Safe Reinforcement Learning with Linear Function Approximation

Sanae Amani 1 Christos Thrampoulidis 2 Lin F. Yang 3
1 University of California, Los Angeles
2University of British Columbia, Vancouver

Abstract

We study safety in RL by first modeling safety as an unknown linear cost function of states and actions, which must always fall below a certain threshold. We then present algorithms, termed SLUCB-QVI and RSLUCB-QVI, for finite-horizon Markov decision processes (MDPs) with linear function approximation. We show that SLUCB-QVI and RSLUCB-QVI, while with no safety violation, achieve a $O(\sqrt{THT})$ regret, nearly matching that of state-of-the-art unsafe algorithms.

Key Assumptions

- $M$ is a linear MDP with feature map $\phi: S \times A \to \mathbb{R}^d$; if for any $h \in [H]$, there exist $d$ unknown measures $\mu_k := (\mu_k(1), \ldots, \mu_k(2))$ over $S$, and unknown vectors $\theta_k, \gamma_k \in \mathbb{R}^d$, such that $E_h(s, a) = \phi_h(s, a)^T \theta_k$, $r_h(s, a) = \phi_h(s, a)^T \gamma_k$, and $c_h(s, a) = \phi_h(s, a)^T \phi_h(s, a)$.
- For all $s \in S$, there exists a known safe action $a(s)$ with known safety measure $c_h(s) = \left(\phi(s, a(s))^T \phi(s, a)\right) < \tau$ for all $h \in [H]$.

SLUCB-QVI and RSLUCB-QVI

- Algorithms for deterministic and randomized policy selection.
- At each episode $k$, the agent uses the cost feedback to compute the non-empty sets $\Gamma_{k+1}$, such that $\Gamma_{k+1} \in \mathbb{R}^d$.
- The agent runs LSVI to compute $Q^*$ which is an upper bound on true $Q$ at each episode $k \in [K]$.

Experiments for SLUCB-QVI

The agent seeks to reach a goal in a 10 x 10 2D map while avoiding dangers. At each time step, the agent can move in four directions, i.e., $A = (a_1 : left, a_2 : right, a_3 : down, a_4 : up)$. With probability 0.9 it moves in the desired direction and with probability 0.05 it moves in either of the orthogonal directions. We implemented RSLUCB-QVI for 10 interaction units (episodes) i.e., $K = 10$ each consisting of 1000 time-steps (horizon), i.e., $H = 1000$. During each interaction unit (episode) and after each move, the agent can end up in one of three kinds of states: 1) goal, resulting in a successful termination of the interaction unit; 2) danger, resulting in a failure and the consequent termination of the interaction unit; 3) safe. The agent receives a reward of 0 for reaching the goal and 0.03 otherwise. We report the average success rate and return over and compare our results with that of CBR proposed by [Turchetta et al., 2020] in which a teacher helps the agent in selecting safe actions by making interventions.

Technical Novelty

Lemma (Optimism in the face of safety constraint): Let $\alpha_0 > 0$ be a constant characterizing the safety constraints, while with high probability it holds that $\tau^k \in \Gamma_{k+1}$ for all $k \in [K]$.

Theoretical Guarantees

Theorem (Regret of SLUCB-QVI and RSLUCB-QVI): They achieve a $O(\sqrt{THT})$ regret, nearly matching that of state-of-the-art unsafe algorithms, where $x := \alpha_0 \frac{1}{\tau^k} - 1$ is a constant characterizing the safety constraints, while with high probability it holds that $\tau^k \in \Gamma_{k+1}$ for all $k \in [K]$.

Contributions

We developed SLUCB-QVI and RSLUCB-QVI, two safe RL algorithms in the setting of finite-horizon linear MDP For these algorithms, we provided sub-linear regret bounds $O(\sqrt{THT})$. We proved that with high probability, they never violate the unknown safety constraints.

References

