Near-Optimal Sample Complexity Bounds for Constrained MDPs

Sharan Vaswani*
Simon Fraser University
vaswani.sharan@gmail.com

Lin F. Yang*
University of California, Los Angeles
linyang@ee.ucla.edu

Csaba Szepesvári
Amii, University of Alberta, DeepMind
szepesva@ualberta.ca

Abstract

In contrast to the advances in characterizing the sample complexity for solving Markov decision processes (MDPs), the optimal statistical complexity for solving constrained MDPs (CMDPs) remains unknown. We resolve this question by providing *minimax* upper and lower bounds on the sample complexity for learning near-optimal policies in a discounted CMDP with access to a generative model (simulator). In particular, we design a model-based algorithm that addresses two settings: (i) *relaxed feasibility*, where small constraint violations are allowed, and (ii) *strict feasibility*, where the output policy is required to satisfy the constraint. For (i), we prove that our algorithm returns an ε-optimal policy with probability $1-\delta$, by making $\tilde{O}\left(\frac{SA\log(1/\delta)}{(1-\gamma)^3\varepsilon^2}\right)$ queries to the generative model, thus matching the sample-complexity for unconstrained MDPs. For (ii), we show that the algorithm's sample complexity is upper-bounded by $\tilde{O}\left(\frac{SA\ \log(1/\delta)}{(1-\gamma)^5\ \varepsilon^2\zeta^2}\right)$ where ζ is the problem-dependent Slater constant that characterizes the size of the feasible region. Finally, we prove a matching lower-bound for the strict feasibility setting, thus obtaining the first near minimax optimal bounds for discounted CMDPs. Our results show that learning CMDPs is as easy as MDPs when small constraint violations are allowed, but inherently more difficult when we demand zero constraint violation.

1 Introduction

Common reinforcement learning (RL) algorithms focus on optimizing an unconstrained objective, and have found applications in games such as Atari [23] or Go [28], robot manipulation tasks [29, 37] or clinical trials [26]. However, many applications require the planning agent to satisfy constraints – for example, in wireless sensor networks [10] where there is a constraint on average power consumption. More generally, in the constrained Markov decision processes (CMDP) framework, the goal is to find a policy that maximizes the value associated with a reward function subject to the policy achieving a return (for a second reward function) that exceeds an apriori determined threshold [3]. There has been substantial work addressing the planning problem to find a near-optimal policy in a known CMDP [8, 7, 30, 24, 1, 35]. However, since the CMDP is unknown in most practical applications, we consider the problem of finding a near-optimal policy in this more challenging setting.

There have been multiple recent approaches to obtain a near-optimal policy in CMDPs in the regret-minimization or PAC-RL settings [13, 38, 9, 19, 31, 22, 36, 12, 15, 16, 11]. These works

*Equal contribution

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

tackle the exploration, estimation and planning problems simultaneously. On the other hand, recent works [16, 33, 6] consider an easier, but even more fundamental problem of obtaining a near-optimal policy with access to a simulator or *generative model* [20, 18, 2]. In particular, these works assume that the transition probabilities in the underlying CMDP are unknown, but the planner has access to a sampling oracle (the generative model) that returns a sample of the next state when given any state-action pair as input. This is the problem setting we consider and *aim to obtain matching upper and lower bounds on the sample complexity of planning in CMDPs with access to a generative model*.

Given a target error $\varepsilon > 0$, the approximate CMDP objective is to return a policy that achieves a cumulative reward within an ε additive error of the optimal policy in the CMDP. Previous work can be classified into two categories based on how it tackles the constraint – for the easier problem that we term *relaxed feasibility*, the policy returned by an algorithm is allowed to violate the constraint by at most ε. On the other hand, for the more difficult *strict feasibility* problem, the returned policy is required to strictly satisfy the constraint and achieve zero constraint violation. Except for the recent works of Wei et al. [33] and Bai et al. [6], most provably efficient approaches including those in the regret-minimization and PAC-RL settings consider the relaxed feasibility setting. For this problem, the best model-based algorithm requires $\tilde{O}\left(\frac{S^2A}{(1-\gamma)^3\varepsilon^2}\right)$ samples to return an ε-optimal policy in an infinite-horizon γ-discounted CMDP with S states and A actions [16], while the best model-free approach requires $\tilde{O}\left(\frac{SA}{(1-\gamma)^5\varepsilon^2}\right)$ samples for achieving the objective [12]. On the other hand, the best known upper bounds for a model-free algorithm in the strict feasibility setting are achieved by Bai et al. [6]. In particular, their algorithm requires $\tilde{O}\left(\frac{SA}{(1-\gamma)^4\varepsilon^2}\right)$ samples [6, Theorem 2] to output an ε-optimal policy. However, their analysis considers normalized (by a $(1-\gamma)$ factor) value functions [6, Eq. 1] that results in reward and constraint value functions to be in the $[0, 1]$ range (compared to the standard $[0, 1/1-\gamma] = O(H)$ range). This difference in the scale of the values prevents a direct comparison of their results to our sample complexity bounds. Subsequently, we show that when appropriately normalized, our sample complexity bounds are better by a factor of $1/1-\gamma$ in both the relaxed feasibility (Section 4) and strict feasibility settings (Section 5).

Importantly, there are no lower bounds characterizing the difficulty of either the relaxed or strict feasibility problems (except in degenerate cases where the constraint is always satisfied and the CMDP problem reduces to an unconstrained MDP). To get an indication of what the optimal bounds might be, it is instructive to compare these results to the unconstrained MDP setting. For unconstrained MDPs with access to a generative model, both model-based [2, 21] and model-free approaches [27] can return an ε-optimal policy within near-optimal $\tilde{\Theta}\left(\frac{SA}{(1-\gamma)^3\varepsilon^2}\right)$ sample-complexity [4]. Hence, compared to the sample-complexity for unconstrained MDPs, the best-known upper-bounds for CMDPs are worse for both the relaxed and strict feasibility settings. However, it is unclear whether solving CMDPs is inherently more difficult than unconstrained MDPs. We resolve these questions for both the relaxed and strict feasibility settings, and make the following contributions.

Generic model-based algorithm: In Section 3, we provide a generic model-based primal-dual algorithm (Algorithm 1) that can be used to achieve both the relaxed and strict feasibility objectives (with appropriate parameter settings). The proposed algorithm requires solving a sequence of unconstrained empirical MDPs using any black-box MDP planner.

Upper-bound on sample complexity under relaxed feasibility: In Section 4, we prove that with a specific set of parameters, Algorithm 1 uses no more than $\tilde{O}\left(\frac{SA}{(1-\gamma)^3\varepsilon^2}\right)$ samples to achieve the relaxed feasibility objective. This improves upon the bounds of HasanzadeZonuzy et al. [16] and matches the lower-bound in the easier unconstrained MDP setting, implying that our bounds are near-optimal. Our result indicates that under relaxed feasibility solving CMDPs is as easy as solving unconstrained MDPs. To the best of our knowledge, these are the first such bounds.

Upper-bound on sample-complexity under strict feasibility: In Section 5, we prove that with a specific set of parameters, Algorithm 1 uses no more than $\tilde{O}\left(\frac{SA}{(1-\gamma)^5\zeta^2\varepsilon^2}\right)$ to achieve the strict feasibility objective. Here $\zeta \in (0, 1/1-\gamma]$ is the problem-dependent *Slater constant* that characterizes the size of the feasible region and influences the difficulty of the problem. Unlike Bai et al. [6], our bounds do not depend on additional (potentially large) problem-dependent quantities.

Lower-bound on sample-complexity under strict feasibility: In Section 7, we prove a matching problem-dependent $\Omega\left(\frac{SA}{(1-\gamma)^5\,\zeta^2\,\varepsilon^2}\right)$ lower bound on the sample-complexity in the strict feasibility setting. Our results thus demonstrate that the proposed model-based algorithm is near minimax optimal. Furthermore, our bounds indicate that under strict feasibility (i) solving CMDPs is inherently more difficult than solving unconstrained MDPs, and (ii) the problem hardness (in terms of the sample-complexity) increases as ζ (and hence the size of the feasible region) decreases. To the best of our knowledge, these are first results characterizing the difficulty of solving CMDPs with access to a generative model and demonstrate a separation between the relaxed and strict feasibility settings.

Overview of techniques: For proving the upper bounds, we use a specific primal-dual algorithm that reduces the CMDP planning problem to solving multiple unconstrained MDPs. Specifically, by using a strong-duality argument, we show that we can obtain an optimal CMDP policy by averaging the optimal policies of a specific sequence of MDPs. For each MDP in this sequence, we use the model-based techniques from Agarwal et al. [2], Li et al. [21] to prove concentration results for data-dependent policies. This allows us to prove concentration for the optimal data-dependent policy in the CMDP, and subsequently bound the sample complexity for both the relaxed and strict feasibility problems. For the lower bound, we modify the MDP hard instances [5, 34] to handle a constraint reward. This makes the resulting gadgets significantly more complex than those required for MDPs, but we show that similar likelihood arguments can be used to prove the lower-bound.

2 Problem Formulation

We consider an infinite-horizon discounted constrained Markov decision process (CMDP) [3] denoted by M, and defined by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, r, c, b, \rho, \gamma \rangle$ where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the action set, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ is the transition probability function, $\rho \in \Delta_{\mathcal{S}}$ is the initial distribution of states and $\gamma \in [0, 1)$ is the discount factor. The primary reward to be maximized is denoted by $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$, whereas the constraint reward is denoted by $c : \mathcal{S} \times \mathcal{A} \to [0, 1]$[2]. If $\Delta_{\mathcal{A}}$ denotes the simplex over the action space, the expected discounted return or *reward value function* of a stationary, stochastic policy[3] $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$ is defined as $V_r^\pi(\rho) = \mathbb{E}_{s_0, a_0, \ldots}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$, where $s_0 \sim \rho, a_t \sim \pi(\cdot|s_t)$, and $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$. For each state-action pair (s, a) and policy π, the reward action-value function is defined as $Q_r^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and satisfies the relation: $V_r^\pi(s) = \langle \pi(\cdot|s), Q_r^\pi(s, \cdot) \rangle$, where $V_r^\pi(s)$ is the reward value function when the starting state is equal to s. Analogously, the *constraint value function* and constraint action-value function of policy π is denoted by $V_c^\pi(\rho)$ and Q_c^π respectively. The CMDP objective is to return a policy that maximizes $V_r^\pi(\rho)$, while ensuring that $V_c^\pi(\rho) \geq b$. Formally,

$$\max_\pi V_r^\pi(\rho) \quad \text{s.t.} \quad V_c^\pi(\rho) \geq b. \tag{1}$$

The optimal stochastic policy for the above CMDP is denoted by π^* and the corresponding reward value function is denoted by $V_r^*(\rho)$. We also define $\zeta := \max_\pi V_c^\pi(\rho) - b$ as the problem-dependent quantity referred to as the Slater constant [12, 6]. The Slater constant is a measure of the size of the feasible region and determines the difficulty of solving Eq. (1).

For simplicity of exposition, we assume that the rewards r and constraint rewards c are known, but the transition matrix $\mathcal{P}$ is unknown and needs to be estimated. We note that assuming the knowledge of the rewards does not affect the leading terms of the sample complexity since learning these is an easier problem compared to the transition matrix [5, 27]. We assume access to a *generative model* or simulator that allows the agent to obtain samples from the $\mathcal{P}(\cdot|s, a)$ distribution for any (s, a). Assuming access to such a generative model, our aim is to characterize the sample complexity required to return a near-optimal policy $\hat{\pi}$ in M. Given a target error $\varepsilon > 0$, we can characterize the performance of policy $\hat{\pi}$ in two ways:

Relaxed feasibility: We require $\hat{\pi}$ to achieve an approximately optimal reward value, while allowing it to have a small constraint violation in M[4]. Formally, we require $\hat{\pi}$ s.t.

$$V_r^{\hat{\pi}}(\rho) \geq V_r^*(\rho) - \varepsilon, \text{ and } V_c^{\hat{\pi}}(\rho) \geq b - \varepsilon. \tag{2}$$

[2]These ranges for r and c are chosen for simplicity. Our results can be easily extended to handle other ranges.

[3]The performance of an optimal policy in a CMDP can always be achieved by a stationary, stochastic policy [3]. On the other hand, for an MDP, it suffices to only consider stationary, deterministic policies [25].

[4]In general, the desired gap in the reward value can be different from the level of constraint violation.

Algorithm 1: Model-based algorithm for CMDPs with generative model

Input: $\mathcal{S}$ (state space), $\mathcal{A}$ (action space), r (rewards), c (constraint rewards), ζ (Slater constant), N (number of samples), b' (constraint RHS), ω (perturbation magnitude), U (projection upper bound), ε_1 (epsilon-net resolution), T (number of iterations), $\lambda_0 = 0$ (initialization).

For each state-action (s, a) pair, collect N samples from $\mathcal{P}(.|s, a)$ and form $\hat{\mathcal{P}}$.

Perturb the rewards to form vector $r_p(s, a) := r(s, a) + \xi(s, a)$ where $\xi(s, a) \sim \mathcal{U}[0, \omega]$.

Form the empirical CMDP $\hat{M} = \langle \mathcal{S}, \mathcal{A}, \hat{\mathcal{P}}, r_p, c, b', \rho, \gamma \rangle$.

Form the epsilon-net $\Lambda = \{0, \varepsilon_1, 2\varepsilon_1, \ldots, U\}$.

for $t \leftarrow 0$ **to** $T - 1$ **do**

  Update the policy by solving an unconstrained MDP: $\hat{\pi}_t = \arg\max \hat{V}^{\pi}_{r_p + \lambda_t c}$.

  Update the dual-variables: $\lambda_{t+1} = \mathcal{R}_{\Lambda}\left[\mathbb{P}_{[0,U]}\left[\lambda_t - \eta\left(\hat{V}^{\hat{\pi}_t}_c(\rho) - b'\right)\right]\right]$.

end

Output: Mixture policy $\bar{\pi}_T = \frac{1}{T}\sum_{t=0}^{T-1} \hat{\pi}_t$.

Strict feasibility We require $\hat{\pi}$ to achieve an approximately optimal reward value, while simultaneously demanding zero constraint violation in M. Formally, we require $\hat{\pi}$ s.t.

$$V^{\hat{\pi}}_r(\rho) \geq V^*_r(\rho) - \varepsilon, \text{ and } V^{\hat{\pi}}_c(\rho) \geq b \tag{3}$$

Next, we describe a general model-based algorithm to handle both these cases, and subsequently instantiate the algorithm for the relaxed feasibility (Section 4) and strict feasibility (Section 5) settings.

3 Methodology

We will use a model-based approach [2, 21, 16] for achieving the objectives in Eq. (2) and Eq. (3). In particular, for each (s, a) pair, we collect N independent samples from $\mathcal{P}(\cdot|s, a)$ and form an empirical transition matrix $\hat{\mathcal{P}}$ such that $\hat{\mathcal{P}}(s'|s, a) = \frac{N(s'|s,a)}{N}$, where $N(s'|s, a)$ is the number of samples that have transitions from (s, a) to s'. These estimated transition probabilities are used to form an empirical CMDP. Due to a technical requirement, (which we will clarify in the Section 6), we require adding a small random perturbation to the rewards in the empirical CMDP[5]. In particular, for each $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we 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 variables. Finally, compared to Eq. (1), we will require solving the empirical CMDP with a constraint right-hand side equal to b'. Note that setting $b' < b$ corresponds to loosening the constraint, while $b' > b$ corresponds to tightening the constraint. This completes the specification of the empirical CMDP $\hat{M}$ that is defined by the tuple $\langle \mathcal{S}, \mathcal{A}, \hat{\mathcal{P}}, r_p, c, b', \rho, \gamma \rangle$. For $\hat{M}$, the corresponding reward value function (and constraint value function) for policy π is denoted as $\hat{V}^{\pi}_{r_p}(\rho)$ (and $\hat{V}^{\pi}_c(\rho)$ respectively). In order to fully instantiate $\hat{M}$, we require setting the values of ω (the magnitude of the perturbation) and b' (the constraint right-hand side). This depends on the specific setting (relaxed vs strict feasibility) and we do this in Sections 4 and 5 respectively. We compute the optimal policy for the empirical CMDP $\hat{M}$ as follows:

$$\hat{\pi}^* \in \arg\max \hat{V}^{\pi}_{r_p}(\rho) \text{ s.t. } \hat{V}^{\pi}_c(\rho) \geq b' \tag{4}$$

In contrast to Agarwal et al. [2], Li et al. [21] that consider model-based approaches for unconstrained MDPs and can solve the resulting empirical MDP using any black-box approach, we will require solving Eq. (4) using a specific primal-dual approach that we outline next. Using this algorithm enables us to prove optimal sample complexity bounds under both relaxed and strict feasibility.

First, observe that Eq. (4) can be written as an equivalent saddle-point problem – $\max_{\pi} \min_{\lambda \geq 0} \left[\hat{V}^{\pi}_{r_p}(\rho) + \lambda\left(\hat{V}^{\pi}_c(\rho) - b'\right)\right]$, where $\lambda \in \mathbb{R}$ corresponds to the Lagrange multiplier for the constraint. The solution to this saddle-point problem is $(\hat{\pi}^*, \lambda^*)$ where $\hat{\pi}^*$ is the optimal empirical policy and λ^* is the optimal Lagrange multiplier. We solve the above saddle-point problem

[5] Similar to MDPs [21], we can instead perturb the Q function while planning in the empirical CMDP.

iteratively, by alternatively updating the policy (primal variable) and the Lagrange multiplier (dual variable). If T is the total number of iterations of the primal-dual algorithm, we define $\hat{\pi}_t$ and λ_t to be the primal and dual iterates for $t \in [T] := \{1, \ldots, T\}$. The primal update at iteration t is given as:

$$\hat{\pi}_t = \arg\max \left[\hat{V}^{\pi}_{r_p}(\rho) + \lambda_t \hat{V}^{\pi}_{c}(\rho)\right] = \arg\max \hat{V}^{\pi}_{r_p + \lambda_t c}. \tag{5}$$

Hence, iteration t of the algorithm requires solving an unconstrained MDP with a reward equal to $r_p + \lambda_t c$. This can be done using any black-box MDP solver such as policy iteration. The algorithm updates the Lagrange multipliers using a gradient descent step and requires projecting and rounding the resulting dual variables. In particular, the dual variables are first projected onto the $[0, U]$ interval, where U is chosen to be an upper-bound on $|\lambda^*|$. After the projection, the resulting iterates are rounded to the closest element in the set $\Lambda = \{0, \varepsilon_1, 2\varepsilon_1, \ldots, U\}$, a one-dimensional epsilon-net (with resolution ε_1) over the dual variables. In Section 6, we will see that constructing such an ε_1-net will enable us to prove concentration results for all $\lambda \in \Lambda$. The dual update at iteration t is given as:

$$\lambda_{t+1} = \mathcal{R}_{\Lambda}\left[\mathbb{P}_{[0,U]}\left[\lambda_t - \eta\left(\hat{V}^{\hat{\pi}_t}_{c}(\rho) - b'\right)\right]\right], \tag{6}$$

where $\mathbb{P}_{[0,U]}[\lambda] = \arg\min_{p \in [0,U]} |\lambda - p|$ projects λ onto the $[0, U]$ interval and $\mathcal{R}_{\Lambda}[\lambda] = \arg\min_{p \in \Lambda} |\lambda - p|$ rounds λ to the closest element in Λ. Since Λ is an epsilon-net, for all $\lambda \in [0, U]$, $|\lambda - \mathcal{R}_{\Lambda}[\lambda]| \leq \varepsilon_1$. Finally, η in Eq. (6) corresponds to the step-size for the gradient descent update. The above primal-dual updates are similar to the dual-descent algorithm proposed in Paternain et al. [24]. The pseudo-code summarizing the entire model-based algorithm is given in Algorithm 1. We note that although Algorithm 1 requires the knowledge of ζ, this is not essential and we can instead use an estimate of ζ. In Appendix F, we show that we can estimate ζ to within a factor of 2 using $\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3 \zeta^2}\right)$ additional queries. Next, we show that the primal-dual updates in Algorithm 1 can be used to solve the empirical CMDP $\hat{M}$. Specifically, we prove the following theorem (proof in Appendix A) that bounds the average optimality gap (in the reward value function) and constraint violation for the mixture policy returned by Algorithm 1.

