Regret Guarantees for Model-Based Reinforcement Learning with Long-Term Average Constraints

Abstract

We consider the problem of constrained Markov Decision Process (CMDP) where an agent interacts with a unichain Markov Decision Process. At every interaction, the agent obtains a reward. Further, there are K cost functions. The agent aims to maximize the long-term average reward while simultaneously keeping the K long-term average costs lower than a certain threshold. In this paper, we propose CMDP-PSRL, a posterior sampling based algorithm using which the agent can learn optimal policies to interact with the CMDP. Further, for MDP with S states, A actions, and diameter D, we prove that following CMDP-PSRL algorithm, the agent can bound the regret of not accumulating rewards from optimal policy by $\tilde{O}(poly(DSA)\sqrt{T})$. Further, we show that the violations for any of the K constraints is also bounded by $\tilde{O}(poly(DSA)\sqrt{T})$. To the best of our knowledge, this is the first work which obtains a $\tilde{O}(\sqrt{T})$ regret bounds for ergodic MDPs with long-term average constraints.

1 INTRODUCTION

Reinforcement Learning and stochastic optimization using Markov Decision Process (MDP) [Puterman, 2014] is being increasingly applied in many domains such as robotics [Levine et al., 2016], recommendation systems [Shani et al., 2005], UAV trajectory optimization [Zhang et al., 2015], etc. Most of these applications aim at maximizing certain metric such as increasing number of clicks in a recommendation system or reducing the time to achieve a certain pose in robotics. Further, there has been significant theoretical analysis and near optimal algorithms using optimistic MDPs [Jaksch et al., 2010], posterior sampling [Agrawal and Jia, 2017, Osband et al., 2013], and policy gradients [Agarwal et al., 2020]. However, many applications, along with optimizing the objective, also aim to satisfy certain constraints.

As a motivating example, consider a wireless sensor network where the devices aim to update a server with sensor values. At time t, the device can choose to send a packet to obtain a reward of 1 unit or to queue the packet and obtain 0 reward. However, communicating a packet results in p_t power consumption. At time t, if the wireless channel condition, s_t, is weak and the device chooses to send a packet, the resulting instantaneous power consumption, p_t, is high. The goal is to send as many packets as possible while keep the average power consumption, $\sum_{t=1}^{T} p_t/T$, within some limit, say C. This environment has state (s_t, q_t) as the channel condition and queue length at time t. To limit the power consumption, the agent may choose to send packets when the channel condition is good or when the queue length grows beyond a certain threshold. The agent aims to learn the policies in an *online manner* which requires efficiently balancing exploration of state-space and exploitation of the estimated system dynamics [Singh et al., 2020].

Similar to the example above, many applications require to keep some costs low while simultaneously maximizing the rewards [Altman, 1999]. Owing to the importance of this problem, in this paper, we consider the problem of constrained Markov Decision Processes (constrained MDP or CMDP). We aim to develop a reinforcement learning algorithm following which an agent can bound the constraint violations and the regret in obtaining the optimal reward to $o(T)$.

The problem setup, where the system dynamics are known, is extensively studied [Altman, 1999]. For a constrained setup, the optimal policy is possibly stochastic [Altman, 1999, Puterman, 2014]. In the domain where the agent learns the system dynamics and aims to learn good policies online, there has been work where to show asymptotic convergence to optimal policies [Gattami et al., 2021], or even provide regret guarantees when the MDP is episodic [Zheng and Ratliff, 2020, Ding et al., 2021]. Recently, [Singh et al.,

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

2020] considered the problem of online optimization of infinite-horizon communicating Markov Decision Processes with long-term average constraints. They propose a mixing based policy where at every time step t, the agent selects an action from the optimal optimistic policy with probability $1-\gamma_t$ or selects an action uniformly at random with probability γ_t. Using a $\gamma_t = \Theta(t^{-1/4})$, they obtain a regret bound of $\tilde{O}\left(S\sqrt{AT^{1.5}}\right)$[1]. Additionally, finding the optimistic policy is a computationally intensive task as the number of optimization variables become $S^3 \times A^2$ for MDP with S states and A actions.

In this paper, we also consider the reinforcement learning an infinite-horizon unichain MDP [Tarbouriech and Lazaric, 2019, Gattami et al., 2021] with long-term average constraints. We eliminate the use of additional exploration using γ_t using the ergodicity of the MDP. The natural ergodicity of the MDP allows us to bound the reward regret of the MDP as $\tilde{O}(poly(DSA)\sqrt{T})$. Additionally, we also bound the constraint regret as $\tilde{O}(poly(DSA)\sqrt{T})$. We propose a posterior sampling based algorithm where we sample the transition dynamics using a Dirichlet distribution [Osband et al., 2013], which achieves this regret bound. To bound the constraints violation, we use the gap between the costs incurred by running a policy for sampled MDP and costs incurred by running a policy for true MDP. Additionally, the posterior sampling approach helps to reduces the optimization variables, to find only the optimal policy for the sampled MDP, to only $S \times A$ variables. Finally, we provide numerical examples where the algorithm converges to the calculated optimal policies. To the best of our knowledge, this is the first work to obtain $O(\sqrt{T})$ regret guarantees for the infinite horizon long-term average constraint setup.

2 RELATED WORK

Stochastic Optimization using Markov Decision Processes has very rich roots [Howard, 1960]. There have been work in understanding convergence of the algorithm to find optimal policies for known MDPs [Bertsekas and Tsitsiklis, 1996, Altman, 1999]. Also, when the MDP is not known, there are algorithms with asymptotic guarantees for learning the optimal policies [Watkins and Dayan, 1992] which maximize an objective without any constraints. Recent algorithms even achieve finite time near-optimal regret bounds with respect to the number of interactions with the environment [Jaksch et al., 2010, Osband et al., 2013, Agrawal and Jia, 2017, Jin et al., 2018]. [Jaksch et al., 2010] uses the optimism principle for minimizing regret for infinite horizon MDPs with bounded diameter. [Osband et al., 2013] extended the analysis of [Jaksch et al., 2010] to posterior sampling for episodic MDPs and bounded the Bayesian regret. [Agrawal and Jia, 2017] uses a posterior sampling based approach and obtains a frequentist regret for the infinite horizon MDPs with bounded diameter.

In many reinforcement learning settings, the agent not only wants to maximize the rewards but also satisfy certain cost constraints [Altman, 1999]. Early works in this area were pioneered by [Altman and Schwartz, 1991]. They provided an algorithm which combined forced explorations and following policies optimized on empirical estimates to obtain an asymptotic convergence. [Borkar, 2005] studied the constrained RL problem using actor-critic and a two time-scale framework [Borkar, 1997] to obtain asymptotic performance guarantees. Very recently, [Gattami et al., 2021] analyzed the asymptotic performance for Lagrangian based algorithms for infinite-horizon long-term average constraints.

Inspired by the finite-time performance analysis of reinforcement learning algorithm for unconstrained problems, there has been a significant thrust in understanding the finite-time performances of constrained MDP algorithms. [Zheng and Ratliff, 2020] considered an episodic CMDP and use an optimism based algorithm to bound the constraint violation as $\tilde{O}(\sqrt{T^{1.5}})$ with high probability. [Kalagarla et al., 2020] also considered the episodic setup to obtain PAC-style bound for an optimism based algorithm. [Ding et al., 2021] considered the setup of H-episode length episodic CMDPs with d-dimensional linear function approximation to bound the constraint violations as $\tilde{O}(d\sqrt{H^5T})$ by mixing the optimal policy with an exploration policy. [Efroni et al., 2020] proposes a linear-programming and primal-dual policy optimization algorithm to bound the regret as $O(S\sqrt{H^3T})$. [Qiu et al., 2020] proposed an algorithm which obtains a regret bound of $\tilde{O}(S\sqrt{AH^2T})$ for the problem of adversarial stochastic shortest path. Compared to these works, we focus on setting with infinite horizon long-term average constraints.

After developing a better understanding of the policy gradient algorithms [Agarwal et al., 2020], there has been theoretical work in the area of model-free policy gradient algorithms for constrained MDP and safe reinforcement learning as well. [Xu et al., 2020] consider an infinite horizon discounted setup with constraints and obtain global convergence using policy gradient algorithms. [Ding et al., 2020] also considers an infinite horizon discounted setup. They use a natural policy gradient to update the primal variable and sub-gradient descent to update the dual variable.

Recently [Singh et al., 2020] considered the setup of infinite-horizon CMDPs with long-term average constraints and obtain a regret bound of $\tilde{O}(T^{1.5})$ using an optimism based algorithm and forced explorations. We consider a similar setting with unichain CMDP and propose a posterior sampling based algorithm to bound the regret as $\tilde{O}(poly(DSA)\sqrt{T})$ using explorations assisted by the ergodicity of the MDP.

[1] $\tilde{O}(\cdot)$ hides the logarithmic terms

3 PROBLEM FORMULATION

We consider an infinite horizon discounted Markov decision process (MDP) $\mathcal{M}$, defined by the tuple $(\mathcal{S}, \mathcal{A}, P, r, c_1, \cdots, c_K)$. $\mathcal{S}$ denotes a finite set of state space with $|\mathcal{S}| = S$, and $\mathcal{A}$ denotes a finite set of actions with $|\mathcal{A}| = A$. $P : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ denotes the probability transition distribution. $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ denotes the instantaneous reward. We also assume that the initial state s_0 follows a distribution ρ. We use $[K] = \{1, 2, \cdots, K\}$ to denote the set of K constraints. $c_k : \mathcal{S} \times \mathcal{A} \to [0, 1]$ denotes cost generated by constraint $k \in [K]$. We use a stochastic policy $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$, which returns the probability of selecting action $a \in \mathcal{A}$ for any given state $s \in \mathcal{S}$.

