Primal-Dual Algorithms for Learning in Constrained Markov Decision Processes

by

Chang Liu

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

Many real-world tasks in fields such as robotics and control can be formulated as constrained Markov decision processes (CMDPs). In CMDPs, the objective is usually to optimize the return while ensuring some constraints being satisfied at the same time. The primal-dual approach is a common technique of addressing CMDPs. It rewrites the original optimization problem of CMDPs into its equivalent Lagrangian form. In this thesis, we deliver an overview of CMDPs and the primal-dual approach, explain several algorithm designs adopting the primal-dual approach under different learning settings in terms of simulator types, and provide analysis of these algorithms.

Preface

Chapter 2 contains original contributions of the author to this thesis. The proof of lemma 2.0.2 is corrected by the author through discussions with my supervisor Csaba Szepesvári. Chapter 2 and chapter 4 contain proofs that are reorganized and explained with interpretations.

Acknowledgement

First I would like to thank my supervisor, Csaba Szepesvári, for his assistance and support in my research endeavors. I feel truly grateful of having the opportunity to work with Csaba. I deeply appreciate all the conversations with him which is always pleasant and helpful. I am also thankful of all the skiing lessons Csaba taught me.

I would also like to thank my friends. I have made many friends over the two years I stayed in the University of Alberta, and their kindness have made my stay in Edmonton joyful and colorful. It was a pleasure to be a part of the RL-theory office. I had a great time with my office mates. In particular, I would like to thank Shuai, Vlad, Wenhao, Kushagra, Johannes, Alex, Zixin, Yongchang, Jiamin, Alireza, and many others. I would also like to thank Yuxin for all the support.

Finally, I would like to thank my family for their love and support.

Contents

Chapter 1

Background

1.1 Introduction

In the recent decades, reinforcement learning (RL) has been widely applied to many real-world problems and has achieved remarkable success across domains such as robotics, gaming, business, and autonomous driving. In common RL studies, an agent interacts with an environment, which is often formulated as a Markov decision process (MDP), and the goal is to optimize an unconstrained value over a period of time, usually the expected cumulative reward in the MDP [Sutton and Barto, 2018]. However, in many real-world applications, it is often not the case that the agent can optimize its objective without any constraint such as efficiency and safety requirements. For example, when we would like to design a robot to carry out some task, it is sensible to place constraints on its power consumption. Problems of this nature, where the agent optimizes a value subject to ensuring that a second value satisfies some constraint, can be formulated as constrained Markov decision processes (CMDPs) [Altman, 1999].

There are a number of directions in the CMDP research. Some early studies assume that the model of CMDP is known to the agent, and aim to find a near-optimal policy to solve the planning problem [Borkar, 2005; Paternain et al., 2019; Achiam et al., 2017; Xu et al., 2020]. As a more general case in real-world problems, many recent studies assume the model is unknown, and therefore require learning of transitions [Vaswani et al., 2022; Wei et al., 2020; Jain et al., 2022; Wei et al., 2021; Ding et al., 2020; Tessler et al., 2019]. Many studies propose policy-based algorithms that are shown to be able to converge in experiments usually with the lack of theoretical guarantees [Yang et al., 2020; Achiam et al., 2017; Stooke et al., 2020; Tessler et al., 2019]. This line of research usually adapts mainstream RL algorithms such as Proximal Policy Optimization (PPO) [Schulman et al., 2017] into constrained setting. Another highly active direction involves proposing near-optimal algorithms that aim to minimize either regret or sample complexity in the probably approximately correct learning (PAC) setting [Brantley et al., 2020; Wachi and Sui, 2020; Vaswani et al., 2022; Jain et al., 2022; Ding et al., 2020]. This line of research develops rigorous mathematical frameworks for analyzing the theoretical guarantees of algorithms with similar designs.

Despite the difference in problem settings, the primal-dual approach is adopted by many of the aforementioned work. The primal-dual approach has a long history and can be seen as early as in Altman [1999]. It rewrites the original CMDP problem into its equivalent Lagrangian form, which is a convex-concave min-max game, or a saddle point problem. In this thesis we focus on a number of primal-dual algorithms under different settings and look into the similarities among the theoretical analysis of these algorithms. We explore CMDPs with global access and online access, and CMDPs with linear function approximations.

1.2 Markov Decision Processes

Consider an infinite-horizon discounted Markov decision process (MDP) $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$. The state space $\mathcal{S}$ and the action space $\mathcal{A}$ are sets of all states and actions. Denote the cardinality by S and A respectively if the set is finite. Each element indexed by $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ of the transition probability $P \in [0, 1]^{S \times A \times S}$ denotes the probability of transitioning from the state s to s' by taking the action a. The reward function r :

$\mathcal{S} \times \mathcal{A} \to [0,1]$ gives the expected reward $r(s,a)$ the learner receives by taking the action a in the state s. For a trajectory $\{s_0, a_0, s_1, a_1, \dots, s_t, a_t, \dots\}$, the cumulative reward, or the return R, received by the learner is defined as the discounted sum of the rewards, i.e., $R = r(s_0, a_0) + \gamma r(s_1, a_1) + \dots + \gamma^t r(s_t, a_t) + \dots$, where $\gamma \in [0, 1)$ is the discount factor. If the initial state is deterministic, we denote it by s_0, and if the initial state is stochastic, we denote the initial state distribution by ρ.

Let π denote a stationary and stochastic policy, defined as a probability distribution over the action space $\mathcal{A}$ given some state $s \in \mathcal{S}$, i.e., $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$, where $\Delta_{\mathcal{A}}$ is the set of all simplex over the action space. Given some policy π, the expected discounted return, or (reward) value function, denoted as $V_r^\pi(s_0)$, is defined as the expected cumulative discounted rewards from the initial state s_0 by following the policy π, i.e., $V_r^\pi(s_0) = \mathbb{E}[\sum_{t=0} \gamma^t r(s_t, a_t)]$, where actions and next states are sampled from the policy π and transition P respectively, i.e., $a_t \sim \pi(\cdot|s_t)$ and $s_{t+1} \sim P(\cdot|s_t, a_t)$. Given some policy π, for any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, the reward action-value function, defined as $Q_r^\pi(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, satisfies $V_r^\pi(s) = \langle \pi(\cdot|s), Q_r^\pi(s, \cdot) \rangle = \sum_a \pi(a|s) Q_r^\pi(s,a)$, where $V_r^\pi(s)$ is the value function when $s_0 = s$.

Fix a policy π. Define $P_\pi \in [0,1]^{S \times S}$ as the probability matrix induced by the policy π. For each entry with index $(s, s') \in \mathcal{S} \times \mathcal{S}$, define $P_\pi(s, s') := \sum_{a \in \mathcal{A}} \pi(a|s) P(s'|s,a)$. Similarly for rewards, define $r_\pi \in [0,1]^S$ as a vector, and for the sth element, define $r_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) r(s,a)$ as the expected reward in the state s induced by the policy π. The above notations and definitions make it convenient to introduce the following affine linear operator.

Define the Bellman operator (also called policy evaluation operator [Szepesvári, 2023c]) T_π induced by the policy π as an affine linear operator $T_\pi : \mathbb{R}^S \to \mathbb{R}^S$,

$$T_\pi v = r_\pi + \gamma P_\pi v, \tag{1.1}$$

where v is an arbitrary vector in $\mathbb{R}^d$. Rewrite eq. (1.1) in an element-wise manner, equivalently we also have for any $s \in \mathcal{S}$,

$$T_\pi v(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left[r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) v(s') \right]. \tag{1.2}$$

Let the vector v in eq. (1.1) be V^π. Then we have $T_\pi V^\pi(s) = \sum_a \pi(a|s)[r(s,a)+\gamma \sum_{s'} P(s'|s,a)V^\pi(s')]$, where the right hand side is exactly $V^\pi(s)$ by Bellman equation. Hence, V^π is the fixed point for the Bellman operator T_π:

$$V^\pi = T_\pi V^\pi. \tag{1.3}$$

Now we have sufficient tools to introduce the performance difference lemma:

Lemma 1.2.1 (Performance difference lemma)**.** *For any value function V^π, and any two memoryless policies π and π',*

$$V^{\pi'} - V^\pi = \left(I - \gamma P_{\pi'}\right)^{-1} \left(T_{\pi'} V^\pi - V^\pi\right).$$

Proof. Note that $V^{\pi'} = \left(I - \gamma P_{\pi'}\right)^{-1} r_{\pi'}$. Hence,

$$\begin{aligned}
V^{\pi'} - V^\pi &= \left(I - \gamma P_{\pi'}\right)^{-1} r_{\pi'} - V^\pi \\
&= \left(I - \gamma P_{\pi'}\right)^{-1} \left(r_{\pi'} - \left(I - \gamma P_{\pi'}\right) V^\pi\right) \\
&= \left(I - \gamma P_{\pi'}\right)^{-1} \left(r_{\pi'} + \gamma P_{\pi'} V^\pi - V^\pi\right) \\
&= \left(I - \gamma P_{\pi'}\right)^{-1} \left[T_{\pi'} V^\pi - V^\pi\right].
\end{aligned}$$

□

The performance difference lemma (or value difference lemma), proposed in Kakade [2003]; Kakade and Langford [2002], is widely used and takes many different yet equivalent forms. lemma 1.2.1 was rediscovered by Szepesvári [2023d]. It provides a method to measure the difference in value by following different policies. Here we introduce an equivalent form of the performance difference lemma. First, we introduce the notion of advantage:

Definition 1.2.1 (Advantage)**.** The advantage of taking action a at state s over some policy π is defined as

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s).$$

The advantage of a deterministic policy π' over another policy π is defined as

$$A^{\pi}\left(s,\pi'\right) = A^{\pi}\left(s,\pi'(s)\right) = Q^{\pi}(s,\pi'(s)) - V^{\pi}(s).$$

Note that the action-value function $Q^{\pi}(s,a)$ is defined such that for any $s \in \mathcal{S}$, $V^{\pi}(s) = \langle \pi(\cdot|s), Q^{\pi}(s,\cdot)\rangle$. Therefore, the advantage of policy π' over π can also be written as:

$$\begin{aligned} A^{\pi}(s,\pi') = \langle A^{\pi}(s,\cdot), \pi'(\cdot|s)\rangle &= \langle Q^{\pi}(s,\cdot), \pi'(\cdot|s)\rangle - \langle Q^{\pi}(s,\cdot), \pi(\cdot|s)\rangle \\ &= \langle Q^{\pi}(s,\cdot), (\pi'(\cdot|s) - \pi(\cdot|s))\rangle. \end{aligned}$$

Now we show the performance difference lemma:

Lemma 1.2.2 (Performance difference lemma [Kakade and Langford, 2002])**.** *For any π, π', and any state $s \in \mathcal{S}$,*

$$V^{\pi'}(s) - V^{\pi}(s) = \frac{1}{1-\gamma}\mathbb{E}_{\pi'}\left[A^{\pi}\left(s',\pi'\right)\right],$$

1.3 Simulator modes and access types

We will assume the learner interacts with a simulator to learn about the underlying MDP before returning a policy. Such simulator is provided to the learner together with an application interface for querying. By different access modes, the simulators can be categorized into three types: global access, local access, and online access [Szepesvári, 2023a; Yin et al., 2022].

Global access, also referred to as generative model or random access in some works [Agarwal et al., 2020; Yin et al., 2022], assumes that all state-action pairs $(s,a) \in \mathcal{S} \times \mathcal{A}$ are made accessible to the learner, and the simulator can be queried with any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. When it is queried so, the simulator will return a random next state s'

sampled from $P(\cdot|s,a)$ and the corresponding reward r.

Local access assumes that the simulator can be queried with states s that were previously observed by the learner with any actions $a \in \mathcal{A}$. When it is queried so, the simulator will return a random next state s' sampled from $P(\cdot|s,a)$ and the corresponding reward r. Local access simulators can be implemented with checkpointing by maintaining data structures that keep track of the visited states.

Online access assumes that the simulator maintains the current state, and can only be either reset to the initial state (or a random sample from the initial state distribution) or transitioned from the current state to the next state when given an action $a \in \mathcal{A}$ by the learner. The associated reward r is revealed to the learner when transitions occur.

1.4 Constrained Markov Decision Processes

Consider an infinite-horizon discounted constrained Markov decision process (CMDP) $M = (\mathcal{S}, \mathcal{A}, P, r, c, b, \gamma)$, where $\mathcal{S}$ is a state space, $\mathcal{A}$ is an action space, P is the transition probability matrix, $r : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the reward function, and $c : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the constraint reward function (also known as utility function), $b \in (0,1)$ is the constraint constant, and $\gamma \in [0,1)$ is the discount factor. In this work, we assume the initial state is deterministic and we will denote it by s_0. Analogous to rewards r, we define V_c and Q_c.

The learner is allowed to interact with the CMDP for some iterations by means of some access type introduced in section 1.3 before the learner is required to return a policy π that is expected to work well on the CMDP. The objective of a learner in CMDP is to return a policy π such that the policy maximizes the reward value function $V_r^\pi(s_0)$, while ensures the requirement on the constraint value function $V_c^\pi(s_0)$, i.e.,

$$\max_\pi V_r^\pi(s_0), \quad \text{s.t.,} \quad V_c^\pi(s_0) \geq b. \tag{1.4}$$

Let π^* denote the optimal stochastic policy for the above CMDP, $V_r^*(s_0)$ be the value function of the optimal policy π^*, and $V_c^*(s_0)$ be the constraint value function of the optimal policy π^*.

The following assumptions and definitions are also useful and oftentimes needed in CMDPs.

Assumption 1.4.1 (Slater condition)**.** *There exists $\gamma > 0$ and a policy π such that $V_c^\pi(s_0) \geq b + \gamma$. In this case, we say the CMDP is feasible with slack $\geq \gamma$.*

Definition 1.4.1 (Slater constant)**.** For a given CMDP M, define the Slater constant $\zeta := \max_\pi V_c^\pi(s_0) - b$ to be a problem-dependent constant.

The Slater condition assumes that there is some slack between the constraint value function V_c^π the learner can achieve and the constraint constant b, and in this sense the Slater constant is a quantity that measures the size of the slack and thus the difficulty of solving the CMDP in eq. (1.4). Note that by definition the Slater constant satisfies that $\zeta \geq \gamma$.

1.5 Primal Dual Approach

The Lagrangian function for the optimization problem eq. (1.4) is the following saddle-point problem

$$\max_{\pi} \min_{\lambda \geq 0} V_r^\pi(s_0) + \lambda(V_c^\pi(s_0) - b), \tag{1.5}$$

where λ is the Lagrange multiplier or the dual variable for the constraint. By Lagrangian duality, (π^*, λ^*) is a solution to the saddle-point problem eq. (1.5), where π^* is the optimal policy for the CMDP and λ^* is the optimal Lagrange multiplier.

The existing approaches [Jain et al., 2022; Vaswani et al., 2022] to solving the above primal-dual saddle-point problem is to solve it iteratively, by alternatively updating the policy (primal variable) and the Lagrange multiplier (dual variable). Let T be the total number of iterations before returning the estimate solutions, and let π_t and λ_t be the primal and dual iterates respectively for some iteration $t \in [T]$.

Chapter 2

Learning in Tabular CMDPs with Global Access

In this chapter, we consider infinite-horizon discounted tabular CMDPs under the global access setting. This chapter also includes the original work of the author (lemma 2.0.2). Let $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, c, b, \gamma)$ be the CMDP, where $\mathcal{S}$ is a state space, $\mathcal{A}$ is an action space, $\mathcal{P}$ is the transition probability matrix, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, and $c : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the constraint reward function, and $\gamma \in [0, 1)$ is the discount factor. Recall that with global access model, the learner can query the simulator with arbitrary state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ to learn about the MDP dynamics. To solve the CMDP in eq. (1.4) under the global access setting, it is natural to adopt the model-based approach. With model-based approach, the learner explicitly calculates and maintains an empirical estimate model and plans with respect to the empirical model. In the tabular case, such an empirical model can be calculated by querying the simulator with every state-action pair multiple times so that it is a good estimate of the true underlying MDP with high probability.

Let $\widehat{M}$ be the empirical model maintained by the learner. Assume the reward and constraint functions are deterministic and known to the learner. Hence the empirical model is essentially the empirical estimate $\widehat{\mathcal{P}}$ of the true transition probability matrix $\mathcal{P}$. To calculate $\widehat{\mathcal{P}}$ and thus instantiate $\widehat{M}$, for each state action pair (s, a), query the simulator N times with (s, a) and collect N random samples of next states. Then for each $s' \in \mathcal{S}$, let $\widehat{\mathcal{P}}(s'|s, a) = \frac{N(s'|s,a)}{N}$, where $N(s'|s, a)$ is the number of s' appearing.

For some technical reason to be explained later, we introduce perturbation to rewards. For each $s \in \mathcal{S}$ and $a \in \mathcal{A}$, define the perturbed rewards $r_p(s, a) = r(s, a) + \xi(s, a)$, where $\xi(s, a) \sim \mathcal{U}[0, \omega]$ are i.i.d. uniform random variable and ω is some parameter to be determined later. Further, let the constraint of the empirical CMDP equal to b'. By setting b' to different values, we can either loosen or tighten the constraint. In particular, setting $b' > b$ means tightening the constraint and $b' < b$ means loosening the constraint. Now, we can define the empirical CMDP $\widehat{M}$ as the tuple $\langle \mathcal{S}, \mathcal{A}, \widehat{\mathcal{P}}, r_p, c, b', s_0, \gamma \rangle$. For the empirical CMDP $\widehat{M}$, let $\widehat{V}^{\pi}_{r_p}(s_0)$ and $\widehat{V}^{\pi}_{c}(s_0)$ be the reward value function and constraint value function for some policy π. To instantiate $\mathcal{M}$, it remains to set the values of ω and b', which is left to the next section.

With the empirical CMDP $\mathcal{M}$ instantiated, the learner needs to solve the following problem

$$\widehat{\pi}^* \in \arg\max \widehat{V}^{\pi}_{r_p}(s_0) \quad \text{ s.t. } \quad \widehat{V}^{\pi}_{c}(s_0) \geq b'. \tag{2.1}$$

As discussed at the beginning of section 1.5, we write the above eq. (2.1) as the equivalent saddle-point problem

$$\max_{\pi} \min_{\lambda \geq 0} \left[\widehat{V}^{\pi}_{r_p}(s_0) + \lambda \left(\widehat{V}^{\pi}_{c}(s_0) - b' \right) \right]. \tag{2.2}$$

Let $(\widehat{\pi}^*, \widehat{\lambda}^*)$ denote the solution to the empirical CMDP eq. (2.1), where $\widehat{\pi}^*$ is the optimal empirical policy and $\widehat{\lambda}^*$ is the optimal Lagrange multiplier. Then again, as discussed in section 1.5, we solve the above problem eq. (2.2) iteratively, by alternatively updating the primal variable and the dual variable, denoted as $\widehat{\pi}_t$ and $\widehat{\lambda}_t$ respectively for some iteration t. Given the dual variable $\widehat{\lambda}_t$ at iteration t, the primal variable $\widehat{\pi}_t$ is updated as:

$$\widehat{\pi}_t = \arg\max_{\pi} \widehat{V}^{\pi}_{r_p}(s_0) + \widehat{\lambda}_t \widehat{V}^{\pi}_{c}(s_0). \tag{2.3}$$

Similar to Paternain et al. [2019]; Jain et al. [2022], the dual variable $\widehat{\lambda}_t$s are updated with gradient descent steps and projection onto a feasible range. However here, after projection, the dual variables are further rounded to the closest element in an epsilon-net $\Lambda = \{0, \varepsilon_1, 2\varepsilon_2, \dots, U\}$ with resolution ε_1, which will be shown to help with the concentration for all $\lambda \in \Lambda$. The parameter U will be determined later. The dual variable $\widehat{\lambda}_t$ is updated as:

$$\widehat{\lambda}_t = \mathcal{R}_\Lambda[\mathbb{P}_{[0,U]}[\widehat{\lambda}_t - \eta(\widehat{V}_c^{\widehat{\pi}_t}(s_0) - b')]], \tag{2.4}$$

where $\mathbb{P}_{[0,U]}[\lambda] = \arg\min_{p \in [0,U]} |\lambda - p|$ is the projection function that projects $\lambda \in \mathbb{R}$ onto the $[0, U]$ interval, and $\mathcal{R}_\Lambda[\lambda] = \arg\min_{p \in \Lambda} |\lambda - p|$ is the rounding function that rounds $\lambda \in \mathbb{R}$ to the closest element in the epsilon-net Λ.

Putting everything together, the model-based algorithm is described in line 1.

