REGAL Revisited: Regularized Reinforcement Learning for Weakly Communicating MDPs

Anh Tuan Tran
Supervisors: Ziping Xu, Prof. Ambuj Tewari

December 2021

1 Introduction and background

Markov Decision Processes (MDPs) are used extensively in artificial intelligence and reinforcement learning to describe how an agent interacts with its environment. MDPs specify what possible states an agent can be in, what its possible actions are, and what the effects of taking those actions will be. By studying MDPs, we can develop techniques and algorithms which can be applied to many real-world problems where an agent is required to act intelligently. Many such algorithms have been developed with the assumption that the conditions of the MDP is known to the agent, i.e. the agent knows all the information it needs about the environment, including the possible states, actions, and effects of taking actions. This project focuses on settings where such information is unknown to the agent: it does not know how the world will behave when it takes a certain action. This makes it necessary for the agent to take actions it is unfamiliar with in order to gradually learn what their effects may be.

There have been several algorithms introduced to solve this problem, most notably UCRL2 [1]. However, the regret bound achieved by UCRL2 is dependent on the diameter of the MDP. The diameter is the maximum average number of steps it takes to get from one state to another state. In MDPs where at least one state is unreachable from any of the other states, the diameter is infinite, making UCRL2 ineffective.

In this project, I will build upon on already developed algorithm called REGAL [2]. This algorithm has good provable theoretical performance and achieves a regret bound which does not depend on the diameter. However, it cannot be implemented into a real computer program because it requires knowledge of certain information which are unavailable. This report presents a modification to the algorithm which makes it implementable, while preserving the provable theoretical performance of the original algorithm.

Algorithm 1 REGularization based Regret Minimizing ALgorithm (REGAL)

for episodes $k = 1, 2, \ldots,$ **do**
  $t_k \leftarrow$ current time
  $\mathcal{M}^k$ is the set of MDPs whose transition function satisfies (5) with $t = t_k$
  Choose $M^k \in \mathcal{M}^k$ to maximize the following criterion over $\mathcal{M}_k$,

$$\lambda^\star(M) - C_k \operatorname{sp}(h^\star(M)) \ .$$

  $\pi^k \leftarrow$ average reward optimal policy for M^k
  Follow π^k until some s, a pair gets visited $N_k(s,a)$ times
end for

2 Modification to REGAL using known deterministic episode length

Algorithm 1 describes the original REGAL algorithm. The algorithm works in episodes. At each episode k, the REGAL constructs a set of plausible MDPs $\mathcal{M}^k$. This is the set of MDPs which the algorithm believes has a high probability of containing the true underlying MDP which accurate describing the environment. From this set, REGAL then chooses the MDP which maximizes $\lambda^*(M) - C_k \text{sp}(h^*(M))$. This criterion maximizes the optimal average reward $\lambda^*(M)$ while minimizes the regularization term $C_k \text{sp}(h^*(M))$. The algorithm then computes the optimal policy for the chosen MDP, and follows that policy until the total number of visits to any state-action pair doubles. C_k is the regularization parameter which is set at the beginning of each episode as:

$$C_k = \frac{2\sum_{s,a} v_k(s,a)\sqrt{\frac{12S\log\frac{2AT}{\delta}}{N_k(s,a)}} + \sqrt{2l_k \log\frac{1}{\delta}}}{l_k}$$

The problematic term is $v_k(s,a)$, the number of times the state-action pair (s,a) is visited in the episode. In order to set C_k, $v_k(s,a)$ needs to be known before the episode begins, which is impossible. We also do not know l_k, because the stopping condition of the algorithm gives us no way to calculate the length of the episode.

In order to solve the issue of the regularization parameter C_k requring unavailable, we set the length of each episode according to some known deterministic function f. This removes the need to guess episode length. To achieve the same bounds as the original algorihm we use $f(k) = \frac{k}{\sqrt{SA}}$.

Algorithm 2: REGularization based Regret Minimizing ALgorithm (REGAL) with known episode length

for episodes $k = 1, 2, \ldots,$ **do**
  $t_k \leftarrow$ current time
  $\mathcal{M}^k$ is the set of MDPs whose transition function satisfies (5) with $t = t_k$
  Choose $M^k \in \mathcal{M}^k$ to maximize the following criterion over $\mathcal{M}^k$:

$$\lambda^*(M) - C_k \text{sp}(h^*(M))$$

  $\pi^k \leftarrow$ average reward optimal policy for $\mathcal{M}^k$
  Follow π^k for $f(k)$ time steps.
end