Note that the a policy π induces a Markov chain over the state space of the MDP. Pertaining to the Markov chains generated by the policies for $\mathcal{M}$, we now state our first assumption on our Markov Decision Process.

Definition 1 (Unichain MDP)**.** *An MDP is called unichain, if for each policy π, the Markov chain induced by π is ergodic, i.e. each state is reachable from any other state.*

Assumption 1. *The MDP $\mathcal{M}$ is an unichain MDP.*

The second definition for the MDP involves the expected time to reach state $s \in \mathcal{S}$ from another state $s' \in \mathcal{S}$.

Definition 2 (Diameter)**.** *Consider the Markov Chain induced by the policy π on the MDP $\mathcal{M}$. Let $T(s'|\mathcal{M}, \pi, s)$ be a random variable that denotes the first time step when this Markov Chain enters state s' starting from state s. Then, the diameter of the MDP $\mathcal{M}$ is defined as:*

$$D(\mathcal{M}) = \max_{s' \neq s} \min_{\pi} \mathbb{E}\left[T(s'|\mathcal{M}, \pi, s)\right] \tag{1}$$

From Assumption 1, we have that the MDP $\mathcal{M}$ has a finite diameter D. After discussing the transition dynamics of the system, we move to the rewards and costs of the MDP $\mathcal{M}$.

Assumption 2. *The reward function $r(s, a)$ and the costs $c_1(s, a), \cdots, c_K(s, a)$ are known to the agent.*

We note that in most of the problems, rewards are engineered. Hence, Assumption 2 is justified in many setups. However, the system dynamics are stochastic and typically not known.

For a policy π, the expected long-term average cost are given by λ_π^k, when the policy π is followed. Also, we denote the average long-term reward using λ_π^R. Formally, λ_π^k and λ_π^R are defined as

$$\lambda_\pi^k = \mathbb{E}_{s_0, a_0, s_1, a_1, \cdots}\left[\lim_{\tau \to \infty} \frac{1}{\tau} \sum_{t=0}^{\tau} c_k\left(s_t, a_t\right)\right] \tag{2}$$

$$\lambda_\pi^R = \mathbb{E}_{s_0, a_0, s_1, a_1, \cdots}\left[\lim_{\tau \to \infty} \frac{1}{\tau} \sum_{t=0}^{\tau} r\left(s_t, a_t\right)\right] \tag{3}$$

$$s_0 \sim \rho_0(s_0),\ a_t \sim \pi(a_t|s_t),\ s_{t+1} \sim P(s_{t+1}|s_t, a_t)$$

For brevity, in the rest of the paper, $\mathbb{E}_{s_t, a_t, s_{t+1}; t \geq 0}[\cdot]$ will be denoted as $\mathbb{E}_{\rho, \pi, P}[\cdot]$, where $s_0 \sim \rho_0(s_0)$, $a_t \sim \pi(s_t|a_t)$, $s_{t+1} \sim P(s_{t+1}|s_t, a_t)$.

The objective is find a policy π^* which is the solution of the following optimization problem.

$$\max_{\pi} \lambda_\pi^R \quad \text{s.t.} \tag{4}$$

$$\lambda_\pi^1 \leq C_1 \tag{C.1}$$

$$\vdots$$

$$\lambda_\pi^K \leq C_K \tag{C.K}$$

where $C_1, \cdots, C_K$ are the bounds on the average costs which the agent needs to satisfy.

Any online algorithm starting with no prior knowledge will require to obtain estimates of transition probabilities P and obtain reward r and costs $c_k, \forall\ k \in [K]$ for each state action pair. Initially, when algorithm does not have good estimates of the model, it accumulates a regret as well as violates constraints as it does not know the optimal policy. We define reward regret $R(T)$ as the difference between the average cumulative reward obtained vs the expected rewards from running the optimal policy π^* for T steps, or

$$R(T) = T\lambda_{\pi^*}^R - \sum_{t=1}^{T} r(s_t, a_t) \tag{5}$$

Additionally, we define constraint regret $R_k(T)$ for each constraint $k \in [K]$ as the gap between the average cost incurred and constraint bounds, or

$$R_k(T) = \left(\sum_{t=1}^{T} c_k(s_t, a_t) - TC_k\right)_+, \tag{6}$$

where $(x)_+ = \max(0, x)$.

In the following section, we present a model-based algorithm to obtain this policy π^*, and reward regret and the constraint regret accumulated by the algorithm.

4 THE CMDP-PSRL ALGORITHM

For infinite horizon optimization problems (or $\tau \to \infty$), we can use steady state distribution of the state to obtain

expected long-term rewards or costs [Puterman, 2014]. We use

$$\lambda_\pi^k = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} c_k(s,a)d_\pi(s,a), \;\; \forall\, k\in[K] \tag{7}$$

$$\lambda_\pi^R = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} r(s,a)d_\pi(s,a) \tag{8}$$

where $d_\pi(s,a)$ is the steady state joint distribution of the state and actions under policy π. Thus, we have the joint optimization problem in the following form

$$\max_{d_\pi(s,a)} \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} r(s,a)d_\pi(s,a) \tag{9}$$

with the following set of constraints,

$$\sum_{a\in\mathcal{A}} d_\pi(s',a) = \sum_{s\in\mathcal{S},a\in\mathcal{A}} P(s'|s,a)d_\pi(s,a) \tag{10}$$

$$\sum_{s\in\mathcal{S},a\in\mathcal{A}} d_\pi(s,a) = 1,\;\; d_\pi(s,a)\geq 0 \tag{11}$$

$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} c_k(s,a)d_\pi(s,a) \leq C_k \;\; \forall\, k\in[K] \tag{12}$$

for all $s'\in\mathcal{S}$, $\forall\, s\in\mathcal{S}$, and $\forall\, a\in\mathcal{A}$. Equation (10) denotes the constraint on the transition structure for the underlying Markov Process. Equation (11) ensures that the solution is a valid probability distribution. Finally, Equation (12) are the constraints for the constrained MDP setup which the policy must satisfy.

Note that arguments in Equation (9) are linear, and the constraints in Equation (10) and Equation (11) are linear, this is a linear programming problem. Since convex optimization problems can be solved in polynomial time [Potra and Wright, 2000], we can use standard approaches to solve Equation (9). After solving the optimization problem, we obtain the optimal policy from the obtained steady state distribution $d^*(s,a)$ as,

$$\pi^*(a|s) = \frac{Pr(a,s)}{Pr(s)} = \frac{d^*(a,s)}{\sum_{a\in\mathcal{A}} d^*(s,a)} \quad \forall\, s\in\mathcal{S} \tag{13}$$

Since we assumed that the CMDP is unichain, the Markov Chain induced from policy π is ergodic. Hence, every state is reachable following the policy π^*, we have $Pr(s)>0$ and Equation (13) is defined for all states $s\in\mathcal{S}$.

Further, since we assumed that the induced Markov Chain is irreducible for all stationary policies, we assume Dirichlet distribution as prior for the state transition probability $P(s'|s,a)$. Dirichlet distribution is also used as a standard prior in literature [Agrawal and Jia, 2017, Osband et al., 2013]. Proposition 1 formalizes the result of the existence of a steady state distribution when the transition probability is sampled from a Dirichlet distribution.

Proposition 1. *For MDP $\widehat{\mathcal{M}}$ with state space $\mathcal{S}$ and action space $\mathcal{A}$, let the transition probabilities $\hat{P}$ come from Dirichlet distribution. Then, any stationary policy π for $\widehat{\mathcal{M}}$ will have a steady state distribution $\hat{d}_\pi$ given as*

$$\hat{d}_\pi(s') = \sum_{s\in\widehat{\mathcal{S}}} \hat{d}_\pi(s)\left(\sum_{a\in\widehat{\mathcal{A}}} \pi(a|s)P(s,a,s')\right) \forall s'\in\widehat{\mathcal{S}}.$$

Proof. Transition probabilities $P(s,a,\cdot)$ follow Dirichlet distribution, and hence they are strictly positive. Further, as the policy $\pi(a|s)$ is a probability distribution on actions conditioned on state, $\pi(a|s)\geq 0$, $\sum_a \pi(a|s)=1$. So, there is a non zero transition probability to reach from state $s\in\widehat{\mathcal{S}}$ to state $s'\in\widehat{\mathcal{S}}$. Further, all the entries of the transition probability matrix are strictly positive. And, hence the Markov Chain is 1) irreducible, as it is possible to reach any state from any other state, and 2) aperiodic, as it is possible to reach any state in a single time step from any other state. Together, we get the existence of the steady-state distribution for any policy π Lawler [2018]. □

The complete constrained posterior sampling based algorithm, which we name CMDP-PSRL, is described in Algorithm 1. The algorithm proceeds in epochs, and a new epoch is started whenever the visitation count in epoch e, $\nu_e(s,a)$, is at least the total visitations before episode e, $N_e(s,a)$, for any state action pair (Line 8). In Line 9, we sample transition probabilities $\tilde{P}$ using the updated posterior and in Line 10, we update the policy using the optimization problem specified in Equation (9).

