Learning in Constrained Markov Decision Processes

Rahul Singh, *Member, IEEE*, Abhishek Gupta, and Ness B. Shroff, *Fellow, IEEE*

***Abstract*—We consider reinforcement learning (RL) in Markov decision processes in which an agent repeatedly interacts with an environment that is modeled by a controlled Markov process. At each time step t, it earns a reward and also incurs a cost vector consisting of M costs. We design model-based RL algorithms that maximize the cumulative reward earned over a time horizon of T time steps while simultaneously ensuring that the average values of the M cost expenditures are bounded by agent-specified thresholds $c_i^{\text{ub}}, i = 1, 2, \ldots, M$. The consideration of the cumulative cost expenditures departs from the existing literature, in that the agent now additionally needs to balance the cost expenses in an online manner while simultaneously performing the exploration–exploitation tradeoff that is typically encountered in RL tasks. This is challenging since the dual objectives of exploration and exploitation necessarily require the agent to expend resources. In order to measure the performance of an RL algorithm that satisfies the average cost constraints, we define an $M + 1$ dimensional regret vector that is composed of its reward regret, and M cost regrets. The reward regret measures the suboptimality in the cumulative reward while the ith component of the cost regret vector is the difference between its ith cumulative cost expense and the expected cost expenditures Tc_i^{ub}. We prove that the expected value of the regret vector is upper-bounded as $\tilde{O}(T^{2/3})$, where T is the time horizon, and $\tilde{O}(\cdot)$ hides factors that are logarithmic in T. We further show how to reduce the regret of a desired subset of the M costs, at the expense of increasing the regrets of rewards and the remaining costs. To the best of our knowledge, ours is the only work that considers nonepisodic RL under average cost constraints and derives algorithms that can *tune the regret vector* according to the agent's requirements on its cost regrets.**

***Index Terms*—Machine learning, Markov decision processes, reinforcement learning.**

Manuscript received 27 February 2022; revised 28 February 2022 and 29 May 2022; accepted 4 July 2022. Date of publication 31 August 2022; date of current version 13 March 2023. This work was supported in part by the NSF under Grant CNS-2112471, Grant CNS-2106933, Grant CNS-2106932, Grant CNS-1955535, Grant CNS-1901057, Grant CNS- 2007231, and Grant CNS-1618520, and in part by an Army Research Office under Grant W911NF2110244. The work of Rahul Singh was supported by the SERB under Grant SRG/2021/002308 and Grant PC 39010B. Recommended by Associate Editor Johanna L. Mathieu. *(Corresponding author: Rahul Singh.)*

Rahul Singh is with the Department of ECE, Indian Institute of Science, Bengaluru 560012, India (e-mail: rahulsingh@iisc.ac.in).

Abhishek Gupta and Ness B. Shroff are with the Department of ECE, Ohio State University, Columbus, OH 43210 USA (e-mail: gupta.706@osu.edu; shroff@ece.osu.edu).

Digital Object Identifier 10.1109/TCNS.2022.3203361

I. Introduction

REINFORCEMENT learning (RL) [1] involves an agent repeatedly interacting with an environment modeled by a Markov decision process (MDP) [2]. More specifically, consider a controlled Markov process [2] s_t, $t = 1, 2, \ldots, T$. At each discrete time t, an agent applies control a_t. State-space and action space are denoted by $\mathcal{S}$ and $\mathcal{A}$, respectively, and are assumed to be finite. The controlled transition probabilities are denoted by $p := \{p(s, a, s') : s, s' \in \mathcal{S}, a \in \mathcal{A}\}$. Thus, $p(s, a, s')$ is the probability that the system state transitions to state s' upon applying action a in state s. The probabilities $p(s, a, s')$ are not known to the agent. At each discrete time $t = 1, 2, \ldots, T$, the agent observes the current state of the environment s_t, applies control action a_t, and earns a reward r_t that is a known function of (s_t, a_t). When the agent applies an action a in the state s, then it earns a reward equal to $r(s, a)$ units. The agent does not know the controlled transition probabilities $p(s, a, s')$ that describe the system dynamics of the environment. The performance of an agent or a RL algorithm is measured by the cumulative rewards that it earns over the time horizon.

However, in many applications, in addition to earning rewards, the agent also incurs costs at each time. The underlying physical constraints impose constraints on its cumulative cost expenditures, so that the agent needs to balance its reward earnings with the cost accretion while also simultaneously learning the choice of optimal decisions, all in an *online manner*. As a motivating example, consider a single-hop wireless network that consists of a wireless node that transmits data packets to a receiver over an unreliable wireless channel. The channel reliability, i.e., the probability that a transmission at time step t is successful, depends upon the instantaneous channel state cs_t and the transmission power a_t. Thus, for example, this probability is higher when the channel is in a good state, or if the transmission is carried out at higher power levels. The transmitter stores packets in a buffer, and its queue length at time t is denoted by Q_t. The wireless node is battery operated, and packet transmission consumes power. Hence, it is desired that the average power consumption is minimal. An appropriate performance metric for networks [3] is the average queue length $\left(\mathbb{E}\sum_{t=1}^{T} Q_t\right)/T$, and hence, it is required that the average queue length stays below a certain threshold. The AP has to choose a_t adaptively so as to minimize the power consumption $\left(\mathbb{E}\sum_{t=1}^{T} a_t\right)/T$ or, equivalently, maximize $\left(\mathbb{E}\sum_{t=1}^{T} -a_t\right)/T$ while simultaneously ensuring that the average queue length is

below a user-specified threshold, i.e., $\left(\mathbb{E}\sum_{t=1}^{T} Q_t\right)/T \leq c^{\text{ub}}$. In this example, the state of the "environment" at time t is given by the queue length and the channel state (Q_t, cs_t). Thus, it might be "optimal" to utilize high transmission power levels only when the instantaneous queue length Q_t is large or the wireless channel's state cs_t is good. Such an adaptive strategy saves energy by transmitting at lower energy levels at other times. Since channel reliabilities are typically not known to the transmitter node, it does not know the transition probabilities $p(s, a, s')$ that describe the controlled Markov process (Q_t, cs_t). Hence, it cannot compute the expectations of the average queue lengths and average power consumption for a fixed control policy, and needs to devise appropriate learning policies to optimize its performance under average-cost constraints. RL algorithms that we propose in this work solve exactly these classes of problems.

Many important network control problems can be solved within the framework of constrained MDPs (CMDPs). For example, Lazar [4] and Hsiao and Lazar [5] utilize CMDPs in order to maximize the throughput offered by a stochastic network, where the network operator wants to simultaneously satisfy constraints on delays, while Nain and Ross [6] design control policies that make dynamic decisions regarding network access in networks shared by different types of traffic. Similarly, the framework of CMDPs has been used in [7] and [8] in order to maximize the timely throughput[1] in stochastic networks. The work [9] addresses the issue of admission control and routing in networks shared by multiple flows in which the goal is to maximize the weighted sum of customers served while simultaneously satisfying constraints on the blocking probability. If the network/system parameters are known, then a CMDP can be posed as a linear program (LP) and solved efficiently. However, in practice, network parameters are seldom known to the network operator, and it needs to design algorithms that "learn" the optimal policies in an "optimal" manner. Our work addresses precisely this issue.

II. Previous Works and Our Contributions

RL Algorithms for unconstrained MDPs: RL problems without constraints are well understood by now. Jaksch et al. [10] develop UCRL2 algorithm using the upper confidence bounds (UCB) strategy [11], while Mete et al. [12] use the reward biased maximum likelihood estimation (RBMLE) approach [13], and Osband and Roy [14] use Thompson sampling. UCRL2 [10] is a popular RL algorithm that has a regret bound of $\tilde{O}(D(p)S\sqrt{AT})$, where $D(p)$ is the diameter [10] of the MDP p; the algorithms proposed in this work are based on UCRL2.

RL Algorithms for Constrained MDPs: The work in [15] is an early work on optimally controlling unknown MDPs under average cost constraints. It utilizes the certainty equivalence (CE) principle, i.e., it applies controls that are optimal under the assumption that the true (but unknown) MDP parameters are equal to the empirical estimates and also occasionally resorts to "forced explorations." This algorithm yields asymptotically (as $T \to \infty$) the same reward rate as the case when the MDP parameters are known. However, analysis is performed under the assumption that the CMDP is *strictly feasible*. Moreover, the algorithm lacks finite-time performance guarantees (bounds on regret). Unlike [15], we do not assume strict feasibility; in fact, we show that the use of *confidence bounds* allows us to get rid of the strict feasibility assumption. Borkar [16] derives a learning scheme based on multitime-scale stochastic approximation [17], in which the task of learning an optimal policy for the CMDP is decomposed into that of learning the optimal value of the dual variables, which correspond to the price of violating the average cost constraints, and that of learning the optimal policy for an unconstrained MDP parameterized by the dual variables. However, the proposed scheme lacks finite-time regret analysis and might suffer from a large regret. Prima facie, this layered decomposition might not be optimal with respect to the sample-complexity of the online RL problem. Recent works [18], [19] have obtained concentration bounds for two time-scale stochastic approximation algorithms, which could be used for deriving regret bounds. The works [20], [21], [22], [23] design policy-search algorithms for constrained RL problems. However unlike our work, they do not utilize the concept of regret vector, and their theoretical guarantees need further research. After the first draft of our work was published online, there appeared a few manuscripts/works that address various facets of learning in CMDPs, and these have some similarities with our work. For example, Qiu et al. [24] consider episodic RL problems with constraints in which the reward function is time-varying. Similarly, Efroni et al. [25] also consider episodic RL in which the state is reset at the beginning of each episode. In contrast, we deal exclusively with non-episodic infinite horizon RL problems. In fact, as we show in our work, the primary difficulty in nonepisodic constrained RL arises due to the fact that it is not possible to simultaneously "control/upper-bound" the reward and M costs during long runs of the controlled Markov process. Consequently, in order to control the regret vector, we make the assumption that the underlying MDP is unichain. However, this problem does not occur in the episodic RL case [24], [25] since the state is reset periodically. Second, unlike the algorithms provided in our work, [24] and [25] do not allow the agent to tune the regret vector. Very recently, we came to know that Chen et al. [26] have derived RL algorithms for CMDPs that have $\tilde{O}(\sqrt{T})$ regret guarantees and, hence, improve upon the bounds derived in this work. Wei et al. [27] also derive model-free learning algorithms for infinite-horizon average reward CMDPs and show that their reward and cost regrets are $\tilde{O}(T^{5/6})$. The work [28] claims to attain $O\sqrt{T}$ regrets for CMDPs; however, unfortunately there seems to be an error in the derivations of their proofs. More specifically, in Lemma 11, they bound the span of the CMDP by diameter. Although this argument works for MDPs [10], it is not true for CMDPs since now not only does the decision maker optimize rewards, but it also has to satisfy cost constraints. Liu et al. [29] consider an episodic setup and derive algorithms, which

[1]Throughput derived from those packets, which reach their destination within their deadline.

have $\tilde{O}\left(\sqrt{K}\right)$ reward regret, with a bounded expected number of constraint violations. Ding et al. [30] propose a primal-dual algorithm for discounted RL for CMDPs and show that its convergence rate is $O\left(1/\sqrt{T}\right)$. The work [31] summarizes recent approaches to RL in CMDPs while Liu et al. [32] apply RL for CMDPs to make dynamic decisions in network slicing applications.

The contributions of this article are summarized as follows.