Algorithm 1: Model-based algorithm for CMDPs with generative model [Vaswani et al., 2022]

```
  Input  : $\mathcal{S}, \mathcal{A}, r, c, \zeta, N, b', \omega, U, \varepsilon_1, T, \widehat{\lambda}_0 = 0$.
1 For each state-action $(s, a)$ pair, collect $N$ samples from $\mathcal{P}(.|s, a)$ and form $\widehat{\mathcal{P}}$
2 Perturb the rewards to form vector $r_p(s, a) = r(s, a) + \xi(s, a)$ where $\xi(s, a) \sim \mathcal{U}[0, \omega]$
3 Form the empirical CMDP $\widehat{M} = \langle \mathcal{S}, \mathcal{A}, \widehat{\mathcal{P}}, r_p, c, b', s_0, \gamma \rangle$
4 Form the epsilon-net $\Lambda = \{0, \varepsilon_1, 2\varepsilon_1, \dots, U\}$
5 for $t \leftarrow 0 \dots T - 1$ do
6     Update the policy by solving an unconstrained MDP: $\widehat{\pi}_t = \arg\max \widehat{V}^{\pi}_{r_p + \widehat{\lambda}_t c}$
7     Update the dual-variables: $\widehat{\lambda}_{t+1} = \mathcal{R}_\Lambda[\mathbb{P}_{[0,U]}[\widehat{\lambda}_t - \eta(\widehat{V}_c^{\widehat{\pi}_t}(s_0) - b')]]$
  Output : Mixture policy $\bar{\pi}_T = \frac{1}{T} \sum_{t=0}^{T-1} \widehat{\pi}_t$.
```

Before analyzing the sample complexity of line 1, we introduce some preliminary results:

Theorem 2.0.1. *(Guarantees for the primal-dual algorithm (Theorem 1 in Vaswani et al. [2022])) For a target error $\varepsilon_{opt} > 0$ and the primal-dual updates in eq. (2.3) and eq. (2.4) with $U > \left|\widehat{\lambda}^*\right|$, $T = \frac{4U^2}{\varepsilon_{opt}^2(1-\gamma)^2}[1 + \frac{1}{(U-\widehat{\lambda}^*)^2}]$, $\eta = \frac{U(1-\gamma)}{\sqrt{T}}$ and $\varepsilon_1 = \frac{\varepsilon_{opt}^2(1-\gamma)^2(U-\widehat{\lambda}^*)}{6U}$, the mixture policy $\bar{\pi}_T = \frac{1}{T}\sum_{t=0}^{T-1} \widehat{\pi}_t$ satisfies*

$$\widehat{V}_{r_p}^{\bar{\pi}_T}(s_0) \geq \widehat{V}_{r_p}^{\widehat{\pi}^*}(s_0) - \varepsilon_{opt} \quad \textit{and} \quad \widehat{V}_c^{\bar{\pi}_T}(s_0) \geq b' - \varepsilon_{opt}.$$

Proof. Similar to Jain et al. [2022], define the dual regret w.r.t. some λ as:

$$R^d(\lambda, T) := \sum_{t=0}^{T-1} \left(\widehat{\lambda}_t - \lambda\right) \left(\widehat{V}_c^{\widehat{\pi}_t}(s_0) - b'\right) \tag{2.5}$$

By primal update in eq. (2.3), for any iteration t, the primal variable $\widehat{\pi}_t$ is the maximizer of $\widehat{V}_{r_p}^{\pi}(s_0) + \widehat{\lambda}_t \widehat{V}_c^{\pi}(s_0)$. Hence, for any π and some t,

$$\widehat{V}_{r_p}^{\widehat{\pi}_t}(s_0) + \widehat{\lambda}_t \widehat{V}_c^{\widehat{\pi}_t}(s_0) \geq \widehat{V}_{r_p}^{\pi}(s_0) + \widehat{\lambda}_t \widehat{V}_c^{\pi}(s_0).$$

Let $\pi = \widehat{\pi}^*$, and note that $\widehat{V}_c^{\widehat{\pi}^*}(s_0) \geq b'$ since $\widehat{\pi}^*$ is a solution to the CMDP $\widehat{M}$,

$$\widehat{V}_{r_p}^{\widehat{\pi}^*}(s_0) - \widehat{V}_{r_p}^{\widehat{\pi}_t}(s_0) \leq \widehat{\lambda}_t \left[\widehat{V}_c^{\widehat{\pi}_t}(s_0) - b'\right]. \tag{2.6}$$

Summing eq. (2.6) over all $t \in \{0, \dots, T-1\}$, dividing by T, and using the definition of the dual regret in eq. (2.5), we have

$$\frac{1}{T} \sum_{t=0}^{T-1} \left[\widehat{V}_{r_p}^{\widehat{\pi}^*}(s_0) - \widehat{V}_{r_p}^{\widehat{\pi}_t}(s_0)\right] + \frac{\lambda}{T} \sum_{t=0}^{T-1} \left(b' - \widehat{V}_c^{\widehat{\pi}_t}(s_0)\right) \leq \frac{R^d(\lambda, T)}{T}.$$

Note that $\bar{\pi}_T = \frac{1}{T} \sum_{t=0}^{T-1} \widehat{\pi}_t$ is the mixture policy, and thus for any reward function l, $\frac{1}{T} \sum_{t=0}^{T-1} \widehat{V}_l^{\widehat{\pi}_t}(s_0) = \widehat{V}_l^{\bar{\pi}_T}(s_0)$.

$$\left[\widehat{V}_{r_p}^{\widehat{\pi}^*}(s_0) - \widehat{V}_{r_p}^{\bar{\pi}_T}(s_0)\right] + \lambda \left(b' - \widehat{V}_c^{\bar{\pi}_T}(s_0)\right) \leq \frac{R^d(\lambda, T)}{T}. \tag{2.7}$$

Then we bound the dual regret $R^d(\lambda, T)$ for some $\lambda \in [0, U]$. Define $\widehat{\lambda}'_{t+1} = \mathbb{P}_{[0,U]}[\widehat{\lambda}_t -$

$\eta(\widehat{V}_c^{\widehat{\pi}_t}(s_0) - b')]$, then

$$\begin{aligned}
\left|\widehat{\lambda}_{t+1} - \lambda\right| &= \left|\mathcal{R}_\Lambda[\widehat{\lambda}'_{t+1}] - \lambda\right| \\
&\leq \left|\mathcal{R}_\Lambda[\widehat{\lambda}'_{t+1}] - \widehat{\lambda}'_{t+1}\right| + \left|\widehat{\lambda}'_{t+1} - \lambda\right| \\
&\leq \varepsilon_1 + \left|\widehat{\lambda}'_{t+1} - \lambda\right|,
\end{aligned}$$

where the first inequality is due to triangle inequality and the second inequality is due to the property of epsilon-net. Squaring both sides,

$$\begin{aligned}
&\left|\widehat{\lambda}_{t+1} - \lambda\right|^2 \\
\leq&\varepsilon_1^2 + \left|\widehat{\lambda}'_{t+1} - \lambda\right|^2 + 2\varepsilon_1 \left|\widehat{\lambda}'_{t+1} - \lambda\right| \\
\leq&\varepsilon_1^2 + 2\varepsilon_1 U + \left|\widehat{\lambda}'_{t+1} - \lambda\right|^2 \\
\leq&\varepsilon_1^2 + 2\varepsilon_1 U + \left|\widehat{\lambda}_t - \eta\left(\widehat{V}_c^{\widehat{\pi}_t}(s_0) - b'\right) - \lambda\right|^2 \\
=&\varepsilon_1^2 + 2\varepsilon_1 U + \left|\widehat{\lambda}_t - \lambda\right|^2 - 2\eta\left(\widehat{\lambda}_t - \lambda\right)\left(\widehat{V}_c^{\widehat{\pi}_t}(s_0) - b'\right) + \eta^2\left(\widehat{V}_c^{\widehat{\pi}_t}(s_0) - b'\right)^2 \\
\leq&\varepsilon_1^2 + 2\varepsilon_1 U + \left|\widehat{\lambda}_t - \lambda\right|^2 - 2\eta\left(\widehat{\lambda}_t - \lambda\right)\left(\widehat{V}_c^{\widehat{\pi}_t}(s_0) - b'\right) + \frac{\eta^2}{(1-\gamma)^2},
\end{aligned}$$

where the first inequality follows directly from above, the second inequality follows because $\lambda \in [0, U]$ and $\widehat{\lambda}'_{t+1}$ is a projection onto $[0, U]$, the third inequality follows because projections are non-expansive, and finally the last inequality follows because for any policy π the constraint value function $\widehat{V}_c^{\pi}(s_0)$ is non-negative and bounded by $1/(1-\gamma)$. Rearranging and dividing by 2η, we get

$$\left(\widehat{\lambda}_t - \lambda\right)\left(\widehat{V}_c^{\widehat{\pi}_t}(s_0) - b'\right) \leq \frac{\varepsilon_1^2 + 2\varepsilon_1 U}{2\eta} + \frac{\left|\widehat{\lambda}_t - \lambda\right|^2 - \left|\widehat{\lambda}_{t+1} - \lambda\right|^2}{2\eta} + \frac{\eta}{2(1-\gamma)^2}.$$

Summing both sides from $t = 0$ to $T - 1$ and noticing that the left side gives the dual regret by definition,

$$\begin{aligned}
R^d(\lambda, T) &\leq T\frac{\varepsilon_1^2 + 2\varepsilon_1 U}{2\eta} + \frac{1}{2\eta}\sum_{t=0}^{T-1}\left[\left|\widehat{\lambda}_t - \lambda\right|^2 - \left|\widehat{\lambda}_{t+1} - \lambda\right|^2\right] + \frac{\eta T}{2(1-\gamma)^2} \\
&\leq T\frac{\varepsilon_1^2 + 2\varepsilon_1 U}{2\eta} + \frac{\left|\widehat{\lambda}_0 - \lambda\right|^2 - \left|\widehat{\lambda}_T - \lambda\right|^2}{2\eta} + \frac{\eta T}{2(1-\gamma)^2} \\
&\leq T\frac{\varepsilon_1^2 + 2\varepsilon_1 U}{2\eta} + \frac{\left|\widehat{\lambda}_0 - \lambda\right|^2}{2\eta} + \frac{\eta T}{2(1-\gamma)^2} \\
&\leq T\frac{\varepsilon_1^2 + 2\varepsilon_1 U}{2\eta} + \frac{U^2}{2\eta} + \frac{\eta T}{2(1-\gamma)^2}.
\end{aligned}$$

Next, we show that assigning η some appropriate value gives us the results claimed in theorem 2.0.1. In particular, setting $\eta = \frac{U(1-\gamma)}{\sqrt{T}}$, we have for any $\lambda \in [0, U]$,

$$R^d(\lambda, T) \leq T^{3/2}\frac{\varepsilon_1^2 + 2\varepsilon_1 U}{2U(1-\gamma)} + \frac{U\sqrt{T}}{1-\gamma}. \tag{2.8}$$

Since $\lambda \in [0, U]$, let $\lambda = 0$, and recall eq. (2.7), we have

$$\begin{aligned}
\widehat{V}_{r_p}^{\widehat{\pi}^*}(s_0) - \widehat{V}_{r_p}^{\bar{\pi}_T}(s_0) &\leq \sqrt{T}\frac{\varepsilon_1^2 + 2\varepsilon_1 U}{2U(1-\gamma)} + \frac{U}{(1-\gamma)\sqrt{T}} \\
&< \sqrt{T}\frac{3\varepsilon_1}{2(1-\gamma)} + \frac{U}{(1-\gamma)\sqrt{T}},
\end{aligned}$$

where the second inequality follows because $\varepsilon_1 < U$. For the constraint violation, there are two cases. The first case is when $b' - \widehat{V}_c^{\bar{\pi}_T}(s_0) \leq 0$. In this case, it also holds that $b' - \varepsilon_{\text{opt}} - \widehat{V}_c^{\bar{\pi}_T}(s_0) \leq 0$, which is what we wanted to show. The second case is when $b' - \widehat{V}_c^{\bar{\pi}_T}(s_0) > 0$. In this case, using the notation $[x]_+ = \max\{x, 0\}$, we have

$$\left[\widehat{V}_{r_p}^{\widehat{\pi}^*}(s_0) - \widehat{V}_{r_p}^{\bar{\pi}_T}(s_0)\right] + U\left[b' - \widehat{V}_c^{\bar{\pi}_T}(s_0)\right]_+ \leq \frac{R^d(U, T)}{T}.$$

Because by assumption it holds that $U > \widehat{\lambda}^*$, lemma 2.0.2 is applicable and gives that

$$\left[b' - \widehat{V}_c^{\bar{\pi}_T}(s_0)\right]_+ \leq \frac{R^d(U,T)}{T\left(U - \widehat{\lambda}^*\right)}.$$

Hence, since $U \in [0, U]$, combining the above display with eq. (2.8) gives

$$\begin{aligned}\left[b' - \widehat{V}_c^{\bar{\pi}_T}(s_0)\right] &\leq \left[b' - \widehat{V}_c^{\bar{\pi}_T}(s_0)\right]_+ \\ &\leq \sqrt{T}\frac{\varepsilon_1^2 + 2\varepsilon_1 U}{2U(1-\gamma)\left(U - \widehat{\lambda}^*\right)} + \frac{U}{\left(U - \widehat{\lambda}^*\right)(1-\gamma)\sqrt{T}} \\ &< \sqrt{T}\frac{3\varepsilon_1}{2(1-\gamma)\left(U - \widehat{\lambda}^*\right)} + \frac{U}{\left(U - \widehat{\lambda}^*\right)(1-\gamma)\sqrt{T}},\end{aligned}$$

where the third inequality follows because because $\varepsilon_1 < U$. Set T such that the second term in both quantities is bounded from above by $\frac{\varepsilon_{\text{opt}}}{2}$. This gives

$$T = T_0 := \frac{4U^2}{\varepsilon_{\text{opt}}^2(1-\gamma)^2}\left[1 + \frac{1}{\left(U - \widehat{\lambda}^*\right)^2}\right]$$

With $T = T_0$, the above expressions can be simplified as follows:

$$\begin{aligned}\widehat{V}_{r_p}^{\widehat{\pi}^*}(s_0) - \widehat{V}_{r_p}^{\bar{\pi}_T}(s_0) &\leq \frac{2U}{(1-\gamma)\varepsilon_{\text{opt}}}\left(1 + \frac{1}{U - \widehat{\lambda}^*}\right)\frac{3\varepsilon_1}{2(1-\gamma)} + \frac{\varepsilon_{\text{opt}}}{2}, \\ \left[b' - \widehat{V}_c^{\bar{\pi}_T}(s_0)\right] &\leq \frac{2U}{(1-\gamma)\varepsilon_{\text{opt}}}\left(1 + \frac{1}{U - \widehat{\lambda}^*}\right)\frac{3\varepsilon_1}{2(1-\gamma)\left(U - \widehat{\lambda}^*\right)} + \frac{\varepsilon_{\text{opt}}}{2}.\end{aligned}$$

Now, set ε_1 such that the first term in both quantities is also bounded from above by $\frac{\varepsilon_{\text{opt}}}{2}$. For this, choose

$$\varepsilon_1 = \frac{\varepsilon_{\text{opt}}^2(1-\gamma)^2\left(U - \widehat{\lambda}^*\right)}{6U}$$

With these values, the algorithm line 1 ensures that

$$\widehat{V}_{r_p}^{\widehat{\pi}^*}(s_0) - \widehat{V}_{r_p}^{\bar{\pi}_T}(s_0) \le \varepsilon_{\text{opt}} \quad \text{and} \quad b' - \widehat{V}_c^{\bar{\pi}_T}(s_0) \le \varepsilon_{\text{opt}}.$$

□

Lemma 2.0.1. *(Lemma B.2 of Jain et al. [2022]) For any $C > \lambda^*$ and any $\widetilde{\pi}$ s.t. $V_r^{\pi^*}(s_0) - V_r^{\widetilde{\pi}}(s_0) +\ C\left[b - V_c^{\widetilde{\pi}}(s_0)\right]_+ \le \beta$, we have $\left[b - V_c^{\widetilde{\pi}}(s_0)\right]_+ \le \frac{\beta}{C-\lambda^*}$.*

Proof. Define $\nu(\tau) = \max_\pi \left\{V_r^\pi(s_0) \mid V_c^\pi(s_0) \ge b + \tau\right\}$ and note that by definition, $\nu(0) = V_r^{\pi^*}(s_0)$ and that ν is a decreasing function for its argument. Let $V_l^{\pi,\lambda}(s_0) = V_r^\pi(s_0) + \lambda\left(V_c^\pi(s_0) - b\right)$. Then, for any policy π s.t. $V_c^\pi(s_0) \ge b + \tau$, we have

$$\begin{aligned} V_l^{\pi,\lambda^*}(s_0) &\le \max_{\pi'} V_l^{\pi',\lambda^*}(s_0) \\ &= V_r^{\pi^*}(s_0) \\ &= \nu(0). \end{aligned}$$

Then we have

$$\begin{aligned} \nu(0) - \tau\lambda^* &\ge V_l^{\pi,\lambda^*}(s_0) - \tau\lambda^* \\ &= V_r^\pi(s_0) + \lambda^*(V_c^\pi(s_0) - b - \tau) \\ &\ge V_r^\pi(s_0), \end{aligned}$$

where the last inequality follows because $V_c^\pi(s_0) \ge b + \tau$. Note that the above inequality holds for any policy π such that $V_c^\pi(s_0) \ge b + \tau$, thus we have

$$\nu(0) - \tau\lambda^* \ge \max_\pi \{V_r^\pi(s_0) | V_c^\pi(s_0) \ge b + \tau\} = \nu(\tau),$$

$$\tau\lambda^* \le \nu(0) - \nu(\tau). \tag{2.9}$$

Recall the notation $[x]_+ = \max\{x, 0\}$, and choose $\widetilde{\tau} = -\left(b - V_c^{\widetilde{\pi}}(s_0)\right)_+$,

$$\begin{aligned}
(C - \lambda^*)\,|\widetilde{\tau}| &= \lambda^* \widetilde{\tau} + C|\widetilde{\tau}| \\
&\leq \nu(0) - \nu(\widetilde{\tau}) + C|\widetilde{\tau}| \\
&= V_r^{\pi^*}(s_0) - V_r^{\widetilde{\pi}}(s_0) + C|\widetilde{\tau}| + V_r^{\widetilde{\pi}}(s_0) - \nu(\widetilde{\tau}) \\
&= V_r^{\pi^*}(s_0) - V_r^{\widetilde{\pi}}(s_0) + C\left(b - V_c^{\widetilde{\pi}}(s_0)\right)_+ + V_r^{\widetilde{\pi}}(s_0) - \nu(\widetilde{\tau}) \\
&\leq \beta + V_r^{\widetilde{\pi}}(s_0) - \nu(\widetilde{\tau}),
\end{aligned}$$

where the first inequality follows from eq. (2.9) and the second inequality is due to the assumption on $\widetilde{\pi}$. Now we bound $\nu(\widetilde{\tau})$:

$$\begin{aligned}
\nu(\widetilde{\tau}) &= \max_{\pi} \left\{ V_r^{\pi}(s_0) \mid V_c^{\pi}(s_0) \geq b - \left(b - V_c^{\widetilde{\pi}}(s_0)\right)_+ \right\} \\
&\geq \max_{\pi} \left\{ V_r^{\pi}(s_0) \mid V_c^{\pi}(s_0) \geq V_c^{\widetilde{\pi}}(s_0) \right\} \\
&\geq V_r^{\widetilde{\pi}}(s_0),
\end{aligned}$$

where the inequality follows from tightening the constraints on π, and the second inequality follows because $\widetilde{\pi}$ immediately falls into the set on the right side. Finally, we have

$$(C - \lambda^*)\,|\widetilde{\tau}| \leq \beta,$$

$$\left(b - V_c^{\widetilde{\pi}}(s_0)\right)_+ \leq \frac{\beta}{C - \lambda^*}.$$

□

The following lemma is a corrected version developed based on the lemma 10 of Vaswani et al. [2022].

