PLUG-IN SAMPLE COMPLEXITY FOR CONSTRAINED REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We present a novel plug-in approach for constrained reinforcement learning that achieves the sample complexity of $\tilde{O}\left(\frac{SAH^4}{\epsilon^2\zeta^2}\right)$ using a generative model. Unlike previous specialized algorithms, our method is general: it requires only black-box access to an optimization oracle that solves the empirical CMDP. The core of our approach is a reward perturbation technique that guarantees the oracle's solution is valid for the original problem.

1 INTRODUCTION

A central problem in reinforcement learning is to find a near-optimal policy for a Markov Decision Process (MDP) with the minimum possible number of samples. We consider this problem in the fundamental *generative model* setting, where an algorithm can query a simulator for any state-action pair to receive a sample of the next state and reward. This setting isolates the core statistical challenge of learning from finite data, abstracting away the difficulties of exploration. For unconstrained MDPs, the sample complexity of learning a near-optimal policy of this problem is now tightly understood to be $\tilde{\Theta}\left(\frac{SA}{(1-\gamma)^3\epsilon^2}\right)$ (Azar et al., 2012; Agarwal et al., 2020a; Li et al., 2024), where $\gamma \in (0, 1)$ is the discount factor, S is the number of states and A is the number of actions for a state.

Among the simplest and most natural algorithms for this task is the **plug-in approach** (also known as empirical risk minimization). This method consists of two distinct stages: first, use the generative model to build an empirical estimate of the transition probabilities and rewards, and second, compute the optimal policy for this empirical model using a black-box planning oracle, as if the estimate were the true system. The appeal of this approach is its striking simplicity and modularity. In the unconstrained setting, it has been proven that this simple plug-in method is, in fact, minimax optimal (Agarwal et al., 2020a; Li et al., 2024), establishing a powerful, general-purpose paradigm.

This paper investigates whether this elegant plug-in paradigm can be extended to the more complex setting of Constrained Markov Decision Processes (CMDPs). In a CMDP, an agent must optimize a reward function while satisfying constraints on one or more auxiliary cost functions. Despite its conceptual appeal, the naive plug-in approach for CMDPs is not known whether it works. The issue is one of feasibility: a black-box planner, oblivious to statistical uncertainty, can easily exploit estimation errors in the empirical model to find a policy that satisfies constraints within the model, but catastrophically violates them in the true, underlying system. A fundamental challenge in CMDPs, in contrast to unconstrained MDPs, is the non-uniqueness of the optimal value function.. Due to this brittleness, state-of-the-art, sample-optimal algorithms for CMDPs have relied on specialized solvers that tightly integrate the estimation and planning phases (HasanzadeZonuzy et al., 2021; Bai et al., 2021; Wei et al., 2021), sacrificing the modularity of the plug-in approach.

This work resolves the tension between the simplicity of the plug-in approach and the demands of constrained optimization. We demonstrate that a simple modification—a carefully designed **reward perturbation** applied to the empirical model—is sufficient to make the plug-in approach both robust and minimax optimal for CMDPs. Our main result shows that after applying this perturbation, any black-box CMDP oracle will return a policy that is provably near-optimal and feasible for the true environment. This restores the plug-in paradigm for the constrained setting, showing that a clean decoupling of learning and planning is indeed possible.

The theoretical foundation for our result is a **novel statistical decoupling argument for CMDPs**. We provide a tight **characterization of the covering number of the space of CMDP value functions, showing it is polynomial**, which allows us to derive a powerful uniform convergence guarantee. We also prove that our perturbation is not merely a proof artifact but is fundamentally necessary, by constructing a simple CMDP instance where the unperturbed plug-in approach is guaranteed to fail. Our contributions include:

- **A General Plug-in Framework**: We present the first plug-in method for CMDPs that achieves the near optimal sample complexity of $\tilde{O}\left(\frac{SAH^4}{\epsilon^2\zeta^2}\right)$ using a generative model, accommodating any black-box CMDP optimization oracle.
- **A Novel Perturbation Technique**: We design and analyze a reward perturbation scheme that robustifies the planning oracle against statistical uncertainty.

Theorem 1. *Assume the Slater's condition number $\check{\zeta}$ (see Definition 1) satisfies that $\check{\zeta}H \geq 20\zeta\epsilon$. With probability $1-\delta$, Algorithm 1 outputs an ϵ-strictly-optimal policy by using $\tilde{O}\left(\frac{SAH^4}{\epsilon^2\zeta^2}\right)$ samples.*

Organization. In Section 2, we introduce basics of CMDP. We present the algorithm and summary technique in Section 3, and provide the proof of Theorem 1 in Section 4.

1.1 Related Works

Constrained Markov Decision Process . Constrained Markov Decision Process (CMDP) (Altman, 2021) is a standard model for addressing safety concerns in reinforcement learning (RL). Many prior works on CMDPs employ a primal-dual approach (Paternain et al., 2019; Ding et al., 2021) to achieve sublinear regret while ensuring bounded constraint violations, though policy gradient algorithms are also studied (Tessler et al., 2018). Efroni et al. (2020) introduces a more stringent metric for hard constraint violation, where only positive constraint violations are accumulated. Efroni et al. (2020) introduces an algorithm that achieves sublinear regret, constraint violations and hard constraint violation.

Plug-in Approaches in Standard RL. In the plug-in solver approach of RL, an empirical model is built by first drawing samples by querying a simulator, and plans in this empirical model via an arbitrary plug-in solver. Due to elegance and flexibility, such approach has been extensively studied in the theory RL. Notable contributions include plug-in solver approach for finite-horizon tabular RL (Agarwal et al., 2020b).

2 Preliminaries

Constrained Markov Decision process (MDP). A constrained MDP $\mathcal{M}$ could be described by a tuple $(\mathcal{S}, \mathcal{A}, H, s_{\text{ini}}, R, C, P, B)$, where $\mathcal{S} \times \mathcal{A}$ is the finite state-action space, H is the planing horizon, s_{ini} is the initial state[1], $R = \{R_{s,a,h}\}$ is the reward function which maps a state-action pair to $\Delta([0,1])$[2], $C = \{C_{s,a,h}\}$ is the violation function which maps a state-action pair to $\Delta([0,1])$, $P = \{P_{s,a,h}\}$ is the transition model which maps a state-action pair to $\Delta(\mathcal{S})$, and B is the threshold of violation.

A policy $\pi = \{\pi_h(s)\}_{(s,h)\in\mathcal{S}\times[H]}$ is a group of mappings from the state space $\mathcal{S}$ to $\Delta(\mathcal{A})$. When a policy π is deterministic, we write $\pi_h(s)$ to denote the unique action at (s,h) following π. Let Π denote the set of all deterministic policies. $V_h^{\pi,P}(R,s)$ denotes the value function. $Q_h^{\pi,P}(R,(s,a))$ denote the Q-function.

Given a transition model P, a policy π and a reward function R, we use $V^{\pi,P}(R) = \{V_h^{\pi,P}(R,s)\}_{(s,h)\in\mathcal{S}\times[H]}$ and $Q^{\pi,P}(R) = \{V_h^{\pi,P}(R,(s,a))\}_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]}$ to denote the cor-

[1] Without loss of generality, we assume s_{ini} is the unique state at the first level.

[2] We use $\Delta(\mathcal{X})$ to denote the set of distributions over $\mathcal{X}$.

responding value function and Q-function respectively, i.e,

$$V_h^{\pi,P}(R,s) = \mathbb{E}_{\pi,P}\left[\sum_{h'=h}^{H} r_{s_h,a_h,h} \mid s_h = s\right],$$

$$Q_h^{\pi,P}(R,(s,a)) = \mathbb{E}_{\pi,P}\left[\sum_{h'=h}^{H} r_{s_h,a_h,h} \mid (s_h,a_h) = (s,a)\right],$$

where $r_{s_h,a_h,h} \sim R_{s_h,a_h,h}$. Then the optimal value function and Q-function are defined as

$$V_h^{\pi,P}(R,s) = \max_{\pi} \mathbb{E}_{\pi,P}\left[\sum_{h'=h}^{H} r_{s_h,a_h,h} \mid s_h = s\right],$$

$$Q_h^{\pi,P}(R,(s,a)) = \max_{\pi} \mathbb{E}_{\pi,P}\left[\sum_{h'=h}^{H} r_{s_h,a_h,h} \mid (s_h,a_h) = (s,a)\right].$$

We define $\mathsf{Opt}(\check{P},\check{R})$ to be the set of optimal deterministic policies with respect to $(\check{P},\check{R})$.

