hat{C}''\sqrt{\Delta_{2d+2}}$$

has the form $\Delta_d \le Y_d + Z\sqrt{\Delta_{2d+2}}$. Expanding this form and using the triangle inequality gives

$$\begin{aligned}
\Delta_0 &\le Y_0 + Z\Delta_2 \le Y_0 + Z\sqrt{Y_2 + Z\sqrt{\Delta_6}} \le Y_0 + Z\sqrt{Y_2} + Z^{3/2}\Delta_6^{1/4} \\
&\le Y_0 + Z\sqrt{Y_2} + Z^{3/2}Y_6^{1/4} + Z^{7/4}\Delta_1^{1/8}4 \le \ldots
\end{aligned}$$

and by doing this up to level $\gamma = \left\lceil \frac{\ln H}{2\ln 2} \right\rceil$, we obtain

$$\Delta_0 \le \sum_{d\in\mathcal{D}\setminus\{\gamma\}} Z^{\frac{2d}{2+d}} Y_d^{\frac{2}{2+d}} + Z^{\frac{2\gamma}{2+\gamma}} \Delta_\gamma^{\frac{2}{2+\gamma}}$$

where $\mathcal{D} = \{0, 2, 6, 14, ...\}$. Note that the exponent of H compared to m is larger in $\hat{C}'_d$ than in $\hat{B}_d$.

Therefore, for sufficiently large m, $\hat{C}'_d$ dominates the other term. More precisely, for

$$m \geq \frac{256H}{9}|\mathcal{K} \times \mathcal{I}| \ln \frac{6}{\delta_P} \tag{C.1}$$

we have $\hat{B}_d \leq \hat{C}'_d$. We can therefore consider $Z = \hat{C}''$ and $Y_d = 2\hat{C}'_d + \hat{A}_d$. Also, since $\hat{C}_d \geq \hat{C}'_d$, we can bound $\Delta_\gamma \leq \hat{A}_\gamma + 2\hat{C}_\gamma$. For notational simplicity, we will use auxiliary variable

$$m_1 = \frac{16|\mathcal{K} \times \mathcal{I}|H^2}{m\epsilon^2} \ln \frac{6}{\delta_P}$$

and get

$$\begin{aligned}
&Z = \hat{C}'' = \sqrt{m_1}\frac{\epsilon}{H} \text{ and}\\
&Y_d = \hat{A}_d + 2\hat{C}'_d = (1/4 + 2\sqrt{m_1}H^d\epsilon \text{ and}\\
&\Delta_\gamma \leq \hat{A}_\gamma + 2\hat{C}_\gamma = (1/4 + 2\sqrt{m_1 H})H^\gamma\epsilon.
\end{aligned}$$

Then

$$(Z^{2d}Y_d^2)^{(2+d)^{-1}} = (m_1^d\epsilon^{2d+2}(1/4 + 2\sqrt{m_1})^2)^{(2+d)^{-1}} = \epsilon(m_1^d\epsilon^d(1/4 + 2\sqrt{m_1})^2)^{(2+d)^{-1}}$$

and

$$(Z^{2\gamma}\Delta_\gamma)^{(2+\gamma)^{-1}} = \left(m_1^\gamma\epsilon^{2\gamma+2}(1/4 + 2\sqrt{m_1 H})^2\right)^{(2+\gamma)^{-1}} = \epsilon\left(m_1^\gamma\epsilon^\gamma(1/4 + 2\sqrt{m_1 H})^2\right)^{(2+\gamma)^{-1}}.$$

Putting these pieces together, we obtain

$$\begin{aligned}
\frac{\Delta_0}{\epsilon} &\le \sum_{d\in\mathcal{D}\setminus\{\gamma\}} (\epsilon m_1)^{\frac{d}{2+d}} \left(\frac{1}{4} + 2\sqrt{m_1}\right)^{\frac{2}{d+2}} + (\epsilon m_1)^{\frac{\gamma}{\gamma+2}} \left(\frac{1}{4} + 2\sqrt{Hm_1}\right)^{\frac{2}{\gamma+2}} \\
&= \frac{1}{4} + 2\sqrt{m_1} + \sum_{d\in\mathcal{D}\setminus\{0,\gamma\}} (\epsilon m_1)^{\frac{d}{2+d}} \left(\frac{1}{4} + 2\sqrt{m_1}\right)^{\frac{2}{d+2}} + (\epsilon m_1)^{\frac{\gamma}{\gamma+2}} \left(\frac{1}{4} + 2\sqrt{Hm_1}\right)^{\frac{2}{\gamma+2}} \\
&\le \frac{1}{4} + 2\sqrt{m_1} + \sum_{d\in\mathcal{D}\setminus\{0,\gamma\}} (\epsilon m_1)^{\frac{d}{2+d}} \left[\left(\frac{1}{4}\right)^{\frac{2}{d+2}} + (2\sqrt{m_1})^{\frac{2}{d+2}}\right] \\
&+ (\epsilon m_1)^{\frac{\gamma}{2+\gamma}} \left[\left(\frac{1}{4}\right)^{\frac{2}{\gamma+2}} + (2\sqrt{Hm_1})^{\frac{2}{\gamma+2}}\right]
\end{aligned}$$

