Model-Free, Regret-Optimal Best Policy Identification in Online CMDPs

Anonymous authors
Paper under double-blind review

Abstract

This paper considers the best policy identification (BPI) problem in online Constrained Markov Decision Processes (CMDPs). We are interested in algorithms that are model-free, have low regret, and identify an optimal policy with a high probability. Existing model-free algorithms for online CMDPs with sublinear regret and constraint violation do not provide any convergence guarantee to an optimal policy and provide only average performance guarantees when a policy is uniformly sampled at random from all previously used policies. In this paper, we develop a new algorithm, named Pruning-Refinement-Identification (PRI), based on a fundamental structural property of CMDPs we discover, called *limited stochasticity*. The property says for a CMDP with N constraints, there exists an optimal policy with *at most* N stochastic decisions.

The proposed algorithm first identifies at which step and in which state a stochastic decision has to be taken and then fine-tunes the distributions of these stochastic decisions. PRI achieves trio objectives: (i) PRI is a model-free algorithm; and (ii) it outputs a near-optimal policy with a high probability at the end of learning; and (iii) in the tabular setting, PRI guarantees $\tilde{\mathcal{O}}(\sqrt{K})$[1] regret and constraint violation, which significantly improves the best existing regret bound $\tilde{\mathcal{O}}(K^{\frac{4}{5}})$ under a model-free algorithm, where K is the total number of episodes.

1 Introduction

In unconstrained reinforcement learning (RL), an agent aims to find the optimal policy that maximizes the accumulated reward by interacting with a stochastic environment. RL has achieved remarkable success in multiple areas, including industrial process optimization, robotics, and gaming. (Rajawat et al., 2023; Abeyruwan et al., 2023; Lindegaard et al., 2023; Liu et al., 2023). However, in many real-world applications, the learned policy must also satisfy a set of constraints. For example, in healthcare applications, we need to optimize patient treatment plans while considering constraints like medication dosage, scheduling of medical procedures, and resource allocation in hospitals. These constrained versions of RL problems can be formulated as Constrained Markov Decision Processes (CMDPs) (Altman, 1999).

Learning in CMDPs has become an active research topic recently. Existing solutions include both model-based (Brantley et al., 2020; Efroni et al., 2020; Singh et al., 2020; Liu et al., 2021; Bura et al., 2021; Ding et al., 2021; Chen et al., 2022) and model-free algorithms (Wei et al., 2022b; Ghosh et al., 2022; Wei et al., 2022a). This paper focuses on model-free approaches for CMDPs due to their computation and memory efficiency. A fundamental drawback of existing model-free algorithms for best policy identification in online CMDPs is that they provide only average performance guarantees for a policy uniformly sampled at random from *all* previously used policies during learning, so they fail to identify a single optimal or a near-optimal policy.[2] Therefore, a natural question arises:

Is it possible to identify an optimal or a near-optimal policy in online CMDPs with the model-free approach with optimal regret?

[1] **Notation:** $f(n) = \tilde{\mathcal{O}}(g(n))$ denotes $f(n) = \mathcal{O}(g(n)\log^k n)$ with $k > 0$. The same applies to $\tilde{\Omega}$.

[2] In this paper, a policy is a mapping from a state at a given step to an action distribution, without any other additional input information. An algorithm that uses multiple policies, e.g. randomly sampling one policy from many policies, is explicitly called a *mixed* policy.

There are two key challenges to answering this question: (i) CMDP problems are typically represented as Linear Programming (LP) problems, resulting in stochastic optimal policies. Model-free online CMDP algorithms often employ the primal-dual approach, utilizing Lagrange multipliers to balance reward maximization and constraint violation. However, these methods yield "greedy policies" for fixed Lagrange multipliers, which aren't necessarily optimal. Consequently, model-free algorithms such as Triple-Q Wei et al. (2021) offer performance guarantees only in terms of averages over various greedy policies determined by different Lagrange multipliers, failing to converge to a single policy. (ii) The best-known regret bound of model-free algorithms for episodic, online CMDPs is $\tilde{\mathcal{O}}(K^{\frac{4}{5}})$ Wei et al. (2022a). It is also known that model-based algorithms can achieve a smaller and order-wise tight regret $\tilde{\mathcal{O}}(\sqrt{K})$ Efroni et al. (2020). The open question is whether a model-free algorithm can reach $\tilde{\mathcal{O}}(\sqrt{K})$ regret in online CMDPs?

This paper tackles both challenges, providing affirmative responses to both questions. We introduce a novel algorithm, Pruning-Refinement-Identification (PRI), rooted in a fundamental CMDP property we unveil—limited stochasticity. This property asserts that for an episodic CMDP with N constraints, there exists an optimal policy making stochastic decisions in at most N step-dependent states out of the HS step-dependent states, where H denotes episode length and S represents possible states at each step.

Based on this insight, PRI consists of three phases. In this first phase (pruning), PRI identifies when and where stochastic decisions are necessary. This defines a set of greedy policies that together approximate a "mixed" optimal policy. The subsequent refinement phase involves learning the weights of these greedy policies. This is done through iterative optimization, utilizing empirical reward and utility value functions. The process refines value function estimates with each iteration, aiming to minimize regret. In the final identification phase, PRI learns the occupancy measure, determining the probability of visiting specific state-action pairs at each step. This information is used to recover a single policy from the near-optimal mixed policy obtain during the refinement phase. The main contributions of this paper include:

- PRI is the first model-free PAC RL algorithm for CMDPs, achieving optimal regret and minimal constraint violation.
- PRI outputs a near-optimal policy with a high probability at the end. The learned policy has $\tilde{\mathcal{O}}(1/\sqrt{K})$ optimality gap with probability $1 - \tilde{\mathcal{O}}(K^{-0.1})$.
- In the tabular setting, PRI guarantees $\tilde{\mathcal{O}}(\sqrt{K})$ regret and constraint violation, which significantly improves the best existing regret bound $\tilde{\mathcal{O}}(K^{\frac{4}{5}})$ under a mode-free algorithm, where K is the total number of episodes. Unlike existing regret bounds, the dominating term in terms of K in the regret bound does not depend on the sizes of the state space and action space.

2 Related Work

Best policy identification in MDPs. For unconstrained MDPs, existing studies on BPI focus on (ϵ, δ)-PAC RL algorithms, i.e., algorithms that identify an ϵ-optimal policy with probability at least $1 - \delta$. Such a learning objective has been considered extensively in discounted and episodic tabular MDPs (Agarwal et al., 2020; Azar et al., 2013; Even-Dar et al., 2006; Domingues et al., 2021; He et al., 2021; Sidford et al., 2018). A recent work Taupin et al. (2022) also studied BPI in linear MDPs, which has a sample complexity of $O(\frac{1}{\epsilon^2})$. In other words, the best regret result it might get is $O(\sqrt{K})$. This paper considers BPI for CMDPs using a model-free approach. To the best of our knowledge, it remains an open question.

Model-based and Model-free algorithms for online CMDPs. As mentioned in the introduction, most existing results on online CMDPs consider regret minimization. For example, Brantley et al. (2020); Efroni et al. (2020); Singh et al. (2020) proposed model-based algorithms for episodic tabular CMDPs. Liu et al. (2021); Bura et al. (2021) proposed efficient algorithms with zero or bounded constraint violation. For model-free algorithms, Wei et al. (2022b) developed Triple-Q that achieves sublinear regret and zero constraint violation in episodic tabular CMDPs. Similar results have been established for linear CMDPs (Ghosh et al., 2022; Ding et al., 2021) and infinite-horizon average CMDPs (Chen et al., 2022; Wei et al., 2022a). However, these existing model-free algorithms for online CMDPs does not converge to an optimal or a near-optimal policy. Note that model-free

Table 1: The Exploration-Exploitation Tradeoff in Episodic CMDPs.

	Algorithm	Regret	Constraint Violation	BPI?
Model-based	OPDOP (Ding et al., 2021)	$\tilde{\mathcal{O}}(H^3\sqrt{S^2AK})$	$\tilde{\mathcal{O}}(H^3\sqrt{S^2AK})$	✗
	OptDual-CMDP (Efroni et al., 2020)	$\tilde{\mathcal{O}}(H^2\sqrt{S^3AK})$	$\tilde{\mathcal{O}}(H^2\sqrt{S^3AK})$	✗
	OptPrimalDual-CMDP (Efroni et al., 2020)	$\tilde{\mathcal{O}}(H^2\sqrt{S^3AK})$	$\tilde{\mathcal{O}}(H^2\sqrt{S^3AK})$	✗
	CONRL (Brantley et al., 2020)	$\tilde{\mathcal{O}}(H^3\sqrt{S^3A^2K})$	$\tilde{\mathcal{O}}(H^3\sqrt{S^3A^2K})$	✗
	OptPess-LP (Liu et al., 2021)	$\tilde{\mathcal{O}}(H^3\sqrt{S^3AK})$	0	✗
	OptPess–PrimalDual (Liu et al., 2021)	$\tilde{\mathcal{O}}(H^3\sqrt{S^3AK})$	$\mathcal{O}(1)$	✗
	OPSRL(Bura et al., 2021)	$\tilde{\mathcal{O}}(\sqrt{S^4H^7AK})$	0	✗
Model-free	Triple-Q(Wei et al., 2022a)	$\tilde{\mathcal{O}}(\frac{1}{\delta}H^4S^{\frac{1}{2}}A^{\frac{1}{2}}K^{\frac{4}{5}})$	0	✗
	PRI	$\tilde{\mathcal{O}}(\sqrt{H^2K})$	$\tilde{\mathcal{O}}(\sqrt{H^2K})$	✓

algorithms have a memory complexity of $O(HSA)$ for maintaining the Q-table while the memory complexity of model-based algorithms is $O(HS^2A)$ for maintaining the transition kernel. Very recently, Moskovitz et al. (2023) considered BPI for online CMDPs. They formulated the CMDP problem as a min-max game and the proposed algorithm converges to a near-optimal policy at the last iteration with optimistic mirror descent. However, the paper does not provide any regret guarantee when learning the near-optimal policy. There are also algorithms for the average-reward CMDP problem, including model-based approaches Agarwal et al. (2021; 2022); Zheng & Ratliff (2020) and model-free approaches Chen et al. (2022); Wei et al. (2022b). These algorithms do not identify the optimal policy at the end. Table 1 summarizes the recent results on online, episodic CMDPs.

3 Problem Formulation

We consider an episodic CMDP, denoted by $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, g^n, n \in [N])$, where $\mathcal{S}$ is the state space ($|\mathcal{S}| = S$), $\mathcal{A}$ is the action space ($|\mathcal{A}| = A$), $\{r_h\}_{h=1}^H, \{g_h^n\}_{h=1}^H, n \in [N]$ are reward, n-th utility functions, and $\mathbb{P} = \{\mathbb{P}_h(\cdot|x, a)\}_{h=1}^H$ are the transition kernels. For simplicity, we assume that in each episode, the agent starts from the same initial state $x_1 = x_{ini}$. It is straightforward to generalize the results to the case when the initial state is sampled with a given distribution but the notation becomes cumbersome. We also assume that $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ and $g_h^n : \mathcal{S} \times \mathcal{A} \to [0, 1]$ are deterministic for notation simplicity. Our results can be easily generalized to random reward/utility signals.

During each episode, the agent interacts with the environment as follows: at each step h, the agent takes action a_h after observing state x_h, receives reward $r_h(x_h, a_h)$ and N utility values $g_h^n(x_h, a_h)$ ($n \in [N]$), and then observes a new state (x_{h+1}), which evolves by following the transition kernel $\mathbb{P}_h(\cdot|x_h, a_h)$. The episode terminates after H steps.