The target of CMDP is to solve the following optimization problem:

$$\max_{\pi} V_1^{\pi,P}(R,s_{\text{ini}}) \quad \text{s. t.} \quad V_1^{\pi,P}(C,s_{\text{ini}}) \le B. \tag{1}$$

We say π is strictly-optimal if π is the optimal solution of equation 1. Let the optimal value of equation 1 be V^*. A policy π is defined as ϵ-strictly-optimal iff $V_1^{\pi,P}(R,s_{\text{ini}}) \ge V^* - \epsilon$ and $V_1^{\pi,P}(C,s_{\text{ini}}) \le B$.

The learning problem. We assume the agent has access to a generative mode and can sample the next state and reward for any state-action pair (Kakade, 2003; Kearns et al., 2002). The sample complexity is defined as the number of samples required to find an ϵ-strictly-optimal policy.

Notations. We use $[N]$ to denote the set $\{1,2,\dots,N\}$ for a natural number N. Let $\mathbf{1}$ denote the S-dimensional all 1 vector. We use Δ^m to denote the m dimensional simplex.

3 Algorithm and Technique Overview

In this section, we present our algorithm and summary the novel techniques.

3.1 Algorithm Description

We present the main algorithm in Algorithm 1. For each triple (s,a,h), the learner first queries N samples and computes the empirical statistics, then adds a perturbation to the empirical statistics. Finally, it computes the optimal policy for the perturbed statistics using linear programming.

Let the N samples of (s,a,h) be $\{(s,a,h,r_i,c_i,s_i)\}_{i=1}^{N}$. We define the empirical statistics as follows.

$$\hat{P}_{s,a,h,s'} = \frac{1}{N}\sum_{i=1}^{H} \mathbb{I}[s_i = s']; \quad \hat{R}_{s,a,h,s'} = \frac{1}{N}\sum_{i=1}^{N} r_i; \quad \hat{C}_{s,a,h,s'} = \frac{1}{N}\sum_{i=1}^{N} c_i. \tag{2}$$

Let $\mathcal{D}$ be the uniform distribution over the set $\{\upsilon, 2\upsilon, \dots, K\upsilon\}$ with $K = \frac{2S^2H^2A^4\cdot A^{2SH}}{\delta}$ and $\upsilon = \frac{\epsilon}{10KH^2}$. For each (s,a,h) triple, a random variable $\varsigma_{s,a,h} \sim \mathcal{D}$ is used to perturb the empirical reward and violation.

Linear Programming for CMDP. A classic method for solving CMDPs relies on linear programming (LP) formulationsAltman (1999). Given a transition model $\check{P}$, a reward $\check{R}$, a violation $\check{C}$ and a threshold $\check{B}$, the LP problem is formulated as

$$\begin{aligned} \max_{d\in[0,1]^{SAH}} \sum_{s,a,h} d_{s,a,h}\check{R}_{s,a,h} \quad \text{s. t.} & \\ \sum_{a} d_{s,a,h+1} = \sum_{s',a'} d_{s',a',h}\check{P}_{s',a,',h}(s) \quad \forall(s,h); & \\ \sum_{s,a} d_{s,a,h} = 1 \quad \forall h; & \\ \sum_{s,a,h} d_{s,a,h}\check{C}_{s,a,h} \leq \check{B}. & \end{aligned} \tag{3}$$

Given a solution d to the above problem, we can recover the corresponding policy by setting $\pi_h(a|s) = \frac{d_{s,a,h}}{\sum_{a'} d_{s,a',h}}$ for all (s,a,h), where we define $\pi_h(\cdot|s)$ arbitrarily if $\sum_{a'} d_{s,a',h}$=0.

We assume that $\mathsf{Linear\ Programming}(\check{P},\check{R},\check{C},\check{B})$ can return the *precise* solution to the above LP problem. The computation of a precise solution generally incurs a high computational cost. To mitigate this, we employ a rounding procedure on the LP parameters, which reduces the cost to polynomial time. Without loss of generality, the parameters $\hat{R}$, $\hat{C}$, B' and $\frac{1}{N}$ are rounded to be integer multiples of some proper scaling factor θ such that $\log(1/\theta)$ is polynomial in $(S,A,H,1/\epsilon,1/\delta)$. In this way, the computational cost to find a precise solution is $\mathrm{poly}(SAH\log(1/\upsilon))$. We refer to Gleixner & Steffy (2020) for more details.

Definition 1. *Let* $\underline{B}^* = \min_\pi V_1^{\pi,P}(C, s_{\text{ini}})$. *The Slater's condition number is defined as* $\check{\zeta} = \frac{B-\underline{B}^*}{H}$.

We fix some ζ and assume $\check{\zeta}H \geq 20\zeta\epsilon$. In the case this assumption holds, our algorithm can successfully return an ϵ-strictly-optimal policy with high probability. Otherwise, it would be statistically hard to estimate $\check{\zeta}$ using $O(\frac{SAH^4}{\zeta^2\epsilon^2})$ queries.

Algorithm 1 Perturbed Linear Programming

```
Input: value threshold $\epsilon$, violation threshold $B$, failure probability $\delta$, condition number $\zeta$
Initialization: $\delta_1 = \frac{\epsilon\delta\zeta^2}{12800S^2AH^5}$, $N = \frac{H^3\log(1/\delta_1)}{1000000\epsilon^2\zeta^2}$, $B' = B - \frac{\zeta\epsilon}{20}$, $\alpha = \upsilon$
for $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ do
    Query $N$ samples for $(s,a,h)$;
    Compute the empirical transition, reward and violation $\hat{P}_{s,a,h,s'}$, $\hat{R}$ and $\hat{C}$ (see equation 2);
end for
for $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ do
    Generate $\varsigma_{s,a,h} \sim \mathcal{D}$;
    $\tilde{R}_{s,a,h} = \hat{R}_{s,a,h} + \varsigma_{s,a,h}$;
    $\tilde{C}_{s,a,h} = \hat{C}_{s,a,h} - \varsigma_{s,a,h}$
    $\tilde{P}_{s,a,h} = (1-\alpha)\hat{P}_{s,a,h} + \frac{\alpha}{S}\mathbf{1}$;
end for
return: $\pi = \mathsf{Linear\ Programming}\Big((\tilde{P} = \{\tilde{P}_{s,a,h}\}_{(s,a,h)}, \tilde{R} = \{\tilde{R}_{s,a,h}\}_{(s,a,h)}, \tilde{C}, B'\Big)$.
```

3.2 Technique Summary

We begin by characterizing the hardness of learning in CMDPs and then present our approach to address it.

Main difficulty in applying LP. A key challenge in CMDPs, unlike their unconstrained counterparts, is the non-uniqueness of the optimal value function. Specifically, equation 3 can have multiple optimal solutions (e.g., d^1, d^2 with induced policies π^1, π^2) that yield vastly different value sequences $\{V_h^{\pi^1,\check{P}}(\check{R},\cdot)\}$ and $\{V_h^{\pi^2,\check{P}}(\check{R},\cdot)\}$. This variability prevents the application of standard uniform concentration arguments. We take the following toy MDP as an example.

Example 1. *Consider an MDP with $H = 2$. There is one state s_{ini} in the first level and S states in the second level. The reward and violations in the first level are 0. Set $P_{s_{\text{ini}},a,1} = \frac{1}{S}\boldsymbol{1}$ for all actions a. For each state s in the second level, there are two actions a^1, a^2 such that $R_{s,a^1,2} = C_{s,a^1,2} = 1$ and $R_{s,a^2,2} = C_{s,a^2,2} = 0$. Set $B = \frac{1}{2}$.*

In the above CMDP, any π satisfying $\frac{1}{S}V_2^{\pi,P}(R,s) = \frac{1}{2}$ is a strictly-optimal policy. Consequently, the class $\{V_2^{\pi,P}(R,\cdot) : \pi \text{ strictly-optimal }\}$ has high complexity. This high complexity can cause naive LP solutions to be sensitive to distributional shifts or estimation errors.

Our solution. We say an action a at (s,h) is constrained-optimal iff there exists some strictly-optimal deterministic policy π such that $\pi_h(s) = a$. A key observation from Example 1 is that the non-uniqueness of the constrained-optimal actions can severely harm the performance of the LP solution. Motivated by this observation, we aim to devise a method for ensuring the uniqueness of the constrained-optimal action.

Breaking ties by perturbation. If two actions a^1 and a^2 are both constrained-optimal, then there exists a $\lambda \in [0,1]$ such that both a^1 and a^2 are optimal actions for the reward function $\lambda\check{R} - (1-\lambda)\check{C}$. Therefore, if the reward $\check{R}$ can be perturbed such that for all $\lambda \in [0,1]$ the optimal action for $\lambda\check{R} - (1-\lambda)\check{C}$ is unique at every (s,h), then the strictly-optimal policy will be unique. However, achieving this is generally impossible due to the adaptivity to select $\lambda \in [0,1]$ such that actions a^1 and a^2 are both optimal. The key insight to overcome this problem is that λ can only be adaptive enough to tie a single pair of actions. This is because making a^1 and a^2 both optimal forces λ to a specific value, rendering it incapable of creating ties between other actions. Through this approach, we establish that with high probability, no more than two strictly-optimal policies exist after perturbation. This leads to an efficient reduction in the complexity of the value function class, ultimately permitting sharper concentration results.