Theorem 1 (Guarantees for the primal-dual algorithm)**.** For a target error $\varepsilon_{\text{opt}} > 0$ and the primal-dual updates in Eq. (5)-Eq. (6) with $U > |\lambda^*|$, $T = \frac{4U^2}{\varepsilon^2_{\text{opt}}(1-\gamma)^2}\left[1 + \frac{1}{(U-\lambda^*)^2}\right]$, $\eta = \frac{U(1-\gamma)}{\sqrt{T}}$ and $\varepsilon_1 = \frac{\varepsilon^2_{\text{opt}}(1-\gamma)^2(U-\lambda^*)}{6U}$, the mixture policy $\bar{\pi}_T := \frac{1}{T}\sum_{t=0}^{T-1} \hat{\pi}_t$ satisfies,

$$\hat{V}^{\bar{\pi}_T}_{r_p}(\rho) \geq \hat{V}^{\hat{\pi}^*}_{r_p}(\rho) - \varepsilon_{\text{opt}} \quad ; \quad \hat{V}^{\bar{\pi}_T}_{c}(\rho) \geq b' - \varepsilon_{\text{opt}}.$$

Hence, with $T = O(1/\varepsilon^2_{\text{opt}})$ and $\varepsilon_1 = O(\varepsilon^2_{\text{opt}})$, the algorithm outputs a policy $\bar{\pi}_T$ that achieves a reward ε_{opt} close to that of the optimal empirical policy $\hat{\pi}^*$, while violating the constraint by at most ε_{opt}. Hence, with sufficient number of iterations T and by choosing a sufficiently small resolution ε_1 for the epsilon-net, we can use the above primal-dual algorithm to approximately solve the problem in Eq. (4). In order to completely instantiate the primal-dual algorithm, we require setting $U > |\lambda^*|$. We will subsequently do this for the the relaxed and strict feasibility settings in Sections 4 and 5 respectively. We note that in contrast to Paternain et al. [24, Theorem 3] that bounds the Lagrangian, Theorem 1 provides explicit bounds on both the reward suboptimality and constraint violation.

We conclude this section by making some observations about the primal-dual algorithm – while the subsequent bounds for both settings heavily depend on using the "best-response" primal update in Eq. (5), the algorithm does not require using the specific form of the dual updates in Eq. (6). Indeed, when used in conjunction with the projection and rounding operations in Eq. (6), we can use any method to update the dual variables (not necessarily gradient descent) provided that it results in an $O\left(T^a + \varepsilon_1 T^b\right)$ (for $a < 1$) bound on the dual regret (see the proof of Theorem 1 for the definition). Next, we specify the values of N, b', ω, ε_1, T, U in Algorithm 1 to achieve the objective in Eq. (2).

4 Upper-bound under Relaxed Feasibility

In order to achieve the objective in Eq. (2) for a target error $\varepsilon > 0$, we require setting $N = \tilde{O}\left(\frac{\log(1/\delta)}{(1-\gamma)^3 \varepsilon^2}\right)$, $b' = b - \frac{3\varepsilon}{8}$ and $\omega = \frac{\varepsilon(1-\gamma)}{8}$. This completely specifies the empirical CMDP $\hat{M}$

and the problem in Eq. (4). In order to specify the primal-dual algorithm, we set $U = O\left(1/\varepsilon\,(1-\gamma)\right)$, $\varepsilon_1 = O\left(\varepsilon^2(1-\gamma)^2\right)$ and $T = O\left(1/(1-\gamma)^4\varepsilon^4\right)$. With these choices, we prove the following theorem in Appendix B and provide a proof sketch below.

Theorem 2. For a fixed $\varepsilon \in (0, 1/1-\gamma]$ and $\delta \in (0,1)$, Algorithm 1 with $N = \tilde{O}\left(\frac{\log(1/\delta)}{(1-\gamma)^3\varepsilon^2}\right)$ samples, $b' = b - \frac{3\varepsilon}{8}$, $\omega = \frac{\varepsilon(1-\gamma)}{8}$, $U = O\left(1/\varepsilon\,(1-\gamma)\right)$, $\varepsilon_1 = O\left(\varepsilon^2(1-\gamma)^2\right)$ and $T = O\left(1/(1-\gamma)^4\varepsilon^4\right)$, returns policy $\bar{\pi}_T$ that satisfies the objective in Eq. (2) with probability at least $1-4\delta$.

Proof Sketch: 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 11 (proved in Appendix B), 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 1, i.e., if

$$\left|V_c^{\bar{\pi}_T}(\rho) - \hat{V}_c^{\bar{\pi}_T}(\rho)\right| \leq \frac{\varepsilon - \varepsilon_{\text{opt}}}{2} \quad ; \quad \left|V_c^{\pi^*}(\rho) - \hat{V}_c^{\pi^*}(\rho)\right| \leq \frac{\varepsilon - \varepsilon_{\text{opt}}}{2}, \tag{7}$$

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

$$V_r^{\pi^*}(\rho) - V_r^{\bar{\pi}_T}(\rho) \leq \frac{2\omega}{1-\gamma} + \varepsilon_{\text{opt}} + \left|V_{r_p}^{\pi^*}(\rho) - \hat{V}_{r_p}^{\pi^*}(\rho)\right| + \left|\hat{V}_{r_p}^{\bar{\pi}_T}(\rho) - V_{r_p}^{\bar{\pi}_T}(\rho)\right|. \tag{8}$$

In order to instantiate the primal-dual algorithm, we require a concentration result for policy π_c^* that maximizes the the constraint value function, i.e. if $\pi_c^* := \arg\max V_c^{\pi}(\rho)$, then we require $\left|V_c^{\pi_c^*}(\rho) - \hat{V}_c^{\pi_c^*}(\rho)\right| \leq \varepsilon + \varepsilon_{\text{opt}}$. In Case 1 of Lemma 9 (proved in Appendix A), 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 1 holds. To guarantee that the primal-dual algorithm outputs an $\frac{\varepsilon}{4}$-approximate policy, we use Theorem 1 to set $T = O\left(\frac{1}{(1-\gamma)^4\varepsilon^4}\right)$ iterations and $\varepsilon_1 = O\left(\varepsilon^2(1-\gamma)^2\right)$. Eq. (8) can then be simplified as,

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

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

$$\begin{aligned}
&\left|V_c^{\pi_c^*}(\rho) - \hat{V}_c^{\pi_c^*}(\rho)\right| \leq \frac{5\varepsilon}{4}\,;\, \left|V_c^{\bar{\pi}_T}(\rho) - \hat{V}_c^{\bar{\pi}_T}(\rho)\right| \leq \frac{3\varepsilon}{8}\,;\, \left|V_c^{\pi^*}(\rho) - \hat{V}_c^{\pi^*}(\rho)\right| \leq \frac{3\varepsilon}{8} \\
&\left|V_{r_p}^{\pi^*}(\rho) - \hat{V}_{r_p}^{\pi^*}(\rho)\right| \leq \frac{\varepsilon}{4}\,;\, \left|\hat{V}_{r_p}^{\bar{\pi}_T}(\rho) - V_{r_p}^{\bar{\pi}_T}(\rho)\right| \leq \frac{\varepsilon}{4}.
\end{aligned} \tag{9}$$

We control such concentration terms for both the constraint and reward value functions in Section 6, and bound the terms in Eq. (9). In particular, we prove that for a fixed $\varepsilon \in (0, 1/1-\gamma]$, using $N \geq \tilde{O}\left(\frac{\log(1/\delta)}{(1-\gamma)^3\,\varepsilon^2}\right)$ samples enssures that the statements in Eq. (9) hold with probability $1-4\delta$. This guarantees that $V_r^{\pi^*}(\rho) - V_r^{\bar{\pi}_T}(\rho) \leq \varepsilon$ and $V_c^{\bar{\pi}_T}(\rho) \geq b - \varepsilon$. □

Hence, the total sample-complexity of achieving the objective in Eq. (2) is $\tilde{O}\left(\frac{SA\log(1/\delta)}{(1-\gamma)^3\varepsilon^2}\right)$. This result improves over the $\tilde{O}\left(\frac{S^2A\log(1/\delta)}{(1-\gamma)^3\varepsilon^2}\right)$ result in HasanzadeZonuzy et al. [16]. Furthermore, our result matches the lower-bound in the easier unconstrained setting [4], implying that our bounds are near-optimal. We conclude that under relaxed feasibility and with access to a generative model, solving constrained MDPs is as easy as solving MDPs. Algorithmically, we do not require constructing an optimistic CMDP like in HasanzadeZonuzy et al. [16]. Instead, we solve the empirical CMDP in Eq. (4) using specific primal-dual updates Eqs. (5) and (6). Note that if the rewards and constraint rewards (corresponding to K constraints) are unknown and need to be estimated, a union bound guarantees that the sample complexity will only increase by a multiplicative $\log(K+1)$ factor [16].

In this setting, when using the *normalized value functions* Bai et al. [6, Corollary 1] prove an $O\left(\frac{H^2SA}{\varepsilon^2\zeta^2}\right)$ bound on the sample complexity. When translated to the standard $O\left(1/1-\gamma\right)$ range, this implies an $O\left(\frac{H^4SA}{\varepsilon^2\zeta^2}\right)$ bound [6, Footnote 6]. In comparison, our result in Theorem 2 has a better dependence on H and does not depend on the Slater constant ζ. Importantly, unlike [6], our result implies that in the relaxed feasibility setting, solving CMDPs is as easy as solving MDPs. In the next section, we instantiate Algorithm 1 for the strict feasibility setting.

5 Upper-bound under Strict Feasibility

Unlike Section 4, since the strict feasibility setting does not allow any constraint violations, it necessitates using a stricter constraint for the empirical CMDP to account for the estimation error in the transition probabilities. Hence, we require setting $b' > b$. Specifically, in order to achieve the objective in Eq. (3) for a target error $\varepsilon > 0$, we require setting $N = \tilde{O}\left(\frac{\log(1/\delta)}{(1-\gamma)^5\zeta^2\varepsilon^2}\right)$,[6] $b' = b + \frac{\varepsilon(1-\gamma)\zeta}{20}$ and $\omega = \frac{\varepsilon(1-\gamma)}{10}$. This completely specifies the empirical CMDP $\hat{M}$ and the problem in Eq. (4). To specify the primal-dual algorithm, we set $U = \frac{4(1+\omega)}{\zeta(1-\gamma)}$, $\varepsilon_1 = O\left(\varepsilon^2(1-\gamma)^4\zeta^2\right)$ and $T = O\left(1/(1-\gamma)^6\zeta^4\varepsilon^2\right)$. With these choices, we prove the following theorem in Appendix C, and provide a proof sketch below.

Theorem 3. For a fixed $\varepsilon \in (0, 1/1-\gamma]$ and $\delta \in (0,1)$, Algorithm 1, with $N = \tilde{O}\left(\frac{\log(1/\delta)}{(1-\gamma)^5\varepsilon^2\zeta^2}\right)$ samples, $b' = b + \frac{\varepsilon(1-\gamma)\zeta}{20}$, $\omega = \frac{\varepsilon(1-\gamma)}{10}$, $U = \frac{4(1+\omega)}{\zeta(1-\gamma)}$, $\varepsilon_1 = O\left(\varepsilon^2(1-\gamma)^4\zeta^2\right)$ and $T = O\left(1/(1-\gamma)^6\zeta^4\varepsilon^2\right)$ returns policy $\bar{\pi}_T$ that satisfies the objective in Eq. (3), with probability at least $1 - 4\delta$.

Proof Sketch: We prove the result for a general $b' = b + \Delta$ for $\Delta > 0$ and primal-dual error $\varepsilon_{\text{opt}} < \Delta$, and subsequently specify Δ (and hence b') and ε_{opt}. In Lemma 12 (proved in Appendix C), we prove 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 1 i.e. if

$$\left|V_c^{\bar{\pi}_T}(\rho) - \hat{V}_c^{\bar{\pi}_T}(\rho)\right| \leq \Delta - \varepsilon_{\text{opt}} \quad ; \quad \left|V_c^{\pi^*}(\rho) - \hat{V}_c^{\pi^*}(\rho)\right| \leq \Delta \tag{10}$$

then (i) policy $\bar{\pi}_T$ satisfies the constraint in M i.e. $V_c^{\bar{\pi}_T}(\rho) \geq b$, and (ii) its suboptimality in M (compared to π^*) can be decomposed as:

$$V_r^{\pi^*}(\rho) - V_r^{\bar{\pi}_T}(\rho) \leq \frac{2\omega}{1-\gamma} + \varepsilon_{\text{opt}} + 2\Delta|\lambda^*| + \left|V_{r_p}^{\pi^*}(\rho) - \hat{V}_{r_p}^{\pi^*}(\rho)\right| + \left|\hat{V}_{r_p}^{\bar{\pi}_T}(\rho) - V_{r_p}^{\bar{\pi}_T}(\rho)\right| \tag{11}$$

In order to upper-bound $|\lambda^*|$, we require a concentration result for policy $\pi_c^* := \arg\max V_c^{\pi}(\rho)$ that maximizes the the constraint value function. In particular, we require $\Delta \in \left(0, \frac{\zeta}{2}\right)$ and $\left|V_c^{\pi_c^*}(\rho) - \hat{V}_c^{\pi_c^*}(\rho)\right| \leq \frac{\zeta}{2} - \Delta$. In Case 2 of Lemma 9 (proved in Appendix A), we show that if this concentration result holds, then we can upper-bound the optimal dual variable $|\lambda^*|$ by $\frac{2(1+\omega)}{\zeta(1-\gamma)}$. Using the above bounds to simplify Eq. (11),

$$V_r^{\pi^*}(\rho) - V_r^{\bar{\pi}_T}(\rho) \leq \frac{2\omega}{1-\gamma} + \varepsilon_{\text{opt}} + \frac{4\Delta(1+\omega)}{\zeta(1-\gamma)} + \left|V_{r_p}^{\pi^*}(\rho) - \hat{V}_{r_p}^{\pi^*}(\rho)\right| + \left|\hat{V}_{r_p}^{\bar{\pi}_T}(\rho) - V_{r_p}^{\bar{\pi}_T}(\rho)\right|.$$

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 $\Delta = \frac{\varepsilon(1-\gamma)\zeta}{40} < \frac{\zeta}{2}$, $\varepsilon_{\text{opt}} = \frac{\Delta}{5} = \frac{\varepsilon(1-\gamma)\zeta}{200} < \frac{\varepsilon}{5}$, and $\omega = \frac{\varepsilon(1-\gamma)}{10} < 1$. We set $U = \frac{8}{\zeta(1-\gamma)}$ for the primal-dual algorithm, ensuring that the $U > |\lambda^*|$ condition required by Theorem 1 holds. In order to guarantee that the primal-dual algorithm outputs an $\frac{\varepsilon(1-\gamma)\zeta}{200}$-approximate policy, we use Theorem 1 to set $T = O\left(\frac{1}{(1-\gamma)^6\zeta^4\varepsilon^2}\right)$

[6] Again, we do not need to know ζ and it can be replaced by the estimator constructed in Section F.

iterations and $\varepsilon_1 = O\left(\varepsilon^2(1-\gamma)^4\zeta^2\right)$. With these values, we can further simplify Eq. (11),

$$V_r^{\pi^*}(\rho) - V_r^{\bar{\pi}_T}(\rho) \leq \frac{3\varepsilon}{5} + \left|V_{r_p}^{\pi^*}(\rho) - \hat{V}_{r_p}^{\pi^*}(\rho)\right| + \left|\hat{V}_{r_p}^{\bar{\pi}_T}(\rho) - V_{r_p}^{\bar{\pi}_T}(\rho)\right|.$$

Putting everything together, in order to guarantee an ε-reward suboptimality for $\bar{\pi}_T$, we require the following concentration results to hold for $\Delta = \frac{\varepsilon(1-\gamma)\zeta}{40}$,

$$\left|V_c^{\bar{\pi}_T}(\rho) - \hat{V}_c^{\bar{\pi}_T}(\rho)\right| \leq \frac{4\Delta}{5}\,;\, \left|V_c^{\pi^*}(\rho) - \hat{V}_c^{\pi^*}(\rho)\right| \leq \Delta\,;\, \left|V_c^{\pi_c^*}(\rho) - \hat{V}_c^{\pi_c^*}(\rho)\right| \leq \frac{19\Delta}{5}$$
$$\left|V_{r_p}^{\pi^*}(\rho) - \hat{V}_{r_p}^{\pi^*}(\rho)\right| \leq \frac{\varepsilon}{5}\,;\, \left|\hat{V}_{r_p}^{\bar{\pi}_T}(\rho) - V_{r_p}^{\bar{\pi}_T}(\rho)\right| \leq \frac{\varepsilon}{5}. \tag{12}$$

We control such concentration terms for both the constraint and reward value functions in Section 6, and bound the terms in Eq. (12). In particular, we prove that for a fixed $\varepsilon \in (0, 1/1-\gamma]$, using $N \geq \tilde{O}\left(\frac{\log(1/\delta)}{(1-\gamma)^5\,\zeta^2\,\varepsilon^2}\right)$ ensures that the statements in Eq. (12) hold with probability $1-4\delta$. This guarantees that $V_r^{\pi^*}(\rho) - V_r^{\bar{\pi}_T}(\rho) \leq \varepsilon$ and $V_c^{\bar{\pi}_T}(\rho) \geq b$. □

Hence, the total sample-complexity of achieving the objective in Eq. (3) is $\tilde{O}\left(\frac{SA\,\log(1/\delta)}{(1-\gamma)^5\,\zeta^2\varepsilon^2}\right)$. Similar to Section 4, in the strict feasibility setting, with the *normalized value functions*, Bai et al. [6] prove an $O\left(\frac{H^2SA}{\varepsilon^2\zeta^2}\right)$ bound on the sample complexity. When translated to the standard $O\left(1/1-\gamma\right)$ scale of the value functions, this implies an $\Omega\left(\frac{H^6SA}{\varepsilon^2\zeta^2}\right)$ bound (see Appendix G for a detailed explanation). In comparison, our result in Theorem 3 has a better dependence on H.