Given a stochastic policy π, which is a collection of H functions $\{\pi_h : \mathcal{S} \times \mathcal{A} \to [0, 1]\}_{h=1}^H$, the agent takes action a with probability $\pi_h(a|x)$ when being in state x at step h . The reward value function of policy π, denoted by $V_h^\pi(x)$, is the expected total reward when starting from an arbitrary state x at step h to the end of the episode: $V_h^\pi(x) = \mathbb{E}_\pi\left[\sum_{i=h}^H r_i(x_i, a_i)\middle| x_h = x\right]$, where the expectation is taken with respect to the policy π and randomness from the transition kernels. Accordingly, the reward Q-function, denoted by $Q_h^\pi(x, a)$, is the expected total reward when the agent starts from an arbitrary action-action pair (x, a) at step h and follows policy π to the end of the episode: $Q_h^\pi(x, a) = r_h(x, a) + \mathbb{E}_\pi\left[\sum_{i=h+1}^H r_i(x_i, a_i)\middle| x_h = x, a_h = a\right]$.

Similarly, we can define the N utility value functions as $W_h^{\pi,n}(x) = \mathbb{E}_\pi\left[\sum_{i=h}^H g_i^n(x_i, a_i)\middle| x_h = x\right]$ and utility Q-functions as $C_h^{\pi,n}(x, a) = g_h^n(x, a) + \mathbb{E}_\pi\left[\sum_{i=h+1}^H g_i^n(x_i, a_i)\middle| x_h = x, a_h = a\right]$. Given the definitions above, we have

$$V_h^\pi(x) = \sum_a \pi_h(a|x)Q_h^\pi(x, a) \quad Q_h^\pi(x, a) = r_h(x, a) + \sum_{x'} \mathbb{P}_h(x'|x, a)V_{h+1}^\pi(x') \tag{1}$$

$$W_h^{\pi,n}(x) = \sum_a \pi_h(a|x)C_h^{\pi,n}(x, a) \quad C_h^{\pi,n}(x, a) = g_h^n(x, a) + \sum_{x'} \mathbb{P}_h(x'|x, a)W_{h+1}^{\pi,n}(x'). \tag{2}$$

The objective of the CMDP is to find an optimal policy that maximizes the expected total reward while making sure the $n-$th expected total utility is no less than $\rho^{(n)}$ for all $n \in [N]$:

$$\pi^* \in \arg\max_{\pi} V_1^{\pi}(x_{ini}) \quad \text{s.t.} \quad W_1^{\pi,n}(x_{ini}) \geq \rho^{(n)} \quad \forall n \in [N]. \tag{3}$$

To avoid triviality, we assume $\rho^{(n)} \in [0, H]$. For simplicity, we use V_1^{π} to represent $V_1^{\pi}(x_{ini})$ and $W_1^{\pi,n}$ to represent $W_1^{\pi,n}(x_{ini})$.

We evaluate an online RL algorithm for CMDP using regret and constraint violation over K episodes:

$$\text{Regret}(K) = KV_1^{\pi^*}(x_{ini}) - \mathbb{E}\left[\sum_{k=1}^{K} V_1^{\pi_k}(x_{ini})\right] \tag{4}$$

$$\text{Violation}^n(K) = K\rho^{(n)} - \mathbb{E}\left[\sum_{k=1}^{K} W_1^{\pi_k,n}(x_{ini})\right] \tag{5}$$

where π_k is the policy used in episode k.

4 PRI (Pruning-Refinement-Identification)

Before formally introducing our algorithm, we first present two structural properties of the optimal solution to the CMDP problem equation 3. These properties have been overlooked in the literature but serve as the foundation of our proposed algorithm. Consider a CMDP problem with N constraints. It is well-known that the problem can be formulated as a linear programming (LP) problem Altman (1999):

$$\max_{\{q_h(x,a)\}} \sum_{h,x,a} q_h(x,a) r_h(x,a) \tag{6}$$

$$\text{s.t.:} \sum_{h,x,a} q_h(x,a) g_h^{(n)}(x,a) \geq \rho^{(n)} \quad \forall n \in [N] \tag{7}$$

$$\sum_a q_{h+1}(x,a) = \sum_{x',a'} \mathbb{P}_h(x|x',a') q_h(x',a') \quad \forall x \in \mathcal{S}, h \in [H] \tag{8}$$

$$\sum_a q_1(x_{ini},a) = 1, \sum_a q_1(x,a) = 0,\ x \neq x_{ini} \tag{9}$$

$$q_h(x,a) \geq 0, \tag{10}$$

where $q_h(x,a)$ denotes the probability that state-action pair (x,a) is visited at step h, called the occupancy measure. Each feasible solution $\{q_h(x,a)\}_{h,x,a}$ to the problem leads to a corresponding Markov policy: $\pi_h(a|x) = \frac{q_h(x,a)}{\sum_a q_h(x,a)}$. In this paper, we call probability distribution $\pi_h(\cdot|x)$ *decision* at state x at step h. So a policy consists of $S \times H$ decisions. A decision $\pi_h(\cdot|x)$ is called *greedy* if $\pi_h(a|x) = 1$ for some $a \in \mathcal{A}$ and stochastic otherwise.

Lemma 1 (Limited Stochasticity). *If $q^* = \{q_h^*(x,a)\}_{h,x,a}$ is an optimal solution to the CMDP problem equation 6-equation 10 and is an extreme point, then there are at most $HS + N$ nonzero values in q^*. This implies that the optimal policy derived from q^* includes at most N stochastic decisions.*

The detailed proof can be found in Appendix B. The following corollary, which is a well-known result, is a direct consequence of the lemma.

Corollary 1. *For unconstrained MDP problems, one of the optimal policies is a greedy policy.*

Given an occupancy measure q and its induced policy π, we define $\mathcal{D}_{h,x}(q) = \{a : q_h(x,a) > 0\}$, which is the set of actions that will be taken with a nonzero probability in state x at step h under the policy π induced by q. Note that if $\pi_h(\cdot|x)$ is a greedy decision, then $|\mathcal{D}_{h,x}(q)| = 1$; and if $\pi(\cdot|x)$ is greedy, then $|\mathcal{D}_{h,x}(q)| > 1$. Let $M = \prod_{h,x} |\mathcal{D}_{h,x}(q)|$, and let π^m represent the mth greedy policy ($m = 1, \cdots, M$) constructed from $\otimes_{h,x}\mathcal{D}_{h,x}(q)$ such that $\pi_h^m(a|x) = 1$ only if $a \in \mathcal{D}_{h,x}(q)$.Note that according to lemma 1, we have $M \leq 2^N$. A greedy policy is a policy under which all decisions are greedy. Next, we will show that a Markov policy is equivalent to a mixed policy of many greedy policies in the following lemma, whose proof can be found in Appendix B.2.

Lemma 2 (Decomposition). *Given any Markov policy π and its corresponding occupancy measure q, there exists a set of M greedy policies and a probability distribution $\{a_m\}_{m=1,\cdots,M}$ such that the mixed policy, which selects a greedy policy π^m at the start of an episode with probability a_m and subsequently follows it, has the same occupancy measure q as the original policy π.*

Online model-free algorithms for CMDPs, such as Triple-Q Wei et al. (2021), guarantee sublinear regret and constraint violation on average but have no convergence guarantee. In fact, Triple-Q continues to adjust the dual variable (virtual queue) based on constraint violation, and when the dual variable is fixed (within a frame), the algorithm reduces to the traditional Q-learning. As suggested in the paper Wei et al. (2021), we can only recover a near-optimal policy by remembering all previous policies and then uniformly sampling one from them for each episode, i.e., a mixed policy of K policies. Therefore, this near-optimal policy is a mixture of many, many greedy policies. More importantly, it is near-optimal only when averaging over a large number of episodes and may be far from optimal in each episode.

Hence, unlike unconstrained MDPs where Q-learning converges to the optimal policy, finding a model-free algorithm that converges to the optimal policy or a near-optimal policy in CMDPs is highly nontrivial and remains to be an open problem. Lemma 1 (limited stochasticity), however, suggests that when the number of constraints, N, is relatively small, solving an unknown CMDP may not differ significantly from solving an unknown MDP. This is because the majority of the decisions, specifically $HS - N$ out of the HS decisions, are greedy and can be learned using traditional algorithms like Q-learning if we can first identify where the stochastic decisions need to be taken. Lemma 2 further suggests that an optimal policy can be decomposed into M greedy policies if all decision types are correctly identified, so we may recover an optimal or a near-optimal policy by evaluating the M greedy policies.

We will first consider the case where *the LP has a unique optimal solution*. Leveraging these two observations from Lemma 1 and 2, we propose a novel three-phase algorithm (Algorithm 1), including policy pruning, policy refinement, and policy identification, called PRI.

The algorithm is presented in Algorithm 1, which includes $\sqrt{K} + 2K$ episodes, $\sqrt{K}$ episodes for pruning, K episodes for refinement and K episodes for identification. In the first phase (policy pruning), we run Triple-Q for $\sqrt{K}$ episodes, we denote $\{\pi_{k,h}\}_{h=1}^{H}$ as the policy used by Triple-Q in the kth episode, and it is a greedy policy.

Remark: For fixed (h, x, a) in the policy pruning phase, we use $\tilde{N}_h(x, a)$ to count the number of episodes in which the greedy policy we follow is $\pi_h(a|x) = 1$, which is the number of greedy policies (among the $\sqrt{K}$ greedy policies) that would take action a if the agent visits state x at step h.

Because of the sub-linear regret and zero violation guaranteed by Triple-Q, we expect that $\frac{N_h(x,a)}{\sqrt{K}}$ is close to zero if $\pi_h^*(a|x) = 0$ and is a non-negligible positive value if otherwise. Therefore, with a high probability, $\tilde{\mathcal{D}}_{h,x} = \mathcal{D}_{h,x}(q^*)$, where $\tilde{\mathcal{D}}_{h,x}$ is gradually updated in Algorithm 1 (Lines 8-10).

After the first phase, PRI obtains M greedy policies. In the second phase, PRI learns the weights $\{\alpha_m\}$ so that a mixed policy that chooses policy π^m with probability α_m is statistically identical to the optimal policy. This is achieved by learning the reward and utility value functions of the greedy policies and then solving an approximated version of the CMDP (Decomposition-Opt equation 11).