4 Analysis

In this section, we present the proof of Theorem 1.

4.1 Basic property of perturbed planning.

We first introduce some basic properties of the perturbed CMDP. We fix the transition model $\check{P}$, the reward $\check{R}$ the violation $\check{C}$ and the threshold $\check{B}$. Let[3] $\varsigma = \{\varsigma_{s,a,h}\}_{(s,a,h)}$ be a group of i.i.d. noise obeying distribution $\mathcal{D}$.

Definition 2 (Optimal actions.). *Given a transition model $\check{P}$, a reward function $\check{R}$, a state-action pair $(s,a)$and $\eta \in [0,H]$, we say the action a is η-optimal iff $Q_h^{*,\check{P}}(\check{R},(s,a)) \geq V_h^{*,\check{P}}(\check{R},s) - \eta$. Especially, we say the action a is optimal iff $Q_h^{*,\check{P}}(\check{R},(s,a)) = V_h^{*,\check{P}}(\check{R},s)$.*

Recall that $\mathsf{Opt}(P,R)$ denotes the set of optimal deterministic policies with transition model P and reward R. For $\lambda \in [0,1]$, let $\pi(\lambda) = \mathsf{Opt}(\check{P}, \lambda\check{R} - (1-\lambda)\check{C} + \varsigma)$ be the set of deterministic optimal policies.

Lemma 1. *With probability $1-\delta$, for all $\lambda \in [0,1]$, one of the two following claims holds:*

- $|\pi(\lambda)| = 1$ *;*
- $|\pi(\lambda)| = 2$. *Furthermore, let $\pi(\lambda) = \{\pi^1, \pi^2\}$. The policies π^1 and π^2 are identical for all state-time step pairs $(s,h) \in \mathcal{S} \times [H]$, differing at exactly one pair (s', h'). That is, $\pi_h^1(s) = \pi_h^2(s)$ for all $(s,h) \neq (s',h')$.*

[3] When the context is clear, we use the subscript $\{\cdot\}_{(s,a,h)}$ to denote $\{\cdot\}_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]}$.

Definition 3 (Constrained optimal value and violation.)**.** *Given a transition model $\check{P}$, a reward function $\check{R}$, a violation function $\check{C}$ and $\lambda \in [0,1]$, define*

$$\mathcal{V}(\check{P},\check{R},\check{C},\lambda) = \left\{ V^{\pi,\check{P}}(\check{R}) \mid \pi \in \mathsf{Opt}(\check{P}, \lambda\check{R} - (1-\lambda)\check{C}) \right\};$$

$$\mathcal{C}(\check{P},\check{R},\check{C},\lambda) = \left\{ V^{\pi,\check{P}}(\check{C}) \mid \pi \in \mathsf{Opt}(\check{P}, \lambda\check{R} - (1-\lambda)\check{C}) \right\}.$$

Definition 4. *For $V^1, V^2 \in \mathbb{R}^{SH}$, we say $V^1 \preccurlyeq V^2$ iff $V^1_{s,h} \le V^2_{s,h}$ for all $(s,h) \in \mathcal{S} \times [H]$.*

Lemma 2. *Fix $\check{P}, \check{R}$ and $\check{C}$. Let $\varsigma = \{\varsigma_{s,a,h}\}_{(s,a,h)}$ be a group of i.i.d. noise obeying distribution $\mathcal{D}$. Let $\tilde{P} = (1-\alpha)\check{P} + \frac{\alpha}{S}\mathbf{1}$, $\tilde{R} = \check{R} + \varsigma$ and $\tilde{C} = \check{C} - \varsigma$. Define $\mathfrak{V} = \{\mathcal{V}(\tilde{P},\tilde{R},\tilde{C},\lambda) \mid \lambda \in [0,1]\}$ and $\mathfrak{C} = \{\mathcal{C}(\tilde{P},\tilde{R},\tilde{C},\lambda) \mid \lambda \in [0,1]\}$. With probability $1-\delta$, "$\preccurlyeq$" is a total order over both $\mathfrak{V}$ and $\mathfrak{C}$.*

Proof. We only prove the conclusion for $\mathfrak{V}$. The conclusion about $\mathfrak{C}$ could be proven in a similar way. It suffices to show that for any $V^1, V^2 \in \mathfrak{V}$, it holds either $V^1 \preccurlyeq V^2$ or $V^2 \preccurlyeq V_1$. Assume $V^1 \in \mathcal{V}(\tilde{P},\tilde{R},\tilde{C},\lambda_1)$ and $V^2 \in \mathcal{V}(\tilde{P},\tilde{R},\tilde{C},\lambda_2)$.

By Lemma 1, with probability $1-\delta$, $|\mathsf{Opt}(\tilde{P}, \lambda\tilde{R} - (1-\lambda)\tilde{C})| = |\mathsf{Opt}(\tilde{P}, \lambda\check{R} - (1-\lambda)\check{C} + \varsigma)| \in \{1,2\}$ for all $\lambda \in [0,1]$.

Case i: $\lambda_1 \neq \lambda_2$. Without loss of generality, we assume $\lambda_1 < \lambda_2$. Assume $V^1 = V^{\pi^1,\tilde{P}}(\tilde{R}), V^2 = V^{\pi^2,\tilde{P}}(\tilde{R})$ for $\pi^1 \in \mathsf{Opt}(\tilde{P}, \lambda_1\tilde{R} - (1-\lambda_1)\tilde{C})$ and $\pi^2 \in \mathsf{Opt}(\tilde{P}, \lambda_2\tilde{R} - (1-\lambda_2)\tilde{C})$.

Fix $(s,h) \in \mathcal{S} \times [H]$. We have that

$$\lambda_1 V_h^{\pi^1,\tilde{P}}(\tilde{R},s) - (1-\lambda_1)V_h^{\pi^1,\tilde{P}}(\tilde{C},s) \ge \lambda_1 V_h^{\pi^2,\tilde{P}}(\tilde{R},s) - (1-\lambda_1)V_h^{\pi^2,\tilde{P}}(\tilde{C},s);$$

$$\lambda_2 V_h^{\pi^2,\tilde{P}}(\tilde{R},s) - (1-\lambda_2)V_h^{\pi^2,\tilde{P}}(\tilde{C},s) \ge \lambda_2 V_h^{\pi^1,\tilde{P}}(\tilde{R},s) - (1-\lambda_2)V_h^{\pi^1,\tilde{P}}(\tilde{C},s).$$

Therefore, we have

$$(\lambda_1(1-\lambda_2) - \lambda_2(1-\lambda_1))V_h^{\pi^1,\tilde{P}}(\tilde{R},s) \ge (\lambda_1(1-\lambda_2) - (1-\lambda_1)\lambda_2)V_h^{\pi^2,\tilde{P}}(\tilde{R},s).$$

It then follows $V_h^{\pi^1,\tilde{P}}(\tilde{R},s) \le V_h^{\pi^2,\tilde{P}}(\tilde{R},s)$.

Case ii: $\lambda_1 = \lambda_2$. Write $\lambda_1 = \lambda_2 = \tilde{\lambda}$. If $|\mathsf{Opt}(\tilde{P}, \lambda(\check{R}+\varsigma) - (1-\lambda)(\check{C}-\varsigma))| = 1\}$, we then have that $V^1 = V^2$. Otherwise, $\mathsf{Opt}(\tilde{P}, \lambda(\check{R}+\varsigma) - (1-\lambda)(\check{C}-\varsigma)) = \{\pi^1, \pi^2\}$ such that $\pi^1_h(s) = \pi^2_h(s)$ for all $(s,h) \neq (s',h')$. By definition, we have that $V^1, V^2 \in \{V^{\pi^1,\tilde{P}}(\tilde{R}), V^{\pi^2,\tilde{P}}(\tilde{R})\}$. It suffices to consider the case that $V^1 = V^{\pi^1,\tilde{P}}(\tilde{R},\cdot)$ and $V^2 = V^{\pi^2,\tilde{P}}(\tilde{R},\cdot)$. For any (s,h), by policy difference lemma (see Lemma 10), we have that[4]