1) We initiate the problem of designing RL algorithms that maximize the cumulative rewards while simultaneously satisfying average cost constraints. We propose an algorithm that we call UCRL for CMDPs, henceforth abbreviated as UCRL-CMDP. UCRL-CMDP is a modification of the popular RL algorithm UCRL2 of the work in [10] that utilizes the principle of optimism in the face of uncertainty (OFU) while making decisions. Since an algorithm that utilizes OFU does not need to satisfy cost constraints (this is briefly discussed at the end of this section), we modify OFU appropriately and derive the principle of *balanced optimism in the face of uncertainty* (BOFU). Under the BOFU principle, at the beginning of each RL episode, the agent has to solve for 1) an MDP and 2) a controller, such that the average costs of a system in which the dynamics are described by 1), and which is controlled using 2), are less than or equal to the cost constraints. This is summarized in Algorithm 1.
2) In order to quantify the finite-time performance of an RL algorithm that has to perform under average cost constraints, we define its $M+1$ dimensional "regret vector" that is composed of its reward regret (8) and M cost regrets (9). More precisely, considering solely the reward regret (as is done in the RL literature) overlooks the cost expenditures. Indeed, we show in Theorem 2 that the reward regret can be made arbitrarily small (with a high probability) at the expense of an increase in the cumulative cost expenditure. Thus, while comparing the performance of two different learning algorithms, we also need to compare their cost expenditures. The reward regret of a learning algorithm is the difference between its reward and the reward of an optimal policy that knows the MDP parameters while the ith cost regret is the difference between the total cost incurred until T time steps, and the budget on the ith expected cost $c_i^{\text{ub}}T$.
3) We ask the following question in the constrained setup: *What is the set of "achievable" $M+1$ dimensional regret vectors?* In Theorem 1, we show that the components of the regret vector of UCRL-CMDP can be bounded as $\tilde{O}\left(T^{2/3}\right)$.
4) We show that the use of BOFU allows us to overcome the shortcomings of the CE approach that were encountered in [15], i.e., there are arbitrarily long time-durations during which the CMDP's system dynamics are described by the current empirical estimates of transition probabilities is infeasible and, hence, the agent is unable to utilize these estimates in order to make control decisions. As a byproduct, BOFU also allows us to get rid of "forced explorations," i.e., employing randomized controls occasionally, which were utilized in [15].
5) Analogous to the unconstrained RL setup, in which one is interested in quantifying a lower bound on the regret of any learning algorithm, we provide a partial characterization of the set of those $M+1$-dimensional regret vectors, which cannot be achieved under any learning algorithm. More specifically, in Theorem 3, we show that a weighted sum of the $M+1$ regrets is necessarily greater than $O(D(p)S\sqrt{AT\log(T)})$, where $D(p)$ is the diameter of the underlying MDP, and S, A is the number of states and control actions, respectively.
6) In many applications, an agent is more sensitive to the cost expenditures of some specific resources compared to the rest, and a procedure to "tune" the $M+1$ dimensional regret vector is essential. In Section VI, we consider the scenario in which the agent can prespecify the desired bounds on each component of the cost regret vector and introduce a modification to the UCRL-CMDP that allows the agent to keep the cost regrets below these bounds.

Failure of OFU in constrained RL problems: Consider a two-state $\mathcal{S}=\{1,2\}$, two-action $\mathcal{A}=\{0,1\}$ MDP in which the controlled transition probabilities $p(1,1,1)=1-\theta$ and $p(1,1,2)=\theta$ are unknown while remaining probabilities are equal to 0.5. Assume that $r(1,a), c(1,a)\equiv 0$ and $r(2,a), c(2,a)\equiv 1$, i.e., reward and cost depend only upon the current state. Assume that $\theta>0.5$, and the average cost threshold satisfies $c^{\text{ub}}<2\theta/(1+2\theta)$. Since state 2 yields reward at the maximum rate, and $\theta>0.5$, this means that the optimal action in state 1 is 1. Let $\hat{\theta}_t$ and ϵ_t denote the empirical estimate of θ, and the radius of the confidence interval, respectively, at time t. Then, UCRL2 sets the optimistic estimate of θ equal to $\hat{\theta}_t+\epsilon_t$ and then implements the control that is optimal when the true parameter value is equal to this estimate. Thus, if $\hat{\theta}_t+\epsilon_t\geq 0.5$, then it chooses action 1 in state 1. Since with a high probability, we have $\hat{\theta}_t+\epsilon_t\geq\theta$, and $\hat{\theta}_t+\epsilon_t\to\theta$ as $T\to\infty$ [10], we have that when the index of the RL episode is sufficiently large, the agent implements action 1 in state 1. Since the average cost of this policy is $2\theta/(1+2\theta)$, this means that UCRL2 violates the average cost constraint.

III. Preliminaries

In our setup, at each time t, the agent earns a reward and also incurs M costs. Reward and cost functions are denoted by $r, \{c_i\}_{i=1}^M, \mathcal{S}\times\mathcal{A}\mapsto\mathbb{R}$. Thus, the instantaneous reward obtained upon taking an action a in the state s is equal to $r(s,a)$ while the ith cost is equal to $c_i(s,a)$. A controlled Markov process in which the agent earns reward and incurs M costs is defined by the tuple $\mathcal{CMP}=(\mathcal{S},\mathcal{A},p,r,c_1,c_2,\ldots,c_M)$. The controlled transition probabilities $p(s,a,s')$ are not known to the agent while the reward and cost functions $r, \{c_i\}_{i=1}^M$ are known to the agent. We will now briefly discuss some notions and results on MDPs. Let $P_{\pi,p,s}^{(t)}$ denote the t-step probability distribution when the policy π is applied to the MDP p and the initial state is

s while $P_{\pi,p}$ is the corresponding stationary measure.[2] For two measures μ_1, μ_2, we let $\|\mu_1 - \mu_2\|_V$ denote the total variation distance [33] between μ_1 and μ_2.

Definition 1: (Unichain MDP) The MDP p is unichain if under any stationary policy there is a single recurrent class. If an MDP is unichain [2], then for the Markov chain induced by any stationary policy π, we have

$$\|P^{(t)}_{\pi,p,s} - P_{\pi,p}\|_{TV} \leq C\rho^t \quad \forall s \in \mathcal{S} \tag{1}$$

where $C > 0, 1 > \rho > 0$ are constants. Let $T_{s,s'}$ denote the time taken by the Markov chain induced by a stationary policy to hit state s', when it starts in state s. The mixing time of an MDP p is defined as $T_M(p) := \max_{\pi,s,s'} \mathbb{E}_{\pi,p} T_{s,s'}$, where the subscript denotes the fact that the expectation is taken with respect to the measure induced by π when it is applied to the MDP p. We will occasionally omit its dependence upon p and denote it by T_M.

Definition 2: (Control Policy) Let $\Delta(\mathcal{A}) := \{\boldsymbol{x} \in \mathbb{R}^{|\mathcal{A}|} : \sum_{i=1}^{|\mathcal{A}|} x_i = 1, x_i \geq 0\}$ be the $|\mathcal{A}|$-simplex and $\mathcal{F}_t$ denote the sigma-algebra [34] generated by the random variables $\{(s_\ell, a_\ell)\}_{\ell=1}^{t-1} \cup s_t$. A stationary policy $\pi : \mathcal{S} \mapsto \Delta(\mathcal{A})$ prescribes randomized controls on the basis of the current state s_t. Thus, under policy π, we have that a_t is chosen according to the probability distribution $\pi(\cdot|s_t)$.

A. Notation

Throughout, we use bold font for denoting vectors; for example, the vector $(x_1, x_2, \ldots, x_N)$ is denoted by $\boldsymbol{x}$. We use $\mathbb{N}$ to denote the set of natural numbers, $\mathbb{R}^M$ to denote the M-dimensional Euclidean space, and $\mathbb{R}^M_+$ to denote non-negative orthant of $\mathbb{R}^M$. Inequalities between two vectors are to be understood componentwise. If $\mathcal{E}$ is an event [34], then $\mathbb{1}(\mathcal{E})$ denotes its indicator function. For a control policy π,[3] $\bar{r}(\pi) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_\pi \sum_{t=1}^T r(s_t, a_t)$, and[4] $\bar{c}_i(\pi) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_\pi \sum_{t=1}^T c_i(s_t, a_t)$. For $\boldsymbol{x} \in \mathbb{R}^N$, we let $\|\boldsymbol{x}\|_1$ denote its 1-norm and $\|\boldsymbol{x}\|_\infty$ be the infinity norm. $\mathbf{0}_M$ denotes the M-dimensional zero vector consisting of all zeros. For $x, y \in \mathbb{R}$, we let $x \vee y := \max\{x, y\}$. Throughout, for $M \in \mathbb{N}$, we abbreviate $[M] := \{1, 2, \ldots, M\}$, $S := |\mathcal{S}|$, $A := |\mathcal{A}|$.

B. Constrained MDPs

We now present some definitions and standard results pertaining to constrained MDPs. These can be found in [35].

Definition 3 (Occupation Measure): Consider the controlled Markov process s_t evolving under the application of a stationary policy π. Its occupation measure $\mu_\pi = \{\mu_\pi(s,a) : (s,a) \in \mathcal{S} \times \mathcal{A}\}$ is defined as $\mu_\pi(s,a) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_\pi(\sum_{t=1}^T \mathbb{1}(s_t = s, a_t = a))$ and describes the average amount of time that the process (s_t, a_t) spends on each possible state–action pair.

Definition 4 (SR(μ)): Consider a vector $\mu = \{\mu(s,a) : (s,a) \in \mathcal{S} \times \mathcal{A}\}$ that satisfies the constraints (6) and (7). Define SR(μ) to be the following stationary randomized policy. When the state s_t is equal to s, the policy chooses the action a with a probability equal to $\frac{\mu(s,a)}{\sum_{a' \in \mathcal{A}} \mu(s,a')}$ if $\sum_{a' \in \mathcal{A}} \mu(s,a') > 0$. However, if $\sum_{a' \in \mathcal{A}} \mu(s,a') = 0$, then the policy takes an action according to some prespecified rule (e.g., implement $a_t = 0$).

Constrained MDP (CMDP): The following dynamic optimization problem is a CMDP [35]:

$$\max_\pi \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}_\pi \sum_{t=1}^T r(s_t, a_t) \tag{2}$$

$$\text{s.t. } \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_\pi \sum_{t=1}^T c_i(s_t, a_t) \leq c_i^{\text{ub}}, i \in [M] \tag{3}$$

where the maximization above is over the class of all history-dependent policies, and c_i^{ub} denotes the desired upper-bound on the average value of ith cost expense. The optimal average reward rate of the CMDP is equal to the optimal value of the above LP and is denoted by $r^\star$.

Linear Programming approach for solving CMDPs: When the controlled transition probabilities $p(s,a,s')$ are known, and p is unichain, an optimal policy for the CMDP (2)–(3) can be obtained by solving the following LP [35]:

$$\max_{\mu=\{\mu(s,a):(s,a)\in\mathcal{S}\times\mathcal{A}\}} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mu(s,a) r(s,a) \tag{4}$$

$$\text{s.t.} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mu(s,a) c_i(s,a) \leq c_i^{\text{ub}}, i \in [M] \tag{5}$$

$$\sum_{a\in\mathcal{A}} \mu(s,a) = \sum_{(s',b)\in\mathcal{S}\times\mathcal{A}} \mu(s',b) p(s',b,s) \ \forall s \in \mathcal{S} \tag{6}$$

$$\mu(s,a) \geq 0 \ \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mu(s,a) = 1. \tag{7}$$

If $\mu^\star$ is a solution of the above LP, then SR($\mu^\star$) solves (2)–(3). Moreover, it can be shown that the average reward and M costs of SR($\mu^\star$) are independent of the initial starting state s_0 if the MDP is unichain [35].