Algorithm 1 CMDP-PSRL

1: **Input:** $\mathcal{S},\mathcal{A},r,c_1,\cdots,c_K$
2: Initialize $N(s,a,s')=1\,\forall(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}$, $\pi_e(a|s)=\frac{1}{|\mathcal{A}|}\;\forall\;(a,s)\in\mathcal{A}\times\mathcal{S}$, $e=0$, $\nu_e(s,a)=N_e(s,a)=0\,\forall(s,a)\in\mathcal{S}\times\mathcal{A}$
3: **for** time index $t=1,2,\cdots$ **do**
4: Observe state s
5: Play action $a\sim\pi(\cdot|s)$
6: Observe rewards $\{r^k\}$ and next state s'
7: $\nu_e(s,a)+=1,\; N(s,a,s')+=1$
8: **if** $\nu_e(s,a)\geq\max(1,N_e(s,a))$ for any s,a **then**
9: $\tilde{P}(s'|a,s)\sim Dir(N(s,a,s'))\;\forall\;(s,a,s')$
10: Solve steady state distribution $d(s,a)$ as the solution of the optimization problem in Equations (9-12)
11: Obtain optimal policy for next epoch, $e+1$, π_{e+1} as

$$\pi_{e+1}(a|s) = \frac{d(s,a)}{\sum_{a\in\mathcal{A}} d(s,a)}$$

12: $e=e+1$
13: $\nu_e(s,a)=0, N_e(s,a)=\sum_{e'}^{e}\nu_{e'}(s,a)\;\forall(s,a)$
14: **end if**
15: **end for**

5 REGRET ANALYSIS

We note that when optimizing for long-term average rewards and long-term average constraints, we want to simultaneously minimize the reward regret and the constraint regrets. Further, if we know the optimal policy π^* before hand, the deviations resulting from the stochasticity of the process can still result in some constraint violations. Also, since we sample a MDP, the policy which is feasible for the MDP may violate constraints on the true MDP. We want to bound this gap between K costs for the two MDPs as well.

We aim to quantify the regret from **(1)** deviation of long-term average rewards of the optimal policy because of incorrect knowledge of the MDP, and **(2)** deviation of the expected rewards and costs from following the optimal policy of the sampled MDP, and **(3)** deviation of the long-term average costs generated by the optimal policy for the sampled MDP on the sampled MDP and the long-term average costs generated by the optimal policy for the sampled MDP on the true MDP.

We now prove the regret bounds for Algorithm 1. We first give the high level ideas used in obtaining the bounds on regret. We divide the regret into regret incurred in each epoch e. Then, we use the posterior sampling lemma (Lemma 1 from Osband et al. [2013]) to obtain the equivalence between the long-term average rewards of the true MDP $\mathcal{M}$ and the long-term average rewards for the optimal value of the sampled MDP $\widehat{\mathcal{M}}$. This allows us to use the results of stationary policies for average reward criteria from Puterman [2014].

Bounding constraint violations requires additional manipulations. Note that the long-term average constraints for the true MDP may not be unique. Hence, we do not know which long-term average constraint to compare with. Further, we obtain a policy which is feasible for the sampled MDP. We now want to bound the constraint violation of this feasible policy for the sampled MDP when applied on the true MDP. To do so, we will use [Ortner et al., 2020, Lemma 5] to bound the difference between average cost of the feasible policy when applied to the sampled MDP and when applied to the true MDP.

We formally state the regret bounds and constraint violation bounds in Theorem 1 which we prove in detail in Appendix A.

Theorem 1. *The expected reward regret $\mathbb{E}[R(T)]$, and the expected constraint regret $\mathbb{E}[R_k(T)]$ of Algorithm 1 are bounded as*

$$\mathbb{E}[R(T)] \leq O\left(poly(DSA)\sqrt{T\log T}\right) \tag{14}$$

$$\mathbb{E}[R_k(T)] \leq O\left(poly(DSA)\sqrt{T\log T}\right) \quad \forall\ k \in [K] \tag{15}$$

Proof Outline. We break the cumulative regret into the regret incurred in each epoch e. We now bound the regret in each epoch by breaking the regret into two terms using the average reward criteria Puterman [2014]. This gives us:

$$\begin{aligned}
&\mathbb{E}[R_T] \\
&= \mathbb{E}\left[\sum_e \sum_{t=t_e}^{t_{e+1}-1} \left(\lambda^R_{\pi^*} - r(s_t, a_t)\right)\right] && (16)\\
&= \mathbb{E}\left[\sum_e \sum_{t=t_e}^{t_{e+1}-1} \left(\tilde{\lambda}^R_{\pi_e} - r(s_t, a_t)\right)\right] && (17)\\
&= \mathbb{E}\left[\sum_e \sum_{t=t_e}^{t_{e+1}-1} \left(\tilde{P}_{\pi_e}\tilde{v}^R_{\pi_e}(s_t) - \tilde{v}^R_{\pi_e}(s_t)\right.\right. \\
&\left.\left. + r_{\pi_e}(s_t) - r(s_t, a_t)\right)\right] && (18)\\
&= \mathbb{E}\left[\sum_e \sum_{t=t_e}^{t_{e+1}-1} \left(\tilde{P}_{\pi_e}\tilde{v}^R_{\pi_e}(s_t) - P_{\pi_e}\tilde{v}^R_{\pi_e}(s_t)\right)\right] \\
&+ \mathbb{E}\left[\sum_e \sum_{t=t_e}^{t_{e+1}-1} \left(P_{\pi_e}\tilde{v}^R_{\pi_e}(s_t) - \tilde{v}^R_{\pi_e}(s_t)\right)\right] \\
&+ \mathbb{E}\left[\sum_e \sum_{t=t_e}^{t_{e+1}-1} \left(r_{\pi_e}(s_t) - r(s_t, a_t)\right)\right] && (19)
\end{aligned}$$

The first term in Equation (19) denotes how far is transition probability matrix for the sampled MDP from the transition probability matrix induced by the policy at episode e for true MDP. We bound this term by bounding the deviation of observed empirical probability from the true transition probability. The second term bounds the deviation of the observed states to the expected states from the transition probability matrix induced by the policy at episode e for true MDP. We bound this term using the Azuma's inequality and by bounding the maximum number of epochs e. The third term is the deviation of the observed rewards from sampling action a_t from policy $\pi_e(\cdot|s_t)$.

Regarding the constraint violations, for each $k \in [K]$, we want to bound,

$$\mathbb{E}\left[R^k_T\right] = \mathbb{E}\left[\left(\sum_{t=1}^{T} c_k(s_t, a_t) - TC_k\right)_+\right] \tag{20}$$

We divide the constraint violation regret into regret over epochs as well. Now, for each epoch, we know that the constraint is satisfied by the policy for the sampled MDP.

This allows us to obtain:

$$\mathbb{E}\left[R_T^k\right]$$
$$= \mathbb{E}\left[\left(\sum_e \sum_{t=t_e}^{t_{e+1}-1} \left(c_k(s_t,a_t) - C_k\right)\right)_+\right] \tag{21}$$
$$= \mathbb{E}\left[\left(\sum_e \sum_{t=t_e}^{t_{e+1}-1} \left(c_k(s_t,a_t) - \tilde{\lambda}_{\pi_e}^k + \tilde{\lambda}_{\pi_e}^k - C_k\right)\right)_+\right] \tag{22}$$
$$= \mathbb{E}\left[\left|\sum_e \sum_{t=t_e}^{t_{e+1}-1} \left(c_k(s_t,a_t) - \tilde{\lambda}_{\pi_e}^k\right)\right|\right]$$
$$+ \mathbb{E}\left[\sum_e \sum_{t=t_e}^{t_{e+1}-1} \left(\tilde{\lambda}_{\pi_e}^k - C_k\right)_+\right] \tag{23}$$

The first term in Equation (23) denotes the difference of the long-term average costs $\tilde{\lambda}_{\pi_e}$ incurred by policy π_e on the sampled MDP with transitions $\tilde{P}$ and the costs $c(s_t, a_t)$ incurred by policy π_e on the true MDP with transitions P. We bound this term using the result of [Ortner et al., 2020, Lemma 5]. The second term is the violation of the constraint k by the feasible policy π_e for the sampled MDP. Since we are working with feasible policy here, this term is trivially 0.

We now bound the span of the long-term average costs and long-term average rewards of the sampled MDP with transition probabilities $\tilde{P}$ with the diameter of the sampled MDP, $\tilde{D}$. Now, again using the posterior sampling lemma [Osband et al., 2013, Lemma 1], we have $\mathbb{E}[\tilde{D}] = D$. To bound the deviation between the sampled transition probabilities and the true transition probabilities, we use result from Weissman et al. [2003] to bound the ℓ_1 distance of the transition probability vector given a state-action pair. Finally, summing over all the epochs with [Jaksch et al., 2010, Lemma 19] will bound the regret as $O(poly(DSA)\sqrt{T\log T})$. □

We note that the fundamental setup of unconstrained optimization ($K = 0$), the bound is loose compared to that of UCRL2 algorithm Jaksch et al. [2010]. This is because we use a stochastic policy instead of a deterministic policy. Recall that the optimal policy for CMDP setup is possibly stochastic Altman [1999].

