OPT2023: 15th Annual Workshop on Optimization for Machine Learning

Safe Posterior Sampling for Constrained MDPs with Bounded Constraint Violation

Krishna C. Kalagarla KALAGARL@USC.EDU
Rahul Jain RAHUL.JAIN@USC.EDU
Pierluigi Nuzzo NUZZO@USC.EDU
Department of Electrical and Computer Engineering, University of Southern California, Los Angeles

Abstract

The model in constrained Markov decision processes (CMDPs) is often unknown and must be learned online while still ensuring the constraint is met, or at least the violation is bounded with time. Some recent papers have made progress on this very challenging problem but either need unsatisfactory assumptions such as the knowledge of a safe policy, or have high cumulative regret. We propose the Safe PSRL algorithm that does not need such assumptions and yet performs very well, both in terms of theoretical regret bounds as well as empirically. The algorithm achieves an efficient tradeoff between exploration and exploitation by use of the posterior sampling principle, and provably suffers only bounded constraint violation by leveraging the idea of pessimism. Our algorithm is based on a primal-dual approach. We establish a sub-linear $\tilde{\mathcal{O}}\left(H^{2.5}\sqrt{|\mathcal{S}|^2|\mathcal{A}|K}\right)$ upper bound on the Bayesian reward objective regret along with a *bounded*, i.e., $\tilde{\mathcal{O}}\left(1\right)$ constraint violation regret over K episodes for an $|\mathcal{S}|$-state, $|\mathcal{A}|$-action, and horizon H CMDP.

1. Introduction

In this paper, we consider the problem of online learning for finite-horizon constrained MDPs (CMDPs) [4]. The transition probability is not known to the agent, thereby requiring the agent to learn about the system dynamics by observing the past states and actions. The performance of this agent is measured by the notion of *cumulative regret*, i.e., the difference between the cumulative reward of the learning agent and that of the optimal policy. This online learning problem thus leads to the well-known trade-off between *exploration* and *exploitation*.

A common approach to balance this exploration-exploitation trade-off is the *'Optimism in the Face of Uncertainty'* (OFU) principle [20] which has been widely used for online learning in MDPs [5, 16–18]. Another alternative for efficient exploration is *posterior sampling* [28]. The advantages of posterior sampling over OFU stem from the fact that (i) known information about the model can be incorporated into the algorithm through the prior distribution, and (ii) posterior sampling algorithms have demonstrated superior empirical performance for online learning over OFU-type algorithms including in the RL setting [24, 25]. Motivated by this superior empirical performance, we utilize posterior sampling for efficient exploration and introduce the `Safe PSRL` algorithm. Our algorithm further uses the primal-dual approach for CMDPs wherein the primal part performs unconstrained MDP planning with a sampled transition probability, and the dual part updates the Lagrangian variable to track the constraint violation.

We achieve bounded constraint violation regret by leveraging the idea of *pessimism*, introduced earlier in the context of constrained bandits [22]. “Pessimism" is achieved by tightening the con-

straint of the CMDP problem in every episode at decreasing levels. By appropriately balancing exploration via posterior sampling and *safe* learning via pessimism, we show that the `Safe PSRL` algorithm achieves sub-linear $\tilde{\mathcal{O}}\left(\frac{H^{2.5}}{\tau - c^0}\sqrt{|\mathcal{S}|^2|\mathcal{A}|K}\right)$ reward regret while achieving bounded, i.e., $\tilde{\mathcal{O}}(1)$-constraint violation regret. Though the regret bounds in this paper are Bayesian in nature, the algorithm shows superior empirical performance in the frequentist sense in comparison to comparable OFU-type algorithms with frequentist regret bounds.

2. Preliminaries

An episodic finite-horizon MDP [26] can be formally defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, s_1, p, r)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, H is the episode length, s_1 is the initial state, $p_h(s'|s,a)$ is the non-stationary transition probability and $r_h(s,a) \in [0,1]$ is the non-stationary reward function. A non-stationary randomized policy $\pi = (\pi_1, \ldots, \pi_H) \in \Pi$ where $\pi_i : \mathcal{S} \to \Delta_{\mathcal{A}}$, maps each state to a probability simplex over the action space $\mathcal{A}$. The value function of a non-stationary randomized policy π, $V_h^\pi(s; r, p)$ at a state $s \in \mathcal{S}$ and time step h is defined as $V_h^\pi(s; r, p) := \mathbb{E}_\pi\left[\sum_{i=h}^H r_i(s_i, a_i)|s_h = s, p\right]$. A finite-horizon constrained MDP (CMDP) [4] is a finite-horizon MDP with a required upper bound on the expectation of a cost function, $\{c, \tau \in (0, H]\}$. The non-stationary cost function is denoted by $c_h(s,a) \in [0,1]$. The total expected reward (cost) of an episode under policy π with respect to the reward (cost) function r (c) is the respective value function from the initial state s_1, i.e., $V_1^\pi(s_1; r, p)(V_1^\pi(s_1; c, p))$. Our objective for this CMDP is to find a policy which maximizes the total expected objective reward under the constraint that the total expected constraint cost is below a desired threshold. The optimal value is denoted by $V^*(s_1; r, p) = V_1^{\pi^*}(s_1; r, p)$ where,

$$\begin{aligned} \pi^* \in \underset{\pi \in \Pi}{\text{argmax}} \quad & V_1^\pi(s_1; r, p) \\ \text{s.t.} \quad & V_1^\pi(c, p) \leq \tau. \end{aligned} \tag{1}$$