where we used the fact that $(a+b)^{\phi} \le a^{\phi} + b^{\phi}$ for $a, b > 0$ and $0 < \phi < 1$. We now bound the $H^{1/(2+\gamma)}$ by using the definition of γ. Since

$$\frac{1}{2+\gamma} = \frac{2\ln 2}{4\ln 2 + \ln H} \le 2\log_H 2$$

and since $H \ge 1$, we have $H^{1/(2+\gamma)} \le 4$. Therefore,

$$\begin{aligned}
\frac{\Delta_0}{\epsilon} &\le \frac{1}{4} + 2\sqrt{m_1} + \sum_{d\in\mathcal{D}\setminus\{0,\gamma\}} (\epsilon m_1)^{\frac{d}{2+d}} \left[\left(\frac{1}{4}\right)^{\frac{2}{d+2}} + (2\sqrt{m_1})^{\frac{2}{d+2}}\right] \\
&+ (\epsilon m_1)^{\frac{\gamma}{2+\gamma}} \left[\left(\frac{1}{4}\right)^{\frac{2}{\gamma+2}} + 4(2\sqrt{m_1})^{\frac{2}{\gamma+2}}\right] \\
&\le \frac{1}{4} + 2\sqrt{m_1} + \sum_{d\in\mathcal{D}\setminus\{0\}} (\epsilon m_1)^{\frac{d}{2+d}} \left[\left(\frac{1}{4}\right)^{\frac{2}{d+2}} + (2\sqrt{m_1})^{\frac{2}{d+2}}\right] \\
&\le \frac{1}{4} + 2\sqrt{m_1} \sum_{i=1}^{\log_2 \gamma} (\epsilon m_1)^{1-2^{-i}} \left[\left(\frac{1}{4}\right)^{2^{-i}} + 4(2\sqrt{m_1})^{2^{-i}}\right] \\
&\le \frac{1}{4} + 2\sqrt{m_1} \sum_{i=1}^{\log_2 \gamma} m_1^{1-2^{-i}} \left[\left(\frac{1}{4}\right)^{2^{-i}} + 4(2\sqrt{m_1})^{2^{-i}}\right].
\end{aligned}$$

In the first inequality, we used the bound for $H^{1/(2+\gamma)}$ and in the second inequality we simplified the expression by noting that all terms are non-negative. In the next step, we re-parameterized the sum. In the final inequality, we used the assumption that $0 < \epsilon \leq 1$ and therefore $\epsilon^{1-2^{-i}} \leq 1$.

$$\frac{\Delta_0}{\epsilon} \leq \frac{1}{4} + 2\sqrt{m_1} + \frac{1}{4}\sum_{i=1}^{\log_2 \gamma} (4m_1)^{1-2^{-i}} + 4\sum_{i=1}^{\log_2 \gamma} (m_1)^{1-2^{-i}} (4m_1)^{2^{-i-1}}$$
$$\leq \frac{1}{4} + 2\sqrt{m_1} + \frac{1}{4}\sum_{i=1}^{\log_2 \gamma} (4m_1)^{1-2^{-i}} + 16\sum_{i=1}^{\log_2 \gamma} \left(\frac{m_1}{4}\right)^{1-2^{-i-1}} .$$

By requiring that

$$m_1 \leq \frac{1}{4}$$

and noting that $1 - 2^{-i} \geq 1/2$ and $1 - 2^{-i-1} \geq 3/4$ for $i \geq 1$, we can bound the expression by

$$\frac{\Delta_0}{\epsilon} \leq \frac{1}{4} + 2\sqrt{m_1} + \frac{1}{4}\log_2 \gamma \sqrt{4m_1} + 16 \log_2 \gamma \left(\frac{m_1}{4}\right)^{3/4} .$$