$$\begin{aligned} V_h^{\pi^1,\tilde{P}}(\tilde{R},s) - V_h^{\pi^2,\tilde{P}}(\tilde{R},s) &= \mathbb{E}_{\pi^1,\tilde{P}}\left[\sum_{\tau=1}^{H}\left((\tilde{P}_{s_\tau,\pi^1_\tau(s_\tau),\tau} - \tilde{P}_{s_\tau,\pi^2_\tau(s_\tau),\tau})^\top V_{\tau+1}^{\pi^2,\tilde{P}}(\tilde{R},\cdot)\right)\right] \\ &\quad + \mathbb{E}_{\pi^1,\tilde{P}}\left[\sum_{\tau=1}^{H}\left(\tilde{R}_{s_\tau,\pi^1_\tau(s_\tau),\tau} - \tilde{R}_{s_\tau,\pi^2_\tau(s_\tau),\tau}\right)\right] \\ &= \mathbb{E}_{\pi^1,\tilde{P}}\left[\mathbb{I}[s_{h'} = s']\right] \cdot \Delta, \end{aligned}$$

where

$$\Delta = \left((\tilde{P}_{s',\pi^1_{h'}(s'),h'} - \tilde{P}_{s',\pi^2_{h'}(s'),h'})^\top V_{h'+1}^{\pi^2,\tilde{P}}(\tilde{R},\cdot)\right) + \left(\tilde{R}_{s',\pi^1_{h'}(s'),h'} - \tilde{R}_{s',\pi^2_{h'}(s'),h'}\right)$$

is the temporal difference at (s',h'). In the case $\Delta \ge 0$, we have that $V^1_{s,h} \ge V^2_{s,h}$ for all $(s,h) \in \mathcal{S} \times [H]$, which means $V^2 \preccurlyeq V^1$. Otherwise we have $V^1 \preccurlyeq V^2$.

The proof is completed.

□

[4] We set $V^{\pi,P}_{H+1}(R,s) = 0$ for any proper (π,P,R,s).xc

Let $\mathbb{N}_{\epsilon_1} \subset$ be $\{0, \epsilon_1, 2\epsilon_1, \dots, L\epsilon_1\}^{SH}$ with $L = \frac{2H}{\epsilon_1}$. For $V = \{V_{s,h}\}_{(s,h)} \in [0, 2H]^{SH}$, define $\mathrm{Proj}_{\mathbb{N}_{\epsilon_1}}(V) = \left\{\left\lfloor \frac{V_{s,h}}{\epsilon_1} \right\rfloor \cdot \epsilon_1\right\}_{(s,h)}$. For $\mathfrak{V} \subset [0, 2H]^{SH}$, define $\mathrm{Proj}_{\mathbb{N}_{\epsilon_1}}(\mathfrak{V}) = \left\{\mathrm{Proj}_{\mathbb{N}_{\epsilon_1}}(V) | V \in \mathfrak{V}\right\}$.

Lemma 3. *Assume $\mathfrak{V}$ is a subset of $[0, 2H]^{SAH}$ and assume "$\preccurlyeq$" in Definition 4 is a total order over $\mathfrak{V}$. Then the size of $\mathrm{Proj}_{\mathbb{N}_{\epsilon_1}}(\mathfrak{V})$ is at most $\frac{2SH^2}{\epsilon_1} + 1$.*

Proof. Note that $V^1 \preccurlyeq V^2$ implies that $\mathrm{Proj}_{\mathbb{N}_{\epsilon_1}}(V^1) \preccurlyeq \mathrm{Proj}_{\mathbb{N}_{\epsilon_1}}(V^2)$ for any $V^1, V^2 \in [0, 2H]^{SAH}$. Along with the assumption that "$\preccurlyeq$" is a total order over $\mathfrak{V}$, "$\preccurlyeq$" is also a total order over $\mathrm{Proj}_{\mathbb{N}_{\epsilon_1}}(\mathfrak{V})$. Since $\mathbb{N}_{\epsilon_1}$ is a finite set and $\mathrm{Proj}_{\mathbb{N}_{\epsilon_1}}(\mathfrak{V}) \subset \mathbb{N}_{\epsilon_1}$. We learn that $\mathrm{Proj}_{\mathbb{N}_{\epsilon_1}}(\mathfrak{V})$ is also finite. Therefore, we can re-write $\mathrm{Proj}_{\mathbb{N}_{\epsilon_1}}(\mathfrak{V})$ as $\{W^1, W^2, \dots, W^m\}$ in an increasing order with respect to "$\preccurlyeq$", where $m = |\mathrm{Proj}_{\mathbb{N}_{\epsilon_1}}(\mathfrak{V})|$. As a result, we have that

$$2SH^2 \geq \|W^m - W^1\|_1 = \sum_{i=1}^{m-1} \|W^{i+1} - W^i\|_1 \geq \epsilon_1 \cdot (m-1),$$

which implies that $m \leq \frac{2SH^2}{\epsilon_1} + 1$. The proof is completed.

□

4.2 Sample Complexity Bounds (Proof of Theorem 1)

Let π be the policy returned by Algorithm 1. The total number of samples is $SAHN = \tilde{O}\left(\frac{SAH^4}{\epsilon^2 \zeta^2}\right)$ It suffices to show that with probability $1 - \delta$, π is an ϵ-strictly-optimal policy. Recall the definition of $\tilde{P}, \tilde{R}, \tilde{C}$ and B' in Algorithm 1.

Concentration event. Choose $\delta_1 = \delta \cdot \frac{\epsilon_1}{16S^2AH^3}$. For $p, v \in \mathbb{R}^S$, we define $\mathbb{V}(p, v) = \sum_s p_s v_s^2 - (p^\top v)^2$. Recall the definition of $\mathcal{V}(\tilde{P}, \tilde{R}, \tilde{C}, \lambda)$ in Definition 3. Define $\mathfrak{V} = \{\mathcal{V}(\tilde{P}, \tilde{R}, \tilde{C}, \lambda) \mid \lambda \in [0, 1]\}$ and $\mathfrak{C} = \{\mathcal{C}(\tilde{P}, \tilde{R}, \tilde{C}, \lambda) \mid \lambda \in [0, 1]\}$. Let $\mathfrak{V}_h = \{\{v_{s,h}\}_{s \in \mathcal{S}} \mid v \in \mathfrak{V}\}$ and $\mathfrak{C}_h = \{\{v_{s,h}\}_{s \in \mathcal{S}} \mid v \in \mathfrak{C}\}$ for $h = 1, 2, \dots, H$. Let $\mathcal{E}_h$ be the event where

$$\left|\left(\hat{P}_{s,a,h} - P_{s,a,h}\right)^\top v\right| \leq 4\sqrt{\frac{\mathbb{V}(P_{s,a,h}, v)\log(1/\delta_1)}{N}} + \frac{4H\log(1/\delta_1)}{N} + 8\sqrt{\frac{\epsilon_1 \log(1/\delta_1)}{N}} + 2\epsilon_1; \tag{4}$$

$$\left|\hat{R}_{s,a,h} - R_{s,a,h}\right| \leq 4\sqrt{\frac{\log(1/\delta_1)}{N}}; \tag{5}$$

$$\left|\hat{C}_{s,a,h} - C_{s,a,h}\right| \leq 4\sqrt{\frac{\log(1/\delta_1)}{N}} \tag{6}$$

for any $v \in \mathfrak{V}_{h+1} \cup \mathfrak{C}_{h+1}$ and any (s, a) pair. Define $\mathcal{E} = \cup_{h=1}^H \mathcal{E}_h$.

Lemma 4. $\mathbb{P}[\mathcal{E}] \geq 1 - 2\delta$.

Proof. By Bernstein's inequality (see Lemma 9), equation 5 and equation 6 holds with probability at least $1 - 4SAH\delta_1$ for all (s, a, h).

Let $\check{\mathfrak{V}}_h = \{\{v_{s,h}\}_{s \in \mathcal{S}} \mid v \in \mathrm{Proj}_{\mathbb{N}_{\epsilon_1}}(\mathfrak{V})\}$ and $\check{\mathfrak{C}}_h = \{\{v_{s,h}\}_{s \in \mathcal{S}} \mid v \in \mathrm{Proj}_{\mathbb{N}_{\epsilon_1}}(\mathfrak{C})\}$. Note that $\mathfrak{V}_{h+1}$ and $\mathfrak{C}_{h+1}$ only depend on the statistics after the h-th layer. By Lemma 2 and Lemma 3, with probability $1 - \delta$, $|\check{\mathfrak{V}}_{h+1}| \leq \mathrm{Proj}_{\mathbb{N}_{\epsilon_1}}(\mathfrak{V}) \leq \frac{2SH^2}{\epsilon_1} + 1$ and $|\check{\mathfrak{C}}_{h+1}| \leq \mathrm{Proj}_{\mathbb{N}_{\epsilon_1}}(\mathfrak{C}) \leq \frac{2SH^2}{\epsilon_1} + 1$.