3. The Learning Problem

We consider the setting where an agent repeatedly interacts with a finite-horizon CMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, s_1, p, r, \{c, \tau\})$ over multiple episodes and with stationary transition probability (i.e., $p_h = p, \forall h$). We employ the Bayesian framework and regard the transition probability p as random with a prior distribution μ_1. The realized transition probability is unknown to the learning agent. We consider settings where transition probability lies in the set Θ_{c_0} with the following property:

Assumption 1 *For all $\hat{p} \in \Theta_{c_0}$, there exists a policy $\pi_0^{\hat{p}}$ such that $V_1^{\pi_0^{\hat{p}}}(c, \hat{p}) \leq c_0 < \tau$.*

Moreover, we assume that the support of the prior distribution μ_1 is a subset of Θ_{c_0} and c_0 is known. The agent interacts with the environment for K episodes, each of length H. In each episode, the agent starts from a state s_1 and chooses a Markov policy π_k determined by the information gathered until that episode. This policy is then executed until the end of the episode, while collecting the rewards and costs. The main objectives of the learning agent are to minimize the Bayesian regrets:

(1) With respect to the reward defined as $\mathfrak{BR}(K; r) := \mathbb{E}\left[\sum_{k=1}^K \left(V_1^{\pi^*}(s_1; r, p) - V_1^{\pi^k}(s_1; r, p)\right)\right]$,

(2) With respect to the constraint defined as $\mathfrak{BR}(K; c) := \mathbb{E}\left[\sum_{k=1}^K \left(V_1^{\pi^k}(s_1; c, p) - \tau\right)\right]$.

4. The Safe PSRL Algorithm

We propose the Safe Posterior Sampling-based Reinforcement Learning (`Safe PSRL`) algorithm for the finite-horizon CMDP model. This algorithm leverages the idea of posterior sampling to balance exploration and exploitation. It also takes a primal-dual approach to handle the constraint cost objective along with reward maximization objective. We further introduce the idea of pessimism [22] to ensure that the cost regret is bounded. This "pessimism" is achieved by considering a "more constrained" CMDP problem as compared to the original problem. This is done by decreasing the constraint threshold by ϵ_k in each episode k. Formally, we consider the objective:

$$\begin{aligned} \max \quad & V_1^{\pi}(r,p) \\ \text{s.t.} \quad & V_1^{\pi}(c,p) \leq \tau - \epsilon_k. \end{aligned} \tag{2}$$

The algorithm starts with the prior distribution μ_1 on the transition probability. Then, at every time step t, the learning agent maintains a posterior distribution μ_t on the unknown transition probability p given by $\mu_t(\Theta) = \mathbb{P}(p \in \Theta | \mathcal{F}_t)$ for any set $\Theta \subseteq \Theta_{c_0}$. Here $\mathcal{F}_t$ is the information available at time t. In parallel, at the beginning of each episode k, transition probability $\hat{p}_k$ is sampled from the posterior distribution μ_{t_k} (where t_k is the time step corresponding to beginning of episode k). We then consider the Lagrangian defined as $L_k(\pi, \lambda) = V_1^{\pi}(r, \hat{p}_k) + \frac{\lambda_k}{\eta_k}\left(\tau - \epsilon_k - V_1^{\pi}(c, \hat{p}_k)\right)$.

The learning agent then chooses a Markov policy π_k (primal update) which maximizes the above Lagrangian. The (dual) parameter λ_k is updated according to the sub-gradient algorithm as: $\lambda_{k+1} = \left(\lambda_k + V_1^{\pi_k}(c, \hat{p}_k) + \epsilon_k - \tau\right)_+$. The agent then applies the policy π_k for the H steps of episode k. The `Safe PSRL` algorithm is summarized next.

Algorithm 1: `Safe-PSRL`

Input: K, μ_1, c_0, τ;
Initialization: $\lambda^1 \leftarrow 0$;
for *episodes* $k = 1, \ldots, K$ **do**
  $K_\epsilon \leftarrow 5$, $\epsilon_k \leftarrow \frac{K_\epsilon |H|^{1.5}\sqrt{|\mathcal{S}|^2|\mathcal{A}|}(\log k|\mathcal{S}||\mathcal{A}|H+1)}{\sqrt{k \log k|\mathcal{S}||\mathcal{A}|H}}$, $\eta_k \leftarrow (\tau - c_0)H\sqrt{k}$, $t_k = (k-1)H + 1$;
  Generate $\hat{p}_k \sim \mu_{t_k}(.)$;
  Compute $\pi_k \in \arg\max_\pi V_1^{\pi}(r - \frac{\lambda_k}{\eta_k}c, \hat{p}_k)$ (Policy Update) ;
  $\lambda_{k+1} \leftarrow \max(0, \lambda_k + V_1^{\pi_k}(c, \hat{p}_k) + \epsilon_k - \tau)$ (Dual Update);
  for $t = (k-1)H + 1, \ldots, kH$ **do**
    Choose action $a_t \sim \pi_k(s_t)$, Observe $s_{t+1} \sim p(.|s_t, a_t)$;
    Update the posterior distribution μ_{t+1} according to Bayes's rule ;
  end
end

Theorem 1 *Suppose Assumption 1 holds, then the regrets of the* `Safe PSRL` *algorithm are:*

$$\mathfrak{BR}(K; r) = \tilde{\mathcal{O}}\left(\frac{H^{2.5}}{\tau - c^0}\sqrt{|\mathcal{S}|^2|\mathcal{A}|K}\right) \quad \mathfrak{BR}(K; c) = \tilde{\mathcal{O}}\left(C''(H - \tau) + H^{1.5}\sqrt{|\mathcal{S}|^2|\mathcal{A}|C''}\right) = \mathcal{O}(1),$$

where $C'' = \mathcal{O}(\frac{H^3|\mathcal{S}|^2|\mathcal{A}|}{(\tau - c^0)^2})$ *is independent of* K.