By requiring that $m_1 \leq 1/64$ and $m_1 \leq (2\log_2 \gamma)^{-2}$ and $m_1 \leq 1/64(\log_2 \gamma)^{-4/3}$, we can assure that $\Delta_0 \leq \epsilon$. Taking all assumptions on m_1 we made above together, we realize that

$$m_1 \leq \left(\frac{1}{8\log_2 \log_2 H}\right)^2 \leq \left(\frac{1}{8\log_2 \gamma}\right)^2$$

is sufficient for them to hold where we used $\log 2_\gamma = \log_2\left(\left\lceil \frac{1}{2}\log_2 H\right\rceil\right) \leq \log_2 \log_2 H$. This gives the following condition on m

$$m \geq 1024(\log_2 \log_2 H)^2 |\mathcal{K} \times \mathcal{I}| \frac{H^2}{\epsilon^2} \ln \frac{6}{\delta_1}$$

which is a stronger condition that one in equation C.1.

By construction of $\iota(s,a)$, we have $\iota(s,a) \leq 2\frac{H}{w_{min}} = \frac{8H^2|S|^2}{\epsilon}$. Also, $\kappa(s,a) \leq \frac{|S|mH}{mw_{min}} = \frac{4|S|^2H^2}{\epsilon}$. Therefore

$$|\mathcal{K} \times \mathcal{I}| \leq \log_2 \frac{4|S|^2H^2}{\epsilon} \log_2 \frac{8H^2|S|^2}{\epsilon} \leq \log_2^2 \frac{8H^2|S|^2}{\epsilon}$$

which let us conclude that

$$m \geq 1024\frac{H^2}{\epsilon^2}(\log_2 \log_2 H)^2 \log_2^2 \frac{8H^2|S|^2}{\epsilon} \ln \frac{6}{\delta_1}$$

is sufficient condition and thus, the statement to show holds. □

Proof of Lemma 18: By Lemma 16, we know that number of episodes where $|X_{\kappa,\iota}| > \kappa$ for some κ, ι is bounded by $6E_{\max}|S||A|m$ with probability at least $1 - \frac{\delta}{8}$. For all other episodes, we have by Lemma 17 that

$$|\tilde{V}_0^{\tilde{\pi}_k(\tilde{\lambda}^{(k)})}(s_0, \tilde{\lambda}^{(k)}) - V_0^{\tilde{\pi}_k(\tilde{\lambda}^{(k)})}(s_0, \tilde{\lambda}^{(k)})| \leq \frac{\epsilon}{5}. \tag{C.2}$$

Using Lemma 15, we get that $M \in \mathcal{M}_k$ for any episode k w.p. at least $1 - \frac{\delta}{8}$. Hence, we get the first result by applying the union bound if m satisfies

$$\begin{aligned} &m \geq 12800\frac{|S|N^2B_\lambda^2H^2}{\epsilon^2}(\log_2 \log_2 H)^2 \log_2^2\Big(\frac{8H^2|S|^2}{\epsilon}\Big) \log \frac{6}{\delta_P} \quad \text{and} \\ &m \geq \frac{30H^2}{\epsilon} \log \frac{10E_{\max}}{\delta}. \end{aligned}$$

From the definitions, we get

$$\log \frac{6}{\delta_P} = \log \frac{60|S|U_{\max}}{\delta} = \log \frac{60|S|^2|A|m}{\delta}.$$

Thus,

$$m \geq 51200\frac{|S|N^2B_\lambda^2H^2}{\epsilon^2}(\log_2 \log_2 H)^2 \log_2^2\Big(\frac{40H^2|S|^2}{\epsilon}\Big) \log \frac{60|S|^2|A|m}{\delta}.$$

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

$$m \geq 102400 \frac{|S|N^2 B_\lambda^2 H^2}{\epsilon^2} (\log_2 \log_2 H)^2 \log_2^2\Big(\frac{40H^2|S|^2}{\epsilon}\Big) \\ \times \log\Big(\frac{2048|S|^3|A|H^2}{\epsilon^2\delta}(\log_2 \log_2 H)^2 \log_2^2\Big(\frac{40H^2|S|^2}{\epsilon}\Big)\Big).$$