In Section 7, we prove a matching lower bound showing that Algorithm 1 is minimax optimal in the strict feasibility setting. In the next section, we give more details for the bounding the concentration terms in Theorem 2 and Theorem 3.

6 Bounding the concentration terms

We have seen that proving Theorem 2 and Theorem 3 require bounding the concentration terms in Eq. (9) and Eq. (12) respectively. In this section, we detail the techniques to achieve these bounds.

Our approach requires reasoning about a general unconstrained MDP $M_\alpha = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \gamma, \alpha)$ with the same state-action space, transition probabilities and discount factor as the CMDP in Eq. (1) but with rewards equal to $\alpha \in [0, \alpha_{\max}]$. Analogously, we define the empirical MDP $\hat{M}_\alpha = (\mathcal{S}, \mathcal{A}, \hat{\mathcal{P}}, \gamma, \alpha)$ where the empirical transition matrix $\hat{\mathcal{P}}$ is the same as that of the empirical CMDP in Eq. (4). Similarly, we define MDP (and its empirical counterpart) $M_\beta = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \gamma, \beta)$ (and $\hat{M}_\beta$) where the rewards $\beta \in [0, \beta_{\max}]$. Note that the rewards α and β are independent of the sampling of the transition matrix. The corresponding value functions for policy π in M_α and $\hat{M}_\alpha$ (and M_β and $\hat{M}_\beta$) are denoted as V_α^π and $\hat{V}_\alpha^\pi$ (and V_β^π and $\hat{V}_\beta^\pi$) respectively, with the optimal value functions denoted as V_α^* and $\hat{V}_\alpha^*$ (and V_β^* and $\hat{V}_\beta^*$) respectively. The action-value function in M_α for policy π and state-action pair (s, a) is denoted as $Q_\alpha^\pi(s, a)$ and analogously for $\hat{M}_\alpha$. For the subsequent technical results, we require that $\hat{M}_\alpha$ satisfy the following gap condition [21]:

Definition 4 (ι-Gap Condition)**.** *MDP $\hat{M}_\alpha$ satisfies the ι-gap condition if $\forall s$, $\hat{V}_\alpha^*(s) - \max_{a':a\neq\hat{\pi}_\alpha^*(s)} \hat{Q}_\alpha^*(s, a') \geq \iota$, where $\hat{\pi}_\alpha^* := \arg\max \hat{V}_\alpha^\pi$ and $\hat{\pi}_\alpha^*(s) = \arg\max_a \hat{Q}_\alpha^*(s, a)$ is the optimal action in state s.*

Intuitively, the gap condition states that there is a unique optimal action at each state and there is a gap between the performance of best action and the second best action. With this gap condition, we use techniques in Li et al. [21] to prove the following lemma in Appendix D.

Lemma 5. *Define $\hat{\pi}_\alpha^* := \arg\max_\pi \hat{V}_\alpha^\pi$. If (i) $\mathcal{E}$ is the event that the ι-gap condition in Definition 4 holds for $\hat{M}_\alpha$ 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 $N \geq \frac{4\,C(\delta)}{1-\gamma}$, then with probability at least $\Pr[\mathcal{E}] - \delta/10$,*

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

Hence, for policy $\hat{\pi}^*_\alpha$, we can obtain a concentration result in another MDP $\hat{M}_\beta$ with an independent reward function β and the same empirical transition matrix $\hat{\mathcal{P}}$.

We wish to use the above lemma for the unconstrained MDP formed at every iteration of the primal update in Eq. (5). In particular, for a given λ_t, we will use Lemma 5 with $\alpha = r_p + \lambda_t c$ and $\beta = r_p$. Doing so will immediately give us a bound on $\left\| V^{\hat{\pi}_t}_{r_p} - \hat{V}^{\hat{\pi}_t}_{r_p} \right\|_\infty$ and hence $\left| V^{\hat{\pi}_t}_{r_p}(\rho) - \hat{V}^{\hat{\pi}_t}_{r_p}(\rho) \right|$. In order to use Lemma 5, we require the unconstrained MDP $\hat{M}_{r_p+\lambda c}$ to satisfy the gap condition in Definition 4 for any $\lambda \in \Lambda$. This is achieved by the perturbation of the rewards in Line 3 of Algorithm 1. Specifically, using Li et al. [21, Lemma 6] with a union-bound over Λ, we prove (in Lemma 16 in Appendix D) that with probability $1 - \delta/10$, $\hat{M}_{r_p+\lambda c}$ satisfies the gap condition in Definition 4 with $\iota = \frac{\omega\, \delta\, (1-\gamma)}{30\, |\Lambda||S||A|^2}$ for every $\lambda \in \Lambda$. This allows us to use Lemma 5 with $\alpha = r_p + \lambda_t c$ for all $t \in [T]$, and $\beta = r_p$ and $\beta = c$. In the following theorem, we obtain a concentration result for each $\hat{\pi}_t$ and hence for the mixture policy $\bar{\pi}_T$.

Theorem 6. 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 $\iota = \frac{\omega\, \delta\, (1-\gamma)\, \varepsilon_1}{30\, U|S||A|^2}$, if $N \geq \frac{4\, C(\delta)}{1-\gamma}$, then for $\bar{\pi}_T$ output by Algorithm 1, with probability at least $1 - \delta/5$,

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

Eqs. (9) and (12) also require proving concentration bounds for fixed (that do not depend on the data) policies π^* and π^*_c. This can be done by directly using Li et al. [21, Lemma 1]. Specifically, we prove the following the lemma in Appendix D.

Lemma 7. *For $\delta \in (0,1)$, $\omega \leq 1$ and $C'(\delta) = 72 \log\left(\frac{4|S| \log(e/1-\gamma)}{\delta}\right)$, if $N \geq \frac{4\, C'(\delta)}{1-\gamma}$ and $B(\delta, N) := \sqrt{\frac{C'(\delta)}{(1-\gamma)^3 N}}$, then with probability at least $1 - 3\delta$,*

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

Using Theorem 6 and Lemma 7, we can bound each term in Eq. (9) and Eq. (12), completing the proof of Theorem 2 and Theorem 3 respectively. In the next section, we prove a lower-bound on the sample-complexity in the strict feasibility setting.

7 Lower-bound under strict feasibility

For a target error of ε, our lower bound construction demonstrates that it is important to estimate the constraint value function to a smaller error equal to $\varepsilon' := \varepsilon(1-\gamma)\zeta$. Intuitively, this is because a small estimation error in the constraint value can incorrectly render the optimal policy infeasible and result in a large $\frac{\text{Range(value function)}}{\zeta} = \frac{1}{(1-\gamma)\zeta}$ suboptimality in the reward value. In Appendix E.1, we detail this intuition in a simplified bandit setting and present the formal CMDP lower-bound below.

We define an algorithm to be (ε, δ)-sound if it outputs a policy $\hat{\pi}$ such that with probability $1 - \delta$, $V^*_r(\rho) - V^{\hat{\pi}}_r(\rho) \leq \varepsilon$ and $V^{\hat{\pi}}_c(\rho) \geq b$ i.e. the algorithm achieves the strict feasibility objective in Eq. (3). We prove a lower bound on the number of samples required by any (ε, δ)-sound algorithm on the CMDP instance in Fig. 1. For this instance, with a specific setting of the rewards and probabilities $p_0 < \bar{p} < p_1$, we prove that any (ε, δ)-sound algorithm requires at least $\Omega\left(\frac{\ln(|\mathcal{S}||\mathcal{A}|/4\delta)}{\varepsilon^2 \zeta^2 (1-\gamma)^5}\right)$ samples to distinguish between M_0 and $M_{i,a}$. In particular, we prove the following theorem in Appendix E.2.

Theorem 8. There exists constants $\gamma_0 \in (1 - 1/\log(|\mathcal{S}|), 1)$, $0 \leq \varepsilon_0 \leq \frac{1}{(1-\gamma)} \min\left\{1, \frac{\gamma}{(1-\gamma)\zeta}\right\}$, $\delta_0 \in (0,1)$, such that, for any $\gamma \in (\gamma_0, 1), \varepsilon \in (0, \varepsilon_0), \delta \in (0, \delta_0)$, any (ε, δ)-sound algorithm requires $\Omega\left(\frac{SA \ln(1/4\delta)}{\varepsilon^2 \zeta^2 (1-\gamma)^5}\right)$ samples from the generative model in the worst case.

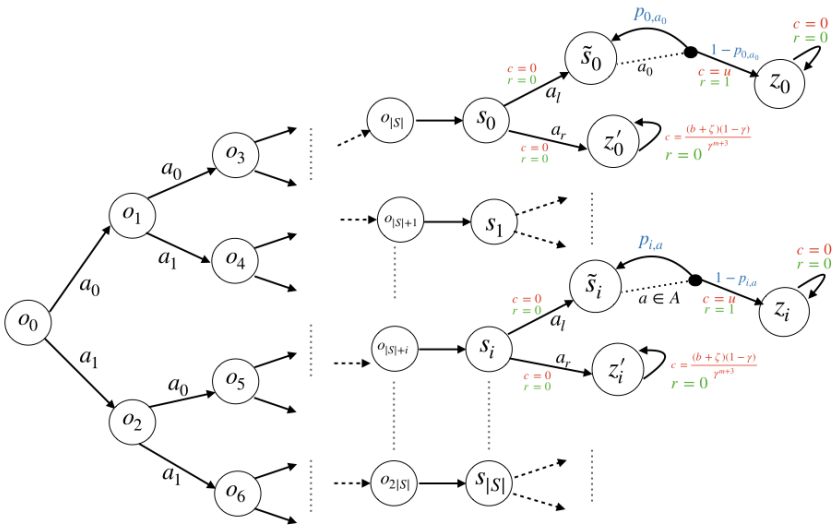

Figure 1: The instance consists of CMDPs with $S = 2^m - 1$ (for some integer $m > 0$) states and A actions. We consider $SA + 1$ CMDPs – M_0 and $M_{i,a}$ ($i \in \{1, \dots S\}$, $a \in \{1, \dots, A\}$) that share the same structure shown in the figure. For each CMDP, o_0 is the fixed starting state and there is a deterministic path of length $m + 1$ from o_0 to each of the $S + 1$ states – s_i (for $i \in \{0, 1, \dots, S\}$). Except for states $\tilde{s}_i$, the transitions in all other states are deterministic. For $i \neq 0$, for action $a \in \mathcal{A}$ in state $\tilde{s}_i$, the probability of staying in $\tilde{s}_i$ is $p_{i,a}$, while that of transitioning to state z_i is $1 - p_{i,a}$. There is only one action a_0 in $\tilde{s}_0$ and the probability of staying in $\tilde{s}_0$ is p_{0,a_0}, while that of transitioning to state z_0 is $1 - p_{0,a_0}$. The CMDPs M_0 and $M_{i,a}$ only differ in the values of $p_{i,a}$. The rewards r and constraint rewards c are the same in all CMDPs and are denoted in green and red respectively.

The above lower bound matches the upper bound in Theorem 3 and proves that Algorithm 1 is near minimax optimal in the strict feasibility setting. It also demonstrates that solving CMDPs under strict feasibility is inherently more difficult than solving unconstrained MDPs or CMDPs in the relaxed feasibility setting. Finally, we can conclude that the problem becomes more difficult (requires more samples) as the Slater constant ζ decreases and the feasible region shrinks.

8 Discussion

We proposed a model-based primal-dual algorithm for planning in CMDPs. Via upper and lower bounds, we proved that our algorithm is near minimax optimal for both the relaxed and strict feasibility settings. Our results demonstrate that solving CMDPs is as easy as MDPs when small constraint violations are allowed, but inherently more difficult when we demand zero constraint violation. Algorithmically, we required a specific primal-dual approach that involved solving a sequence of MDPs. In contrast, model-based approaches for MDPs [2, 21] allow the use of any black-box planner. It is possible to obtain an $O\left(\frac{H^3 S^2 A}{\varepsilon^2}\right)$ sample complexity for a black-box CMDP planner in the relaxed feasibility setting. However, the $O(S^2 A)$ dependence in the bound implies that we need to estimate all entries in the transition probability matrix, and is therefore loose in the special case of unconstrained MDPs [2, 21]. In the future, we aim to extend our near-optimal sample complexity results to black-box CMDP solvers.

Acknowledgements

We would like to thank Reza Babanezhad and Arushi Jain for helpful feedback on the paper. Csaba Szepesvári gratefully acknowledges the funding from Natural Sciences and Engineering Research Council (NSERC) of Canada, "Design.R AI-assisted CPS Design" (DARPA) project and the Canada CIFAR AI Chairs Program for Amii. Lin Yang is supported in part by DARPA grant HR00112190130. This work was partly done while Lin Yang was visiting Deepmind.

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, pages 22–31. PMLR, 2017.

[2] Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020.

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

[4] Mohammad Gheshlaghi Azar, Rémi Munos, and Bert Kappen. On the sample complexity of reinforcement learning with a generative model. *arXiv preprint arXiv:1206.6461*, 2012.

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

[6] Qinbo Bai, Amrit Singh Bedi, Mridul Agarwal, Alec Koppel, and Vaneet Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. *arXiv preprint arXiv:2109.06332*, 2021.

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

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

[9] Kianté Brantley, Miroslav Dudik, Thodoris Lykouris, Sobhan Miryoosefi, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings. *arXiv preprint arXiv:2006.05051*, 2020.

[10] Chiara Buratti, Andrea Conti, Davide Dardari, and Roberto Verdone. An overview on wireless sensor networks technology and evolution. *Sensors*, 9(9):6869–6896, 2009.

[11] Yi Chen, Jing Dong, and Zhaoran Wang. A primal-dual approach to constrained markov decision processes. *arXiv preprint arXiv:2101.10895*, 2021.

[12] Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3304–3312. PMLR, 2021.

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

[14] Fei Feng, Wotao Yin, and Lin F. Yang. Does knowledge transfer always help to learn a better policy? *CoRR*, abs/1912.02986, 2019. URL http://arxiv.org/abs/1912.02986.

[15] Ather Gattami, Qinbo Bai, and Vaneet Aggarwal. Reinforcement learning for constrained markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, pages 2656–2664. PMLR, 2021.

[16] Aria HasanzadeZonuzy, Dileep M. Kalathil, and Srinivas Shakkottai. Model-based reinforcement learning for infinite-horizon discounted constrained markov decision processes. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 2519–2525. ijcai.org, 2021.

[17] Arushi Jain, Sharan Vaswani, Reza Babanezhad, Csaba Szepesvari, and Doina Precup. Towards painless policy optimization for constrained mdps. *arXiv preprint arXiv:2204.05176*, 2022.

[18] Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom), 2003.

[19] Krishna Chaitanya Kalagarla, Rahul Jain, and Pierluigi Nuzzo. A sample-efficient algorithm for episodic finite-horizon MDP with constraints. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, pages 8030–8037. AAAI Press, 2021.

[20] Michael Kearns and Satinder Singh. Finite-sample convergence rates for q-learning and indirect algorithms. *Advances in neural information processing systems*, pages 996–1002, 1999.

[21] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in neural information processing systems*, 33:12861–12872, 2020.

[22] Sobhan Miryoosefi and Chi Jin. A simple reward-free approach to constrained reinforcement learning. In *International Conference on Machine Learning*, pages 15666–15698. PMLR, 2022.

[23] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[24] Santiago Paternain, Luiz FO Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. *arXiv preprint arXiv:1910.13393*, 2019.

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

[26] Andrew J Schaefer, Matthew D Bailey, Steven M Shechter, and Mark S Roberts. Modeling medical treatment using markov decision processes. In *Operations research and health care*, pages 593–612. Springer, 2005.

[27] Aaron Sidford, Mengdi Wang, Xian Wu, Lin F Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 5192–5202, 2018.

[28] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[29] Jie Tan, Tingnan Zhang, Erwin Coumans, Atil Iscen, Yunfei Bai, Danijar Hafner, Steven Bohez, and Vincent Vanhoucke. Sim-to-real: Learning agile locomotion for quadruped robots. *arXiv preprint arXiv:1804.10332*, 2018.

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

[31] Akifumi Wachi and Yanan Sui. Safe reinforcement learning in constrained markov decision processes. In *International Conference on Machine Learning*, pages 9797–9806. PMLR, 2020.

[32] Mengdi Wang. Randomized linear programming solves the markov decision problem in nearly linear (sometimes sublinear) time. *Mathematics of Operations Research*, 45(2):517–546, 2020.

[33] Honghao Wei, Xin Liu, and Lei Ying. A provably-efficient model-free algorithm for constrained markov decision processes. *arXiv preprint arXiv:2106.01577*, 2021.

[34] Chenjun Xiao, Ilbin Lee, Bo Dai, Dale Schuurmans, and Csaba Szepesvari. On the sample complexity of batch reinforcement learning with policy-induced data. *arXiv preprint arXiv:2106.09973*, 2021.

[35] Tengyu Xu, Yingbin Liang, and Guanghui Lan. Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, pages 11480–11491. PMLR, 2021.

[36] Tiancheng Yu, Yi Tian, Jingzhao Zhang, and Suvrit Sra. Provably efficient algorithms for multi-objective competitive rl. In *International Conference on Machine Learning*, pages 12167–12176. PMLR, 2021.

[37] Andy Zeng, Shuran Song, Johnny Lee, Alberto Rodriguez, and Thomas Funkhouser. Tossingbot: Learning to throw arbitrary objects with residual physics. *IEEE Transactions on Robotics*, 36 (4):1307–1319, 2020.