5. Regret Analysis

A key property of posterior sampling [24] is the posterior sampling lemma.

Lemma 2 *For any function f, we have $\mathbb{E}[f(\hat{p}_t)] = \mathbb{E}[f(p)]$ where p is the transition probability and $\hat{p}_t$ is the sampled transition probability from the posterior distribution μ_t at time t.*

The following is a restatement [24] of the sub-linear regret bound achieved when using posterior sampling for unconstrained finite horizon MDPs.

Lemma 3 *The Bayesian regret of the PSRL algorithm for unconstrained MDPs is given by* $\sum_{k=1}^{K} \mathbb{E}\left[V_1^{\pi^k}(c,p) - V_1^{\pi^k}(c,\hat{p}_k)\right] \leq H^{1.5}\sqrt{30|\mathcal{S}|^2|\mathcal{A}|K\log(|\mathcal{S}||\mathcal{A}|KH)} + 2H.$

5.1. Cost Constraint Violation Analysis

We can decompose the constraint violation regret $\mathfrak{BR}(K;c)$ as follows:

$$\mathbb{E}\left[\sum_{k=1}^{K}\left(V_1^{\pi^k}(c,p) - \tau\right)\right] = \sum_{k=1}^{K}\mathbb{E}\left[V_1^{\pi^k}(c,p) - V_1^{\pi^k}(c,\hat{p}_k)\right] + \sum_{k=1}^{K}\mathbb{E}\left[V_1^{\pi^k}(c,\hat{p}_k) - \tau\right]$$

$$\leq \sum_{k=1}^{K}\mathbb{E}\left[V_1^{\pi^k}(c,p) - V_1^{\pi^k}(c,\hat{p}_k)\right] + \sum_{k=1}^{K}\mathbb{E}\left[\lambda_{k+1} - \lambda_k - \epsilon_k\right] \text{ (by dual update rule of algorithm)}$$

$$= \sum_{k=1}^{K}\mathbb{E}\left[V_1^{\pi^k}(c,p) - V_1^{\pi^k}(c,\hat{p}_k)\right] + \mathbb{E}\left[\lambda_{K+1}\right] - \sum_{k=1}^{K}\epsilon_k \tag{3}$$

$$\leq H^{1.5}\sqrt{30|\mathcal{S}|^2|\mathcal{A}|K\log(|\mathcal{S}||\mathcal{A}|KH)} + 2H + \mathbb{E}\left[\lambda_{K+1}\right] - \sum_{k=1}^{K}\epsilon_k \tag{4}$$

where the last upper bound follows by use of Lemma 3 to upper bound the first term in (3). We next upper bound the dual parameter $\mathbb{E}[\lambda_{K+1}]$ by the use of Lyapunov-drift analysis [22].

Lemma 4

$$\mathbb{E}\left[\lambda_{K+1}\right] \leq \frac{1}{\zeta}\log\frac{11H^2}{3\rho^2} + H + \sum_{1}^{C''}\epsilon_k + C''(H-\tau) + \frac{4(H^2 + \epsilon_{K+1}^2 + \eta_{K+1}H)}{(\tau - c^0)}. \tag{5}$$

where $C'' = \frac{80H^3|\mathcal{S}|^2|\mathcal{A}|}{(\tau-c_0)^2}$, $\rho = -(\tau - c_0)/4$ *and* $\zeta = \rho/(H^2 + H\rho/3)$.

Next, we bound the $\sum_k \epsilon_k$ term:

$$\sum_{k=1}^{K}\epsilon_k \geq \int_1^{K+1}\epsilon_u \mathrm{d}u \geq 10H^{1.5}\sqrt{|\mathcal{S}|^2|\mathcal{A}|Klog|\mathcal{S}||\mathcal{A}|HK} - 10H^{1.5}\sqrt{|\mathcal{S}|^2|\mathcal{A}|log|\mathcal{S}||\mathcal{A}|H}. \tag{6}$$

Thus, putting together (4), (5) and (6), the leading terms of $\tilde{\mathcal{O}}(\sqrt{K})$ cancel out and we get $\mathfrak{BR}(K;c) = \tilde{\mathcal{O}}\left(C''(H-\tau) + H^{1.5}\sqrt{|\mathcal{S}|^2|\mathcal{A}|C''}\right) = \tilde{\mathcal{O}}(1)$.