C. Learning Algorithms and Regret Vector

We will develop RL algorithms to solve the finite-time horizon version of the CMDP (2)–(3) when the probabilities $p(s,a,s')$ are not known to the agent. A learning policy π chooses action a_t on the basis of past operational history of the system. In order to measure the performance of a learning algorithm, we define its reward and cost regrets. The "cumulative reward regret" until time T, denoted by $\Delta^{(R)}(T)$, is defined as

$$\Delta^{(R)}(T) := r^\star T - \sum_{t=1}^T r(s_t, a_t) \tag{8}$$

where $r^\star$ is the optimal average reward of the CMDP (2)–(3) when controlled transition probabilities $p(s,a,s')$ are known. Note that $r^\star$ is the optimal value of the LP (4)–(7). The "cumulative cost regret" for the ith cost until time T is denoted by

[2]Under the assumption that a unique stationary measure exists.
[3]In case limit does not exist, lim should be replaced by lim inf.
[4]In case limit does not exist, lim should be replaced by lim sup.

Algorithm 1: UCRL-CMDP

Input: State-space $\mathcal{S}$, Action-space $\mathcal{A}$, Confidence parameter δ, Time horizon T

Initialize: Set $t := 1$, and observe the initial state s_1.

for Episodes $k = 1, 2, \ldots$ **do**

Initialize Episode k:

1) Set the start time of episode k, $\tau_k := t$. For all state-action tuples $(s, a) \in \mathcal{S} \times \mathcal{A}$, initialize the number of visits within episode k, $n_k(s, a) = 0$.
2) For all $(s, a) \in \mathcal{S} \times \mathcal{A}$ set $N_{\tau_k}(s, a)$, i.e., the number of visits to (s, a) prior to episode k. Also, set the transition counts $N_{\tau_k}(s, a, s')$ for all $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$.
3) Compute the empirical estimate $\hat{p}_t$ of the MDP as in (10).

Compute Policy $\tilde{\pi}_k$:

1) Let $\mathcal{C}_{\tau_k}$ be the set of plausible MDPs as in (11).
2) Solve (12)–(16) to obtain $\tilde{\pi}_k$.
3) In case (12)–(16) is infeasible, choose $\tilde{\pi}_k$ to be some predetermined policy (chosen at time $t = 0$).

Implement $\tilde{\pi}_k$:

while $t - \tau_{k_t} < \lceil T^\alpha \rceil$ **do**

1) Sample a_t according to the distribution $\tilde{\pi}_k(\cdot|s_t)$. Observe reward $r(s_t, a_t)$, and observe next state s_{t+1}.
2) Update $n_k(s_t, a_t) = n_k(s_t, a_t) + 1$.
3) Set $t := t + 1$.

end while

end for

$\Delta^{(i)}(T)$, and is defined as

$$\Delta^{(i)}(T) := \sum_{t=1}^{T} c_i(s_t, a_t) - c_i^{\text{ub}}\, T. \tag{9}$$

IV. UCRL-CMDP: A Learning Algorithm for CMDPs

We propose UCRL-CMDP to adaptively control an unknown CMDP. It is depicted in Algorithm 1. UCRL-CMDP maintains empirical estimates of the each transition probability $p(s, a, s')$ as follows:

$$\hat{p}_t(s, a, s') = \begin{cases} \frac{N_t(s,a,s')}{N_t(s,a)} \text{ if } N_t(s, a) > 0 \\ \frac{1}{S}, \text{ otherwise} \end{cases} \tag{10}$$

where $N_t(s, a)$ and $N_t(s, a, s')$ denote the number of visits to (s, a) and (s, a, s') until t, respectively.

Confidence Intervals: Additionally, it also maintains confidence interval $\mathcal{C}_t$ associated with the estimate $\hat{p}_t$ as follows:

$$\mathcal{C}_t := \Bigg\{ p' : \sum_{s' \in \mathcal{S}} p'(s, a, s') = 1 \, \forall (s, a), p'(s, a, s') \geq 0$$
$$|p'(s, a, s') - \hat{p}_t(s, a, s')| \leq \epsilon_t(s, a) \quad \forall (s, a) \Bigg\} \tag{11}$$

where $\epsilon_t(s, a) := \sqrt{\frac{2 \log(T^b |\mathcal{S}||\mathcal{A}|)}{N_t(s,a) \vee 1}}$, $b > 1$ is an agent-specified constant.

Episode: UCRL-CMDP proceeds in episodes and utilizes a single stationary control policy within an episode. Each episode is of duration $\lceil T^\alpha \rceil$ steps.[5] Let τ_k denote the start time of episode k. kth episode is denoted by $\mathcal{E}_k := \{\tau_k, \tau_k + 1, \ldots, \tau_{k+1} - 1\}$, and comprises of $\tau_{k+1} - \tau_k$ consecutive time-steps. Denote by k_t the index of the ongoing episode at time t. At the beginning of $\mathcal{E}_k$, the agent solves the following *constrained* optimization problem in which the decision variables are 1) occupation measure $\mu = \{\mu(s, a) : (s, a) \in \mathcal{S} \times \mathcal{A}\}$ of the controlled process, and 2) "candidate" MDP p'

$$\max_{\mu, p'} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mu(s, a) r(s, a) \tag{12}$$

$$\text{s.t.} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mu(s, a) c_i(s, a) \leq c_i^{\text{ub}}, i \in [M] \tag{13}$$

$$\sum_{a \in \mathcal{A}} \mu(s, a) = \sum_{(s', b)} \mu(s', b) p'(s', b, s) \; \forall s \in \mathcal{S} \tag{14}$$

$$\mu(s, a) \geq 0 \; \forall (s, a), \sum_{(s,a)} \mu(s, a) = 1 \tag{15}$$

$$p' \in \mathcal{C}_{\tau_k}. \tag{16}$$

The maximization with respect to p' denotes that the agent is optimistic regarding the belief of the "true" (but unknown) MDP p while that with respect to μ ensures that the agent optimizes its control strategy for this optimistic MDP. The constraints (13) ensure that the cost expenditures do not exceed the thresholds $\{c_i^{\text{ub}}\}_{i=1}^M$ and hence ensure that the agent also balances the cost expenses while being optimistic with respect to the rewards about the choice of the MDP thereby taking a balanced approach to optimism when the underlying MDP parameters are unknown. If the constraints (13) were absent, we would recover the UCRL2 algorithm of [10] that is based on the OFU principle [11]. However, as is shown in Section II, the OFU principle might fail when it is applied to learning the optimal controls for CMDPs. Indeed, as is shown in the example of Section II, the limiting average cost is greater than the threshold value of cost. The BOFU principle proposed in this work is a natural extension of the OFU principle to the case when the agent has to satisfy certain constraints on costs, in addition to maximizing the rewards. In case the problem (12)–(16) is feasible, let $(\tilde{\mu}_k, \tilde{p}_k)$ denote a solution. The agent then chooses a_t according to $\text{SR}(\tilde{\mu}_k)$ within $\mathcal{E}_k$. However, in case (12)–(16) is infeasible, the agent implements an arbitrary stationary control policy that has been chosen at time $t = 0$. In summary, it implements a stationary controller within $\mathcal{E}_k$, which is denoted by $\tilde{\pi}_k$. We make the following assumptions on the MDP p while analyzing UCRL-CMDP.

Assumption 1:

1) The MDP $p = \{p(s, a, s') : s, s' \in \mathcal{S}, a \in \mathcal{A}\}$ is unichain. Thus, under any stationary policy π, we have

$$\|P_{\pi,p,s}^{(t)} - P_{\pi,p}\|_{TV} \leq C \rho^t, \; t = 1, 2, \ldots, s \in \mathcal{S} \tag{17}$$

where $C > 0, 0 \leq \rho < 1$.

[5] If $x \in \mathbb{R}$, we let $\lceil x \rceil$ be the least integer greater than or equal to x.

2) The CMDP (2)–(3) is feasible.
3) Without loss of generality, we assume that the magnitude of rewards and costs are upper-bounded by 1, and hence, $r^\star \leq 1$ as well as $\{c_i^{\text{ub}}\}_{i=1}^M$ can be taken to be less than 1.

We establish the following bound on the regrets of UCRL-CMDP. It is proved in the next section.

Theorem 1: Consider the UCRL-CMDP (Algorithm 1) applied with $\delta = 1/T^{1/3}, \alpha = 1/3$ to an MDP p that satisfies Assumption 1. The reward and cost regrets can be bounded as follows:

$$\mathbb{E}\Delta^{(R)}(T), \mathbb{E}\Delta^{(i)}(T),\ i \in [M] \leq 4T_M(p)\sqrt{2\log\left(T^b|\mathcal{S}||\mathcal{A}|\right)} \times \left(\left(\sqrt{2}+1\right)\sqrt{SAT} + T^{2/3}\sqrt{\log\left(SAT^{4/3}\right)}\right) + \frac{C\lceil T^{2/3}\rceil}{1-\rho} + T^{2/3} + \frac{2}{T^{2b-2}|\mathcal{S}||\mathcal{A}|}. \tag{18}$$

A detailed proof is provided in Section V. Over here, we only provide a proofsketch.

Proofsketch: We show that the proposed algorithm can be interpreted as an "index policy" in which it assigns an index (20) to each policy that is calculated on the basis of operational history and then plays the policy with the highest index. We use this characterization in order to analyze the behavior of the algorithm on the "good set," $\mathcal{G}$ (21) on which the following two occur: 1) concentration of the empirical estimate of p, and 2) the number of times (s, a) is visited is proportional to the number of times those set of policies are implemented under which (s, a) is visited with a positive probability. In Lemma 1 and Lemma 2, we show that $\mathcal{G}$ occurs with a high probability; since the regret on $\mathcal{G}^c$ is bounded as $O(T)$, it suffices to analyze the algorithm on $\mathcal{G}$. Lemma 7 shows that the instantaneous regrets depend on the radius of the confidence ball. The behavior of the radius of the ball upon playing a suboptimal policy is then used to complete the proof.

Remark 1: In comparison with the $\tilde{O}(\sqrt{T})$ regret bounds for unconstrained RL, our bounds for the constrained case are $\tilde{O}(T^{2/3})$. The reason for this is that the proof techniques of [10] cannot be applied. More specifically, for the former case, one is able to relate the diameter $D(p)$ of the MDP to a bound on the span of the relative value function $h_k(\cdot)$, of the optimistic MDP obtained during the kth episode[6] as follows: Suppose that $h_k(s) - h_k(s') > D(p)$, then one would obtain a contradiction since we can construct a policy for the extended MDP, which starts in state s' and reaches s in $D(p)$ steps (in expectation), so that the "missed rewards" on account of starting in s' (as opposed to starting in s) is upper-bounded by $D(p)$. Indeed, one could always choose the true transition probabilities p at every step, and implement a policy, which takes from s' to s in $D(p)$ steps (that such a policy exists, follows from definition of the diameter). However, in the case of CMDP, the agent is not only maximizing the rewards, but also making sure that the cost expenditures are below their respective thresholds, i.e., it is solving a multiobjective optimization problem and it is not clear how to convert these multiobjective criteria to a scalar objective function. One could argue that consideration of the Lagrangian would allow us to "scalarize" this problem, so that we could derive an upper-bound on the span of the bias function associated with the extended MDP that maximizes $r(x(t), u(t)) + \sum_i \lambda_i c_i(x(t), u(t))$. However, this result will then depend upon the values of Lagrange multipliers λ_i, $i = 1, 2, \ldots, M$, and in order for such upper-bounds to be useful, we would have to derive bounds on these multipliers. It is not clear how such a bound could be derived. In order to overcome this difficulty, we instead view UCRL-CMDP as an index policy, derive upper and lower bounds on the indices of stationary policies, and upper-bound the number of times "suboptimal policies" are played.

V. Proof of Theorem 1

We begin by introducing few notations. If $\mathcal{B}$ denotes a subset of $\mathcal{S}$, then we let $\Pi_{\mathcal{B}}$ be the set of those stationary policies for which $P_{\pi,p}(s) > 0$ for all $s \in \mathcal{B}$. Let $\mathcal{B}_\pi$ denote the set of states for which $P_{\pi,p}(s) > 0$. We now derive a few preliminary results that are used while proving the main result. The following result can be shown by an application of Azuma–Hoeffding inequality [36].