[38] Liyuan Zheng and Lillian Ratliff. Constrained upper confidence reinforcement learning. In *Learning for Dynamics and Control*, pages 620–629. PMLR, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [N/A]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Supplementary material

Organization of the Appendix

A Proofs for primal-dual algorithm
B Proof of Theorem 2
C Proof of Theorem 3
D Concentration proofs
E Lower Bound
F Estimating ζ
G Comparison to Bai et al. [6]

A Proofs for primal-dual algorithm

Theorem 1 (Guarantees for the primal-dual algorithm)**.** For a target error $\varepsilon_{\text{opt}} > 0$ and the primal-dual updates in Eq. (5)-Eq. (6) with $U > |\lambda^*|$, $T = \frac{4U^2}{\varepsilon_{\text{opt}}^2(1-\gamma)^2}\left[1 + \frac{1}{(U-\lambda^*)^2}\right]$, $\eta = \frac{U(1-\gamma)}{\sqrt{T}}$ and $\varepsilon_1 = \frac{\varepsilon_{\text{opt}}^2(1-\gamma)^2\,(U-\lambda^*)}{6U}$, the mixture policy $\bar{\pi}_T := \frac{1}{T}\sum_{t=0}^{T-1}\hat{\pi}_t$ satisfies,

$$\hat{V}_{r_p}^{\bar{\pi}_T}(\rho) \geq \hat{V}_{r_p}^{\hat{\pi}^*}(\rho) - \varepsilon_{\text{opt}} \quad ; \quad \hat{V}_c^{\bar{\pi}_T}(\rho) \geq b' - \varepsilon_{\text{opt}}.$$

Proof. We will define the dual regret w.r.t λ as the following quantity, $R^d(\lambda, T)$.

$$R^d(\lambda, T) := \sum_{t=0}^{T-1}(\lambda_t - \lambda)\,(\hat{V}_c^{\hat{\pi}_t}(\rho) - b') \tag{13}$$

Using the primal update in Eq. (5), for any π,

$$\hat{V}_{r_p}^{\hat{\pi}_t}(\rho) + \lambda_t \hat{V}_c^{\hat{\pi}_t}(\rho) \geq \hat{V}_{r_p}^{\pi}(\rho) + \lambda_t \hat{V}_c^{\pi}(\rho) \tag{14}$$

Substituting $\pi = \hat{\pi}^*$, we have,

$$\hat{V}_{r_p}^{\hat{\pi}^*}(\rho) - \hat{V}_{r_p}^{\hat{\pi}_t}(\rho) \leq \lambda_t[\hat{V}_c^{\hat{\pi}_t}(\rho) - \hat{V}_c^{\hat{\pi}^*}(\rho)] \tag{15}$$

Since $\hat{\pi}^*$ is a solution to the empirical CMDP, $\hat{V}_c^{\hat{\pi}^*}(\rho) \geq b'$, we have

$$\hat{V}_{r_p}^{\hat{\pi}^*}(\rho) - \hat{V}_{r_p}^{\hat{\pi}_t}(\rho) \leq \lambda_t[\hat{V}_c^{\hat{\pi}_t}(\rho) - b'] \tag{16}$$

Averaging from $t = 0$ to $T - 1$, and using the definition of the dual regret in Eq. (13),

$$\frac{1}{T}\left[\sum_{t=0}^{T-1}[\hat{V}_{r_p}^{\hat{\pi}^*}(\rho) - \hat{V}_{r_p}^{\hat{\pi}_t}(\rho)]\right] \leq \frac{R^d(0, T)}{T} \tag{17}$$

This allows us to get a handle on the average optimality gap in terms of the average dual regret. For the second part, starting from the definition of the dual regret,

$$\sum_{t=0}^{T-1}(\lambda_t - \lambda)\,(\hat{V}_c^{\hat{\pi}_t}(\rho) - b') = R^d(\lambda, T)$$

Dividing by T, and using Eq. (16),

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

The above statement holds for all λ. If $A \leq 0$, it implies that $\frac{1}{T}\left[\sum_{t=0}^{T-1}(b' - \hat{V}_c^{\hat{\pi}_t}(\rho))\right] \leq 0$, which is the desired bound on the constraint violation. In this case, we can set $\lambda = 0$ and this recovers Eq. (17).

If $A > 0$, we will set $\lambda = U$. We define $\bar{\pi}_T = \frac{1}{T}\sum_{t=0}^{T-1}\hat{\pi}_t$ as a mixture policy. Using the linearity of $\hat{V}_{r_p}^{\pi}(\rho)$ and $\hat{V}_c^{\pi}(\rho)$, i.e. $\frac{1}{T}\sum_{t=0}^{T-1}\hat{V}_{r_p}^{\hat{\pi}_t}(\rho) = \hat{V}_{r_p}^{\bar{\pi}_T}(\rho)$, and $\frac{1}{T}\sum_{t=0}^{T-1}\hat{V}_c^{\hat{\pi}_t}(\rho) = \hat{V}_c^{\bar{\pi}_T}(\rho)$, we obtain the following equation,

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

Since $(b' - \hat{V}_c^{\bar{\pi}_T}(\rho)) > 0$,

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

where, $[x]_+ = \max\{x, 0\}$

Using Lemma 10 for $U > \lambda^*$,

$$\left[b' - \hat{V}_c^{\bar{\pi}_T}(\rho)\right]_+ \leq \frac{R^d(U,T)}{T(U-\lambda^*)} \tag{18}$$

It remains to bound the dual regret for the gradient descent procedure for updating λ_t. For this, we bound the distance $|\lambda_{t+1} - \lambda|$ for a general $\lambda \in [0, U]$. Defining $\lambda'_{t+1} := \mathbb{P}_{[0,U]}[\lambda_t - \eta\,(\hat{V}_c^{\hat{\pi}_t}(\rho) - b')]$,

$$\begin{aligned}
|\lambda_{t+1} - \lambda| &= |\mathcal{R}_\Lambda[\lambda'_{t+1}] - \lambda| = |\mathcal{R}_\Lambda[\lambda'_{t+1}] - \lambda'_{t+1} + \lambda'_{t+1} - \lambda| \leq |\mathcal{R}_\Lambda[\lambda'_{t+1}] - \lambda'_{t+1}| + |\lambda'_{t+1} - \lambda| \\
&\leq \varepsilon_1 + |\lambda'_{t+1} - \lambda| \\
&\qquad \text{(Since } |\lambda - \mathcal{R}_\Lambda[\lambda]| \leq \varepsilon_1 \text{ for all } \lambda \in [0,U] \text{ because of the epsilon-net.)}
\end{aligned}$$

Squaring both sides,

$$\begin{aligned}
|\lambda_{t+1} - \lambda|^2 &= \varepsilon_1^2 + |\lambda'_{t+1} - \lambda|^2 + 2\varepsilon_1|\lambda'_{t+1} - \lambda| \leq \varepsilon_1^2 + 2\varepsilon_1 U + |\lambda'_{t+1} - \lambda|^2 \\
&\qquad \text{(Since } \lambda,\, \lambda'_{t+1} \in [0,U],) \\
&\leq \varepsilon_1^2 + 2\varepsilon_1 U + |\lambda_t - \eta\,(\hat{V}_c^{\hat{\pi}_t}(\rho) - b') - \lambda|^2 \\
&\qquad \text{(Since projections are non-expansive)} \\
&= \varepsilon_1^2 + 2\varepsilon_1 U + |\lambda_t - \lambda|^2 - 2\eta\,(\lambda_t - \lambda)\,(\hat{V}_c^{\hat{\pi}_t}(\rho) - b') + \eta^2(\hat{V}_c^{\hat{\pi}_t}(\rho) - b')^2 \\
&\leq \varepsilon_1^2 + 2\varepsilon_1 U + |\lambda_t - \lambda|^2 - 2\eta\,(\lambda_t - \lambda)\,(\hat{V}_c^{\hat{\pi}_t}(\rho) - b') + \frac{\eta^2}{(1-\gamma)^2}
\end{aligned}$$

Rearranging and dividing by 2η,

$$(\lambda_t - \lambda)\,(\hat{V}_c^{\hat{\pi}_t}(\rho) - b') \leq \frac{\varepsilon_1^2 + 2\varepsilon_1 U}{2\eta} + \frac{|\lambda_t - \lambda|^2 - |\lambda_{t+1} - \lambda|^2}{2\eta} + \frac{\eta}{2(1-\gamma)^2}$$

Summing from $t = 0$ to $T - 1$ and using the definition of the dual regret,

$$R^d(\lambda, T) \leq T\,\frac{\varepsilon_1^2 + 2\varepsilon_1 U}{2\eta} + \frac{1}{2\eta}\sum_{t=0}^{T-1}[|\lambda_t - \lambda|^2 - |\lambda_{t+1} - \lambda|^2] + \frac{\eta T}{2(1-\gamma)^2}$$

Telescoping and bounding $|\lambda_0 - \lambda|$ by U,

$$\leq T\,\frac{\varepsilon_1^2 + 2\varepsilon_1 U}{2\eta} + \frac{U^2}{2\eta} + \frac{\eta T}{2(1-\gamma)^2}$$

Setting $\eta = \frac{U(1-\gamma)}{\sqrt{T}}$,

$$R^d(\lambda, T) \leq T^{3/2}\,\frac{\varepsilon_1^2 + 2\varepsilon_1 U}{2U(1-\gamma)} + \frac{U\sqrt{T}}{1-\gamma} \tag{19}$$

Using Eq. (19) with the expressions for the optimality gap (Eq. (17)) and constraint violation (Eq. (18)), we obtain the following bounds.

For the reward optimality gap, since $\lambda = 0 \in [0, U]$,

$$\hat{V}_{r_p}^{\hat{\pi}^*}(\rho) - \hat{V}_{r_p}^{\bar{\pi}_T}(\rho) \leq \sqrt{T}\,\frac{\varepsilon_1^2 + 2\varepsilon_1 U}{2U(1-\gamma)} + \frac{U}{(1-\gamma)\sqrt{T}} < \sqrt{T}\,\frac{3\varepsilon_1}{2(1-\gamma)} + \frac{U}{(1-\gamma)\sqrt{T}}$$

(Since $\varepsilon_1 < U$)

For the constraint violation, since $U \in [0, U]$,

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

(Since $\varepsilon_1 < U$)

Let us set T s.t. that the second term in both quantities is bound by $\frac{\varepsilon_{\text{opt}}}{2}$,

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

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

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

We want to set ε_1 s.t. the first term in both quantities is also bounded by $\frac{\varepsilon_{\text{opt}}}{2}$,

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

With these values, the algorithm ensures that,

$$\hat{V}_{r_p}^{\hat{\pi}^*}(\rho) - \hat{V}_{r_p}^{\bar{\pi}_T}(\rho) \leq \varepsilon_{\text{opt}} \quad ; \quad \left[b' - \hat{V}_c^{\bar{\pi}_T}(\rho)\right] \leq \varepsilon_{\text{opt}}.$$

□

Lemma 9 (Bounding the dual variable)**.** *The objective Eq. (4) satisfies strong duality. Defining* $\pi_c^* := \arg\max V_c^{\pi}(\rho)$. *We consider two cases: (1) If* $b' = b - \varepsilon'$ *for* $\varepsilon' > 0$ *and event* $\mathcal{E}_1 = \left\{\left|\hat{V}_c^{\pi_c^*}(\rho) - V_c^{\pi_c^*}(\rho)\right| \leq \frac{\varepsilon'}{2}\right\}$ *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* $\mathcal{E}_2 = \left\{\left|\hat{V}_c^{\pi_c^*}(\rho) - V_c^{\pi_c^*}(\rho)\right| \leq \frac{\zeta}{2} - \Delta\right\}$ *holds, then* $\lambda^* \leq \frac{2(1+\omega)}{\zeta(1-\gamma)}$.

Proof. Writing the empirical CMDP in Eq. (4) in its Lagrangian form,

$$\hat{V}_{r_p}^{\hat{\pi}^*}(\rho) = \max_{\pi}\min_{\lambda \geq 0} \hat{V}_{r_p}^{\pi}(\rho) + \lambda[\hat{V}_c^{\pi}(\rho) - b']$$

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 [24], implying,

$$= \min_{\lambda \geq 0}\max_{\pi} \hat{V}_{r_p}^{\pi}(\rho) + \lambda[\hat{V}_c^{\pi}(\rho) - b']$$

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

$$= \max_{\pi} \hat{V}_{r_p}^{\pi}(\rho) + \lambda^*\left[\hat{V}_c^{\pi}(\rho) - b'\right]$$

Define $\pi_c^* := \arg\max V_c^\pi(\rho)$ and $\hat{\pi}_c^* := \arg\max \hat{V}_c^\pi(\rho)$

$$\geq \hat{V}_{r_p}^{\hat{\pi}_c^*}(\rho) + \lambda^* \left[\hat{V}_c^{\hat{\pi}_c^*}(\rho) - b'\right]$$
$$= \hat{V}_{r_p}^{\hat{\pi}_c^*}(\rho) + \lambda^* \left[\left(\hat{V}_c^{\hat{\pi}_c^*}(\rho) - V_c^{\pi_c^*}(\rho)\right) + (V_c^{\pi_c^*}(\rho) - b) + (b - b')\right]$$

By definition, $\zeta = V_c^{\pi_c^*}(\rho) - b$

$$= \hat{V}_{r_p}^{\hat{\pi}_c^*}(\rho) + \lambda^* \left[\left(\hat{V}_c^{\hat{\pi}_c^*}(\rho) - \hat{V}_c^{\pi_c^*}(\rho)\right) + \left(\hat{V}_c^{\pi_c^*}(\rho) - V_c^{\pi_c^*}(\rho)\right) + \zeta + (b - b')\right]$$

By definition of $\hat{\pi}_c^*$, $\left(\hat{V}_c^{\hat{\pi}_c^*}(\rho) - \hat{V}_c^{\pi_c^*}(\rho)\right) \geq 0$

$$\hat{V}_{r_p}^{\hat{\pi}^*}(\rho) \geq \hat{V}_{r_p}^{\hat{\pi}_c^*}(\rho) + \lambda^* \left[\zeta + (b - b') - \left|\hat{V}_c^{\pi_c^*}(\rho) - V_c^{\pi_c^*}(\rho)\right|\right]$$

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

$$\hat{V}_{r_p}^{\hat{\pi}^*}(\rho) \geq \hat{V}_{r_p}^{\hat{\pi}_c^*}(\rho) + \lambda^* \left[\zeta + \varepsilon' - \left|\hat{V}_c^{\pi_c^*}(\rho) - V_c^{\pi_c^*}(\rho)\right|\right]$$

If the event $\mathcal{E}_1$ holds, $\left|\hat{V}_c^{\pi_c^*}(\rho) - V_c^{\pi_c^*}(\rho)\right| \leq \frac{\varepsilon'}{2}$, implying, $\left|\hat{V}_c^{\pi_c^*}(\rho) - V_c^{\pi_c^*}(\rho)\right| < \zeta + \frac{\varepsilon'}{2}$, then,

$$\geq \hat{V}_{r_p}^{\hat{\pi}_c^*}(\rho) + \lambda^* \frac{\varepsilon'}{2}$$
$$\implies \lambda^* \leq \frac{2}{\varepsilon'}[\hat{V}_{r_p}^{\hat{\pi}^*}(\rho) - \hat{V}_{r_p}^{\hat{\pi}_c^*}(\rho)] \leq \frac{2(1+\omega)}{\varepsilon'(1-\gamma)}$$

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

$$\hat{V}_{r_p}^{\hat{\pi}^*}(\rho) \geq \hat{V}_{r_p}^{\hat{\pi}_c^*}(\rho) + \lambda^* \left[\zeta - \Delta - \left|\hat{V}_c^{\pi_c^*}(\rho) - V_c^{\pi_c^*}(\rho)\right|\right]$$

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

$$\geq \hat{V}_{r_p}^{\hat{\pi}_c^*}(\rho) + \lambda^* \frac{\zeta}{2}$$
$$\implies \lambda^* \leq \frac{2}{\zeta}[\hat{V}_{r_p}^{\hat{\pi}^*}(\rho) - \hat{V}_{r_p}^{\hat{\pi}_c^*}(\rho)] \leq \frac{2(1+\omega)}{\zeta(1-\gamma)}$$

□

Lemma 10 (Lemma B.2 of Jain et al. [17]). *For any $C > \lambda^*$ and any $\tilde{\pi}$ s.t. $\hat{V}_{r_p}^{\hat{\pi}^*}(\rho) - \hat{V}_{r_p}^{\tilde{\pi}}(\rho) + C[b - \hat{V}_c^{\tilde{\pi}}(\rho)]_+ \leq \beta$, we have $[b - \hat{V}_c^{\tilde{\pi}}(\rho)]_+ \leq \frac{\beta}{C - \lambda^*}$.*

Proof. Define $\nu(\tau) = \max_\pi\{V_r^\pi(\rho) \mid V_c^\pi(\rho) \geq b + \tau\}$ and note that by definition, $\nu(0) = V_r^{\hat{\pi}^*}(\rho)$ and that ν is a decreasing function for its argument.

Let $V_l^{\pi,\lambda}(\rho) = V_r^\pi(\rho) + \lambda(V_c^\pi(\rho) - b)$. Then, for any policy π s.t. $V_c^\pi(\rho) \geq b + \tau$, we have

$$\begin{aligned}
V_l^{\pi,\lambda^*}(\rho) &\leq \max_{\pi'} V_l^{\pi',\lambda^*}(\rho) \\
&= V_r^{\hat{\pi}^*}(\rho) && \text{(by strong duality)} \\
&= \nu(0) && \text{(from above relation)} \\
\implies \nu(0) - \tau\lambda^* &\geq V_l^{\pi,\lambda^*}(\rho) - \tau\lambda^* = V_r^\pi(\rho) + \lambda^* \underbrace{(V_c^\pi(\rho) - b - \tau)}_{\text{Non-negative}} \\
\implies \nu(0) - \tau\lambda^* &\geq \max_\pi\{V_r^\pi(\rho) \mid V_c^\pi(\rho) \geq b + \tau\} = \nu(\tau)\,. \\
\implies \tau\lambda^* &\leq \nu(0) - \nu(\tau)\,. && (20)
\end{aligned}$$

Now we choose $\tilde{\tau} = -(b - V_c^{\tilde{\pi}}(\rho))_+$.

$$\begin{aligned}
(C - \lambda^*)|\tilde{\tau}| &= \lambda^* \tilde{\tau} + C|\tilde{\tau}| && \text{(since } \tilde{\tau} \le 0\text{)} \\
&\le \nu(0) - \nu(\tilde{\tau}) + C|\tilde{\tau}| && \text{(Eq. (20))} \\
&= V_r^{\hat{\pi}^*}(\rho) - V_r^{\tilde{\pi}}(\rho) + C|\tilde{\tau}| + V_r^{\tilde{\pi}}(\rho) - \nu(\tilde{\tau}) && \text{(definition of } \nu(0)\text{)} \\
&= V_r^{\hat{\pi}^*}(\rho) - V_r^{\tilde{\pi}}(\rho) + C(b - V_c^{\tilde{\pi}}(\rho))_+ + V_r^{\tilde{\pi}}(\rho) - \nu(\tilde{\tau}) \\
&\le \beta + V_r^{\tilde{\pi}}(\rho) - \nu(\tilde{\tau})\,.
\end{aligned}$$

Now let us bound $\nu(\tilde{\tau})$:

$$\begin{aligned}
\nu(\tilde{\tau}) &= \max_{\pi}\{V_r^{\pi}(\rho) \mid V_c^{\pi}(\rho) \ge b - (b - V_c^{\tilde{\pi}}(\rho))_+\} \\
&\ge \max_{\pi}\{V_r^{\pi}(\rho) \mid V_c^{\pi}(\rho) \ge V_c^{\tilde{\pi}}(\rho)\} && \text{(tightening the constraint)}
\end{aligned}$$

$$\nu(\tilde{\tau}) \ge V_r^{\tilde{\pi}}(\rho) \implies (C - \lambda^*)|\tilde{\tau}| \le \beta \implies (b - V_c^{\tilde{\pi}}(\rho))_+ \le \frac{\beta}{C - \lambda^*}$$

□

B Proof of Theorem 2

Theorem 2. For a fixed $\varepsilon \in (0, 1/1-\gamma]$ and $\delta \in (0,1)$, Algorithm 1 with $N = \tilde{O}\left(\frac{\log(1/\delta)}{(1-\gamma)^3\varepsilon^2}\right)$ samples, $b' = b - \frac{3\varepsilon}{8}$, $\omega = \frac{\varepsilon(1-\gamma)}{8}$, $U = O\left(1/\varepsilon\,(1-\gamma)\right)$, $\varepsilon_1 = O\left(\varepsilon^2(1-\gamma)^2\right)$ and $T = O\left(1/(1-\gamma)^4\varepsilon^4\right)$, returns policy $\bar{\pi}_T$ that satisfies the objective in Eq. (2) with probability at least $1 - 4\delta$.