Lemma 2.0.2. *For any $C > \widehat{\lambda}^*$ and any $\widetilde{\pi}$ s.t. $\widehat{V}_{r_p}^{\widehat{\pi}^*}(s_0) - \widehat{V}_{r_p}^{\widetilde{\pi}}(s_0) +\, C\left[b' - \widehat{V}_c^{\widetilde{\pi}}(s_0)\right]_+ \leq \beta$, we have $\left[b' - \widehat{V}_c^{\widetilde{\pi}}(s_0)\right]_+ \leq \frac{\beta}{C - \widehat{\lambda}^*}$.*

Proof. Define $\nu(\tau) = \max_{\pi} \left\{ \widehat{V}_{r_p}^{\pi}(s_0) \mid \widehat{V}_c^{\pi}(s_0) \geq b' + \tau \right\}$ and note that by definition, $\nu(0) =$

$\widehat{V}_{r_p}^{\widehat{\pi}^*}(s_0)$ and that ν is a decreasing function for its argument. Let $\widehat{V}_l^{\pi,\lambda}(s_0) = \widehat{V}_{r_p}^{\pi}(s_0) + \lambda\left(\widehat{V}_c^{\pi}(s_0) - b'\right)$. Then, for any policy π s.t. $\widehat{V}_c^{\pi}(s_0) \geq b' + \tau$, we have

$$\begin{aligned}
\widehat{V}_l^{\pi,\widehat{\lambda}^*}(s_0) &\leq \max_{\pi'} \widehat{V}_l^{\pi',\widehat{\lambda}^*}(s_0) \\
&= \widehat{V}_{r_p}^{\widehat{\pi}^*}(s_0) \\
&= \nu(0).
\end{aligned}$$

Then we have

$$\begin{aligned}
\nu(0) - \tau\widehat{\lambda}^* &\geq \widehat{V}_l^{\pi,\widehat{\lambda}^*}(s_0) - \tau\widehat{\lambda}^* \\
&= \widehat{V}_{r_p}^{\pi}(s_0) + \widehat{\lambda}^*(\widehat{V}_c^{\pi}(s_0) - b' - \tau) \\
&\geq \widehat{V}_{r_p}^{\pi}(s_0),
\end{aligned}$$

where the last inequality follows because $\widehat{V}_c^{\pi}(s_0) \geq b' + \tau$. Note that the above inequality holds for any policy π such that $\widehat{V}_c^{\pi}(s_0) \geq b' + \tau$, thus we have

$$\nu(0) - \tau\widehat{\lambda}^* \geq \max_{\pi}\{\widehat{V}_{r_p}^{\pi}(s_0) | \widehat{V}_c^{\pi}(s_0) \geq b' + \tau\} = \nu(\tau),$$

$$\tau\widehat{\lambda}^* \leq \nu(0) - \nu(\tau). \tag{2.10}$$

Recall the notation $[x]_+ = \max\{x, 0\}$, and choose $\widetilde{\tau} = -\left(b' - \widehat{V}_c^{\widetilde{\pi}}(s_0)\right)_+$,

$$\begin{aligned}
\left(C - \widehat{\lambda}^*\right)|\widetilde{\tau}| &= \widehat{\lambda}^*\widetilde{\tau} + C|\widetilde{\tau}| \\
&\leq \nu(0) - \nu(\widetilde{\tau}) + C|\widetilde{\tau}| \\
&= \widehat{V}_{r_p}^{\pi^*}(s_0) - \widehat{V}_{r_p}^{\widetilde{\pi}}(s_0) + C|\widetilde{\tau}| + \widehat{V}_{r_p}^{\widetilde{\pi}}(s_0) - \nu(\widetilde{\tau}) \\
&= \widehat{V}_{r_p}^{\widehat{\pi}^*}(s_0) - \widehat{V}_{r_p}^{\widetilde{\pi}}(s_0) + C\left(b' - \widehat{V}_c^{\widetilde{\pi}}(s_0)\right)_+ + \widehat{V}_{r_p}^{\widetilde{\pi}}(s_0) - \nu(\widetilde{\tau}) \\
&\leq \beta + \widehat{V}_{r_p}^{\widetilde{\pi}}(s_0) - \nu(\widetilde{\tau}),
\end{aligned}$$

where the first inequality follows from eq. (2.10) and the second inequality is due to the assumption on $\widetilde{\pi}$. Now we bound $\nu(\widetilde{\tau})$:

$$\begin{aligned}\nu(\widetilde{\tau}) &= \max_{\pi} \left\{ \widehat{V}^{\pi}_{r_p}(s_0) \mid \widehat{V}^{\pi}_{c}(s_0) \geq b' - \left(b' - \widehat{V}^{\widetilde{\pi}}_{c}(s_0) \right)_+ \right\} \\ &\geq \max_{\pi} \left\{ \widehat{V}^{\pi}_{r_p}(s_0) \mid \widehat{V}^{\pi}_{c}(s_0) \geq \widehat{V}^{\widetilde{\pi}}_{c}(s_0) \right\} \\ &\geq \widehat{V}^{\widetilde{\pi}}_{r_p}(s_0),\end{aligned}$$

where the inequality follows from tightening the constraints on π, and the second inequality follows because $\widetilde{\pi}$ immediately falls into the set on the right side. Finally, we have

$$\left(C - \widehat{\lambda}^* \right) |\widetilde{\tau}| \leq \beta,$$

$$\left(b' - \widehat{V}^{\widetilde{\pi}}_{c}(s_0) \right)_+ \leq \frac{\beta}{C - \widehat{\lambda}^*}.$$

□

Lemma 2.0.3. *(Lemma 9 of Vaswani et al. [2022]) The objective eq.* (1.4) *satisfies strong duality. Defining* $\pi^*_c := \arg\max_\pi V^\pi_c(s_0)$*. We consider two cases: (1) If* $b' = b - \varepsilon'$ *for* $\varepsilon' > 0$ *and event* $\mathcal{E}_1 = \left\{ \left| \widehat{V}^{\pi^*_c}_c(s_0) - V^{\pi^*_c}_c(s_0) \right| \leq \frac{\varepsilon'}{2} \right\}$ *holds, then* $\lambda^* \leq \frac{2(1+\omega)}{\varepsilon'(1-\gamma)}$ *and (2) If* $b' = b + \Delta$ *for* $\Delta \in \left(0, \frac{\zeta}{2}\right)$ *and event* $\mathcal{E}_2 = \left\{ \left| \widehat{V}^{\pi^*_c}_c(s_0) - V^{\pi^*_c}_c(s_0) \right| \leq \frac{\zeta}{2} - \Delta \right\}$ *holds, then* $\lambda^* \leq \frac{2(1+\omega)}{\zeta(1-\gamma)}$.

Proof. Writing the empirical CMDP in eq. (1.4) in its Lagrangian form,

$$\widehat{V}^{\widehat{\pi}^*}_{r_p}(s_0) = \max_{\pi} \min_{\lambda \geq 0} \widehat{V}^{\pi}_{r_p}(s_0) + \lambda \left[\widehat{V}^{\pi}_{c}(s_0) - b' \right]$$

Using the linear programming formulation of CMDPs in terms of the state-occupancy measures μ, we know that both the objective and the constraint are linear functions of μ, and strong duality holds w.r.t μ. Since μ and π have a one-one mapping, we can switch the min and the max [Paternain et al., 2019], implying,

$$= \min_{\lambda \geq 0} \max_{\pi} \widehat{V}^{\pi}_{r_p}(s_0) + \lambda \left[\widehat{V}^{\pi}_{c}(s_0) - b' \right]$$

Since λ^* is the optimal dual variable for the empirical CMDP in (4),

$$= \max_{\pi} \widehat{V}^{\pi}_{r_p}(s_0) + \lambda^* \left[\widehat{V}^{\pi}_{c}(s_0) - b'\right]$$

Define $\pi^*_c := \arg\max V^{\pi}_c(s_0)$ and $\widehat{\pi}^*_c := \arg\max \widehat{V}^{\pi}_c(s_0)$

$$\begin{aligned}
&\geq \widehat{V}^{\widehat{\pi}^*_c}_{r_p}(s_0) + \lambda^* \left[\widehat{V}^{\widehat{\pi}^*_c}_{c}(s_0) - b'\right] \\
&= \widehat{V}^{\widehat{\pi}^*_c}_{r_p}(s_0) + \lambda^* \left[\left(\widehat{V}^{\widehat{\pi}^*_c}_{c}(s_0) - V^{\pi^*_c}_{c}(s_0)\right) + \left(V^{\pi^*_c}_{c}(s_0) - b\right) + (b - b')\right]
\end{aligned}$$

By definition, $\zeta = V^{\pi^*_c}_c(s_0) - b$

$$= \widehat{V}^{\widehat{\pi}^*_c}_{r_p}(s_0) + \lambda^* \left[\left(\widehat{V}^{\widehat{\pi}^*_c}_{c}(s_0) - \widehat{V}^{\pi^*_c}_{c}(s_0)\right) + \left(\widehat{V}^{\pi^*_c}_{c}(s_0) - V^{\pi^*_c}_{c}(s_0)\right) + \zeta + (b - b')\right]$$

By definition of $\widehat{\pi}^*_c$, $\left(\widehat{V}^{\widehat{\pi}^*_c}_{c}(s_0) - \widehat{V}^{\pi^*_c}_{c}(s_0)\right) \geq 0$

$$\widehat{V}^{\widehat{\pi}^*}_{r_p}(s_0) \geq \widehat{V}^{\widehat{\pi}^*_c}_{r_p}(s_0) + \lambda^* \left[\zeta + (b - b') - \left|\widehat{V}^{\pi^*_c}_{c}(s_0) - V^{\pi^*_c}_{c}(s_0)\right|\right]$$

1) If $b' = b - \varepsilon'$ for $\varepsilon' > 0$. Hence,

$$\widehat{V}^{\widehat{\pi}^*}_{r_p}(s_0) \geq \widehat{V}^{\widehat{\pi}^*_c}_{r_p}(s_0) + \lambda^* \left[\zeta + \varepsilon' - \left|\widehat{V}^{\pi^*_c}_{c}(s_0) - V^{\pi^*_c}_{c}(s_0)\right|\right]$$

If the event $\mathcal{E}_1$ holds, $\left|\widehat{V}^{\pi^*_c}_{c}(s_0) - V^{\pi^*_c}_{c}(s_0)\right| \leq \frac{\varepsilon'}{2}$, implying, $\left|\widehat{V}^{\pi^*_c}_{c}(s_0) - V^{\pi^*_c}_{c}(s_0)\right| < \zeta + \frac{\varepsilon'}{2}$, then,

$$\begin{aligned}
&\geq \widehat{V}^{\widehat{\pi}^*_c}_{r_p}(s_0) + \lambda^* \frac{\varepsilon'}{2} \\
\implies \lambda^* &\leq \frac{2}{\varepsilon'} \left[\widehat{V}^{\widehat{\pi}^*}_{r_p}(s_0) - \widehat{V}^{\widehat{\pi}^*_c}_{r_p}(s_0)\right] \leq \frac{2(1+\omega)}{\varepsilon'(1-\gamma)}
\end{aligned}$$

2) If $b' = b + \Delta$ for $\Delta \in \left(0, \frac{\zeta}{2}\right)$. Hence,

$$\widehat{V}^{\widehat{\pi}^*}_{r_p}(s_0) \geq \widehat{V}^{\widehat{\pi}^*_c}_{r_p}(s_0) + \lambda^* \left[\zeta - \Delta - \left|\widehat{V}^{\pi^*_c}_{c}(s_0) - V^{\pi^*_c}_{c}(s_0)\right|\right]$$

If the event $\mathcal{E}_2$ holds, $\left|\widehat{V}_c^{\pi_c^*}(s_0) - V_c^{\pi_c^*}(s_0)\right| \leq \frac{\zeta}{2} - \Delta$ for $\Delta < \frac{\zeta}{2}$, then,

$$\begin{aligned} &\geq \widehat{V}_{r_p}^{\widehat{\pi}_c^*}(s_0) + \lambda^* \frac{\zeta}{2} \\ \implies \lambda^* &\leq \frac{2}{\zeta}\left[\widehat{V}_{r_p}^{\widehat{\pi}^*}(s_0) - \widehat{V}_{r_p}^{\widehat{\pi}_c^*}(s_0)\right] \leq \frac{2(1+\omega)}{\zeta(1-\gamma)} \end{aligned}$$

□

We analyze Algorithm line 1 under both the relaxed feasibility and strict feasibility cases [Vaswani et al., 2022].

For CMDPs with relaxed feasibility, the requirement for the learner to meet the constraint is relaxed. The learner is required to return a policy $\widehat{\pi}$ with an approximately optimal value and a small constraint violation in M, i.e.,

$$V_r^{\widehat{\pi}}(s_0) \geq V_r^*(s_0) - \varepsilon, \text{ and } V_c^{\widehat{\pi}}(s_0) \geq b - \varepsilon. \tag{2.11}$$

For CMDPs with strict feasibility, the requirement for the learner to meet the constraint is strict. The learner is required to return a policy $\widehat{\pi}$ with an approximately optimal value and no constraint violation in M, i.e.,

$$V_r^{\widehat{\pi}}(s_0) \geq V_r^*(s_0) - \varepsilon, \text{ and } V_c^{\widehat{\pi}}(s_0) \geq b. \tag{2.12}$$

2.1 Relaxed feasibility

Theorem 2.1.1. *(Theorem 2 of Vaswani et al. [2022]) For a fixed $\varepsilon \in (0, 1/(1-\gamma)]$ and $\delta \in (0,1)$, Algorithm line 1 with $N = \widetilde{O}(\frac{\log(1/\delta)}{(1-\gamma)^3\varepsilon^2})$ samples, $b' = b - \frac{3\varepsilon}{8}, \omega = \frac{\varepsilon(1-\gamma)}{8}, U = O(1/\varepsilon(1-\gamma)), \varepsilon_1 = O(\varepsilon^2(1-\gamma)^2)$ and $T = O(1/(1-\gamma)^4\varepsilon^4)$, returns policy $\bar{\pi}_T$ that satisfies the objective in eq.* (2.11) *with probability at least* $1 - 4\delta$.

Proof. We prove the result for a general primal-dual error $\varepsilon_{\text{opt}} < \varepsilon$ and $b' = b - \frac{\varepsilon - \varepsilon_{\text{opt}}}{2}$, and subsequently specify ε_{opt} and hence b'. In lemma 2.1.1, we show that if the constraint value functions are sufficiently concentrated (the empirical value function is close to the ground truth value function) for both the optimal policy π^* in M and the mixture policy $\bar{\pi}_T$ returned by Algorithm line 1, i.e., if

$$\left|V_c^{\bar{\pi}_T}(s_0) - \widehat{V}_c^{\bar{\pi}_T}(s_0)\right| \leq \frac{\varepsilon - \varepsilon_{\text{opt}}}{2}; \quad \left|V_c^{\pi^*}(s_0) - \widehat{V}_c^{\pi^*}(s_0)\right| \leq \frac{\varepsilon - \varepsilon_{\text{opt}}}{2},$$

then (i) policy $\bar{\pi}_T$ violates the constraint in M by at most ε, i.e., $V_c^{\bar{\pi}_T}(s_0) \geq b - \varepsilon$, and (ii) its suboptimality in M (compared to π^*) can be decomposed as:

$$V_r^{\pi^*}(s_0) - V_r^{\bar{\pi}_T}(s_0) \leq \frac{2\omega}{1-\gamma} + \varepsilon_{\text{opt}} + \left|V_{r_p}^{\pi^*}(s_0) - \widehat{V}_{r_p}^{\pi^*}(s_0)\right| + \left|\widehat{V}_{r_p}^{\bar{\pi}_T}(s_0) - V_{r_p}^{\bar{\pi}_T}(s_0)\right|. \quad (2.13)$$

In order to instantiate the primal-dual algorithm, we require a concentration result for policy π_c^* that maximizes the constraint value function, i.e. if $\pi_c^* := \arg\max V_c^{\pi}(s_0)$, then we require $\left|V_c^{\pi_c^*}(s_0) - \widehat{V}_c^{\pi_c^*}(s_0)\right| \leq \varepsilon + \varepsilon_{\text{opt}}$. In Case 1 of lemma 2.0.3, we show that if this concentration result holds, then we can upper-bound the optimal dual variable $|\lambda^*|$ by $\frac{2(1+\omega)}{(\varepsilon+\varepsilon_{\text{opt}})(1-\gamma)}$. With these results in hand, we can instantiate all the algorithm parameters except N (the number of samples required for each state-action pair). In particular, we set $\varepsilon_{\text{opt}} = \frac{\varepsilon}{4}$ and hence $b' = b - \frac{3\varepsilon}{8}$, and $\omega = \frac{\varepsilon(1-\gamma)}{8} < 1$. Setting $U = \frac{32}{5\varepsilon(1-\gamma)}$ ensures that the $U > |\lambda^*|$ condition required by theorem 2.0.1 holds. To guarantee that the primal-dual algorithm outputs an $\frac{\varepsilon}{4}$-approximate policy, we use theorem 2.0.1, recall that $|\lambda^*| \leq C := \frac{16}{5\varepsilon(1-\gamma)^2}$ and $U = 2C$, and set

$$\begin{aligned} T &= \frac{4U^2}{\varepsilon_{\text{opt}}^2(1-\gamma)^2}\left[1 + \frac{1}{(U-\lambda^*)^2}\right] = \frac{64}{\varepsilon^2(1-\gamma)^2}\left[1 + \frac{1}{(U-\lambda^*)^2}\right] \leq \frac{256}{\varepsilon^2(1-\gamma)^2}[C^2+1] \\ &< \frac{512}{\varepsilon^2(1-\gamma)^2}C^2 = \frac{512}{\varepsilon^2(1-\gamma)^2}\frac{256}{25\varepsilon^2(1-\gamma)^2}. \end{aligned}$$

Thus,

$$T = O\left(\frac{1}{(1-\gamma)^4\varepsilon^4}\right).$$

Using theorem 2.0.1, we need to set ε_1,

$$\varepsilon_1 = \frac{\varepsilon_{\text{opt}}^2 (1-\gamma)^2 \left(U - \lambda^*\right)}{6U} = \frac{\varepsilon^2 (1-\gamma)^2 \left(U - \lambda^*\right)}{96U} \leq \frac{\varepsilon^2(1-\gamma)^2}{96}.$$

Thus,

$$\varepsilon_1 = O\left(\varepsilon^2(1-\gamma)^2\right).$$

With our choices of the value of ε_{opt} and ω, eq. (2.13) can then be simplified as

$$V_r^{\pi^*}(s_0) - V_r^{\bar{\pi}_T}(s_0) \leq \frac{\varepsilon}{2} + \left|V_{r_p}^{\pi^*}(s_0) - \widehat{V}_{r_p}^{\pi^*}(s_0)\right| + \left|\widehat{V}_{r_p}^{\bar{\pi}_T}(s_0) - V_{r_p}^{\bar{\pi}_T}(s_0)\right|.$$

Putting everything together, in order to guarantee an ε-reward suboptimality for $\bar{\pi}_T$, we require that:

$$\begin{aligned} &\left|V_c^{\pi_c^*}(s_0) - \widehat{V}_c^{\pi_c^*}(s_0)\right| \leq \frac{5\varepsilon}{4};\ \left|V_c^{\bar{\pi}_T}(s_0) - \widehat{V}_c^{\bar{\pi}_T}(s_0)\right| \leq \frac{3\varepsilon}{8};\ \left|V_c^{\pi^*}(s_0) - \widehat{V}_c^{\pi^*}(s_0)\right| \leq \frac{3\varepsilon}{8} \\ &\left|V_{r_p}^{\pi^*}(s_0) - \widehat{V}_{r_p}^{\pi^*}(s_0)\right| \leq \frac{\varepsilon}{4};\ \left|\widehat{V}_{r_p}^{\bar{\pi}_T}(s_0) - V_{r_p}^{\bar{\pi}_T}(s_0)\right| \leq \frac{\varepsilon}{4}. \end{aligned} \tag{2.14}$$

We control such concentration terms for both the constraint and reward value functions in section 2.3, and bound the terms in eq. (2.14). In particular, we prove that for a fixed $\varepsilon \in (0, 1/1-\gamma]$, using $N \geq \widetilde{O}\left(\frac{\log(1/\delta)}{(1-\gamma)^3\varepsilon^2}\right)$ samples ensures that the statements in eq. (2.14) hold with probability $1-4\delta$. This guarantees that $V_r^{\pi^*}(s_0) - V_r^{\bar{\pi}_T}(s_0) \leq \varepsilon$ and $V_c^{\bar{\pi}_T}(s_0) \geq b - \varepsilon$.