At each round of the second phase (policy refinement), the following optimization with M optimization variables is solved.

$$\textbf{Decomposition-Opt:} \quad \max_{\{a_m\}_{m=1}^{M}} \sum_{m=1}^{M} \alpha_m \bar{V}_1^{\pi^m}$$
$$\text{s.t.:} \left| \sum_{m=1}^{M} \alpha_m \bar{W}_1^{\pi^m,n} - \rho^{(n)} \right| \leq \sqrt{\frac{H^2 \log\left((t-1)\epsilon' K\right)}{\epsilon'(t-1)\sqrt{K}}} \quad \forall n, \tag{11}$$
$$\sum_m \alpha_m = 1, \alpha_m \geq \epsilon' \quad \forall m.$$

Algorithm 1 PRI

```
Phase 1: Policy Pruning
Initialize $\tilde{N}_h(x,a) = 0$ for all $h$, $x$ and $a$. Other initialization is the same as in Triple-Q.
for $k = 1, \cdots, \sqrt{K}$ do
    For all $(h,x,a)$, $\tilde{N}_h(x,a) \leftarrow \tilde{N}_h(x,a) + \pi_{k,h}(a|x)$.
    Execute Triple-Q for one episode.
for all $(h,x)$ do
    Initialize $\tilde{\mathcal{D}}_{h,x} = \emptyset$
    for all $a \in \mathcal{A}$ do
        $\tilde{\mathcal{D}}_{h,x} \leftarrow \tilde{\mathcal{D}}_{h,x} \bigcup \{a\}$ if $\frac{\tilde{N}_h(x,a)}{\sqrt{K}} \geq \frac{\epsilon}{2}$.
Obtain $M$ greedy policies from $\{\tilde{\mathcal{D}}_{h,x}\}_{h,x}$ where $M = \prod_{h,x} |\tilde{\mathcal{D}}_{h,x}|$.
Phase 2: Policy Refinement
if $M = 1$ then
    Output the greedy policy.
else
    Set $\hat{V}_1^{\pi^m} = 0, \hat{W}_1^{\pi^m,n} = 0$, and $a_m = \frac{1}{M}$ for all $n$ and $m$.
    for round $t = 1, \cdots, \sqrt{K}$ do
        for $m = 1, \cdots, M$ do
            for $k = 1, \cdots, \alpha_m \sqrt{K}$ do
                Execute greedy policy $\pi^m$ for one episode.
                if $k \leq \epsilon' \sqrt{K}$ then
                    Set $\hat{V}_1^{\pi^m} \leftarrow \hat{V}_1^{\pi^m} + V_{k,1}^{\pi^m}$ and $\hat{W}_1^{\pi^m,n} \leftarrow \hat{W}_1^{\pi^m,n} + W_{k,1}^{\pi^m,n}$ for all $n$, where
                    $V_{k,1}^{\pi^m}$ and $W_{k,1}^{\pi^m,n}$ are the total reward and utility of type $n$ received in the $k$th episode.
            Set $\bar{V}_1^{\pi^m} = \frac{\hat{V}_1^{\pi^m}}{t\epsilon'\sqrt{K}}$ and $\bar{W}_1^{\pi^m,n} = \frac{\hat{W}_1^{\pi^m,n}}{t\epsilon'\sqrt{K}}$ for all $n$.
        Update $\{\alpha_m\}$ by solving Decomposition-Opt equation 11.
Phase 3: Policy Identification
Initialize $N_h(x,a) = 0$ for all $h$, $x$ and $a$.
for $t = 1, \cdots, \sqrt{K}$ do
    for $m = 1, \cdots, M$ do
        for $k = 1, \cdots, \alpha_m \sqrt{K}$ do
            for $h = 1, \cdots, H$ do
                Take action $a_h$ given by policy $\pi^m$, i.e. $\pi^m(a_h|x_h) = 1$.
                $N_h(x_h, a_h) \leftarrow N_h(x_h, a_h) + 1$.
For all $(h,x,a)$, set $\tilde{\pi}_h(a|x) = \frac{N_h(x,a)}{\sum_{\tilde{a} \in \mathcal{A}} N_h(\tilde{a},x)}$.
Output policy $\tilde{\pi}$.
```

After learning sufficiently accurate $\{\alpha_m\}$ in the second phase, PRI learns the occupancy measure under the mixed policy defined by $\{\alpha_m\}$ and constructs a Markov policy $\tilde{\pi}$ based on the learned occupancy measure.

In the next section, we will show that PRI guarantees $\mathcal{O}(\sqrt{K})$ regret and constraint violation and outputs a near-optimal policy $\tilde{\pi}$ with a high probability. An informal statement of the main results is presented below. The formal statements of the theorems and the proofs will be presented in the next section.

Main Results: Assume the LP defined by equation 6-equation 10 has a unique solution. With a high probability, PRI yields policy $\tilde{\pi}$ such that

- $\{(h,x,a) : \tilde{\pi}_h(a|x) > 0\} = \{(h,x,a) : \pi_h^*(a|x) > 0\}$,
- PRI guarantees $\mathcal{O}(\sqrt{K})$ regret and constraint violation over the $\sqrt{K} + 2K$. episodes, and
- $|\tilde{\pi}_h(a|x) - \pi_h^*(a|x)| = \mathcal{O}(1/\sqrt{K})$ for all (h,x,a), and $\tilde{\pi}_h(a|x) = \pi_h^*(a|x)$ if $\pi_h^*(a|x) \in \{0,1\}$.

Remark: If the LP has more than one solution, we will introduce a multi-solution pruning algorithm to the policy pruning phase of PRI to resolve the issue. The algorithm and the analysis can be found in Section 6.

5 Main Results

In this section, we provide our main results assuming that the LP associated with the CMDP problem has a unique solution. This assumption can be relaxed and the results can be found in Section 6. Letπ^* be the unique optimal policy and $\{q_h^{\pi^*}(x,a)\}$ is the corresponding occupancy measure. Furthermore, let $\{\pi^m\}$ $(m=1,\cdots,M)$ be the set of greedy policies associated with the optimal policy as defined in Lemma 2, and $\{a_m^*\}$ the associated weights. We also make the following additional assumptions.

Assumption 1. *The ϵ and ϵ' used in PRI satisfy $q_h^{\pi^*}(x,a) \geq \epsilon$ for any (h,x,a) such that $\pi_h^*(a|x) > 0$, and $\min_m a_m^* \geq \epsilon' > 0$.*

Assumption 2. *There exist two positive constants c_v and c_w such that given a feasible occupancy measure q^π to the LP and the corresponding reward value function and utility value function V^π and $W^{\pi,n}$, we have either $V_1^{\pi^*} - V_1^\pi \geq c_v||q^{\pi^*} - q^\pi||_1$ or for some $n \in [N]$, $W_1^{\pi^*,n} - W_1^{\pi,n} \geq c_w||q^{\pi^*} - q^\pi||_1$, where $\|\cdot\|_1$ is the $L1$-norm.*

Recall a feasible occupancy measure defines a unique Markov policy. The assumption above states that when a policy's occupancy measure is different from that of the unique optimal policy, then either the reward value function or one of the utility reward functions should also be different from that under the optimal policy.

Assumption 3. *Under any greedy policy π, for all x and h, we have $\Pr(x_h = x) = \sum_{x',a'} q_{h-1}^\pi(x',a')\mathbb{P}_h(x|x',a') > p_{\min}$.*

This assumption above says all states should be visited with a non-negligible probability under any greedy policy. It is worth noting that this assumption can be removed if we apply the extension version of PRI, which is stated in section 6. To prove our main result, we first recall the regret and constraint violation guaranteed under Triple-Q Wei et al. (2022a) in the following lemma.

Lemma 3. *For sufficiently large K, over K episodes, Triple-Q guarantees $\tilde{\mathcal{O}}(K^{0.8})$ regret and zero constraint violation, and furthermore,*

$$\Pr\left(K\rho^{(n)} - \sum_{k=1}^{K} W_1^{\pi_k,n} \leq 0\right) = 1 - \mathcal{O}\left(\frac{1}{K^2}\right). \tag{12}$$

Noting that we use Triple-Q in the pruning phase, the assumptions for Triple-Q like Slater's condition are still required. A brief review can be found in the Appendix A. We remark that PRI can be viewed as a "meta-algorithm" that builds on any model-free CMDP algorithm with sublinear regret and constraint violation. In the following theorem, we show that PRI can correctly classify stochastic and greedy decisions with a high probability after the pruning phase.

Theorem 1 (Pruning)**.** *Let $\mathcal{D}^* = \{(h,x,a) : \pi_h^*(a|x) > 0\}$ and $\tilde{\mathcal{D}} = \left\{(h,x,a) : \frac{\tilde{N}_h(x,a)}{\sqrt{K}} \geq \frac{\epsilon}{2}\right\}$. Under Assumptions 1 and 3, after policy pruning, we have*

$$\Pr\left(\tilde{\mathcal{D}}_{h,x} = \mathcal{D}_{h,x}(q^*),\ \forall (h,x)\right) = 1 - \tilde{\mathcal{O}}\left(K^{-0.1}\right). \tag{13}$$

The detailed proof is deferred to Appendix C. Note that since the pruning phase includes $\sqrt{K}$ episodes, the regret and constraint violation are both bounded by $H\sqrt{K}$.

The following theorem shows that the regret and constraint violation during the refinement phase are both $\tilde{\mathcal{O}}(\sqrt{K})$. Note that $\tilde{\mathcal{O}}(\sqrt{K})$ regret and constraint violation imply that the learned mixed policy is close to optimal. The proof can be found in Appendix D.

Theorem 2 (Refinement)**.** *Assume $\tilde{\mathcal{D}} = \mathcal{D}^*$ after policy pruning. Under Assumption 1 to 3, with probability $1 - \tilde{\mathcal{O}}(\frac{1}{\sqrt{K}})$, the regret and constraint violation during the policy refinement phase are both $\tilde{\mathcal{O}}(H\sqrt{K})$.*

The refinement phase learns a near optimal mixed policy, which is a combination of M greedy policies for $M \leq 2^N$. In the following theorem we show that the identification phase is to find a single near-optimal policy by using the occupancy measure of the mixed policy. The proof can be found in Appendix E.

Theorem 3 (Identification)**.** *Assume $\tilde{\mathcal{D}} = \mathcal{D}^*$ after policy pruning. Under Assumption 1 to 3, with probability $1 - \tilde{\mathcal{O}}(\frac{1}{K})$, the regret and constraint violation during the policy identification phase are both $\mathcal{O}(\sqrt{K})$. Furthermore, $|\tilde{\pi}_h(a|x) - \pi^*_h(a|x)| = \mathcal{O}(\frac{1}{\sqrt{K}})$ if $0 < \pi^*_h(a|x) < 1$ and $\tilde{\pi}_h(a|x) = \pi^*_h(a|x)|$ if $\pi^*_h(a|x) \in \{0, 1\}$.*

By summarizing the results from the three theorems above, we have the regret and the constraint violation over the $\sqrt{K} + 2K$ episodes are $\tilde{\mathcal{O}}(H\sqrt{K})$ with probability $1 - \tilde{\mathcal{O}}(\frac{1}{K^{0.1}})$. Consider the regret, the pruning phase includes $\sqrt{K}$ episodes, resulting in at most $H\sqrt{K}$ regret. Theorem 2 and Theorem 3 show that the regret in the refinement and identification phases are both $\tilde{\mathcal{O}}(H\sqrt{K})$. Note that the order-wise bounds are independent of S and A, unlike those in the literature. However, there is an implicit dependence on S and A as the results hold only when K is sufficiently large and how large K needs to be depends on S and A.

6 Extension to CMDPs with Multiple Optimal Solutions

In this section, we consider the case where the optimal policy is not unique so the LP has multiple optimal solutions. Here, the RL agent's objective is to learn one of these optimal policies. According to Lemma 1, an optimal solution associated with an extreme point of the LP involves no more than $HS + N$ stochastic decisions. Additionally, any optimal policy can be viewed as a combination of the optimal policies associated with the extreme points. We define the set of optimal policies as Π^* and the subset associated with extreme points as $\Pi^{*,e}$. We expand our assumptions to the case of multiple solutions as follows.

Assumption 4. *The ϵ used PRI satisfies $\min_{(h,x,a):\pi_h(a|x)\neq 0} q^\pi_h(x, a) \geq \epsilon \quad \forall \pi \in \Pi^{*,e}$.*

Assumption 5. *Given any occupancy measure q' and the induced Markov policy π', there exists an optimal policy $\pi^* \in \Pi^*$ such that $V^{\pi^*}_1 - V^{\pi'}_1 \geq c_v||q^{\pi^*} - q'||_1$ or for some n, $W^{\pi^*,n}_1 - W^{\pi',n}_1 \geq c_w||q^{\pi^*} - q'||_1$, where c_v and c_w are two positive constants.*

Note that if π' is an optimal policy, then the assumption holds trivially with $\pi^* = \pi'$. Under Assumptions 3-5, the following theorem shows that a unique optimal policy is identified after Multi-Solution Pruning. The algorithm result in at most $H^2SAK^{0.25}$ regret and constraint violation.