Lemma 1: Define $\mathcal{G}_1 := \{p \in \mathcal{C}_{\tau_k}\ \forall k = 1, 2, \ldots, K\}$. Then, $\mathbb{P}(\mathcal{G}_1) \geq 1 - \frac{1}{T^{b-1-(1-\alpha)}}$.

Lemma 2: Let $n_k(s, a)$ denote the number of visits to (s, a) during $\mathcal{E}_k$, and $\beta > 1/2$ satisfy $2\beta - \alpha = 1$. Define

$$\mathcal{G}_2 := \left\{\sum_{k=1}^K \frac{n_k(s,a) - \mathbb{E}(n_k(s,a)|\mathcal{F}_{\tau_k})}{\sqrt{N_k(s,a)}} \leq T^\beta\sqrt{\log\left(\frac{SAT}{\delta}\right)}\ \forall (s,a) \in \mathcal{S}\times\mathcal{A}\right\} \tag{19}$$

where K is the total number of episodes. We have $\mathbb{P}(\mathcal{G}_2) \geq 1 - \frac{\delta}{T}$.

Proof: Note that $\frac{n_k(s,a) - \mathbb{E}(n_k(s,a)|\mathcal{F}_{\tau_k})}{\sqrt{N_k(s,a)}}$, $k = 1, 2, \ldots, K$ is a martingale difference sequence. Furthermore, since the duration of each episode is $\lceil T^\alpha\rceil$, and $\sqrt{N_k(s,a)} \geq 1$, we have $\frac{n_k(s,a) - \mathbb{E}(n_k(s,a)|\mathcal{F}_{\tau_k})}{\sqrt{N_k(s,a)}} \leq \lceil T^\alpha\rceil$. By applying Azuma–Hoeffding's inequality to this martingale difference sequence, we get that the probability of the event $\sum_{k=1}^K \frac{n_k(s,a) - \mathbb{E}(n_k(s,a)|\mathcal{F}_{\tau_k})}{\sqrt{N_k(s,a)}} \geq T^\beta\sqrt{\log\left(\frac{SAT}{\delta}\right)}$ can be upper-bounded by $\exp\left(-\frac{T^{2\beta}}{T^{1-\alpha}T^{2\alpha}}\log\frac{SAT}{\delta}\right) = \exp\left(-T^{2\beta-(1+\alpha)}\log\frac{SAT}{\delta}\right)$. Since $2\beta - (1+\alpha) = 0$, the above bound reduces to $\frac{\delta}{SAT}$. The proof then follows by using union bound for all state–action pairs (s, a). ■

Lemma 3: If $s \in \mathcal{B}_{\pi_k}$, then[7]

$$\mathbb{E}\left(n_k(s,a)|\mathcal{F}_{\tau_k}\right) \geq \left\lfloor\frac{\lceil T^\alpha\rceil}{2T_M(p)}\right\rfloor \times \frac{\pi_k(a|s)}{2}.$$

[6] See [10] for more details.

[7] For $x \in \mathbb{R}$, we let $\lfloor x\rfloor$ be the greatest integer less than or equal to x.

Proof: Within this proof, we use T_M to denote $T_M(p)$. Since we have $\mathbb{E}_{\pi,p}T_{s',s} \leq T_M \; \forall s' \in \mathcal{S}$, it follows from Markov's inequality that the probability with which s_t does not hit the state s in $2T_M$ steps, is less than $1/2$, or equivalently the state s is visited at least once with a probability greater than $1/2$, which yields us $\min_{s' \in \mathcal{S}} \mathbb{E}_\pi \left(\sum_{t=1}^{2T_M} \mathbb{1}\{s_t = s\} | s_0 = s' \right) \geq 1/2$. The proof is then completed by dividing the total time of $\lceil T^\alpha \rceil$ steps in an episode into "mini-episodes" of $2T_M$ steps each, and noting that $n_k(s,a)$ is equal to the sum of the number of visits to (s,a) during each such miniepisode. ■

We begin by giving an equivalent characterization of the UCRL-CMDP rule. At each τ_k, it assigns an index $\mathcal{I}_k(\pi)$ to each stationary policy π as follows:

$$\mathcal{I}_k(\pi) := \max_{\theta \in \mathcal{C}_{\tau_k}} \left\{ \bar{r}(\pi,\theta) : \bar{c}_i(\pi,\theta) \leq c_i^{\text{ub}}, \, i \in [M] \right\}. \quad (20)$$

In case the above optimization problem is infeasible, i.e., $\bar{c}_i(\pi,\theta) > c_i^{\text{ub}} \; \forall \theta \in \mathcal{C}_{\tau_k}$ for some i, then the policy is assigned an index of $-\infty$. UCRL-CMDP implements a policy with the largest index during $\mathcal{E}_k$.

Define the "good set"

$$\mathcal{G} := \mathcal{G}_1 \cap \mathcal{G}_2. \quad (21)$$

Lemma 4: On the set $\mathcal{G}$, we have the following for $\theta \in \mathcal{C}_{\tau_k}$:

$$|\bar{r}(\pi,p) - \bar{r}(\pi,\theta)|, |\bar{c}_i(\pi,p) - \bar{c}_i(\pi,\theta)|, \, i \in [M]$$
$$\leq 2 \max_s \sum_{a \in \mathcal{A}} \pi(a|s) \epsilon_{\tau_k}(s,a). \quad (22)$$

Proof: Note that $P^{(1)}_{\pi,p,s}$ is the vector of transition probabilities from state s of the Markov chain that results when the policy π is applied to the MDP p. Consider an MDP $\theta \in \mathcal{C}_{\tau_k}$. Since on $\mathcal{G}$, we have $p \in \mathcal{C}_{\tau_k}$, we have $\|P^{(1)}_{\pi,\hat{p}_{\tau_k},s} - P^{(1)}_{\pi,p,s}\|_1, \|P^{(1)}_{\pi,\hat{p}_{\tau_k},s} - P^{(1)}_{\pi,\theta,s}\|_1 \leq \sum_{a \in \mathcal{A}} \pi(a|s) \epsilon_{\tau_k}(s,a)$, where $\pi(a|s)$ is the probability with which the policy implements a in state s. From triangle inequality, we have that $\|P^{(1)}_{\pi,p,s} - P^{(1)}_{\pi,\theta,s}\|_1 \leq 2 \sum_{a \in \mathcal{A}} \pi(a|s) \epsilon_{\tau_k}(s,a)$. Equation (22) then follows from [37], Corollary 3.1]. ■

For a stationary policy π, we say $r^\star - \bar{r}(\pi,p)$ is its instantaneous reward regret, and $\bar{c}_i(\pi,p) - c_i^{\text{ub}}$ is its instantaneous cost regret for the ith cost. We now show that if the instantaneous reward regret, or an instantaneous cost regret of a policy is greater than a certain threshold, this threshold depends upon the radius of the confidence ball at time τ_k, then it is not played during $\mathcal{E}_k$. For a stationary policy π, define

$$\delta_k(\pi) := 2 \max_{s \in \mathcal{B}_\pi} \sum_{a \in \mathcal{A}} \pi(a|s) \epsilon_{\tau_k}(s,a).$$

Consider the following two possibilities.

Case A) $\bar{c}_i(\pi,p) > c_i^{\text{ub}} + \delta_k(\pi)$ for some i: From (22), we have that $|\bar{c}_i(\pi,p) - \bar{c}_i(\pi,\theta)| \leq \delta_k(\pi)$, which implies $\bar{c}_i(\pi,\theta) > c_i^{\text{ub}}$ for all $\theta \in \mathcal{C}_{\tau_k}$. Thus, $\mathcal{I}_k(\pi) = -\infty$.

Case B) From (22), we have that $|\bar{r}(\pi,p) - \bar{r}(\pi,\theta)| \leq \delta_k(\pi)$ for all $\theta \in \mathcal{C}_{\tau_k}$, so that the index $\mathcal{I}_k(\pi)$ is bounded by $\bar{r}(\pi,p) + \delta_k(\pi)$.

The following result summarizes this discussion.

Lemma 5: Let π be a stationary randomized policy. On the set $\mathcal{G}$, we have that $\mathcal{I}_k(\pi) = -\infty$ if $\bar{c}_i(\pi,p) > c_i^{\text{ub}} + \delta_k(\pi)$, for some $i \in [M]$. Also, $\mathcal{I}_k(\pi) \leq \bar{r}(\pi,p) + \delta_k(\pi)$.

We now show that if a stationary policy is feasible for the MDP p, i.e., $\bar{c}_i(\pi,p) \leq c_i^{\text{ub}} \; \forall i$, then its index $\mathcal{I}_k(\pi)$ is lower-bounded by $\bar{r}(\pi,p)$.

Lemma 6: If π is feasible for the true MDP, i.e., it satisfies $\bar{c}_i(\pi,p) \leq c_i^{\text{ub}} \; \forall i \in [M]$, then on $\mathcal{G}$, its index satisfies $\mathcal{I}_k(\pi) \geq \bar{r}(\pi,p)$. With π set equal to the policy that solves the CMDP $\max_\pi \{\bar{r}(\pi,p) : \bar{c}_i(\pi,p) \leq c_i^{\text{ub}} \; \forall i \in [M]\}$, we obtain that the index of an optimal policy is greater than $r^\star$.

Proof: Note that on the set $\mathcal{G}$, the true MDP p always belongs to $\mathcal{C}_{\tau_k}$. If $\bar{c}_i(\pi,p) \leq c_i^{\text{ub}} \; \forall i \in [M]$, we have $\mathcal{I}_k(\pi) = \max_{\theta \in \mathcal{C}_{\tau_k}} \{\bar{r}(\pi,\theta) : \bar{c}_i(\pi,\theta) \leq c_i^{\text{ub}}, \, i \in [M]\} \geq \bar{r}(\pi,p)$. ■

Upon combining Lemma 5 and Lemma 6, we obtain the following result.

Lemma 7: On the set $\mathcal{G}$, the instantaneous regrets during $\mathcal{E}_k$ can be bounded by $\delta_k(\pi_k)$.

Proof: We begin by bounding cost regrets. Consider a stationary policy π. In case $\bar{c}_i(\pi,p) > c_i^{\text{ub}} + \delta_k(\pi_k)$, then it follows from Lemma 5 that $\mathcal{I}_k(\pi) = -\infty$. However, it is shown in Lemma 6 that there is a policy $\tilde{\pi}$, which is feasible for the true MDP, and whose index is greater than $r^\star$. In case the index of π is less than the index of $\tilde{\pi}$, the policy π would not be played by UCRL-CMDP. This means that in order for π to be a candidate to be played during $\mathcal{E}_k$, we must have $\bar{c}_i(\pi,p) \leq c_i^{\text{ub}} + \delta_k(\pi_k)$, or equivalently the instantaneous cost regret of π must be bounded by $\delta_k(\pi_k)$.

In order to bound the reward regret, we note that it was shown in Lemma 6 that the index of an optimal policy is greater than $r^\star$, and since the index $\mathcal{I}_k(\pi_k)$ must be greater than or equal to the index of an optimal policy, we must have $\mathcal{I}_k(\pi_k) \geq r^\star$. From Lemma 5, we have $\mathcal{I}_k(\pi_k) \leq \bar{r}(\pi,p) + \delta_k(\pi_k)$. Upon combining these inequalities, we obtain $\bar{r}(\pi,p) + \delta_k(\pi_k) \geq r^\star$, or $\bar{r}(\pi,p) \geq r^\star - \delta_k(\pi_k)$. This shows that the instantaneous reward regret $r^\star - \bar{r}(\pi,p)$ is bounded by $\delta_k(\pi_k)$. ■

We now use the result on instantaneous regrets in order to bound the cumulative regrets of UCRL-CMDP.