6 EVALUATION OF THE PROPOSED ALGORITHM

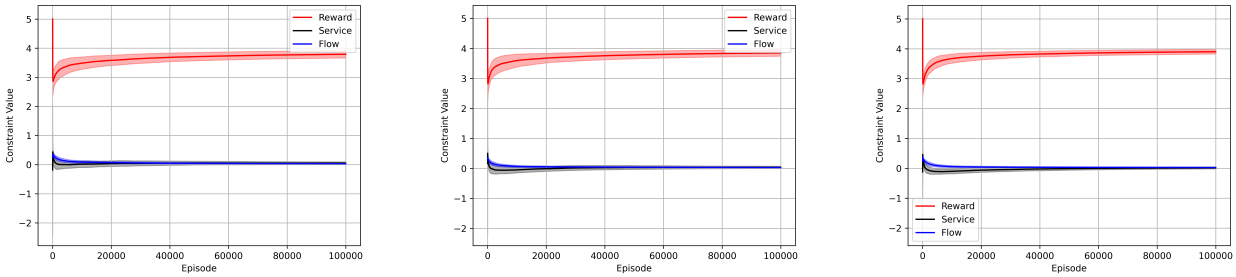
To validate the performance of the proposed CDMP-PSRL algorithm and the understanding of our analysis, we run the simulation on the flow and service control in a single-serve queue, which is introduced in [Altman and Schwartz, 1991]. A discrete-time single-server queue with a buffer of finite size L is considered in this case. The number of the customer waiting in the queue is considered as the state in this problem and thus $|S| = L + 1$. Two kinds of the actions, service and flow, are considered in the problem and control the number of customers together. The action space for service is a finite subset A in $[a_{min}, a_{max}]$, where $0 < a_{min} \leq a_{max} < 1$. Given a specific service action a, the service a customer is successfully finished with the probability b. If the service is successful, the length of the queue will reduce by 1. Similarly, the space for flow is also a finite subsection B in $[b_{min}, b_{max}]$. In contrast to the service action, flow action will increase the queue by 1 with probability b if the specific flow action b is given. Also, we assume that there is no customer arriving when the queue is full. The overall action space is the Cartesian product of the A and B. According to the service and flow probability, the transition probability can be computed and is given in the Table 1.

Define the reward function as $r(s, a, b)$ and the constraints for service and flow as $c^1(s, a, b)$ and $c^2(s, a, b)$, respectively. Define the stationary policy for service and flow as π_a and π_b, respectively. Then, the problem can be defined as

$$\begin{aligned} \max_{\pi_a,\pi_b} \quad & \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} r(s_t, \pi_a(s_t), \pi_b(s_t)) \\ s.t. \quad & \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} c^1(s_t, \pi_a(s_t), \pi_b(s_t)) \geq 0 \\ & \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} c^2(s_t, \pi_a(s_t), \pi_b(s_t)) \geq 0 \end{aligned} \tag{24}$$

According to the discussion in [Altman and Schwartz, 1991], we define the reward function as $r(s, a, b) = 5 - s$, which is an decreasing function only dependent on the state. It is reasonable to give higher reward when the number of customer waiting in the queue is small. For the constraint function, we define $c^1(s, a, b) = -10a + 6$ and $c^2 = -8 * (1 - b)^2 + 2$, which are dependent only on service and flow action, respectively. Higher constraint value is given if the probability for the service and flow are low and high, respectively.

In the simulation, the length of the buffer is set as $L = 5$. The service action space is set as $[0.2, 0.4, 0.6, 0.8]$ and the flow action space is set as $[0.5, 0.6, 0.7, 0.8]$. We use the length of horizon $T = 10^5$ and run 100 independent simulations of the proposed CMDP-PSRL algorithm. The result is shown in the Figure 2. The average values of the cumulative reward and the constraint functions are shown in the solid lines. Also, we plot the standard deviation around the mean value in the shadow to show the random error. It is found that the cumulative reward convergences to about 4. The service and flow constraints converge to 0 as expected. In order to compare this result to the optimal, we assume that



(a) $M = 1$ (b) $M = 4$ (c) $M = 16$

Figure 1: Performance of the proposed CMDP-PSRL algorithm when triggering a new episode for $\nu_e(s,a) \geq \max(1, N_e(s,a)/M)$. We note that the performance of the proposed algorithm empirically improves when we trigger a new episode early enough.

Table 1: Transition probability of the queue system

Current State	$P(x_{t+1} = x_t - 1)$	$P(x_{t+1} = x_t)$	$P(x_{t+1} = x_t + 1)$
$1 \leq x_t \leq L-1$	$a(1-b)$	$ab + (1-a)(1-b)$	$(1-a)b$
$x_t = L$	a	$1-a$	0
$x_t = 0$	0	$1-b(1-a)$	$b(1-a)$

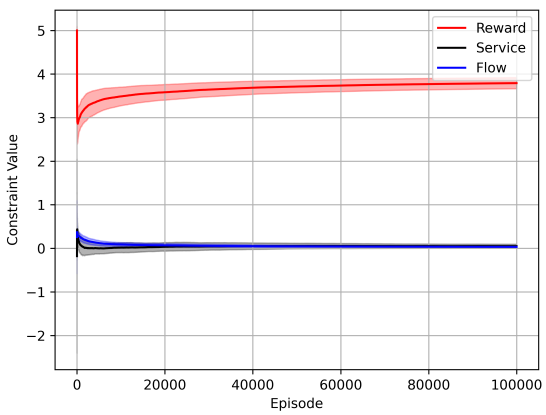


Figure 2: Performance of the proposed CMDP-PSRL algorithm on a flow and service control problem for a single queue. The average constraint violations become zero as the algorithm proceeds.

the full information of the transition dynamics is known and then use Linear Programming to solve the problem. The optimal cumulative reward from LP is shown to be 4.08. We note that the reward of the proposed CMDP-PSRL algorithm becomes closer the optimal reward as the algorithm proceeds.

We also experiment with our algorithm to improve the empirical performance. We note that the CMDP-PSRL algorithm uses a new policy after every epoch. The new policy is generated from a larger number of samples and hence it's performance is closer to the true MDP. We present the empirical results in Figure 1 for different trigger rates of the epoch. However, one cannot trigger a new policy after every time-step as this result in a large regret because the deviation of the value of the expected state and the observed state becomes large as suggested by the proof of Theorem 1 in Appendix A.

7 CONCLUSION

This paper, considers the setup of reinforcement learning in unichain infinite-horizon constrained Markov Decision Processes with K long-term average constraint. We propose a posterior sampling based algorithm, CMDP-PSRL, which proceeds in epochs. At every epoch, we sample a new CMDP and generate a solution for the constraint optimization problem. A major advantage of the posterior sampling based algorithm over an optimistic approach is, that it reduces the complexity of solving for the optimal solution of the constraint problem. We also study the proposed CMDP-PSRL algorithm from regret perspective. We bound the regret of the reward collected by the CMDP-PSRL algorithm as $O(poly(DSA)\sqrt{T \log T})$. Further, we bound the gap between the long-term average costs of the sampled MDP and the true MDP to bound the K constraint violations as $O(poly(DSA)\sqrt{T \log T})$. Finally, we evaluate the proposed CMDP-PSRL algorithm on a flow control

problem for single queue and show that the proposed algorithm performs empirically well. This paper is the first work which obtains a $\tilde{O}(\sqrt{T})$ regret bounds for ergodic MDPs with long-term average constraints. A model-free algorithm that obtains similar regret bounds for infinite horizon long-term average constraints remains an open problem.

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 64–66. PMLR, 09–12 Jul 2020. URL http://proceedings.mlr.press/v125/agarwal20a.html.

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pages 1184–1194, 2017.

E. Altman and A. Schwartz. Adaptive control of constrained markov chains. *IEEE Transactions on Automatic Control*, 36(4):454–462, 1991. doi: 10.1109/9.75103.

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

D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-dynamic programming.* Athena Scientific, Belmont, MA, 1996.

Vivek S. Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5):291–294, 1997. ISSN 0167-6911. doi: https://doi.org/10.1016/S0167-6911(97)90015-3. URL https://www.sciencedirect.com/science/article/pii/S0167691197900153.

V.S. Borkar. An actor-critic algorithm for constrained markov decision processes. *Systems & Control Letters*, 54(3):207–213, 2005. ISSN 0167-6911. doi: https://doi.org/10.1016/j.sysconle.2004.08.007. URL https://www.sciencedirect.com/science/article/pii/S0167691104001276.

Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. *Advances in Neural Information Processing Systems*, 33, 2020.

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

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

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

R. A. Howard. *Dynamic Programming and Markov Processes*. MIT Press, Cambridge, MA, 1960.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/d3b1fb02964aa64e257f9f26a31f72cf-Paper.pdf.

Krishna C Kalagarla, Rahul Jain, and Pierluigi Nuzzo. A sample-efficient algorithm for episodic finite-horizon mdp with constraints. *arXiv preprint arXiv:2009.11348*, 2020.

Gregory F Lawler. *Introduction to stochastic processes*. Chapman and Hall/CRC, 2018.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

Ronald Ortner, Pratik Gajane, and Peter Auer. Variational regret bounds for reinforcement learning. In *Uncertainty in Artificial Intelligence*, pages 81–90. PMLR, 2020.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

Florian A Potra and Stephen J Wright. Interior-point methods. *Journal of Computational and Applied Mathematics*, 124(1-2):281–302, 2000.

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

Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. Upper confidence primal-dual reinforcement learning for cmdp with adversarial loss. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15277–15287. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/ae95296e27d7f695f891cd26b4f37078-Paper.pdf.

Guy Shani, David Heckerman, Ronen I Brafman, and Craig Boutilier. An mdp-based recommender system. *Journal of Machine Learning Research*, 6(9), 2005.

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

Jean Tarbouriech and Alessandro Lazaric. Active exploration in markov decision processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 974–982. PMLR, 2019.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3):279–292, May 1992. ISSN 1573-0565. doi: 10.1007/BF00992698. URL https://doi.org/10.1007/BF00992698.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. 2003.

Tengyu Xu, Yingbin Liang, and Guanghui Lan. A primal approach to constrained policy optimization: Global optimality and finite-time analysis. *arXiv preprint arXiv:2011.05869*, 2020.

Baochang Zhang, Zhili Mao, Wanquan Liu, and Jianzhuang Liu. Geometric reinforcement learning for path planning of uavs. *Journal of Intelligent & Robotic Systems*, 77(2): 391–409, 2015.

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

A PROOF FOR REGRET BOUNDS

We now prove the regret bounds of Algorithm 1. We use the set of variables and symbols described in Table 2 for our analysis.