For bounding the concentration terms for $\bar{\pi}_T$ in eq. (2.14), we use theorem 2.3.1 with $U = \frac{32}{5\varepsilon(1-\gamma)}$, $\omega = \frac{\varepsilon(1-\gamma)}{8}$ and $\varepsilon_1 = \frac{\varepsilon^2(1-\gamma)^2}{96}$. In this case, $\iota = \frac{\omega\delta(1-\gamma)\varepsilon_1}{30U|S||A|^2} = O\left(\frac{\delta\varepsilon^4(1-\gamma)^4}{SA^2}\right)$ and

$$C(\delta) = 72\log\left(\frac{16(1+U+\omega)SA\log(e/1-\gamma)}{(1-\gamma)^2\iota\delta}\right) = O\left(\log\left(\frac{S^2A^3}{\delta^2\varepsilon^5(1-\gamma)^7}\right)\right).$$

With this value of $C(\delta)$, in order to satisfy the concentration bounds for $\bar{\pi}_T$, we require that

$$2\sqrt{\frac{C(\delta)}{N\cdot(1-\gamma)^3}} \leq \frac{\varepsilon}{4} \implies N \geq O\left(\frac{C(\delta)}{(1-\gamma)^3\varepsilon^2}\right)$$

We use the lemma 2.3.3 to bound the remaining concentration terms for π^* and π_c^* in eq. (2.14). In this case, for $C'(\delta) = 72 \log\left(\frac{4S\log(e/1-\gamma)}{\delta}\right)$, we require that,

$$2\sqrt{\frac{C'(\delta)}{N\cdot(1-\gamma)^3}} \leq \frac{\varepsilon}{4} \implies N \geq O\left(\frac{C'(\delta)}{(1-\gamma)^3\varepsilon^2}\right)$$

Hence, if $N \geq \widetilde{O}\left(\frac{\log(1/\delta)}{(1-\gamma)^3\varepsilon^2}\right)$, the bounds in eq. (2.14) are satisfied, completing the proof. □

Lemma 2.1.1. *(Lemma 11 of Vaswani et al. [2022]) For $b' = b - \frac{\varepsilon - \varepsilon_{opt}}{2}$, if (i) $\varepsilon_{opt} < \varepsilon$, and (ii) the following conditions are satisfied,*

$$\left|V_c^{\bar{\pi}_T}(s_0) - \widehat{V}_c^{\bar{\pi}_T}(s_0)\right| \leq \frac{\varepsilon - \varepsilon_{opt}}{2}; \quad \left|V_c^{\pi^*}(s_0) - \widehat{V}_c^{\pi^*}(s_0)\right| \leq \frac{\varepsilon - \varepsilon_{opt}}{2},$$

where $\pi_c^* := \arg\max V_c^\pi(s_0)$, then

- *(a) policy $\bar{\pi}_T$ violates the constraint by at most ε, i.e. $V_c^{\bar{\pi}_T}(s_0) \geq b - \varepsilon$ and*
- *(b) its optimality gap can be bounded as:*

$$V_r^{\pi^*}(s_0) - V_r^{\bar{\pi}_T}(s_0) \leq \frac{2\omega}{1-\gamma} + \varepsilon_{opt} + \left|V_{r_p}^{\pi^*}(s_0) - \widehat{V}_{r_p}^{\pi^*}(s_0)\right| + \left|\widehat{V}_{r_p}^{\bar{\pi}_T}(s_0) - V_{r_p}^{\bar{\pi}_T}(s_0)\right|.$$

Proof. For (a), from theorem 2.0.1, we know that,

$$\widehat{V}_c^{\bar{\pi}_T}(s_0) \geq b' - \varepsilon_{\text{opt}},$$

then

$$\begin{aligned} V_c^{\bar{\pi}_T}(s_0) &\geq V_c^{\bar{\pi}_T}(s_0) - \widehat{V}_c^{\bar{\pi}_T}(s_0) + b' - \varepsilon_{\text{opt}} \\ &\geq -\left|V_c^{\bar{\pi}_T}(s_0) - \widehat{V}_c^{\bar{\pi}_T}(s_0)\right| + b' - \varepsilon_{\text{opt}}. \end{aligned}$$

By definition of relaxed feasibility, we require $\bar{\pi}_T$ to violate the constraint in the true CMDP M by at most ε, i.e., $V_c^{\bar{\pi}_T}(s_0) \geq b - \varepsilon$. From the above equation, a sufficient condition for ensuring this is,

$$-\left|V_c^{\bar{\pi}_T}(s_0) - \widehat{V}_c^{\bar{\pi}_T}(s_0)\right| + b' - \varepsilon_{\text{opt}} \geq b - \varepsilon,$$

meaning that we require

$$\left|V_c^{\bar{\pi}_T}(s_0) - \widehat{V}_c^{\bar{\pi}_T}(s_0)\right| \leq (b' - b) - \varepsilon_{\text{opt}} + \varepsilon.$$

Plugging in the value of $b' = b - \frac{\varepsilon - \varepsilon_{\text{opt}}}{2}$, we have

$$\left|V_c^{\bar{\pi}_T}(s_0) - \widehat{V}_c^{\bar{\pi}_T}(s_0)\right| \leq \frac{\varepsilon - \varepsilon_{\text{opt}}}{2},$$

which indeed holds by our assumption.

For (b), let π^* be the solution to eq. (1.4). Then we show that π^* is feasible for the constrained problem in eq. (2.1), i.e., $\widehat{V}_c^{\pi^*}(s_0) \geq b'$. We have

$$V_c^{\pi^*}(s_0) \geq b \Longrightarrow \widehat{V}_c^{\pi^*}(s_0) \geq b - \left|V_c^{\pi^*}(s_0) - \widehat{V}_c^{\pi^*}(s_0)\right|.$$

Since we require $\widehat{V}_c^{\pi^*}(s_0) \geq b'$, using the above equation, a sufficient condition to ensure this is

$$b - \left|V_c^{\pi^*}(s_0) - \widehat{V}_c^{\pi^*}(s_0)\right| \geq b',$$

meaning that we require

$$\left|V_c^{\pi^*}(s_0) - \widehat{V}_c^{\pi^*}(s_0)\right| \leq b - b'.$$

Since $b' = b - \frac{\varepsilon - \varepsilon_{\text{opt}}}{2}$, we require that

$$\left|V_c^{\pi^*}(s_0) - \widehat{V}_c^{\pi^*}(s_0)\right| \leq \frac{\varepsilon - \varepsilon_{\text{opt}}}{2}.$$

Given that the above statements hold, we can decompose the suboptimality in the reward value function as follows:

$$
\begin{aligned}
& V_r^{\pi^*}(s_0) - V_r^{\bar{\pi}_T}(s_0) \\
=& V_r^{\pi^*}(s_0) - V_{r_p}^{\pi^*}(s_0) + V_{r_p}^{\pi^*}(s_0) - V_r^{\bar{\pi}_T}(s_0) \\
=& \left[V_r^{\pi^*}(s_0) - V_{r_p}^{\pi^*}(s_0)\right] + V_{r_p}^{\pi^*}(s_0) - \widehat{V}_{r_p}^{\pi^*}(s_0) + \widehat{V}_{r_p}^{\pi^*}(s_0) - V_r^{\bar{\pi}_T}(s_0) \\
\leq& \left[V_r^{\pi^*}(s_0) - V_{r_p}^{\pi^*}(s_0)\right] + \left[V_{r_p}^{\pi^*}(s_0) - \widehat{V}_{r_p}^{\pi^*}(s_0)\right] + \widehat{V}_{r_p}^{\widehat{\pi}^*}(s_0) - V_r^{\bar{\pi}_T}(s_0) \\
=& \left[V_r^{\pi^*}(s_0) - V_{r_p}^{\pi^*}(s_0)\right] + \left[V_{r_p}^{\pi^*}(s_0) - \widehat{V}_{r_p}^{\pi^*}(s_0)\right] + \left[\widehat{V}_{r_p}^{\widehat{\pi}^*}(s_0) - \widehat{V}_{r_p}^{\bar{\pi}_T}(s_0)\right] + \widehat{V}_{r_p}^{\bar{\pi}_T}(s_0) - V_r^{\bar{\pi}_T}(s_0) \\
=& \underbrace{\left[V_r^{\pi^*}(s_0) - V_{r_p}^{\pi^*}(s_0)\right]}_{\text{Perturbation Error}} + \underbrace{\left[V_{r_p}^{\pi^*}(s_0) - \widehat{V}_{r_p}^{\pi^*}(s_0)\right]}_{\text{Concentration Error}} + \underbrace{\left[\widehat{V}_{r_p}^{\widehat{\pi}^*}(s_0) - \widehat{V}_{r_p}^{\bar{\pi}_T}(s_0)\right]}_{\text{Primal-Dual Error}} \\
& + \underbrace{\left[\widehat{V}_{r_p}^{\bar{\pi}_T}(s_0) - V_{r_p}^{\bar{\pi}_T}(s_0)\right]}_{\text{Concentration Error}} + \underbrace{\left[V_{r_p}^{\bar{\pi}_T}(s_0) - V_r^{\bar{\pi}_T}(s_0)\right]}_{\text{Perturbation Error}},
\end{aligned}
$$

where the inequality follows from the optimality of $\widehat{\pi}^*$ and since we have ensured that π^* is feasible for eq. (2.1). For a perturbation magnitude equal to ω, we use lemma 2.3.2 to bound both perturbation errors by $\omega/(1-\gamma)$. Using theorem 2.0.1 to bound the primal-dual error by ε_{opt},

$$
V_r^{\pi^*}(s_0) - V_r^{\bar{\pi}_T}(s_0) \leq \frac{2w}{1-\gamma} + \varepsilon_{\text{opt}} + \underbrace{[V_{r_p}^{\pi^*}(s_0) - \widehat{V}_{r_p}^{\pi^*}(s_0)]}_{\text{Concentration Error}} + \underbrace{[\widehat{V}_{r_p}^{\bar{\pi}_T} - V_{r_p}^{\bar{\pi}_T}(s_0)]}_{\text{Concentration Error}}.
$$

□

2.2 Strict feasibility

Theorem 2.2.1. *(Theorem 3 of Vaswani et al. [2022]) For a fixed* $\varepsilon \in (0, 1/1-\gamma]$ *and* $\delta \in (0,1)$*, Algorithm line 1, with* $N = \widetilde{O}\left(\frac{\log(1/\delta)}{(1-\gamma)^5\varepsilon^2\zeta^2}\right)$ *samples,* $b' = b + \frac{\varepsilon(1-\gamma)}{10}$*,* $U = \frac{4(1+\omega)}{\zeta(1-\gamma)}$*,* $\varepsilon_1 = O(\varepsilon^2(1-\gamma)^4\zeta^2)$ *and* $T = O(1/(1-\gamma)^6\zeta^4\varepsilon^2)$ *returns policy* $\bar{\pi}_T$ *that satisfies the objective in eq.* (2.12)*, with probability at least* $1-4\delta$*.*

The proof mostly adopts similar methodology to the proof of relaxed feasibility.

2.3 Concentration results

Definition 2.3.1 (ι-Gap Condition). *MDP $\widehat{M}_\alpha$ satisfies the ι-gap condition if $\forall s$, $\widehat{V}^*_\alpha(s) - \max_{a':a\neq\widehat{\pi}^*_\alpha(s)} \widehat{Q}^*_\alpha(s,a') \geq \iota$, where $\widehat{\pi}^*_\alpha := \arg\max \widehat{V}^\pi_\alpha$ and $\widehat{\pi}^*_\alpha(s) = \arg\max_a \widehat{Q}^*_\alpha(s,a)$ is the optimal action in state s.*

Lemma 2.3.1. *(Lemma 5 of Vaswani et al. [2022]) Define $\widehat{\pi}^*_\alpha := \arg\max_\pi \widehat{V}^\pi_\alpha$. If (i) $\mathcal{E}$ is the event that the ι-gap condition in definition 2.3.1 holds for $\widehat{M}_\alpha$ and (ii) for $\delta \in (0,1)$ and $C(\delta) = 72\log\left(\frac{16\alpha_{\max}SA\log(e/1-\gamma)}{(1-\gamma)^2\iota\delta}\right)$, the number of samples per state-action pair is $N \geq \frac{4C(\delta)}{1-\gamma}$, then with probability at least $Pr[\mathcal{E}] - \delta/10$,*

$$\left\|\widehat{V}_\beta^{\widehat{\pi}^*_\alpha} - V_\beta\widehat{\pi}^*_\alpha\right\|_\infty \leq \sqrt{\frac{C(\delta)}{N\cdot(1-\gamma)^3}}\,\|\beta\|_\infty\,.$$

Lemma 2.3.2. *For any policy π, we have*

$$\left\|V^\pi_r(s_0) - V^\pi_{r_p}(s_0)\right\|_\infty \leq \frac{\omega}{1-\gamma} \quad ; \quad \left\|\widehat{V}^\pi_r(s_0) - \widehat{V}^\pi_{r_p}(s_0)\right\|_\infty \leq \frac{\omega}{1-\gamma}$$

Proof. For policy $\pi, V^\pi_r(s_0) = (I-\gamma P_\pi)^{-1}\, r^\pi$ and $V^\pi_{r_p}(s_0) = (I-\gamma P_\pi)^{-1}\, r^\pi_p$.

$$V^\pi_r(s_0) - V^\pi_{r_p}(s_0) = (I-\gamma P_\pi)^{-1}\left[r^\pi - r^\pi_p\right]$$
$$\implies \left\|V^\pi_r(s_0) - V^\pi_{r_p}(s_0)\right\|_\infty \leq \left\|(I-\gamma P_\pi)^{-1}\right\|_1 \left\|r^\pi - r^\pi_p\right\|_\infty$$

Since $\left\|(I-\gamma P_\pi)^{-1}\right\|_1 \leq \frac{1}{1-\gamma}$ and $\left\|r^\pi - r^\pi_p\right\|_\infty \leq \omega$

$$\left\|V^\pi_r(s_0) - V^\pi_{r_p}(s_0)\right\|_\infty \leq \frac{\omega}{1-\gamma}.$$

The same argument can be used to bound $\left\|\widehat{V}^\pi_r(s_0) - \widehat{V}^\pi_{r_p}(s_0)\right\|_\infty$ completing the proof. □

Theorem 2.3.1. *(Theorem 6 of Vaswani et al. [2022]) For $\delta \in (0,1), \omega \leq 1$ and $C(\delta) = 72\log\left(\frac{16(1+U+\omega)SA\log(e/1-\gamma)}{(1-\gamma)^2\iota\delta}\right)$ where $\iota = \frac{\omega\delta(1-\gamma)\varepsilon_1}{30U|S||A|^2}$, if $N \geq \frac{4C(\delta)}{1-\gamma}$, then for $\bar{\pi}_T$ output by Algo-*

rithm theorem 2.0.1, with probability at least $1-\delta/5$,

$$\left|V_{r_p}^{\bar{\pi}_T}(s_0) - \widehat{V}_{r_p}^{\bar{\pi}_T}(s_0)\right| \leq 2\sqrt{\frac{C(\delta)}{N\cdot(1-\gamma)^3}}; \quad \left|V_{c}^{\bar{\pi}_T}(s_0) - \widehat{V}_{c}^{\bar{\pi}_T}(s_0)\right| \leq \sqrt{\frac{C(\delta)}{N\cdot(1-\gamma)^3}}.$$

Proof. Since by definition $\bar{\pi}_T = \frac{1}{T}\sum_{t=0}^{T-1}\widehat{\pi}_t$ is a mixture policy, we have

$$\begin{aligned}\left|V_{r_p}^{\bar{\pi}_T}(s_0) - \widehat{V}_{r_p}^{\bar{\pi}_T}(s_0)\right| &= \left|\frac{1}{T}\sum_{t=0}^{T-1}\left[V_{r_p}^{\widehat{\pi}_t}(s_0) - \widehat{V}_{r_p}^{\widehat{\pi}_t}(s_0)\right]\right| \leq \frac{1}{T}\sum_{t=0}^{T-1}\left|V_{r_p}^{\widehat{\pi}_t}(s_0) - \widehat{V}_{r_p}^{\widehat{\pi}_t}(s_0)\right| \\ &\leq \frac{1}{T}\sum_{t=0}^{T-1}\left\|V_{r_p}^{\widehat{\pi}_t} - \widehat{V}_{r_p}^{\widehat{\pi}_t}\right\|_\infty.\end{aligned}$$

Recall that $\widehat{M}_{r+\widehat{\lambda}_t c}$ satisfies the gap condition with $\iota = \frac{\omega\delta}{30|\Lambda||S||A|^2}$ for every $\widehat{\lambda}_t \in \Lambda$. Since $|\Lambda| = \frac{U}{\varepsilon_1}$, $\iota = \frac{\omega\delta(1-\gamma)\varepsilon_1}{30U|S||A|^2}$. Since $\widehat{\pi}_t := \arg\max_\pi \widehat{V}^\pi_{r_p+\widehat{\lambda}_t c}$, we use lemma 2.3.1 with $\alpha = r_p + \widehat{\lambda}_t c$ and $\beta = r_p$, and obtain the following result. For $N \geq \frac{4C(\delta)}{1-\gamma}$, for each $t \in [T]$, with probability at least $1-\delta/5$,

$$\left\|V_{r_p}^{\widehat{\pi}_t} - \widehat{V}_{r_p}^{\widehat{\pi}_t}\right\|_\infty \leq \sqrt{\frac{C(\delta)}{N\cdot(1-\gamma)^3}}(1+\omega) \leq 2\sqrt{\frac{C(\delta)}{N\cdot(1-\gamma)^3}}.$$

Using the above relations,

$$\left|V_{r_p}^{\bar{\pi}_T}(s_0) - \widehat{V}_{r_p}^{\bar{\pi}_T}(s_0)\right| \leq 2\sqrt{\frac{C(\delta)}{N\cdot(1-\gamma)^3}}.$$

Similarly, invoking lemma 2.3.1 with $\alpha = r_p+\widehat{\lambda}_t c$ and $\beta = c$ gives the bound on $\left|V_{c}^{\bar{\pi}_T}(s_0) - \widehat{V}_{c}^{\bar{\pi}_T}(s_0)\right|$.

□

Lemma 2.3.3. *For* $\delta \in (0,1), \omega \leq 1$ *and* $C'(\delta) = 72\log\left(\frac{4|S|\log(e/1-\gamma)}{\delta}\right)$, *if* $N \geq \frac{4C'(\delta)}{1-\gamma}$ *and* $B(\delta,N) := \sqrt{\frac{C'(\delta)}{(1-\gamma)^3 N}}$, *then with probability at least* $1-3\delta$,

$$\left|V_{r_p}^{\pi^*}(s_0) - \widehat{V}_{r_p}^{\pi^*}(s_0)\right| \leq 2B(\delta,N); \left|V_{c}^{\pi^*}(s_0) - \widehat{V}_{c}^{\pi^*}(s_0)\right| \leq B(\delta,N); \left|V_{c}^{\pi^*_c}(s_0) - \widehat{V}_{c}^{\pi^*_c}(s_0)\right| \leq B(\delta,N).$$

Chapter 3

Learning in Linear CMDPs with Global Access

In this section, we consider inifinite-horizon discounted CMDPs with linear function approximation under the global access setting. Similar to chapter 2, let $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, c, b, \rho, \gamma)$ be the infinite-horizon γ-discounted CMDP. With the primal-dual approach, we have the primal variable π_t and the dual variable λ_t for iteration t. Hence, the learner will generate a sequence of policies $\{\pi_0, \pi_1, \ldots, \pi_{T-1}\}$ and Lagrange multipliers $\{\lambda_0, \lambda_1, \ldots, \lambda_{T-1}\}$ when the algorithm eventually returns a policy after T iterations. To measure the performance of the learning algorithm on maximizing reward value function and satisfying requirements for constraint value function, define the average optimality gap (OG) and the average constraint violation (CV) [Jain et al., 2022] as

$$\begin{aligned}
\mathrm{OG} &:= \frac{1}{T}\sum_{t=0}^{T-1}[V_r^*(\rho) - V_r^{\pi_t}(\rho)], \\
\mathrm{CV} &:= \frac{1}{T}\sum_{t=0}^{T-1}[b - V_c^{\pi_t}(\rho)]_+.
\end{aligned}$$

Further, we introduce some more notations for the learner's estimates. For some policy π, denote $\widehat{Q}_r^\pi$ and $\widehat{Q}_c^\pi$ as the estimate reward action-value function and estimate constraint action-value function respectively. For now, there is no need of knowing how the learner estimates the two action-value functions. Similar to section 1.4, we can introduce the estimate reward value function as $\widehat{V}_r^\pi = \langle \pi(\cdot|s), \widehat{Q}_r^\pi(s,\cdot)\rangle$ and the estimate constraint value function as $\widehat{V}_c^\pi = \langle \pi(\cdot|s), \widehat{Q}_c^\pi(s,\cdot)\rangle$. Define the primal regret and dual regret as

$$\begin{aligned}\mathcal{R}^p\left(\pi^*, T\right) &:= \mathbb{E}_{s\sim\nu_{\rho,\pi^*}} \sum_{t=0}^{T-1} \langle \pi^*(\cdot|s) - \pi_t(\cdot|s), (\widehat{Q}_r^{\pi_t}(s,\cdot) + \lambda_t \widehat{Q}_c^{\pi_t}(s,\cdot)\rangle, \\ \mathcal{R}^d(\lambda, T) &:= \sum_{t=0}^{T-1} (\lambda_t - \lambda)\,(\widehat{V}_c^{\pi_t}(\rho) - b),\end{aligned} \tag{3.1}$$

where $\nu_{\rho,\pi^*} \in \Delta_{\mathcal{S}}$ is the discounted occupancy measure. With the above quantities defined, we have the following theorem.