Proof of Theorem 1: We will only derive upper-bound on the reward regret, since the bound on cost regrets can be derived by following similar steps. Now, $\mathbb{E}\left(\sum_{t \in \mathcal{E}_k} r^\star - r(s_t,a_t)\right) = \mathbb{E}\left(\mathbb{E}\left\{\sum_{t \in \mathcal{E}_k} r^\star - \bar{r}(\pi_k,p) + \bar{r}(\pi_k,p) - r(s_t,a_t) \middle| \mathcal{F}_{\tau_k}\right\}\right)$. It follows from (17) that we have $\mathbb{E}\{\sum_{t \in \mathcal{E}_k} \bar{r}(\pi_k,p) - r(s_t,a_t) | \mathcal{F}_{\tau_k}\} \leq \frac{C}{1-\rho}$, and hence the expected regret during $\mathcal{E}_k$ can be bounded by $\mathbb{E}\left(\mathbb{E}\left\{\sum_{t \in \mathcal{E}_k} r^\star - \bar{r}(\pi_k,p) \middle| \mathcal{F}_{\tau_k}\right\}\right) + \frac{C}{1-\rho}$. Let $\Delta_k^{(R)} := \mathbb{E}\left\{\sum_{t \in \mathcal{E}_k} r^\star - \bar{r}(\pi_k,p) | \mathcal{F}_{\tau_k}\right\}$ denote the regret incurred during the kth episode. Thus, the cumulative expected regret can be bounded as follows:

$$\mathbb{E}\Delta^{(R)}(T) \leq \mathbb{E}\left(\sum_{k=1}^{K} \Delta_k^{(R)}\right) + K\frac{C}{1-\rho} \quad (23)$$

where K is the total number of episodes. Henceforth, we will focus on bounding the first term $\sum_{k=1}^{K} \Delta_k^{(R)}$ in the r.h.s. above. This is bounded separately on the sets $\mathcal{G}, \mathcal{G}_1^c, \mathcal{G}_2^c$.

We begin by bounding $\sum_{k=1}^{K}\Delta_k^{(R)}$ on $\mathcal{G}$. Since from Lemma 7, the instantaneous regret on $\mathcal{G}$ during $\mathcal{E}_k$ can be bounded by $\delta_k(\pi_k)$, we have

$$\begin{aligned}
&\Delta_k^{(R)} \le \delta_k(\pi_k)|\mathcal{E}_k| \\
&\le 4T_M \sum_{(s,a):s\in\mathcal{B}_{\pi_k}} \mathbb{E}(n_k(s)|\mathcal{F}_k)\frac{\pi_k(a|s)\sqrt{2\log\left(T^b|\mathcal{S}||\mathcal{A}|\right)}}{\sqrt{N_k(s,a)}} \\
&\quad + 4T_M \sum_{(s,a):s\in\mathcal{B}_{\pi_k}} \left\{\frac{|\mathcal{E}_k|}{2T_M}\frac{1}{2} - \mathbb{E}(n_k(s)|\mathcal{F}_k)\right\} \\
&\quad \times \quad \frac{\pi_k(a|s)\sqrt{2\log\left(T^b|\mathcal{S}||\mathcal{A}|\right)}}{\sqrt{N_k(s,a)}} \\
&\le 4T_M \\
&\quad \times \sum_{(s,a):s\in\mathcal{B}_{\pi_k}} \mathbb{E}(n_k(s)|\mathcal{F}_k)\frac{\pi_k(a|s)\sqrt{2\log\left(T^b|\mathcal{S}||\mathcal{A}|\right)}}{\sqrt{N_k(s,a)}}
\end{aligned} \tag{24}$$

where the last inequality follows from Lemma 3. We will now bound the term $\sum_{k=1}^{K}\sum_{(s,a):s\in\mathcal{B}_{\pi_k}}\frac{\mathbb{E}(n_k(s)|\mathcal{F}_k)\pi_k(a|s)}{\sqrt{N_k(s,a)}}$. We have

$$\begin{aligned}
&\sum_{k=1}^{K}\frac{\mathbb{E}(n_k(s)|\mathcal{F}_k)\pi_k(a|s)}{\sqrt{N_k(s,a)}} = \sum_{k=1}^{K}\frac{\mathbb{E}(n_k(s,a)|\mathcal{F}_k)}{\sqrt{N_k(s,a)}} \\
&= \sum_{k=1}^{K}\frac{n_k(s,a)}{\sqrt{N_k(s,a)}} + \sum_{k=1}^{K}\frac{\mathbb{E}(n_k(s,a)|\mathcal{F}_k) - n_k(s,a)}{\sqrt{N_k(s,a)}}.
\end{aligned} \tag{25}$$

As is shown in [10], p. 1578], the term $\sum_{k=1}^{K}\sum_{(s,a)}\frac{n_k(s,a)}{\sqrt{N_k(s,a)}}$ can be bounded by $\left(\sqrt{2}+1\right)\sqrt{SAT}$ on each sample path while from (19), we have that on $\mathcal{G}_2$, the term $\sum_{k=1}^{K}\frac{\mathbb{E}(n_k(s,a)|\mathcal{F}_k)-n_k(s,a)}{\sqrt{N_k(s,a)}}$ is bounded by $T^{\beta}\sqrt{\log\left(\frac{SAT}{\delta}\right)}$. It follows from (25) and the discussion above that on $\mathcal{G}$, we have $\sum_{(s,a):s\in\mathcal{B}_{\pi_k}}\sum_{k=1}^{K}\frac{\mathbb{E}(n_k(s)|\mathcal{F}_k)\pi_k(a|s)}{\sqrt{N_k(s,a)}} \le \left(\sqrt{2}+1\right)\sqrt{SAT} + T^{\beta}\log^{1/2}(SAT/\delta)$. Upon summing (24) over episodes, and using the above inequality, we obtain that the regret on $\mathcal{G}$ can be bounded as follows:

$$\begin{aligned}
&\sum_{k=1}^{K}\Delta_k^{(R)} \le 4T_M\sqrt{2\log\left(T^b|\mathcal{S}||\mathcal{A}|\right)} \\
&\times\left(\left(\sqrt{2}+1\right)\sqrt{SAT} + T^{\beta}\log^{1/2}\left(SAT/\delta\right)\right).
\end{aligned} \tag{26}$$

This completes the analysis on $\mathcal{G}$.

We now analyze the regret on $\mathcal{G}_2^c$. From Lemma 2, the probability of $\mathcal{G}_2^c$ is bounded by δ. On $\mathcal{G}_2^c$, the sample path regret $\sum_{k=1}^{K}\Delta_k^{(R)}$ can be trivially bounded by T, so that its contribution to the expected regret is bounded by δT.

To analyze the regret on $\mathcal{G}_1^c$, we note that if the confidence ball $\mathcal{C}_{\tau_k}$ at time τ_k fails, then the regret during $\mathcal{E}_k$ can be bounded by the duration of $\mathcal{E}_k$. Since $\tau_{k+1}-\tau_k = \lceil T^{\alpha}\rceil$, the regret during $\mathcal{E}_k$ is bounded by $\lceil T^{\alpha}\rceil$. From Lemma 1, we have that the probability with which confidence ball fails at time t is upper-bounded by $\frac{2}{T^{2b-1}|\mathcal{S}||\mathcal{A}|}$. Hence, the expected regret from the failure of ball (in case an episode starts at t) at time t is bounded by $\frac{2\lceil T^{\alpha}\rceil}{T^{2b-1}|\mathcal{S}||\mathcal{A}|}$, so that the cumulative expected regret is bounded by $\frac{2}{T^{2b-2}|\mathcal{S}||\mathcal{A}|}$. ■

VI. Learning Under-Bounds on Cost Regret

The upper-bounds for the regrets of UCRL-CMDP in Theorem 1 are the same for reward and M costs regrets. However, in many practical applications, an agent is more sensitive to overutilizing certain specific costs, as compared to the other costs. Thus, in this section, we derive algorithms that enable the agent to tune the upper-bounds on the regrets of different costs. We also quantify the reward regret of these algorithms.

A. Modified UCRL-CMDP

Throughout this section, we assume that p satisfies the following condition.

Assumption 2: For the MDP p, there exists a stationary policy under which the average costs are strictly below the thresholds $\{c_i^{\text{ub}} : i = 1, 2, \ldots, M\}$. More precisely, there exists an $\epsilon > 0$ and a stationary policy $\pi_{\text{feas.}}$ such that we have $\bar{c}_i(\pi_{\text{feas.}}) < c_i^{\text{ub}} - \epsilon \quad \forall i \in [M]$. Define

$$\eta := \min_{i\in[M]}\left\{c_i^{\text{ub}} - \epsilon - \bar{c}_i(\pi_{\text{feas.}})\right\}. \tag{27}$$

The modified algorithm maintains empirical estimates $\hat{p}_t$ and confidence intervals $\mathcal{C}_t$ (11) in exactly the same manner as UCRL-CMDP (see Algorithm 1) does. It also proceeds in episodes of duration $\lceil T^{\alpha}\rceil$ steps and uses a single stationary control policy within an episode. However, at the beginning of each episode k, it solves an optimization problem, which is a modification of (12)–(16). More concretely, the cost constraints (13) are replaced by the following modified constraints:

$$\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\mu(s,a)c_i(s,a) \le c_i^{\text{ub}} - d_i,\ i \in [M]$$

where $d_i := b_i\epsilon, \quad i \in [M]$, and the parameters $b_i \in (0,1), i \in [M]$ are chosen by the agent. If this problem is feasible, let $\tilde{\mu}_k$ be an optimal occupation measure obtained by solving it. In this case, the agent implements $\text{SR}(\tilde{\mu}_k)$ within $\mathcal{E}_k$. However, in case the problem is infeasible, then it implements a stationary controller that has been chosen at time $t = 0$. We derive upper-bounds on the regrets of the modified UCRL-CMDP algorithm in the following result.

Theorem 2: Consider the modified UCRL-CMDP algorithm with $\delta = 1/T^{1/3}$, $\alpha = 1/3$ applied to an MDP p that satisfies Assumption 1 and Assumption 2. Then, the expected reward regret can be upper-bounded as follows:

$$\begin{aligned}
&\mathbb{E}\Delta^{(R)}(T) \\
&\le 4T_M\left(\left(\sqrt{2}+1\right)\sqrt{SAT} + T^{2/3}\sqrt{\log\left(SAT^{4/3}\right)}\right) \\
&+ \frac{C\left\lceil T^{2/3}\right\rceil}{1-\rho} + T^{2/3} + \frac{2}{T^{2b-2}|\mathcal{S}||\mathcal{A}|} + zT
\end{aligned} \tag{28}$$

where $z = (\max_i b_i)\frac{\hat{\eta}}{\eta}\epsilon$, η is as in (27) and

$$\hat{\eta} := \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} r(s,a) - \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} r(s,a). \quad (29)$$

The expected cost regret can be upper-bounded as follows:

$$\mathbb{E}\Delta^{(i)}(T) \leq 4T_M\sqrt{2\log\left(T^b|\mathcal{S}||\mathcal{A}|\right)} \times \left(\left(\sqrt{2}+1\right)\sqrt{SAT} + T^{2/3}\sqrt{\log\left(SAT^{4/3}\right)}\right) + \frac{C\left\lceil T^{2/3}\right\rceil}{1-\rho} + T^{2/3} + \frac{2}{T^{2b-2}|\mathcal{S}||\mathcal{A}|} - b_i\epsilon T, i \in [M]. \quad (30)$$

Remark 2: Note that the prefactor in the $O(T)$ term in (28) depends upon ϵ linearly, and this quantity can be tuned by the agent. When $\epsilon = T^{-1/3}$, then $\mathbb{E}\Delta^{(R)}(T)$ can be bounded as $O\left(T^{2/3}\right)$.

