# A Survey of Constraint Formulations in Safe Reinforcement Learning


Akfumi Wachi[1], Xun Shen[2], Yanan Sui[3]

[1]LY Corporation [2]Osaka University [3]Tshinghua University


## Our Contributions

- Provide a comprehensive survey focusing on constraint formulations in safe reinforcement learning and introduce representative algorithms for each formulation.
- Discuss the relationships between various constraint formulations by defining three theoretical notions: transformability, generalizability, and conservative approximation.
- Present theoretical results demonstrating that two problems exist, termed Identical or More General Safe RL (IoMG-SafeRL) problems, into which other common problems can be either transformed or conservatively approximated.
- Bridge the gaps between the safe RL problems with appropriate algorithms by organizing existing research focusing on constraint formulation.

## Safe Reinforcement Learning

- Safe reinforcement learning (RL) is a promising paradigm for applying RL algorithms to real-world applications.
- Safe RL is typically modeled as constrained Markov decision processes (CMDPs).

$$\mathcal{M} \cup \mathcal{C} \coloneqq \underbrace{\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, H, r, \gamma_r, \rho \rangle}_{\text{Standard MDP } (\mathcal{M})} \cup \mathcal{C},$$

where $\mathcal{S} \coloneqq \{s\}$ is a state space, $\mathcal{A} \coloneqq \{a\}$ is an action space, and $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the state transition, where $\mathcal{P}(s' \mid s, a)$ is the probability of transition from state $s$ to state $s'$ when action $a$ is taken. $H \in \mathbb{Z}_+$ is the (fixed) finite length of each episode, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, $\gamma_r \in [0, 1)$ is the discount factor for the reward, and $\rho \in \Delta(\mathcal{S})$ is the initial state distribution.

- Given a policy $\pi \in \Pi$, the value function is defined as

$$V_r^\pi(\rho) \coloneqq \mathbb{E}_\pi \left[ \sum_{h'=h}^{H} \gamma_r^{h'} r(s_{h'}, a_{h'}) \,\middle|\, s_h = s \right].$$

Since the initial state $s_0$ is sampled from $\rho$, we slightly abuse the notation and define

$$V_r^\pi(\rho) \coloneqq \mathbb{E}_{s \sim \rho} \left[ V_{r,0}^\pi(s) \right].$$

- Given a constraint tuple $\mathcal{C}$, a policy must be within the feasible policy space

$$\widehat{\Pi} \coloneqq \{ \pi \in \Pi \mid f_\mathcal{C}(\pi) \le 0 \}.$$

- The optimal policy $\pi^\star$ is defined as

$$\arg\max_{\pi \in \Pi} V_r^\pi(\rho) \quad \text{subject to} \quad f_\mathcal{C}(\pi) \le 0.$$

- Due to the diversity of safety constraint representations and little discussion on their interrelations, it is not easy to understand safe RL research systematically.

**Step 1: Problem Formulation**

$\max_{\pi \in \Pi} V_r^\pi(\rho)$ subject to [Safety Constraint]

- Typical RL objective
- Diverse Safety Constraint representations

**Step 2: Policy Optimization**

- Either use an existing algorithm suitable for the problem setup or develop a new algorithm

Figure 1. A typical sequence of safe RL based on constrained criteria.

## Theoretical Relations among Common Constraint Formulations of Safe RL

- Provide theoretical relations between each constraint representation.

