Analysis of Learning in Constrained MDPs

Aayush Rajesh
200070001

1 Introduction

A traditional setting of a problem in Reinforcement Learning involves interacting with an environment. At every time instant t, the agent takes an action a_t based on the current state s_t, earns a reward $r(s_t, a_t)$, and finds themselves in a new state s_{t+1}. The environment is modeled as a Markov Decision Process (MDP), however, since the nature of the underlying environment is unknown to the agent, so are the transition probabilities $p(s_t, a_t, s_{t+1})$ for moving from a particular state to the next based on the action taken. The goal of the agent is to maximize the total reward earnings by controlling the actions taken at each state.

We can make a slight modification to this problem setup, by having the agent incur a cost for every action taken. In this case, the agent tries to maximize the total reward while attempting to limit the total cost expenditure below a certain system-defined threshold. This corresponds to the environment being modeled by a Constrained MDP.

This report primarily summarizes the work done in [1], which provides an algorithm for learning in a specific setup of constrained MDPs and proves the optimality of the regret incurred by the algorithm. Also detailed in the report are some previous algorithms and results, which are used in obtaining some results about the aforementioned algorithm.

2 Previous Work

2.1 UCRL2 Algorithm

The UCRL2 algorithm, introduced in [2], is described as an algorithm that implements the notion of "optimism in the face of uncertainty" (OFU). The algorithm maintains a set of statistically likely true descriptions of the underlying MDP and makes an optimistic assumption by treating one of these MDP descriptions as the true nature of the environment. The algorithm then implements a control policy that is optimal for the chosen MDP description.

The UCRL2 algorithm makes use of upper confidence bounds in defining the set of plausible MDPs. Only those MDPs with rewards and transition probabilities within a certain bound from the empirical estimates of these values occupy the set.

The algorithm proceeds as follows. In each episode k beginning at time instant τ_k, we initialize the number of visits in the episode $n_k(s,a) \coloneqq 0$ for all state-action pairs $(s,a) \in \mathcal{S} \times \mathcal{A}$. Let $N_k(s,a)$ denote the number of visits before the episode k. Then, we can obtain the empirical reward estimates and transition probabilities as

$$\hat{r}_k(s,a) = \frac{1}{\max\{1, N_k(s,a)\}} \cdot \sum_{t=1}^{\tau_k - 1} r(s,a) \mathbb{1}_{s_t = s,\, a_t = a}$$

$$\hat{p}_k(s'|a,s) = \frac{1}{\max\{1, N_k(s,a)\}} \cdot \sum_{t=1}^{\tau_k - 1} \mathbb{1}_{s_t = s,\, a_t = a,\, s_{t+1} = s'}$$

The set $\mathcal{M}_k$ of plausible MDPs are those with transition probabilities $\tilde{p}(\cdot|s,a)$ and rewards $\tilde{r}(s,a)$ close to the empirical estimates. That is, they satisfy

$$|\tilde{r}(s,a) - \hat{r}_k(s,a)| \leq \epsilon_k(s,a)$$
$$\|\tilde{p}(\cdot|s,a) - \hat{p}_k(\cdot|a,s)\|_1 \leq \epsilon'_k(s,a),$$

where $\epsilon_k(s,a)$ and $\epsilon'_k(s,a)$ are values specified by parameters. From this set, we can find an "optimistic" MDP $\tilde{M}_k$ and a corresponding near-optimal policy $\tilde{\pi}_k$. This can be done using an algorithm such as extended value iteration (see Appendix).

We now execute the policy $\tilde{\pi}_k$ during the episode, while regularly updating $n_k(s,a)$. The episode ends when we encounter a state for which we perform the action specified by the policy as many times during the episode as we had done before the episode, i.e. when $\exists s \in \mathcal{S}$ such that $n_k(s, \tilde{\pi}_k(s)) = N_k(s, \tilde{\pi}_k(s))$.

It is important to note why the UCRL2 algorithm fails in the case if Constrained MDP problems. Due to the fact that the goal of the policy in Constrained MDPs is to both maximize rewards while satisfying cost constraints, an algorithm utilizing OFU need not satisfy the cost constraints posed in the problem. The authors highlight this fact using a simple example which is depicted in Figure 1.