5.2. Reward Objective Regret Analysis

Let $\pi^{\epsilon_k,*}$ be the optimal policy for the pessimistic optimization problem (2) and $\pi^{\epsilon_k,\hat{p}_k}$ be the optimal policy for a similar pessimistic optimization problem, but where the transition probability is the sampled $\hat{p}_k$ instead of the true p. We then decompose the reward regret term $\mathfrak{BR}(K;r)$ as follows:

$$\sum_{k=1}^{K}\mathbb{E}\left[V_1^{\pi^*}(r,p)-V_1^{\pi^k}(r,p)\right]=\sum_{k=1}^{C''-1}\mathbb{E}\left[V_1^{\pi^*}(r,p)-V_1^{\pi^k}(r,p)\right]+\sum_{k=C''}^{K}\mathbb{E}\left[V_1^{\pi^*}(r,p)-V_1^{\pi^k}(r,p)\right]$$

$$\leq C''H+\sum_{k=C''}^{K}\mathbb{E}\left[V_1^{\pi^*}(r,p)-V_1^{\pi^{\epsilon_k,*}}(r,p)\right]+\sum_{k=C''}^{K}\mathbb{E}\left[V_1^{\pi^{\epsilon_k,*}}(r,p)-V_1^{\pi^{\epsilon_k,\hat{p}_k}}(r,\hat{p}_k)\right]$$

$$+\sum_{k=C''}^{K}\mathbb{E}\left[V_1^{\pi^{\epsilon_k,\hat{p}_k}}(r,\hat{p}_k)-V_1^{\pi_k}(r,\hat{p}_k)\right]+\sum_{k=C''}^{K}\mathbb{E}\left[V_1^{\pi_k}(r,\hat{p}_k)-V_1^{\pi_k}(r,p)\right]\text{ (splitting into four parts)}$$

$$\leq C''H+\sum_{k=C''}^{K}\mathbb{E}\left[V_1^{\pi^*}(r,p)-V_1^{\pi^{\epsilon_k,*}}(r,p)\right]+0\quad\text{(by the posterior sampling property in Lemma 2)}$$

$$+\sum_{k=C''}^{K}\mathbb{E}\left[V_1^{\pi^{\epsilon_k,\hat{p}_k}}(r,\hat{p}_k)-V_1^{\pi_k}(r,\hat{p}_k)\right]+\sum_{k=C''}^{K}\mathbb{E}\left[V_1^{\pi_k}(r,\hat{p}_k)-V_1^{\pi_k}(r,p)\right]$$

$$\leq C''H+\sum_{k=C''}^{K}\mathbb{E}\left[V_1^{\pi^*}(r,p)-V_1^{\pi^{\epsilon_k,*}}(r,p)\right]+\sum_{k=C''}^{K}\mathbb{E}\left[V_1^{\pi^{\epsilon_k,\hat{p}_k}}(r,\hat{p}_k)-V_1^{\pi_k}(r,\hat{p}_k)\right]$$

$$+H^{1.5}\sqrt{30|\mathcal{S}|^2|\mathcal{A}|K\log(|\mathcal{S}||\mathcal{A}|KH)}+2H(\text{ by the regret bound in Lemma 3})\tag{7}$$

Lemma 5 *[21] The first summation term above can be bounded as* $\sum_{k=C''}^{K}\mathbb{E}\left[V_1^{\pi^*}(r,p)-V_1^{\pi^{\epsilon_k,*}}(r,p)\right]\leq\sum_{k=C''}^{K}\frac{\epsilon_k H}{\tau-c^0}=\tilde{\mathcal{O}}\left(\frac{H^{2.5}}{\tau-c^0}\sqrt{|\mathcal{S}|^2|\mathcal{A}|K}\right)$.

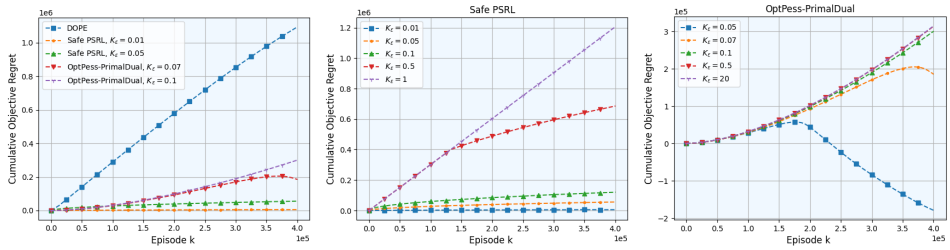
By optimality of π_k and the update rule of the dual parameter λ_k, we can prove the following lemma:

Lemma 6 $\sum_{k=C''}^{K}\mathbb{E}\left[V_1^{\pi^{\epsilon_k,\hat{p}_k}}(r,\hat{p}_k)-V_1^{\pi_k}(r,\hat{p}_k)\right]=\tilde{\mathcal{O}}\left(\frac{H}{\tau-c^0}\sqrt{K}\right)$

Now, putting together (7), Lemma 5 and Lemma 6, we get that $\mathfrak{BR}(K;r)=\tilde{\mathcal{O}}\left(\frac{H^{2.5}}{\tau-c^0}\sqrt{|\mathcal{S}|^2|\mathcal{A}|K}\right)$.