Variable	Definition	Description
$\mathcal{M}$		True CMDP
$P(s, a, s')$		Transition probabilities for landing in state s' from state-action pair (s, a) for $\mathcal{M}$
$r(s, a)$		reward obtained by taking action a in state s for $\mathcal{M}$
$N_e(s, a, s')$		Visitation count of (s, a, s') pair till epoch e
$\hat{P}_e(s, a, s')$	$\frac{N_e(s,a,s')}{\sum_{s''} N(s,a,s'')}$	Estimated probability of landing in s' from s, a at epoch e
$\tilde{P}(s, a, \cdot)$	$Dir(N_e(s, a, \cdot)$	Sampled transition probabilities
$\tilde{\mathcal{M}}$		Sampled CMDP with transition probabilities as $\tilde{P}$

Table 2: Glossary of the variables and symbols used in the analysis

Note that the optimal return will be unique, however the policy achieving the optimal return may not be unique. We first describe a high level approach used in obtaining the bounds on regret and then we formally obtain the required result.

We perform the following sequence of steps to bound the reward regret:

- Bound the rewards obtained by the agent from the average reward of the optimal policy, by break the total regret into regret of epochs.
- Bound the total number of epochs of Algorithm 1 (Lemma 2).
- **Equivalence of the expected reward:** Use Lemma 1 to obtain the equivalence between the expected return of an optimal policy of the true CMDP $\mathcal{M}$ and the value of an optimal policy of the sampled CMDP $\tilde{\mathcal{M}}_e$.
- Use average reward-criteria from Puterman [2014] to split the regret in each epoch into:
  1. **Deviation from expected rewards of policy:** Deviation of the observed rewards and the obtained rewards. We will use Azuma-Hoeffdings inequality (Lemma 3) to bound this term.
  2. **Deviation from true transition probabilities:** Regret because of not knowing the true transition probabilities. We will use results from Lemma 4 to bound the ℓ_1 distance of the estimated probabilities to true transition probabilities.
  3. **Deviation from expected next state:** Deviation of the gain of the observed state from the expected gain of the next state. We will again use Azuma-Hoeffdings inequality (Lemma 3).
- Finally, we will sum up the regret incurred in each epoch, over all epochs, to bound the regret.

Since the optimal policy is not unique, the expected costs may be different for each of the feasible policy. This prohibits to use the equivalence of the expected cost. To bound the constraint regret for any constraint, we perform the following steps:

- Bound the cost incurred by by the agent from the average reward of the optimal policy, by breaking the total regret into regret of epochs.
- Bound the total number of epochs of Algorithm 1 (Lemma 2).
- Break the constraint regret into two parts. First is deviation of incurred cost from running the optimal policy for the sampled CMDP $\tilde{\mathcal{M}}_e$ on the true CMDP $\mathcal{M}$ and the long-term average costs for $\tilde{\mathcal{M}}_e$, or $c(s_t, a_t) - \tilde{\lambda}_{\tilde{\pi}_e}$. Second is the constraint violation by the long-term average costs for $\tilde{\mathcal{M}}_e^k$ or $C_k - \tilde{\lambda}^k_{\tilde{\pi}_e}$.

- Similar to reward regret, bound:
  1. **Deviation from expected costs of policy:** Deviation of the observed rewards and the obtained rewards. We will use Azuma-Hoeffdings inequality (Lemma 3) to bound this term.
  2. **Deviation from true transition probabilities:** Regret because of not knowing the true transition probabilities. We will use results from Lemma 4 to bound the ℓ_1 distance of the estimated probabilities to true transition probabilities.
  3. **Deviation from expected next state:** Deviation of the gain of the observed state from the expected gain of the next state. We will again use Azuma-Hoeffdings inequality (Lemma 3).
- Finally, we will sum up the violations incurred in each epoch, over all epochs, to bound the constraint violations.

We first divide the regret into regret incurred in each epoch e. We compute the regret incurred by the optimal policy $\tilde{\pi}$ for the sampled CMDP $\widehat{\mathcal{M}}$. To bound the regret incurred by the optimal policy $\tilde{\pi}$ for the sampled CMDP $\widehat{\mathcal{M}}$ we use the results of stationary policies for average reward criteria from Puterman [2014].

Now we follow the proof style from Jaksch et al. [2010] to obtain the required results. We first state some auxiliary lemmas required for completion of the proof.

Lemma 1 (Posterior Sampling, Lemma 1 in Osband et al. [2013])**.** *For any $\sigma(H_t)$-measurable function g, if P follows distribution ϕ, then for transition probabilities $\tilde{P}$ sampled from ϕ we have,*

$$\mathbb{E}\left[g(P)|\sigma(H_t)\right] = \mathbb{E}\left[g(\tilde{P})|\sigma(H_t)\right] \tag{25}$$

The next lemma bounds the number of time the algorithm samples a transition matrix and generates a new policy.

Lemma 2. *The total number of epochs E for the CDMP-PSRL Algorithm 1 up to step $T \geq SA$ is upper bounded as*

$$E \leq 1 + AS + AS\log_2\left(\frac{T}{SA}\right) \tag{26}$$

Proof. The proof follows similar to the proof of [Jaksch et al., 2010, Proposition 18].

□

Corollary 1. *The total number of epochs E for the modified CDMP-PSRL Algorithm 1 up to step $T \geq SA$, where epochs are triggered when $\nu_e(s,a) \geq N_e(s,a)/M$, is upper bounded as*

$$E \leq 1 + AS + AS\log_{(1+\frac{1}{M})}\left(\frac{T}{SA}\right) \leq 1 + AS + MAS\log_2\left(\frac{T}{SA}\right) \tag{27}$$

The second lemma is the Azuma-Hoeffding's inequality, which we use to bound Martingale difference sequences.

Lemma 3 (Azuma-Hoeffding's Inequality)**.** *Let $X_1, \cdots, X_n$ be a Martingale difference sequence such that $|X_i| \leq c$ for all $i \in \{1, 2, \cdots, n\}$, then,*

$$\mathbb{P}\left(|\sum_{i=1}^{n} X_i| \geq \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2}{2nc^2}\right) \tag{28}$$

Corollary 2. *Let $X_1, \cdots, X_n$ be a Martingale difference sequence such that $|X_i| \leq c$ for all $i \in \{1, 2, \cdots, n\}$, then,*

$$\mathbb{E}\left(|\sum_{i=1}^{n} X_i|\right) \leq \mathcal{O}(c\sqrt{n\log n}) \tag{29}$$

We can now use the Azuma-Hoeffding's inequality to upper bound the expected value of the expected value of the absolute value of the sum of the n terms of the Martingale difference sequence $\{X_i\}_{i=1}^{n}$.

Proof.

$$\mathbb{E}\left(|\sum_{i=1}^{n} X_i|\right) \leq c\sqrt{n\log n}\mathbb{P}\left(|\sum_{i=1}^{n} X_i| \leq c\sqrt{n\log n}\right) + cn\mathbb{P}\left(|\sum_{i=1}^{n} X_i| \geq c\sqrt{n\log n}\right) \tag{30}$$

$$\leq c\sqrt{n\log n} + cn\left(2\exp\left(-\frac{c^2 n\log n}{2nc^2}\right)\right) \tag{31}$$

$$= c\sqrt{n\log n} + cn\left(2\exp\left(-\frac{\log n}{2}\right)\right) \tag{32}$$

$$= c\sqrt{n\log n} + cn\frac{2}{\sqrt{n}} \tag{33}$$

$$= 3c\sqrt{n\log n} \tag{34}$$

$$\tag{35}$$

where Equation (31) follows by putting $\epsilon = c\sqrt{n\log n}$ in Equation (28). □

We also want to bound the deviation of the estimates of the estimated transition probabilities of the Markov Decision Processes $\mathcal{M}$. For that we use ℓ_1 deviation bounds from Weissman et al. [2003]. Consider, the following event,

$$\mathcal{E}_t = \left\{\|\hat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1 \leq \sqrt{\frac{14S\log(2AT)}{\max\{1, n(s,a)\}}} \forall (s,a) \in \mathcal{S}\times\mathcal{A}\right\} \tag{36}$$

where $n = \sum_{t'=1}^{t} \mathbf{1}_{\{s_{t'}=s, a_{t'}=a\}}$. Then we have, the following lemma:

Lemma 4. *The probability that the event $\mathcal{E}_t$ fails to occur us upper bounded by $\frac{1}{20t^6}$.*

Proof. From the result of Weissman et al. [2003], the ℓ_1 distance of a probability distribution over S events with n samples is bounded as:

$$\mathbb{P}\left(\|P(\cdot|s,a) - \hat{P}(\cdot|s,a)\|_1 \geq \epsilon\right) \leq (2^S - 2)\exp\left(-\frac{n\epsilon^2}{2}\right) \leq (2^S)\exp\left(-\frac{n\epsilon^2}{2}\right) \tag{37}$$

This, for $\epsilon = \sqrt{\frac{2}{n(s,a)}\log(2^S 20SAt^7)} \leq \sqrt{\frac{14S}{n(s,a)}\log(2At)} \leq \sqrt{\frac{14S}{n(s,a)}\log(2AT)}$ gives,

$$\mathbb{P}\left(\|P(\cdot|s,a) - \hat{P}(\cdot|s,a)\|_1 \geq \sqrt{\frac{14S}{n(s,a)}\log(2At)}\right) \leq (2^S)\exp\left(-\frac{n(s,a)}{2}\frac{2}{n(s,a)}\log(2^S 20SAt^7)\right) \tag{38}$$

$$= 2^S\frac{1}{2^S 20SAt^7} \tag{39}$$

$$= \frac{1}{20ASt^7} \tag{40}$$

We sum over the all the possible values of $n(s,a)$ till t time-step to bound the probability that the event $\mathcal{E}_t$ does not occur as:

$$\sum_{n(s,a)=1}^{t} \frac{1}{20SAt^7} \leq \frac{1}{20SAt^6} \tag{41}$$

Finally, summing over all the s, a, we get,

$$\mathbb{P}\left(\|P(\cdot|s,a) - \hat{P}(\cdot|s,a)\|_1 \geq \sqrt{\frac{14S}{n(s,a)}\log(2At)}\ \forall s,a\right) \leq \frac{1}{20t^6} \tag{42}$$