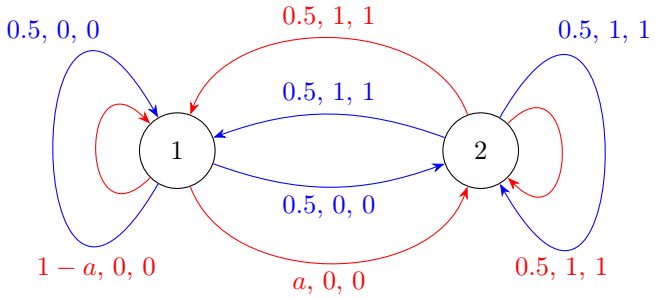

Figure 1: Two-state two-action MDP with action 0 in blue and action 1 in red.

Here, different actions are shown in two different colors. Labeling every transition is the transition probability, reward, and cost respectively. Suppose the unknown variable $a > 0.5$, and we have an average cost constraint of $C < 2a/(1+2a)$. Since only state 2 yields rewards, an optimistic policy may involve taking action 1 at state 1 (since this action transitions to state 2 with higher probability), and any action at state 2 (since both are identical for this state). Under this policy, the stationary distribution comes out to be occupying state 1 with probability $1/(1+2a)$, and occupying state 2 with probability $2a/(1+2a)$. However, this means that our average cost expenditure (which is the same as the probability of state 2) is $2a/(1+2a) > C$. Therefore, following OFU need not satisfy the constraints of the problem.

3 Preliminaries

3.1 Problem Setup and Notation

The problem being considered involves the agent earning a reward and incurring M costs as a result of the action taken at time instant t. The reward and cost functions are denoted by $r(\cdot|\cdot)$ and $\{c_i(\cdot|\cdot)\}_{i=1}^{M}$ respectively, and are mappings from $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Since we wish to impose constraints on our M costs, let c_i^{ub} denote the upper bound on the average value of the i^{th} cost expenditure.

Therefore, the Controlled Markov Process can be completely specified by the tuple $\mathcal{CMP} = (\mathcal{S}, \mathcal{A}, p, r, c_1, c_2, \ldots, c_M)$. In the learning setup, the agent is un-

aware of the transition probabilities p, but the reward r and cost c are known.

It is important to take note of how the problem takes shape when the transition probabilities p are known. In that case, a constrained MDP poses the following optimization problem:

$$\max_{\pi} \liminf_{T\to\infty} \frac{1}{T} \sum_{i=1}^{T} r(s_t, a_t) \tag{1}$$

$$\text{s.t.} \limsup_{T\to\infty} \sum_{i=1}^{T} c_i(s_t, a_t) \leq c_i^{ub} \quad \forall i \in [M]. \tag{2}$$

Note that we use lim sup and lim inf in case the limit does not exist. This optimization problem can be solved using the linear program:

$$\max_{\mu} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mu(s,a) r(s,a), \tag{3}$$

$$\text{s.t.} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mu(s,a) c_i(s,a) \leq c_i^{ub} \quad \forall i \in [M], \tag{4}$$

$$\sum_{a\in\mathcal{A}} \mu(s,a) = \sum_{(s',a)\in\mathcal{S}\times\mathcal{A}} \mu(s',a) p(s',a,s) \quad \forall s \in \mathcal{S}, \tag{5}$$

$$\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mu(s,a) = 1, \quad \mu(s,a) \geq 0 \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A}. \tag{6}$$

If μ^* satisfies the linear program, then the policy $SR(\mu^*)$ -which picks action a at state s with probability $\mu^*(s,a)/\sum_{a'} \mu^*(s,a')$, and follows fixed rule when the quantity is undefined - solves the constrained MDP constraints (1)-(2).

3.2 Definitions

Definition 1 (Control Policy)**.** Let the $|\mathcal{A}|$-simplex $\Delta(\mathcal{A})$ be defined as

$$\Delta(\mathcal{A}) \coloneqq \left\{ \mathbf{x} \in \mathbb{R}^{|\mathcal{A}|} : \sum_{i=1}^{|\mathcal{A}|} x_i = 1, x_i \geq 0 \right\}.$$

Then, a stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ chooses the action a_t in a state s_t according to the distribution $p(\cdot|s_t)$.