By Lemma 9, with probability $1 - 8SH^2\delta_1/\epsilon_1$, it holds that

$$\left|\left(\hat{P}_{s,a,h} - P_{s,a,h}\right)^\top v\right| \leq 4\sqrt{\frac{\mathbb{V}(P_{s,a,h}, v)\log(1/\delta_1)}{N}} + \frac{4H\log(1/\delta_1)}{N}$$

for all $v \in \check{\mathfrak{V}}_{h+1} \cup \check{\mathfrak{C}}_{h+1}$.

Assume this event holds. For any $v \in \mathfrak{V}_{h+1} \cup \mathfrak{C}_{h+1}$, there exists $v' \in \check{\mathfrak{V}}_{h+1} \cup \check{\mathfrak{C}}_{h+1}$ such that $\|v - v'\|_\infty \le \epsilon_1$. As a result, we have that

$$\begin{aligned} &\left| \left(\hat{P}_{s,a,h} - P_{s,a,h} \right)^\top v \right| \\ &\le \left| \left(\hat{P}_{s,a,h} - P_{s,a,h} \right)^\top v' \right| + 2\epsilon_1 \\ &\le 4\sqrt{\frac{\mathbb{V}(P_{s,a,h}, v') \log(1/\delta_1)}{N}} + \frac{4H \log(1/\delta_1)}{N} + 2\epsilon_1 \\ &\le 4\sqrt{\frac{\mathbb{V}(P_{s,a,h}, v) \log(1/\delta_1)}{N}} + \frac{4H \log(1/\delta_1)}{N} + 8\sqrt{\frac{\epsilon_1 \log(1/\delta_1)}{N}} + 2\epsilon_1. \end{aligned} \tag{7}$$

Putting all together, we learn that $\mathbb{P}[\mathcal{E}] \ge 1 - 16S^2AH^3\delta_1/\epsilon_1 - \delta \ge 1 - 2\delta$.

□

We continue the analysis conditioned on $\mathcal{E}$.

Lemma 5. *There exists some $\lambda \in [0, 1]$, such that*

$$V_h^{\pi,\tilde{P}}(\lambda\tilde{R} - (1-\lambda)\tilde{C}, s) = V_h^{*,\tilde{P}}(\lambda\tilde{R} - (1-\lambda)\tilde{C}, s) \tag{8}$$

for all $(s, h) \in \mathcal{S} \times [H]$.

Proof. Since π is the optimal solution to the following optimization problem,

$$\max_{\pi'} V_1^{\pi',\tilde{P}}(\tilde{R}, s_{\text{ini}}), \text{ such that } V_1^{\pi',\tilde{P}}(\tilde{C}, s_{\text{ini}}) \le B', \tag{9}$$

there exists $\lambda \in [0, 1]$ such that

$$V_1^{\pi,\tilde{P}}(\lambda\tilde{R} - (1-\lambda)\tilde{C}, s_{\text{ini}}) = V_1^{*,\tilde{P}}(\lambda\tilde{R} - (1-\lambda)\tilde{C}, s_{\text{ini}}).$$

Recalling that $\tilde{P}_{s,a,h} = (1-\alpha)\hat{P}_{s,a,h} + \frac{\alpha}{S}\mathbf{1}$, we learn that $\mathbb{E}_{\pi,\tilde{P}}[\mathbb{I}[s_h = s]] \ge \frac{\alpha}{S}$ for all $(s, h) \in \mathcal{S} \times [H]$. As a result, $V_h^{\pi,\tilde{P}}(\lambda\tilde{R} - (1-\lambda)\tilde{C}) = V_h^{*,\tilde{P}}(\lambda\tilde{R} - (1-\lambda)\tilde{C}, s)$ for all $(s, h) \in \mathcal{S} \times [H]$.

□

Denote the value fo λ in Lemma 5 as λ^*. Recall Definition 3 for the constrained optimal value set. Then π is an (possibly stochastic) optimal policy with respect to the reward $\lambda^*\tilde{R} - (1-\lambda^*)\tilde{C}$ and the transition model $\tilde{P}$.

To facilitate the analysis, we have the following technical lemmas. We defer to proofs to Appendix C due to space limitations. Recall that $\epsilon_3 = \frac{\epsilon\zeta}{8H}$.

Lemma 6. *With probability $1 - \delta$, for any $\lambda \in [0, 1]$, $|V^{*,\tilde{P}}(\lambda\tilde{R} - (1-\lambda)\tilde{C}, s_{\text{ini}}) - V^{*,P}(\lambda R - (1-\lambda)C, s_{\text{ini}})| \le \epsilon_3$.*

Assume the events in Lemma 6 holds.

Lemma 7. *It holds that*

$$|V_1^{\pi,P}(R, s_{\text{ini}}) - V_1^{\pi,\tilde{P}}(\tilde{R}, s_{\text{ini}})| \le \epsilon_3; \tag{10}$$

$$|V_1^{\pi,P}(C, s_{\text{ini}}) - V_1^{\pi,\tilde{C}}(C, s_{\text{ini}})| \le \epsilon_3. \tag{11}$$

Lemma 8. *Recall that $\underline{B}^* = \min_\pi V_1^{\pi,P}(C, s_{\text{ini}})$. If $V_1^{\pi,\tilde{C}}(\tilde{C}, s_{\text{ini}}) = B'$, it then holds that $\lambda^* \ge (1-\lambda^*) \cdot \frac{B' - \underline{B}^* - \epsilon_3}{H}$.*

Proof. By setting $\lambda = 0$ in Lemma 6, we learn that there exits a policy $\underline{\pi}$ such that $V^{\underline{\pi},\tilde{P}}(\tilde{C}) \leq \underline{B}^* + \epsilon_3$. We then prove by contradiction. Assume that $\lambda^* < (1-\lambda^*)\frac{B'-\underline{B}^*-\epsilon_3}{H+HK\upsilon}$. Then we have that

$$\begin{aligned} V_1^{\underline{\pi},\tilde{P}}(\lambda^*\tilde{R} - (1-\lambda^*)\tilde{C}) &\geq -(1-\lambda^*)(\underline{B}^* + \epsilon_3) \\ &> \lambda^*(H + HK\upsilon) - (1-\lambda^*)B' \\ &\geq V_1^{\pi,\tilde{P}}(\lambda^*\tilde{R} - (1-\lambda^*)\tilde{C}), \end{aligned}$$

which contradicts to the optimality of π. □

By Lemma 6 , we have that

$$\lambda^* V_1^{\pi,\tilde{P}}(\tilde{R}, s_{\text{ini}}) - (1-\lambda^*)V_1^{\pi,P}(\tilde{C}, s_{\text{ini}}) \geq V^{*,P}(\lambda^* R - (1-\lambda^*)C) - \epsilon_3. \tag{12}$$

We continue the analysis in two cases.

Case i: $V_1^{\pi,\tilde{C}}(\tilde{C}, s_{\text{ini}}) = B'$**.** By Lemma 7, we have that $B - 2\epsilon_3 = B' - \epsilon_3 \leq V^{\pi,P}(C) \leq B' + \epsilon_3 \leq B$, and

$$\lambda^* V_1^{\pi,P}(R, s_{\text{ini}}) - (1-\lambda^*)V_1^{\pi,P}(C, s_{\text{ini}}) \geq \lambda^* V_1^{\pi',P}(R, s_{\text{ini}}) - (1-\lambda^*)V_1^{\pi',P}(C, s_{\text{ini}}) - 3\epsilon_3 \tag{13}$$

for any π'. As a result, we obtain

$$V_1^{\pi,P}(R, s_{\text{ini}}) \geq V_1^{\pi',P}(R, s_{\text{ini}}) - \frac{1}{\lambda^*} \cdot \Big((1-\lambda^*)(V_1^{\pi',P}(C, s_{\text{ini}}) - B) + 5\epsilon_3\Big) \tag{14}$$

for any π. It then follows that, for any π' such that $V_1^{\pi',P}(C, s_{\text{ini}}) \leq B$, $V_1^{\pi',P}(R, s_{\text{ini}}) \leq V_1^{\pi,P}(R, s_{\text{ini}}) + \frac{5\epsilon_3}{\lambda^*}$. Using Lemma 8 and the fact that $5\epsilon_3 \cdot \frac{H+B-\underline{B}^*-2\epsilon_3}{B-\underline{B}^*-2\epsilon_3} \leq \epsilon$, we learn that π is an ϵ-strictly-optimal policy.