Theorem 3.0.1. *(Theorem 3.1 of Jain et al. [2022]) Assuming that* $\left\|Q_r^{\pi_t} - \widehat{Q}_r^{\pi_t}\right\|_\infty \leq \widetilde{\varepsilon}$ *and* $\left\|Q_c^{\pi_t} - \widehat{Q}_c^{\pi_t}\right\|_\infty \leq \widetilde{\varepsilon}$*, for a generic algorithm producing a sequence of polices* $\{\pi_0, \pi_1, \ldots, \pi_{T-1}\}$ *and dual variables* $\{\lambda_0, \lambda_1, \ldots, \lambda_{T-1}\}$ *such that for all* t, λ_t *is constrained to lie in the* $[0, U]$ *where* $U > \lambda^*, OG$ *and* CV *can be bounded as:*

$$\begin{aligned}OG &\leq \frac{\mathcal{R}^p\left(\pi^*, T\right) + (1-\gamma)\mathcal{R}^d(0,T)}{(1-\gamma)T} + \widetilde{\varepsilon} g(U), \\ CV &\leq \frac{\mathcal{R}^p\left(\pi^*, T\right) + (1-\gamma)\mathcal{R}^d(U,T)}{\left(U-\lambda^*\right)(1-\gamma)T} + \frac{\widetilde{\varepsilon} g(U)}{\left(U-\lambda^*\right)},\end{aligned}$$

where $g(U) := \left[\frac{1+U}{1-\gamma} + U\right]$.

Proof. We will begin with bounding the value differences in the Lagrangian using lemma 1.2.1. Let $T_{\pi^*}^r$ and $T_{\pi^*}^c$ be the Bellman operators of the optimal policy for the reward and cost respectively. Then,

$$\left[V_r^{\pi^*} - V_r^{\pi_t}\right] + \lambda_t\left[V_c^{\pi^*} - V_c^{\pi_t}\right] = \left(I - \gamma P_{\pi^*}\right)^{-1}\left[\left[T_{\pi^*}^r V_r^{\pi_t} - V_r^{\pi_t}\right] + \lambda_t\left[T_{\pi^*}^c V_c^{\pi_t} - V_c^{\pi_t}\right]\right]$$

Let M_π be the state-action operator applied Q functions such that $M_\pi(Q)(s) = \sum_a \pi(a|s)Q(s,a)$. Observe that $T_{\pi^*}^r V_r^{\pi_t} = M_{\pi^*} Q_r^{\pi_t}$ and $V_r^{\pi_t} = M_{\pi_t} Q_r^{\pi_t}$. The expressions for the constraint rewards are analogous. Rewriting the above expression,

$$\begin{aligned}\left[V_r^{\pi^*} - V_r^{\pi_t}\right] + \lambda_t \left[V_c^{\pi^*} - V_c^{\pi_t}\right] &= (I - \gamma P_{\pi^*})^{-1} \left[\left[M_{\pi^*} Q_r^{\pi_t} - M_{\pi_t} Q_r^{\pi_t}\right] + \lambda_t \left[M_{\pi^*} Q_c^{\pi_t} - M_{\pi_t} Q_c^{\pi_t}\right]\right] \\ &= (I - \gamma P_{\pi^*})^{-1} \left[\left[M_{\pi^*} - M_{\pi_t}\right] \left[Q_r^{\pi_t} + \lambda_t Q_c^{\pi_t}\right]\right] \\ &= (I - \gamma P_{\pi^*})^{-1} \left[\left[M_{\pi^*} - M_{\pi_t}\right] \left[\widehat{Q}_r^{\pi_t} + \lambda_t \widehat{Q}_c^{\pi_t}\right]\right] \\ &+ \underbrace{(I - \gamma P_{\pi^*})^{-1} \left[\left[M_{\pi^*} - M_{\pi_t}\right] \left[Q_r^{\pi_t} - \widehat{Q}_r^{\pi_t} + \lambda_t \left(Q_c^{\pi_t} - \widehat{Q}_c^{\pi_t}\right)\right]\right]}_{\text{Error}}\end{aligned}$$

Let us first bound the maximum norm of the "Error" term,

$$\begin{aligned}\|\text{ Error }\|_\infty &= \left\|(I - \gamma P_{\pi^*})^{-1} \left[\left[M_{\pi^*} - M_{\pi_t}\right] \left[Q_r^{\pi_t} - \widehat{Q}_r^{\pi_t} + \lambda_t \left(Q_c^{\pi_t} - \widehat{Q}_c^{\pi_t}\right)\right]\right]\right\|_\infty \\ &\leq \frac{1}{1-\gamma} \left\|\left[Q_r^{\pi_t} - \widehat{Q}_r^{\pi_t} + \lambda_t \left(Q_c^{\pi_t} - \widehat{Q}_c^{\pi_t}\right)\right]\right\|_\infty \\ &\leq \frac{1}{1-\gamma} \left\|Q_r^{\pi_t} - \widehat{Q}_r^{\pi_t}\right\|_\infty + \frac{\lambda_t}{1-\gamma} \left\|Q_c^{\pi_t} - \widehat{Q}_c^{\pi_t}\right\|_\infty .\end{aligned}$$

By assumption, $\left\|Q_r^{\pi_t} - \widehat{Q}_r^{\pi_t}\right\|_\infty, \left\|Q_c^{\pi_t} - \widehat{Q}_c^{\pi_t}\right\|_\infty \leq \varepsilon$.

$$\|\text{ Error }\|_\infty \leq \frac{\varepsilon}{1-\gamma} \left(1 + \lambda_t\right).$$

Since the dual variables are projected onto the $[0, U]$ interval, $\lambda_t \leq U$, implying that

$$\|\text{ Error }\|_\infty \leq \frac{\varepsilon}{1-\gamma}(1 + U)$$

Substituting in this bound on the error, using the convention that left-multiplication by a measure means integration with respect to it,

$$\begin{aligned}&\left[V_r^{\pi^*}(\rho) - V_r^{\pi_t}(\rho)\right] + \lambda_t \left[V_c^{\pi^*}(\rho) - V_c^{\pi_t}(\rho)\right] \\ \leq &\rho \left(I - \gamma P_{\pi^*}\right)^{-1} \left[\left[M_{\pi^*} - M_{\pi_t}\right] \left[\widehat{Q}_r^{\pi_t} + \lambda_t \widehat{Q}_c^{\pi_t}\right]\right] + \frac{\varepsilon}{1-\gamma}(1 + U) \\ \leq &\frac{1}{1-\gamma} \nu_{\rho,\pi^*} \left[\left[M_{\pi^*} - M_{\pi_t}\right] \left[\widehat{Q}_r^{\pi_t} + \lambda_t \widehat{Q}_c^{\pi_t}\right]\right] + \frac{\varepsilon}{1-\gamma}(1 + U)\end{aligned}$$

where $\nu_{\rho,\pi^*} = (1-\gamma)\rho\,(I-\gamma P_{\pi^*})^{-1}$ is the discounted probability measure over the states obtained when starting from ρ and following π^*. Summing from $t=0$ to $T-1$ and dividing by T.

$$\begin{aligned}
&\frac{1}{T}\nu_{\rho,\pi^*}\sum_{t=0}^{T-1}\left[\left[V_r^{\pi^*}(\rho)-V_r^{\pi_t}(\rho)\right]+\lambda_t\left[V_c^{\pi^*}(\rho)-V_c^{\pi_t}(\rho)\right]\right]\\
\leq&\frac{\nu_{\rho,\pi^*}}{(1-\gamma)T}\sum_{t=0}^{T-1}\left[\left[\mathcal{M}_{\pi^*}-\mathcal{M}_{\pi_t}\right]\left[\widehat{Q}_r^{\pi_t}+\lambda_t\widehat{Q}_c^{\pi_t}\right]\right]+\frac{\varepsilon}{1-\gamma}(1+U)
\end{aligned}$$

Now, observe that

$$\begin{aligned}
&\nu_{\rho,\pi^*}\sum_{t=0}^{T-1}\left[\left[\mathcal{M}_{\pi^*}-\mathcal{M}_{\pi_t}\right]\left[\widehat{Q}_r^{\pi_t}+\lambda_t\widehat{Q}_c^{\pi_t}\right]\right]\\
=&\sum_{t=0}^{T-1}\langle\pi^*(\cdot|s)-\pi_t(\cdot|s),\widehat{Q}_r^{\pi_t}(s,\cdot)+\lambda_t\widehat{Q}_c^{\pi_t}(s,\cdot)\rangle_{s\sim\nu_{\rho,\pi^*}}\\
=&\mathcal{R}^p\left(\pi^*,T\right)
\end{aligned}$$

Putting everything together,

$$\frac{1}{T}\sum_{t=0}^{T-1}\left[V_r^{\pi^*}(\rho)-V_r^{\pi_t}(\rho)\right]+\frac{1}{T}\sum_{t=0}^{T-1}\lambda_t\left[V_c^{\pi^*}(\rho)-V_c^{\pi_t}(\rho)\right]\leq\frac{\mathcal{R}^p\left(\pi^*,T\right)}{(1-\gamma)T}+\frac{\varepsilon}{1-\gamma}(1+U) \quad (3.2)$$

The above result bounds the sub-optimality in the Lagrangian. Next, we will see how this result implies a bound on the sub-optimality in the objective and the constraint violation. To bound the reward sub-optimality, we will upper bound the negative of the second term on the left-hand side in the above equation, i.e., we upper bound $\frac{1}{T}\sum_{t=0}^{T-1}\lambda_t\left[V_c^{\pi_t}(\rho)-V_c^{\pi^*}(\rho)\right]$.

We have,

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\lambda_t\left[V_c^{\pi_t}(\rho)-V_c^{\pi^*}(\rho)\right] &\leq \frac{1}{T}\sum_{t=0}^{T-1}\lambda_t\left[V_c^{\pi_t}(\rho)-b\right]\\ &= \frac{1}{T}\sum_{t=0}^{T-1}\lambda_t\left[V_c^{\pi_t}(\rho)-\widehat{V}_c^{\pi_t}(\rho)\right]+\frac{1}{T}\sum_{t=0}^{T-1}\lambda_t\left[\widehat{V}_c^{\pi_t}(\rho)-b\right]\\ &= \frac{1}{T}\sum_{t=0}^{T-1}\lambda_t\left[V_c^{\pi_t}(\rho)-\widehat{V}_c^{\pi_t}(\rho)\right]+\frac{\mathcal{R}^d(0,T)}{T}\\ &\leq U\varepsilon+\frac{\mathcal{R}^d(0,T)}{T}.\end{aligned} \tag{3.3}$$

Using Eqs.eq. (3.2) and eq. (3.3),

$$\mathrm{OG}=\frac{1}{T}\sum_{t=0}^{T-1}\left[V_r^{\pi^*}(\rho)-V_r^{\pi_t}(\rho)\right]\leq\frac{\mathcal{R}^p\left(\pi^*,T\right)+(1-\gamma)\mathcal{R}^d(0,T)}{(1-\gamma)T}+\frac{\varepsilon}{1-\gamma}(1+U)+U\varepsilon$$

This proves the first part of the theorem. We now bound the constraint violation. For an arbitrary λ,

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\left[\left(\lambda_t-\lambda\right)\left(V_c^{\pi_t}(\rho)-b\right)\right]\\ =&\frac{1}{T}\sum_{t=0}^{T-1}\left[\left(\lambda_t-\lambda\right)\left(V_c^{\pi_t}(\rho)-\widehat{V}_c^{\pi_t}(\rho)\right)\right]+\frac{1}{T}\sum_{t=0}^{T-1}\left[\left(\lambda_t-\lambda\right)\left(\widehat{V}_c^{\pi_t}(\rho)-b\right)\right]\\ =&\frac{1}{T}\sum_{t=0}^{T-1}\left[\left(\lambda_t-\lambda\right)\left(V_c^{\pi_t}(\rho)-\widehat{V}_c^{\pi_t}(\rho)\right)\right]+\frac{\mathcal{R}^d(\lambda,T)}{T},\end{aligned}$$

implying

$$\frac{1}{T}\sum_{t=0}^{T-1}\left[\left(\lambda_t-\lambda\right)\left(V_c^{\pi_t}(\rho)-b\right)\right]\leq U\varepsilon+\frac{\mathcal{R}^d(\lambda,T)}{T} \tag{3.4}$$

Adding eq. (3.4) and eq. (3.2) and reordering the terms gives

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^{\pi^*}(\rho)-V_r^{\pi_t}(\rho)\right)+\frac{\lambda}{T}\sum_{t=0}^{T-1}\left(b-V_c^{\pi_t}(\rho)\right)\\
&\leq \frac{1}{T}\sum_{t=0}^{T-1}\underbrace{\lambda_t\left(b-V_c^{\pi^*}(\rho)\right)}_{\leq 0 \text{ since } V_c^{\pi^*}(\rho)\geq b.}+\underbrace{\frac{\mathcal{R}^p\left(\pi^*,T\right)+(1-\gamma)\mathcal{R}^d(\lambda,T)}{(1-\gamma)T}+\frac{\varepsilon}{1-\gamma}(1+U)+U\varepsilon}_{h(\lambda)}\\
\Longrightarrow &\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^{\pi^*}(\rho)-V_r^{\pi_t}(\rho)\right)+\frac{\lambda}{T}\sum_{t=0}^{T-1}\left(b-V_c^{\pi_t}(\rho)\right)\leq h(\lambda)
\end{aligned}$$

We consider two cases: (i) if $\sum_{t=0}^{T-1}\left(b-V_c^{\pi_t}(\rho)\right)\geq 0$, we set $\lambda=U$, else, if (ii) $\sum_{t=0}^{T-1}\left(b-V_c^{\pi_t}(\rho)\right)<0$, we set $\lambda=0$. Using these choices, and since $\mathcal{R}^d(\lambda,T)$ is linearly increasing in λ,

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^{\pi^*}(\rho)-V_r^{\pi_t}(\rho)\right)+\frac{U}{T}\left[\sum_{t=0}^{T-1}\left(b-V_c^{\pi_t}(\rho)\right)\right]_+\leq h(U)$$

Now take the policy π' such that $V_r^{\pi^*}(\rho)-V_r^{\pi'}(\rho)=\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^{\pi^*}(\rho)-V_r^{\pi_t}(\rho)\right)$ and $V_c^{\pi^*}(\rho)-V_c^{\pi'}(\rho)=\frac{1}{T}\sum_{t=0}^{T-1}\left(b-V_c^{\pi_t}(\rho)\right)$. Then,

$$\left[V_r^{\pi^*}(\rho)-V_r^{\pi'}(\rho)\right]+U\left[b-V_c^{\pi'}(\rho)\right]_+\leq h(U)$$

Using lemma 2.0.1 with $C=U>\lambda^*$ and $\beta=h(U)$, we get

$$\begin{aligned}
\mathrm{CV}=\frac{1}{T}\left[\sum_{t=0}^{T-1}b-V_c^{\pi_t}(\rho)\right]_+&=\left[b-V_c^{\pi'}(\rho)\right]_+\\
\leq\frac{h(U)}{U-\lambda^*}&=\frac{\mathcal{R}^p\left(\pi^*,T\right)+(1-\gamma)\mathcal{R}^d(U,T)}{\left(U-\lambda^*\right)(1-\gamma)T}+\frac{1}{\left(U-\lambda^*\right)}\left[\frac{\varepsilon}{(1-\gamma)}(1+U)+U\varepsilon\right],
\end{aligned}$$

□

3.1 Coin betting

theorem 3.0.1 gives upper bounds of both the average optimality gap and the average constraint violation by the primal and dual regret defined in eq. (3.1). Note that minimizing the primal regret and dual regret is equivalent to solving an online linear optimization problem. Further, coin-betting, proposed by Orabona and Pál [2016], is a parameter-free algorithm designed for solving online linear optimization. Adopting a parameter-free online linear optimization algorithm can help eliminate the sensitivity of the algorithm to the values of hyper-parameters. To instantiate the coin-betting algorithm (Algorithm 2 in Orabona and Pál [2016]), we need to define one more variable w_t, which works as the "wealth" in the coin-betting setting.

$$w_{t+1}(s,a) = \frac{\sum_{i=0}^{t} \widetilde{A}_l^i(s,a)}{(t+1)+T/2} \left(1 + \sum_{i=0}^{t} \widetilde{A}_l^i(s,a) w_i(s,a)\right), \tag{3.5}$$

where given the primal variable π_t, $\widetilde{A}_l^t(s,a)$ is defined as

$$\widetilde{A}_l^t(s,a) = \begin{cases} \widehat{A}_l^t(s,a) & \text{if } w_t(s,a) > 0, \\ \left[\widehat{A}_l^t(s,a)\right]_+ & \text{if } w_t(s,a) \leq 0, \end{cases}$$

where $\widehat{A}_l^t(s,a)$ is defined as

$$\widehat{A}_l^t(s,a) = \frac{1-\gamma}{1+U} \left[\widehat{Q}_l^t(s,a) - \left\langle \widehat{Q}_l^t(s,\cdot), \pi_t(\cdot \mid s)\right\rangle\right].$$

Equivalently, $\widetilde{A}_l^t(s,a)$ can also be defined in a closed form:

$$\widetilde{A}_l^t(s,a) = \widehat{A}_l^t(s,a)\mathcal{I}\left\{w_t(s,a) > 0\right\} + \left[\widehat{A}_l^t(s,a)\right]_+ \mathcal{I}\left\{w_t(s,a) \leq 0\right\},$$

where $\mathcal{I}\{\omega\}$ is the indicator function that takes value 1 when the event ω holds true and 0 otherwise. Now it still remains to determine how to update π_t. We update the primal variable (policy π_t) using the same update rule in the coin-betting algorithm, i.e.,

$$\pi_{t+1}(a|s) = \begin{cases} \pi_0(a|s), & \text{if } \sum_a \pi_0(a|s)\left[w_{t+1}(s,a)\right]_+ = 0, \\ \frac{\pi_0(a|s)[w_{t+1}(s,a)]_+}{\sum_{a'} \pi_0(a'|s)[w_{t+1}(s,a')]_+}, & \text{otherwise}, \end{cases} \tag{3.6}$$

The update rule for the dual variable λ_t adopts the update rule of the "fraction" variable in the coin-betting algorithm Continuous Coin Betting (COCOB) proposed in Orabona and Tommasi [2017]. In particular, we have

$$\begin{aligned} \lambda_{t+1} &= \lambda_0 - \beta_t \left[\frac{1}{1-\gamma} - \sum_{i=0}^{t}(\lambda_i - \lambda_0)(\widehat{V}_c^{\pi_i}(\rho) - b)\right], \\ \beta_t &= (1-\gamma)\left(2\sigma\left(\frac{2\sum_{i=0}^{t}\left(\widehat{V}_c^{\pi_i}(\rho) - b\right)}{\frac{1}{1-\gamma} + \sum_{i=0}^{t}\left|\widehat{V}_c^{\pi_i}(\rho) - b\right|} - 1\right)\right), \end{aligned} \tag{3.7}$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$. With these updates for the primal and dual variables, it can be shown that the results of Orabona and Tommasi [2017] give us the following upper bounds for the primal regret and dual regret:

$$\mathcal{R}^p\left(\pi^*, T\right) \leq \frac{3(1+U)}{1-\gamma}\sqrt{T}\sqrt{1 + \mathrm{KL}\left(\pi_0 \| \pi^*\right)},$$

$$\mathcal{R}^d(\lambda, T) \leq \frac{1}{1-\gamma} + \left\|\lambda - \lambda^0\right\| \sqrt{\left(\frac{1}{(1-\gamma)^2} + \frac{G_T}{1-\gamma}\right)\Gamma_T},$$

where $\mathrm{KL}\left(\pi_0 || \pi^*\right) = \mathbb{E}_{s\sim\nu_{\rho,\pi^*}} \mathrm{KL}\left(\pi_0(\cdot \mid s) \| \pi^*(\cdot \mid s)\right), \Gamma_T = \log\left(1 + \left(G_T(1-\gamma)+1\right)^2 \|\lambda - \lambda^0\|^2\right)$ and $G_T = \sum_{i=0}^{T}\left|\widehat{V}_c^{\pi_i}(\rho) - b\right| = O(T)$.