Definition 2 (Unichain MDP)**.** The MDP p is said to be unichain if under any stationary distribution π the policy-induced Markov chain has a single recurrent class. If the MDP is unichain, then under any stationary distribution π the induced Markov chain satisfies

$$\|P_{\pi,p,s}^{(t)} - P_{\pi,p}\|_V \leq C\rho^t \quad \forall s \in \mathcal{S},$$

with $C > 0$, $\rho \in [0, 1]$ being constants and the norm being the total variational distance (maximum difference between probabilities that two distributions assign to the same event). Here, $P^{(t)}_{\pi,p,s}$ is the t-step probability distribution obtained by following policy π in the MDP p with initial state s, and $P_{\pi,p}$ is the stationary distribution under the policy π.

Definition 3 (Occupation Measure)**.** For a controlled Markov process under policy π the occupation measure $\mu_\pi = \{\mu_\pi(s, a) : (s, a) \in \mathcal{S} \times \mathcal{A}\}$ describes the average amount of time each state-action pair occurs during the evolution of the process. More formally,

$$\mu_\pi(s, a) \coloneqq \lim_{T\to\infty} \frac{1}{T} \mathbb{E}_\pi \left[\sum_{t=1}^{T} \mathbb{1}\{s_t = a, a_t = a\} \right].$$

Definition 4 (Regret)**.** The performance of the learning algorithm is defined by the reward and cost regrets. If r^* is the optimal average reward for the constrained MDP (1)-(2), the cumulative reward and cost regrets until time T are

$$\Delta^{(R)}(T) \coloneqq r^* T - \sum_{t=1}^{T} r(s_t, a_t),$$

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

Definition 5 (Diameter)**.** The diameter $D(p)$ of an MDP p is the time it takes to move from a state s to another state s' by following an appropriate policy. Alternatively, it is defined as

$$D(p) \coloneqq \max_{s,s'} \min_{\pi} \mathbb{E}_{\pi,p} T_{s,s'}.$$

The diameter of an MDP p with state space $\mathcal{S}$ and action space $\mathcal{A}$ satisfied $D(p) \geq \log_{|\mathcal{A}|} |\mathcal{S}|$. In fact, in [2], it is shown that the regret incurred by the UCRL2 algorithm is dependent on the diameter of the MDP as $\tilde{\mathcal{O}}(D|\mathcal{S}|\sqrt{|\mathcal{A}|T})$. This is due to the fact that the analysis in [2] considers *communicating* MDPs, i.e. those with a finite diameter D.

4 UCRL-CMDP Algorithm

We have seen how the UCRL2 algorithm, which implements OFU, fails to learn an optimal policy in the constrained MDP case. Now, we can describe the UCRL-CMDP algorithm, which is based on the UCRL2 algorithm, but has minor differences to make it compatible with the extra cost constraints being imposed.

Maintaining $\hat{p}_k(s, a, s')$ as the empirical estimate upto episode k for the transition probability, which is set to uniform distribution if the state-action pair

has never been visited prior to the k^{th} episode, i.e. when $N_k(s,a) = 0$. Much like the UCRL2 algorithm, we maintain a "confidence interval" associated with this estimate $\hat{p}_t$, which is the set of plausible MDPs with transition probabilities close to the empirical estimate. More formally, the confidence interval is defined as:

$$\mathcal{C}_k \coloneqq \{p' : \sum_{s' \in \mathcal{S}} p'(s,a,s') = 1,\, p'(s,a,s') \geq 0,\\ |p'(s,a,s') - \hat{p}_k(s,a,s')| \leq \epsilon_k(s,a),\, \forall (s,a) \in \mathcal{S} \times \mathcal{A}\},$$

where the interval size is explicitly defined using an agent specified constant $b > 1$ as

$$\epsilon_k(s,a) \coloneqq \sqrt{\frac{2\log(T^b|\mathcal{S}||\mathcal{A}|)}{\min(N_k(s,a),1)}}.$$

After initializing the required counts and obtaining the confidence interval at the beginning of episode k at time $t = \tau_k$, the agent solves the following constrained optimization problem described by Equations (3)-(6), with the added maximization over MDPs p' satisfying $p' \in \mathcal{C}_{\tau_k}$.

Maximizing with respect to p' expresses the optimism of the agent, on top of choosing the optimal policy which is conveyed through the maximization over μ. In case the problem is feasible and has a solution given by $(\tilde{\mu}_k, \tilde{p}_k)$, then the algorithm chooses actions within the k^{th} episode according to $SR(\tilde{\mu}_k)$. If not, then the algorithm executes a pre-determined policy for the duration of the episode.