Theorem 4. *Under Assumption 4 and 5, with probability $1 - \mathcal{O}(1/K^{0.02})$, for sufficiently large K, multi-solution pruning outputs a unique optimal policy with at most N stochastic decisions. The regret and constraint violation during multi-solution pruning are bounded by $H^2SAK^{0.25}$ with probability one.*

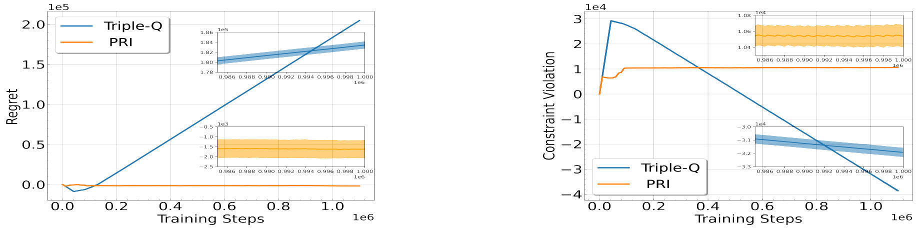
More discussions and the detailed proof are deferred to Appendix F.1 due to the page limit. We note that adding this multi-solution pruning to PRI only increases the regret and constraint violation by $HSAK^{0.25}$ which is order-wise smaller than $\tilde{\mathcal{O}}(H\sqrt{K})$ in terms of K. Therefore, the regret and constraint violation remains to be $\tilde{\mathcal{O}}(H\sqrt{K})$ for sufficiently large K. The learned policy is a near optimal policy with $\tilde{O}(1/\sqrt{K})$ gap with probability $1 - \tilde{O}(K^{-0.02})$.

7 Experiments

Synthetic CMDP

This section presents numerical evaluations of the proposed algorithm. We first evaluated our algorithm for a synthetic CMDP with a single constraint. The transition kernels, rewards, and utilities are chosen such that the problem has a unique optimal solution and satisfies Assumption 3. The objective is to maximize the cumulative reward while guaranteeing that the cumulative utility is at least 2. Comparison between Triple-Q and PRI can be found in Figure 1. Experiment details can be found in the Appendix H.1.

We can observe that PRI converges significantly faster than Triple-Q. Remarkably, both regret and constraint violation level off at the beginning of policy refinement after approximately $110,000$ episodes. However, the regret of Triple-Q continues to increase sublinearly. PRI significantly outperforms Triple-Q on regret. At the end of the $1,100,000$ episodes, Triple-Q has a regret of 2.05×10^5 and constraint violation of -3.86×10^4. In contrast, the regret and constraint violation under PRI are -1.73×10^3 and 1.06×10^4, respectively. Thus, the regret is significantly lower than Triple-Q. Since the full CMDP model is given, we can obtain the optimal solution by using the linear programming approach. The cumulative reward and cumulative utility we get for our learned policy are 1.57301, and 2.00008, which are very close to the optimal solution 1.57306 and 2.



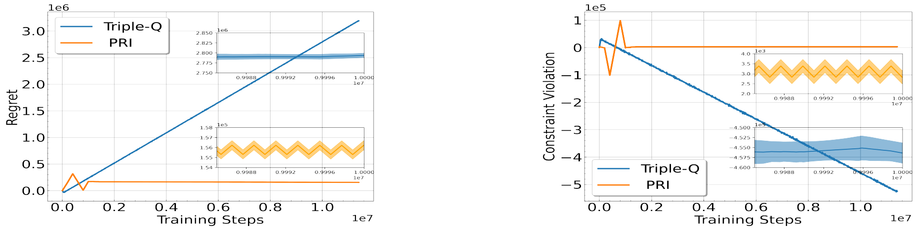
(a) Regret

(b) Constraint Violation

Figure 1: Results for a synthetic CMDP with a unique solution, the shaded region represents the 95% confidence interval.

Grid-world

In our second experiment, which is a grid-world environment (refer to Appendix H.2 for details), we compared Triple-Q with PRI, and the results are shown in Figure 2. This problem has multiple optimal policies. Therefore, we used the extended PRI with multi-solution-pruning. PRI consists of $200,000$ episodes for the initial phase, followed by $200,000$ episodes for each multi-solution pruning phase. Both policy refinement and policy identification phases include $5,000,000$ episodes each. For reference, we ran Triple-Q for the same number of episodes. The outcomes concerning regret and constraint violation are visualized in Figure 2a and 2b. We can observe that Triple-Q has a regret of 3.19×10^6 and a constraint violation of -5.26×10^5, whereas PRI achieves 1.54×10^5 regret and 2.98×10^3 constraint violation, indicating substantially lower regret with PRI.



(a) Regret

(b) Constraint Violation

Figure 2: Results for the grid world environment, the shaded region represents the 95% confidence interval.

8 Conclusions

In this paper, we developed a model-free, regret-optimal algorithm for online CMDPs, called PRI. The algorithm is based on a fundamental observation that for a CMDP with N constraints, there exists an optimal policy that includes at most N stochastic decisions. In the tabular setting, PRI guarantees $\tilde{O}(\sqrt{K})$ regret and constraint violation and the bounds are independent of the size of state and action spaces for sufficiently large K. The same result holds when the violation cannot be canceled across episodes with minor modifications of the algorithm. The details can be found in Appendix G.

REFERENCES

Saminda Wishwajith Abeyruwan, Laura Graesser, David B D'Ambrosio, Avi Singh, Anish Shankar, Alex Bewley, Deepali Jain, Krzysztof Marcin Choromanski, and Pannag R Sanketi. i-sim2real: Reinforcement learning of robotic policies in tight human-robot interaction loops. In *Conference on Robot Learning*, pp. 212–224. PMLR, 2023.

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. In *Advances Neural Information Processing Systems (NeurIPS)*, pp. 20095–20107. Curran Associates Inc., 2020.

Mridul Agarwal, Qinbo Bai, and Vaneet Aggarwal. Markov decision processes with long-term average constraints. *arXiv preprint arXiv:2106.06680*, 2021.

Mridul Agarwal, Qinbo Bai, and Vaneet Aggarwal. Regret guarantees for model-based reinforcement learning with long-term average constraints. In *Uncertainty in Artificial Intelligence*, pp. 22–31. PMLR, 2022.

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

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

Kianté Brantley, Miro Dudik, Thodoris Lykouris, Sobhan Miryoosefi, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings. In *Advances Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 16315–16326. Curran Associates, Inc., 2020.

Archana Bura, Aria HasanzadeZonuzy, Dileep Kalathil, Srinivas Shakkottai, and Jean-Francois Chamberland. Safe exploration for constrained reinforcement learning with provable guarantees. *arXiv preprint arXiv:2112.00885*, 2021.

Liyu Chen, Rahul Jain, and Haipeng Luo. Learning infinite-horizon average-reward markov decision process with constraints. In *Int. Conf. Machine Learning (ICML)*, pp. 3246–3270. PMLR, 2022.

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, volume 130, pp. 3304–3312. PMLR, 2021.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pp. 578–598. PMLR, 2021.

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

Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Provably efficient model-free constrained rl with linear function approximation. In *NeurIPS*, 2022.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for discounted mdps. *Advances in Neural Information Processing Systems*, 34:22288–22300, 2021.

Marius Lindegaard, Hjalmar Jacob Vinje, and Odin Aleksander Severinsen. Intrinsic rewards from self-organizing feature maps for exploration in reinforcement learning. *arXiv preprint arXiv:2302.04125*, 2023.

Pengsen Liu, Jizhe Zhou, and Jiancheng Lv. Exploring the first-move balance point of go-moku based on reinforcement learning and monte carlo tree search. *Knowledge-Based Systems*, 261: 110207, 2023.

Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained MDPs. In *Advances Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Ted Moskovitz, Brendan O'Donoghue, Vivek Veeriah, Sebastian Flennerhag, Satinder Singh, and Tom Zahavy. Reload: Reinforcement learning with optimistic ascent-descent for last-iterate convergence in constrained mdps. *arXiv preprint arXiv:2302.01275*, 2023.

Anand Singh Rajawat, SB Goyal, Chetan Chauhan, Pradeep Bedi, Mukesh Prasad, and Tony Jan. Cognitive adaptive systems for industrial internet of things using reinforcement algorithm. *Electronics*, 12(1):217, 2023.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. *Advances in Neural Information Processing Systems*, 31, 2018.

Rahul Singh, Abhishek Gupta, and Ness B Shroff. Learning in markov decision processes under constraints. *arXiv preprint arXiv:2002.12435*, 2020.

Jerome Taupin, Yassir Jedra, and Alexandre Proutiere. Best policy identification in linear mdps. *arXiv preprint arXiv:2208.05633*, 2022.

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

Honghao Wei, Xin Liu, and Lei Ying. Triple-Q: a model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation. In *Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, 2022a.

Honghao Wei, Xin Liu, and Lei Ying. A provably-efficient model-free algorithm for infinite-horizon average-reward constrained markov decision processes. In *AAAI Conf. Artificial Intelligence*, February 2022b.

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

A Review of Triple-Q

In this section, we briefly review Triple-Q for CMDPs. The design of Triple-Q is based on the primal-dual approach in optimization. The notions used in this section may be slightly abused, but we will make sure the definitions are clear. Consider the case with only one constraint for simplification. Given a Lagrange multiplier λ, we consider the Lagrangian of the problem from a given initial state x_1 :

$$\begin{aligned} & \max_{\pi} V_1^{\pi}(x_1) + \lambda \left(W_1^{\pi}(x_1) - \rho\right) \\ = & \max_{\pi} \mathbb{E}\left[\sum_{h=1}^{H} r_h(x_h, \pi_h(x_h)) + \lambda g_h(x_h, \pi_h(x_h))\right] - \lambda\rho, \end{aligned}$$

which is an unconstrained MDP with reward $r_h(x_h, \pi_h(x_h)) + \lambda g_h(x_h, \pi_h(x_h))$ at step h. Assuming we solve the unconstrained MDP and obtain the optimal policy, denoted by π^*_λ, we can then update the dual variable (the Lagrange multiplier) using a gradient method:

$$\lambda \leftarrow \left(\lambda + \rho - \mathbb{E}\left[W_1^{\pi^*_\lambda}(x_1)\right]\right)^+. \tag{14}$$

While primal-dual is a standard approach, analyzing the finite-time performance, such as regret or sample complexity, is particularly challenging. Triple is designed as a two-time scale algorithm for addressing the trade-off between regret and constraint violations.

- At each step, Triple-Q updates two the Q-values for (x_{h-1}, a_{h-1}) after observing (s_h, a_h), reward $r_h(x_h, a_h)$ and utility $g_h(x_h, a_h)$ in a fast time scale. In particular,

  $$\begin{aligned} Q_{h-1}(x_{h-1}, a_{h-1}) &\leftarrow (1-\alpha_t)Q_{h-1}(x_{h-1}, a_{h-1}) + \alpha_t\left(r_{h-1}(x_{h-1}, a_{h-1}) + V_h(x_h) + b_t\right) \\ C_{h-1}(x_{h-1}, a_{h-1}) &\leftarrow (1-\alpha_t)C_{h-1}(x_{h-1}, a_{h-1}) + \alpha_t\left(g_{h-1}(x_{h-1}, a_{h-1}) + W_h(x_h) + b_t\right) \end{aligned}$$

- at the end of each frame, the virtual queue length is updated in a slow time scale manner as $\left(Z + \rho + \epsilon - \frac{\bar{C}}{K^\alpha}\right)^+$, where $\bar{C}$ is the average of all the Q-values of the utility function at the initial state-action (x_1, a_1).