Regret bound

Consider an episode $k \in G$, we have

$$\lambda_k^* - C_k \text{sp}(h_k^*) \geq \lambda^* - C_k \text{sp}(h^*) \tag{1}$$

$$\begin{aligned}
\Delta_k &= \sum_{s,a} v_k(s,a)[\lambda^* - r(s,a)] \\
&= \sum_{s,a} v_k(s,a)[\lambda^* - \lambda_k^* + \lambda_k^* - r(s,a)] \\
&\leq \sum_{s,a} v_k(s,a)[\lambda_k^* - C_k \text{sp}(h_k^*) + C_k \text{sp}(h^*) - r(s,a)] \\
&= \sum_{s,a} v_k(s,a)[\lambda_k^* - r(s,a)] - C_k \sum_{s,a} v_k(s,a)[\text{sp}(h_k^*) - \text{sp}(h^*)] \\
&= \mathbf{v}_k(\lambda_k^* \mathbf{e} - \mathbf{r}_k) - C_k \sum_{s,a} v_k(s,a)[\text{sp}(h_k^*) - \text{sp}(h^*)] \\
&= \mathbf{v}_k(\tilde{\mathbf{P}}_k - \mathbf{I}) h_k^* - C_k \sum_{s,a} v_k(s,a)[\text{sp}(h_k^*) - \text{sp}(h^*)] \\
&\quad \text{[using equation (7) of [2]]} \\
&= \mathbf{v}_k(\tilde{\mathbf{P}}_k - \mathbf{P}_k + \mathbf{P}_k - \mathbf{I}) h_k^* - C_k \sum_{s,a} v_k(s,a)[\text{sp}(h_k^*) - \text{sp}(h^*)] \\
&\leq \|\mathbf{v}_k(\tilde{\mathbf{P}}_k - \mathbf{P}_k)\|_1 \text{sp}(h_k^*) + \mathbf{v}_k(\mathbf{P}_k - \mathbf{I}) h_k^* - C_k \sum_{s,a} v_k(s,a)[\text{sp}(h_k^*) - \text{sp}(h^*)]
\end{aligned} \tag{2}$$

From the proof of Lemma 10 of [1], replacing Hoeffding-Azuma with Bernstein's in-

equality we have

$$\sum_{k\in G} \mathbf{v}_k(\mathbf{P}_k - \mathbf{I})h_k^* \leq \sqrt{2\sum_{k\in G} \mathrm{sp}(h_k^*)^2 l_k \log(1/\delta)} + \max_{k\in G} \mathrm{sp}(h_k^*)(m + \log(1/\delta)) \tag{3}$$

If $k \in G$ then $M, M_k \in \mathcal{M}$, and we also have

$$\|\mathbf{v}_k(\tilde{\mathbf{P}}_k - \mathbf{P}_k)\|_1 \leq 2\sum_{s,a} v_k(s,a)\sqrt{\frac{12S\log(2AT/\delta)}{N_k(s,a)}} \tag{4}$$

Plugging equations 3 and 4 into 2, we get

$$\begin{aligned}\sum_{k\in G} \Delta_k \leq \sum_{k\in G} \mathrm{sp}(h_k^*) &\left(2\sum_{s,a} v_k(s,a)\sqrt{\frac{12S\log(2AT/\delta)}{N_k(s,a)}} + \sqrt{2l_k\log(1/\delta)} + C_k\sum_{s,a} v_k(s,a)\right) \\ &+ \sum_{k\in G} C_k l_k \mathrm{sp}(h^*) + (\mathrm{sp}(h^*) + \max_{k\in G}\frac{1}{C_k})(m + \log(1/\delta))\end{aligned} \tag{5}$$

Let $\overline{T} = \max_k l_k \leq \sqrt{\frac{T}{SA}}$ (because we choose $f(k) = l_k = \frac{k}{\sqrt{SA}}$), Lemma 4 of [3] gives

$$\sum_{s,a}\sum_{k} \frac{\nu_k(x)}{\sqrt{\max(1, N_k(x))}} \leq \sqrt{SAT} + \sqrt[4]{SAT} \tag{6}$$

Thus choosing

$$C_k = \frac{2\sqrt{12S\log(2AT/\delta)}(\sqrt{SAT} + \sqrt[4]{SAT}) + \sqrt{2l_k\log(1/\delta)}}{l_k} \tag{7}$$

gives

$$\begin{aligned}\sum_{k\in G} \Delta_k &\leq \mathrm{sp}(h^*)(2\sqrt{12S\log(2AT/\delta)}(\sqrt{SAT} + \sqrt[4]{SAT}) + \sqrt{2l_k\log(1/\delta)}) \\ &\quad + (\mathrm{sp}(h^*) + \max_{k\in G}\frac{1}{C_k})(m + \log(1/\delta)) \\ &= O(\mathrm{sp}(h^*)S\sqrt{AT\log(AT/\delta)})\end{aligned} \tag{8}$$

This modified REGAL algorithm achieves the same regret bound as the original. The C_k parameter no longer requires knowledge of $v_k(s,a)$, while l_k is known at the start of the episode.