On the other hand,

$$E_{\max} = \log_2 |S| \log_2 \frac{40|S|H^2}{\epsilon} \leq \log_2^2 \frac{40|S|H^2}{\epsilon}$$

and

$$\log \frac{10E_{\max}}{\delta} = \log \frac{10 \log_2 |S| \log_2(40|S|H^2/\epsilon)}{\delta} \leq \log \frac{10 \log_2^2(40|S|H^2/\epsilon)}{\delta} \\ \leq \log \frac{|S|^4|A|H^2}{\epsilon\delta}.$$

Setting

$$m = 102400 \frac{|S|N^2 B_\lambda^2 H^2}{\epsilon^2} (\log_2 \log_2 H)^2 \log_2^2\Big(\frac{40H^2|S|^2}{\epsilon}\Big) \\ \times \log\Big(\frac{2048|S|^4|A|H^2}{\epsilon^2\delta}(\log_2 \log_2 H)^2 \log_2^2\Big(\frac{40H^2|S|^2}{\epsilon}\Big)\Big). \tag{C.3}$$

is therefore a valid choice for m to ensure that with probability at least $1 - \frac{\delta}{4}$, there are at most

$$6E_{\max}|S||A|m = 614400 \frac{|S|^2|A|N^2 B_\lambda^2 H^2}{\epsilon^2} (\log_2 \log_2 H)^2 \log_2^2\Big(\frac{4|S|H^2}{\epsilon}\Big) \log_2^2\Big(\frac{8H^2|S|^2}{\epsilon}\Big) \\ \times \log\Big(\frac{2048|S|^4|A|H^2}{\epsilon^2\delta}(\log_2 \log_2 H)^2 \log_2^2\Big(\frac{8H^2|S|^2}{\epsilon}\Big)\Big)$$

sub-optimal episodes. Finally the proof concludes by the definition of dual function $D(\lambda)$. □.

Proof of Lemma 19: Suppose we have n number of samples from each (s, a) which $n \geq 2592|S|^2H^2 \log 4/\delta_1$. For a given $\lambda \in [0, B_\lambda]$, we use Lemma 2 of [33] with adjustment. For any

$s \in S$, first

$$\tilde{V}_0^{\tilde{\pi}(\lambda)}(s,\lambda) \geq \tilde{V}_0^{\pi^*(\lambda)}(s,\lambda) \geq V_0^{\pi^*(\lambda)}(s,\lambda) - \sqrt{\frac{32|S|H^3N^2B_\lambda^2 \log 4/\delta_1}{n}} \tag{C.4}$$

w.p. at least $1 - 3|S|^2|A| \log 4/\delta_1$. Next

$$V_0^{\pi^*(\lambda)}(s,\lambda) \geq V_0^{\tilde{\pi}(\lambda)}(s,\lambda) \geq \tilde{V}_0^{\tilde{\pi}(\lambda)}(s,\lambda) - \sqrt{\frac{32|S|H^3N^2B_\lambda^2 \log 4/\delta_1}{n}} \tag{C.5}$$

w.p. at least $1 - 3|S|^2|A|H \log 4/\delta_1$. Combining the two inequalities (C.4) and (C.5) leads us to

$$\|\tilde{V}_0^{\tilde{\pi}(\lambda)}(\lambda) - V_0^{\pi^*}(\lambda)\|_\infty \leq \sqrt{\frac{32|S|H^3N^2B_\lambda^2 \log 4/\delta_1}{n}}$$

w.p. at least $1 - 6|S|^2|A|H\delta_1$. Now if we put $\frac{\epsilon}{5} = \sqrt{\frac{32|S|H^3N^2B_\lambda^2 \log 4/\delta_1}{n}}$ and $\frac{\delta}{4} = 6|S|^2|A|H\delta_1$, we get that $n << |S|mH$, which concludes that if we have $|S|mH$ number of samples from each (s,a), we get

$$\mathbb{P}(\|\tilde{V}_0^{\tilde{\pi}(\lambda)}(\lambda) - V_0^{\pi^*}(\lambda)\|_\infty \leq \frac{\epsilon}{5}) \geq 1 - \frac{\delta}{4}.$$