VII. Proof of Theorem 2

Proof closely follows the proof of Theorem 1, hence we point out only the key differences. The modified UCRL-CMDP algorithm assigns the following modified index[8] to policy π:

$$\mathcal{I}_k(\pi) := \max_{\theta\in\mathcal{C}_{\tau_k}} \left\{\bar{r}(\pi,\theta) : \bar{c}_i(\pi,\theta) \leq c_i^{\text{ub}} - d_i,\ i \in [M]\right\}.$$

If for some i we have $\bar{c}_i(\pi,\theta) > c_i^{\text{ub}} - d_i \quad \forall\theta \in \mathcal{C}_{\tau_k}$, then we set $\mathcal{I}_k(\pi) = -\infty$.

The proof of next result is omitted since it is similar to that of Lemma 5.

Lemma 8: Let π be a stationary randomized policy. On the set $\mathcal{G}$ we have that $\mathcal{I}_k(\pi) = -\infty$ if $\bar{c}_i(\pi,p) > c_i^{\text{ub}} - d_i + \delta_k(\pi)$ for some $i \in [M]$. Also, $\mathcal{I}_k(\pi) \leq \bar{r}(\pi,p) + \delta_k(\pi)$.

The following result allows us to derive bounds on the instantaneous regrets.

Lemma 9: If a stationary policy π satisfies $\bar{c}_i(\pi,p) \leq c_i^{\text{ub}} - d_i,\ \forall i \in [M]$, then on $\mathcal{G}$ its index satisfies $\mathcal{I}_k(\pi) \geq \bar{r}(\pi,p)$. With π set equal to the policy which solves the CMDP $\max_\pi \bar{r}(\pi,p)$ such that $\bar{c}_i(\pi,p) \leq c_i^{\text{ub}} - d_i \quad \forall i \in [M]$, on $\mathcal{G}$ the index of such a policy satisfies $\mathcal{I}_k(\pi) \geq r^\star - z$, where z is as in Theorem 2.

Proof: We note that on the set $\mathcal{G}$, the true MDP p always belongs to $\mathcal{C}_{\tau_k}$. Since $\bar{c}_i(\pi,p) \leq c_i^{\text{ub}} - d_i \quad \forall i \in [M]$ this means that the index of π satisfies

$$\mathcal{I}_k(\pi) = \max_{\theta\in\mathcal{C}_{\tau_k}} \left\{\bar{r}(\pi,\theta) : \bar{c}_i(\pi,\theta) \leq c_i^{\text{ub}} - d_i,\ i \in [M]\right\} \geq \bar{r}(\pi,p).$$

It follows from Lemma 14 that the optimal value of the CMDP $\max_\pi \bar{r}(\pi,p)$, such that $\bar{c}_i(\pi,p) \leq c_i^{\text{ub}} - d_i\ \forall i \in [M]$, is greater than or equal to $r^\star - z$. Hence, it follows from the discussion above that the index of the policy, which is optimal for this CMDP is greater than or equal to $r^\star - z$. ■

As earlier, we bound the regret on the sets $\mathcal{G}, \mathcal{G}_1^c$ and $\mathcal{G}_2^c$ separately. On $\mathcal{G}$, the regret is bounded by the time spent playing suboptimal policies.

[8]To avoid introducing unnecessary notation, we continue to use the same notation for denoting the modified indices as that used to denote the indices of UCRL-CMDP.

Lemma 10: For the modified UCRL-CMDP algorithm, on the set $\mathcal{G}$ the instantaneous reward regret during $\mathcal{E}_k$ can be bounded by $\delta_k(\pi_k) + z$ while the instantaneous cost regret associated with the ith cost can be bounded by $\delta_k(\pi_k) - d_i$.

Proof: Consider a stationary policy π for which $\bar{c}_i(\pi,p) > c_i^{\text{ub}} - d_i + \delta_k(\pi_k)$. It follows from Lemma 8 that the index of this policy satisfies $\mathcal{I}_k(\pi) = -\infty$. However, it is shown in Lemma 9 that there is a policy $\tilde{\pi}$ that has index greater than $r^\star - z$. Since $\mathcal{I}_k(\pi)$ is less than the index of $\tilde{\pi}$, π will not be played by UCRL-CMDP during $\mathcal{E}_k$. Thus, in order π to be a c means that $\bar{c}_i(\pi,p) \leq c_i^{\text{ub}} - d_i + \delta_k(\pi_k)$, which shows that the instantaneous cost regret is bounded by $\delta_k(\pi_k) - d_i$.

In order to bound the instantaneous reward regret, note that it was shown in Lemma 9 that there is a policy with index greater than $r^\star - z$. Hence, the index of π_k is necessarily greater than $r^\star - z$. Since from Lemma 8, we have that the index of π_k is upper-bounded by $\bar{r}(\pi,p) + \delta_k(\pi_k)$, we must have $\bar{r}(\pi,p) + \delta_k(\pi_k) \geq r^\star - z$, or equivalently $r^\star - \bar{r}(\pi,p) \leq \delta_k(\pi_k) + z$. This shows that the instantaneous reward regret is bounded by $\delta_k(\pi_k) + z$. ■

Proof of Theorem 2: Since the proof closely follows that of Theorem 1, we only point out the key differences. The regret decomposition result (23) holds for reward as well cost regrets. Similarly, the regrets on $\mathcal{G}_2^c$ and $\mathcal{G}_1^c$ can be bounded by δT and $\frac{2}{T^{2b-2}|\mathcal{S}||\mathcal{A}|}$, respectively. The only difference from the proof of Theorem 1 arises while bounding the terms $\sum_k \Delta_k^{(R)}$ and $\sum_k \Delta_k^{(i)}$. It follows from Lemma 10 that the bound on $\sum_k \Delta_k^{(R)}$ differs from (24) by an additional term zT, and similarly, the bound on $\sum_k \Delta_k^{(i)}$ differs from the earlier bound by an additional term $\epsilon b_i T$. The proof is then completed by summing the bounds on regrets over the sets $\mathcal{G}, \mathcal{G}_1^c, \mathcal{G}_2^c$. ■

VIII. Achievable Regret Vectors

Let $\boldsymbol{\lambda} \geq \mathbf{0}_M$. Consider the Lagrangian relaxation of (2)–(3)

$$\mathcal{L}(\boldsymbol{\lambda};\pi) := \liminf_{T\to\infty} \frac{\mathbb{E}_\pi \sum_{t=1}^T r(s_t,a_t) + \boldsymbol{\lambda}\cdot\left(\boldsymbol{c}^{\text{ub}} - \boldsymbol{c}(s_t,a_t)\right)}{T}$$

where $\boldsymbol{c}(s_t,a_t)$ is the vector that consists of costs $c_i(s_t,a_t), i \in [M]$. Consider its associated dual function [38], $\mathcal{D}(\boldsymbol{\lambda}) := \max_\pi \mathcal{L}(\boldsymbol{\lambda};\pi)$, and the dual problem

$$\min_{\boldsymbol{\lambda}\geq\mathbf{0}} \mathcal{D}(\boldsymbol{\lambda}). \quad (31)$$

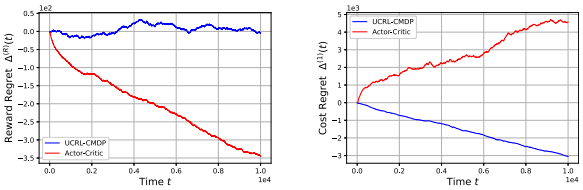
Define the diameter $D(p)$ of MDP p as follows, $D(p) := \max_{s,s'} \min_\pi \mathbb{E}_{\pi,p} T_{s,s'}$. $D(p)$ is finite if p is communicating [2].

Theorem 3: There is a problem instance such that the regrets $\Delta^{(R)}(T), \{\Delta^{(i)}(T)\}_{i=1}^M$ under any learning algorithm ϕ satisfy

$$\mathbb{E}_\phi\Delta^{(R)}(T) + \sum_{i=1}^M \lambda_i^\star \mathbb{E}_\phi\Delta^{(i)}(T) \geq .015\cdot\sqrt{D(p)SAT} \quad (32)$$

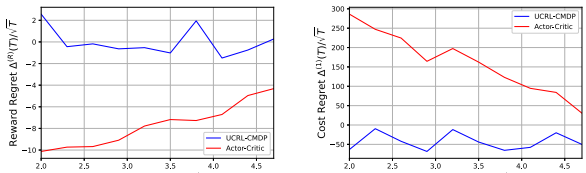
where $\boldsymbol{\lambda}^\star$ is an optimal solution of the dual problem (31), and subscript denotes that expectation is taken with respect to probability measure induced by ϕ.

Proof: We begin by considering an auxiliary reward maximization problem that involves the same MDP p, but in which the reward received at time t by the agent is equal to



(a) (b)

Fig. 1. Plot of the (a) reward regret and (b) cost regret for the network in which the desired delay is $c^{\text{ub}} = 4.5$.



(a) (b)

Fig. 2. Plot of the normalized (a) reward regret and (b) cost regret, as the desired delay c^{ub} is varied.

$r(s_t, a_t) + \boldsymbol{\lambda} \cdot (\boldsymbol{c}^{\text{ub}} - \boldsymbol{c}(s_t, a_t))$ instead of $r(s_t, a_t)$. However, there are no average cost constraints in the auxiliary problem. Let ϕ' be a history-dependent policy for this auxiliary problem. Denote its optimal reward by $r^\star(\boldsymbol{\lambda})$. Then, the regret for cumulative rewards collected by ϕ' in the auxiliary problem is given by $r^\star(\boldsymbol{\lambda})\, T - \mathbb{E}_{\phi'}[\sum_{t=1}^{T} r(s_t, a_t) + \boldsymbol{\lambda} \cdot (\boldsymbol{c}^{\text{ub}} - \boldsymbol{c}(s_t, a_t))]$. It follows from [10], Th. 5] that the controlled transition probabilities $p(s, a, s')$ of the underlying MDP can be chosen so that this regret is greater than $0.015\sqrt{D(p)SAT}$, i.e., $r^\star(\boldsymbol{\lambda})\, T - \mathbb{E}_{\phi'}\left[\sum_{t=1}^{T} r\left(s_t, a_t\right) + \boldsymbol{\lambda} \cdot \left(\boldsymbol{c}\left(s_t, a_t\right) - \boldsymbol{c}^{\text{ub}}\right)\right] \geq 0.015\sqrt{D(p)SAT}$. We observe that any valid learning algorithm for the constrained problem is also a valid algorithm for the auxiliary problem. Thus, if ϕ is a learning algorithm for the problem with average cost constraints, then we have

$$r^\star(\boldsymbol{\lambda})\, T - \mathbb{E}_{\phi}\left[\sum_{t=1}^{T} r(s_t, a_t) + \sum_{i=1}^{M} \lambda_i \left(c_i^{\text{ub}} - c_i\left(s_t, a_t\right)\right)\right] \geq 0.015\sqrt{D(p)SAT}. \tag{33}$$

We now substitute (34) in the above to obtain $\mathbb{E}_\phi\, \Delta^{(R)}(T) + \sum_{i=1}^{M} \lambda_i\, \mathbb{E}_\phi\, \Delta^{(i)}(T) \geq 0.015\sqrt{D(p)SAT} + r^\star T - r^\star(\boldsymbol{\lambda})\, T$. Since the r.h.s. is maximized for values of $\boldsymbol{\lambda}$, which are optimal for the dual problem (31), we set it equal to $\boldsymbol{\lambda}^\star$, and then use Lemma 11 in order to obtain $\mathbb{E}_\phi\, \Delta^{(R)}(T) + \sum_{i=1}^{M} \lambda_i \mathbb{E}_\phi\, \Delta^{(i)}(T) \geq 0.015\sqrt{D(p)SAT}$. ■

IX. Simulation Results

We compare the performance of the proposed UCRL-CMDP algorithm with the actor-critic algorithm for CMDPs that was proposed in [16]. Actor-critic algorithms are a popular class of online learning algorithms [39], [40], [41] that are based on multitime-scale stochastic approximation [42]. Note that since the algorithms proposed in [24] and [25] are for an episodic RL setup, we do not compare the performance of UCRL-CMDP with algorithms proposed therein.