Proof. We fill in the details required for the proof sketch in the main paper. Proceeding according to the proof sketch, we first detail the computation of T and ε_1 for the primal-dual algorithm. Recall that $U = \frac{32}{5\varepsilon\,(1-\gamma)}$ and $\varepsilon_{\text{opt}} = \frac{\varepsilon}{4}$. Using Theorem 1, we need to set

$$T = \frac{4U^2}{\varepsilon_{\text{opt}}^2\,(1-\gamma)^2}\left[1 + \frac{1}{(U - \lambda^*)^2}\right] = \frac{64}{\varepsilon^2(1-\gamma)^2}\left[1 + \frac{1}{(U - \lambda^*)^2}\right]$$

Recall that $|\lambda^*| \le C := \frac{16}{5\varepsilon\,(1-\gamma)}$ and $U = 2C$. Simplifying,

$$\begin{aligned}
&\le \frac{256}{\varepsilon^2(1-\gamma)^2}\left[C^2 + 1\right] < \frac{512}{\varepsilon^2(1-\gamma)^2}C^2 = \frac{512}{\varepsilon^2(1-\gamma)^2}\,\frac{256}{25\varepsilon^2\,(1-\gamma)^2} \\
\implies\ & T = O\left(1/\varepsilon^4(1-\gamma)^4\right).
\end{aligned}$$

Using Theorem 1, we need to set ε_1,

$$\begin{aligned}
\varepsilon_1 &= \frac{\varepsilon_{\text{opt}}^2(1-\gamma)^2\,(U - \lambda^*)}{6U} = \frac{\varepsilon^2(1-\gamma)^2\,(U - \lambda^*)}{96U} \le \frac{\varepsilon^2(1-\gamma)^2}{96} \\
\implies\ & \varepsilon_1 = O\left(\varepsilon^2(1-\gamma)^2\right).
\end{aligned}$$

For bounding the concentration terms for $\bar{\pi}_T$ in Eq. (9), we use Theorem 6 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, $\iota = \frac{\omega\,\delta\,(1-\gamma)\,\varepsilon_1}{30\,U|S||A|^2} = O\left(\frac{\delta\varepsilon^4\,(1-\gamma)^4}{SA^2}\right)$ and

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

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

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

We use the Lemma 7 to bound the remaining concentration terms for π^* and π_c^* in Eq. (9). In this case, for $C'(\delta) = 72\log\left(\frac{4S\log(e/1-\gamma)}{\delta}\right)$, we require that,

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

Hence, if $N \ge \tilde{O}\left(\frac{\log(1/\delta)}{(1-\gamma)^3\,\varepsilon^2}\right)$, the bounds in Eq. (9) are satisfied, completing the proof. □

Lemma 11 (Decomposing the suboptimality)**.** *For $b' = b - \frac{\varepsilon - \varepsilon_{opt}}{2}$, if (i) $\varepsilon_{opt} < \varepsilon$, and (ii) the following conditions are satisfied,*

$$\left|V_c^{\bar{\pi}_T}(\rho) - \hat{V}_c^{\bar{\pi}_T}(\rho)\right| \leq \frac{\varepsilon - \varepsilon_{opt}}{2}\ ;\ \left|V_c^{\pi^*}(\rho) - \hat{V}_c^{\pi^*}(\rho)\right| \leq \frac{\varepsilon - \varepsilon_{opt}}{2}\ ;\ \left|\hat{V}_c^{\pi_c^*}(\rho) - V_c^{\pi_c^*}(\rho)\right| \leq \frac{\zeta}{4}$$

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

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

Proof. From Theorem 1, we know that,

$$\hat{V}_c^{\bar{\pi}_T}(\rho) \geq b' - \varepsilon_{\text{opt}} \implies V_c^{\bar{\pi}_T}(\rho) \geq V_c^{\bar{\pi}_T}(\rho) - \hat{V}_c^{\bar{\pi}_T}(\rho) + b' - \varepsilon_{\text{opt}} \geq -\left|V_c^{\bar{\pi}_T}(\rho) - \hat{V}_c^{\bar{\pi}_T}(\rho)\right| + b' - \varepsilon_{\text{opt}}$$

Since we require $\bar{\pi}_T$ to violate the constraint in the true CMDP by at most ε, we require $V_c^{\bar{\pi}_T}(\rho) \geq b - \varepsilon$. From the above equation, a sufficient condition for ensuring this is,

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

meaning that we require $\left|V_c^{\bar{\pi}_T}(\rho) - \hat{V}_c^{\bar{\pi}_T}(\rho)\right| \leq (b' - b) - \varepsilon_{\text{opt}} + \varepsilon$.

In the subsequent analysis, we will require π^* to be feasible for the constrained problem in Eq. (4), meaning that we require $\hat{V}_c^{\pi^*}(\rho) \geq b'$. Since π^* is the solution to Eq. (1), we know that,

$$V_c^{\pi^*}(\rho) \geq b \implies \hat{V}_c^{\pi^*}(\rho) \geq b - \left|V_c^{\pi^*}(\rho) - \hat{V}_c^{\pi^*}(\rho)\right|$$

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

$$b - \left|V_c^{\pi^*}(\rho) - \hat{V}_c^{\pi^*}(\rho)\right| \geq b' \text{meaning that we require} \left|V_c^{\pi^*}(\rho) - \hat{V}_c^{\pi^*}(\rho)\right| \leq b - b'.$$

Since $b' = b - \frac{\varepsilon - \varepsilon_{\text{opt}}}{2}$, we require the following statements to hold:

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

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

$$\begin{aligned}
&V_r^{\pi^*}(\rho) - V_r^{\bar{\pi}_T}(\rho) \\
&= V_r^{\pi^*}(\rho) - V_{r_p}^{\pi^*}(\rho) + V_{r_p}^{\pi^*}(\rho) - V_r^{\bar{\pi}_T}(\rho) \\
&= [V_r^{\pi^*}(\rho) - V_{r_p}^{\pi^*}(\rho)] + V_{r_p}^{\pi^*}(\rho) - \hat{V}_{r_p}^{\pi^*}(\rho) + \hat{V}_{r_p}^{\pi^*}(\rho) - V_r^{\bar{\pi}_T}(\rho) \\
&\leq [V_r^{\pi^*}(\rho) - V_{r_p}^{\pi^*}(\rho)] + [V_{r_p}^{\pi^*}(\rho) - \hat{V}_{r_p}^{\pi^*}(\rho)] + \hat{V}_{r_p}^{\hat{\pi}^*}(\rho) - V_r^{\bar{\pi}_T}(\rho) \\
&\qquad\qquad \text{(By optimality of } \hat{\pi}^* \text{ and since we have ensured that } \pi^* \text{ is feasible for Eq. (4))} \\
&= [V_r^{\pi^*}(\rho) - V_{r_p}^{\pi^*}(\rho)] + [V_{r_p}^{\pi^*}(\rho) - \hat{V}_{r_p}^{\pi^*}(\rho)] + [\hat{V}_{r_p}^{\hat{\pi}^*}(\rho) - \hat{V}_{r_p}^{\bar{\pi}_T}(\rho)] + \hat{V}_{r_p}^{\bar{\pi}_T}(\rho) - V_r^{\bar{\pi}_T}(\rho) \\
&= \underbrace{[V_r^{\pi^*}(\rho) - V_{r_p}^{\pi^*}(\rho)]}_{\text{Perturbation Error}} + \underbrace{[V_{r_p}^{\pi^*}(\rho) - \hat{V}_{r_p}^{\pi^*}(\rho)]}_{\text{Concentration Error}} + \underbrace{[\hat{V}_{r_p}^{\hat{\pi}^*}(\rho) - \hat{V}_{r_p}^{\bar{\pi}_T}(\rho)]}_{\text{Primal-Dual Error}} + \underbrace{[\hat{V}_{r_p}^{\bar{\pi}_T}(\rho) - V_{r_p}^{\bar{\pi}_T}(\rho)]}_{\text{Concentration Error}} + \underbrace{[V_{r_p}^{\bar{\pi}_T}(\rho) - V_r^{\bar{\pi}_T}(\rho)]}_{\text{Perturbation Error}}
\end{aligned}$$

For a perturbation magnitude equal to ω, we use Lemma 15 to bound both perturbation errors by $\frac{\omega}{1-\gamma}$. Using Theorem 1 to bound the primal-dual error by ε_{opt},

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

□

C Proof of Theorem 3

Theorem 3. For a fixed $\varepsilon \in (0, 1/1-\gamma]$ and $\delta \in (0,1)$, Algorithm 1, with $N = \tilde{O}\left(\frac{\log(1/\delta)}{(1-\gamma)^5 \varepsilon^2 \zeta^2}\right)$ samples, $b' = b + \frac{\varepsilon(1-\gamma)\zeta}{20}$, $\omega = \frac{\varepsilon(1-\gamma)}{10}$, $U = \frac{4(1+\omega)}{\zeta(1-\gamma)}$, $\varepsilon_1 = O\left(\varepsilon^2(1-\gamma)^4\zeta^2\right)$ and $T = O\left(1/(1-\gamma)^6\zeta^4\varepsilon^2\right)$ returns policy $\bar{\pi}_T$ that satisfies the objective in Eq. (3), with probability at least $1 - 4\delta$.

Proof. We fill in the details required for the proof sketch in the main paper. Proceeding according to the proof sketch, we first detail the computation of T and ε_1 for the primal-dual algorithm. Recall that $U = \frac{8}{\zeta(1-\gamma)}$, $\Delta = \frac{\varepsilon(1-\gamma)\zeta}{40}$ and $\varepsilon_{\text{opt}} = \frac{\Delta}{5}$. Using Theorem 1, we need to set

$$T = \frac{4U^2}{\varepsilon_{\text{opt}}^2 (1-\gamma)^2}\left[1 + \frac{1}{(U-\lambda^*)^2}\right] = \frac{100}{\Delta^2(1-\gamma)^2}\left[1 + \frac{1}{(U-\lambda^*)^2}\right]$$

Recall that $|\lambda^*| \leq C := \frac{4}{\zeta(1-\gamma)}$ and $U = 2C$. Simplifying,

$$\leq \frac{400}{\Delta^2(1-\gamma)^2}\left[C^2+1\right] < \frac{800}{\Delta^2(1-\gamma)^2}C^2 = \frac{800}{\Delta^2(1-\gamma)^2}\frac{16}{\zeta^2(1-\gamma)^2}$$

$$\implies T \leq \frac{800 \cdot 1600}{\varepsilon^2\zeta^2(1-\gamma)^4}\frac{16}{\zeta^2(1-\gamma)^2} = O\left(1/\varepsilon^2\zeta^4(1-\gamma)^6\right).$$

Using Theorem 1, we need to set ε_1,

$$\varepsilon_1 = \frac{\varepsilon_{\text{opt}}^2(1-\gamma)^2(U-\lambda^*)}{6U} = \frac{\Delta^2(1-\gamma)^2(U-\lambda^*)}{150U} \leq \frac{\Delta^2(1-\gamma)^2}{150}$$

$$\implies \varepsilon_1 \leq \frac{\varepsilon^2\zeta^2(1-\gamma)^4}{150 \cdot 1600} = O\left(\varepsilon^2\zeta^2(1-\gamma)^4\right).$$

For bounding the concentration terms for $\bar{\pi}_T$ in Eq. (12), we use Theorem 6 with $U = \frac{8}{\zeta(1-\gamma)}$, $\omega = \frac{\varepsilon(1-\gamma)}{10}$ and $\varepsilon_1 = \frac{\varepsilon^2\zeta^2(1-\gamma)^4}{150\cdot 1600}$. In this case, $\iota = \frac{\omega\,\delta\,(1-\gamma)\,\varepsilon_1}{30\,U|S||A|^2} = O\left(\frac{\delta\varepsilon^3\zeta^3(1-\gamma)^7}{SA^2}\right)$ and

$$C(\delta) = 72\log\left(\frac{16(1+U+\omega)\,SA\log(e/1-\gamma)}{(1-\gamma)^2\,\iota\,\delta}\right) = O\left(\log\left(\frac{S^2A^3}{(1-\gamma)^{10}\delta^2\varepsilon^3\zeta^4}\right)\right).$$

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

$$2\sqrt{\frac{C(\delta)}{N\cdot(1-\gamma)^3}} \leq \frac{\Delta}{5} \implies N \geq O\left(\frac{C(\delta)}{(1-\gamma)^3\,\Delta^2}\right) \geq O\left(\frac{C(\delta)}{(1-\gamma)^5\,\zeta^2\,\varepsilon^2}\right)$$

We use the Lemma 7 to bound the remaining concentration terms for π^* and π_c^* in Eq. (12). In this case, for $C'(\delta) = 72\log\left(\frac{4S\log(e/1-\gamma)}{\delta}\right)$, we require that,

$$2\sqrt{\frac{C'(\delta)}{N\cdot(1-\gamma)^3}} \leq \frac{\Delta}{5} \implies N \geq O\left(\frac{C'(\delta)}{(1-\gamma)^3\,\Delta^2}\right) \geq O\left(\frac{C'(\delta)}{(1-\gamma)^5\,\zeta^2\,\varepsilon^2}\right)$$

Hence, if $N \geq \tilde{O}\left(\frac{\log(1/\delta)}{(1-\gamma)^5\,\zeta^2\,\varepsilon^2}\right)$, the bounds in Eq. (12) are satisfied, completing the proof. □

Lemma 12 (Decomposing the suboptimality). *For a fixed $\Delta > 0$ and $\varepsilon_{opt} < \Delta$, if $b' = b + \Delta$, then the following conditions are satisfied,*

$$\left|V_c^{\bar{\pi}_T}(\rho) - \hat{V}_c^{\bar{\pi}_T}(\rho)\right| \leq \Delta - \varepsilon_{opt}\,;\ \left|V_c^{\pi^*}(\rho) - \hat{V}_c^{\pi^*}(\rho)\right| \leq \Delta$$

then (a) policy $\bar{\pi}_T$ satisfies the constraint i.e. $V_c^{\bar{\pi}_T}(\rho) \geq b$ and (b) its optimality gap can be bounded as:

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

Proof. Compared to Eq. (4), we define a slightly modified CMDP problem by changing the constraint RHS to b'' for some b'' to be specified later. We denote its corresponding optimal policy as $\tilde{\pi}^*$. In particular,

$$\tilde{\pi}^* \in \arg\max_{\pi} \hat{V}_{r_p}^{\pi}(\rho) \text{ s.t. } \hat{V}_c^{\pi}(\rho) \geq b'' \tag{21}$$

From Theorem 1, we know that,

$$\hat{V}_c^{\bar{\pi}_T}(\rho) \geq b' - \varepsilon_{\text{opt}} \implies V_c^{\bar{\pi}_T}(\rho) \geq V_c^{\bar{\pi}_T}(\rho) - \hat{V}_c^{\bar{\pi}_T}(\rho) + b' - \varepsilon_{\text{opt}} \geq -\left|V_c^{\bar{\pi}_T}(\rho) - \hat{V}_c^{\bar{\pi}_T}(\rho)\right| + b' - \varepsilon_{\text{opt}}$$

Since we require $\bar{\pi}_T$ to satisfy the constraint in the true CMDP, we require $V_c^{\bar{\pi}_T}(\rho) \geq b$. From the above equation, a sufficient condition for ensuring this is,

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

meaning that we require $\left|V_c^{\bar{\pi}_T}(\rho) - \hat{V}_c^{\bar{\pi}_T}(\rho)\right| \leq (b' - b) - \varepsilon_{\text{opt}}$.

In the subsequent analysis, we will require π^* to be feasible for the constrained problem in Eq. (21). This implies that we require $\hat{V}_c^{\pi^*}(\rho) \geq b''$. Since π^* is the solution to Eq. (1), we know that,

$$V_c^{\pi^*}(\rho) \geq b \implies \hat{V}_c^{\pi^*}(\rho) \geq b - \left|V_c^{\pi^*}(\rho) - \hat{V}_c^{\pi^*}(\rho)\right|$$

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

$$b - \left|V_c^{\pi^*}(\rho) - \hat{V}_c^{\pi^*}(\rho)\right| \geq b'' \text{meaning that we require } \left|V_c^{\pi^*}(\rho) - \hat{V}_c^{\pi^*}(\rho)\right| \leq b - b''.$$

Hence we require the following statements to hold:

$$\left|V_c^{\bar{\pi}_T}(\rho) - \hat{V}_c^{\bar{\pi}_T}(\rho)\right| \leq (b' - b) - \varepsilon_{\text{opt}} \quad ; \quad \left|V_c^{\pi^*}(\rho) - \hat{V}_c^{\pi^*}(\rho)\right| \leq b - b''.$$

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

$$\begin{aligned}
V_r^{\pi^*}(\rho) - V_r^{\bar{\pi}_T}(\rho) &= V_r^{\pi^*}(\rho) - V_{r_p}^{\pi^*}(\rho) + V_{r_p}^{\pi^*}(\rho) - V_r^{\bar{\pi}_T}(\rho) \\
&= [V_r^{\pi^*}(\rho) - V_{r_p}^{\pi^*}(\rho)] + [V_{r_p}^{\pi^*}(\rho) - \hat{V}_{r_p}^{\pi^*}(\rho)] + \hat{V}_{r_p}^{\pi^*}(\rho) - V_r^{\bar{\pi}_T}(\rho) \\
&\leq [V_r^{\pi^*}(\rho) - V_{r_p}^{\pi^*}(\rho)] + [V_{r_p}^{\pi^*}(\rho) - \hat{V}_{r_p}^{\pi^*}(\rho)] + \hat{V}_{r_p}^{\tilde{\pi}^*}(\rho) - V_r^{\bar{\pi}_T}(\rho) \\
&\quad \text{(By optimality of } \hat{\pi}^* \text{ and since we have ensured that } \pi^* \text{ is feasible for Eq. (21))} \\
&= [V_r^{\pi^*}(\rho) - V_{r_p}^{\pi^*}(\rho)] + [V_{r_p}^{\pi^*}(\rho) - \hat{V}_{r_p}^{\pi^*}(\rho)] + [\hat{V}_{r_p}^{\tilde{\pi}^*}(\rho) - \hat{V}_{r_p}^{\hat{\pi}^*}(\rho)] + \hat{V}_{r_p}^{\hat{\pi}^*}(\rho) - V_r^{\bar{\pi}_T}(\rho) \\
&= [V_r^{\pi^*}(\rho) - V_{r_p}^{\pi^*}(\rho)] + [V_{r_p}^{\pi^*}(\rho) - \hat{V}_{r_p}^{\pi^*}(\rho)] + [\hat{V}_{r_p}^{\tilde{\pi}^*}(\rho) - \hat{V}_{r_p}^{\hat{\pi}^*}(\rho)] + \hat{V}_{r_p}^{\hat{\pi}^*}(\rho) - V_r^{\bar{\pi}_T}(\rho) \\
&= [V_r^{\pi^*}(\rho) - V_{r_p}^{\pi^*}(\rho)] + [V_{r_p}^{\pi^*}(\rho) - \hat{V}_{r_p}^{\pi^*}(\rho)] + [\hat{V}_{r_p}^{\tilde{\pi}^*}(\rho) - \hat{V}_{r_p}^{\hat{\pi}^*}(\rho)] + [\hat{V}_{r_p}^{\hat{\pi}^*}(\rho) - \hat{V}_{r_p}^{\bar{\pi}_T}(\rho)] \\
&\quad + \hat{V}_{r_p}^{\bar{\pi}_T}(\rho) - V_r^{\bar{\pi}_T}(\rho) \\
&= \underbrace{[V_r^{\pi^*}(\rho) - V_{r_p}^{\pi^*}(\rho)]}_{\text{Perturbation Error}} + \underbrace{[V_{r_p}^{\pi^*}(\rho) - \hat{V}_{r_p}^{\pi^*}(\rho)]}_{\text{Concentration Error}} + \underbrace{[\hat{V}_{r_p}^{\tilde{\pi}^*}(\rho) - \hat{V}_{r_p}^{\hat{\pi}^*}(\rho)]}_{\text{Sensitivity Error}} + \underbrace{[\hat{V}_{r_p}^{\hat{\pi}^*}(\rho) - \hat{V}_{r_p}^{\bar{\pi}_T}(\rho)]}_{\text{Primal-Dual Error}} \\
&\quad + \underbrace{[\hat{V}_{r_p}^{\bar{\pi}_T}(\rho) - V_{r_p}^{\bar{\pi}_T}(\rho)]}_{\text{Concentration Error}} + \underbrace{[V_{r_p}^{\bar{\pi}_T}(\rho) - V_r^{\bar{\pi}_T}(\rho)]}_{\text{Perturbation Error}}
\end{aligned}$$