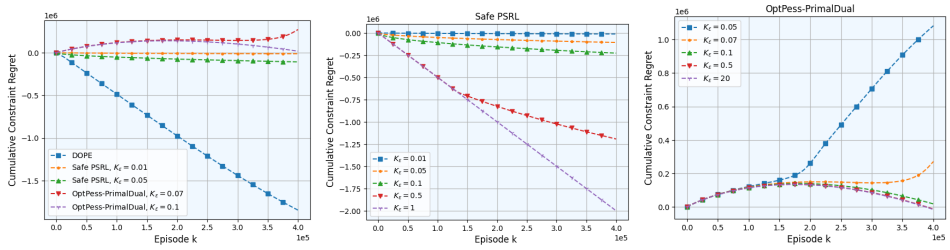
6. Experimental Results

We now evaluate the empirical performance of the `Safe PSRL` algorithm and compare it with the state-of-the-art `DOPE` algorithm [8] and the `OptPess-PrimalDual` algorithm [21]. We consider the setting of a media streaming service [8] and evaluate the cumulative regret for the `Safe PSRL`, `DOPE`, and `OptPess-PrimalDual` algorithms. The transition probability is fixed and not sampled from a prior distribution (i.e., the evaluation is not Bayesian in nature, but frequentist). We further scale the ϵ_k parameters of the `Safe PSRL` and `OptPess-PrimalDual` algorithm, by varying the coefficient of the ϵ_k parameters denoted by K_ϵ, to control the pessimism. The performance of our algorithm is also compared against the `DOPE` algorithm, which requires a known safe policy. We choose the optimal policy of the given CMDP with a tighter constraint threshold $c_0 = 1$ as



(*a*) (*b*) (*c*)

Figure 1: Cumulative objective regret.



(*a*) (*b*) (*c*)

Figure 2: Cumulative constraint regret.

the safe policy. Figure 1(a) shows that the `Safe PSRL` algorithm significantly outperforms the `DOPE` and `OptPess-PrimalDual` algorithms in terms of objective regret. At the same time, it ensures that the constraint regret is negative for almost all of the episodes, as shown by Fig. 2(a). The results show that the constraint is satisfied in almost all of the episodes, which is stronger than the theoretical guarantee for `Safe PSRL`. Further, though the `OptPess-PrimalDual` algorithm appears to perform better than the `DOPE` algorithm in Fig. 1(a) in terms of objective regret, it has very high constraint regret, as shown in Fig. 2(a). On the other hand, `DOPE` satisfies the constraint in every episode and exhibits very low constraint regret. We further evaluate `Safe PSRL` for various values of K_ϵ and note that, in all instances, the constraint regret is negative for almost all of the episodes, as shown by Fig. 2(b). Moreover, the objective regret in Fig. 1(b) increases as the levels of *pessimism* expressed by K_ϵ increase. Therefore, for suitable levels of pessimism, `Safe PSRL` algorithm ensures low objective regret while satisfying the constraint objective. Differently, Fig. 2(c) shows that the `OptPess-PrimalDual` algorithm is unable to achieve low regret even at high levels of pessimism. Considering Fig. 1(c) and Fig. 2(c) together, we see that the algorithm achieves low objective regret at the expense of exploding constraint regret. Overall, `Safe PSRL` is able to achieve superior objective regret performance while satisfying the constraint for almost all the episodes. This result is further achieved without the knowledge of a safe policy.

Acknowledgments

This research was supported in part by the National Science Foundation under Awards 1846524, 2139982 and 2025732, the Office of Naval Research under Award N00014-20-1-2258, the Okawa Research Grant, and the USC Center for Autonomy and Artificial Intelligence.

References

[1] Mridul Agarwal, Qinbo Bai, and Vaneet Aggarwal. Regret guarantees for model-based reinforcement learning with long-term average constraints. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

[2] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

[3] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135. PMLR, 2013.

[4] Eitan Altman. *Constrained Markov Decision Processes*, volume 7. CRC Press, 1999.

[5] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org, 2017.

[6] Qinbo Bai, Amrit Singh Bedi, Mridul Agarwal, Alec Koppel, and Vaneet Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3682–3689, 2022.

[7] Kianté Brantley, Miro Dudik, Thodoris Lykouris, Sobhan Miryoosefi, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings. *Advances in Neural Information Processing Systems*, 33:16315–16326, 2020.

[8] Archana Bura, Aria Hasanzadezonuzy, Dileep Kalathil, Srinivas Shakkottai, and Jean-Francois Chamberland. Dope: Doubly optimistic and pessimistic exploration for safe reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

[9] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24:2249–2257, 2011.

[10] Liyu Chen, Rahul Jain, and Haipeng Luo. Learning infinite-horizon average-reward markov decision processes with constraints. *arXiv preprint arXiv:2202.00150*, 2022.

[11] 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*, 33:8378–8390, 2020.

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

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

[14] Aria HasanzadeZonuzy, Archana Bura, Dileep Kalathil, and Srinivas Shakkottai. Learning with safety constraints: Sample complexity of reinforcement learning for constrained mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7667–7674, 2021.

[15] Mehdi Jafarnia-Jahromi, Rahul Jain, and Ashutosh Nayyar. Online learning for unknown partially observable mdps. *arXiv preprint arXiv:2102.12661*, 2021.

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

[17] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

[18] Krishna C Kalagarla, Rahul Jain, and Pierluigi Nuzzo. A Sample-Efficient Algorithm for Episodic Finite-Horizon MDP with Constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8030–8037, 2021.

[19] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer, 2012.

[20] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[21] Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained mdps. *Advances in Neural Information Processing Systems*, 34:17183–17193, 2021.

[22] Xin Liu, Bin Li, Pengyi Shi, and Lei Ying. An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. *Advances in Neural Information Processing Systems*, 34:24075–24086, 2021.

[23] Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International Conference on Machine Learning*, pages 2701–2710. PMLR, 2017.

[24] Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

[25] Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision processes: A thompson sampling approach. In *Advances in Neural Information Processing Systems*, pages 1333–1342, 2017.