3.2 Approximate action-value function realizability

So far we have derived results on generic CMDP algorithms adopting primal-dual approach and some preliminary analysis on coin-betting type algorithms, yet we have made no assumption of CMDPs. It is not always feasible to require a learner to perform well on every CMDPs. It is reasonable to restrict the set of CMDPs that the learner is required to work well on. To achieve this, we can make assumptions on CMDPs that assign some good properties. We expect the learner to perform well where the assumptions hold true. One of the most common assumptions one can make is that the value functions or the action value functions can be linearly parameterized by some feature map. Weaker assumptions claim that the value functions or the action value functions can be approximately linearly parameterized. Specifically, for the latter assumption, let $\Phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ be a feature map with feature vectors $\phi(s, a)$ for the state-action pair (s, a), we have

Assumption 3.2.1. *[Approximate universal linear Q^π-realizability] Given an CMDP M and the feature map Φ, for any memoryless policy π,*

$$\inf_{\theta \in \mathbb{R}^d} \|Q_r^\pi - \Phi\theta\|_\infty \leq \varepsilon_b, \tag{3.8}$$

$$\inf_{\theta \in \mathbb{R}^d} \|Q_c^\pi - \Phi\theta\|_\infty \leq \varepsilon_b. \tag{3.9}$$

The linear function approximation assumption claims that for any memoryless policy π, the best approximators for the action value function and the constraint action value function have a uniform error of at most ε_b. Planning under the approximate universal action value function realizability assumption is well studied [Szepesvári, 2023b]. The key objective of planning in MDPs with linear function approximation is to solve eq. (3.8) by approximating the minimizer with some estimate $\widehat{\theta}_r$. To calculate such $\widehat{\theta}_r$, one way is to maintain a set $\mathcal{C} \subseteq \mathcal{S} \times \mathcal{A}$ of state-action pairs, for each state-action pair $(s, a) \in \mathcal{C}$, rollout m trajectories from $z = (s, a)$ for a fixed number of steps H following policy π. Then, calculate the estimate action value function $q_r(z)$ as the average of the truncated cumulative discounted rewards generated by all m trajectories, i.e.,

$$q_r(z) = \frac{1}{m} \sum_{i=1}^{m} \sum_{t=1}^{H} \gamma^t r(s_t^i, a_t^i),$$

where $s_1^i = s$ and $a_1^i = a$ for all $i = 1, \dots, m$. Thus, we can denote q_r as a $|\mathcal{C}|$-dimensional vector that stores all the estimate action values of state-action pairs $z \in \mathcal{C}$ in the set. Let $\omega : \mathcal{C} \to (0, \infty)$ be some weighting function. Without loss of generality, we assume that $\sum_{z \in \mathcal{C}} \omega(z) = 1$. Then we apply the generalized least squares to solve $\widehat{\theta}_r^\pi$:

$$\widehat{\theta}_r^\pi = \arg\min_\theta \sum_{z \in \mathcal{C}} \omega(z)[\langle \theta, \phi(z) \rangle - q_r(z)]^2. \tag{3.10}$$

Define the weighted moment matrix G_ω as

$$G_\omega = \sum_{z \in \mathcal{C}} \omega(z) \phi(z) \phi(z)^\top,$$

which is nonsingular and thus we can solve eq. (3.10) as

$$\widehat{\theta}_r^\pi = G_\omega^{-1} \sum_{z \in \mathcal{C}} \omega(z) q_r(z) \phi(z).$$

Now we elaborate more on the set $\mathcal{C}$, which is known as the coreset. Due to the linear structure of the action value function, there is no need of keeping all state-action pairs in the coreset. However in order to extrapolate well to state-action pairs outside of the coreset $\mathcal{C}$ while keeping the coreset as small as possible for efficiency, we need to carefully design what should be included in the coreset. This problem is known as the G-optimal design. For some coreset $\mathcal{C}$ and some weighting function ω, we have the following result:

Lemma 3.2.1. *(Lemma 5.2 of Jain et al. [2022]) For policy π, any distribution ω and subset $\mathcal{C}$, if we use m trajectories to estimate the action value function for each $(s, a) \in \mathcal{C}$, and solve eq. (3.10) to compute $\widehat{\theta}_r^\pi$, then for any $(s, a) \in (\mathcal{S} \times \mathcal{A})$ pair, the error $\left| \left\langle \phi(s, a), \widehat{\theta}_r^\pi \right\rangle - Q_r^\pi \right|$ can be upper-bounded by*

$$\varepsilon_b \left(1 + \|\phi(s, a)\|_{G_\omega^\dagger} \right) + \frac{\|\phi(s, a)\|_{G_\omega^\dagger}^\dagger}{1 - \gamma} \sqrt{\frac{\log(2|\mathcal{C}|/\delta)}{2m}},$$

where, $G_\omega = \sum_{(s,a)\in\mathcal{C}} \omega(s,a)\phi(s,a)\phi(s,a)^\top$ and $A^\dagger$ is pseudoinverse of A.

The proof of this lemma can be found in Szepesvári [2023b]; Jain et al. [2022]. To control the extrapolation error at points (s,a) outside of the coreset $\mathcal{C}$, we need to carefully choose the coreset $\mathcal{C}$ and weighting function ω. To do so, we introduce the Kiefer-Wolfowitz theorem:

Theorem 3.2.1 (Kiefer-Wolfowitz [Kiefer and Wolfowitz, 1960])**.** *Let $\mathcal{Z}$ be finite. Let $\varphi : \mathcal{Z} \to \mathbb{R}^d$ be such that the underlying feature matrix Φ is rank d. There exists a set $\mathcal{C} \subseteq \mathcal{Z}$ and a distribution $\omega : \mathcal{C} \to [0,1]$ over this set, i.e. $\sum_{z'\in\mathcal{C}} \omega(z') = 1$, such that*

1. *$|\mathcal{C}| \le d(d+1)/2$,*
2. *$\sup_{z\in\mathcal{Z}} \|\varphi(z)\|_{G_\omega^{-1}} \le \sqrt{d}$,*
3. *In the previous line, the inequality is achieved with equality and the value of $\sqrt{d}$ is best possible under all possible choices of $\mathcal{C}$ and ω.*

If such $\mathcal{C}$ is found, then the function approximation error $\widetilde{\varepsilon}$ can be bounded by

$$\widetilde{\varepsilon} \le \varepsilon_b(1+\sqrt{d}) + \frac{\sqrt{d}}{1-\gamma}\sqrt{\frac{\log(2d(d+1)/\delta)}{2m}}.$$

Now we are ready to present the Coin-Betting Politex algorithm in line 2, and combine theorem 3.0.1 with lemma 3.2.1, we bound the average optimality gap and the average constraint violation in the following theorem:

Theorem 3.2.2. *Under Assumption assumption 3.2.1, OG and CV of line 2 can be bounded as:*

$$OG \le \frac{\left(\frac{3(1+U)\sqrt{1+KL(\pi_0\|\pi^*)}}{1-\gamma} + \Psi\right)}{(1-\gamma)\sqrt{T}} + \frac{\widetilde{\varepsilon}(1+2U)}{1-\gamma},$$

$$CV \le \frac{\zeta\left(\frac{3(1+U)\sqrt{1+KL(\pi_0\|\pi^*)}}{1-\gamma} + \Psi\right)}{\sqrt{T}} + \zeta\widetilde{\varepsilon}(1+2U),$$

where $U = \frac{2}{\zeta(1-\gamma)}$, $\widetilde{\varepsilon} = \varepsilon_b(1+\sqrt{d}) + \frac{\sqrt{d}}{1-\gamma}\sqrt{\frac{\log(2d(d+1)/\delta)}{2m}}$ and $\Psi = 4U\sqrt{\log((T+1)U)} + 1$.

Algorithm 2: Coin-Betting Politex [Jain et al., 2022]

```
Input : $\pi_0$ (policy initialization), $\lambda_0$ (dual variable initialization), $m$ (Number of trajectories), $T$ (Number of iterations), Feature map $\Phi$.
Compute coreset $\mathcal{C}$ and distribution $\omega$
Solve the unconstrained problem $\max_\pi \widehat{V}_c^\pi(\rho)$ to estimate the Slater constant $\zeta$ and set $U = \frac{2}{\zeta(1-\gamma)}$
for $t \leftarrow 0 \ldots T-1$ do
    For every $(s,a) \in \mathcal{C}$, use $m$ trajectories starting from $(s,a)$ using policy $\pi_t$ and estimate the action-value functions $q_r(s,a)$ and $q_c(s,a)$
    Compute and store $\widehat{\theta}_r^{\pi_t}$ and $\widehat{\theta}_c^{\pi_t}$ using eq. (3.10).
    for every $s$ encountered in the trajectory generated by $\pi_t$, and for every $a$ do
        Compute $\widehat{Q}_r^{\pi_t}(s,a) = \langle \widehat{\theta}_r^{\pi_t}, \phi(s,a) \rangle$,
            $\widehat{Q}_c^{\pi_t}(s,a) = \langle \widehat{\theta}_c^{\pi_t}, \phi(s,a) \rangle, \quad \text{and} \quad \widehat{Q}_l^{\pi_t}(s,a) = \widehat{Q}_r^{\pi_t}(s,a) + \lambda_t \widehat{Q}_c^{\pi_t}(s,a).$
        Update $\pi_{t+1}(a|s)$ using eq. (3.6)
    Compute $\widehat{V}_c^{\pi_t}(\rho)$, update $\lambda_{t+1}$ using eq. (3.7)
```

Chapter 4

Learning in Episodic Linear CMDPs with Online Access

In this section, we study time-inhomogenous episodic CMDPs with linear function approximation. Let $M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, c, b)$ be the CMDP. We have $\mathcal{S}$ and $\mathcal{A}$ as the state and action space, H as the fixed finite horizon of each episode, $\mathbb{P} = \{\mathbb{P}_h\}_1^H$ as the time-inhomogeneous transition probability, $r = \{r_h\}_1^H$ as the time-inhomogeneous reward function, $c = \{c_h\}_1^H$ as the time-inhomogeneous constraint reward function, and note that now we require $b \in (0, H]$ so that the problem is feasible and not trivial. Since the value functions for rewards and constraint rewards are analogous in the following analysis, we use l to denote either rewards r or constraint rewards c. Similar to our definitions in Section section 1.2, we define the value functions for episodic CMDPs. For some policy π, define the value function $V_{l,h}^{\pi}$ as the expected cumulative rewards starting from the state s:

$$V_{l,h}^{\pi}(s) = \mathbb{E}\left[\sum_{i=h}^{H} l_i(s_i, a_i) \mid s_h = s\right],$$

where the expectation is taken over the randomness of the trajectory $\{s_i, a_i\}_h^H$ generated by following policy π. Define the action value function $Q_{l,h}^{\pi}$ as the expected cumulative rewards starting from taking the action a in the state s:

$$Q^{\pi}_{l,h}(s,a) = \mathbb{E}\left[\sum_{i=h}^{H} l_i(s_i,a_i) \mid s_h = s, a_h = a\right].$$

For brevity, we introduce a notion $\mathbb{P}_h V^{\pi}_{l,h+1}(s,a)$ to denote the expected values of the next state s_{h+1} if the learner takes action a in the current state s. Specifically, we define $\mathbb{P}_h V^{\pi}_{l,h}$ as:

$$\mathbb{P}_h V^{\pi}_{l,h+1}(s,a) = \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)} V^{\pi}_{l,h+1}(s').$$

Now, the Bellman equation can be written as:

$$Q^{\pi}_{l,h}(s,a) = r_h(s,a) + \mathbb{P}_h V^{\pi}_{l,h+1}(s,a), \tag{4.1}$$

and

$$V^{\pi}_{l,h}(s) = \sum_{a \in \mathcal{A}} Q^{\pi}_{l,h}(s,a)\pi(a|s) = \left\langle Q^{\pi}_{l,h}(s,\cdot), \pi(\cdot|s) \right\rangle. \tag{4.2}$$

Further, we assume that the transition probability and the (constraint) reward function can be well represented by linear functions in some feature maps.

4.1 Linear CMDPs

As discussed in section 3.2, to scale up the state and action space, we often rely on all kinds of assumptions on MDPs. Apart from the (action) value function realizability assumptions, another common assumption on MDPs is that the transition probability (and in some cases the reward function as well) can be represented or approximated by some functions. One of the most simple function approximations gives us the linear kernel MDPs model. MDPs with linear function approximation [Ayoub et al., 2020; Cai et al., 2020; Jin et al., 2020; Ding et al., 2020; Zhou et al., 2021] is an assumption on MDPs that the transition and reward can be linearly parameterized by some feature maps of state-action pairs. There exist several different yet incomparable assumptions on how transitions are linearly parameterized, including linear mixture MDPs (also known as linear kernel MDPs) [Ayoub et al., 2020; Zhou et al., 2021; Cai et al., 2020; Ding et al., 2020] and linear MDPs [Jin et al., 2020]. In particular, we consider linear mixture MDPs in this section and thus have the following assumption:

Assumption 4.1.1 (Linear mixture MDPs). *MDP $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$ is a linear mixture MDP with a kernel feature map $\psi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$, if for any $h \in [H]$, there exists a vector $\theta_h \in \mathbb{R}^d$ with $\|\theta_h\|_2 \leq \sqrt{d}$ such that for any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,*

$$\mathbb{P}_h(s'|s, a) = \langle \psi(s, a, s'), \theta_h \rangle.$$

Moreover, we assume that for any function $V : \mathcal{S} \to [0, H]$, $\left\| \int_{\mathcal{S}} \psi(s, a, s') V(s') ds' \right\|_2 \leq \sqrt{d_1} H$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, and $\max\{d_1, d_2\} \leq d$.

For CMDPs, we make the same assumptions about the transitions and additionally we make some assumptions that rewards and constraint rewards can also be linearly parameterized by a different feature mapping.

Assumption 4.1.2 (Linear mixture CMDPs). *The CMDP $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, c, b)$ is a linear MDP with a kernel feature map $\psi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^{d_1}$ and a value feature map $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_2}$, if for any $h \in [H]$, there exists a vector $\theta_h \in \mathbb{R}^{d_1}$ with $\|\theta_h\|_2 \leq \sqrt{d_1}$ such that for any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,*

$$\mathbb{P}_h(s'|s, a) = \langle \psi(s, a, s'), \theta_h \rangle,$$

and there exists a feature map $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_2}$ and vectors $\theta_{r,h}, \theta_{c,h} \in \mathbb{R}^{d_2}$ such that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$r_h(s, a) = \langle \varphi(s, a), \theta_{r,h} \rangle, \quad \text{and } c_h(s, a) = \langle \varphi(s, a), \theta_{c,h} \rangle,$$

where $\max\{\|\theta_{r,h}\|_2, \|\theta_{c,h}\|_2\} \leq \sqrt{d_2}$. Moreover, we assume that for any function $V : \mathcal{S} \to$

$[0, H]$, $\left\| \int_{\mathcal{S}} \psi(s, a, s') V(s') ds' \right\|_2 \leq \sqrt{d_1} H$ *for all* $(s, a) \in \mathcal{S} \times \mathcal{A}$, *and* $\max\{d_1, d_2\} \leq d$.

The objective of the learner is to solve the following CMDP:

$$\max_{\pi} V_{r,1}^{\pi}(s_1), \quad \text{s.t.,} \quad V_{c,1}^{\pi}(s_1) \geq b. \tag{4.3}$$

To measure the performance of the learner on some CMDP M over K episodes of interaction in terms of maximizing the value and ensuring the constraint, we define the regret and constraint violation as:

$$\begin{aligned} \text{Regret}(K) &= \sum_{k=1}^{K} \left(V_{r,1}^{\pi^*}(x_1) - V_{r,1}^{\pi^k}(x_1) \right) \\ \text{Violation } (K) &= \left[\sum_{k=1}^{K} \left(b - V_{c,1}^{\pi^k}(x_1) \right) \right]_+ . \end{aligned} \tag{4.4}$$

Now to adopt the primal-dual approach, we rewrite the CMDP objective eq. (4.3) as the equivalent saddle-point problem

$$\max_{\pi} \min_{\lambda \geq 0} V_{r,1}^{\pi}(s_1) + \lambda \left(V_{c,1}^{\pi}(s_1) - b \right). \tag{4.5}$$

Again to solve the above saddle-point problem, we iteratively update the primal variable and dual variable over episodes. In the kth episode, denote the primal variable by π^k and denote the dual variable by λ^k. Given the primal and dual variables at the end of the last episode, the $k-1$th episode, one can update the primal variable π^k by directly solving the following policy optimization problem:

$$\max_{\pi} V_{r,1}^{\pi}(s_1) + \lambda^{k-1}(V_{c,1}^{\pi}(s_1) - b).$$

However, to solve the above problem requires the help of some policy optimization algorithms, which is computationally expensive and infeasible under the online setting. Hence, we need to come up with a new method to avoid the expensive computation.

4.2 Policy improvement and policy evaluation

Basically, we still follow the standard procedures of policy iterations to iteratively update our policies π_k, which includes two steps: policy improvement and policy evaluation. The policy evaluation step computes the value function $Q^{\pi^{k-1}}$ of the policy π^{k-1} for each episode k. In many cases, it suffices to compute only a good estimate of the true value function. The policy improvement step usually computes a greedy policy π^k with respect to the value function $Q^{\pi^{k-1}}$. Thus, iteratively conducting the policy evaluation and policy improvement steps generates a sequence of policies $\{\pi^k\}$.

4.2.1 Policy improvement

To study the performance of such policy iteration algorithms, we need to relate polices in the sequence. The tool to achieve this is the performance difference lemma. Note that lemma 1.2.2 works under the infinite-horizon discounted MDPs. Analogously, we define the time-inhomogeneous advantage as $A_h^\pi(s,a) = Q_h^\pi(s,a) - V_h^\pi(s)$, and thus have the performance difference lemma for finite-horizon MDPs:

$$V^{\pi'}(s) - V^\pi(s) = \mathbb{E}_{\pi'}\left[\sum_{h=1}^{H} A_h^\pi(s', \pi')\right] = \mathbb{E}_{\pi'}\left[\sum_{h=1}^{H} \langle Q^\pi(s,\cdot), (\pi' - \pi)(\cdot|s)\rangle\right]. \qquad (4.6)$$

With the performance difference lemma eq. (4.6), we now can relate the value function of any policy $V_{l,1}^\pi(s_1)$ to the value function $V_{l,1}^{\pi^{k-1}}(s_1)$ of the policy at the end of $k-1$th iteration π^{k-1}:

$$V_{l,1}^\pi(s_1) = V_{l,1}^{\pi^{k-1}}\left(s_1^k\right) + \mathbb{E}_{\pi^{k-1}}\left[\sum_{h=1}^{H} \left\langle Q_{l,h}^\pi(s_h,\cdot), \left(\pi_h - \pi_h^{k-1}\right)(\cdot \mid s_h)\right\rangle\right]. \qquad (4.7)$$

Note that eq. (4.7) provides us with a way to represent the value function $V_{l,1}^\pi(s_1)$ of any policy π. To calculate $V_{l,1}^\pi(s_1)$ in practice, we need to replace the expectation with a sample trajectory and thus obtain an approximation $L_l^{k-1}(\pi)$:

$$L_l^{k-1}(\pi) = V_{l,1}^{k-1}\left(s_1\right) + \sum_{h=1}^{H} \left\langle Q_{l,h}^{k-1}\left(s_h, \cdot\right), \left(\pi_h - \pi_h^{k-1}\right)\left(\cdot \mid s_h\right)\right\rangle, \tag{4.8}$$

where $V_{l,1}^{k-1}\left(s_1\right)$ and $Q_{l,h}^{k-1}\left(s_h, \cdot\right)$ are estimate value functions calculated by the learner in each episode. Finally, with the ability to obtain an approximation of the value function of any policy π, we are ready to present the update rule for the primal variable. As discussed above, we are oftentimes incapable of solving the problem in eq. (4.5), and instead we adopt an online mirror descent style of policy update by solving the following problem:

$$\max_{\pi} L_r^{k-1}(\pi) - \lambda^{k-1}\left(b - L_c^{k-1}(\pi)\right) \frac{1}{\alpha} \sum_{h=1}^{H} D\left(\pi_h\left(\cdot \mid_h\right) \mid \widetilde{\pi}_h^{k-1}\left(\cdot \mid s_h\right)\right), \tag{4.9}$$

where $\alpha > 0$ is a trade-off parameter, $D(\pi|\widetilde{\pi}^{k-1})$ is the Kullback–Leibler divergence (KL divergence), i.e.,

$$D(\pi(\cdot|s)|\pi'(\cdot|s)) = \sum_{a \in \mathcal{A}} \pi(a|s) \log\left(\frac{\pi(a|s)}{\pi'(a|s)}\right),$$

and

$$\widetilde{\pi}_h^{k-1}\left(\cdot \mid s_h\right) = (1-\theta)\pi_h^{k-1}\left(\cdot \mid s_h\right) + \theta\, \mathrm{Unif}(\mathcal{A})$$

is a policy mixing π^{k-1} with a uniform policy $\mathrm{Unif}(\mathcal{A})$ with some weight $\theta \in (0, 1]$. Mixing the uniform policy with policy π^k is important for it ensures the absolute continuity of π with respect to $\widetilde{\pi}^k$. The absolute continuity is required by the definition of the KL divergence which states that $\pi = 0$ wherever $\widetilde{\pi} = 0$. Clearing out all the terms that are irrelevant to π, the problem in eq. (4.9) can be solved by

$$\arg\max_{\pi} \sum_{h=1}^{H} \left\langle \left(Q_{r,h}^{k-1} + \lambda^{k-1} Q_{c,h}^{k-1}\right)\left(s_h, \cdot\right), \pi_h\left(\cdot|s_h\right)\right\rangle \frac{1}{\alpha} \sum_{h=1}^{H} D\left(\pi_h\left(\cdot \mid s_h\right) \mid \widetilde{\pi}_h^{k-1}\left(\cdot \mid s_h\right)\right). \tag{4.10}$$

Note that there is no dependence among different states and actions in the above update rules. Hence for any time step $h \in [H]$, we can equivalently update the policy by each state-action pair (s, a) by a closed form rule:

$$\pi_h^k(\cdot \mid \cdot) \propto \widetilde{\pi}_h^{k-1}(\cdot \mid \cdot) \exp\left(\alpha\left(Q_{r,h}^{k-1} + \lambda^{k-1} Q_{c,h}^{k-1}\right)(\cdot, \cdot)\right). \tag{4.11}$$