The algorithm only needs to know the values of H, A, S and K, and no other problem-specific values are needed. Furthermore, Triple-Q includes updates of two Q-functions per step: one for Q_h and one for C_h; and one simple virtual queue update per frame. So its computational complexity is similar to SARSA.

B Proofs of the Technical Lemmas

B.1 Proof of Lemma 1 (Limited Stochasticity)

Lemma 1. *If $q^* = \{q_h^*(x,a)\}_{h,x,a}$ is an optimal solution to the CMDP problem equation 6-equation 10 and is an extreme point, then there are at most $HS + N$ nonzero values in q^*. This implies that the optimal policy derived from q^* includes at most N stochastic decisions.*

Proof. The LP has HSA decision variables $\{q_h(s,a)\}$ in total. So at an extreme point, at least HSA constraints become tight. In other words, at least HSA constraints become equalities under solution q^*. Since there are only $HS + N$ constraints defined in equation 7-equation 9, at least

$$HSA - HS - N = HS(A-1) - N$$

constraints in equation 10 become tight (equality) under q^*. Therefore, there are at least $HS(A-1) - N$ zeros in q^* or at most $HS + N$ nonzero values in q^*.

Now suppose the optimal policy obtained from q^* has less than $HS - N$ greedy decisions. Then q^* would have at least

$$HS - N - 1 + 2(N+1) = HS + N + 1$$

nonzero values because each greedy decision requires one nonzero $q_h(x,a)$ and each stochastic decision requires at least two nonzero $q_h(x,a)$. This leads to a contradiction. □

B.2 Proof of Lemma 2 (Decomposition)

Lemma 2. *Given any Markov policy π and its corresponding occupancy measure q, there exists a set of M greedy policies and a probability distribution $\{a_m\}_{m=1,\cdots,M}$ such that the mixed policy, which selects a greedy policy π^m at the start of an episode with probability a_m and subsequently follows it, has the same occupancy measure q as the original policy π.*

Proof. To simplify the notation, we will prove the lemma for the case where $|\mathcal{D}_{h,x}(q)| \in \{1, 2\}$, i.e., any stochastic decision takes two possible actions and assume $\mathcal{A} = \{0, 1\}$. The extension to the general case is trivial.

Under a Markov policy $\{\pi_h\}_{h=1}^{H}$, the actions are independently chosen given state x and step h. Suppose we will execute the Markov policy for K episodes. We will generate K matrices $\{B_k\}_{k=1}^{K}$ of size $H \times S$ such that $B_k(h, x)$ is a realization of a random variable with distribution $\pi_h(\cdot|x)$. All these values are independently generated. Now to execute policy π at episode k, at state x and step h, the agent takes action a such that $B_k(h, x) = a$. This is statistically the same as sampling an action using $\pi_h(\cdot|x)$ when reaching state x at step h.

We note that each binary matrix B_k corresponds to a greedy policy from the M greedy policies and vice versa. Furthermore, the binary matrix associated with greedy policy π^m is generated with probability

$$\alpha_m = \prod_{h,x} \left(\sum_{a \in \mathcal{D}_{h,x}(q)} \pi_h(a|x) \pi_h^m(a|x) \right),$$

because

$$\sum_{a \in \mathcal{D}_{h,x}(q)} \pi_h(a|x) \pi_h^m(a|x) = \sum_{a \in \mathcal{D}_{h,x}(q)} \pi_h(a|x) \mathbb{I}(\pi_h^m(a|x) = 1),$$

which is the probability that action selected by the greedy policy π^m is also selected under policy π. Therefore, if we consider a mixed policy that chooses policy π^m with probability a_m, then it is statistically the same as policy π and has the same occupancy measure q. □

C Proof of Theorem 1 (Pruning)

In this part, we are going to show the detailed proof of Pruning.

Theorem 1. *Let $\mathcal{D}^* = \{(h, x, a) : \pi_h^*(a|x) > 0\}$ and $\tilde{\mathcal{D}} = \left\{(h, x, a) : \frac{\tilde{N}_h(x,a)}{\sqrt{K}} \geq \frac{\epsilon}{2}\right\}$. Under Assumptions 1 to 3, after policy pruning, we have*

$$\Pr\left(\tilde{\mathcal{D}}_{h,x} = \mathcal{D}_{h,x}(q^*),\ \forall (h, x)\right) = 1 - \tilde{\mathcal{O}}\left(K^{-0.1}\right).$$

Proof. At the end of the first phase, i.e., $\sqrt{K}$ episodes, we consider a mixed policy $\hat{\pi}$ that selects the policy used in the kth episode, π_k, with probability $1/\sqrt{K}$. We assume that all constraints are satisfied under $\hat{\pi}$, which occurs with probability $1 - \mathcal{O}(K^{-2})$. The reward value function of the policy $\hat{\pi}$ is

$$V_1^{\hat{\pi}} = \frac{1}{\sqrt{K}} \sum_{k=1}^{\sqrt{K}} V_1^{\pi_k} \tag{15}$$

and $V^{\pi^*} - V_1^{\hat{\pi}} \geq 0$ because the constraints are satisfied under $\hat{\pi}$. Note that policy $\hat{\pi}$ is not a Markov policy. We next prove that the occupancy measure induced by policy $\hat{\pi}$ is a valid solution to the LP problem:

$$\begin{aligned} \sum_a q_h^{\hat{\pi}}(x, a) &= \sum_a \frac{1}{\sqrt{K}} \sum_{k=1}^{\sqrt{K}} q_h^{\pi_k}(x, a) \\ &= \frac{1}{\sqrt{K}} \sum_{k=1}^{\sqrt{K}} \sum_a q_h^{\pi_k}(x, a) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{K}} \sum_{k=1}^{\sqrt{K}} \sum_{x',a'} q_{h-1}^{\pi_k}(x',a') \mathbb{P}_h(x|x',a') \\
&= \sum_{x',a'} \mathbb{P}_h(x|x',a') \frac{1}{\sqrt{K}} \sum_{k=1}^{\sqrt{K}} q_{h-1}^{\pi_k}(x',a') \\
&= \sum_{x',a'} \mathbb{P}_h(x|x',a') q_{h-1}^{\hat{\pi}}(x',a').
\end{aligned}
$$

Besides, it is easy to verify that $\forall h, x, a, q_h^{\hat{\pi}}(x,a) \geq 0$. Thus the policy $\hat{\pi}$ is a valid policy for the LP problem.

Recall that $\mathcal{D}_{h,x}(q) = \{a : q_h(x,a) > 0\}$. We have $\tilde{\mathcal{D}}_{h,x}(q) = \mathcal{D}_{h,x}(q^*), \ \forall (h,x)$ is equivalent to $\mathcal{D}^* = \tilde{\mathcal{D}}$.

We further define event

$$
\mathcal{E} = \left\{ \exists (h,x,a) \in \mathcal{D}^*, \frac{\tilde{N}_h(x,a)}{\sqrt{K}} < \frac{\epsilon}{2} \right\},
$$

i.e., the event that at the end of the pruning phase, the algorithm eliminates an action used by the optimal policy. Note that π_k's are greedy policies ($\pi_{k,h}(a|x) \in \{0,1\}$). Therefore, we have

$$
\tilde{N}_h(x,a) = \sum_{k=1}^{\sqrt{K}} \pi_{k,h}(a|x).
$$

Thus, assuming this event $\mathcal{E}$ occurs, we can obtain

$$
\begin{aligned}
& q_h^{\hat{\pi}}(x,a) \\
=& \frac{1}{\sqrt{K}} \sum_{k=1}^{\sqrt{K}} q_h^{\pi_k}(x,a) \\
=& \frac{1}{\sqrt{K}} \sum_{k=1}^{\sqrt{K}} \left(\sum_{x',a'} q_{h-1}^{\pi_k}(x',a') \mathbb{P}_h(x|x',a') \right) \pi_{k,h}(a|x) \\
\leq& \frac{1}{\sqrt{K}} \sum_{k=1}^{\sqrt{K}} \pi_{k,h}(a|x) \\
=& \frac{1}{\sqrt{K}} \tilde{N}_h(x,a) \\
<& \frac{\epsilon}{2}.
\end{aligned}
$$

According to Assumption 1, we have

$$
q_h^{\pi^*}(x,a) \geq \epsilon \quad \forall (h,x,a) \in \mathcal{D}^*, \tag{16}
$$

which implies $||q^{\hat{\pi}} - q^*||_1 \geq \frac{\epsilon}{2}$. According to Assumption 2, we have either

$$
V_1^{\pi^*} - V_1^{\hat{\pi}} \geq c_v \frac{\epsilon}{2} \quad \text{(Case 1)} \tag{17}
$$

or

$$
W_1^{\pi^*,n} - W_1^{\hat{\pi},n} \geq c_w \frac{\epsilon}{2} \text{ for some } n \quad \text{(Case 2)}. \tag{18}
$$

Therefore, we have

$$
\begin{aligned}
\Pr(\mathcal{E}) \leq& \Pr\left(\sqrt{K} \left(V_1^{\pi^*} - V^{\hat{\pi}} \right) \geq c_v \frac{\epsilon}{2} \sqrt{K} \right) \\
&+ \Pr\left(\exists n \in [N], \sqrt{K} \left(W_1^{\pi^*,n} - W_1^{\hat{\pi},n} \right) \geq c_w \frac{\epsilon}{2} \sqrt{K} \right).
\end{aligned}
$$

Based on Lemma 3's result on regret and the Markov inequality, we have

$$\Pr\left(\sqrt{K}\left(V_1^{\pi^*} - V^{\hat{\pi}}\right) \geq c_v \frac{\epsilon}{2}\sqrt{K}\right) \leq \frac{c_1 \sqrt{K}^{0.8}}{c_v \frac{1}{2}\epsilon\sqrt{K}} = \frac{2c_1}{c_v \epsilon K^{0.1}}.$$

From Lemma 3's result on constraint violation, the high probability bound, we have

$$\begin{aligned}&\Pr\left(\exists n \in [N], \sqrt{K}\left(W_1^{\pi^*,n} - W_1^{\hat{\pi},n}\right) \geq c_w \frac{\epsilon}{2}\sqrt{K}\right)\\ \leq &\Pr\left(\exists n \in [N], \sqrt{K}\left(\rho^{(n)} - W_1^{\hat{\pi},n}\right) \geq c_w \frac{\epsilon}{2}\sqrt{K}\right)\\ = &\mathcal{O}\left(\frac{1}{K}\right).\end{aligned}$$

Thus for sufficiently large $\sqrt{K}$, $\Pr\left(\mathcal{E}\right) = \mathcal{O}\left(K^{-0.1}\right)$, i.e., with probability $1 - \mathcal{O}\left(K^{-0.1}\right)$, we have

$$\mathcal{D}^* \subseteq \tilde{\mathcal{D}}.$$

Now define event $\mathcal{E}' = \left\{\exists(h, x, a) \notin \mathcal{D}^*, \frac{N_h(x,a)}{\sqrt{K}} \geq \frac{\epsilon}{2}\right\}$. Similar to equation 16 and based on Assumption 3, we can obtain

$$q_h^{\hat{\pi}}(x, a) = \frac{1}{\sqrt{K}} \sum_{k=1}^{\sqrt{K}} q_h^{\pi_k}(x, a) \geq p_{\min} \frac{1}{\sqrt{K}} \tilde{N}_h(x, a) \geq \frac{\epsilon p_{\min}}{2}. \tag{19}$$

Since $q_h^{\pi^*}(x, a) = 0$ for $(h, x, a) \notin \mathcal{D}^*$, $||q^{\hat{\pi}} - q^{\pi^*}||_1 \geq \frac{\epsilon p_{\min}}{2}$. Similar to the analysis on $\mathcal{E}$, we obtain

$$\Pr(\mathcal{E}') = \mathcal{O}\left(K^{-0.1}\right). \tag{20}$$

In other words, with probability $1 - \mathcal{O}\left(K^{-0.1}\right)$, we have

$$\tilde{\mathcal{D}} \subseteq \mathcal{D}^*,$$

which completes the proof. □

D Proof of Theorem 2 (Refinement)

Theorem 2. *Assume $\tilde{\mathcal{D}} = \mathcal{D}^*$ after policy pruning. Under Assumptions 1 and 3, with probability $1 - \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{K}}\right)$, the regret and constraint violation during the policy refinement phase are both $\mathcal{O}\left(H\sqrt{K}\right)$.*

Proof. Recall that in Lemma 2, we have shown that there exists a mixed policy of M greedy policies defined by $\mathcal{D}^*_{h,x}$ that has the same occupancy measure as that under the optimal policy, and $\{\alpha^*_m\}$ are the associated weights.