[26] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994. ISBN 0471619779.

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

[28] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[29] Honghao Wei, Xin Liu, and Lei Ying. Triple-q: A model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation. In *International Conference on Artificial Intelligence and Statistics*, pages 3274–3307. PMLR, 2022.

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

Appendix A. Proofs

A.1. Proof of Lemma 4

We restate the following lemma [22] which states the Lyapunov-drift conditions for the boundedness of a random process.

Lemma 7 *[22] Consider a random process $S(t)$ with a Lyapunov function $\Phi(k)$ such that $\Phi(0) = 0$ and $\Delta(k) = \Phi(k+1) - \Phi(k)$ is the Lyapunov drift. Given an increasing sequence $\{\varphi_k\}$ and constants ρ and ν_{max} with $0 < \rho \leq \nu_{max}$, if the expected drift $\mathbb{E}[\Delta(k)|S(k) = s]$ satisfies the following conditions:*
(i) There exists constants $\rho > 0$ and $\varphi_k > 0$ s.t. $\mathbb{E}[\Delta(k)|S(k) = s] \leq -\rho$ when $\Phi(k) \geq \varphi_k$, and
(ii) $|\Phi(k+1) - \Phi(k)| \leq \nu_{\max}$ holds with probability 1, then

$$\mathbb{E}\left[e^{\zeta\Phi(t)}\right] \leq \mathbb{E}\left[e^{\zeta\Phi_0}\right] + \frac{2e^{\zeta(\nu_{\max}+\varphi_t)}}{\zeta\rho}, \quad \textit{where } \zeta = \rho/(\nu_{\max}^2 + \nu_{\max}\rho/3).$$

We divide the episodes into two parts, i.e. $k < C''$ and $k \geq C''$ where $C'' = \frac{80H^3|\mathcal{S}|^2|\mathcal{A}|}{(\tau-c_0)^2}$. We can clearly see that for $k \geq C''$, we have $\epsilon_k \leq \frac{\tau-c_0}{2}$. Thus, for $k \geq C''$, Problem (2) is feasible for all $\hat{p_k} \in \Theta_{c_0}$ by Assumption 1. For $k \geq C''$, we show that the Lyapunov function $\Phi(\lambda) = \lambda$ satisfies the conditions of Lemma 7 and thus provide a bound on the exponential moment of the dual variable λ.

Lemma 8 *For $k \geq C''$, when $\lambda \geq \varphi_k$, we have, $\mathbb{E}[\lambda_{k+1} - \lambda_k|\lambda_k = \lambda] \leq \rho$ and $|\lambda_{k+1} - \lambda_k| \leq H$ with probability 1, where $\varphi_k := 4(H^2 + \epsilon_k^2 + \eta_k H)/(\tau - c^0)$ and $\rho := -(\tau - c_0)/4$. Thus, we have,*

$$\mathbb{E}\left[e^{\zeta\lambda_{K+1}}\right] \leq \mathbb{E}\left[e^{\zeta\lambda_{C''}}\right] + \frac{2e^{\zeta(H+\varphi_{K+1})}}{\zeta\rho}, \quad \textit{where } \zeta = \rho/(H^2 + H\rho/3).$$

Proof Now for $k \geq C''$, consider:

$$\begin{aligned}
\frac{{\lambda_{k+1}}^2}{2} - \frac{{\lambda_k}^2}{2} &= \lambda_k(\lambda_{k+1} - \lambda_k) + \frac{1}{2}(\lambda_{k+1} - \lambda_k)^2 \\
&= \lambda_k(V_1^{\pi_k}(c, \hat{p}_k) + \epsilon_k - \tau) + \frac{1}{2}(V_1^{\pi_k}(c, \hat{p}_k) + \epsilon_k - \tau)^2 \\
&= \lambda_k(V_1^{\pi_k}(c, \hat{p}_k) + \epsilon_k - \tau) - \eta_k V_1^{\pi_k}(r, \hat{p}_k) + \eta_k V_1^{\pi_k}(r, \hat{p}_k) + \frac{1}{2}(V_1^{\pi_k}(c, \hat{p}_k) + \epsilon_k - \tau)^2 \\
&\leq \lambda_k(V_1^{\pi_k}(c, \hat{p}_k) + \epsilon_k - \tau) - \eta_k V_1^{\pi_k}(r, \hat{p}_k) + \eta_k H + \frac{1}{2}(V_1^{\pi_k}(c, \hat{p}_k) + \epsilon_k - \tau)^2 \\
&\leq \lambda_k(V_1^{\pi_k}(c, \hat{p}_k) + \epsilon_k - \tau) - \eta_k V_1^{\pi_k}(r, \hat{p}_k) + \eta_k H + (V_1^{\pi_k}(c, \hat{p}_k) - \tau)^2 + \epsilon_k^2 \\
&\text{(Using } \frac{(a+b)^2}{2} \leq a^2 + b^2\text{)} \\
&\leq \lambda_k(V_1^{\pi_k}(c, \hat{p}_k) + \epsilon_k - \tau) - \eta_k V_1^{\pi_k}(r, \hat{p}_k) + \eta_k H + H^2 + \epsilon_k^2 \\
&\leq \lambda_k(V_1^{\pi_0^{\hat{p}_k}}(c, \hat{p}_k) + \epsilon_k - \tau) - \eta_k V_1^{\pi_0^{\hat{p}_k}}(r, \hat{p}_k) + \eta_k H + H^2 + \epsilon_k^2 \\
&\text{(By optimality of } \pi_k \text{ in primal update)} \\
&\leq \lambda_k(c_0 + \epsilon_k - \tau) + \eta_k H + H^2 + \epsilon_k^2
\end{aligned}$$

$$\leq -\frac{\lambda_k(\tau - c_0)}{2} + \eta_k H + H^2 + \epsilon_k^2$$
$$(\text{ as for } k \geq C'', \epsilon_k \leq \frac{(\tau - c_0)}{2})$$