For a perturbation magnitude equal to ω, we use Lemma 15 to bound both perturbation errors by $\frac{\omega}{1-\gamma}$. Using Theorem 1 to bound the primal-dual error by ε_{opt},

$$\leq \frac{2\omega}{1-\gamma} + \varepsilon_{\text{opt}} + \underbrace{[V_{r_p}^{\pi^*}(\rho) - \hat{V}_{r_p}^{\pi^*}(\rho)]}_{\text{Concentration Error}} + \underbrace{[\hat{V}_{r_p}^{\tilde{\pi}^*}(\rho) - \hat{V}_{r_p}^{\hat{\pi}^*}(\rho)]}_{\text{Sensitivity Error}} + \underbrace{[\hat{V}_{r_p}^{\bar{\pi}_T}(\rho) - V_{r_p}^{\bar{\pi}_T}(\rho)]}_{\text{Concentration Error}}$$

Since $b' = b + \Delta$ and setting $b'' = b - \Delta$, we use Lemma 13 to bound the sensitivity error term,

$$V_r^{\pi^*}(\rho) - V_r^{\bar{\pi}_T}(\rho) \leq \frac{2\omega}{1-\gamma} + \varepsilon_{\text{opt}} + 2\Delta\lambda^* + \underbrace{[V_{r_p}^{\pi^*}(\rho) - \hat{V}_{r_p}^{\pi^*}(\rho)]}_{\text{Concentration Error}} + \underbrace{[\hat{V}_{r_p}^{\bar{\pi}_T}(\rho) - V_{r_p}^{\bar{\pi}_T}(\rho)]}_{\text{Concentration Error}}$$

With these values of b' and b'', we require the following statements to hold,

$$\left|V_c^{\bar{\pi}_T}(\rho) - \hat{V}_c^{\bar{\pi}_T}(\rho)\right| \leq \Delta - \varepsilon_{\text{opt}} \quad ; \quad \left|V_c^{\pi^*}(\rho) - \hat{V}_c^{\pi^*}(\rho)\right| \leq \Delta.$$

□

Lemma 13 (Bounding the sensitivity error)**.** *If $b' = b + \Delta$ and $b'' = b - \Delta$ in Eq. (4) and Eq. (21) such that,*

$$\hat{\pi}^* \in \arg\max_{\pi} \hat{V}_{r_p}^{\pi}(\rho) \text{ s.t. } \hat{V}_c^{\pi}(\rho) \geq b + \Delta$$

$$\tilde{\pi}^* \in \arg\max_{\pi} \hat{V}_{r_p}^{\pi}(\rho) \text{ s.t. } \hat{V}_c^{\pi}(\rho) \geq b - \Delta\,,$$

then the sensitivity error term can be bounded by:

$$\left|\hat{V}_{r_p}^{\hat{\pi}^*}(\rho) - \hat{V}_{r_p}^{\tilde{\pi}^*}(\rho)\right| \leq 2\Delta\lambda^*.$$

Proof. Writing the empirical CMDP in Eq. (4) in its Lagrangian form,

$$\begin{aligned}\hat{V}_{r_p}^{\hat{\pi}^*}(\rho) &= \max_{\pi}\min_{\lambda\geq 0} \hat{V}_{r_p}^{\pi}(\rho) + \lambda[\hat{V}_c^{\pi}(\rho) - (b+\Delta)] \\ &= \min_{\lambda\geq 0}\max_{\pi} \hat{V}_{r_p}^{\pi}(\rho) + \lambda[\hat{V}_c^{\pi}(\rho) - (b+\Delta)]\end{aligned}$$

(By strong duality Lemma 9)

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

$$\begin{aligned}&= \max_{\pi} \hat{V}_{r_p}^{\pi}(\rho) + \lambda^*\left[\hat{V}_c^{\pi}(\rho) - (b+\Delta)\right] \\ &\geq \hat{V}_{r_p}^{\tilde{\pi}^*}(\rho) + \lambda^*\left[\hat{V}_c^{\tilde{\pi}^*}(\rho) - (b+\Delta)\right] \qquad \text{(The relation holds for } \pi = \tilde{\pi}^*.)\end{aligned}$$

Since $\hat{V}_c^{\tilde{\pi}^*}(\rho) \geq b - \Delta$,

$$\begin{aligned}&\hat{V}_{r_p}^{\hat{\pi}^*}(\rho) \geq \hat{V}_{r_p}^{\tilde{\pi}^*}(\rho) - 2\lambda^*\Delta \\ \implies &\hat{V}_{r_p}^{\tilde{\pi}^*}(\rho) - \hat{V}_{r_p}^{\hat{\pi}^*}(\rho) \leq 2\Delta\lambda^*\end{aligned}$$

Since the CMDP in Eq. (21) (with $b'' = b - \Delta$) is a less constrained problem than the one in Eq. (4) (with $b' = b + \Delta$), $\hat{V}_{r_p}^{\tilde{\pi}^*}(\rho) \geq \hat{V}_{r_p}^{\hat{\pi}^*}(\rho)$, and hence,

$$\left|\hat{V}_{r_p}^{\tilde{\pi}^*}(\rho) - \hat{V}_{r_p}^{\hat{\pi}^*}(\rho)\right| \leq 2\Delta\lambda^*.$$

□

D Concentration proofs

Lemma 5. *Define $\hat{\pi}^*_\alpha := \arg\max_\pi \hat{V}^\pi_\alpha$. If (i) $\mathcal{E}$ is the event that the ι-gap condition in Definition 4 holds for $\hat{M}_\alpha$ 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 $N \geq \frac{4\,C(\delta)}{1-\gamma}$, then with probability at least $\Pr[\mathcal{E}] - \delta/10$,*

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

Proof. Since the policy $\hat{\pi}^*_\alpha$ depends on the sampling, we can not directly apply the standard concentration results to bound $\left\|\hat{V}^{\hat{\pi}^*_\alpha}_\beta - \hat{V}^{\hat{\pi}^*_\alpha}_\beta\right\|_\infty$. We thus seek to apply a critical lemma established in Li et al. [21]. It begins with introducing a sequence of vectors for a general data-dependent policy π and reward $beta$, defined recursively as

$$V^{\pi,(0)}_\beta = (I-\gamma P^\pi)^{-1}\beta^\pi \quad \text{and} \quad V^{\pi,(l)}_\beta = (I-\gamma P^\pi)^{-1}\sqrt{\mathrm{Var}_{P_\pi}(V^{\pi,(l-1)}_\beta)},\ \forall l \geq 1.$$

In their Lemma 2 (restated below), they show that if certain concentration relation can be established between the empirical and ground truth MDP, then $\left\|\hat{V}^\pi_\beta - \hat{V}^\pi_\beta\right\|_\infty$ can be bounded.

Lemma 14 (Lemma 2 of [21])**.** *For a data-dependent policy π, suppose there exists a $\nu_1 \geq 0$ such that $\{V^{\pi,(l)}_\beta\}$ obeys*

$$\left|(\hat{\mathcal{P}}_\pi - \mathcal{P}_\pi)V^{\pi,(l)}_\beta\right| \leq \sqrt{\frac{\nu_1}{N}}\sqrt{\mathrm{Var}_{\mathcal{P}_\pi}[V^{\pi,(l)}_\beta]} + \frac{\nu_1\left\|V^{\pi,(l)}_\beta\right\|_\infty}{N},\ \textit{for all } 0 \leq l \leq \log\left(\frac{e}{1-\gamma}\right).$$

Suppose that $N \geq \frac{16e^2}{1-\gamma}\nu_1$. Then

$$\left\|\hat{V}^\pi_\beta - V^\pi_\beta\right\|_\infty \leq \frac{6}{1-\gamma}\cdot\sqrt{\frac{\nu_1}{N(1-\gamma)}}\,\|\beta\|_\infty\,.$$

To use this lemma for $\pi = \hat{\pi}^*_\alpha$, we will need to establish Bernstein-type bounds on $(\mathcal{P}_{s,a} - \hat{\mathcal{P}}_{s,a})V^{\hat{\pi}^*_\alpha,(l)}_\beta$ for all (s,a) and integer $0 \leq l \leq \log(e/(1-\gamma))$. Since $\hat{\pi}^*_\alpha$ depends on the sampling, a direct concentration bound is not possible. Instead, we will first bound $(\mathcal{P}_{s,a} - \hat{\mathcal{P}}_{s,a})V^{\pi,(l)}_\beta$ for all $\pi \in \Pi_{s,a}$, where $\Pi_{s,a}$ is a random set independent of $\hat{\mathcal{P}}$, and then show that $\hat{\pi}^* \in \Pi_{s,a}$ with good probability.

First, we describe the construction of $\Pi_{s,a}$. We will follow the ideas in Agarwal et al. [2] and Li et al. [21], and construct an absorbing empirical MDP $\hat{M}_{s,a}$, which is the same as the original empirical MDP, but state-action pair (s,a) is absorbing, i.e., $\hat{\mathcal{P}}'(s'|s,a) = 1$ if and only if $s' = s$. The reward for (s,a) is equal to u. We define $\hat{V}^\pi_{s,a,\alpha,u}$ and $\hat{Q}^\pi_{s,a,\alpha,u}$ to be the value function and Q-function of policy π for $\hat{M}_{s,a}$ with reward function α, and define $\hat{\pi}^*_{s,a,\alpha,u}$ to be the optimal policy i.e. $\hat{\pi}^*_{s,a,\alpha,u} = \arg\max \hat{V}^\pi_{s,a,\alpha,u}$. We will use the shorthand – $\hat{V}^*_{\alpha,u} := \max \hat{V}^\pi_{s,a,\alpha,u}$ and $\hat{Q}^*_{\alpha,u} := \max \hat{Q}^\pi_{s,a,\alpha,u}$.

We consider a grid,

$$U_{s,a} = \{0, \pm\iota(1-\gamma)/2, \pm 2\iota(1-\gamma)/2, \pm 3\iota(1-\gamma)/2\ldots, \pm\alpha_{\max}\},$$

and define $\Pi_{s,a} = \{\hat{\pi}^*_{s,a,\alpha,u} : u \in U_{s,a}\}$. Then $\Pi_{s,a}$ is a random set independent of $\hat{\mathcal{P}}_{s,a}$. Let $L = \{0, 1, \ldots, \lceil\log(e/(1-\gamma))\rceil\}$. Then, by Lemma 19, we have, with probability at least $1 - \delta/|S|/|A|$, for all $\pi \in \Pi_{s,a}$ and $l \in L$

$$\left|(\mathcal{P}_{s,a} - \hat{\mathcal{P}}_{s,a})\cdot V^{\pi,(l)}_\beta\right| \leq \sqrt{\frac{2\log(4|U_{s,a}||L||S||A|/\delta)}{N}}\sqrt{\mathrm{Var}_{P_{s,a}}(V^{\pi,(l)}_\beta)} + \frac{2\log(4|U_{s,a}||S||A||L|/\delta)\left\|V^{\pi,(l)}_\beta\right\|_\infty}{3N},$$

which we denote as event $\mathcal{E}_{s,a}$.

Next, we show that if $u^* = \hat{Q}^*_\alpha(s,a) - \gamma \hat{V}^*_\alpha(s)$, then $\hat{\pi}^*_{s,a,\alpha,u^*} = \hat{\pi}^*_\alpha$:
(1) If $\pi^*(s) = a$, it straightforward to verify that

$$\hat{V}^*_\alpha(s) = \hat{Q}^*_\alpha(s,a) = u^* + \gamma \hat{V}^*_\alpha(s) \geq r(s,a') + \hat{\mathcal{P}}(\cdot|s',a')^\top \hat{V}^*_\alpha, \quad \forall a' \neq a.$$

(2) If $\pi^*(s) \neq a$, then

$$\hat{V}^*_\alpha(s) = \max_{a'} \hat{Q}^*_\alpha(s,a') = \max_{a'} (r(s,a') + \hat{\mathcal{P}}(\cdot|s,a')^\top \hat{V}^*_\alpha) = r(s,\hat{\pi}^*(s)) + \hat{\mathcal{P}}(\cdot|s,\hat{\pi}^*(s))^\top \hat{V}^*_\alpha.$$

(3) For $s' \neq s$, we have

$$\hat{V}^*_\alpha(s') = \hat{Q}^*_\alpha(s',\hat{\pi}^*(s')) = r(s',\hat{\pi}^*(s')) + \hat{\mathcal{P}}(\cdot|s',\hat{\pi}^*(s'))^\top \hat{V}^*_\alpha = \max_{a'} (r(s,a') + \hat{\mathcal{P}}(\cdot|s',a')^\top \hat{V}^*_\alpha).$$

Therefore, $\hat{Q}^*_\alpha(s',a')$ and $\hat{\pi}^*_\alpha$ satisfies the Bellman equations in the absorbing MDP; consequently, we have $\hat{Q}^{\hat{\pi}^*_\alpha}_{s,a,\alpha,u} = \hat{Q}^*_{\alpha,u} = \hat{Q}^*_\alpha$ and $\hat{V}^{\hat{\pi}^*_\alpha}_{s,a,\alpha,u} = \hat{V}^*_{\alpha,u} = \hat{V}^{\hat{\pi}^*_\alpha}_\alpha$ and $\hat{\pi}^*_\alpha$ is an optimal policy in the absorbing MDP.

Moreover, suppose event $\mathcal{E}$ happens, then $\hat{Q}^*_\alpha$ satisfies the ι-gap condition. By Lemma 17, for all $|u - u^*| \leq \iota(1-\gamma)/2$, we have

$$\hat{\pi}^*_\alpha = \hat{\pi}^*_{s,a,\alpha,u^*} = \hat{\pi}^*_{s,a,\alpha,u}.$$

Thus, if $\mathcal{E}$ happens, then $\hat{\pi}^*_\alpha = \hat{\pi}^*_{s,a,\alpha,u}$ for some $u \in U_{s,a}$ and thus $\hat{\pi}^* \in \Pi_{s,a}$. On $\mathcal{E} \cap \mathcal{E}_{s,a}$, we have, for all $l \in L$,

$$\left|(\mathcal{P}_{s,a} - \hat{\mathcal{P}}_{s,a}) \cdot V^{\hat{\pi}^*_\alpha,(l)}_\beta\right| \leq \sqrt{\frac{2\log(4|U_{s,a}||L||S||A|/\delta)}{N}} \sqrt{\mathrm{Var}_{P_{s,a}}(V^{\hat{\pi}^*_\alpha,(l)}_\beta)} + \frac{2\log(4|U_{s,a}||S||A||L|/\delta)\left\|V^{\hat{\pi}^*_\alpha,(l)}_\beta\right\|_\infty}{3N}.$$

By a union bound over all (s,a), we have, with probability at least $\Pr[\mathcal{E} \cap (\cap_{s,a}\mathcal{E}_{s,a})] \geq \Pr[\mathcal{E}] - \delta$, for all (s,a), and $l \in L$,

$$\left|(\mathcal{P}_{s,a} - \hat{\mathcal{P}}_{s,a}) \cdot V^{\hat{\pi}^*_\alpha,(l)}_\beta\right| \leq \sqrt{\frac{2\log(4|U_{s,a}||L||S||A|/\delta)}{N}} \sqrt{\mathrm{Var}_{P_{s,a}}(V^{\hat{\pi}^*_\alpha,(l)}_\beta)} + \frac{2\log(4|U_{s,a}||S||A||L|/\delta)\left\|V^{\hat{\pi}^*_\alpha,(l)}_\beta\right\|_\infty}{3N}.$$

Let $\nu_1 = 2\log(4|U_{s,a}||S||A||L|/\delta)$ and apply Lemma 14, we arrive at,

$$\left\|\hat{V}^{\hat{\pi}^*}_\beta - V^{\hat{\pi}^*}_\beta\right\|_\infty \leq \frac{6}{1-\gamma} \cdot \sqrt{\frac{\nu_1}{N(1-\gamma)}} \|\beta\|_\infty$$

provided $N \geq \frac{16e^2}{1-\gamma}\nu_1$. For instantiating ν_1, note that $|U_{s,a}| = \frac{4\alpha_{\max}}{(1-\gamma)^2\iota}$, $|L| = \log(e/1-\gamma)$. Hence, $\nu_1 = 2\log\left(\frac{16\alpha_{\max} SA \log(e/1-\gamma)}{(1-\gamma)^2\,\iota\,\delta}\right)$.