□

Lemma 5 (Bounded Span of CMDP)**.** *For an CMDP $\mathcal{M}$, and for any stationary policy π with average reward ρ, the difference of bias of any two states s, and s', is upper bounded by the diameter of the CMDP D as:*

$$V(s) - V(s') \leq D \; \forall \; s, s' \in \mathcal{S}. \tag{43}$$

Proof. Consider two states $s, s' \in \mathcal{S}$ such that $V(s) \geq V(s')$. Also, let τ be a random variable defined as:

$$\tau = \min\{t \geq 1 : s_t = s', s_1 = s\} \tag{44}$$

Then, we have

$$\begin{aligned}
V(s) &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} V^{(t)}(s) && (45)\\
\implies \mathbb{E}^{\pi'}\left[V(s)\right] &= \mathbb{E}^{\pi'}\left[\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} V^{(t)}(s)\right] && (46)\\
\implies V(s) &= \mathbb{E}^{\pi'}\left[\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} V^{(t)}(s)\right] && (47)\\
&\leq \mathbb{E}^{\pi'}\left[\lim_{T\to\infty} \frac{1}{T} \sum_{t=\tau}^{T} V^{(t-\tau)}(s') + \rho\tau\right] && (48)\\
&\leq \mathbb{E}^{\pi'}\left[V(s') + \rho\tau\right] && (49)\\
&\leq V(s') + \rho\mathbb{E}^{\pi'}\left[\tau\right] && (50)\\
\implies V(s) - V(s') &\leq \rho\mathbb{E}^{\pi'}\left[\tau\right] && (51)\\
&\leq \rho D && (52)
\end{aligned}$$

where $V^{(t)}(s)$ is the cumulative reward value obtained till t time steps. Equation (52) follows from choosing policy π' which achieves the diameter. Now, for $\rho \leq 1$, we get the required result. □

Corollary 3 (Bounded Cost Span of CMDP)**.** *For an CMDP $\mathcal{M}$, any stationary policy π, and constraint k with average cost ρ_k, the difference of bias of any two states s, and s', is upper bounded by the diameter of the CMDP D as:*

$$V_k(s) - V_k(s') \leq D \; \forall \; s, s' \in \mathcal{S}. \tag{53}$$

After proving all the auxiliary lemmas, we are now ready to prove the main theorem.

Theorem 2. *The expected regret $\mathbb{E}\left[R(T)\right]$, and constraint violation regret $\mathbb{E}\left[R_k(T)\right]$ of Algorithm 1 for CMDP with Dirichlet priors and expected diameter D are bounded as:*

$$\mathbb{E}\left[R(T)\right] \leq \tilde{O}\left(DAS\sqrt{AT}\right) \tag{54}$$

$$\mathbb{E}\left[R_k(T)\right] \leq \tilde{O}\left(DAS\sqrt{AT}\right) \tag{55}$$

Proof. We start from the definition of regret $R(T)$ in Equation (5). Note that the algorithm proceeds in epochs. Hence, we can bound the regret of each epoch and then summing over all the epochs to bound the total cumulative regret for which the

algorithm runs. We obtain the following set of equations:

$$\mathbb{E}\left[R(T)\right] = \mathbb{E}\left[T\lambda_{\pi^*} - \sum_{t=0}^{T} r^k(s_t, a_t)\right] \tag{56}$$

$$= \mathbb{E}\left[\sum_{t=0}^{T} \left(\lambda_{\pi^*} - r^k(s_t, a_t)\right)\right] \tag{57}$$

$$= \mathbb{E}\left[\sum_{e=1}^{E} \sum_{t=t_e}^{t_{e+1}-1} \left(\lambda_{\pi^*} - r^k(s_t, a_t)\right)\right] \tag{58}$$

$$\leq \mathbb{E}\left[\sum_{e=1}^{E} \sum_{t=t_e}^{t_{e+1}-1} \left(\lambda_{\pi^*} - r^k(s_t, a_t)\right)\right] \tag{59}$$

$$\leq \sum_{e=1}^{E} \mathbb{E}\left[\Delta_e\right] \tag{60}$$

where Δ_e is the regret in each epoch e, and E is the total number of epochs. Using Lemma 2, we have that the total number of epochs E bounded by $SA\log(T/(SA))$.

We now bound the expected regret $\mathbb{E}\left[|\Delta_e|\right]$ for all epoch e using the gain-bias relationship as:

$$\mathbb{E}\left[\Delta_e\right] = \mathbb{E}\left[\sum_{t=t_e}^{t_{e+1}-1} \left(\lambda_{\pi^*} - r^k(s_t, a_t)\right)\right] \tag{61}$$

$$= \mathbb{E}\left[\sum_{t=t_e}^{t_{e+1}-1} \left(\tilde{\lambda}^k_{\tilde{\pi}} - r^k(s_t, a_t)\right)\right] \tag{62}$$

$$= \mathbb{E}\left[\sum_{t=t_e}^{t_{e+1}-1} \left((\tilde{P}_{\tilde{\pi}_e}\tilde{V}^k_{\tilde{\pi}_e})(s_t) - \tilde{V}^k_{\tilde{\pi}_e}(s_t) + r^k_{\tilde{\pi}_e}(s_t) - r^k(s_t, a_t)\right)\right] \tag{63}$$

$$= \mathbb{E}\left[\sum_{t=t_e}^{t_{e+1}-1} \left((\tilde{P}_{\tilde{\pi}_e}\tilde{V}^k_{\tilde{\pi}_e})(s_t) - \tilde{V}^k_{\tilde{\pi}_e}(s_t)\right)\right] + \mathbb{E}\left[\sum_{t=t_e}^{t_{e+1}-1} \left(r^k_{\tilde{\pi}_e}(s_t) - r^k(s_t, a_t)\right)\right] \tag{64}$$

$$= R_1(e) + R_2(e) \tag{65}$$

where Equation (62) follows from Lemma 1, and Equation (63) follows from the Theorem 8.2.6 of Puterman [2014].

The $R_1(e)$ term denotes how far are we from the optimal policy. The optimal policy depends on the accuracy of our model. Hence, we would see the effect of number of samples $N_e(s, a)$ collected to create the model. We now bound the $R_1(e)$ term as follows:

$$R_1(e) = \mathbb{E}\left[\sum_{t=t_e}^{t_{e+1}-1} \left((\tilde{P}_{\tilde{\pi}_e}\tilde{V}^k_{\tilde{\pi}_e})(s_t) - \tilde{V}^k_{\tilde{\pi}_e}(s_t)\right)\right] \tag{66}$$

$$= \mathbb{E}\left[\sum_{t=t_e}^{t_{e+1}-1} \left((\tilde{P}_{\tilde{\pi}_e}\tilde{V}^k_{\tilde{\pi}_e})(s_t) - (P_{\tilde{\pi}_e}\tilde{V}^k_{\tilde{\pi}_e})(s_t) + (P_{\tilde{\pi}_e}\tilde{V}^k_{\tilde{\pi}_e})(s_t) - \tilde{V}^k_{\tilde{\pi}_e}(s_t)\right)\right] \tag{67}$$

$$= \mathbb{E}\left[\sum_{t=t_e}^{t_{e+1}-1} \left((\tilde{P}_{\tilde{\pi}_e}\tilde{V}^k_{\tilde{\pi}_e})(s_t) - (P_{\tilde{\pi}_e}\tilde{V}^k_{\tilde{\pi}_e})(s_t)\right)\right] + \mathbb{E}\left[\sum_{t=t_e}^{t_{e+1}-1} \left((P_{\tilde{\pi}_e}\tilde{V}^k_{\tilde{\pi}_e})(s_t) - \tilde{V}^k_{\tilde{\pi}_e}(s_t)\right)\right] \tag{68}$$

$$= \mathbb{E}\left[R_s(e)\right] + \mathbb{E}\left[R_g(e)\right] \tag{69}$$

where R_s term refers to the regret incurred because of following optimal policy for sampled CMDP instead of the true CMDP, and R_t refers to the gap between the expected number of visitations for a state action pair and the true visitations.

We now bound the three terms and their summations over the epochs.