Case ii: $V_1^{\pi,\tilde{P}}(\tilde{C}, s_{\text{ini}}) < B'$**.** In this case, we learn that $V_1^{\pi,\tilde{P}}(\tilde{R}, s_{\text{ini}}) = V_1^{*,\tilde{P}}(\tilde{R}, s_{\text{ini}})$. By Lemma 6 and Lemma 7, we have that

$$\begin{aligned} &V_1^{\pi,P}(C, s_{\text{ini}}) \leq B' + \epsilon_3 = B; \\ &V_1^{\pi,P}(R, s_{\text{ini}}) \geq V_1^{\pi,\tilde{P}}(\tilde{R}, s_{\text{ini}}) - \epsilon_3 \geq V^{*,P}(R, s_{\text{ini}}) - 2\epsilon_3. \end{aligned}$$

Noting that $2\epsilon_3 \leq \epsilon$, π is an ϵ-strictly-optimal policy.

The proof is completed by rescaling δ as $\delta/3$.

5 Conclusion

We introduce a novel plug-in approach for constrained RL that achieves the sample complexity of via a generative model. Our method is general, requiring only black-box access to an optimization oracle for the empirical CMDP. A key innovation is a reward perturbation technique that ensures the oracle's solution is almost uniquely determined. Two promising directions for future work are the extension of our results to CMDPs with multiple constraints and an investigation into the robustness of the proposed plug-in algorithm.

References

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *The 33rd Annual Conference on Learning Theory*, pp. 67–83, 2020a.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in Markov decision processes. In *The 33rd Annual Conference on Learning Theory*, pp. 64–66, 2020b.

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

Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

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.

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.

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*, pp. 3304–3312. PMLR, 2021.

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

Ambros Gleixner and Daniel E Steffy. Linear programming using limited-precision oracles. *Mathematical Programming*, 183(1):525–554, 2020.

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

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

Michael Kearns, Yishay Mansour, and Andrew Y Ng. A sparse sampling algorithm for near-optimal planning in large markov decision processes. *Machine Learning*, 49(2-3):193–208, 2002.

Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Operations Research*, 72(1):203–221, 2024.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *The 22nd Annual Conference on Learning Theory*, 2009.

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

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

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

Statement about usage of large language models. We use large language models to help refine and improve our writing.

A Parameter Settings

Given the input parameter $\zeta, \epsilon, S, A, H, B$ and δ, we set $\epsilon_3 = \frac{\zeta\epsilon}{20}$, $\epsilon_2 = \frac{\epsilon_3\zeta}{40H^2}$, $\epsilon_1 = \frac{\epsilon_3}{10H}$, $K = 2S^2H^2A^4 \cdot A^{2SH} \cdot \frac{1}{\delta}$, $\upsilon = \frac{\epsilon}{10KH^2}$, $B' = B - \epsilon_3$, $\alpha = \upsilon$ and $N = \frac{H^3\log(1/\delta_1)}{1000000\epsilon^2\zeta^2}$, $\delta_1 = \frac{\delta\epsilon_2}{16S^2AH^3}$.

B Technical Lemmas

B.1 Concentration Inequalities

Lemma 9 (Theorem 4 in Maurer & Pontil (2009))**.** *Let $Z, Z_1, ..., Z_n$ ($n \geq 2$) be i.i.d. random variables with values in $[0,1]$ and let $\delta > 0$. Define $\bar{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$ and $\hat{V}_n = \frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z})^2$. Then we have*

$$\mathbb{P}\left[\left|\mathbb{E}\left[Z\right] - \frac{1}{n}\sum_{i=1}^{n} Z_i\right| > \sqrt{\frac{2\hat{V}_n \ln(2/\delta)}{n-1}} + \frac{7\ln(2/\delta)}{3(n-1)}\right] \leq \delta.$$

B.2 Properties of MDPs

Lemma 10 (Policy difference lemma)**.** *Given a transition model $\check{P}$, a reward function $\check{R}$ and two policies π^1, π^2, it holds that*

$$V_h^{\pi^1,\check{P}}(\check{R}, s) - V_h^{\pi^2,\check{P}}(\check{R}, s)$$
$$= \mathbb{E}_{\pi^1,\check{P}}\left[\sum_{\tau=h}^{H} \mathbb{E}_{a_\tau^i \sim \pi_\tau^i(\cdot|s_\tau), i=1,2}\left[\check{R}_{s_\tau,a_\tau^1,\tau} - \check{R}_{s_\tau,a_\tau^2,\tau} + (\check{P}_{s_\tau,a_\tau^1,\tau} - \check{P}_{s_\tau,a_\tau^2,\tau})^\top V_{\cdot,\tau+1}^{\pi,\check{P}}(\check{R})\right] |s_h = s\right]$$

for any proper (s,h).

C Missing Proofs

C.1 Proof of Lemma 1

Given $(s,h) \in \mathcal{S} \times [H]$ and $\lambda \in [0,1]$, define that (s,h,λ) is *proper* if the optimal action at (s,h) with respect to the reward $\lambda\check{R} - (1-\lambda)\check{C} + \varsigma$ is unique. Otherwise, we say (s,h,λ) is *improper*. Moreover, we say (s,h,λ,a,a') is *improper* for two actions (a,a') iff both a and a' are optimal actions at (s,h) with respect to the reward $\lambda\check{R} - (1-\lambda)\check{C} + \varsigma$.

We claim that

Lemma 11. *With probability $1-\delta$, for any $\lambda \in [0,1]$ there exists at most one state-level pair (s,h) such that (s,h,λ) is* improper.

The conclusion follows easily by Lemma 11.

C.2 Proof of Lemma 11

Define $\mathcal{G}$ to be the event that there exists some λ such that the size of $\{(s,h)|(s,h,\lambda)$ is *improper*$\}$ is at least 2. So it suffices to bound $\mathbb{P}[\mathcal{G}]$.

For $z_1 = (s_1, h_1, a_1^1, a_1^2)$ and $z_2 = (s_2, h_2, a_2^1, a_2^2)$, define $\mathcal{G}(z_1, z_2)$ to be the event that there exists $\lambda \in [0,1]$ such that $(s_1, h_1, \lambda, a_1^1, a_1^2)$ and $(s_2, h_2, \lambda, a_2^1, a_2^2)$ are both *improper*. Let $Z = \mathcal{S} \times [H] \times \mathcal{A} \times \mathcal{A}$.

Therefore,

$$\mathbb{P}\left[\mathcal{G}\right] \le \sum_{z_1,z_2\in\mathcal{Z},z_1\neq z_2} \mathbb{P}\left[\mathcal{G}(z_1,z_2)\right] \le S^2H^2A^4 \max_{z_1,z_2\in\mathcal{Z},z_1\neq z_2} \mathbb{P}\left[\mathcal{G}(z_1,z_2)\right]. \tag{15}$$

Fix a pair $(z_1, z_2) \in \mathcal{Z}^2$ such that $z_1 \neq z_2$.

Without loss of generality, we assume that $h_1 \le h_2$. For a deterministic policy π and $z = (s, h, a^1, a^2) \in \mathcal{S} \times [H] \times \mathcal{A}^2$, define

$$\begin{aligned}\Lambda(z) &= \left\{\lambda \in [0,1] \mid (s,h,\lambda,a^1,a^2) \textit{ improper}\right\};\\ \Lambda(z,\pi) &= \left\{\lambda \in [0,1] \mid Q_h^{\pi,\check{P}}(\lambda\check{R} - (1-\lambda)\check{C} + \varsigma, (s,a^1)) = Q_h^{\pi,\check{P}}(\lambda\check{R} - (1-\lambda)\check{C} + \varsigma, (s,a^2)\right\}.\end{aligned} \tag{16}$$

It then follows that $\Lambda(z) = \cup_{\pi\in\Pi}\Lambda(z,\pi)$.

Lemma 12. *Fix* $\pi \in \Pi$. *With probability* $1 - \frac{1}{2|\mathcal{Z}|^2|\Pi|^2} \cdot \delta$*,* $\Lambda(z_2, \pi)$ *is either a single-point set or an empty set.*

Lemma 13. *Fix* $\pi, \pi' \in \Pi$. *With probability* $1 - \frac{1}{|\mathcal{Z}|^2|\Pi|^2}\delta$*,* $\Lambda(z_2,\pi) \cap \Lambda(z_1,\pi') = \emptyset$*.*

By Lemma 13, we learn that

$$\mathbb{P}\left[\Lambda(z_1,\pi') \cap \Lambda(z_2,\pi) \neq \emptyset\right] \le \frac{1}{|\mathcal{Z}|^2|\Pi|^2} \cdot \delta$$

for any $\pi, \pi' \in \Pi$.

Therefore, we have that

$$\mathbb{P}[\mathcal{G}(z_1,z_2)] \le \sum_{\pi,\pi'\in\Pi} \mathbb{P}\left[\Lambda(z_1,\pi') \cap \Lambda(z_2,\pi) \neq \emptyset\right] \le \frac{1}{|\mathcal{Z}|^2}\delta. \tag{17}$$

Combining equation 17 with equation 15, we have that

$$\mathbb{P}[\mathcal{G}] \le \delta;$$

The proof is completed.