Recall that the policy refinement consists of $\sqrt{K}$ rounds. Let $\{\alpha_{t,m}\}_{m=1,\cdots,M}$ be the weights used in round t. Then in the tth round, greedy policy π^m is used for $\alpha_{t,m}\sqrt{K}$ episodes, where $\alpha_{t,m}$ is the optimal solution to Decomposition-Opt equation 11.

First, we will bound the estimation errors of the reward and utility value functions. Recall that PRI uses $\epsilon'\sqrt{K}$ episodes in each round to estimate the reward value function and the utility value functions instead of all episodes because $\{\alpha_{t,m}\}$ are random variables correlated with the estimated value functions from the previous round. At the beginning of round t, we have $(t-1)\epsilon'\sqrt{K}$ samples from the previous round. Indexing the samples by k', we have

$$\bar{W}_1^{\pi^m,n} = \frac{\sum_{k'=1}^{(t-1)\epsilon'\sqrt{K}} W_{k',1}^{\pi^m,n}}{(t-1)\epsilon'\sqrt{K}}. \tag{21}$$

Define

$$\delta W_1^{\pi^m,n} = \bar{W}_1^{\pi^m,n} - W_1^{\pi^m,n} \tag{22}$$

$$= \frac{\sum_{k'=1}^{(t-1)\epsilon'\sqrt{K}} \left(W_{k',1}^{\pi^m,n} - W_1^{\pi^m,n} \right)}{(t-1)\epsilon'\sqrt{K}}. \tag{23}$$

Since $W_{k',1}^{\pi^m,n} \in [0, H]$ are i.i.d. random variables, by the Azuma-Hoeffding inequality, we have

$$\Pr\left(\left|\delta W_1^{\pi^m,n}\right| \leq \sqrt{\frac{2H^2 \log\left((t-1)\epsilon'\sqrt{K}\right)}{\epsilon'(t-1)\sqrt{K}}} \right) \tag{24}$$

$$\geq 1 - \frac{2}{(t-1)\epsilon'\sqrt{K}}. \tag{25}$$

Similarly, defining

$$\delta V_1^{\pi^m} = \bar{V}_1^{\pi^m} - V_1^{\pi^m} = \frac{\sum_{k'=1}^{(t-1)\epsilon' K^{\alpha}} \left(V_{k',1}^{\pi^m} - V_1^{\pi^m} \right)}{(t-1)\epsilon'\sqrt{K}},$$

we have

$$\Pr\left(\left|\delta V_1^{\pi^m}\right| \leq \sqrt{\frac{2H^2 \log\left((t-1)\epsilon'\sqrt{K}\right)}{\epsilon'(t-1)\sqrt{K}}} \right) \tag{26}$$

$$\geq 1 - \frac{2}{(t-1)\epsilon'\sqrt{K}}. \tag{27}$$

Therefore, with probability at least $1 - \frac{2M}{(t-1)\epsilon'\sqrt{K}}$,

$$\begin{aligned}
\left| \sum_{m=1}^{M} \alpha_m^* \bar{W}_1^{\pi_m,n} - \rho^{(n)} \right| &= \left| \sum_{m=1}^{M} \alpha_m^* \left(W_1^{\pi_m,n} + \delta W_1^{\pi_m,n} \right) - \rho^{(n)} \right| \\
&\leq \sum_{m=1}^{M} \alpha_m^* \left|\delta W_1^{\pi_m,n}\right| \\
&\leq \sqrt{\frac{2H^2 \log\left((t-1)\epsilon'\sqrt{K}\right)}{\epsilon'(t-1)\sqrt{K}}}.
\end{aligned}$$

In other words, $\{a_m^*\}$ is a feasible solution to Decomposition-Opt equation 11 with a high probability, which implies that Decomposition-Opt equation 11 has a solution with a high probability.

We now consider $\{a_{t,m}\}$ and the regret and constraint violation in round t. If $\{a_m^*\}$ is a feasible solution to Decomposition-Opt equation 11, then

$$\begin{aligned}
&\sum_{m=1}^{M} \alpha_{t,m} \sqrt{K} V_1^{\pi^m} \\
=&\sum_{m=1}^{M} \alpha_{t,m} \sqrt{K} \left(\bar{V}_1^{\pi^m} - \delta V_1^{\pi^m} \right) \\
\geq&\sqrt{K} \left(\sum_{m=1}^{M} \alpha_{t,m} \bar{V}_1^{\pi^m} \right) - \sqrt{K} \max_m \left|\delta V_1^{\pi^m}\right|
\end{aligned}$$

$$
\begin{aligned}
&\geq_{(a)}\sqrt{K}\left(\sum_{m=1}^{M}\alpha_m^*\left(V_1^{\pi^m}+\delta V_1^{\pi^m}\right)\right)-\sqrt{K}\max_m|\delta V_1^{\pi^m}|\\
&\geq\sqrt{K}\left(\sum_{m=1}^{M}\alpha_m^*V_1^{\pi^m}\right)-2\sqrt{K}\max_m|\delta V_1^{\pi^m}|\\
&=\sqrt{K}V_1^{\pi^*}-2\sqrt{K}\max_m|\delta V_1^{\pi^m}|\\
&\geq\sqrt{K}\left(V_1^{\pi^*}-2\sqrt{\frac{2H^2\log\left((t-1)\epsilon'\sqrt{K}\right)}{\epsilon'(t-1)\sqrt{K}}}\right),
\end{aligned}
$$

where (a) holds because $\{a_{t,m}\}_m$ is the optimal solution to Decomposition-Opt equation 11. In other words, with a high probability, the regret is bounded by

$$
2\sqrt{\frac{2\sqrt{K}H^2\log\left((t-1)\sqrt{K}\right)}{\epsilon'(t-1)}}. \tag{28}
$$

Thus, with probability

$$
\prod_{t=2}^{\sqrt{K}}\left(1-\frac{2}{(t-1)\epsilon'\sqrt{K}}\right)\geq 1-\frac{2\log K}{\epsilon'\sqrt{K}} \tag{29}
$$

regret in round t is bounded by equation 28 for all t. Therefore, the regret during policy refinement is bounded by

$$
\begin{aligned}
&2H\sqrt{K}+\sum_{t=3}^{\sqrt{K}}2\sqrt{\frac{2\sqrt{K}H^2\log\left((t-1)\sqrt{K}\right)}{\epsilon'(t-1)}}\\
\leq&2H\sqrt{K}+2\sqrt{\frac{2KH^2\log K}{\epsilon'}}\sum_{t=3}^{\sqrt{K}}\sqrt{\frac{1}{(t-1)\sqrt{K}}}\\
\leq&2H\sqrt{K}+2\sqrt{\frac{2\sqrt{K}H^2\log K}{\epsilon'}}\int_{t=1}^{\sqrt{K}}\sqrt{\frac{1}{t}}dt\\
\leq&2H\sqrt{K}+2\sqrt{\frac{2KH^2\log K}{\epsilon'}}\\
=&\mathcal{O}(H\sqrt{K}).
\end{aligned}
$$

The analysis is the same for the constraint violation. □

E PROOF OF THEOREM 3 (IDENTIFICATION)

Theorem 3. *Assume $\tilde{\mathcal{D}}=\mathcal{D}^*$ after policy pruning. Under Assumptions 1 and 3, with probability $1-\tilde{\mathcal{O}}\left(\frac{1}{K}\right)$, the regret and constraint violation during the policy identification phase are both $\mathcal{O}\left(\sqrt{K}\right)$. Furthermore, $|\tilde{\pi}_h(a|x)-\pi_h^*(a|x)|=\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$ if $0<\pi_h^*(a|x)<1$ and $\tilde{\pi}_h(a|x)=\pi_h^*(a|x)|$ if $\pi_h^*(a|x)\in\{0,1\}$.*

Proof. Consider the $\{a_m\}$ obtained at the end of the refinement phase, and the mixed policy $\hat{\pi}$ defined by $\{a_m\}$. According to the proof of Theorem 2, we have with probability $1-\mathcal{O}(K^{-1})$,

$$
V_1^{\hat{\pi}}=\sum_{m=1}^{M}\alpha_mV_1^{\pi^m}
$$

$$\geq \left(V_1^{\pi^*} - 2\sqrt{\frac{H^2 \log\left((t-1)\epsilon' K\right)}{\epsilon' K}} \right) \tag{30}$$

$$W_1^{\hat{\pi},n} = \sum_{m=1}^{M} \alpha_m W_1^{\pi^m,n}$$

$$\geq \left(W_1^{\pi^*,n} - 2\sqrt{\frac{H^2 \log\left((t-1)\epsilon' K\right)}{\epsilon' K}} \right) \quad \forall n. \tag{31}$$

Therefore, the regret and constraint violation during the identification phase, which includes K episodes, are both $\mathcal{O}(H\sqrt{K})$.