4.1 Analysis of UCRL-CMDP

Finding bounds on the cumulative reward and cost regrets is crucial for analyzing the performance of a learning algorithm. While it is expected that the regrets scale up with time, we would prefer a sub-linear growth in the regret to ensure better scalability of the algorithm.

For the analysis of the algorithm, we shall assume that the underlying MDP p is unichain, and the rewards, costs, and upper bounds on costs all lie in the range $[0,1]$. Furthermore, we shall assume that the constrained MDP optimization problem posed in (1)-(2) is feasible.

Theorem 1. For the UCRL-CMDP algorithm applied to an MDP satisfying the above assumption, the cumulative reward regret $\Delta^{(R)}(T)$ and cumulative cost regrets $\Delta^{(i)}(T) \quad \forall i \in [M]$ are bounded by $\tilde{\mathcal{O}}(T^{2/3})$.

Clearly it is evident that the UCRL-CMDP algorithm achieves sub-linear regret. It does not, however, perform better than the UCRL2 algorithm, which can be

shown to have regret bounds of $\tilde{\mathcal{O}}(\sqrt{T})$. The difference arises from the fact that we are required to solve a multi-objective optimization problem.

We can go about showing the exact bounds on the performance of UCRL-CMDP by analyzing the behavior on a "good set" and bounding the number of times a sub-optimal policy is followed for an episode. For the following results, we take $\delta = T^{-1/3}$ and $\alpha = 1/3$.

Lemma 1. Let $\mathcal{G}_1$ be the event that the MDP p lies in the optimistic set for each episode, i.e,

$$\mathcal{G}_1 \coloneqq \{p \in \mathcal{C}_{\tau_k},\, \forall k \in [M]\}.$$

Then,

$$\mathbb{P}\{\mathcal{G}_1\} \geq 1 - \frac{1}{T^{\alpha+b-2}}$$

Lemma 2. Define the event $\mathcal{G}_2$ as

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

where K is the total number of episodes, $\mathcal{F}_{\tau_k}$ is the history of the MDP up to the episode, and β satisfies $2\beta - \alpha = 1$. Then, $\mathbb{P}\{\mathcal{G}_2\} \geq 1 - \frac{\delta}{T}$.

Lemma 2 follows by noting that each of the terms in the summation is a martingale difference term. Therefore, we can lower bound the probability of the converse event for a particular state action pair using Azuma-Hoeffding inequality [3], and taking union bound over all state-action pairs.

Lemma 3. For all states s such that $P_{\pi_k,p}(s) > 0$,

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

where T_M denotes the "mixing time", the maximum expected time it would take to hit one state by starting in another state. $\lceil T^{\alpha} \rceil$ denotes the duration of an episode.

The above Lemma follows by using Markov Inequality to lower bound the probability of hitting a state s in $2T_M$ steps by $1/2$. This is then multiplied by the probability of taking the action a in state s ($\pi_k(a|s)$), and the number of such durations of $2T_M$ in the episode of duration $\lceil T^{\alpha} \rceil$.

The UCRL-CMDP algorithm can be thought of as implementing an "index policy" that assigns an index to each stationary policy as follows,

$$\mathcal{I}_k(\pi) \coloneqq \max_{p' \in \mathcal{C}_{\tau_k}} \left\{ \overline{r}(\pi, p') : \overline{c}_i(\pi, p') \leq c_i^{ub} \right\}$$

For any policy not satisfying the cost constraints, the associated index is $-\infty$.

Lemma 4. For $p' \in \mathcal{C}_{\tau_k}$ on the "good set" $\mathcal{G} = \mathcal{G}_1 \cap \mathcal{G}_2$,

$$|\overline{r}(\pi, p) - \overline{r}(\pi, p')|, |\overline{c}_i(\pi, p) - \overline{c}_i(\pi, p')| \leq 2 \max_s \sum_{a \in \mathcal{A}} \pi_k(a|s) \epsilon_{\tau_k}(s, a) \coloneqq \delta_k(\pi).$$

The Lemma follows from the definition of the confidence interval $\mathcal{C}_{\tau_k}$, and using the triangle inequality on the distance between distributions of $\hat{p}_{\tau_k}$ and p, and $\hat{p}_{\tau_k}$ and p'. This can also be observed from the factor of 2 on the RHS of the inequality.