Finally the proof concludes by the definition of dual function $D(\lambda)$. □

Proof of Lemma 20 : For any $k \in [0, K]$:

$$\begin{aligned}
&\mathbb{E}[\|x_{k+1} - x^*\|^2|x_k] = \mathbb{E}[\|\Pi_{\mathcal{X}}(x_k - \alpha\tilde{g}(x_k)) - x^*\|^2|x_k] \\
&\leq \mathbb{E}[\|x_k - \alpha\tilde{g}(x_k) - x^*\|^2|x_k] = \|x_k - x^*\|^2 + \alpha^2\mathbb{E}[\|\tilde{g}\|^2|x_k] - 2\alpha\mathbb{E}[\tilde{g}^T(x_k)(x_k - x^*)|x_k] \\
&\leq \|x_k - x^*\|^2 + \alpha^2\mathbb{E}[\|\tilde{g}\|^2|x_k] - 2\alpha(g(x_k) - g(x^*))
\end{aligned}$$

First inequality is due to the fact that distance from projection of point x_{k+1} to x^* is smaller than the distance from x_{k+1} to x^*. And the last inequality yields from the the fact that $\tilde{g}$ is unbiased noisy subgradient of $g(\cdot)$. Now, we take expectation w.r.t. x_k:

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq \mathbb{E}[\|x_k - x^*\|^2] + \alpha^2 \mathbb{E}[\|\tilde{g}\|^2] - 2\alpha \mathbb{E}[g(x_k) - g(x^*)]$$

Now, we apply this procedure recursively from K to 1 and use the facts that $\|x\| \leq B_1, \forall x \in \mathcal{X}$ and $\|\tilde{g}\| \leq B_2$:

$$\mathbb{E}[\|x_{K+1} - x^*\|^2] \leq B_1^2 + K\alpha^2 B_2^2 - 2\alpha \mathbb{E}[\sum_{k=1}^{K} g(x_k) - Kg(x^*)]$$

Because $\mathbb{E}[\|x_{K+1} - x^*\|^2]$ is non-negative

$$\mathbb{E}[\frac{1}{K}\sum_{k=1}^{K} g(x_k) - g(x^*)] \leq \frac{B_1^2 + K\alpha^2 B_2^2}{2K\alpha}.$$

Since $g(\cdot)$ is a convex function, then $g(\frac{1}{K}\sum_{k=1}^{K} x_k)) \leq \frac{1}{K}\sum_{k=1}^{K} g(x_k)$. Hence,

$$\mathbb{E}[g(\frac{1}{K}\sum_{k=1}^{K} x_k) - g(x^*)] \leq \frac{B_1^2 + K\alpha^2 B_2^2}{2K\alpha}.$$

Now, if we choose $\alpha = \frac{B_1}{B_2\sqrt{K}}$, we get

$$\mathbb{E}[g(\frac{1}{K}\sum_{k=1}^{K} x_k) - g(x^*)] \leq \frac{B_1 B_2}{\sqrt{K}}. \tag{C.6}$$

□

Proof of Proposition 2 : For any given λ' and λ from $[0, B_\lambda]$

$$\begin{aligned}
\tilde{D}_k(\lambda') &= \max_{\pi, M' \in \mathcal{M}_{\lambda'}} L'_k(\pi, \lambda') \geq \tilde{V}_0^{\tilde{\pi}_k(\lambda)}(s_0) + \sum_i \lambda'_i(\bar{C}_i - \tilde{C}_{i,0}^{\tilde{\pi}_k(\lambda)}(s_0)) \\
&= \tilde{V}_0^{\tilde{\pi}_k(\lambda)}(s_0) + \sum_i \lambda_i(\bar{C}_i - \tilde{C}_{i,0}^{\tilde{\pi}_k(\lambda)}(s_0)) + \sum_i (\lambda'_i - \lambda_i)(\bar{C}_i - \tilde{C}_{i,0}^{\tilde{\pi}_k(\lambda)}(s_0)) \\
&= \tilde{D}_k(\lambda) + \sum_i (\lambda'_i - \lambda_i)(\bar{C}_i - \tilde{C}_{i,0}^{\tilde{\pi}_k(\lambda)}(s_0)).
\end{aligned}$$

It shows that the vector $[\tilde{C}_{i,0}^{\tilde{\pi}_k(\lambda)}(s_0)]_i$ is subgradient of $\tilde{D}_k(\lambda)$. Now, if we take $\mathbb{E}[\cdot|F_{\epsilon/5}]$ we have:

$$\tilde{D}(\lambda') \geq \tilde{D}(\lambda) + (\lambda' - \lambda)^T \mathbb{E}[[\bar{C}_i - \tilde{C}_{i,0}^{\tilde{\pi}_k(\lambda)}(s_0)]_i | F_{\epsilon/5}],$$

we get that $[\bar{C}_i - \tilde{C}_{i,0}^{\tilde{\pi}_k(\lambda)}(s_0)]_i$ is stochastic subgradient of $\tilde{D}(\lambda)$ by definition. Hence, we can apply Lemma 20 and get the result with K and α specified by (4.24). To certify the choice of those parameters, we provide the bound on λ and $[\tilde{C}_{i,0}^{\tilde{\pi}_k(\lambda)}(s_0)]_i$.