When both equation 30 and equation 31 hold, under Assumption 2, we have

$$\|q^{\hat{\pi}} - q^{\pi^*}\|_1 = \mathcal{O}(1/\sqrt{K}).$$

For any (h, x, a) such that $0 < \pi_h^*(x|a) < 1$,

$$\mathbb{E}\left[\sum_{k=1}^{a_m K} \mathbb{I}(x_{k,h} = x, a_{k,h} = a)\right] = \alpha_m q_h^{\pi^m}(x, a)K,$$

which implies that

$$\Pr\left(\left|\sum_{k=1}^{a_m K} \mathbb{I}_{(x_{k,h}=x, a_{k,h}=a)} - \alpha_m q_h^{\pi^m}(x, a)K\right| \leq \sqrt{K \log K}\right)$$
$$= 1 - \mathcal{O}\left(\frac{1}{K}\right)$$

according to the Azuma-Hoeffding inequality. Define event

$$\Phi = \left\{ \left| \frac{\sum_{k=1}^{K} \mathbb{I}_{(x_{k,h}=x, a_{k,h}=a)}}{K} - \sum_{m} \alpha_m q_h^{\pi^m}(x, a) \right| \leq M\sqrt{\frac{\log K}{K}} \right\}. \tag{32}$$

We have

$$\Pr\left(\Phi\right) = 1 - \mathcal{O}\left(\frac{1}{K}\right).$$

Define $\tilde{q}_h(x, a) = \frac{N_h(x,a)}{K}$, which is the empirical occupant measure under policy $\hat{\pi}$. Note that

$$N_h(x, a) = \sum_{k=1}^{K} \mathbb{I}(x_{k,h} = x, a_{k,h} = a)$$

and

$$q_h^{\hat{\pi}}(x, a) = \sum_{m} \alpha_m q_h^{\pi^m}(x, a).$$

Therefore, we have with probability $1 - \mathcal{O}(1/K)$,

$$\|\tilde{q} - q^{\pi^*}\|_1 = \mathcal{O}(1/\sqrt{K}),$$

which implies that

$$\|\tilde{\pi} - \pi^*\|_1 = \mathcal{O}(1/\sqrt{K}).$$

Furthermore, since $\tilde{\mathcal{D}} = \mathcal{D}^*$, we immediately have $\tilde{\pi}_h(a|x) = \pi_h^*(a|x)|$ if $\pi_h^*(a|x) \in \{0, 1\}$. □

F Extension to CMDPs with multiple optimal solutions

When the optimal solution to the CMDP is not unique, or the RL agent does not know whether the CMDP has a unique solution or not, the agent adds Multi-Solution Pruning after the pruning phase in PRI to keep one and only one optimal policy belonging to $\Pi^{*,e}$. Recall that after the pruning phase, the action space for state x and step h, denoted by $\mathcal{A}_{h,x}$, is limited to $\mathcal{A}_{h,x} = \tilde{D}_{h,x}$. The key idea of the multi-solution pruning algorithm is to evaluate each stochastic decision (h', x') such that $|\mathcal{A}_{h',x'}| > 1$. The algorithm first decides whether some of the actions in $\mathcal{A}_{h',x'}$ can be removed, e.g., a', while retaining at least one optimal policy with the following action space:

$$\otimes_{(h,x)\neq(h',x')}\mathcal{A}_{h,x} \otimes (\mathcal{A}_{h',x'} \setminus \{a'\}).$$

This is done by running Triple-Q with the above action space for $K^{0.25}$ episodes. If the regret is small, then with a high probability, at least one of the optimal policies is retained so we can remove action a' from $\mathcal{A}_{h',x'}$. The detailed algorithm is presented in Algorithm 2.

If the regret is large, then any optimal policy in $\otimes_{(h,x)}\mathcal{A}_{h,x}$ has to use action a' in state x' at step h'. Mulit-Solution Pruning next determines whether using a' alone is sufficient, i.e., whether an optimal policy is retained in the following action space

$$\otimes_{(h,x)\neq(h',x')}\mathcal{A}_{h,x} \otimes (\mathcal{A}_{h',x'} = \{a'\}).$$

This is again done by running Triple-Q with the above action space for $K^{0.25}$ episodes. If the regret is small, then with a high probability, one optimal policy takes a greedy decision at (h', x') with action a'; otherwise, the algorithm keeps a' in $\mathcal{A}_{h',x'}$ and moves to a different action in $\mathcal{A}_{h',x'}$. Note that we use Triple-Q for $K^{0.25}$ episodes each time, instead of $\sqrt{K}$ episodes, because it is easier to learn whether an optimal policy exists than learning the actual optimal policy.

Algorithm 2 Multi-Solution Pruning

```
Set $v^*$ to be the average cumulative reward received over the $\sqrt{K}$ episodes under the policy pruning phase of PRI.
Set flag$(h,x) \leftarrow 0$ for all $(h,x)$ such that $|\tilde{\mathcal{D}}_{h,x}| > 1$.
while $\exists(h,x)$ such that $|\tilde{\mathcal{D}}_{h,x}| > 1$ and flag$(h,x) = 0$ do
    Select $(h', x')$ such that $|\tilde{\mathcal{D}}_{h',x'}| > 1$ and flag$(h', x') = 0$ with the smallest $h'$. Ties are broken arbitrarily.
    flag$(h', x') \leftarrow 1$
    for $a' \in \tilde{\mathcal{D}}_{h',x'}$ do
        Reset Triple-Q and run it for $K^{0.25}$ episodes with $\tilde{D}_{h,x}$ $((h,x) \neq (h',x'))$ and $\tilde{D}_{h',x'}\setminus\{a'\}$ as the action spaces while counting $\tilde{N}_h(x,a)$ as in policy pruning. Record the average cumulative reward $\tilde{v}$ and average cumulative utilities $\tilde{w}^n$.
        if $v^* - \tilde{v} \leq \frac{2}{K^{0.03}}$ and $\tilde{w}^n \geq \rho^{(n)}$ for all $n$ then
            Update $\tilde{\mathcal{D}}_{h,x} = \left\{a : \frac{\tilde{N}_h(x,a)}{K^{0.25}} \geq \frac{\epsilon}{2}\right\}$ for all $(h,x)$.
        else
            Run Triple-Q for $K^{0.25}$ episodes with action space $\tilde{D}_{h,x}$ for $(h,x) \neq (h',x')$ and $\{a'\}$ for $(h', x')$. Record the average cumulative reward $\tilde{v}$
            if $v^* - \tilde{v} \leq \frac{2}{K^{0.03}} + 2\sqrt{\frac{2H^2 \log K^{0.25}}{K^{0.25}}}$ and $\tilde{w}^n \geq \rho^{(n)}$ for all $n$ then
                Update $\tilde{\mathcal{D}}_{h,x} = \left\{a : \frac{\tilde{N}_h(x,a)}{K^{0.25}} \geq \frac{\epsilon}{2}\right\}$ for all $(h,x)$.
```

F.1 Proof of Theorem 4 (multiple-solution pruning)

Theorem 4. *Under Assumption 4 and 5, with probability* $1 - \mathcal{O}(1/K^{0.02})$, *for sufficiently large* K, *multi-solution pruning outputs a unique optimal policy with at most* N *stochastic decisions. The regret and constraint violation during multi-solution pruning are bounded by* $H^2SAK^{0.25}$ *with probability one.*

Proof. We first consider the $K^{0.25}$ episodes after a' is removed from $\mathcal{A}_{h',x'}$. Consider the case that there still exists an optimal policy after removing the action. In this case, we will show that $v^* - \tilde{v} \leq \frac{2}{K^{0.03}}$ with a high probability. Define

$$\bar{V}_1 = \frac{1}{K^{0.25}} \sum_{k=1}^{K^{0.25}} V_1^{\pi_k},$$

where π_k is the policy used in the kth episode.

Note that

$$v^* - \tilde{v} = v^* - V_1^{\pi^*} + V_1^{\pi^*} - \bar{V}_1 + \bar{V}_1 - \tilde{v}.$$

We next bound the three terms $v^* - V_1^{\pi^*}$, $V_1^{\pi^*} - \bar{V}_1$, and $\bar{V}_1 - \tilde{v}$ individually.

Let $v_{k,1}$ be the cumulative reward received in episode k and $V_1^{\pi_k}$ be the reward value function. Note that

$$X_\tau = \sum_{k=1}^{\tau} (v_{k,1} - V_1^{\pi_k})$$

is a Martingale. By Azuma's inequality, we have

$$\Pr\left(\left|\tilde{v} - \bar{V}_1\right| \leq \sqrt{\frac{2H^2 \log K^{0.25}}{K^{0.25}}}\right) \geq 1 - \frac{1}{2K^{0.25}}. \tag{33}$$

A similar argument yields that

$$\Pr\left(\left|v^* - \bar{V}_1^{\pi^*}\right| \leq \sqrt{\frac{2H^2 \log K^{0.5}}{K^{0.5}}}\right) \geq 1 - \frac{1}{2K^{0.5}}. \tag{34}$$

We next bound $V_1^{\pi^*} - \bar{V}_1$ based on Lemma 3 and the Markov inequality. First, based on the Markov inequality, we have

$$\Pr\left(V_1^{\pi^*} - \bar{V}_1 \geq K^{-0.03} \middle| V_1^{\pi^*} \geq \bar{V}_1\right) \leq \frac{\mathbb{E}\left[V_1^{\pi^*} - \bar{V}_1 \middle| V_1^{\pi^*} \geq \bar{V}_1\right]}{K^{-0.03}}.$$

Note that we have

$$\begin{aligned}
\mathbb{E}\left[V_1^{\pi^*} - \bar{V}_1 \middle| V_1^{\pi^*} \geq \bar{V}_1\right] =& \frac{\mathbb{E}\left[V_1^{\pi^*} - \bar{V}_1\right] - \Pr\left(V_1^{\pi^*} < \bar{V}_1\right) \mathbb{E}\left[V_1^{\pi^*} - \bar{V}_1 \middle| V_1^{\pi^*} < \bar{V}_1\right]}{\Pr\left(V_1^{\pi^*} \geq \bar{V}_1\right)} && (35)\\
\leq& \frac{\frac{c_1 K^{0.2}}{K^{0.25}} + H \Pr\left(V_1^{\pi^*} < \bar{V}_1\right)}{1 - \Pr\left(V_1^{\pi^*} < \bar{V}_1\right)} && (36)
\end{aligned}$$

and

$$\begin{aligned}
& \Pr\left(V_1^{\pi^*} - \bar{V}_1 \geq K^{-0.03}\right) && (37)\\
=& \Pr\left(V_1^{\pi^*} - \bar{V}_1 \geq K^{-0.03} \middle| V_1^{\pi^*} \geq \bar{V}_1\right) \Pr\left(V_1^{\pi^*} \geq \bar{V}_1\right) && (38)\\
& + \Pr\left(V_1^{\pi^*} - \bar{V}_1 \geq K^{-0.03} \middle| V_1^{\pi^*} < \bar{V}_1\right) \Pr\left(V_1^{\pi^*} < \bar{V}_1\right) && (39)\\
\leq& \Pr\left(V_1^{\pi^*} - \bar{V}_1 \geq K^{-0.03} \middle| V_1^{\pi^*} \geq \bar{V}_1\right) \left(1 - \Pr\left(V_1^{\pi^*} < \bar{V}_1\right)\right) + \Pr\left(V_1^{\pi^*} < \bar{V}_1\right) && (40)\\
=& K^{0.03} \left(\frac{c_1 K^{0.2}}{K^{0.25}} + H \Pr\left(V_1^{\pi^*} < \bar{V}_1\right)\right) + \Pr\left(V_1^{\pi^*} < \bar{V}_1\right). && (41)
\end{aligned}$$

Note that when the constraints are satisfied, we have $V_1^{\pi^*} \geq \bar{V}_1$. Therefore, according Lemma 3, $\Pr\left(V_1^{\pi^*} < \bar{V}_1\right) = \mathcal{O}(K^{-0.5})$, which implies that

$$\Pr\left(V_1^{\pi^*} - \bar{V}_1 \geq K^{-0.03}\right) = \mathcal{O}\left(K^{-0.02}\right). \tag{42}$$

Combining the inequality above with inequalities equation 33 and equation 34, we can conclude that

$$\Pr\left(v^* - \tilde{v} \leq 2K^{-0.03}\right) = 1 - \mathcal{O}\left(K^{-0.02}\right). \tag{43}$$

Based on Lemma 3's result on constraint violation, we obtain

$$\Pr\left(v^* - \tilde{v} \leq 2K^{-0.03}, \tilde{w}^n \geq \rho^{(n)}\ \forall n\right) = 1 - \mathcal{O}\left(K^{-0.02}\right), \tag{44}$$