Now for $\lambda \geq \varphi_k$ where $\varphi_k := 4(H^2 + \epsilon_k^2 + \eta_k H)/(\tau - c^0)$, we have:

$$\begin{aligned}
\mathbb{E}\left[\lambda_{k+1} - \lambda_k | \lambda_k = \lambda\right] &\leq \mathbb{E}\left[\frac{\lambda_{k+1}^2 - \lambda_k^2}{2\lambda_k} | \lambda_k = \lambda\right] (\text{Using } x - y \leq \frac{x^2 - y^2}{2y}, \text{for } y > 0) \\
&= \frac{1}{\lambda}\mathbb{E}\left[\frac{\lambda_{k+1}^2 - \lambda_k^2}{2} | \lambda_k = \lambda\right] \\
&\leq \frac{1}{\lambda}\mathbb{E}\left[-\frac{\lambda_k(\tau - c_0)}{2} + \eta_k H + H^2 + \epsilon_k^2 | \lambda_k = \lambda\right] \\
&= -\frac{(\tau - c_0)}{2} + \frac{\eta_k H + H^2 + \epsilon_k^2}{\lambda} \\
&\leq -\frac{(\tau - c_0)}{2} + \frac{(\tau - c_0)}{4} \\
&= -\frac{(\tau - c_0)}{4} := \rho
\end{aligned}$$

Further, $|\lambda_{k+1} - \lambda_k| = |V_1^{\pi_k}(c, \hat{p}_k) + \epsilon_k - \tau| \leq H$ with probability 1. Thus, by lemma 7, we have :

$$\mathbb{E}\left[e^{\zeta\lambda_{K+1}}\right] \leq \mathbb{E}\left[e^{\zeta\lambda_{C''}}\right] + \frac{2e^{\zeta(H+\varphi_{K+1})}}{\zeta\rho}, \tag{8}$$

where $\zeta = \rho/(H^2 + H\rho/3)$. ■

The above inequality (8) can be simplified as follows:

$$\begin{aligned}
&\implies e^{\zeta\mathbb{E}[\lambda_{K+1}]} \leq \mathbb{E}\left[e^{\zeta\lambda_{C''}}\right] + \frac{2e^{\zeta(H+\varphi_{K+1})}}{\zeta\rho} \text{ (By Jensen's inequality)} \\
&\implies \mathbb{E}\left[\lambda_{K+1}\right] \leq \frac{1}{\zeta}\log\left[\mathbb{E}\left[e^{\zeta\lambda_{C''}}\right] + \frac{2e^{\zeta(H+\varphi_{K+1})}}{\zeta\rho}\right]
\end{aligned}$$

Further,

$$\begin{aligned}
\lambda_{C''} &\leq \lambda_1 + \sum_1^{C''-1} (V_1^{\pi^k}(c, \hat{p}_k) + \epsilon_k - \tau)_+ \\
&\leq \sum_1^{C''} \epsilon_k + C''(H - \tau) := \lambda_{C''}^{\max}
\end{aligned}$$

Continuing,

$$\mathbb{E}\left[\lambda_{K+1}\right] \leq \frac{1}{\zeta}\log\left[e^{\zeta\lambda_{C''}^{\max}} + \frac{2e^{\zeta(H+\varphi_{K+1})}}{\zeta\rho}\right]$$

$$\leq \frac{1}{\zeta}\log\left[e^{\zeta\lambda_{C''}^{\max}} + \frac{8H^2 e^{\zeta(H+\varphi_{K+1})}}{3\rho^2}\right] \text{ (Using } \zeta \geq \frac{3(\tau-c_0)}{13H^2})$$
$$\leq \frac{1}{\zeta}\log\left[\frac{11H^2}{3\rho^2}e^{\zeta\left(H+\varphi_{K+1}+\lambda_{C''}^{\max}\right)}\right]$$
$$= \frac{1}{\zeta}\log\frac{11H^2}{3\rho^2} + H + \varphi_{K+1} + \lambda_{C''}^{\max}$$
$$= \frac{1}{\zeta}\log\frac{11H^2}{3\rho^2} + H + \sum_{1}^{C''}\epsilon_k + C''(H-\tau) + \frac{4(H^2+\epsilon_{K+1}^2+\eta_{K+1}H)}{(\tau-c^0)}$$