Experiment Setup: We consider the single-hop wireless network that was discussed in Section I. For simplicity, we let the action set $\mathcal{A}$ be binary and take the channel state to be static. Thus, $a_t = 0$ means no packet was transmitted while $a_t = 1$ means one packet is attempted. Let $A_t \in \{0, 1, 2, 3\}$ be the number of packet arrivals, which are assumed i.i.d. across times with mass function (0.65, 0.2, 0.1, 0.05) for the experiments shown in Figs. 1 and 2. Queue length evolves as $Q_{t+1} = (Q_t + A_t - D_t)^+ \wedge B$, $t = 0, 1, 2, \ldots$, where B is the capacity of the buffer[9] while D_t is the number that is delivered to destination at time t. In our experiments, we use $B = 6$ and take the channel reliability as 0.9.

Actor-Critic Algorithm for CMDPs: Let $a(n) = 1/n$, $b(n) = 1/(n \log n)$ and $c(n) = 1/(n \log^2 n)$. Let $\mathcal{Q} := \{\boldsymbol{x} \in \mathbb{R}^{|\mathcal{A}|-1} : x_i \geq 0\, \forall i, \sum_{j=1}^{|\mathcal{A}|-1} x_j \leq 1\}$ denote the simplex of subprobability vectors. Let $\Gamma(\cdot)$ denote the map that projects a vector onto $\mathcal{Q}$. It begins by replacing the original constrained MDP by an unconstrained one by imposing a penalty upon constraint violation. The instantaneous reward for this modified MDP is equal to $r(s_t, a_t) - \tilde{\lambda}_t(c(s_t, a_t) - c^{\text{ub}})$ where $\tilde{\lambda}_t \geq 0$ is the price associated with the constraint violation. In order to solve this unconstrained MDP, the algorithm keeps an estimate of the value function $V_t : \mathcal{S} \mapsto \mathbb{R}$, which is updated as $V_{t+1}(s) = V_t(s) + a(N_t(s))\mathbb{1}\{s_t = s\}(r(s, u_t) + \tilde{\lambda}_t c(s, u_t) - V_t(s) - V_t(s^\star) + V_t(s_{t+1}))$, where $s^\star$ is a designated state. Let $\pi_t(a|s)$ denote the probability with which action a is implemented in state s at time t. Let $a^\star$ be a designated action. These probabilities are generated as follows. The algorithm maintains vectors $\hat{\pi}_t(s) = \{\hat{\pi}_t(a|s) : a \in \mathcal{A}\}$, $s \in \mathcal{S}$, and updates it as $\hat{\pi}_{t+1}(s) = \Gamma(\hat{\pi}_t(s) + \star), t = 1, 2, \ldots,$ where, $\star = \sum_{a \neq a^\star} b(N_t(s, a)) \times \mathbb{1}\{s_t = s, a_t = a\}\hat{\pi}_t(s, a) \times [V_t(s) + V_t(s^\star) - r(s, a) + \tilde{\lambda}_t c(s, a) - V_t(s_{t+1})]e_j$, where e_a is the unit vector with a 1 in the place corresponding to action a.[10] The probability for action $a^\star$ is computed as $\hat{\pi}_t(a^\star|s) = 1 - \sum_{a \neq a^\star} \hat{\pi}_t(a|s)$. The action probabilities π_t are then generated from $\hat{\pi}_t$ as $\pi_t(a|s) = (1 - \epsilon_t)\hat{\pi}_t(a|s) + \frac{\epsilon_t}{|\mathcal{A}|}$, $a \in \mathcal{A}$, where $\epsilon_t \to 0$. Finally, the price $\tilde{\lambda}_t$ is updated as $\tilde{\lambda}_{t+1} = [\tilde{\lambda}_t + \gamma_t(c(s_t, a_t) - c^{\text{ub}})]^+$, where c^{ub} is the threshold on average queue length. In our experiments, we use $s^\star = B$, $a^\star = 0$ and $\epsilon_t = 1/t$.

Results: Fig. 1 compares the cumulative regrets incurred by these algorithms. We observe that the reward regret as well as the cost regret of UCRL-CMDP are low. We observe the following drawback of the actor-critic algorithm's performance that the cost regret is prohibitively high. We then vary the budget c^{ub} on the average queue length. These results are shown in Fig. 2. Once again, we make a similar observation that UCRL-CMDP is effective in balancing both, the reward regret $\Delta^{(R)}(t)$ and the cost regret $\Delta^{(1)}(t)$, whereas the actor-critic algorithm yields a high cost regret. In both of these experiments the probability vector of arrivals was held fixed at (0.65, 0.2, 0.1, 0.05). We vary this probability vector, and plot the regrets in Fig. 3. Once again,

[9]For $x \in \mathbb{R}$, we let $(x)^+ := \max\{x, 0\}$, $x \wedge B := \min\{x, B\}$

[10]We enumerate the available actions as $1, 2, \ldots, |\mathcal{A}|$.

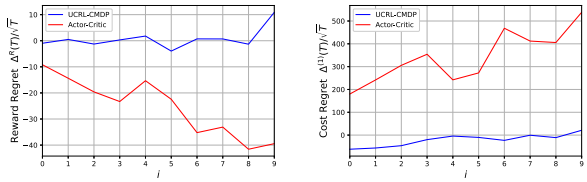


Fig. 3. Plot of the (a) reward regret and (b) cost regret, as the probability distribution of the arrivals is varied. The probability vector of A_t is equal to $(0.65-0.02i, 0.2, 0.1+0.01i, 0.05+0.01i)$, where the parameter i is varied from 0 to 9. The desired delay c^{ub} is held fixed at 4.5, and channel reliability at 0.9.

UCRL-CMDP outperforms the actor-critic algorithm. Though the reward regret of actor-critic algorithm is lower than that of the UCRL-CMDP algorithms, this occurs at the expense of an undesirable much larger cost regret. In contrast, the reward regret as well as the cost regret of UCRL-CMDP is low. All plots are obtained after averaging over 100 runs.

X. Conclusion

In this work, we initiate a study to develop learning algorithms that simultaneously control all the components of the regret vector while controlling unknown MDPs. We devised algorithms that are able to tune different components of the cost regret vector and also obtained a nonachievability result that characterizes those regret vectors that cannot be achieved under any learning rule. In our work, we assume that the underlying MDP is unichain. An interesting research problem is to characterize the set of achievable regret vectors under the weaker assumption that the underlying MDP is communicating.

Appendix A
Results Used in the Proof of Theorem 3

Lemma 11: Consider the dual problem (31) associated with the CMDP (2)–(3), and let $\boldsymbol{\lambda}^\star$ be a solution of the dual problem. If Assumption 2 holds true, then $\mathcal{D}(\boldsymbol{\lambda}^\star) = r^\star$.

Proof: Under Assumption 2, the CMDP (2)–(3) is strictly feasible, so that Slater's condition [43] is satisfied, and consequently strong duality holds true. Thus, if $\boldsymbol{\lambda}^\star$ solves the dual problem (31), we then have that $\mathcal{D}(\boldsymbol{\lambda}^\star) = r^\star$. ■

Lemma 12: Let $\boldsymbol{\lambda} \geq \mathbf{0}_M$ and ϕ be a learning algorithm for the problem of maximizing cumulative rewards under average cost constraints. We then have the following:

$$\mathbb{E}_\phi \sum_{t=1}^{T} (r(s_t,a_t) + \sum_{i=1}^{M} \lambda_i (c_i^{\text{ub}} - c_i(s_t,a_t))) = r^\star T - \mathbb{E}_\phi \Delta^{(R)}(T) - \sum_{i=1}^{M} \lambda_i \mathbb{E}_\phi \Delta^{(i)}(T). \quad (34)$$

Proof: We have, $\mathbb{E}_\phi \sum_{t=1}^{T}(r(s_t,a_t) + \sum_{i=1}^{M} \lambda_i (c_i^{\text{ub}} - c_i(s_t,a_t))) = \mathbb{E}_\phi \sum_{t=1}^{T} r(s_t,a_t) + \sum_{i=1}^{M} \lambda_i \mathbb{E}_\phi \sum_{t=1}^{T} (c_i^{\text{ub}} - c_i(s_t,a_t)) = r^\star T - \mathbb{E}_\phi \Delta^{(R)}(T) - \sum_{i=1}^{M} \lambda_i \mathbb{E}_\phi \Delta^{(i)}(T)$. ■

Appendix B
Perturbation Analysis of CMDPs

We derive some results on the variations in the value of optimal reward of the CMDP (2)–(3) as a function of the cost budgets $\boldsymbol{c}^{\text{ub}}$. Consider a vector $\hat{\boldsymbol{c}}^{\text{ub}}$ of cost budgets that satisfies

$$c_i^{\text{ub}} - \epsilon \leq \hat{c}_i^{\text{ub}} \leq c_i^{\text{ub}} \quad \forall i \in [M] \quad (35)$$

where $\epsilon > 0$. Now consider the following CMDP in which the upper-bounds on the average costs are equal to $\{\hat{c}_i^{\text{ub}}\}_{i=1}^M$

$$\max_\pi \liminf_{T\to\infty} \frac{1}{T} \mathbb{E}_\pi \sum_{t=1}^{T} r(s_t,a_t) \quad (36)$$

$$\text{s.t. } \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}_\pi \sum_{t=1}^{T} c_i(s_t,a_t) \leq \hat{c}_i^{\text{ub}}, i \in [1,M]. \quad (37)$$

Lemma 13: Let the MDP p satisfy Assumption 1 and Assumption 2. Let $\boldsymbol{\lambda}^\star$ be an optimal dual variable/Lagrange multiplier associated with the CMDP (36)–(37). Then, $\boldsymbol{\lambda}^\star$ satisfies $\sum_{i=1}^{M} \lambda_i^\star \leq \frac{\hat{\eta}}{\eta}$, where the constant η is as in (27) while $\hat{\eta}$ is as in (29).

Proof: Within this proof, we let $\pi^\star(\hat{c}^{\text{ub}})$ denote an optimal stationary policy for (36)–(37). Recall that the policy $\pi_{\text{feas.}}$ that was defined in Assumption 2 satisfies $\bar{c}_i(\pi_{\text{feas.}}) \leq c_i^{\text{ub}} - \eta$. We have

$$\begin{aligned}
&\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} r(s,a) \geq \bar{r}(\pi^\star(\hat{c}^{\text{ub}})) \\
&= \bar{r}(\pi^\star(\hat{c}^{\text{ub}})) + \sum_{i=1}^{M} \lambda_i^\star \left(\hat{c}_i^{\text{ub}} - \bar{c}_i(\pi^\star(\hat{c}^{\text{ub}}))\right) \\
&\geq \bar{r}(\pi_{\text{feas.}}) + \sum_{i=1}^{M} \lambda_i^\star \left(\hat{c}^{\text{ub}} - \bar{c}(\pi_{\text{feas.}})\right) \\
&\geq \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} r(s,a) + \sum_{i=1}^{M} \lambda_i^\star \left(\hat{c}^{\text{ub}} - \bar{c}(\pi_{\text{feas.}})\right) \\
&\geq \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} r(s,a) + \eta \sum_{i=1}^{M} \lambda_i^\star
\end{aligned}$$

where the second inequality follows since a policy that is optimal for the problem (36)–(37) maximizes the Lagrangian $\bar{r}(\pi) + \sum_{i=1}^{M} \lambda_i (\hat{c}_i^{\text{ub}} - \bar{c}_i(\pi))$ when the Lagrange multiplier $\boldsymbol{\lambda}$ is set equal to $\boldsymbol{\lambda}^\star$ [38]. Rearranging the above inequality yields the desired result. ■

Lemma 14: Let the MDP p satisfy Assumption 1 and Assumption 2. If $r^\star(\hat{\boldsymbol{c}}^{\text{ub}})$ denotes optimal reward value of (36)–(37), and $r^\star$ is optimal reward of problem (2)–(3), then we have that $r^\star - r^\star(\hat{\boldsymbol{c}}^{\text{ub}}) \leq (\max_{i\in[1,M]}\{c_i^{\text{ub}} - \hat{c}_i^{\text{ub}}\})\frac{\hat{\eta}}{\eta}$, where $\hat{\eta}$ is as in (29), η is as in (27), and $\hat{\boldsymbol{c}}$ satisfies (35).