First, we bound λ :

$$\|\lambda\|_2 = \sqrt{\sum_{i=1}^{N} \lambda_i^2} \leq \sqrt{N} B_\lambda \leq \sqrt{N} \frac{H}{\bar{C}_{\min}}.$$

Next, we bound $[\tilde{C}_{i,0}^{\tilde{\pi}_k(\lambda)}(s_0)]_i$:

$$\|[\tilde{C}_{i,0}^{\tilde{\pi}_k(\lambda)}(s_0)]_i\|_2 = \sqrt{\sum_{i=1}^{N} (\tilde{C}_{i,0}^{\tilde{\pi}_k(\lambda)}(s_0))^2} \leq \sqrt{N}(H + \bar{C}_{\max}).$$

Substituting these bounds for B_1 and B_2 in Lemma 20 would get

$$|\tilde{D}(\tilde{\lambda}^{(k)}) - \tilde{D}(\tilde{\lambda}^*)| \leq \frac{\epsilon}{5}. \tag{C.7}$$

Now, consider the following

$$\begin{aligned}
\tilde{D}_k(\tilde{\lambda}^{(k)}) - \tilde{D}(\tilde{\lambda}^*) &= \tilde{D}_k(\tilde{\lambda}^{(k)}) - D(\tilde{\lambda}^{(k)}) + D(\tilde{\lambda}^{(k)}) - \tilde{D}(\tilde{\lambda}^{(k)}) + \tilde{D}(\tilde{\lambda}^{(k)}) - \tilde{D}(\tilde{\lambda}^*) \\
&\leq \frac{\epsilon}{5} + D(\tilde{\lambda}^{(k)}) - \tilde{D}(\tilde{\lambda}^{(k)}) + \tilde{D}(\tilde{\lambda}^{(k)}) - \tilde{D}(\tilde{\lambda}^*) \\
&\leq \frac{\epsilon}{5} + \frac{\epsilon}{5} + \tilde{D}(\tilde{\lambda}^{(k)}) - \tilde{D}(\tilde{\lambda}^*) \\
&\frac{\epsilon}{5} + \frac{\epsilon}{5} + \frac{\epsilon}{5} = \frac{3\epsilon}{5}.
\end{aligned}$$

w.p. at least $1 - \frac{\delta}{4}$. The first inequality is due to Lemma 18. The second is according to inequality

(4.29) and Lemma 19. Finally, the last line is due to (C.7). Hence, the proof is complete. □

Proof of Lemma 21: For a given $0 < \epsilon, \delta < 1$, we get that for any $\lambda \in [0, B_\lambda]$ Algorithm 6 gives

$$\mathbb{P}(|\tilde{D}(\lambda) - D(\lambda)| \leq \frac{\epsilon}{5}) \geq 1 - \frac{\delta}{4}. \tag{C.8}$$

according to Lemma 19.Now, for $\tilde{\lambda}^*$ we have

$$\tilde{D}(\tilde{\lambda}^*) \leq \tilde{D}(\lambda^*) \leq D(\lambda^*) + \frac{\epsilon}{5} \tag{C.9}$$

w.p. at least $1 - \frac{\delta}{4}$. The first inequality is due to the definition of $\tilde{\lambda}^*$, and the second inequality is true according to (C.8). Finally, consider

$$\begin{aligned}\tilde{D}_k(\tilde{\lambda}^{(k)}) - D(\lambda^*) &= \tilde{D}_k(\tilde{\lambda}^{(k)}) - \tilde{D}(\tilde{\lambda}^*) + \tilde{D}(\tilde{\lambda^*}) - D(\lambda^*) \\ &\leq \frac{3\epsilon}{5} + \frac{\epsilon}{5} = \frac{4\epsilon}{5}\end{aligned}$$

w.p. at least $1 - \frac{3\delta}{4}$ according to proposition 2 and (C.9). Hence, the proof is complete. □