Using Lemma 4, we can make a comment on the nature of policies played during the episode k. Consider the threshold $\delta_k(\pi_k)$. If $\overline{c}_i(\pi_k, p) > c_i^{ub} + \delta_k(\pi_k)$, put together with $\overline{c}_i(\pi_k, p) \leq \overline{c}_i(\pi_k, p')\delta_k(\pi_k)$ from Lemma 4, we have that $\overline{c}_i(\pi_k, p') > c_i^{ub}$, i.e. associated index for all p' is $-\infty$.

Furthermore, if $|\overline{r}(\pi_k, p) - \overline{r}(\pi_k, p')| \leq \delta_k(\pi_k)$, then by the definition of the index $\mathcal{I}_k(\pi_k)$, it is upper bounded by $\overline{r}(\pi_k, p) + \delta_k(\pi_k)$. However, since this bound is greater than $\overline{r}(\pi_k, p)$, which in itself lies in the set of indices (for feasible MDPs p), we have that the index of a policy is lower bounded by $\overline{r}(\pi_k, p)$.

Lemma 5. On the "good set" $\mathcal{G}$, the instantaneous (single step) cost and reward regret can be bounded by $\delta_k(\pi_k)$.

Proof. Consider a stationary policy π being played. For the cost regrets, if $\overline{c}_i(\pi, p) > c_i^{ub} + \delta_k(\pi)$, then $\mathcal{I}_k(\pi) = -\infty$. However, we know the existence of a feasible policy $\tilde{\pi}$ whose index is greater than $\overline{r}(\tilde{\pi}, p)$, which is greater than $-\infty$. Therefore, for π to be played, we must have the instantaneous cost regret being upper bounded by $\delta_k(\pi)$.

Now, we move on to the reward regrets. The index of an optimal policy must always be greater than or equal to the optimal average reward r^*. However, we also have shown an upper bound on the index of the policy being played, i.e. the policy with the highest index. Therefore, $\overline{r}(\pi, p) + \delta_k(\pi) > r^*$, meaning the instantaneous regret is upper bounded by $\delta_k(\pi)$. □

With the above results on hand, we can provide a sketch proof of Theorem 1. Lemma 1 and Lemma 2 guarantee that the "good set" occurs with high probability, and it suffices to analyze the regrets incurred by the algorithm on this set. From Lemma 5, we know that the instantaneous regrets are bounded by $\delta_k(\pi_k)$. Therefore, the total regret is bounded by this quantity multiplied by the length of the episode. By bounding these quantities appropriately, we obtain that the reward and cost regrets are $\tilde{\mathcal{O}}(T^\beta)$. However, from the condition on β from Lemma 2, we have that $\beta = 2/3$, giving us regret of order $\tilde{\mathcal{O}}(T^{2/3})$. The exact details of bounding the regret are fleshed out in [1].

4.2 Achievable Regret Vectors

Considering the technicality of the analysis on regret in order to evaluate the performance of the algorithm, a natural question that arises is that of the use

of Lagrange multipliers. Given that we are required to solve the multi-objective optimization problem in (1), we could have used Lagrange multipliers to scalarize the problem. However, in that case, our derived bounds would be in terms of the multipliers $\{\lambda_i\}_{i=1}^{M}$. This would require us to further derive bounds on the multipliers, which is not a very straightforward task.

The real advantage of using Lagrange multipliers is evident in the attempt to characterize the set of achievable regret vectors. Consider a Lagrangian form of the constraints (1)-(2),

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

where boldface quantities represent vectors. The dual function for this expression is $\mathcal{D}(\lambda) \coloneqq \max_\pi \mathcal{L}(\boldsymbol{\lambda};\pi)$ with the dual problem being

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

Note that it has been shown that $\boldsymbol{\lambda^*}$ solving the dual problem satisfies $\mathcal{D}(\boldsymbol{\lambda^*}) = r^*$.