C.2.1 Proof of Lemma 12

Conditioned on $\{\varsigma_{s,a,h}\}_{(s,a)\in\mathcal{S}\times\mathcal{A},h\ge h_2+1}$, the set $\Lambda(z_2,\pi)$ is determined by $\varsigma_{s_2,a_2^1,h_2}$ and $\varsigma_{s_2,a_2^2,h_2}$. More precisely, $\lambda \in \Lambda(z_2,\pi)$ implies that

$$\left|\lambda L_1 + L_2 + \varsigma_{s_2,a_2^1,h_2} - \varsigma_{s_2,a_2^2,h_2}\right| = 0, \tag{18}$$

where

$$\begin{aligned}L_1 &= \left(Q_{h_2}^{\pi,\check{P}}(\check{R}+\check{C},(s_2,a_2^1)) - Q_{h_2}^{\pi,\check{P}}(\check{R}+\check{C},(s_2,a_2^2))\right);\\ L_2 &= \left(Q_{h_2}^{\pi,\check{P}}(\check{C}-\varsigma,(s_2,a_2^1)) - Q_{h_2}^{\pi,\check{P}}(\check{C}-\varsigma,(s_2,a_2^2))\right).\end{aligned}$$

In the case $L_1 = 0$, for fixed $\varsigma_{s_2,a_2^1,h_2}$, there is at most one value for $\varsigma_{s_2,a_2^2,h_2}$ such that equation 18 holds. As a result, $\mathbb{P}[\Lambda(z_2,\pi) = \emptyset] \ge 1 - \frac{1}{K}$.

In the case $L_1 \neq 0$, we learn that $\Lambda(z_2,\pi) = \{\frac{-L_2 - \varsigma_{s_2,a_2^1,h_2} + \varsigma_{s_2,a_2^2,h_2}}{L_1}\} \cap [0,1]$ is either a single-point set or an empty set.

C.2.2 Proof of Lemma 13

Assume $h_2 \ge h_1$. With probability $1 - \frac{1}{2|\mathcal{Z}|^2|\Pi|^2}$, $\Lambda(z_2,\pi)$ is either a single-point set or an empty set. We assume $\Lambda(z_2,\pi)$ is a single-point set $\{\overline{\lambda}\}$ and consider the randomness due to $\varsigma_{s_1,a_1^1,h_1}$ and $\varsigma_{s_1,a_1^2,h_1}$.

Case i: $(s_1, h_1) \neq (s_2, h_2)$**.** By definition, for $\lambda \in \Lambda(z_1, \pi')$ it holds that

$$\lambda L1 + L_2 + \varepsilon_{s_1, a_1^1, h_1} - \varsigma_{s_1, a_1^2, h_1} = 0,$$

where

$$L_1 = \left(Q_{h_1}^{\pi', \check{P}}(\check{R} + \check{C}, (s_1, a_1^1)) - Q_{h_1}^{\pi', \check{P}}(\check{R} + \check{C}, (s_1, a_1^2)) \right);$$
$$L_2 = \left(Q_{h_1}^{\pi', \check{P}}(\check{C} - \varsigma, (s_1, a_1^1)) - Q_{h_1}^{\pi', \check{P}}(\check{C} - \varsigma, (s_1, a_1^2)) \right).$$

In the case $L_1 = 0$, with probability $1 - \frac{1}{K}$, $\Lambda(z_1, \pi') = 0$. In the case $L_1 \neq 0$, with probability $1 - \frac{1}{K}$,

$$\overline{\lambda} L1 + L_2 + \varepsilon_{s_1, a_1^1, h_1} - \varsigma_{s_1, a_1^2, h_1} \neq 0,$$

which means $\Lambda(z_1, \pi') \cap \Lambda(z_2, \pi) = \emptyset$.

ii: $(s_1, h_1) = (s_2, h_2)$**.** Write $(s_1, h_1) = (s_2, h_2) = (s, h)$. Without loss of generality, we assume that $a_1^1 \notin \{a_2^1, a_2^2\}$. By definition, for $\lambda \in \Lambda(z_1, \pi')$ it holds that

$$\lambda L1 + L_2 + \varepsilon_{s_1, a_1^1, h_1} - \varsigma_{s_1, a_1^2, h_1} = 0,$$

where

$$L_1 = \left(Q_{h_1}^{\pi', \check{P}}(\check{R} + \check{C}, (s_1, a_1^1)) - Q_{h_1}^{\pi', \check{P}}(\check{R} + \check{C}, (s_1, a_1^2)) \right);$$
$$L_2 = \left(Q_{h_1}^{\pi', \check{P}}(\check{C} - \varsigma, (s_1, a_1^1)) - Q_{h_1}^{\pi', \check{P}}(\check{C} - \varsigma, (s_1, a_1^2)) \right).$$

In the case $L_1 = 0$, with probability $1 - \frac{1}{K}$, $\Lambda(z_1, \pi') = 0$. In the case $L_1 \neq 0$, with probability $1 - \frac{1}{K}$ (where we only consider the randomness in $\varsigma_{s_1, a_1^1, h_1}$,

$$\overline{\lambda} L1 + L_2 + \varepsilon_{s_1, a_1^1, h_1} - \varsigma_{s_1, a_1^2, h_1} \neq 0,$$

which means $\Lambda(z_1, \pi') \cap \Lambda(z_2, \pi) = \emptyset$.

The proof is completed.

C.3 Proof of Lemma 6

Let $\epsilon_2 = \frac{\epsilon_3 \zeta}{40H^2}$. Let Λ_{ϵ_2} be an ϵ_2-net of the interval $[0, 1]$ with size at most $\frac{1}{\epsilon_2} + 1$. For $\lambda \in \Lambda_{\epsilon_2}$, we define $\{\tilde{U}_h^*(s)\}_{s \in \mathcal{S}, h \in [H]}$ be the optimal value function with reward as $\lambda \tilde{R} - (1 - \lambda)\tilde{C}$ and transition model $\tilde{P}$. We also define $\{U_h^*(s)\}_{s \in \mathcal{S}, h \in [H]}$ be the optimal value function with reward as $\lambda R - (1 - \lambda) C$ and transition model P.

Define $\mathcal{E}^1(\lambda)$ be the event that

$$\left| \left(\hat{P}_{s,a,h} - P_{s,a,h} \right)^\top \tilde{U}_{h+1}^* \right| \leq 4 \sqrt{\frac{\mathbb{V}(P_{s,a,h}, \tilde{U}_{h+1}^*) \log(1/\delta_1)}{N}} + \frac{4H \log(1/\delta_1)}{N};$$
$$\left| \left(\hat{P}_{s,a,h} - P_{s,a,h} \right)^\top U_{h+1}^* \right| \leq 4 \sqrt{\frac{\mathbb{V}(P_{s,a,h}, U_{h+1}^*) \log(1/\delta_1)}{N}} + \frac{4H \log(1/\delta_1)}{N}$$

for all (s, a, h). By Lemma 9, $\mathcal{E}^1(\lambda)$ holds with probability $1 - 8SAH\delta_1$. Taking union bound over $\lambda \in \Lambda(\epsilon_2)$, we learn that $\mathcal{E}^1 := \cap_{\lambda \in \Lambda_{\epsilon_2}} E^1(\lambda)$ holds with probability at least $1 - \frac{16SAH}{\epsilon_2}\delta_1 \geq 1 - \delta$.

We continue the proof conditioned on $\mathcal{E}^1$.

Define $\tilde{F}(\lambda) = V^{*, \tilde{P}}(\lambda \tilde{R} - (1 - \lambda)\tilde{C}, s_{\text{ini}})$ and $F(\lambda) = V^{*, P}(\lambda R - (1 - \lambda) C, s_{\text{ini}})$. Then $|\tilde{F}(\lambda_1) - \tilde{F}(\lambda_2)| \leq 4H|\lambda_1 - \lambda_2|$ and $|F(\lambda_1) - F(\lambda_2)| \leq 2H|\lambda_1 - \lambda_2|$. So it suffices to show that for any $\lambda \in \Lambda_{\epsilon_2}$, $|\tilde{F}(\lambda) - F(\lambda)| \leq \frac{\epsilon_3}{8}$.

Fix $\lambda \in \Lambda_{\epsilon_2}$. Let the two deterministic policies to reach $\tilde{F}(\lambda)$ and $F(\lambda)$ be respectively $\tilde{\pi}^*$ and π^*. Let $\{\tilde{U}_h^*(s)\}$ denote the optimal value function with reward $\lambda \tilde{R} - (1 - \lambda)\tilde{C}$ and transition model $\tilde{P}$.