3 Addressing several issues for REGAL paper

3.1 Definition issue

$\mathcal{M}^k$, the set of MDPs which contains the true MDP with high probability, does not exclusively contain weakly communicating MDPs. Therefore, the optimization step in REGAL is ill-defined because there is no guarantee that all MDPs in $\mathcal{M}^k$ have constant gain. Furthermore, there may exist several optimal MDPs which maximizes the regularized gain of REGAL. To address this, the set $\mathcal{M}^k$ should be redefined to only contain weakly communicating MDPs, and in the case of multiple optimal MDPs, the MDP with the largest span $sp(h^*(M))$.

3.2 Problem with Theorem 3 proof

Fruit et al claims the following proof issue in the proof of theorem 3, in which the term

$$c\sqrt{\Sigma_{s,a} v_{k,j}(s,a)} - C_{k,j}\Sigma_{s,a} v_{k,j}(s,a)$$

cannot be eliminated in equation 16 because it is not less than or equal to 0:

$$\begin{aligned} \Sigma_{s,a} v_{k,j}(s,a) &\leq 2^j \\ -C_{k,j} = -\frac{c}{\sqrt{\Sigma_{s,a} v_{k,j}(s,a)}} &\leq -\frac{c}{\sqrt{2^j}} \leq 0 \qquad\qquad c \geq 0 \\ -C_{k,j}\Sigma_{s,a} v_{k,j}(s,a) &\geq -C_{k,j} 2^j \end{aligned}$$

However, the REGAL paper does not use any inequality to eliminate this term. Instead, the term

$$c\sqrt{\Sigma_{s,a} v_{k,j}(s,a)} - C_{k,j}\Sigma_{s,a} v_{k,j}(s,a)$$

is set to 0, by definition of $C_{k,j} = -\frac{c}{\sqrt{\Sigma_{s,a} v_{k,j}(s,a)}}$.
Thus, their claim in incorrect.

4 Next steps: The maximization problem

This section explores methods for implementing the maximization in Algorithm 2. The problem is to find an implementable algorithm to choose $M^k \in \mathcal{M}^k$ which maximizes the following value over $\mathcal{M}^k$:

$$\lambda^*(M) - C_k \mathrm{sp}(h^*(M))$$

We can construct an extended MDP based on extreme transition dynamics. For each transition $P(x)$, the extreme transition probability which encourages visiting s is

$$P(x)^{s+} = P(x) - W_P(\cdot|x) + \mathbb{1}_s \Sigma_j W_P(j|x)$$

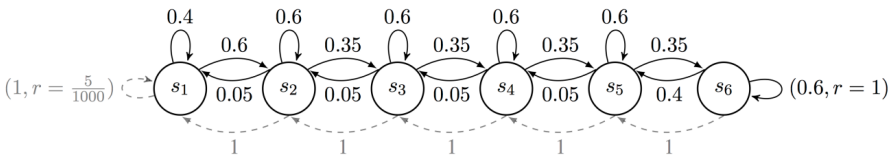

Figure 1: The Riverswim MDP

The extended MDP combines all potential MDP $M^k \in \mathcal{M}^k$ into a single MDP, with extended action space $\mathcal{A}'$. For each action a in $\mathcal{A}$, there are corresponding actions $a' \in \mathcal{A}'$, such that $a' = (a, s)$, where s is the state onto which all the uncertainty is placed.

To prove that the policy which maximizes the penalized reward in the extended MDP corresponds to the policy which maximizes the penalized reward in the original MDP, we have to find a bound on $\text{sp}(h^*(M)) - \text{sp}(\widetilde{h}(M))$.

5 Evaluation

To make the problem tractable, I implemented a version which only considers a discrete set of MDPs. Each MDP in this set uses only the most extreme values of the transition dynamics (maximum or minimum). An early implementation allowed each state-action pair to select its own next state to prioritize in the extreme transition dynamics, however, the runtime was excessively long. Instead, every state-action pair now has to prioritize the same next state.

I also implemented several other popular algorithms for MDPs to compare their performance against REGAL. They are evaluated on the Riverswim MDP, shown in Figure 1. The agent starts at state s1. It can choose to go left, which is always successful. Alternatively, it can choose to move right, which has a 0.35 probability of moving it right, 0.6 probability of staying still, and 0.05 probability of moving to the left. The rightmost state s6 has the highest reward of 1, while the leftmost state has very low reward.

In Figure 2, we see that REGAL performs about the same as UCRL2, however they are both outperformed by Thompson sampling and SCAL. However, when the Riverswim is modified so that the leftmost state has a much high reward of 0.35, while the rightmost state still has reward 1.0, REGAL performs better than the other algorithms (Figure 3). By increasing the reward of the leftmost state, the span of the bias vector of the MDP is reduced, which suggests that the modified REGAL algorithm can outperform other algorithms in low span environments.