Deviation from expected rewards of policy:

Note that the $R_2(e)$ term is the deviation of observed reward $r^k(s_t, a_t)$ from the expected reward of the policy $\tilde{\pi}_e$, $r_{\tilde{\pi}_e^k(s_t)}$. Further, $r_{\tilde{\pi}_e^k(s_t)} = \mathbb{E}[r^k(s_t, a_t)|s_t]$, and $r^k(s_t, a_t) - r_{\tilde{\pi}_e^k(s_t)}$ is a zero mean Martingale adapted to filtration $\{\sigma(H_k)\}_{k=t_e}^{t_{e+1}}$. Hence, we can bound the R_2 term using Azuma-Hoeffding's Lemma as,

$$
\begin{aligned}
R_2(e) &= \mathbb{E}\left[\sum_{t=t_e}^{t_{e+1}-1} \left(r^k_{\tilde{\pi}_e}(s_t) - r^k(s_t, a_t)\right)\right] && (70)\\
&\leq \sqrt{2(t_{e+1}-t_e)\log(t_{e+1}-t_e)}\,\mathbb{P}\left(\sum_{t=t_e}^{t_{e+1}-1} \left(r^k_{\tilde{\pi}_e}(s_t) - r^k(s_t, a_t)\right) < \sqrt{2(t_{e+1}-t_e)\log(t_{e+1}-t_e)}\right)\\
&\quad + (t_{e+1}-t_e)\,\mathbb{P}\left(\sum_{t=t_e}^{t_{e+1}-1} \left(r^k_{\tilde{\pi}_e}(s_t) - r^k(s_t, a_t)\right) \geq \sqrt{2(t_{e+1}-t_e)\log(t_{e+1}-t_e)}\right) && (71)\\
&\leq \sqrt{2(t_{e+1}-t_e)\log(t_{e+1}-t_e)} + \frac{1}{(t_{e+1}-t_e)^2}(t_{e+1}-t_e) && (72)\\
&\leq \sqrt{2(t_{e+1}-t_e)\log(t_{e+1}-t_e)} + \frac{1}{(t_{e+1}-t_e)} && (73)\\
&\leq \sqrt{2(t_{e+1}-t_e)\log T} + \frac{1}{(t_{e+1}-t_e)} && (74)
\end{aligned}
$$

Here, Equation (72) is obtained from Equation (71) by bounding the first probability term using the upper bound of 1 and by bounding the second probability term using the Azuma-Hoeffding's inequality (Equation (28)). We can now sum over Equation (74) over all epochs and bound the total deviation using Cauchy-Schwarz inequality as:

$$
\begin{aligned}
\sum_{e=1}^{E} R_2(e) &= \sum_{e=1}^{E}\left(\sqrt{2(t_{e+1}-t_e)\log T} + \frac{1}{(t_{e+1}-t_e)}\right) && (75)\\
&\leq \sqrt{2\left(\sum_{e=1}^{E}\right)\sum_{e=1}^{E}(t_{e+1}-t_e)\log T} + \sum_{e=1}^{E}\frac{1}{(t_{e+1}-t_e)} && (76)\\
&\leq \sqrt{2(E)(\log T)\sum_{e=1}^{E}(t_{e+1}-t_e)} + E\max_{e\in\{1,2,\cdots,E\}}\frac{1}{(t_{e+1}-t_e)} && (77)\\
&\leq \sqrt{\left(2SA\log_2\left(\frac{T}{SA}\right)\right)(T\log T)} + E && (78)\\
&\leq (\log_2 T)\sqrt{2SAT} + SA\log_2 T && (79)
\end{aligned}
$$

where Equation (78) follows from the fact that $t_{e+1} - t_e \geq 1$, and the the value of E comes from the Jaksch et al. [2010].

Deviation from expected rewards of policy:

We first consider the case where the estimated probability distribution lies in some neighborhood of the true distribution. Particularly, for all s, a we construct the set of probability distributions $P'(\cdot|s,a)$,

$$
\mathcal{P}_t = \left\{P' : \|\hat{P}(\cdot|s,a) - P'(\cdot|s,a)\|_1 \leq \sqrt{\frac{14S\log(2AT)}{\max\{1, n(s,a)\}}}\right\} \tag{80}
$$

where $n = \sum_{t'}^{t} \mathbf{1}_{\{s_{t'}=s, a_{t'}=a\}}$. Using the construction of the set $\mathcal{P}_t$, we can now define the events $\mathcal{E}_t$, and $\tilde{\mathcal{E}}_t$ as:

$$
\begin{aligned}
\mathcal{E}_t &= \{P \in \mathcal{P}_t\} && (81)\\
\tilde{\mathcal{E}}_t &= \left\{\tilde{P} \in \mathcal{P}_t\right\} && (82)
\end{aligned}
$$

Further, note that $\mathcal{P}_t$ is $\sigma(H_t)$ measurable and hence from Lemma 1 we have $\mathbb{P}(\tilde{\mathcal{E}}_t) = \mathbb{P}(\tilde{P} \in \mathcal{P}_t) = \mathbb{P}(P \in \mathcal{P}_t) = \mathbb{P}(\mathcal{E}_t)$.

We now consider the case when the event $\mathcal{E}_t$ occurs whenever we sample the transition probabilities to update the policy at time step t_e. The expected value of $R_s(e)$ when event $\mathcal{E}_{t_e}$ and $\tilde{\mathcal{E}}_{t_e}$ holds is:

$$\mathbb{E}\left[R_s(e)|\mathcal{E}_{t_e}, \tilde{\mathcal{E}}_{t_e}\right] = \mathbb{E}\left[\sum_{t=t_e}^{t_{e+1}-1} \left((\tilde{P}_{\tilde{\pi}_e}\tilde{V}^k_{\tilde{\pi}_e})(s_t) - (P_{\tilde{\pi}_e}\tilde{V}^k_{\tilde{\pi}_e})(s_t)\right) \middle| \mathcal{E}_{t_e}, \tilde{\mathcal{E}}_{t_e}\right] \tag{83}$$

$$= \mathbb{E}\left[\sum_{t=t_e}^{t_{e+1}-1} \sum_{a_t\in\mathcal{A}} \tilde{\pi}_e(a_t|s_t) \sum_{s_{t+1}\in\mathcal{S}} \left((\tilde{P}(s_{t+1}|s_t,a_t)\tilde{V}^k_{\tilde{\pi}_e})(s_t) - (P(s_{t+1}|s_t,a_t)\tilde{V}^k_{\tilde{\pi}_e})(s_t)\right) \middle| \mathcal{E}_{t_e}, \tilde{\mathcal{E}}_{t_e}\right] \tag{84}$$

$$= \mathbb{E}\left[\sum_{t=t_e}^{t_{e+1}-1} \sum_{a_t\in\mathcal{A}} \tilde{\pi}_e(a_t|s_t) \sum_{s_{t+1}\in\mathcal{S}} \left(\tilde{P}(s_{t+1}|s_t,a_t) - P(s_{t+1}|s_t,a_t)\right) \tilde{V}^k_{\tilde{\pi}_e}(s_t) \middle| \mathcal{E}_{t_e}, \tilde{\mathcal{E}}_{t_e}\right] \tag{85}$$

$$\leq \mathbb{E}\left[\sum_{t=t_e}^{t_{e+1}-1} \sum_{a_t\in\mathcal{A}} \tilde{\pi}_e(a_t|s_t) \sum_{s_{t+1}\in\mathcal{S}} \left(\tilde{P}(s_{t+1}|s_t,a_t) - P(s_{t+1}|s_t,a_t)\right) \tilde{D} \middle| \mathcal{E}_{t_e}, \tilde{\mathcal{E}}_{t_e}\right] \tag{86}$$

$$\leq \mathbb{E}\left[\sum_{t=t_e}^{t_{e+1}-1} \sum_{a_t\in\mathcal{A}} \tilde{D}\|\tilde{P}(s_{t+1}|s_t,a_t) - P(s_{t+1}|s_t,a_t)\|_1 \middle| \mathcal{E}_{t_e}, \tilde{\mathcal{E}}_{t_e}\right] \tag{87}$$

$$\leq \mathbb{E}\left[\tilde{D}\sum_{t=t_e}^{t_{e+1}-1} \sum_{a_t\in\mathcal{A}} \sqrt{\frac{14S\log(2AT)}{\max\{1, N_e(s_t,a_t)\}}} \middle| \mathcal{E}_{t_e}, \tilde{\mathcal{E}}_{t_e}\right] \tag{88}$$

$$\leq D\sum_{t=t_e}^{t_{e+1}-1} \sum_{a_t\in\mathcal{A}} \sqrt{\frac{14S\log(2AT)}{\max\{1, N_e(s_t,a_t)\}}} \tag{89}$$

$$= D\sum_{a\in\mathcal{A}}\left(\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \nu_e(s,a)\sqrt{\frac{14S\log(2AT)}{\max\{1, N_e(s,a)\}}}\right) \tag{90}$$

where Equation (86) follows from the fact that the value of $\max_s \tilde{V}^k_{\tilde{\pi}_e}(s) - \min_s \tilde{V}^k_{\tilde{\pi}_e}(s)$ is bounded by $\tilde{D}$ for any stationary policy with maximum reward 1 (Lemma 5) and translating $\tilde{V}^k_{\tilde{\pi}_e}(s)$ does not change the gain $\lambda_{\tilde{\pi}_e}$ of the stationary policy $\tilde{\pi}_e$. Equation (87) comes from the fact that $\tilde{\pi}(a_t|s_t) \leq 1$. Further, we have $\mathbb{E}[\tilde{D}] = D$ from Lemma 1. Now, summing over the epochs, we get the total regret from choosing the optimal policy for the sampled transition probabilities instead of the true transition probabilities as:

$$\sum_{e=1}^{E}\mathbb{E}\left[R_s(e)|\mathcal{E}_{t_e}, \tilde{\mathcal{E}}_{t_e}\right] = \sum_{e=1}^{E} D\sum_{a\in\mathcal{A}}\left(\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \nu_e(s,a)\sqrt{\frac{14S\log(2AT)}{\max\{1, N_e(s,a)\}}}\right) \tag{91}$$

$$= \sum_{a\in\mathcal{A}} D\sqrt{14S\log(2AT)}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{e=1}^{E}\frac{\nu_e(s,a)}{\sqrt{\max\{1, N_e(s,a)\}}} \tag{92}$$

$$\leq (\sqrt{2}+1)\sum_{a\in\mathcal{A}} D\sqrt{14S\log(2AT)}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sqrt{N(s,a)} \tag{93}$$

$$\leq (\sqrt{2}+1)\sum_{a\in\mathcal{A}} D\sqrt{14S\log(2AT)}\sqrt{SA\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} N(s,a)} \tag{94}$$

$$= (\sqrt{2}+1)AD\sqrt{14S\log(2AT)}\sqrt{SAT} \tag{95}$$

$$= (\sqrt{2}+1)ASD\sqrt{14AT\log(2AT)} \tag{96}$$

where Equation (93) follows from [Jaksch et al., 2010, Lemma 19].