Define $\tilde{X} = \lambda \tilde{R} - (1-\lambda)\tilde{C}$ and $X = \lambda R - (1-\lambda)C$. By policy difference lemma (Lemma 10), we have

$$
\begin{aligned}
&\tilde{F}(\lambda) - F(\lambda) \\
&= \mathbb{E}_{\pi^*,P}\left[\sum_{h=1}^{H}\left(\tilde{X}_{s_h,\tilde{\pi}^*_h(s_h),h} - X_{s_h,\pi^*_h(s_h),h} + \left(\tilde{P}_{s_h,\tilde{\pi}^*_h(s_h),h} - P_{s_h,\pi^*_h(s_h),h}\right)^\top \tilde{U}^*_{h+1}\right)\right] \\
&\geq \mathbb{E}_{\pi^*,P}\left[\sum_{h=1}^{H}\left(\tilde{X}_{s_h,\pi^*_h(s_h),h} - X_{s_h,\pi^*_h(s_h),h} + \left(\tilde{P}_{s_h,\pi^*_h(s_h),h} - P_{s_h,\pi^*_h(s_h),h}\right)^\top \tilde{U}^*_{h+1}\right)\right] \\
&\geq -\left(2K\upsilon H + 8H\sqrt{\frac{\log(1/\delta_1)}{N}} + 8\frac{H^2\log(1/\delta_1)}{N} + 8\mathbb{E}_{\pi^*,P}\left[\sum_{h=1}^{H}\sqrt{\frac{\mathbb{V}(P_{s_h,a_h,h},\tilde{U}^*_{h+1})\log(1/\delta_1)}{N}}\right]\right) \\
&\geq -\left(2K\upsilon H + 8H\sqrt{\frac{\log(1/\delta_1)}{N}} + 8\frac{H^2\log(1/\delta_1)}{N} + 8\mathbb{E}_{\pi^*,P}\left[\sqrt{\frac{H\sum_{h=1}^{H}\mathbb{V}(P_{s_h,a_h,h},\tilde{U}^*_{h+1})\log(1/\delta_1)}{N}}\right]\right) \\
&\geq -\left(2K\upsilon H + 8H\sqrt{\frac{\log(1/\delta_1)}{N}} + 8\frac{H^2\log(1/\delta_1)}{N} + 8\sqrt{\mathbb{E}_{\pi^*,P}\left[\frac{H\sum_{h=1}^{H}\mathbb{V}(P_{s_h,a_h,h},\tilde{U}^*_{h+1})\log(1/\delta_1)}{N}\right]}\right) \\
&\geq -\frac{\epsilon_3}{8}.
\end{aligned}
$$

In the last line, we use the following bound for $\mathbb{E}_{\pi^*,P}\left[\sum_{h=1}^{H}\mathbb{V}(P_{s_h,a_h,h},\tilde{U}^*_{h+1})\right]$.

$$
\begin{aligned}
&\mathbb{E}_{\pi^*,P}\left[\sum_{h=1}^{H}\mathbb{V}(P_{s_h,a_h,h},\tilde{U}^*_{h+1})\right] \\
&= \mathbb{E}_{\pi^*,P}\left[\sum_{h=1}^{H}\left((\tilde{U}^*_{h+1}(s_{h+1}))^2 - (P^\top_{s_h,a_h,h}\tilde{U}^*_{h+1})^2\right)\right] \\
&=\leq (2H + 2HK\upsilon)\mathbb{E}_{\pi^*,P}\left[\sum_{h=1}^{H}\left|\tilde{U}^*_h(s_h) - P^\top_{s_h,a_h,h}\tilde{U}^*_{h+1}\right|\right] \\
&\leq (2H + 2HK\upsilon)^2 + (2H + 2HK\upsilon)\mathbb{E}_{\pi^*,P}\left[\sum_{h=1}^{H}\left|(\tilde{P}_{s_h,a_h,h} - P_{s_h,a_h,h})^\top\tilde{U}^*_{h+1}\right|\right] \\
&\leq (2H + 2HK\upsilon)^2 + (2H + 2HK\upsilon)\mathbb{E}_{\pi^*,P}\left[\sum_{h=1}^{H}4\sqrt{\frac{\mathbb{V}(P_{s_h,a_h,h},\tilde{U}^*_{h+1})\log(1/\delta_1)}{N}} + \frac{4H^2\log(1/\delta_1)}{N} + H^2\alpha\right] \\
&\leq 6H^2 + \frac{3H^{\frac{3}{2}}}{\sqrt{N}}\sqrt{\mathbb{E}_{\pi^*,P}\left[\sum_{h=1}^{H}\mathbb{V}(P_{s_h,a_h,h},\tilde{U}^*_{h+1})\right]} \\
&\leq 9H^2.
\end{aligned}
$$

Using similar arguments, we can show the other side $\tilde{F}(\lambda) - F(\lambda) \leq \frac{\epsilon_3}{8}$.

C.4 Proof of Lemma 7

Let $\{\tilde{U}_h(s)\}$ denote the value function following π with reward $\tilde{R}$ and transition $\tilde{P}$. Direct computation gives that

$$
\begin{aligned}
&\left|V_1^{\pi,P}(R,s_{\mathrm{ini}}) - V_1^{\pi,\tilde{R}}(C,s_{\mathrm{ini}})\right| \\
&= \left|\mathbb{E}_{\pi,P}\left[\sum_{h=1}^{H}\left(R_{s_h,a_h,h} - \tilde{R}_{s_h,a_h,h} + \left(P_{s_h,a_h,h} - \tilde{P}_{s_h,a_h,h}\right)^{\top}\tilde{U}_{h+1}\right)\right]\right| \\
&\le \mathbb{E}_{\pi,P}\left[\sum_{h=1}^{H}\left(Kv + 2\sqrt{\frac{\log(1/\delta_1)}{N}} + \sum_{h=1}^{H}\left(4\sqrt{\frac{\mathbb{V}(P_{s_h,a_h,h},\tilde{U}_{h+1})\log(1/\delta_1)}{N}} + \frac{4H\log(1/\delta_1)}{N}\right)\right)\right] \\
&\qquad\qquad + 8H\sqrt{\frac{\epsilon_1\log(1/\delta_1)}{N}} + 2H\epsilon_1 + H^2\alpha \qquad (19) \\
&\le HKv + 6H\sqrt{\frac{\log(1/\delta_1)}{N}} + 4\sqrt{\frac{H\log(1/\delta_1)}{N}}\cdot\sqrt{\mathbb{E}_{\pi,P}\left[\sum_{h=1}^{H}\mathbb{V}(P_{s_h,a_h,h},\tilde{U}_{h+1})\right]} + \frac{1}{4}\epsilon_3 \\
&\le \epsilon_3. \qquad (20)
\end{aligned}
$$

Here equation 19 is by the definition of the good event $\mathcal{E}$, and equation 20 holds by the fact that

$$
\begin{aligned}
&\mathbb{E}_{\pi,P}\left[\sum_{h=1}^{H}\mathbb{V}(P_{s_h,a_h,h},\tilde{U}_{h+1})\right] \\
&= \mathbb{E}_{\pi,P}\left[\sum_{h=1}^{H}\left((\tilde{U}_{h+1}(s_{h+1}))^2 - (P_{s_h,a_h,h}^{\top}\tilde{U}_{h+1})^2\right)\right] \\
&\le (2H+2HKv)\cdot\mathbb{E}\left[\sum_{h=1}^{H}\left|\tilde{U}_h(s_h) - P_{s_h,a_h,h}^{\top}\tilde{U}_{h+1}\right|\right] \\
&\le (2H+2HKv)^2 + (2H+2HKv)\cdot\mathbb{E}_{\pi,P}\left[\left|(\tilde{P}_{s_h,a_h,h} - P_{s_h,a_h,h})^{\top}\tilde{U}_{h+1}\right|\right] \\
&\le (2H+2HKv)^2 + (2H+2HKv)\cdot\mathbb{E}_{\pi,P}\left[4\sqrt{\frac{\mathbb{V}(P_{s_h,a_h,h},\tilde{U}_{h+1})\log(1/\delta_1)}{N}} + \frac{4H^2\log(1/\delta_1)}{N} + H^2\alpha\right] \\
&\le 6H^2 + \frac{3H^{\frac{3}{2}}}{N}\cdot\sqrt{\mathbb{E}_{\pi,P}\left[\sum_{h=1}^{H}\mathbb{V}(P_{s_h,a_h,h},\tilde{U}_{h+1})\right]} \\
&\le 9H^2.
\end{aligned}
$$

In a similar way, we could show that $|V_1^{\pi,P}(C,s_{\mathrm{ini}}) - V_1^{\pi,\tilde{C}}(C,s_{\mathrm{ini}})| \le \epsilon_3$.

The proof is completed.