On the other hand, if no optimal policy exists after removing a', then we have $\forall \pi \in \Pi^{*,e}$, $q^{\pi}_{h'}(x',a') \geq \epsilon$. Let π'' be an optimal policy with action spaces

$$\otimes_{(h,x)\neq(h',x')}\mathcal{A}_{h,x} \otimes (\mathcal{A}_{h',x'} \setminus \{a'\}),$$

and suppose all constraints are satisfied under π''. Note that π'' is *not* an optimal policy for the original problem. We first have

$$v^* - \tilde{v} = v^* - V_1^{\pi^*} + V_1^{\pi^*} - V_1^{\pi''} + V_1^{\pi''} - \bar{V}_1 + \bar{V}_1 - \tilde{v}$$

Based on Assumptions 4 and 5, we have

$$\begin{aligned} V_1^{\pi^*} - V_1^{\pi''} &\geq c_v||q^{\pi^*} - q^{\pi''}||_1 \\ &\geq c_v|q^{\pi^*}_{h'}(x',a') - q^{\pi''}_{h'}(x',a')| \\ &\geq c_v\epsilon. \end{aligned}$$

This holds because all constraints are satisfied under π'' under our assumption. Similar to the first case, we have

$$\Pr\left(\left|\tilde{v} - \bar{V}_1\right| \leq \sqrt{\frac{2H^2\log K^{0.25}}{K^{0.25}}}\right) \geq 1 - \frac{1}{2K^{0.25}}. \tag{45}$$

and

$$\Pr\left(\left|v^* - \bar{V}_1^{\pi^*}\right| \leq \sqrt{\frac{2H^2\log K^{0.5}}{K^{0.5}}}\right) \geq 1 - \frac{1}{2K^{0.5}}. \tag{46}$$

Furthermore, according to Lemma 3,

$$\Pr\left(V_1^{\pi''} - \bar{V}_1 \geq 0\right) \geq 1 - \mathcal{O}\left(\frac{1}{K^{0.5}}\right) \tag{47}$$

because the constraint is satisfied with probability $1 - \mathcal{O}\left(\frac{1}{K^{0.5}}\right)$.

Summarizing the results above, using the union bound, we conclude that with probability $1 - \mathcal{O}\left(K^{-0.02}\right)$, we have

$$v^* - \tilde{v} \geq c_v\epsilon - 2\sqrt{\frac{2H^2\log K^{0.25}}{K^{0.25}}} > 2K^{-0.03}$$

for sufficiently large K if an optimal policy is not retained. If none of the policies formed by action spaces

$$\otimes_{(h,x)\neq(h',x')}\mathcal{A}_{h,x} \otimes (\mathcal{A}_{h',x'} \setminus \{a'\})$$

can satisfy the constraints, then it can be easily shown that $\tilde{w}^n < \rho^{(n)}$ with probability $1-\mathcal{O}(K^{-0.25})$ for some n.

If a' is deemed to be necessary, Mulit-Solution Pruning next determines whether using a' alone is sufficient, i.e., can stochastic decision (h',x') become greedy without losing optimality? The algorithm runs Triple-Q with action space $\{a'\}$ for (h',x'). If there exists an optimal policy π with $\pi_{h'}(a'|x') = 1$, then following the same analysis above, we have with probability at least $1 - \mathcal{O}\left(K^{-0.02}\right)$,

$$v^* - \tilde{v} \leq \frac{2}{K^{0.03}}.$$

Otherwise, according to the Assumption 4, $\forall \pi^* \in \Pi^{*,e}$, there exists another action $a'' \neq a'$ such that $q^{\pi^*}_{h'}(x',a'') \geq \epsilon$. Because any optimal policy can be represented as a linear combination of those policies in $\Pi^{*,e}$, we have that for any optimal policy π^*, $\sum_{a\neq a'} q^{\pi^*}_{h'}(x',a) \geq \epsilon$. Letting π'' be an optimal policy with action spaces

$$\otimes_{(h,x)\neq(h',x')}\mathcal{A}_{h,x} \otimes (\mathcal{A}_{h',x'} = \{a'\}),$$

which satisfies all constraints, we have

$$\sum_{a \neq a'} q_{h'}^{\pi^*}(x', a) - q_{h'}^{\pi''}(x', a) \geq \epsilon.$$

Thus, according to Assumption 5,

$$V_1^{\pi*} - V_1^{\pi''} \geq c_v ||q^{\pi*} - q^{\pi''}||_1 \geq c_v \epsilon$$

because π'' satisfies all constraints.

The remaining analysis is identical to case when a' is removed from the action space $\mathcal{A}_{h',x'}$. If none of the policies formed by action spaces

$$\otimes_{(h,x)\neq(h',x')} \mathcal{A}_{h,x} \otimes (\mathcal{A}_{h',x'} = \{a'\})$$

can satisfy the constraints, then it can be easily shown that $\tilde{w}^n < \rho^{(n)}$ with probability $1 - \mathcal{O}(K^{-0.25})$ for some n.

After the algorithm goes through all action space $\mathcal{A}_{h,x}$, with probability $1 - \mathcal{O}(1/K^{0.02})$, we obtain action space

$$\otimes_{(h,x)} \mathcal{A}_{h,x}$$

such that none of the stochastic decision can be reduced to a greedy decision without losing optimality. Since any optimal policy can be written as a linear combination of optimal policies associated with extreme points, and any combination of two optimal policy only increases the number of stochastic decisions. Therefore, we conclude that the optimal policy induced by

$$\otimes_{(h,x)} \mathcal{A}_{h,x}$$

is an extreme point and is unique. Besides, it is easy to verify that the regret and constraint violation of multiple solution pruning are bounded by $H^2SAK^{0.25}$ because it takes at most $HSAK^{0.25}$ episodes to finish the algorithm. □

G Extension to Constraint Violation without Episode-Wise Cancellation

While the constraint violation defined in (5) allows the cancellation across episodes, we can guarantee $O(\sqrt{K} \log K)$ violation when the cancellation is not allowed, as defined in (5) in Efroni et al. (2020), by making some minor modification to the algorithm. In particular, in the policy refinement and identification phases, instead of using the M greedy policy in a round-robin fashion, we can use a mixed policy that chooses policy m with probability α_m. With that modification, we will have

- The Pruning phase consists at most $\sqrt{K}$ episodes, so resulting in at most $O(\sqrt{K})$ violation.
- During the refinement phase, based on inequality (27), we have with a high probability that the mixed policy used in the tth iteration has a regret bounded by $O\left(\frac{\log K}{(t-1)\sqrt{K}}\right)$ for $t \geq 2$. The same result holds for constraint violation, i.e., with a high probability, the constraint violation of the policy used in the tth iteration is bounded by $O\left(\frac{\log K}{(t-1)\sqrt{K}}\right)$. The tth iteration includes $\sqrt{K}$ episodes and the refinement phase includes $\sqrt{K}$ iterations. Therefore, with a high probability, the total violation without canceling across episodes is $O(\sqrt{K} \log K)$.
- The mixed policy used in the policy identification phase has a constraint violation bound of $O\left(\frac{\log K}{\sqrt{K}}\right)$. The policy is used for K episodes, so the violation is bounded by $O(\sqrt{K} \log K)$.

Therefore, the total violation is bounded by $\tilde{O}(\sqrt{K})$ by using definition (5) in Efroni et al. (2020).

H Simulations

H.1 Synthetic CMDP

In the systehtic CMDP, we choose $|\mathcal{S}| = 3$, $|\mathcal{A}| = 3$, $H = 3$. The detailed parameters of the CMDP in the first experiment are shown in Table 2, 3 and 4.

We executed the pruning phase (Triple-Q) over $100,000$ episodes, followed by the refinement phase over $1,000,000$ episodes. Since the problem has only one constraint, the optimal policy has only one stochastic decision, which can be decided by evaluating the frequencies of two greedy policies. Thus, phase 3 is not necessary for this specific environment.

Table 2: Transition Kernels (the rows represent (previous state, action) and the columns represent (step, next state)).

	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)	(3,1)	(3,2)	(3,3)
(1,1)	0.3112981	0.35107633	0.27041442	0.42626645	0.04822746	0.14663183	0.4031534	0.19783729	0.39831431
(1,2)	0.23314339	0.32491141	0.48360071	0.24246185	0.19021328	0.43972054	0.26457139	0.21435897	0.26256243
(1,3)	0.45555851	0.32401226	0.24598487	0.3312717	0.76155926	0.41364763	0.33227521	0.58780374	0.33912326
(2,1)	0.32676574	0.35320112	0.1300059	0.35453348	0.32114495	0.40817113	0.1762648	0.30097191	0.48437535
(2,2)	0.11092341	0.28034838	0.45655888	0.23441632	0.2847394	0.235718	0.17239783	0.37273618	0.08000908
(2,3)	0.56231085	0.3664505	0.41343525	0.4110502	0.39411565	0.35611087	0.65133738	0.32629191	0.43561556

Table 3: Rewards (the rows represent (state, action) and the columns represent step.)

	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)	(3,1)	(3,2)	(3,3)
1	0.5507979	0.70814782	0.29090474	0.51082761	0.89294695	0.89629309	0.12558531	0.20724388	0.0514672
2	0.44080984	0.02987621	0.45683322	0.64914405	0.27848728	0.6762549	0.59086282	0.02398188	0.55885409
3	0.25925245	0.4151012	0.28352508	0.69313792	0.44045372	0.15686774	0.54464902	0.78031476	0.30636353

Table 4: Utilities (the rows represent (state, action) and the columns represent step.)

	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)	(3,1)	(3,2)	(3,3)
1	0.22195788	0.38797126	0.93638365	0.97599542	0.67238368	0.90283411	0.84575087	0.37799404	0.09221701
2	0.6534109	0.55784076	0.36156476	0.2250545	0.40651992	0.46894025	0.26923558	0.29179277	0.4576864
3	0.86053391	0.5862529	0.28348786	0.27797751	0.45462208	0.20541034	0.20137871	0.51403506	0.08722937

H.2 Grid World

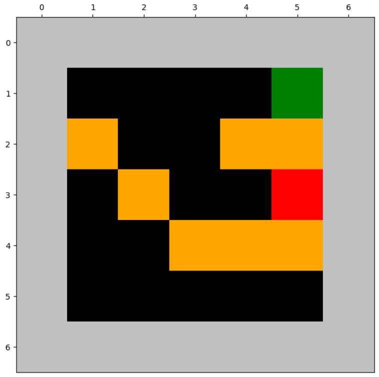


Figure 3: Grid World

As shown in Figure 3, the task of the agent is to go from the red grid point to the green grid point. The black grid points are the *safe* points over which the agent can move, and the yellow grid points are obstacles. Moving over an obstacle incurs a penalty of one. The constraint is that the agent can incur only an average cost of 0.5 or less. The agent can take six steps at maximum. The reward associated with reaching the destination is 1, and the rewards for other locations, after six steps, are the Euclidean distance from the location to the destination (normalized by the longest distance). At each grid point, the agent has five actions to choose from: up, down, left, right, and stay, except at the boundary. The goal is to maximize the reward subject to the constraint.

During the experiment, we observed that policy pruning is much more efficient than the theoretical worst case. For this specific environment, the optimal policy should have $6 \times 5 \times 5 + 1 = 155$ nonzero $\pi_h^*(a|x)$'s. After the first phase (Triple-Q), we have roughly 200 (step, state, action) triples (here "roughly" considers the difference among different trials with different random seeds), associated with stochastic decisions to check and prune. Except for the two “necessary” decisions, which are stochastic decisions in the optimal policy, for all trials, the algorithm only checked two candidate triples and eliminated the rest candidate triples in the process.