A.2. Proof of Lemma 6

Proof

$$\sum_{k=C''}^{K}\mathbb{E}\left[V_1^{\pi^{\epsilon_k},\hat{p}_k}(r,\hat{p}_k) - V_1^{\pi_k}(r,\hat{p}_k)\right] = \sum_{k=C''}^{K}\mathbb{E}\left[\frac{\lambda_k}{\eta_k}\left(V_1^{\pi^{\epsilon_k},\hat{p}_k}(c,\hat{p}_k) - V_1^{\pi_k}(c,\hat{p}_k)\right)\right]$$
$$+ \sum_{k=C''}^{K}\mathbb{E}\left[\left(V_1^{\pi^{\epsilon_k},\hat{p}_k}(r,\hat{p}_k) - \frac{\lambda_k}{\eta_k}V_1^{\pi^{\epsilon_k},\hat{p}_k}(c,\hat{p}_k)\right)\right] - \sum_{k=C''}^{K}\mathbb{E}\left[\left(V_1^{\pi^{k}}(r,\hat{p}_k) - \frac{\lambda_k}{\eta_k}V_1^{\pi^{k}}(c,\hat{p}_k)\right)\right]$$
$$\leq \sum_{k=C''}^{K}\mathbb{E}\left[\frac{\lambda_k}{\eta_k}\left(V_1^{\pi^{\epsilon_k},\hat{p}_k}(c,\hat{p}_k) - V_1^{\pi_k}(c,\hat{p}_k)\right)\right] + 0 \quad \text{(By optimality of } \pi_k \text{ in primal update)}$$
$$\leq \sum_{k=C''}^{K}\mathbb{E}\left[\frac{\lambda_k}{\eta_k}\left(\tau - \epsilon_k - V_1^{\pi_k}(c,\hat{p}_k)\right)\right]$$
$$\leq \sum_{k=C''}^{K}\mathbb{E}\left[\frac{1}{\eta_k}((\lambda_k(\lambda_{k+1}-\lambda_k)+\tau^2)\right] \text{(By update rule for } \lambda_k)$$
$$\leq \mathbb{E}\left[\sum_{k=C''}^{K}\frac{1}{\eta_k}(\frac{\lambda_k^2}{2} - \frac{\lambda_{k+1}^2}{2}) + \sum_{k=C''}^{K}\frac{1}{2\eta_k}(\lambda_{k+1}-\lambda_k)^2 + \sum_{k=C''}^{K}\frac{\tau^2}{\eta_k}\right]$$
$$\leq \mathbb{E}\left[\frac{(\lambda_{C''})^2}{2\eta_{C''}}\right] + \sum_{k=C''}^{K}\frac{H^2}{2\eta_k} + \sum_{k=C''}^{K}\frac{H^2}{\eta_k} \text{(As } \eta_k \text{ increases with } k)$$
$$\leq \frac{(\sum_{k=1}^{C''}\epsilon_k + C''(H-\tau))^2}{2\eta_{C''}} + \frac{3H}{2}\sum_{C''}^{K}\frac{1}{(\tau-c_0)\sqrt{k}}$$
$$= \tilde{\mathcal{O}}\left(\frac{H}{\tau-c^0}\sqrt{K}\right)$$

■

Appendix B. Experiment Setup

We consider the setting of a media streaming service [8] from a wireless base station. The base station provides the streaming service at two different speeds. These speeds follow independent

Bernoulli distributions denoted by parameters $\mu_1 = 0.9$ and $\mu_2 = 0.1$, with μ_1 corresponding to the faster service. The data packets arriving at the device are stored in a buffer and sent out according to a Bernoulli random process with mean γ. The buffer size s_h evolves as $s_{h+1} = \min\left(\max\left(0, s_h + A_h - B_h\right), N\right)$ where A_h is the number of packet arrivals, B_h is the number of packet departures, and $N = 10$ is the maximum size of the buffer. The device desires to minimize the cost of running out of packets, i.e., an empty buffer, while restricting the use of the faster service. We model this scenario as a finite horizon CMDP with the state representing the buffer size and actions $\{1, 2\}$ denoting the choice of speed. We set the objective cost as $r(s, a) = \mathbb{1}\{s = 0\}$ and the constraint cost as $c(s, a) = \mathbb{1}\{a = 1\}$. The episode length H is 10 and the threshold τ is 5.

The algorithms are evaluated over $K = 400,000$ episodes. They are carried out 10 times and averaged to obtain the regret plots. All the experiments are performed on a 2019 MacBook Pro with 1.4 GHz Quad-Core Intel Core i5 processor and 16GB RAM.

Appendix C. Related Work

Posterior (or Thompson) sampling goes back to the work of [28], but attracted less attention for several decades until empirical evidence [9] showed its superior performance for online learning. Recently, it has been widely applied to various settings like multi-armed bandits [2, 3, 19], MDPs [23–25] and POMDPs [15].

OFU-based algorithms have been widely used for efficient learning in CMDPs, e.g., in the setting of PAC performance guarantees for finite-horizon CMDPs [14, 18], or to provide regret bounds for CMDPs in the finite-horizon setting [7, 13] and infinite-horizon average cost setting [27]. Policy gradient algorithms for CMDPs [11, 12] have also been studied. However, these algorithms do not provide bounded or zero constraint violation guarantees. Recently, some OFU-based approaches for *safe* learning with bounded or zero constraint violation guarantees have been proposed [6, 10, 21, 29, 30]. But these either assume the transition model is known, or assume that a safe policy is known to the algorithm (and can be used by it), e.g., in [8, 21]. The OptPess-PrimalDual algorithm in [21] is the closest comparable algorithm to our `Safe PSRL` algorithm.

While the use of the posterior sampling principle for constrained RL problems is under-explored (despite the promise of better empirical performance), [1] indeed introduces a PSRL algorithm for CMDPs but for the average setting. Moreover, it only achieves a $\tilde{\mathcal{O}}(\sqrt{K})$ constraint violation regret which is worse than our $\tilde{\mathcal{O}}(1)$ bound.