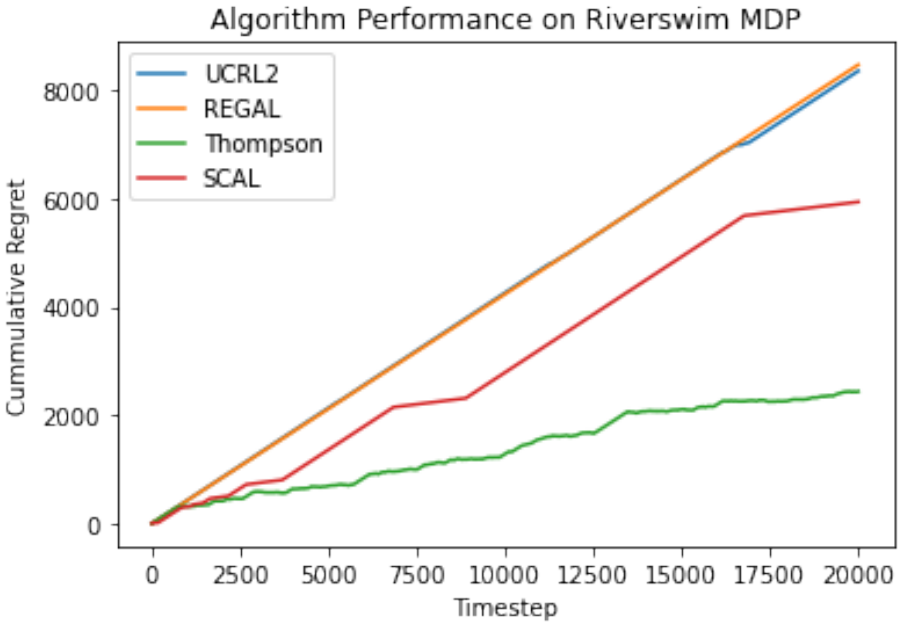

Figure 2: Cumulative regret achieved by several algorithms in the Riverswim MDP environment.

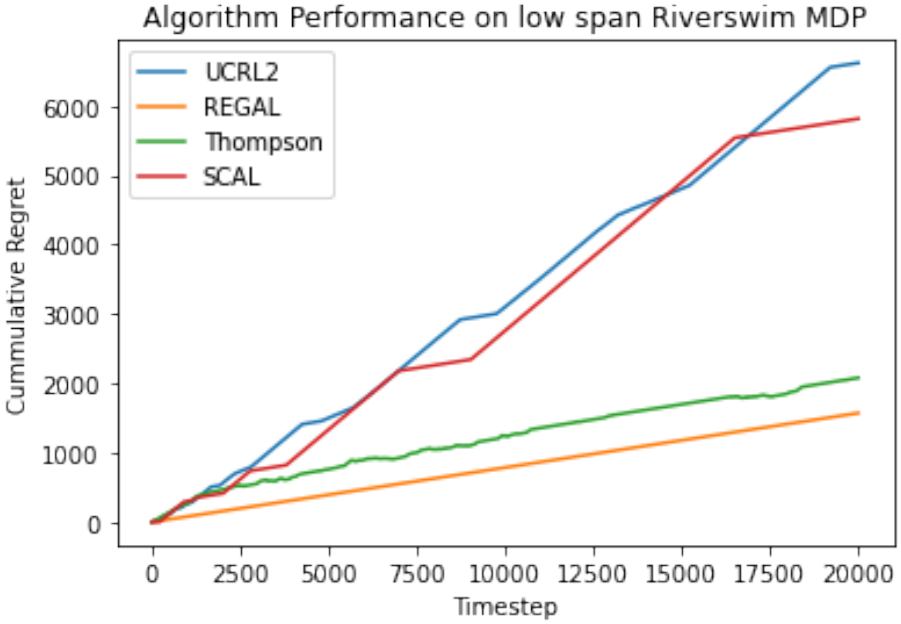

Figure 3: Cumulative regret achieved by several algorithms in the low span Riverswim MDP environment.

6 Applications

Many real world environments are complex and messy, and thus can be considered unknown. Agents which interact with the real world do not fully know what effects their actions will have on the environment. By continuing to improve on algorithms which deal with unknown MDPs, we will be able to develop better and better autonomous systems which can behave more intelligently in unknown or uncertain environments.

References

[1] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning.," *Journal of Machine Learning Research*, vol. 11, no. 4, 2010.

[2] P. L. Bartlett and A. Tewari, "Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps," *arXiv preprint arXiv:1205.2661*, 2012.

[3] Z. Xu and A. Tewari, *Reinforcement learning in factored mdps: Oracle-efficient algorithms and tighter regret bounds for the non-episodic setting*, 2020. arXiv: 2002.02302 [stat.ML].