Safe Reinforcement Learning with Linear Function Approximation

Sanae Amani [1] Christos Thrampoulidis [2] Lin F. Yang[1]

[1] University of California, Los Angeles

[2]University 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 $\tilde{\mathcal{O}}\left(\kappa\sqrt{d^3H^3T}\right)$ regret, nearly matching that of state-of-the-art unsafe algorithms.

Problem Formulation

- **Finite horizon MDP:** $M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, c)$, $\mathcal{S}$: known state set, $\mathcal{A}$: known action set, H: known episode's length, $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$: unknown transition probabilities, $r = \{r_h\}_{h=1}^H$: unknown reward functions, and $c = \{c_h\}_{h=1}^H$: unknown cost functions.
- **Safety Constraint:** When being in state s_h^k, at episode k and time-step $h \in [H]$, the agent must select a *safe* policy π_h^k such that
  - if π_h^k is deterministic:
  $$c_h(s_h^k, \pi_h^k(s_h^k)) \le \tau.$$
  - if π_h^k is randomized:
  $$\mathbb{E}_{a \sim \pi_h^k(s_h^k)} c_h(s_h^k, a) \le \tau.$$

Set of safe policies
$$\Pi^{\text{safe}} := \left\{\pi : \pi_h(s) \in \Gamma_h^{\text{safe}}(s),\ \forall (s,h) \in \mathcal{S} \times [H]\right\}.$$

- if the policies are deterministic
$$\Gamma_h^{\text{safe}}(s) := \{a \in \mathcal{A} : c_h(s,a) \le \tau\}.$$
- if the policies are randomized
$$\Gamma_h^{\text{safe}}(s) := \{\theta \in \Delta_{\mathcal{A}} : \mathbb{E}_{a\sim\theta} c_h(s,a) \le \tau\}.$$

Goal

Let
$$V_h^*(s) = \sup_{\pi \in \Pi^{\text{safe}}} V_h^\pi(s),\ \forall (s,h) \in \mathcal{S} \times [H]$$
.

The agent's goal is to keep
$$R_K := \sum_{k=1}^K V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k).$$
as small as possible, while π^k are safe, i.e., $\pi^k \in \Pi^{\text{safe}}$ for all $k \in [K]$ with high probability.

Key Assumptions

1. M is a linear MDP with feature map $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, if for any $h \in [H]$, there exist d unknown measures $\boldsymbol{\mu}_h^* := [\mu_h^{*(1)}, \ldots, \mu_h^{*(d)}]^\top$ over $\mathcal{S}$, and unknown vectors $\boldsymbol{\theta}_h^*, \gamma_h^* \in \mathbb{R}^d$ such that $\mathbb{P}_h(.|s,a) = \left\langle \boldsymbol{\mu}_h^*(.), \boldsymbol{\phi}(s,a)\right\rangle$, $r_h(s,a) = \langle \boldsymbol{\theta}_h^*, \boldsymbol{\phi}(s,a)\rangle$, and $c_h(s,a) = \left\langle \gamma_h^*, \boldsymbol{\phi}(s,a)\right\rangle$.
2. For all $s \in \mathcal{S}$, there exists a known safe action $a_0(s)$ with known safety measure $\tau_h(s) := \left\langle \boldsymbol{\phi}\left(s, a_0(s)\right), \gamma_h^*\right\rangle < \tau$ 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 $\Gamma_h^k(s)$, such that
$$\mathbb{P}\left(\Gamma_h^k(s) \subset \Gamma_h^{\text{safe}}(s),\ \forall (s,h,k) \in \mathcal{S} \times [H] \times [K]\right) \ge 1 - \delta.$$
- The agent runs LSVI to compute Q^k which is an upper bound on true Q at each episode $k \in [K]$.

Experiments for RSLUCB-QVI

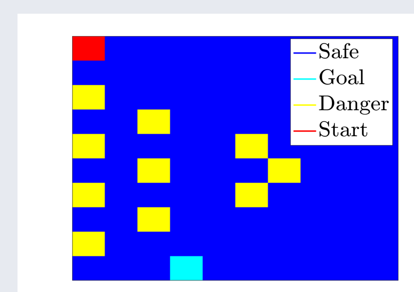

Figure: Map

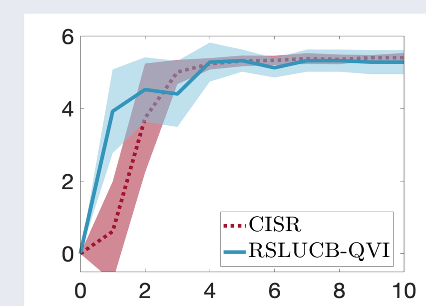

Figure: Average Return

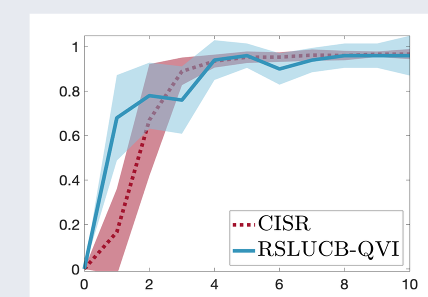

Figure: Success Rate

The agent seeks to reach a goal in a 10×10 2D map while avoiding dangers. At each time step, the agent can move in four directions, i.e., $\mathcal{A} = \{a_1 : \text{left}, a_2 : \text{right}, a_3 : \text{down}, a_4 : \text{up}\}$. 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 return of 6 for reaching the goal and 0.01 otherwise. We report the average of success rate and return over and compare our results with that of CISR proposed by [Turchetta et al., 2020] in which a teacher helps the agent in selecting safe actions by making interventions.

