Long-term Safe Reinforcement Learning with Binary Feedback

Akfumi Wachi[1], Wataru Hashimoto[2], Kazumune Hashimoto[2]

[1]LY Corporation [2]Osaka University

Our Contributions

- Propose an algorithm called LoBiSaRL.
- LoBiSaRL enables us to solve safe RL problems with binary feedback and unknown, stochastic state transition while guaranteeing the satisfaction of long-term safety constraints.
- Theoretically show that future safety can be pessimistically characterized by 1) inevitable randomness due to the stochastic state transition and 2) divergence between the current policy and a reference policy to stabilize the state.
- Empirically demonstrate the effectiveness of the LoBiSaRL compared with several baselines.

Safe Reinforcement Learning

- Safe reinforcement learning (RL) is a promising paradigm for applying RL algorithms to real-world applications (Garcıa and Fernández, 2015).
- Safe RL is beneficial in safety-critical decision-making problems, such as autonomous driving, healthcare, and robotics, where safety issues must be incorporated to prevent RL policies from posing risks to humans or objects.
- Safe RL has received significant attention in recent years as a crucial issue of RL during both the learning and execution phases (Amodei et al., 2016).

Existing Safe RL Approaches

- Safe RL is typically formulated as constrained policy optimization problems where the expected cumulative reward is maximized while guaranteeing or encouraging the satisfaction of safety constraint.
- Satisfying safety constraints almost surely or with high probability received less attention to date.

- Several previous work on safe RL aimed to guarantee safety at every time step with high probability, even during the learning process.
- However, existing work has room for improvement regarding strong assumptions.

	State transition		Safety	Additional assumption(s)
	Known	D/S		
Wachi and Sui (2020)	Yes	D	GP	-
Amani et al. (2021)	Linear	S	Linear	Known safe policy
Wachi et al. (2021)	Yes	D	GLM	-
Bennett et al. (2023)	No	S	GLM	Known safe policy
LoBiSaRL (Ours)	No	S	GLM	Lipschitz continuity & conservative policy

Table 1. Comparison among existing work regarding their assumptions on a state transition, safety function, and others (D means deterministic state transition, and S means stochastic state transition).

Problem Formulation

Consider episodic finite-horizon CMDPs, which are defined as a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, P, T, r, g, s_1)$, where $\mathcal{S}$ is a state space $\{s\}$,$\mathcal{A}$ is an action space $\{a\}$, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is an unknown, stochastic state transition function to map a state-action pair to a probability distribution over the next states, $T \in \mathbb{Z}_+$ is a fixed length of each episode, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a reward function, $g : \mathcal{S} \times \mathcal{A} \to \{0, 1\}$ is an unknown binary safety function, and $s_1 \in \mathcal{S}$ is the initial state.

Goal. This paper aims to obtain the optimal policy $\pi^\star : \mathcal{S} \to \mathcal{A}$ to maximize the value function $V_t^\pi(s_t)$ under the following safety constraint such that

$$\max_\pi V_t^\pi(s_t) \ \text{ s.t. } \ \Pr\Big\{g(s_\tau, a_\tau) = 1 \ \ \forall \tau \in [t, T]\Big\} \geq 1 - \delta.$$

Remark. It is quite hard to satisfy the aforementioned constraint. It is because even if the agent executes an action a_t at time t and state s_t such that

$$\Pr\Big\{g(s_t, a_t) = 1\Big\} \geq 1 - \delta, \tag{1}$$

there may *not* be any viable action at $s_{t+1} \sim P(s_t, a_t)$ and further future states.

Difficulties and Assumptions

Difficulty 1. If the binary safety function does not exhibit any regularity, it is impossible to infer the safety of state-action pairs.

Assumption 1. There exists a known feature mapping function $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^m$, unknown coefficient vectors $\boldsymbol{w}^\star \in \mathbb{R}^m$, and a fixed, strictly increasing (inverse) link function $\mu : \mathbb{R} \to [0, 1]$ such that

$$\mathbb{E}[\, g(s, a) \mid s, a \,] = \mu\big(f^\star(s, a)\big), \tag{2}$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, where $f^\star : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a linear predictor defined as

$$f^\star(s, a) := \langle \boldsymbol{\phi}(s, a), \boldsymbol{w}^\star \rangle, \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}. \tag{3}$$

Difficulty 2. The state transition is stochastic and unknown a priori. Hence, to satisfy the long-term safety constraint, we must explicitly incorporate the stochasticity of the state transition and its influence on future safety.