We now consider the other case, where the event in Equation (80) does not occur. We already bounded the probability of this

event in Lemma 4 using result from Weissman et al. [2003]. In particular, we have:

$$\sum_{e=1}^{E} \mathbb{E}\left[R_s(e)|\mathcal{E}_{t_e}^c \cup \tilde{\mathcal{E}}_{t_e}^c\right] \leq \sum_{e=1}^{E} \sum_{s,a} \nu_k(s,a) \mathbb{P}(\mathcal{E}_{t_e}^c) \tag{97}$$

$$\leq \sum_{e=1}^{E} t_k \mathbb{P}(\mathcal{E}_{t_e}^c \cup \tilde{\mathcal{E}}_{t_e}^c) \tag{98}$$

$$\leq \sum_{t=1}^{T} t \left(\mathbb{P}(\mathcal{E}_t^c) + \mathbb{P}(\tilde{\mathcal{E}}_{t_e}^c)\right) \tag{99}$$

$$\leq 2 \sum_{t=1}^{T} t \mathbb{P}(\mathcal{E}_t^c) \tag{100}$$

$$\leq 2 \sum_{t=1}^{T^{1/4}} t \mathbb{P}(\mathcal{E}_t^c) + 2 \sum_{t=T^{1/4}+1}^{T} t \mathbb{P}(\mathcal{E}_t^c) \tag{101}$$

$$\leq 2 \sum_{t=1}^{T^{1/4}} t.1 + 2 \sum_{t=T^{1/4}+1}^{T} t \frac{1}{t^6} \tag{102}$$

$$\leq 2\sqrt{T} + 2 \int_{t=T^{1/4}}^{\infty} \frac{1}{t^5} \tag{103}$$

$$\leq 2\sqrt{T} + 2 \frac{1}{4T} \tag{104}$$

$$\leq 4\sqrt{T} \tag{105}$$

where Equation (98) follows from the fact that $\sum_{s,a} \nu_k(s,a) \geq \sum_{s,a} N_k(s,a) = t_k$. Further, Equation (102) follows from Lemma 4.

This analysis gives the bound on the sum of $\mathbb{E}[R_s(e)]$ as:

$$\sum_{e=1}^{E} \mathbb{E}\left[R_s(e)\right] = \sum_{e=1}^{E} \mathbb{E}\left[R_s(e)|\mathcal{E}_{t_e}\right] \mathbb{P}(\mathcal{E}_{t_e}) + \sum_{e=1}^{E} \mathbb{E}\left[R_s(e)|\mathcal{E}_{t_e}^c\right] \mathbb{P}(\mathcal{E}_{t_e}^c) \tag{106}$$

$$\leq \sum_{e=1}^{E} \mathbb{E}\left[R_s(e)|\mathcal{E}_{t_e}\right] + \sum_{e=1}^{E} \mathbb{E}\left[R_s(e)|\mathcal{E}_{t_e}^c\right] \tag{107}$$

$$\leq (\sqrt{2}+1) ASD \sqrt{14AT \log(2AT)} + 2\sqrt{T} \tag{108}$$

Deviation from expected next state:

Lastly, $R_g(e)$ term denotes the deviation from landing in a state s_{t+1} from the expected value of the gains from all the

possible s_{t+1} from s_t, a_t. We bound the $R_g(e)$ as:

$$\mathbb{E}[R_g(e)] = \mathbb{E}\left[\sum_{t=t_e}^{t_{e+1}-1}\left((P_{\tilde{\pi}_e}\tilde{V}^k_{\tilde{\pi}_e})(s_t) - \tilde{V}^k_{\tilde{\pi}_e}(s_t)\right)\right] \tag{109}$$

$$= \mathbb{E}\left[V^k_{\tilde{\pi}_e}(s_{t_{e+1}}) - V^k_{\tilde{\pi}_e}(s_{t_e}) + \sum_{t=t_e}^{t_{e+1}-1}\left((P_{\tilde{\pi}_e}\tilde{V}^k_{\tilde{\pi}_e})(s_t) - \tilde{V}^k_{\tilde{\pi}_e}(s_{t+1})\right)\right] \tag{110}$$

$$\leq \mathbb{E}\left[V^k_{\tilde{\pi}_e}(s_{t_{e+1}}) - V^k_{\tilde{\pi}_e}(s_{t_e})\right] + \mathbb{E}\left[\sum_{t=t_e}^{t_{e+1}-1}\left((P_{\tilde{\pi}_e}\tilde{V}^k_{\tilde{\pi}_e})(s_t) - \tilde{V}^k_{\tilde{\pi}_e}(s_{t+1})\right)\right] \tag{111}$$

$$\leq \mathbb{E}\left[\tilde{D}\right] + \mathbb{E}\left[\sum_{t=t_e}^{t_{e+1}-1}\left((P_{\tilde{\pi}_e}\tilde{V}^k_{\tilde{\pi}_e})(s_t) - \tilde{V}^k_{\tilde{\pi}_e}(s_{t+1})\right)\right] \tag{112}$$

$$\leq D + \mathbb{E}\left[\sum_{t=t_e}^{t_{e+1}-1}\left((P_{\tilde{\pi}_e}\tilde{V}^k_{\tilde{\pi}_e})(s_t) - \tilde{V}^k_{\tilde{\pi}_e}(s_{t+1})\right)\right] \tag{113}$$

Note that the second term in Equation (112), $(P_{\tilde{\pi}_e}\tilde{V}^k_{\tilde{\pi}_e})(s_t) - \tilde{V}^k_{\tilde{\pi}_e}(s_{t+1})$, is a zero mean Martingale process adapted to filtration $\{\sigma(H_k)\}_{k=t_e}^{t_{e+1}}$. Also the difference $(P_{\tilde{\pi}_e}\tilde{V}^k_{\tilde{\pi}_e})(s_t) - \tilde{V}^k_{\tilde{\pi}_e}(s_{t+1})$ is bounded by $\tilde{D}$ for all t. Hence, using Azuma-Hoeffding's inequality, and $\mathbb{E}[\tilde{D}] = D$ gives us:

$$\mathbb{E}[R_g(e)] \leq D + D\sqrt{2(t_{e+1} - t_e)\log(t_{e+1} - t_e)} + \frac{1}{t_{e+1} - t_e} \tag{114}$$

Now, summing Equation (114) for all epochs we get.

$$\sum_{e=1}^{E}\mathbb{E}\left[R_g(e)\right] = \sum_{e=1}^{E}\left(D + D\sqrt{2(t_{e+1} - t_e)\log(t_{e+1} - t_e)} + \frac{1}{t_{e+1} - t_e}\right) \tag{115}$$

$$= ED + D\sqrt{2E\sum_{e=1}^{E}(t_{e+1} - t_e)\log(t_{e+1} - t_e)} + E\max_e \frac{1}{t_{e+1} - t_e} \tag{116}$$

$$= DSA\log_2\left(\frac{T}{SA}\right) + D\sqrt{2SA\log\left(\frac{T}{SA}\right)T\log T} + SA\log_2\left(\frac{T}{SA}\right) \tag{117}$$

Summing over all the possible sources of regret, we now have,

$$\begin{aligned}\mathbb{E}\left[R(T)\right] &= \sum_{e=1}^{E}\mathbb{E}\left[\Delta_e\right] \\ &= \sum_{e=1}^{E} R_1(e) + \mathbb{E}[R_s(e)] + \mathbb{E}[R_g(e)] \\ &= \tilde{O}\left(SAD\sqrt{AT}\right)\end{aligned} \tag{118}$$

We can now proceed to bound the constraint regret for constraint k. Let $\tilde{\lambda}^k_{\tilde{\pi}_e}$ be the average cost for constraint k, following policy $\tilde{\pi}_e$ in epoch e for the sampled CMDP $\tilde{\mathcal{M}}$. We have the following set of equations:

$$\mathbb{E}\left[R_k(T)\right] = \mathbb{E}\left[\sum_{t=0}^{T}\left(C_k - c_k(s_t, a_t)\right)\right] \tag{119}$$

$$= \mathbb{E}\left[\sum_{e=1}^{E}\sum_{t=t_e}^{t_{e+1}-1}\left(C_k - c_k(s_t, a_t)\right)\right] \tag{120}$$

$$= \mathbb{E}\left[\sum_{e=1}^{E}\sum_{t=t_e}^{t_{e+1}-1}\left(C_k - \tilde{\lambda}^k_{\tilde{\pi}_e} + \tilde{\lambda}^k_{\tilde{\pi}_e} - c_k(s_t, a_t)\right)\right] \tag{121}$$

$$\leq \mathbb{E}\left[\sum_{e=1}^{E}\sum_{t=t_e}^{t_{e+1}-1}\left(\tilde{\lambda}^k_{\tilde{\pi}_e} - c_k(s_t, a_t)\right)\right] \tag{122}$$

$$= \sum_{e=1}^{E}\mathbb{E}\left[\sum_{t=t_e}^{t_{e+1}-1}\left(\tilde{\lambda}^k_{\tilde{\pi}_e} - c_k(s_t, a_t)\right)\right] \tag{123}$$

$$\leq \tilde{O}\left(SAD\sqrt{AT}\right) \tag{124}$$

where Equation (122) follows from the fact that policy $\tilde{\pi}_e$ is feasible for the CMDP $\tilde{\mathcal{M}}_e$. Equation (124) follows from replacing the reward with costs in the analysis of $R(T)$.

Combining the two regrets gives the result as in the statement of the theorem. □