Experiments for SLUCB-QVI

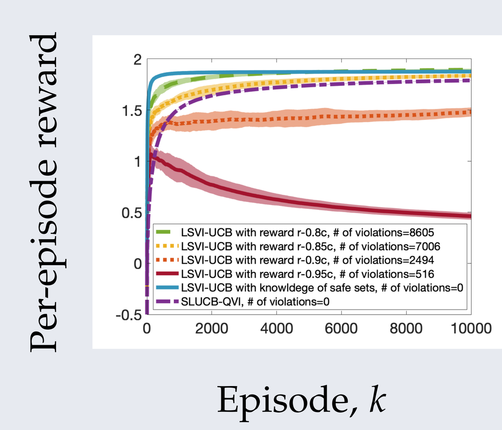

Figure: Comparison of SLUCB-QVI to the unsafe state-of-the-art verifying that: 1) when LSVI-UCB [Jin et al., 2020] has knowledge of γ_h^*, it outperforms SLUCB-QVI (without knowledge of γ_h^*) as expected; 2) when LSVI-UCB does not know γ_h^* (as is the case for SLUCB-QVI) and its goal is to maximize $r - \lambda' c$ instead of r, larger λ' leads to smaller per-episode reward and number of constraint violations while the number of constraint violations for SLUCB-QVI is zero.

Technical Novelty

The linear structure of the MDP allows us to parametrize $Q_h^*(s,a)$ by a linear form. Thus, a natural idea to estimate $Q_h^*(s,a)$ is to solve least-squares problem for $\mathbf{w}_h^*$. In fact, the agent computes $Q_h^k(s,a)$ defined as
$$Q_h^k(s,a) = \min\left\{\left\langle \mathbf{w}_h^k, \boldsymbol{\phi}(s,a)\right\rangle + \underbrace{\kappa_h(s)\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}}_{\text{exploration bonus}}, H\right\},$$

1. β that encourages enough exploration regarding the uncertainty about r and $\mathbb{P}$
2. **Main Challenge:** $\kappa_h(s) > 1$ that encourages enough exploration regarding the uncertainty about c.

We make use of standard analysis of unsafe bandits and MDPs [Abbasi-Yadkori et al., 2011] and [Jin et al., 2020] to define β, appropriately quantifying $\kappa_h(s)$ is the main challenge the presence of safety constraints brings to the analysis of SLUCB-QVI compared to the unsafe LSVI-UCB.

Technical Novelty

Lemma (Optimism in the face of safety constraint): Let $\kappa_h(s) := \frac{2H}{\tau - \tau_h(s)} + 1$. Then, with appropriate choice of β, it holds that
$$\mathbb{P}\left(V_h^*(s) \le V_h^k(s), \forall (s,h,k) \in \mathcal{S} \times [H] \times [K]\right) \ge 1 - \delta.$$

Finally, the agent runs an upper confidence bound (UBC) decision rule

- deterministic policy:
$$\text{play } a_h^k = \arg\max_{a \in \Gamma_h^k(s_h^k)} Q_h^k(s_h^k, a),$$
- randomized policy:
$$\text{play } a_h^k \sim \arg\max_{\theta \in \Gamma_h^k(s_h^k)} \mathbb{E}_{a\sim\theta}\left[Q_h^k(s_h^k, a)\right].$$

Theoretical Guarantees

Theorem (Regret of SLUCB-QVI and RSLUCB-QVI): They achieve a $\tilde{\mathcal{O}}\left(\kappa\sqrt{d^3H^3T}\right)$ regret, nearly matching that of state-of-the-art unsafe algorithms, where
$$\kappa := \arg\max_{h,s} \frac{2H}{\tau - \tau_h(s)} + 1$$
is a constant characterizing the safety constraints, while with high probability it holds that $\pi^k \in \Pi^{\text{safe}}$ 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 $\tilde{\mathcal{O}}\left(\kappa\sqrt{d^3H^3T}\right)$. We proved that with high probability, they never violate the unknown safety constraints.

References

[Abbasi-Yadkori et al., 2011] Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320.

[Amani et al., 2019] Amani, S., Alizadeh, M., and Thrampoulidis, C. (2019). Linear stochastic bandits under safety constraints. In *Advances in Neural Information Processing Systems*, pages 9252–9262.

[Jin et al., 2020] Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143.

[Turchetta et al., 2020] Turchetta, M., Kolobov, A., Shah, S., Krause, A., and Agarwal, A. (2020). Safe reinforcement learning via curriculum induction. *arXiv preprint arXiv:2006.12136*.