Assumption 2. For all $s, \bar{s} \in \mathcal{S}$ and $a, \bar{a} \in \mathcal{A}$, the feature mapping function $\boldsymbol{\phi}(\cdot, \cdot)$ is Lipschitz continuous with a constant $L_\phi \in \mathbb{R}_+$; that is,

$$\|\boldsymbol{\phi}(s, a) - \boldsymbol{\phi}(\bar{s}, \bar{a})\|_2 \leq L_\phi \cdot d_{\mathcal{SA}}((s, a), (\bar{s}, \bar{a})), \tag{4}$$

where $d_{\mathcal{SA}}(\cdot, \cdot)$ is a distance metric on $\mathcal{S} \times \mathcal{A}$. For ease of exposition, we assume that $d_{\mathcal{SA}}$ satisfies $d_{\mathcal{SA}}((s, a), (\bar{s}, \bar{a})) = d_{\mathcal{S}}(s, \bar{s}) + d_{\mathcal{A}}(a, \bar{a})$.

Assumption 3. Let $L_\sharp \in \mathbb{R}_+$ be a positive scalar. There exists a known $L_\sharp$-Lipschitz continuous policy $\pi^\sharp : \mathcal{S} \to \mathcal{A}$ such that, for any states $s, \bar{s} \in \mathcal{S}$,

$$d_{\mathcal{A}}(\pi^\sharp(s) - \pi^\sharp(\bar{s})) \leq L_\sharp \cdot d_{\mathcal{S}}(s, \bar{s}). \tag{5}$$

Also, with a positive scalar $\eta \in \mathbb{R}_+$, for any policy $\pi : \mathcal{S} \to \mathcal{A}$, the following inequality holds for all $s \in \mathcal{S}$:

$$\max_{s' \sim P(\cdot \mid s, \pi(s))} d_{\mathcal{S}}(s, s') \leq \bar{d} + \eta \cdot d_{\mathcal{A}}(\pi(s), \pi^\sharp(s)).$$

To guarantee long-term safety, it is important to properly tune the maximum divergence from the conservative policy (MDCP).

Characterizing Safety

We first obtain the lower bounds of the safety linear predictor $f^\star$:

$$\ell(s_t, a_t) := \max\big(\ell_{\text{GLM}}(s_t, a_t), \ell_{\text{Lipschitz}}(s_t, a_t)\big), \tag{6}$$

where $\ell_{\text{GLM}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $\ell_{\text{Lipschitz}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ are pessimistic safety linear predictors inferred by GLM and Lipshitz continuity, defined as

$$\ell_{\text{GLM}}(s_t, a_t) := \langle \boldsymbol{\phi}(s_t, a_t), \hat{\boldsymbol{w}} \rangle - \beta \cdot \|\boldsymbol{\phi}(s_t, a_t)\|_{W_n^{-1}},$$
$$\ell_{\text{Lipschitz}}(s_t, a_t) := f^\sharp(s_1) - L_1 \left\{ L_2 t + L_3 X_1^{t-1} + x_t \right\},$$

Note: L_1, L_2 and L_3 are Lipschitz constants.

Theoretical Result

Theorem 1. Suppose, at state s_t, the agent executes the action a_t while tuning the MDCPs $x_t, x_{t+1}, \ldots, x_T$ so that the following inequality holds.

$$\ell(s_t, a_t) - L_1 \left\{ L_2(T - t) - (L_3 - 1)x_t + L_3 X_{t+1}^{T-1} + x_T \right\} \geq z \tag{7}$$

Set $\delta := 1 - (1 - \mu(z))^{T-t}$. Then, we have

$$\Pr\Big\{g(s_\tau, a_\tau) = 1 \ \ \forall \tau \in [t, T]\Big\} \geq 1 - \delta, \quad \forall t \in [T],$$

— i.e. the long-term safety constraint is satisfied — with a high probability.

Experiments

- Grid-world environment with random reward and safety.
- Instantaneous agent is much safer than Random, Unsafe, Linear baselines but sometimes violates the safety constraint.
- As for safety, LoBiSaRL is the only algorithm to guarantee the satisfaction of the safety constraint in the long run.
- LoBiSaRL is often too conservative and the performance in terms of reward is worse than Instantaneous.

	Reward	Unsafe actions
Random	0.32 ± 0.24	23.2 ± 10.3
Unsafe	$\mathbf{1.00 \pm 0.00}$	26.8 ± 13.6
Linear	0.73 ± 0.13	18.3 ± 5.7
Instantaneous	0.86 ± 0.10	3.3 ± 2.2
LoBiSaRL (Ours)	0.76 ± 0.12	$\mathbf{0.0 \pm 0.0}$

Table 2. Experimental results. Reward is normalized with respect to Unsafe agent.

References

Amani, S. et al. (2021). Safe reinforcement learning with linear function approximation. In *ICML*.

Amodei, D. et al. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Bennett, A. et al. (2023). Provable safe reinforcement learning with binary feedback. In *AISTAT*.

Garcıa, J. and Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *JMLR*, 16(1):1437–1480.

Wachi, A. et al. (2021). Safe policy optimization with local generalized linear function approximations. In *NeurIPS*.

Wachi, A. and Sui, Y. (2020). Safe reinforcement learning in constrained Markov decision processes. In *ICML*.

akifumi.wachi@lycorp.co.jp