Proof: As discussed in Section III-B, a CMDP can be posed as an LP. Since under Assumption 2, both the CMDPs (2)–(3) and (36)–(37) are strictly feasible, we can use the strong duality property of LPs [38] in order to conclude that the optimal value of the primal and the dual problems for both the CMDPs are

equal. Thus,

$$r^\star = \sup_{\pi} \inf_{\lambda} \ \bar{r}(\pi) + \sum_{i=1}^{M} \lambda_i \left(c_i^{\text{ub}} - \bar{c}_i(\pi)\right) \tag{38}$$

$$r^\star(\hat{\boldsymbol{c}}^{\text{ub}}) = \sup_{\pi} \inf_{\lambda} \ \bar{r}(\pi) + \sum_{i=1}^{M} \lambda_i \left(\hat{c}_i^{\text{ub}} - \bar{c}_i(\pi)\right). \tag{39}$$

Let $\pi^{(1)}, \pi^{(2)}$ and $\lambda^{(1)}, \lambda^{(2)}$ denote optimal policies and vector consisting of optimal dual variables for the two CMDPs. It then follows from (38) and (39) that, $r^\star \leq \bar{r}\left(\pi^{(1)}\right) + \sum_{i=1}^{M} \lambda_i^{(2)} \left(c_i^{\text{ub}} - \bar{c}_i\left(\pi^{(1)}\right)\right)$, and $r^\star\left(\hat{\boldsymbol{c}}^{\text{ub}}\right) \geq \bar{r}\left(\pi^{(1)}\right) + \sum_{i=1}^{M} \lambda_i^{(2)} (\hat{c}_i^{\text{ub}} - \bar{c}_i\left(\pi^{(1)}\right))$. Subtracting the second inequality from the first yields $r^\star - r^\star\left(\boldsymbol{c}^{\text{ub}}\right) \leq \sum_{i=1}^{M} \lambda_i^{(2)} \left(c_i^{\text{ub}} - \hat{c}_i^{\text{ub}}\right) \leq (\max_{i\in[1,M]} \left\{c_i^{\text{ub}} - \hat{c}_i^{\text{ub}}\right\}) \left(\sum_{i=1}^{M} \lambda_i^{(2)}\right) \leq \left(\max_{i\in[1,M]} \left\{c_i^{\text{ub}} - \hat{c}_i^{\text{ub}}\right\}\right) \frac{\hat{\eta}}{\eta}$, where the last inequality follows from Lemma 13. ■

REFERENCES

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning - An Introduction* (Adaptive Computation and Machine Learning). Cambridge, MA, USA: MIT Press, 1998. [Online]. Available: http://www.worldcat.org/oclc/37293240

[2] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ, USA: Wiley, 2014.

[3] L. I. Sennott, *Stochastic Dynamic Programming and the Control of Queueing Systems*, vol. 504. Hoboken, NJ, USA: Wiley, 2009.

[4] A. Lazar, "Optimal flow control of a class of queueing networks in equilibrium," *IEEE Trans. Autom. Control*, vol. AC-28, no. 11, pp. 1001–1007, Nov. 1983.

[5] M.-T. T. Hsiao and A. A. Lazar, "Optimal decentralized flow control of Markovian queueing networks with multiple controllers," *Perform. Eval.*, vol. 13, no. 3, pp. 181–204, 1991.

[6] P. Nain and K. Ross, "Optimal priority assignment with hard constraint," *IEEE Trans. Autom. Control*, vol. AC-31, no. 10, pp. 883–888, Oct. 1986.

[7] R. Singh and P. R. Kumar, "Throughput optimal decentralized scheduling of multihop networks with end-to-end deadline constraints: Unreliable links," *IEEE Trans. Autom. Control*, vol. 64, no. 1, pp. 127–142, Jan. 2019.

[8] R. Singh and P. R. Kumar, "Adaptive CSMA for decentralized scheduling of multi-hop networks with end-to-end deadline constraints," *IEEE/ACM Trans. Netw.*, vol. 29, no. 3, pp. 1224–1237, Jun. 2021.

[9] E. A. Feinberg and M. I. Reiman, "Optimality of randomized trunk reservation," *Probability Eng. Informational Sci.*, vol. 8, no. 4, pp. 463–489, 1994.

[10] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," *J. Mach. Learn. Res.*, vol. 11, no. Apr, pp. 1563–1600, 2010.

[11] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2020.

[12] A. Mete, R. Singh, X. Liu, and P. R. Kumar, "Reward biased maximum likelihood estimation for reinforcement learning," in *Proc. 3rd Conf. Learn. Dyn. Control*, 2021, pp. 815–827.

[13] P. R. Kumar and A. Becker, "A new family of optimal adaptive controllers for Markov chains," *IEEE Trans. Autom. Control*, vol. AC-27, no. 1, pp. 137–146, Feb. 1982.

[14] D. R. I. Osband and B. V. Roy, "(More) Efficient reinforcement learning via posterior sampling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3003–3011.

[15] E. Altman and A. Schwartz, "Adaptive control of constrained Markov chains," *IEEE Trans. Autom. Control*, vol. 36, no. 4, pp. 454–462, Apr. 1991.

[16] V. S. Borkar, "An actor-critic algorithm for constrained Markov decision processes," *Syst. Control Lett.*, vol. 54, no. 3, pp. 207–213, 2005.

[17] V. S. Borkar, "Stochastic approximation with two time scales," *Syst. Control Lett.*, vol. 29, no. 5, pp. 291–294, 1997.

[18] V. S. Borkar and S. Pattathil, "Concentration bounds for two time scale stochastic approximation," in *Proc. 56th Annu. Allerton Conf. Commun., Control, Comput.*, 2018, pp. 504–511.

[19] V. S. Borkar, "A concentration bound for contractive stochastic approximation," *Syst. Control Lett.*, vol. 153, 2021, Art. no. 104947.

[20] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 22–31.

[21] Y. Liu, J. Ding, and X. Liu, "IPO: Interior-point policy optimization under constraints," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 4940–4947.

[22] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward constrained policy optimization," in *Proc. 7th Int. Conf. Learn. Representations*, 2019.

[23] E. Uchibe and K. Doya, "Constrained reinforcement learning from intrinsic and extrinsic rewards," in *Proc. IEEE 6th Int. Conf. Develop. Learn.*, 2007, pp. 163–168.

[24] S. Qiu, X. Wei, Z. Yang, J. Ye, and Z. Wang, "Upper confidence primal-dual reinforcement learning for CMDP with adversarial loss," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 15277–15287, 2020.

[25] Y. Efroni, S. Mannor, and M. Pirotta, "Exploration-exploitation in constrained MDPs," 2020, *arXiv:2003.02189*.

[26] L. Chen, R. Jain, and H. Luo, "Learning infinite-horizon average-reward Markov decision processes with constraints," in *Proc. 39th Int. Conf. Mach. Learn.*, 2022, pp. 3204–3245.

[27] H. Wei, X. Liu, and L. Ying, "A provably-efficient model-free algorithm for infinite-horizon average-reward constrained Markov decision processes," in *Proc. 36th AAAI Conf. Artif. Intell.*, 2022.

[28] M. Agarwal, Q. Bai, and V. Aggarwal, "Markov decision processes with long-term average constraints," 2021, *arXiv:2106.06680*.

[29] T. Liu, R. Zhou, D. Kalathil, P. R. Kumar, and C. Tian, "Learning policies with zero or bounded constraint violation for constrained MDPs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 17183–17193.

[30] D. Ding, K. Zhang, T. Basar, and M. Jovanovic, "Natural policy gradient primal-dual method for constrained Markov decision processes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 8378–8390.

[31] Y. Liu, A. Halev, and X. Liu, "Policy learning with constraints in model-free reinforcement learning: A survey," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, 2021, pp. 4508–4515.

[32] Y. Liu, J. Ding, Z.-L. Zhang, and X. Liu, "Clara: A constrained reinforcement learning based resource allocation framework for network slicing," in *Proc. IEEE Int. Conf. Big Data*, 2021, pp. 1427–1437.

[33] C. Villani, *Optimal Transport: Old and New*, vol. 338. Berlin, Germany: Springer, 2008.

[34] S. Resnick, *A Probability Path*. Berlin, Germany: Springer, 2019.

[35] E. Altman, *Constrained Markov Decision Processes*. London, U.K.: Chapman and Hall, Mar. 1999.

[36] K. Azuma, "Weighted sums of certain dependent random variables," *Tohoku Math. J., 2nd Ser.*, vol. 19, no. 3, pp. 357–367, 1967.

[37] A. Y. Mitrophanov, "Sensitivity and convergence of uniformly ergodic Markov chains," *J. Appl. Probability*, vol. 42, no. 4, pp. 1003–1014, 2005.

[38] D. P. Bertsekas, *Nonlinear Programming*, vol. 48, no. 3. New York, NY, USA: Taylor & Francis, 1997.

[39] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 1008–1014.

[40] J. Peters and S. Schaal, "Natural actor-critic," *Neurocomputing*, vol. 71, no. 7–9, pp. 1180–1190, 2008.

[41] V. R. Konda and V. S. Borkar, "Actor-critic–type learning algorithms for Markov decision processes," *SIAM J. Control Optim.*, vol. 38, no. 1, pp. 94–123, 1999.

[42] V. S. Borkar, *Stochastic Approximation: A. Dynamical Systems Viewpoint*, vol. 48. Berlin, Germany: Springer, 2009.

[43] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

Rahul Singh (Member, IEEE) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology Kanpur, Kanpur, India, in 2009, the M.S. degree in electrical engineering from the University of Notre Dame, South Bend, IN, USA, in 2011, and the Ph.D. degree in electrical and computer engineering from Texas A&M University, College Station, TX, USA, in 2015.

He is currently an Assistant Professor with ECE Department, Indian Institute of Science (IISc), Bangalore, India. His research interests include stochastic control, machine learning, and networks.

Dr. Singh was a runner-up for the Best Paper Award of ACM MobiHoc 2020 for his article.

Abhishek Gupta received the B.Tech. degree in aerospace engineering from IIT Bombay, Mumbai, India, in 2009, and the M.S. and Ph.D. degrees in aerospace engineering from the University of Illinois at Urbana-Champaign (UIUC), Champaign, IL, USA, in 2014.

He is currently an Assistant Professor with ECE Department, The Ohio State University, Columbus, OH, USA. His research interests include stochastic control theory, probability theory, and game theory with applications to transportation markets, electricity markets, and cybersecurity of control systems.

Ness B. Shroff (Fellow, IEEE) received the Ph.D. degree from Columbia University, New York, NY, USA, in 1994.

He is currently the Institute Director of the NSF AI Institute for designing future edge networks and distributed intelligence (AI-EDGE).

Dr. Shroff holds the Ohio Eminent Scholar Chaired Professorship of Networking and Communications at The Ohio State University, Columbus, OH, USA. He is a recipient of numerous best paper awards for his research and has been noted as a Highly Cited Researcher by Thomson Reuters in 2014 and 2015. He was also a recipient of the IEEE INFOCOM Achievement Award in 2014. He is currently the Steering Committee Chair for ACM Mobihoc and the Editor in Chief of the IEEE/ACM TRANSACTIONS ON NETWORKING.