Theorem 2. There exists an underlying MDP p for which the reward and cost regrets under any policy π satisfy

$$\mathbb{E}_\pi \Delta^{(R)}(T) + \sum_{i=1}^{M} \lambda_i^* \mathbb{E}_\pi \Delta^{(i)}(T) \geq 0.015\sqrt{D(p)|\mathcal{S}||\mathcal{A}|T}.$$

Proof. Consider a regular MDP (without cost constraints) having reward function $r(s_t,a_t) + \boldsymbol{\lambda}\cdot(\mathbf{c^{ub}} - \mathbf{c}(s_t,a_t))$. Then, the optimal average reward for this MDP is $r^*(\boldsymbol{\lambda})$. From results on communicating MDPs in [2], we can choose the nature of the underlying MDP p such that $r^*(\boldsymbol{\lambda})T - \mathbb{E}_\pi \sum_{t=1}^{T} r(s_t,a_t) + \boldsymbol{\lambda}\cdot(\mathbf{c^{ub}} - \mathbf{c}(s_t,a_t)) \geq 0.015\sqrt{D(p)|\mathcal{S}||\mathcal{A}|T}$. Adding and subtracting r^*T on both sides and rearranging the terms, we obtain

$$\mathbb{E}_\pi \Delta^{(R)}(T) + \sum_{i=1}^{M} \lambda_i^* \mathbb{E}_\pi \Delta^{(i)}(T) \geq 0.015\sqrt{D(p)|\mathcal{S}||\mathcal{A}|T} + r^*T - r*(\boldsymbol{\lambda})T.$$

However, from the fact that $\boldsymbol{\lambda^*}$ satisfies $r^*(\boldsymbol{\lambda}) = \mathcal{D}(\boldsymbol{\lambda^*}) = r^*$ due to the definition of the reward function, we obtain the expression in the Theorem. □

5 Algorithm Performance

Consider an experimental setup of a single-hop wireless network consisting of a node transmitting data packets over an unreliable channel. The transmitting node has control over the transmission power, which is treated as the action a_t. The probability that the transmission successfully goes through is higher

when the transmission occurs at higher power levels. A suitable cost metric for this setup is the queue length at the transmitter. denoted by Q_t, which evolves as $Q_{t+1} = \min\{(Q_t + A_t - D_t)^+, B\}$, where A_t and D_t denotes arrivals and departures respectively, and B is the buffer size.

An optimal algorithm would prefer to minimize power consumption, or in terms of reward, maximize $(\mathbb{E}\sum -a_t)/T$ while ensuring that average queue length lies below a threshold $(\mathbb{E}\sum Q_t)/T \leq c^{ub}$.

The paper considers the comparison of the performance of UCRL-CMDP with a standard Actor-Critic algorithm in this setup. The Actor-Critic algorithm yields a high cost regret when simulated with specific parameter values. While the reward regret of the UCRL-CMDP algorithm is higher than that of the Actor-Critic algorithm, it also yields much lower cost regrets. The algorithm is thus effective in balancing both kinds of regrets, at the cost of marginally sub-optimal reward regrets. This earns the algorithm its name — "balanced optimism in the face of uncertainty", or BOFU.

6 Appendix

6.1 Extended Value Iteration

While value iteration plays a policy for a given MDP, in the UCRL2 (as well as UCRL-CMDP algorithm, we are also required to choose an optimistic MDP from the set of plausible MDPs. This can be achieved using an algorithm known as extended value iteration.

The algorithm performs the following iteration on the value function $V_i(\cdot)$ and normalized value function $V_i'(\cdot)$ for all $s \in \mathcal{S}$:

$$V_0(s) = 0,$$
$$V_i(s) = \max_{a\in\mathcal{A}} \left\{ \tilde{r}_k(s,a) + \max_{p\in\mathcal{M}_k} \left\{ \sum_{s'\in\mathcal{S}} p(s')V_i(s') \right\} \right\}.$$

The termination condition for the algorithm is when the change in state value function is nearly uniform and close to the average reward.

The external maximization provides us with the optimal policy, while the internal maximization chooses the optimistic MDP from the plausible set. Since the internal maximization is over a convex polytope, we are guaranteed that the algorithm converges in a finite number of iterations.

References

[1] R. Singh, A. Gupta, and N. B. Shroff, "Learning in constrained markov decision processes," *IEEE Transactions on Control of Network Systems*, vol. 10, no. 1, pp. 441–453, 2023.

[2] P. Auer, T. Jaksch, and R. Ortner, "Near-optimal regret bounds for reinforcement learning," in *Advances in Neural Information Processing Systems* (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.), vol. 21, Curran Associates, Inc., 2008.

[3] K. Azuma, "Weighted sums of certain dependent random variables," *Tohoku Mathematical Journal*, vol. 19, no. 3, pp. 357 – 367, 1967.