□

Lemma 15. *For any policy π, we have*

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

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

$$V^\pi_r(\rho) - V^\pi_{r_p}(\rho) = (I - \gamma P_\pi)^{-1}[r^\pi - r^\pi_p]$$

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

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

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

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

Theorem 6. 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 $\iota = \frac{\omega\, \delta\, (1-\gamma)\, \varepsilon_1}{30\, U|S||A|^2}$, if $N \geq \frac{4\, C(\delta)}{1-\gamma}$, then for $\bar{\pi}_T$ output by Algorithm 1, with probability at least $1-\delta/5$,

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

Proof.

$$\left|V_{r_p}^{\bar{\pi}_T}(\rho) - \hat{V}_{r_p}^{\bar{\pi}_T}(\rho)\right| = \left|\frac{1}{T}\sum_{t=0}^{T-1}\left[V_{r_p}^{\pi_t}(\rho) - \hat{V}_{r_p}^{\pi_t}(\rho)\right]\right| \leq \frac{1}{T}\sum_{t=0}^{T-1}\left|V_{r_p}^{\pi_t}(\rho) - \hat{V}_{r_p}^{\pi_t}(\rho)\right|$$
$$\leq \frac{1}{T}\sum_{t=0}^{T-1}\left\|V_{r_p}^{\pi_t} - \hat{V}_{r_p}^{\pi_t}\right\|_\infty$$

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

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

Using the above relations,

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

Similarly, invoking Lemma 5 with $\alpha = r_p + \lambda_t c$ and $\beta = c$ gives the bound on $\left|V_{c}^{\bar{\pi}_T}(\rho) - \hat{V}_{c}^{\bar{\pi}_T}(\rho)\right|$. □

Lemma 7. *For $\delta \in (0,1)$, $\omega \leq 1$ and $C'(\delta) = 72 \log\left(\frac{4|S| \log(e/1-\gamma)}{\delta}\right)$, if $N \geq \frac{4\, C'(\delta)}{1-\gamma}$ and $B(\delta, N) := \sqrt{\frac{C'(\delta)}{(1-\gamma)^3 N}}$, then with probability at least $1-3\delta$,*

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

Proof. Since π^* and π_c^* are fixed policies independent of the sampling, we can directly use Li et al. [21, Lemma 1]. □

D.1 Helper lemmas

Lemma 16. *With probability at least $1-\delta/10$, for every $\lambda \in \Lambda$, $\hat{M}_{r_p+\lambda c}$ satisfies the gap condition in Definition 4 with $\iota = \frac{\omega\, \delta\, (1-\gamma)}{30\, |\Lambda||S||A|^2}$.*

Proof. Using Lemma 18 with a union bound over Λ gives the desired result. □

Lemma 17. *Let π_α^* and $\pi_{\alpha'}^*$ be two optimal policies for an MDP with rewards α and α' respectively. Suppose Q_α^* satisfies the ι-gap condition. Then, for all α' with $\|\alpha' - \alpha\|_\infty < \iota(1-\gamma)/2$, we have*

$$\pi_\alpha^* = \pi_{\alpha'}^*.$$

Proof. Since $\|\alpha' - \alpha\|_\infty < \iota(1-\gamma)/2$, we thus have

$$\|Q^*_\alpha - Q^*_{\alpha'}\|_\infty < \frac{\iota(1-\gamma)}{2(1-\gamma)} = \frac{\iota}{2}.$$

Note that, for all s,

$$\begin{aligned} Q^*_\alpha(s, \pi^*_{\alpha'}(s)) &> Q^*_{\alpha'}(s, \pi^*_{\alpha'}(s)) - \frac{\iota}{2} \\ &\geq Q^*_{\alpha'}(s, \pi^*_\alpha(s)) - \frac{\iota}{2}, \\ &> Q^*_\alpha(s, \pi^*_\alpha(s)) - \iota \\ &> Q^*_\alpha(s, a'). \quad \forall a' \neq \pi^*_\alpha(s). \end{aligned}$$

Hence, $\pi^*_{\alpha'}(s) \notin \{a' : a' \neq \pi^*_\alpha(s)\}$ for all $s \in S$, and consequently $\pi^*_\alpha = \pi^*_{\alpha'}$. □

Lemma 18 (Lemma 6 of Li et al. [21])**.** *Consider the MDP $M = (S, A, P, \gamma, r_p)$ with randomly perturbed rewards r_p ($r_p(s,a) = r(s,a) + \xi(s,a)$ where $\xi(s,a) \sim \mathcal{U}[0, \omega]$ and $r(s,a) \in \mathbb{R}$). If optimal Q-function is denoted as $Q^*_{r_p}$ and $\pi^*_{r_p} := \arg\max V^\pi_{r_p}$ is the optimal deterministic policy, then with probability at least $1-\delta$, for some $\delta \in (0,1]$, we have, for all s and $a \neq \pi^*_{r'}(s)$*

$$Q^*_{r_p}(s, \pi^*_{r_p}(s)) \geq Q^*_{r_p}(s,a) + \frac{\omega\delta(1-\gamma)}{3|S||A|^2}$$

i.e. M satisfies the gap condition in Definition 4 with $\iota = \frac{\omega\delta(1-\gamma)}{3|S||A|^2}$.

Lemma 19 (Bernstein inequality)**.** *Fix a state s, an action a. Let $\delta \geq 0$. Then, for any fixed vector V, with probability greater than $1-\delta$, it holds that,*

$$\left|(\mathcal{P}_{s,a} - \hat{\mathcal{P}}_{s,a}) \cdot V\right| \leq \sqrt{\frac{2\log(4/\delta)}{N}} \sqrt{Var_{P_{s,a}}(V)} + \frac{2\log(4/\delta)\,\|V\|_\infty}{3N}.$$

E Lower Bound

In this section, we first present the lower-bound in the simplified bandit setting (Appendix E.1), and then present the formal CMDP lower bound in Appendix E.2.

E.1 Bandit Setting

Consider the 2-arm bandit setting where arm-1 has a mean reward H and mean constraint reward $b - \zeta$ and arm-2 has a mean reward 0 and mean constraint reward $b + \zeta$. For the constraint RHS equal to b (implying that the Slater constant is ζ), the ground truth optimal policy plays each arm with equal probability to achieve a reward value $H/2$ (with no constraint violation). However, suppose there is an error ε' in the estimation of the constraint reward in arm-2 (we estimate it to be $b + \zeta - \varepsilon'$); then even if everything else is exactly estimated, the empirical optimal policy has to play action 2 with probability $\zeta/(2\zeta - \varepsilon') \approx 1/2 + \varepsilon'/(4\zeta)$ to satisfy the constraint, which gives a value $H/2 - H\varepsilon'/(4\zeta)$. This results in an $H\varepsilon'/(4\zeta)$ error of the final policy. To obtain an ε-correct policy without constraint violation, one needs to set $\varepsilon' = \varepsilon\zeta/H$.
Combining this with the lower-bound results for unconstrained MDPs, this implies that we require a sample-complexity of $\Omega(H^3SA/\varepsilon'^2) = \Omega(H^5SA/\varepsilon^2\zeta^2)$. This intuition is formalized for CMDPs in the lower-bound construction below.

E.2 Proof of Theorem 8

Theorem 8. There exists constants $\gamma_0 \in (1 - 1/\log(|\mathcal{S}|), 1)$, $0 \le \varepsilon_0 \le \frac{1}{(1-\gamma)} \min\left\{1, \frac{\gamma}{(1-\gamma)\zeta}\right\}$, $\delta_0 \in (0, 1)$, such that, for any $\gamma \in (\gamma_0, 1), \varepsilon \in (0, \varepsilon_0), \delta \in (0, \delta_0)$, any (ε, δ)-sound algorithm requires $\Omega\left(\frac{SA\ln(1/4\delta)}{\varepsilon^2\zeta^2(1-\gamma)^5}\right)$ samples from the generative model in the worst case.

Proof of Theorem 8. Without loss of generality, we assume $|\mathcal{S}| = 2^m - 1$ for some integer m. In what follows, we first introduce our hard instance. Note that some of the states of this instance may have less than $|\mathcal{A}|$ actions. This is without loss of generality as one can easily duplicate actions to make each state having exactly $|\mathcal{A}|$ actions.

Hard Instance. We will consider the basic gadget defined in Figure 2. We will make $|\mathcal{S}| + 1$ copies of this gadget. In the i-th gadet for $i = 0, 1, \ldots, |\mathcal{S}|$, there is an input state s_i with 2 actions a_l, a_r. Playing a_l deterministically transitions to $\tilde{s}_i$ with reward and constraint 0. Playing a_r goes to z'_i with reward and constraint 0. In state $\tilde{s}_i$, there are $|\mathcal{A}|$ actions (only 1 action a_0 on $\tilde{s}_0$). Playing action $a \in \mathcal{A}$, this state transitions to state z_i with probability $1 - p_{i,a}$ and self loop with probability $p_{i,a}$, where $p_{i,a}$ will be determined shortly. The reward at this state is 1 and constraint reward is u to be determined. In state z'_i, there is only one action, whose reward is 0 and constraint is $(b + \zeta)(1 - \gamma)/\gamma^{m+3}$.

Lastly, we consider $2|\mathcal{S}| + 1$ routing state $o_0, o_1, \ldots$ that form a binary tree, i.e., in each of these routing state, there are two actions a_0 and a_1. The state-action pair (o_i, a_j) transitions to state o_{2i+j+1} for $i < |\mathcal{S}|$. Each state $o_{|\mathcal{S}|+k}$ transitions deterministically to the gate state s_k for $k = 0, 1, \ldots |\mathcal{S}|$. Rewards and constraint are all 0 for these states. Note that, for any state s_k, there is a unique deterministic path of length $m + 1$ connecting o_0. Hence we require $\gamma \ge 1 - 1/(cm)$ for some constant c and hence $\gamma^{\Theta(m)} = \Theta(1)$.

Note that this instance modifies the MDP instance in [14]. Some parameters chosen are also adapted from there.

Hypotheses. With the above defined hard instance structure, we now define a family of hypotheses on the probability transitions. Later we will show that any sound algorithm would be able to test the hypothesises but would require a large number of samples. Let $q_0, q_1, q_2 \in (0, 1)$ be some parameters to be determined. We define,

- *Null case* MDP M_0: $p_{0,a_0} = q_1$, and $p_{i,a} = q_0$ for all $i \in [|\mathcal{S}|]$ and $a \in \mathcal{A}$.
- *Alternative case* MDP $M_{i,a}$: $p_{0,a_0} = q_1$, $p_{j,a'} = q_0$ for all $(j, a') \neq (i, a)$, and $p_{i,a} = q_2$.

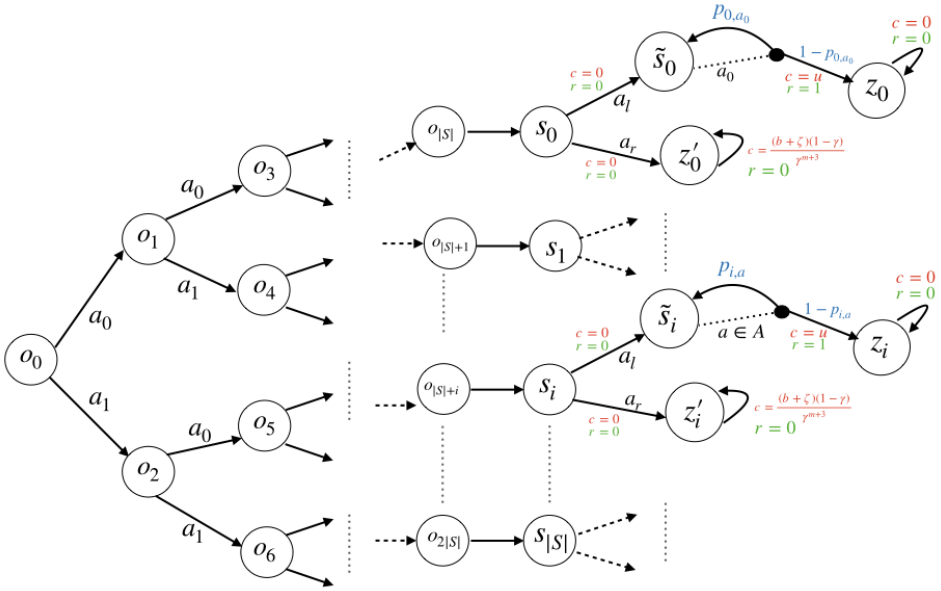

Figure 2: The lower-bound instance consists of CMDPs with $S = 2^m - 1$ (for some integer $m > 0$) states and A actions. We consider $SA + 1$ CMDPs – M_0 and $M_{i,a}$ ($i \in \{1, \dots S\}$, $a \in \{1, \dots, A\}$) that share the same structure shown in the figure. For each CMDP, o_0 is the fixed starting state and there is a deterministic path of length $m + 1$ from o_0 to each of the $S + 1$ states – s_i (for $i \in \{0, 1, \dots, S\}$). Except for states $\tilde{s}_i$, the transitions in all other states are deterministic. For $i \neq 0$, for action $a \in \mathcal{A}$ in state $\tilde{s}_i$, the probability of staying in $\tilde{s}_i$ is $p_{i,a}$, while that of transitioning to state z_i is $1 - p_{i,a}$. There is only one action a_0 in $\tilde{s}_0$ and the probability of staying in $\tilde{s}_0$ is p_{0,a_0}, while that of transitioning to state z_0 is $1 - p_{0,a_0}$. The CMDPs M_0 and $M_{i,a}$ only differ in the values of $p_{i,a}$. The rewards r and constraint rewards c are the same in all CMDPs and are denoted in green and red respectively for each state and action.

Note that all these MDPs $M_0, M_1 \dots$ share exactly the same graph structure, with only the transition probability changes. Moreover, $M_{i,a}$ differs from M_0 only on state-action pair $(\tilde{s}_i, a)$.

Optimal Policies. Now we specify q_0, q_1 and q_2 and check the optimal policies of each hypothesis. Let $\varepsilon' = (1 - \gamma)\zeta\varepsilon$,

$$q_0 = \frac{1 - c_1(1 - \gamma)}{\gamma}, q_1 = q_0 + \alpha_1, \text{ and } q_2 = q_0 + \alpha_2,$$

and

$$\alpha_1 = \frac{c_2(1 - \gamma q_0)^2 \varepsilon'}{\gamma}, \quad \text{and} \quad \alpha_2 = \frac{c_3(1 - \gamma q_0)^2 \varepsilon'}{\gamma}$$

for some absolute constants $c_1, c_2, c_3 > 0$ and that $\alpha_1/q_0 \in (0, 1/2)$, $\alpha_1/(1 - q_0) \in (0, 1/2)$, $\alpha_2/q_0 \in (0, 1/2)$, and $\alpha_2/(1 - q_0) \in (0, 1/2)$. We choose these parameters such that

$$\frac{1}{c_1(1 - \gamma)} = \frac{1}{1 - \gamma q_0} < \frac{1}{1 - \gamma q_1} < \frac{1}{1 - \gamma q_2} = \frac{1}{c_1(1 - \gamma)} + c_4 \varepsilon'$$

for some constants $c_4 > 0$ and that

$$\left| \frac{1}{1 - \gamma q_1} - \frac{1}{1 - \gamma q_0} \right| = \Theta(\varepsilon'), \text{ and } \left| \frac{1}{1 - \gamma q_1} - \frac{1}{1 - \gamma q_2} \right| = \Theta(\varepsilon').$$

Note that, for the reward values, if $\zeta = \Theta(1)$, then these actions only differ by $\Theta(\varepsilon') \ll \varepsilon$. A correct algorithm would not need to distinguish these actions. Yet, once the constraints are concerns, we will show shortly that these actions do differ because of the constraint values.

With these parameters, we can derive the optimal CMDP policy for M_0 and each $M_{i,a}$. For any policy π, we denote the state occupancy measure as μ^π, i.e.,

$$\mu^\pi(s,a) = \sum_z \rho(z) \sum_{t=1}^{\infty} \gamma^{t-1} \sum_z \Pr_\pi[s_t = s, a_t = a | s_1 = z] \tag{22}$$

where ρ is the initial distribution with $\rho(o_0) = 1$ and $\rho(s) = 0$ for all $s \neq o_0$. This occupancy measure describes the discounted reachablity from o_0 to an arbitrary state action pair. The reward value and constraint value can be written as

$$V^\pi = \sum_{s,a} \mu^\pi(s,a) r(s,a), \quad \text{and} \quad V_c^\pi = \sum_{s,a} \mu^\pi(s,a) c(s,a).$$

Note that, given a state-occupancy measure μ, a policy can be specified as $\pi_\mu(a|s) = \mu(s,a)/\sum_{a'} \mu(s,a')$. We can use the LP formulation for the CMDP as follows

$$\begin{aligned}
&\max \sum_{s,a} \mu(s,a) r(s,a) \text{ subject to ,} \\
&\forall s: \ \sum_a \mu(s,a) = \rho(s) + \gamma \sum_{s',a'} P(s|s',a') \mu(s',a'), \\
&\sum_{s,a} \mu^\pi(s,a) c(s,a) \geq b, \\
&\mu(s,a) \geq 0.
\end{aligned} \tag{23}$$

Due to the structure of the CMDPs, we can further simply the structure of the constraints. In particular, we have

$$\forall i: \quad \sum_a \mu(\tilde{s}_i, a) = \gamma \mu(s_i, a_l) + \gamma \sum_a p_{i,a} \mu(\tilde{s}_i, a) \tag{24}$$

let $\bar{\nu} = \sum_i \mu(s_i, a_r)$. We then have

$$\sum_i \mu(s_i, a_l) + \bar{\nu} = \gamma^{m+2},$$

and

$$\mu(z_i', a) = \frac{\gamma \mu(s_i, a_r)}{1-\gamma}.$$

Consider M_0, let $\mu_0 := \mu(s_0, a_l)$, $\mu_0^c := \sum_{i>0} \mu(s_i, a_l)$ we have

$$\sum_{i>0,a} \mu(\tilde{s}_i, a) = \frac{\gamma \mu_0^c}{1-\gamma q_0}, \quad \text{and} \quad \mu(\tilde{s}_0, a_0) = \frac{\gamma \mu_0}{1-\gamma q_1}.$$

The LP can be rewritten as

$$\begin{aligned}
\max \quad & \frac{\gamma \mu_0}{1-\gamma q_1} + \frac{\gamma \mu_0^c}{1-\gamma q_0} \\
\text{s.t.} \ & \frac{\gamma \mu_0 \cdot u}{1-\gamma q_1} + \frac{\gamma \mu_0^c \cdot u}{1-\gamma q_0} + \frac{\bar{\nu} \cdot (b+\zeta)}{\gamma^{m+2}} \geq b, \ \mu_0 + \mu_0^c + \bar{\nu} = \gamma^{m+2}, \text{ and } \mu(s,a) \geq 0, \ \forall (s,a).
\end{aligned}$$

Here, we specify the values of u as,

$$u = \frac{(1-\gamma q_0)(b-x)}{\gamma^{m+3}},$$

for some $x = \Theta(\zeta)$, such that,

$$\frac{\gamma^{m+3} u}{1-\gamma q_0} = b - x < \frac{\gamma^{m+3} u}{1-\gamma q_1} = b - x + \varepsilon_1' < \frac{\gamma^{m+3} u}{1-\gamma q_2} = b - x + \varepsilon_2' < b,$$

where $c'\varepsilon' \leq \varepsilon_1' \leq \varepsilon_1' + c''\varepsilon' \leq \varepsilon_2' \leq c'''\varepsilon'$ for some positive constants c', c'', c''' determined by c_1, c_2, c_3. For this value of u, the maximum value of the constraint value function $\max V_c^\pi$ is $b + \zeta$, implying that ζ is the Slater constant for all these CMDPs.