4.2.2 Policy evaluation

Now that we have established the update rule for the policy improvement step, it remains to show how to calculate the estimate value functions used in the update rules, which is the policy evaluation step. With the linear mixture CMDPs assumption assumption 4.1.2, we can use least squares to estimate the value functions $V_{l,1}^{\pi^k}$. For brevity, we denote $V_{l,1}^{\pi^k}$ by $V_{l,1}^k$. The Bellman equation states that

$$Q_{r,h}^k(s_h, a_h) = r_h(s_h, a_h) + \mathbb{P}_h V_{r,h+1}^k(s_h, a_h).$$

To estimate the value function $Q_{r,h}^k$, we estimate the second part $\mathbb{P}_h V_{r,h+1}^k$ by $(\phi_{r,h}^k)^\top w_{r,h}^k$, which can be seen in many linear mixture MDPs works [Ayoub et al., 2020; Zhou et al., 2021], where $w_{r,h}^k$ can be solved via least squares. Specifically, we solve the following regularized least squares for $w_{r,h}^k$:

$$w_{r,h}^k = \arg\min_w \sum_{\tau=1}^{k-1}\left(V_{r,h+1}^\tau\left(s_{h+1}^\tau\right) - \phi_{r,h}^\tau\left(s_h^\tau, a_h^\tau\right)^\top w\right)^2 + \lambda\|w\|_2^2, \tag{4.12}$$

where

$$\phi_{r,h}^\tau(\cdot, \cdot) := \int_{\mathcal{S}} \psi\left(\cdot, \cdot, s'\right) V_{r,h+1}^\tau\left(s'\right) ds',$$

and $V_{r,h+1}^\tau(\cdot) = \left\langle Q_{r,h+1}^\tau(\cdot, \cdot), \pi_{h+1}^\tau(\cdot \mid \cdot)\right\rangle_{\mathcal{A}}$ for $h \in [H-1]$. Analogously, we estimate $\mathbb{P}_h V_{c,h}^k$ by $(\phi_{c,h}^k)^\top w_{c,h}^k$, solved from:

$$w_{c,h}^k = \arg\min_w \sum_{\tau=1}^{k-1}\left(V_{c,h+1}^\tau\left(s_{h+1}^\tau\right) - \phi_{c,h}^\tau\left(s_h^\tau, a_h^\tau\right)^\top w\right)^2 + \lambda\|w\|_2^2. \tag{4.13}$$

Further, we estimate r_h by $\varphi^\top u_{r,h}^k$ in which we solve $u_{r,h}^k$ from:

$$u_{r,h}^k = \arg\min_u \sum_{\tau=1}^{k-1} \left(r_h\left(s_{h+1}^\tau\right) - \varphi\left(s_h^\tau, a_h^\tau\right)^\top u\right)^2 + \lambda\|u\|_2^2. \tag{4.14}$$

Analogously we estimate c_h by $\varphi^\top u_{c,h}^k$ in which we solve $u_{c,h}^k$ from:

$$u_{c,h}^k = \arg\min_u \sum_{\tau=1}^{k-1} \left(c_h\left(s_{h+1}^\tau\right) - \varphi\left(s_h^\tau, a_h^\tau\right)^\top u\right)^2 + \lambda\|u\|_2^2. \tag{4.15}$$

Algorithm 3: Optimistic Primal-Dual Proximal Policy Optimization [Ding et al., 2020]

```
Input  : Let $\left\{Q_{r,h}^0, Q_{c,h}^0\right\}_{h=1}^H$ be zero functions, $\{\pi_h^0\}_{h\in[H]}$ be uniform distributions on
         $\mathcal{A}, V_{c,1}^0$ be $b, \lambda^0$ be $0, \chi$ be $2H/\gamma, \alpha, \eta > 0, \theta \in (0,1]$.
for episode $k = 1, \ldots, K+1$ do
    Set the initial state $s_1^k = s_1$. for step $h = 1, 2, \ldots, H$ do
        Mix the policy
            $\widetilde{\pi}_h^{k-1}(\cdot \mid \cdot) \leftarrow (1-\theta)\pi_h^{k-1}(\cdot \mid \cdot) + \theta\,\mathrm{Unif}(\mathcal{A})$
        Update the policy
            $\pi_h^k(\cdot \mid \cdot) \propto \widetilde{\pi}_h^{k-1}(\cdot \mid \cdot)\mathrm{e}^{\left(\alpha\left(Q_{r,h}^{k-1}+\lambda^{k-1}Q_{c,h}^{k-1}\right)(\cdot,\cdot)\right)}$
        Take an action $a_h^k \sim \pi_h^k\left(\cdot \mid s_h^k\right)$ and receive reward and utility
          $r_h\left(s_h^k, a_h^k\right), c_h\left(s_h^k, a_h^k\right)$
        Observe the next state $s_{h+1}^k$
    Update the dual variable $\lambda^k$ by
        $\lambda^k \leftarrow \mathrm{Proj}_{[0,\chi]}\left(\lambda^{k-1} + \eta\left(b - V_{c,1}^{k-1}\left(s_1\right)\right)\right)$
    Estimate the action-value or value functions $\left\{Q_{r,h}^k(\cdot,\cdot), Q_{c,h}^k(\cdot,\cdot), V_{c,h}^k(\cdot)\right\}_{h=1}^H$ via
        $\mathrm{LSTD}\left(\left\{s_h^\tau, a_h^\tau, r_h\left(s_h^\tau, a_h^\tau\right), c_h\left(s_h^\tau, a_h^\tau\right)\right\}_{h,\tau=1}^{H,k}\right)$
```

Finally, to encourage exploration, we add upper confidence bound (UCB) type bonus to our estimates of (constraint) value functions and (constraint) reward functions so that

Algorithm 4: Least-Squares Temporal Difference (LSTD) with UCB exploration [Ding et al., 2020]

```
Input  : $\{s_h^\tau, a_h^\tau, r_h\left(s_h^\tau, a_h^\tau\right), c_h\left(s_h^\tau, a_h^\tau\right)\}_{h,\tau=1}^{H,k}$.
1  Set $\left\{V_{r,H+1}^k, V_{c,H+1}^k\right\}$ be zero functions and $\lambda = 1, \beta = O\left(\sqrt{dH^2\log(dT/p)}\right)$
2  for step $h = H, H-1, \ldots, 1$ do
3      $\Lambda_{l,h}^k \leftarrow \sum_{\tau=1}^{k-1} \phi_{l,h}^\tau\left(s_h^\tau, a_h^\tau\right) \phi_{l,h}^\tau\left(s_h^\tau, a_h^\tau\right)^\top + \lambda I$
4      $w_{l,h}^k \leftarrow \left(\Lambda_{l,h}^k\right)^{-1} \sum_{\tau=1}^{k-1} \phi_{l,h}^\tau\left(s_h^\tau, a_h^\tau\right) V_{l,h+1}^\tau\left(s_{h+1}^\tau\right)$
5      $\phi_{l,h}^k(\cdot,\cdot) \leftarrow \int_{\mathcal{S}} \psi\left(\cdot,\cdot,s'\right) V_{l,h+1}^k\left(s'\right) ds'$
6      $\Gamma_{l,h}^k(\cdot,\cdot) \leftarrow \beta\left(\phi_{l,h}^k(\cdot,\cdot)^\top \left(\Lambda_{l,h}^k\right)^{-1} \phi_{l,h}^k(\cdot,\cdot)\right)^{1/2}$
7      $\Lambda_h^k \leftarrow \sum_{\tau=1}^{k-1} \varphi\left(s_h^\tau, a_h^\tau\right) \varphi\left(s_h^\tau, a_h^\tau\right)^\top + \lambda I$
8      $u_{l,h}^k \leftarrow \left(\Lambda_h^k\right)^{-1} \sum_{\tau=1}^{k-1} \varphi\left(s_h^\tau, a_h^\tau\right) l_h\left(s_h^\tau, a_h^\tau\right)$
9      $\Gamma_h^k(\cdot,\cdot) \leftarrow \beta\left(\varphi(\cdot,\cdot)^\top \left(\Lambda_h^k\right)^{-1} \varphi(\cdot,\cdot)\right)^{1/2}$
10     $Q_{l,h}^k(\cdot,\cdot) \leftarrow \min\left(\varphi(\cdot,\cdot)^\top u_{l,h}^k + \phi_{l,h}^k(\cdot,\cdot)^\top w_{l,h}^k + \left(\Gamma_h^k + \Gamma_{l,h}^k\right)(\cdot,\cdot), H-h+1\right)^+$
11     $V_{l,h}^k(\cdot) \leftarrow \left\langle Q_{l,h}^k(\cdot,\cdot), \pi_h^k(\cdot \mid \cdot)\right\rangle_{\mathcal{A}}$.
Output: $\left\{Q_{l,h}^k(\cdot,\cdot), V_{l,h}^k(\cdot,\cdot)\right\}_{h=1}^H$
```

the estimates are optimistic with high probability. Denote the bonus for (constraint) value functions and (constraint) reward functions by $\Gamma_{l,h}^k$ and Γ_h^k respectively. Specifically, the bonus are given by:

$$\Gamma_h^k = \beta\left(\varphi^\top \left(\Lambda_h^k\right)^{-1} \varphi\right)^{1/2}, \tag{4.16}$$

and

$$\Gamma_{l,h}^k = \beta\left(\left(\phi_{l,h}^k\right)^\top \left(\Lambda_{l,h}^k\right)^{-1} \phi_{l,h}^k\right)^{1/2}, \tag{4.17}$$

where $\beta > 0$ is the coefficient.

Putting everything together, we have the main algorithm in Algorithm line 3. For each episode, it performs policy improvement in line line 3 and line 4 at each time step, and then it performs the standard dual variable update in line line 7 with gradient descent. At the end of each episode, the main algorithm performs the policy evaluation to calculate estimates by calling to the least-squares temporal difference procedure in Algorithm algorithm 4. The LSTD procedure implements the policy evaluation step by solving all the least-squares problems defined in Section section 4.2.2. The closed form expression is given by fundamental regularized linear regression.

Theorem 4.2.1. *(Theorem 1 of Ding et al. [2020]) Let Assumptions assumption 1.4.1 and assumption 4.1.2 hold. Fix* $p \in (0, 1)$*. We set* $\alpha = \sqrt{\log|\mathcal{A}|/(H^2K)}$, $\beta = C_1\sqrt{dH^2\log(dT/p)}$, $\eta = 1/\sqrt{K}$, $\theta = 1/K$, *and* $\lambda = 1$ *in Algorithm line 3, where* C_1 *is an absolute constant. Suppose* $\log|\mathcal{A}| = O(d^2\log^2(dT/p))$*. Then, with probability* $1-p$*, the regret and the constraint violation satisfy*

$$\textit{Regret}(K) \leq CdH^{2.5}\sqrt{T}\log\left(\frac{dT}{p}\right)$$

$$\textit{Violation}(K) \leq C'dH^{2.5}\sqrt{T}\log\left(\frac{dT}{p}\right)$$

where C *and* C' *are absolute constants.*

To prove theorem 4.2.1 we need some preliminary analysis and results. First of all, we decompose the regret to break it down to terms that are easier to analyze. Recall the definition of the regret in eq. (4.4):

$$\text{Regret}(K) = \sum_{k=1}^{K}\left(V_{r,1}^{\pi^*}(s_1) - V_{r,1}^{\pi^k}(s_1)\right).$$

It is not easy to directly relate the values $V_{r,1}^{\pi^k}$ of policies π^ks to the value of the optimal policy π^*, but the algorithm design of Algorithm line 3 makes it easier to relate the values $V_{r,1}^{\pi^k}$ to the estimate value functions $V_{r,1}^k$ returned by Algorithm algorithm 4, where the policies π^k are updated according to the estimate values. Hence, we add and subtract the estimate values:

$$\text{Regret}(K) = \underbrace{\sum_{k=1}^{K}\left(V_{r,1}^{\pi^*}(s_1) - V_{r,1}^{k}(s_1)\right)}_{\text{(R.I)}} + \underbrace{\sum_{k=1}^{K}\left(V_{r,1}^{k}(s_1) - V_{r,1}^{\pi^k}(s_1)\right)}_{\text{(R.II)}}. \tag{4.18}$$

We use the following lemma to expand term (R.I):

Lemma 4.2.1. *For reward l being either reward r or constraint reward c,*

$$\begin{aligned}\sum_{k=1}^{K}\left(V_{l,1}^{\pi^*}(s_1)-V_{l,1}^{k}(s_1)\right)=&\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\left\langle Q_{l,h}^{k}\left(s_h,\cdot\right),\pi_h^*\left(\cdot\mid s_h\right)-\pi_h^k\left(\cdot\mid s_h\right)\right\rangle\right]\\&+\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\iota_{l,h}^{k}\left(s_h,a_h\right)\right],\end{aligned}\tag{4.19}$$

where $\iota_{l,h}^k:=l_h+\mathbb{P}_hV_{l,h+1}^k-Q_{l,h}^k$.

To prove lemma 4.2.1, we need to use the Bellman equations in form of eq. (4.1) and eq. (4.2) to rewrite the left hand side as

$$V_{l,h}^{\pi^*}(s)-V_{l,h}^{k}(s)=\left\langle\mathbb{P}_h\left(V_{l,h+1}^{\pi^*}-V_{l,h+1}^{k}\right)(s,\cdot),\pi_h^*(\cdot\mid s)\right\rangle+\left\langle\iota_{l,h}^{k}(s,\cdot),\pi_h^*(\cdot\mid s)\right\rangle+\xi_h^k(s),$$

where $\xi_h^k(s):=\langle Q_h^k(s,\cdot),\pi_h^*(\cdot|s)-\pi_h^k(\cdot|s)\rangle$. Note that the above equation is recursive. Recursively expanding the expression, where for the first term the terminal case is $V_{r,H+1}=0$ for any policy, gives us the desired result.

To expand (R.II) we need to introduce some filtrations. For any $k\in[K],h\in[H]$, define $\mathcal{F}_{h,1}^k$ as a σ-algebra generated by state-action sequences:

$$\{(s_i^\tau,a_i^\tau)\}_{(\tau,i)\in[k-1]\times[H]}\bigcup\left\{\left(s_i^k,a_i^k\right)\right\}_{i\in[h]}.$$

Additionally, for any $k\in[K],h\in[H]$, define $\mathcal{F}_{h,2}^k$ as a σ-algebra generated by state-action sequences with an extra next state:

$$\{(s_i^\tau,a_i^\tau)\}_{(\tau,i)\in[k-1]\times[H]}\bigcup\left\{\left(s_i^k,a_i^k\right)\right\}_{i\in[h]}\bigcup\left\{s_{h+1}^k\right\}.$$

Let

$$\begin{aligned}N_{r,h,1}^k&:=\left\langle\left(Q_{r,h}^k-Q_{r,h}^{\pi^k}\right)\left(s_h^k,\cdot\right),\pi_h^k(\cdot,s_h^k)\right\rangle-\left(Q_{r,h}^k-Q_{r,h}^{\pi^k}\right)\left(s_h^k,a_h^k\right),\\N_{r,h,2}^k&:=\left(\mathbb{P}_hV_{r,h+1}^k-\mathbb{P}_hV_{r,h+1}^{\pi^k}\right)\left(s_h^k,a_h^k\right)-\left(V_{r,h+1}^k-V_{r,h+1}^{\pi^k}\right)\left(s_{h+1}^k\right),\end{aligned}$$

and define the martingale sequence $M^k_{r,h,m}$ as:

$$M^k_{r,h,m} = \sum_{\tau=1}^{k-1}\sum_{i=1}^{H}\left(N^\tau_{r,i,1} + N^\tau_{r,i,2}\right) + \sum_{i=1}^{h-1}\left(N^k_{r,i,1} + N^k_{r,i,2}\right) + \sum_{\ell=1}^{m} N^k_{r,h,\ell}. \tag{4.20}$$

Note that by the definition in eq. (4.20), in particular we have

$$M^K_{r,H,2} = \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{\ell=1}^{2} N^k_{r,h,\ell} = \sum_{k=1}^{K}\sum_{h=1}^{H}(N^k_{r,h,1} + N^k_{r,h,2}).$$

Now, we introduce the following lemma to expand term (R.II):

Lemma 4.2.2.

$$\sum_{k=1}^{K}\left(V^k_{r,1}(s_1) - V^{\pi^k}_{r,1}(s_1)\right) = -\sum_{k=1}^{K}\sum_{h=1}^{H}\iota^k_{r,h}(s^k_h, a^k_h) + M^K_{r,H,2}. \tag{4.21}$$

Lemma 4.2.3. *[Policy Improvement: Primal-Dual Mirror Descent Step [Ding et al., 2020]] Let Assumption assumption 1.4.1 and Assumption assumption 4.1.1 hold. In Algorithm line 3, if we set $\alpha = \sqrt{\log|\mathcal{A}|}/\left(H^2\sqrt{K}\right)$ and $\theta = 1/K$, then*

$$\begin{aligned}&\sum_{k=1}^{K}\left(V^{\pi^*}_{r,1}(s_1) - V^k_{r,1}(s_1)\right) + \sum_{k=1}^{K}\lambda^k\left(V^{\pi^*}_{c,1}(s_1) - V^k_{c,1}(s_1)\right)\\ &\leq C_2 H^{2.5}\sqrt{T\log|\mathcal{A}|} + \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\iota^k_{r,h}(s_h, a_h)\right] + \lambda^k\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[\iota^k_{c,h}(s_h, a_h)\right].\end{aligned} \tag{4.22}$$

where C_2 is an absolute constant and $T = HK$.

Proof. Recall that line line 4 of Algorithm line 3 follows a solution π^k to the following problem,

$$\max_\pi \sum_{h=1}^{H}\left\langle Q^{k-1}_{r,h} + \lambda^{k-1}Q^{k-1}_{c,h}, \pi_h\right\rangle - \frac{1}{\alpha}\sum_{h=1}^{H} D\left(\pi_h \mid \widetilde{\pi}^{k-1}_h\right) \tag{4.23}$$

where we use the shorthand $\left\langle Q^{k-1}_{r,h} + \lambda^{k-1}Q^{k-1}_{c,h}, \pi_h\right\rangle$ for $\left\langle\left(Q^{k-1}_{r,h} + \lambda^{k-1}Q^{k-1}_{c,h}\right)(s_h,\cdot), \pi_h(\cdot \mid s_h)\right\rangle$ and the shorthand $D\left(\pi_h \mid \widetilde{\pi}^{k-1}_h\right)$ for $D\left(\pi_h(\cdot \mid s_h) \mid \widetilde{\pi}^{k-1}_h(\cdot \mid s_h)\right)$ if dependence on the state-action sequence $\{s_h, a_h\}_{h=1}^{H}$ is clear from context. We note that eq. (4.23) is in form of a mirror descent problem in lemma 4.2.7. We can apply the pushback property with $x^* = \pi_h^k, y = \widetilde{\pi}_h^{k-1}$ and $z = \pi_h^*$,

$$\begin{aligned}
&\sum_{h=1}^{H} \left\langle Q_{r,h}^{k-1} + \lambda^{k-1} Q_{c,h}^{k-1}, \pi_h^k \right\rangle - \frac{1}{\alpha} \sum_{h=1}^{H} D\left(\pi_h^k \mid \widetilde{\pi}_h^{k-1}\right) \\
\geq &\sum_{h=1}^{H} \left\langle Q_{r,h}^{k-1} + \lambda^{k-1} Q_{c,h}^{k-1}, \pi_h^* \right\rangle - \frac{1}{\alpha} \sum_{h=1}^{H} D\left(\pi_h^* \mid \widetilde{\pi}_h^{k-1}\right) + \frac{1}{\alpha} \sum_{h=1}^{H} D\left(\pi_h^* \mid \pi_h^k\right).
\end{aligned}$$

Equivalently, we write the above inequality as follows,

$$\begin{aligned}
&\sum_{h=1}^{H} \left\langle Q_{r,h}^{k-1}, \pi_h^* - \pi_h^{k-1} \right\rangle + \lambda^{k-1} \sum_{h=1}^{H} \left\langle Q_{c,h}^{k-1}, \pi_h^* - \pi_h^{k-1} \right\rangle \\
\leq &\sum_{h=1}^{H} \left\langle Q_{r,h}^{k-1} + \lambda^{k-1} Q_{c,h}^{k-1}, \pi_h^k - \pi_h^{k-1} \right\rangle - \frac{1}{\alpha} \sum_{h=1}^{H} D\left(\pi_h^k \mid \widetilde{\pi}_h^{k-1}\right) \\
+ &\frac{1}{\alpha} \sum_{h=1}^{H} D\left(\pi_h^* \mid \widetilde{\pi}_h^{k-1}\right) - \frac{1}{\alpha} \sum_{h=1}^{H} D\left(\pi_h^* \mid \pi_h^k\right).
\end{aligned} \tag{4.24}$$