Thus, for M_0, the solution is,

$$\mu_0 = \frac{\gamma^{m+2}\zeta}{\zeta + x - \varepsilon_1'}, \quad \mu_0^c = 0, \quad \text{and} \quad \bar{\nu} = \frac{(x - \varepsilon_1')\gamma^{m+2}}{\zeta + x - \varepsilon_1'}.$$

Note that this implies the policy deterministically choose a path to reach state s_0, and then plays action a_l with probability μ_0/γ^{m+2}. The optimal value in this case is

$$V^*_{M_0}(o_0) = \frac{\zeta}{\zeta + x - \varepsilon_1'} \cdot \frac{\gamma^{m+3}}{1 - \gamma q_1}.$$

Similarly, for $M_{i,a}$ with $i \geq 1$, let $\mu_{i,a} := \mu(s_i, a)$, $\mu^c_{i,a} := \sum_{i'>0,(i',a')\neq(i,a)} \mu(s_{i'}, a')$, the LP can be written as

$$\begin{aligned}
&\max \quad \frac{\gamma\mu_0}{1-\gamma q_1} + \frac{\gamma\mu^c_{i,a}}{1-\gamma q_0} + \frac{\gamma\mu_{i,a}}{1-\gamma q_2} \\
&\text{s.t. } \frac{\gamma\mu_0 u}{1-\gamma q_1} + \frac{\gamma\mu^c_{i,a} u}{1-\gamma q_0} + \frac{\gamma\mu_{i,a} u}{1-\gamma q_2} + \frac{\bar{\nu}\cdot(b+\zeta)}{\gamma^{m+2}} \geq b,\ \mu_0 + \mu^c_{i,a} + \mu_{i,a} + \bar{\nu} = \gamma^{m+2}, \\
&\text{and } \mu(s,a) \geq 0,\ \forall (s,a).
\end{aligned}$$

the solution is

$$\mu_{i,a} = \frac{\zeta\gamma^{m+2}}{\zeta + x - \varepsilon_2'},\ \mu^c_{i,a} = \mu_0 = 0, \quad \text{and} \quad \bar{\nu} = \frac{(x - \varepsilon_2')\gamma^{m+2}}{\zeta + x - \varepsilon_2'},$$

i.e., the policy chooses a path to reach state s_i deterministically and with optimal value

$$V^*_{M_{i,a}}(o_0) = \frac{\zeta}{\zeta + x - \varepsilon_2'} \cdot \frac{\gamma^{m+3}}{1 - \gamma q_2}.$$

Lastly, we shall check the gap of the value functions.

$$\left| V^*_{M_{i,a}}(o_0) - V^*_{M_0}(o_0) \right| \geq \left(\frac{\zeta}{\zeta + x - \varepsilon_2'} - \frac{\zeta}{\zeta + x - \varepsilon_1'} \right) \cdot \frac{\gamma^{m+3}}{1 - \gamma q_1} = \frac{\varepsilon''}{\zeta + x} \cdot \frac{\gamma^{m+3}}{1 - \gamma q_1} \geq c_7 \varepsilon.$$

where $\varepsilon'' \geq c_8 \varepsilon'$ for some constants c_7, c_8 determined by c_1, c_2, c_3 and x. Thus the error in V^π_c is amplified by a factor of $\Theta[(1-\gamma)^{-1}\zeta^{-1}]$.

Implications of Soundness: Near-Optimal Policies. Let $\mathcal{K}$ be a (ε, δ)-sound algorithm, i.e., on input any CMDP with a generative model, it outputs a policy, which is ε-optimal with probability at least $1 - \delta$. We thus define the event

$$\mathcal{E}_0 = \left\{ \mathcal{K} \text{ outputs policy } \pi \text{ such that } \mu^\pi(s_0, a_l) \geq \frac{\zeta\gamma^{m+2}}{(\zeta + x - \varepsilon_1'/2)} \right\},$$

i.e., this event requires the output policy reaching s_0 and play action a_l with sufficiently high probability. We now measure the probability of $\mathcal{E}_0$ on different input CMDPs. Due to the soundness, it must be the case that

$$\Pr_{M_{i,a}}[\mathcal{E}_0] < \delta.$$

If not, on $\mathcal{E}_0$, $\mu^\pi(s_0, a_l) \geq \frac{\zeta\gamma^{m+2}}{\zeta + x - \varepsilon_1'/2}$, we can then compute the best possible $V^\pi_{M_{i,a}}(o_0)$ as solving the following LP,

$$\begin{aligned}
&\max \quad \frac{\gamma\mu_0}{1-\gamma q_1} + \frac{\gamma\mu^c_{i,a}}{1-\gamma q_0} + \frac{\gamma\mu_{i,a}}{1-\gamma q_2} \\
&\text{s.t. } \frac{\gamma\mu_0 u}{1-\gamma q_1} + \frac{\gamma\mu^c_{i,a} u}{1-\gamma q_0} + \frac{\gamma\mu_{i,a} u}{1-\gamma q_2} + \frac{\bar{\nu}\cdot(b+\zeta)}{\gamma^{m+2}} \geq b,\ \mu_0 + \mu^c_{i,a} + \mu_{i,a} + \bar{\nu} = \gamma^{m+2}, \\
&\mu_0 \geq \frac{\zeta\gamma^{m+2}}{(\zeta + x - \varepsilon_1'/2)} \text{ and } \mu(s,a) \geq 0,\ \forall (s,a).
\end{aligned}$$

Plugging the values of $p_{i,a}$, and due to $q_0 < q_1 < q_2$, we obtain $\mu_0 = \frac{\zeta\gamma^{m+2}}{\zeta+x-\varepsilon_1'/2}$, $\mu_{i,a}^c = 0$ and,

$$\mu_{i,a} \cdot (b - x + \varepsilon_2') + \mu_0 \cdot (b - x + \varepsilon_1') + \bar{\nu}(b + \zeta) = b\gamma^{m+2}, \mu_{i,a} + \mu_0 + \bar{\nu} = \gamma^{m+2}.$$

Hence,

$$\mu_{i,a} = \gamma^{m+2} \cdot \frac{\zeta - \mu_0(\zeta + x - \varepsilon_1')}{\zeta + x - \varepsilon_2'} \leq \frac{c_9\varepsilon_1'\zeta \cdot \gamma^{m+2}}{2(\zeta + x - \varepsilon_1'/2)(\zeta + x - \varepsilon_2)}$$

for some constant c_9 depending on c_1, c_2, c_3, x, and

$$V^\pi(o_0) \leq \frac{\zeta}{\zeta + x - \varepsilon_1'/2} \cdot \frac{\gamma^{m+3}}{1 - \gamma q_1} + \frac{c_9\varepsilon_1'\zeta}{2(\zeta + x - \varepsilon_1'/2)(\zeta + x - \varepsilon_2)} \cdot \frac{\gamma^{m+3}}{1 - \gamma q_2}$$

Note that

$$V^*_{M_{i,a}}(o_0) = \frac{\zeta}{\zeta + x - \varepsilon_2'} \cdot \frac{\gamma^{m+3}}{1 - \gamma q_2}$$

and

$$V^*_{M_{i,a}}(o_0) - V^\pi_{M_{i,a}}(o_0) \geq \left(\frac{\zeta}{\zeta + x - \varepsilon_2'} - \frac{\zeta}{\zeta + x - \varepsilon_1'/2} - \frac{c_9\varepsilon_1'\zeta}{2(\zeta + x - \varepsilon_1'/2)(\zeta + x - \varepsilon_2')}\right) \cdot \frac{\gamma^{m+3}}{1 - \gamma q_2} \geq \varepsilon$$

for some appropriately chosen c_1, c_2, c_3, x, which is a contradiction of the (ε, δ)-soundness.

Implications of Soundness: Expectation on Null Hypothesis. Let $N_{i,a}$ be the number of samples the algorithm $\mathcal{K}$ takes on state-action (s_i, a). Next we show that, $\mathbb{E}[N_{i,a}]$ has to be big on M_0.

Lemma 20. *Let $t_* = \frac{c_{10}\log\delta^{-1}}{(1-\gamma)^3\varepsilon'^2}$ for some constant c_{10}. For any (ε, δ)-sound algorithm $\mathcal{K}$, for any (i, a), we have*

$$\mathbb{E}_{M_0}(N_{i,a}) \geq t_*.$$

Proof. Suppose $\mathbb{E}_{M_0}(N_{i,a}) < t_*$, then we aim to show a contradiction: $\Pr_{M_{i,a}}[\mathcal{E}_0] \geq \delta$. Similar to the proof above, since $\mathcal{K}$ is (ε, δ)-sound, it must be the case that

$$\Pr_{M_0}[\mathcal{E}_0] \geq 1 - \delta.$$

We now consider the likelihood ratio

$$\Pr_{M_{i,a}}[\mathcal{E}_0] / \Pr_{M_0}[\mathcal{E}_0].$$

For any realization of the empirical samples, consider the samples the algorithm takes as $\tau = \{(s_{i,a}^{(1)}, s_{i,a}^{(2)}, \ldots, s_{i,a}^{(N_{i,a})}) : (i, a) \in [|\mathcal{S}|] \times [\mathcal{A}]\}$. Let us define $N_{s',s,a}$ as the number of samples from $(s, a) \to s'$. By Markov property, since the only difference of the probability matrix between $M_{i,a}$ and M_0 is on $p_{i,a}$, we have

$$\begin{aligned}\frac{\Pr_{M_{i,a}}[\tau]}{\Pr_{M_0}[\tau]} &= \frac{q_2^{N_{\tilde{s}_i,\tilde{s}_i,a}}(1 - q_2)^{N_{z_i,\tilde{s}_i,a}}}{q_0^{N_{\tilde{s}_i,\tilde{s}_i,a}}(1 - q_0)^{N_{z_i,\tilde{s}_i,a}}} = \left(\frac{q_2}{q_0}\right)^{N_{\tilde{s}_i,\tilde{s}_i,a}} \cdot \left(\frac{1 - q_2}{1 - q_0}\right)^{N_{z_i,\tilde{s}_i,a}} \\ &= \left(1 + \frac{\alpha_2}{q_0}\right)^{Nq_0 - \Delta} \cdot \left(1 - \frac{\alpha_2}{1 - q_0}\right)^{N(1-q_0)+\Delta}\end{aligned}$$

where $\Pr_M[\tau]$ denotes the probability of $\mathcal{K}$ taking the samples τ in CMDP M, $\Delta = N_{i,a}q_0 - N_{\tilde{s}_i,\tilde{s}_i,a}$, and $N = N_{i,a}$. By a similar derivation of Lemma 5 in [14] (page 15-19), on the following event,

$$\mathcal{E}_{i,a}' = \left\{N_{i,a} \leq 10t_*, \text{ and } |N_{\tilde{s}_i,\tilde{s}_i,a} - N_{i,a}q_0| \leq \sqrt{20(1 - q_0)q_0N_{i,a}}\right\}$$

we have

$$\frac{\Pr_{M_{i,a}}[\tau]}{\Pr_{M_0}[\tau]} \geq 4\delta$$

provided appropriately chosen c_1, c_2, c_3, x, c_{10}. By Markov inequality and Doob's inequality (e.g. Lemma 7-8 of [14]), we have

$$\Pr_{M_0}[\mathcal{E}'_{i,a}] \geq 1 - \frac{1}{10} - \frac{1}{10} = \frac{4}{5}.$$

We are able to compute the probability of $\mathcal{E}_0$ on $M_{i,a}$ as follows:

$$\begin{aligned}\Pr_{M_{i,a}}[\mathcal{E}_0] &= \sum_{\tau \in \mathcal{E}_0} \Pr_{M_{i,a}}[\tau] \geq \sum_{\tau \in \mathcal{E}_0 \cap \mathcal{E}'_{i,a}} \Pr_{M_{i,a}}[\tau] \\ &= \sum_{\tau \in \mathcal{E}_0 \cap \mathcal{E}'_{i,a}} \frac{\Pr_{M_{i,a}}[\tau]}{\Pr_{M_0}[\tau]} \cdot \Pr_{M_0}[\tau] \geq 4\delta \sum_{\tau \in \mathcal{E}_0 \cap \mathcal{E}'_{i,a}} \Pr_{M_0}[\tau] \geq 3\delta,\end{aligned}$$

provided $\delta \leq c_{11}$ for some absolute constant c_{11}, hence a contradiction of soundness. □

Wrapping up. Hence, if the algorithm is (ε, δ)-sound for all $\{M_{i,a}\}$, it must be the case that

$$\mathbb{E}_{M_0}[N_{i,a}] \geq t_*, \forall (i,a) \in [|\mathcal{S}|] \times \mathcal{A}.$$

By linearity of expectation, we have

$$\mathbb{E}_{M_0}\left[\sum_{i,a} N_{i,a}\right] \geq |\mathcal{S}||\mathcal{A}| t_*.$$

Since $\varepsilon' = \varepsilon(1-\gamma)\zeta$, $\mathbb{E}_{M_0}\left[\sum_{i,a} N_{i,a}\right] \geq \frac{c_{10}|\mathcal{S}||\mathcal{A}| \log \delta^{-1}}{(1-\gamma)^5 \zeta^2 \varepsilon^2}$, which completes the proof. □

F Estimating ζ

In this section, we show that ζ can be estimated up to error 0.2ζ using small number of samples. The formal guarantee is provided in the following theorem.

Theorem 21. There exists an algorithm that, with probability at least $1-\delta$, halts, takes

$$O\left(\frac{c_{max}\log\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\delta\zeta}\right)}{(1-\gamma)^3\zeta^2}\right)$$

samples per state-action pair, and outputs an estimator $\hat{\zeta}$, such that

$$|\hat{\zeta}-\zeta| \leq 0.2\zeta.$$

Proof. Let $\zeta_i = 2^{-i}/(1-\gamma)$ for $i = 0, 1, 2, \ldots$, and

$$N_i = \frac{c_{\max} C_i(\delta)}{(1-\gamma)^3\zeta_i^2},$$

where

$$C_i(\delta) = c' \log\left(2\frac{|\mathcal{S}||\mathcal{A}|i^2}{(1-\gamma)\zeta_i\delta}\right)$$

for some constant c'. We start by running the algorithm in [21] for N_i samples per state-action pair on the MDP with c as reward , for $i = 0, 1, \ldots$ and stop if the following is satisfied:

$$|\hat{V}^*_{c,i}(\rho) - b| \geq 9\zeta_i,$$

where $\hat{V}^*_{c,i}$ is the empirical optimal value function obtained for using N_i samples. Then we output $\hat{\zeta} = \hat{V}^*_{c,i}(\rho)$.

Next we show that the algorithm halts. Let $\mathcal{E}_i$ be event for iteration i,

$$|\hat{V}^*_{c,i}(\rho) - V^*_c(\rho)| \leq \zeta_i.$$

Thus, by Theorem 1 of [21],

$$\Pr[\mathcal{E}_i] \geq 1 - \frac{\delta}{2i^2}.$$

Next, let i^* be such that $0.05\zeta \leq \zeta_{i^*} < 0.1\zeta$. Hence, if $\mathcal{E}_{i^*}$ happens, then

$$|\hat{V}^*_{c,i^*}(\rho) - V^*_c(\rho)| \leq \zeta_{i^*}$$

and

$$|V^*_c(\rho) - b| - |\hat{V}^*_{c,i^*}(\rho) - V^*_c(\rho)| \leq |\hat{V}^*_{c,i^*}(\rho) - b|.$$

Hence, on $\mathcal{E}_{i^*}$,

$$|\hat{V}^*_{c,i^*}(\rho) - b| \geq \zeta - \zeta_{i^*} \geq 0.9\zeta \geq 9\zeta_{i^*}$$

and the algorithm halts at least before iteration i^*.

Next, suppose the algorithm halts at $i \leq i^*$, then on $\mathcal{E}_i$, we have

$$|\hat{V}^*_{c,i}(\rho) - V^*_c(\rho)| \leq \zeta_i,$$

$$|\hat{V}^*_{c,i}(\rho) - b| \geq 9\zeta_i \geq 9|\hat{V}^*_{c,i}(\rho) - V^*_c(\rho)|,$$

and

$$|(|\hat{V}^*_{c,i}(\rho) - b| - |V^*_{c,i}(\rho) - b|)| \leq |\hat{V}^*_{c,i}(\rho) - V^*_c(\rho)| \leq |\hat{V}^*_{c,i}(\rho) - b|/9.$$

Note that $\zeta = |V^*_c(\rho) - b|$, we have

$$|\hat{\zeta} - \zeta| \leq \hat{\zeta}/9 \implies \zeta \geq 8/9\hat{\zeta} \text{ and } |\hat{\zeta} - \zeta| \leq \zeta/8.$$

Thus, on event $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \ldots \mathcal{E}_{i^*}$, which happens with probability at least

$$1 - \sum_{i=1}^{i^*} \frac{\delta}{2i^2} \geq 1 - \frac{\pi^2 \delta}{12},$$

we have $|\hat{\zeta} - \zeta| \leq \zeta/8$, proving the correctness.

We now consider the overall sample complexity. Suppose $\mathcal{E}$ happens, then the number of samples consumed is upper bounded by

$$\sum_{i=1}^{i^*} N_i \leq \frac{c'' c_{max} C_{i^*}(\delta)}{(1-\gamma)^3 \zeta^2} = \frac{c'' c' c_{max} \log\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\delta\zeta}\right)}{(1-\gamma)^3 \zeta^2},$$

for some constant c'', completing the proof. □

G Comparison to Bai et al. [6]

For a target error of ε, our lower-bound construction in the strict feasibility setting (Section 7) shows that it is important to estimate the constraint value function to a smaller error equal to ε'. Intuitively, this is because a small estimation error in the constraint value can (incorrectly) render the optimal policy infeasible and result in a $\frac{\text{Range(value function)}}{\zeta}$ suboptimality gap.

For the theoretical results in [6], the value functions are normalized and hence Range(value function) $= 1$. Hence, a small constraint violation can result in an $\frac{1}{\zeta}$ suboptimality gap, and the constraint value function needs to be estimated to a smaller error equal to $\varepsilon' := \varepsilon\zeta$. Combining this with the standard results for the primal-dual algorithm for unconstrained MDPs [32] for *normalized value functions*, this implies a sample-complexity of $O(H^2SA/\varepsilon'^2) = O(H^2SA/\varepsilon^2\zeta^2)$ which is the sample complexity reported in [6, Theorem 2].

On the other hand, if we scale the value function to lie to the standard $O\left(1/1-\gamma\right)$ range, a small constraint violation can result in an $\frac{1}{(1-\gamma)\zeta}$ suboptimality gap, and the constraint value function needs to be estimated to a smaller error equal to $\varepsilon' := \varepsilon(1-\gamma)\zeta$. This rescaling affects the sample complexity for the primal-dual algorithm for unconstrained MDPs [32]. Specifically, for value functions in the $[0, 1/1-\gamma]$ range, unconstrained MDPs require $O\left(\frac{H^4SA}{\varepsilon^2}\right)$ samples to estimate the value function to an error ε. Since we require a smaller error ε' in the strict feasibility setting, this implies an $O\left(\frac{H^4SA}{\varepsilon'^2}\right) = O\left(\frac{H^6SA}{\varepsilon^2\zeta^2}\right)$ sample complexity stated in Section 5.