By taking expectation $\mathbb{E}_{\pi^*}$ on both sides of eq. (4.24) over the random state-action sequence $\{(s_h, a_h)\}_1^{H}$*ting from s_1, and applying lemma 4.2.1, we have

$$\begin{aligned}
&\left(V_{r,1}^{\pi^*}(s_1) - V_{r,1}^{k-1}(s_1)\right) + \lambda^{k-1} \left(V_{c,1}^{\pi^*}(s_1) - V_{c,1}^{k-1}(s_1)\right) \\
\leq &\sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[\left\langle Q_{r,h}^{k-1} + \lambda^{k-1} Q_{c,h}^{k-1}, \pi_h^k - \pi_h^{k-1} \right\rangle\right] - \frac{1}{\alpha} \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[D\left(\pi_h^k \mid \widetilde{\pi}_h^{k-1}\right)\right] \\
&+ \frac{1}{\alpha} \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[D\left(\pi_h^* \mid \widetilde{\pi}_h^{k-1}\right) - D\left(\pi_h^* \mid \pi_h^k\right)\right] \\
&+ \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[\iota_{r,h}^{k-1}(s_h, a_h)\right] + \lambda^{k-1} \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[\iota_{c,h}^{k-1}(s_h, a_h)\right].
\end{aligned} \tag{4.25}$$

The rest is to bound the right-hand side of the above inequality. By the Hölder's inequality and the Pinsker's inequality, we first have

$$
\begin{aligned}
&\sum_{h=1}^{H}\left\langle Q_{r,h}^{k-1}+\lambda^{k-1}Q_{c,h}^{k-1}, \pi_h^k-\pi_h^{k-1}\right\rangle-\frac{1}{\alpha}\sum_{h=1}^{H}D\left(\pi_h^k \mid \widetilde{\pi}_h^{k-1}\right)\\
&=\sum_{h=1}^{H}\left\langle Q_{r,h}^{k-1}+\lambda^{k-1}Q_{c,h}^{k-1}, \pi_h^k-\widetilde{\pi}_h^{k-1}\right\rangle-\frac{1}{\alpha}\sum_{h=1}^{H}D\left(\pi_h^k \mid \widetilde{\pi}_h^{k-1}\right)\\
&\quad+\sum_{h=1}^{H}\left\langle Q_{r,h}^{k-1}+\lambda^{k-1}Q_{c,h}^{k-1}, \widetilde{\pi}_h^{k-1}-\pi_h^{k-1}\right\rangle\\
&\leq\sum_{h=1}^{H}\left(\left\|Q_{r,h}^{k-1}+\lambda^{k-1}Q_{c,h}^{k-1}\right\|_\infty\left\|\pi_h^k-\widetilde{\pi}_h^{k-1}\right\|_1-\frac{1}{2\alpha}\left\|\pi_h^k-\widetilde{\pi}_h^{k-1}\right\|_1^2\right)\\
&\quad+\sum_{h=1}^{H}\left\|Q_{r,h}^{k-1}+\lambda^{k-1}Q_{c,h}^{k-1}\right\|_\infty\left\|\widetilde{\pi}_h^{k-1}-\pi_h^{k-1}\right\|_1.
\end{aligned}
$$

Then, using the square completion,

$$
\begin{aligned}
&\left\|Q_{r,h}^{k-1}+\lambda^{k-1}Q_{c,h}^{k-1}\right\|_\infty\left\|\pi_h^k-\widetilde{\pi}_h^{k-1}\right\|_1-\frac{1}{2\alpha}\left\|\pi_h^k-\widetilde{\pi}_h^{k-1}\right\|_1^2\\
&=-\frac{1}{2\alpha}\left(\alpha\left\|Q_{r,h}^{k-1}+\lambda^{k-1}Q_{c,h}^{k-1}\right\|_\infty-\left\|\pi_h^k-\widetilde{\pi}_h^{k-1}\right\|_1\right)^2+\frac{\alpha}{2}\left\|Q_{r,h}^{k-1}+\lambda^{k-1}Q_{c,h}^{k-1}\right\|_\infty^2\\
&\leq\frac{\alpha}{2}\left\|Q_{r,h}^{k-1}+\lambda^{k-1}Q_{c,h}^{k-1}\right\|_\infty^2,
\end{aligned}
$$

where we drop off the first quadratic term for the inequality, and $\left\|\widetilde{\pi}_h^{k-1}-\pi_h^{k-1}\right\|_1 \leq \theta$, we have

$$
\begin{aligned}
&\sum_{h=1}^{H}\left\langle Q_{r,h}^{k-1}+\lambda^{k-1}Q_{c,h}^{k-1}, \pi_h^k-\pi_h^{k-1}\right\rangle-\frac{1}{\alpha}\sum_{h=1}^{H}D\left(\pi_h^k \mid \pi_h^{k-1}\right)\\
&\leq\frac{\alpha}{2}\sum_{h=1}^{H}\left\|Q_{r,h}^{k-1}+\lambda^{k-1}Q_{c,h}^{k-1}\right\|_\infty^2+\theta\sum_{h=1}^{H}\left\|Q_{r,h}^{k-1}+\lambda^{k-1}Q_{c,h}^{k-1}\right\|_\infty\\
&\leq\frac{\alpha(1+\chi)^2H^3}{2}+\theta(1+\chi)H^2,
\end{aligned}
\tag{4.26}
$$

where the last inequality is due to $\left\|Q_{r,h}^{k-1}\right\|_\infty \leq H$, a result from line line 11 in Algorithm algorithm 4, and $0 \leq \lambda^{k-1} \leq \chi$. Taking the same expectation $\mathbb{E}_{\pi^*}$ as previously on both sides of eq. (4.26) and substituting it into the left-hand side of eq. (4.25) yield,

$$\begin{aligned}
&\left(V_{r,1}^{\pi^*}\left(s_1\right)-V_{r,1}^{k-1}\left(s_1\right)\right)+\lambda^{k-1}\left(V_{c,1}^{\pi^*}\left(s_1\right)-V_{c,1}^{k-1}\left(s_1\right)\right) \\
&\leq \frac{\alpha(1+\chi)^2 H^3}{2}+\theta(1+\chi) H^2+\frac{1}{\alpha} \sum_{h=1}^H \mathbb{E}_{\pi^*}\left[D\left(\pi_h^* \mid \widetilde{\pi}_h^{k-1}\right)-D\left(\pi_h^* \mid \pi_h^k\right)\right] \\
&\quad+\sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\iota_{r,h}^{k-1}\left(s_h, a_h\right)\right]+\lambda^{k-1} \sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\iota_{c,h}^{k-1}\left(s_h, a_h\right)\right] \\
&\leq \frac{\alpha(1+\chi)^2 H^3}{2}+\theta(1+\chi) H^2+\frac{\theta H \log |\mathcal{A}|}{\alpha}+\frac{1}{\alpha} \sum_{h=1}^H \mathbb{E}_{\pi^*}\left[D\left(\pi_h^* \mid \pi_h^{k-1}\right)-D\left(\pi_h^* \mid \pi_h^k\right)\right] \\
&\quad+\sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\iota_{r,h}^{k-1}\left(s_h, a_h\right)\right]+\lambda^{k-1} \sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\iota_{c,h}^{k-1}\left(s_h, a_h\right)\right] .
\end{aligned} \tag{4.27}$$

where in the second inequality we note the fact that $D\left(\pi_h^* \mid \widetilde{\pi}_h^{k-1}\right)-D\left(\pi_h^* \mid \pi_h^{k-1}\right) \leq \theta \log |\mathcal{A}|$ from lemma 4.2.8.

We note that λ^0 is initialized to be zero. By taking a telescoping sum of both sides of eq. (4.27) from $k=1$ to $k=K+1$ and shifting the index k by one, we have

$$\begin{aligned}
&\sum_{k=1}^K\left(V_{r,1}^{\pi^*}\left(s_1\right)-V_{r,1}^k\left(s_1\right)\right)+\sum_{k=1}^K \lambda^k\left(V_{c,1}^{\pi^*}\left(s_1\right)-V_{c,1}^k\left(s_1\right)\right) \\
&\leq \frac{\alpha(1+\chi)^2 H^3(K+1)}{2}+\theta(1+\chi) H^2(K+1)+\frac{\theta H(K+1) \log |\mathcal{A}|}{\alpha}+\frac{H \log |\mathcal{A}|}{\alpha} \\
&\quad+\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\iota_{r,h}^k\left(s_h, a_h\right)\right]+\lambda^k \sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\pi^*}\left[\iota_{c,h}^k\left(s_h, a_h\right)\right] .
\end{aligned}$$

where we ignore $-\alpha^{-1} \sum_{h=1}^H \mathbb{E}_{\pi^*}\left[D\left(\pi_h^* \mid \pi_h^{K+1}\right)\right]$ and utilize

$$D\left(\pi_h^* \mid \pi_h^0\right)=\sum_{a \in \mathcal{A}} \pi_h^*\left(a \mid s_h\right) \log \left(|\mathcal{A}| \pi_h^*\left(a \mid s_h\right)\right) \leq \log |\mathcal{A}|$$

where π_h^0 is uniform over $\mathcal{A}$ and we ignore $\sum_{a \in \mathcal{A}} \pi_h^*\left(a \mid s_h\right) \log \left(\pi_h^*\left(a \mid s_h\right)\right)$ that is nonpositive. Finally, we take $\chi:=H / \gamma$ and α, θ in the lemma to complete the proof. □

Lemma 4.2.4. *Let Assumption assumption 1.4.1 and Assumption assumption 4.1.1 hold. In*

Algorithm line 3, if we set $\alpha = \sqrt{\log|\mathcal{A}|}/\left(H^2\sqrt{K}\right), \eta = 1/\sqrt{K}$, *and* $\theta = 1/K$, *then*

$$\text{Regret}(K) = C_3 H^{2.5}\sqrt{T\log|\mathcal{A}|} + \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^*}\left[\iota_{r,h}^{k}\left(s_h, a_h\right)\right] - \iota_{r,h}^{k}\left(s_h^k, a_h^k\right)\right) + M_{r,H,2}^{K}, \quad (4.28)$$

where C_3 *is an absolute constant.*

The proof to lemma 4.2.4 first bounds the second term of the left hand side in eq. (4.22) and then combines it with lemma 4.2.3 to obtain the desired results. The first part is constructed based on the update rule of the dual variables λ^ks and the fact that the dual variables are projections and thus bounded. We have seen similar techniques in the proof to theorem 2.0.1. We show the first part of the proof in details: by the dual update in Algorithm line 3, we have

$$\begin{aligned}
0 &\leq \left(\lambda^{K+1}\right)^2 \\
&= \sum_{k=1}^{K+1}\left(\left(\lambda^k\right)^2 - \left(\lambda^{k-1}\right)^2\right) \\
&= \sum_{k=1}^{K+1}\left(\text{Proj}_{[0,\chi]}\left(\lambda^{k-1} + \eta\left(b - V_{c,1}^{k-1}\left(s_1\right)\right)\right)\right)^2 - \left(\lambda^{k-1}\right)^2 \\
&\leq \sum_{k=1}^{K+1}\left(\lambda^{k-1} + \eta\left(b - V_{c,1}^{k-1}\left(s_1\right)\right)\right)^2 - \left(\lambda^{k-1}\right)^2 \\
&\leq \sum_{k=1}^{K+1} 2\eta\lambda^{k-1}\left(V_{c,1}^{\pi^*}\left(s_1\right) - V_{c,1}^{k-1}\left(s_1\right)\right) + \eta^2\left(b - V_{c,1}^{k-1}\left(s_1\right)\right)^2,
\end{aligned}$$

where we use the feasibility of π^* in the last inequality. Since $\lambda^0 = 0$ and $\left|b - V_{c,1}^{k-1}\left(x_1\right)\right| \leq H$, the above inequality implies that

$$-\sum_{k=1}^{K}\lambda^k\left(V_{c,1}^{\pi^*}\left(s_1\right) - V_{c,1}^{k}\left(s_1\right)\right) \leq \sum_{k=1}^{K+1}\frac{\eta}{2}\left(b - V_{c,1}^{k-1}\left(s_1\right)\right)^2 \leq \frac{\eta H^2(K+1)}{2}.$$

We note that with lemma 4.2.4, it remains to bound two more terms on the right hand side of eq. (4.28) to get our regret bound. We bound each of the two terms in the following two lemmas.

Lemma 4.2.5. *Let Assumption assumption 4.1.1 hold. Fix* $p \in (0,1)$*. If we set* $\beta = C_1\sqrt{dH^2\log(dT/p)}$ *in Algorithm line 3, then with probability* $1-p/2$ *it holds that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^*}\left[\iota_{r,h}^k\left(s_h,a_h\right)\right]-\iota_{r,h}^k\left(s_h^k,a_h^k\right)\right)\leq 4C_1\sqrt{2d^2H^3T\log(K+1)\log\left(\frac{dT}{p}\right)}$$

where C_1 *is an absolute constant and* $T = HK$*.*

Recall that $\iota_{r,h}^k(s,a) := r_h(s,a) + \mathbb{P}_h V_{r,h+1}^k(s,a) - Q_{r,h}^k(s,a)$ is defined as the model prediction error at state-action pair (s_h, a_h), incurred partially by the error in transition estimates. Note that due to UCB style bonus in Algorithm algorithm 4, in particular $(\Gamma_{l,h}^k + \Gamma_h^k)$, with high probability this error can be bounded as such: for any state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, for any $k \in [K], h \in [H]$, with probability $1-p/2$,

$$-2(\Gamma_h^k + \Gamma_{r,h}^k)(s,a) \leq \iota_{r,h}^k(s,a) \leq 0, \tag{4.29}$$

where p appears in the UCB bonus coefficient $\beta := C_1\sqrt{dH^2\log(dT/p)}$ in eq. (4.16) and eq. (4.17). Hence, the proof to lemma 4.2.5 first relates the model prediction error $\iota_{r,h}^k$ to the UCB bonus terms $(\Gamma_h^k + \Gamma_{r,h}^k)(s,a)$, and then bounding these bonus terms by elliptical potential lemma (lemma 4.2.9) completes the proof. Finally, we use the last lemma below to bound the last term in eq. (4.28).

Lemma 4.2.6. *Fix* $p \in (0,1)$*. In Algorithm line 3, it holds with probability* $1-p/2$ *that*

$$\left|M_{r,H,2}^K\right| \leq 4\sqrt{H^2T\log\left(\frac{4}{p}\right)},$$

where $T = HK$*.*

The proof to lemma 4.2.6 applies the Azuma-Hoeffding inequality to the martingale sequence $M_{r,H,2}^K$ which is bounded by showing that $N_{r,h,1}^k$ and $N_{r,h,2}^k$ are bounded for all

$k \in [K], h \in [H]$.

Lemma 4.2.7. *(Pushback property of KL-divergence [Wei et al., 2020]). Let $f : \Delta \to \mathbb{R}$ be a concave function where Δ is a probability simplex in $\mathbb{R}^d$. Let Δ^o be the interior of Δ. Let $x^* = \operatorname{argmax}_{x \in \Delta} f(x) - \alpha^{-1} D(x, y)$ for a fixed $y \in \Delta^o$ and $\alpha > 0$. Then, for any $z \in \Delta$,*

$$f\left(x^*\right) - \frac{1}{\alpha} D\left(x^*, y\right) \geq f(z) - \frac{1}{\alpha} D(z, y) + \frac{1}{\alpha} D\left(z, x^*\right).$$

Lemma 4.2.8. *(Bounded KL-divergence Difference [Wei et al., 2020]). Let π_1, π_2 be two probability distributions in $\Delta(\mathcal{A})$. Let $\widetilde{\pi}_2 = (1-\theta)\pi_2 + \mathbf{1}\theta/|\mathcal{A}|$ where $\theta \in (0, 1]$. Then,*

$$D\left(\pi_1 \mid \widetilde{\pi}_2\right) - D\left(\pi_1 \mid \pi_2\right) \leq \theta \log |\mathcal{A}|.$$

Moreover, we have an uniform bound, $D\left(\pi_1 \mid \widetilde{\pi}_2\right) \leq \log(|\mathcal{A}|/\theta)$.

Lemma 4.2.9. *(Elliptical Potential Lemma [Abbasi-yadkori et al., 2011]) Let $\{\phi_t\}_{t=1}^{\infty}$ be a sequence of functions in $\mathbb{R}^d$ and $\Lambda_0 \in \mathbb{R}^{d \times d}$ be a positive definite matrix. Let $\Lambda_t = \Lambda_0 + \sum_{i=1}^{t-1} \phi_i \phi_i^\top$. Assume $\|\phi_t\|_2 \leq 1$ and $\lambda_{\min}\left(\Lambda_0\right) \geq 1$. Then for any $t \geq 1$ it holds that*

$$\log\left(\frac{\det\left(\Lambda_{t+1}\right)}{\det\left(\Lambda_1\right)}\right) \leq \sum_{i=1}^{t} \phi_i^\top \Lambda_i^{-1} \phi_i \leq 2 \log\left(\frac{\det\left(\Lambda_{t+1}\right)}{\det\left(\Lambda_1\right)}\right).$$

Bibliography

Y. Abbasi-yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, 2011.

J. Achiam, D. Held, A. Tamar, and P. Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 2017.

A. Agarwal, S. Kakade, and L. F. Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Proceedings of Thirty Third Conference on Learning Theory*, 2020.

E. Altman. *Constrained Markov Decision Processes*. CRC Press, 1999.

A. Ayoub, Z. Jia, C. Szepesvári, M. Wang, and L. Yang. Model-based reinforcement learning with value-targeted regression. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

V. Borkar. An actor-critic algorithm for constrained markov decision processes. *Systems and Control Letters*, 2005.

K. Brantley, M. Dudík, T. Lykouris, S. Miryoosefi, M. Simchowitz, A. Slivkins, and W. Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

Q. Cai, Z. Yang, C. Jin, and Z. Wang. Provably efficient exploration in policy optimization. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

D. Ding, X. Wei, Z. Yang, Z. Wang, and M. R. Jovanović. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, 2020.

A. Jain, S. Vaswani, R. B. Harikandeh, C. Szepesvári, and D. Precup. Towards painless policy optimization for constrained MDPs. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of Thirty Third Conference on Learning Theory*, 2020.

S. Kakade and J. Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, 2002.

S. M. Kakade. On the sample complexity of reinforcement learning. Phd thesis, 2003.

J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 1960.

F. Orabona and D. Pál. Coin betting and parameter-free online learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016.

F. Orabona and T. Tommasi. Training deep networks without learning rates through coin betting. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

S. Paternain, L. F. O. Chamon, M. Calvo-Fullana, and A. Ribeiro. Constrained reinforcement learning has zero duality gap. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms, 2017.

A. Stooke, J. Achiam, and P. Abbeel. Responsive safety in reinforcement learning by pid lagrangian methods. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

R. Sutton and A. Barto. *Reinforcement Learning, second edition: An Introduction.* MIT Press, 2018.

C. Szepesvári. Rl theory lecture notes: Online planning - part i., 2023a. URL `https://rltheory.github.io/lecture-notes/planning-in-mdps/lec5/`.

C. Szepesvári. Rl theory lecture notes: Approximate policy iteration., 2023b. URL `https://rltheory.github.io/lecture-notes/planning-in-mdps/lec8/`.

C. Szepesvári. Rl theory lecture notes: The fundational theorem., 2023c. URL `https://rltheory.github.io/lecture-notes/planning-in-mdps/lec2/`.

C. Szepesvári. Rl theory lecture notes: Policy iteration., 2023d. URL `https://rltheory.github.io/lecture-notes/planning-in-mdps/lec4/`.

C. Tessler, D. J. Mankowitz, and S. Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2019.

S. Vaswani, L. Yang, and C. Szepesvári. Near-optimal sample complexity bounds for constrained MDPs. In *Advances in Neural Information Processing Systems*, 2022.

A. Wachi and Y. Sui. Safe reinforcement learning in constrained markov decision processes. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

H. Wei, X. Liu, and L. Ying. A provably-efficient model-free algorithm for constrained markov decision processes, 2021.

X. Wei, H. Yu, and M. J. Neely. Online primal-dual mirror descent under stochastic constraints. In *Abstracts of the 2020 SIGMETRICS/Performance Joint International Conference on Measurement and Modeling of Computer Systems*, 2020.

T. Xu, Y. Liang, and G. Lan. Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, 2020.

T.-Y. Yang, J. Rosca, K. Narasimhan, and P. J. Ramadge. Projection-based constrained policy optimization. In *International Conference on Learning Representations*, 2020.

D. Yin, B. Hao, Y. Abbasi-Yadkori, N. Lazić, and C. Szepesvári. Efficient local planning with linear function approximation. In *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, 2022.

D. Zhou, Q. Gu, and C. Szepesvári. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Proceedings of Thirty Fourth Conference on Learning Theory*, 2021.