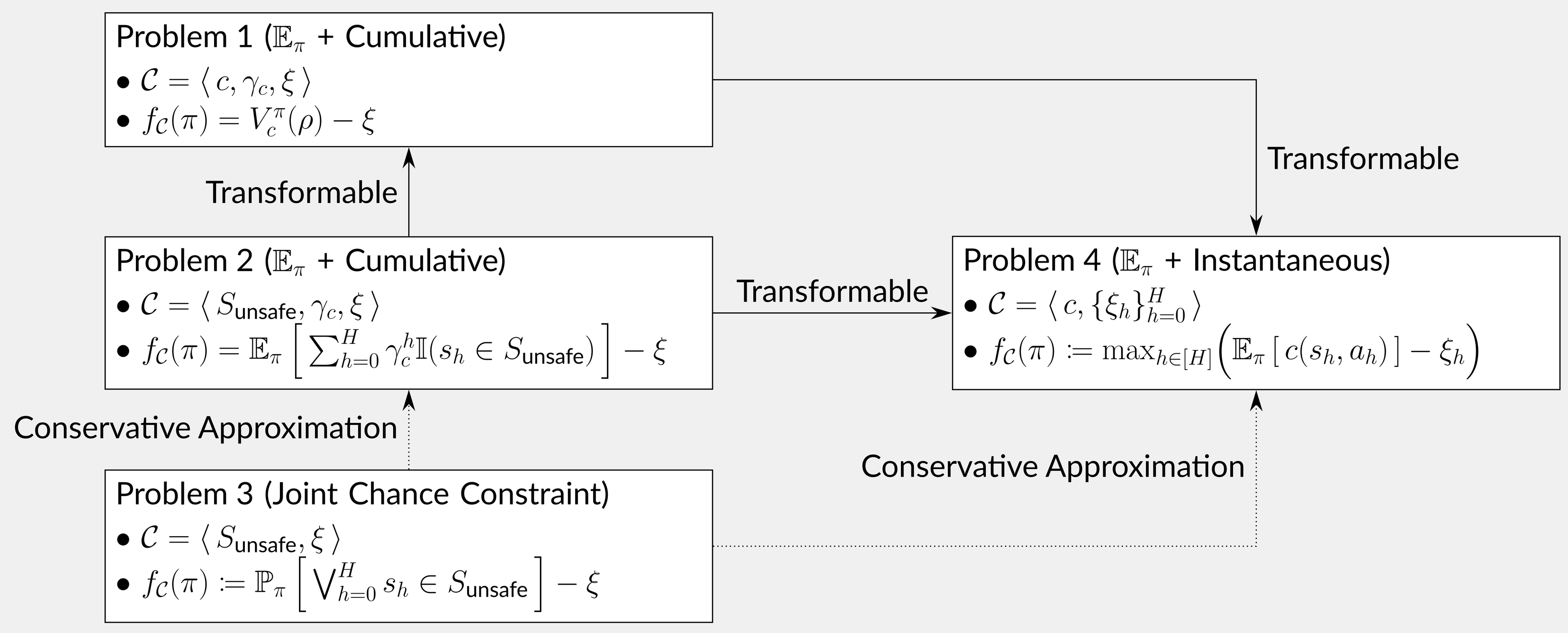


Figure 2. Relations among common safe RL formulations based on $\mathbb{E}_\pi$ and the one with chance constraints.

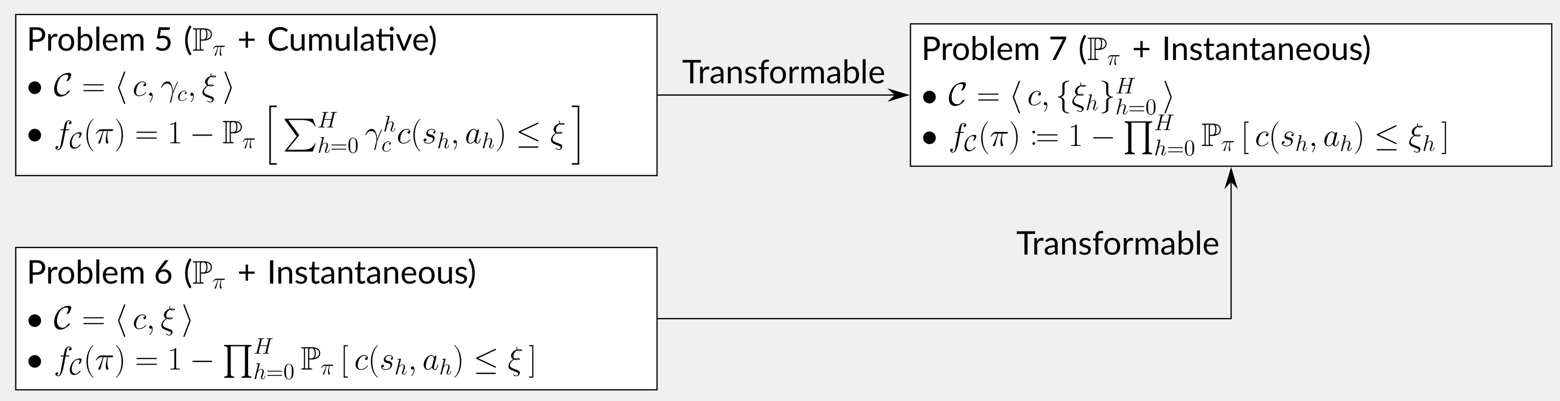


Figure 3. Relations among common safe RL formulations based on $\mathbb{P}_\pi$ (i.e., almost-surely constraints).

## Review on Diverse Constraint Formulations and Representative Existing Papers

- List of representative papers and algorithms associated with each constraint formulation of safe RL.

| Problem | Type | Representative Work | Algorithm | Theoretical Guarantee | | Open Source Software (OSS) |
|---|---|---|---|---|---|---|
| | | | | Optimality | Safety | |
| Problem 1 | Online | [Achiam *et al.*, 2017] | CPO | – | – | A, SSA, FSRL, SafePO, OmniSafe |
| | | [Ray *et al.*, 2019] | TRPO-Lagrangian | – | – | A, SSA, FSRL, SafePO, OmniSafe |
| | | | PPO-Lagrangian | – | – | A, SSA, FSRL, SafePO, OmniSafe |
| | | [Tessler *et al.*, 2019] | RCPO | – | – | A, SafePO, OmniSafe |
| | | [Liu *et al.*, 2020] | IPO | – | – | A, OmniSafe |
| | | [Yang *et al.*, 2020] | PCPO | – | – | A, SafePO, OmniSafe |
| | | [Stooke *et al.*, 2020] | PID-Lagrangian | – | – | A, SafePO, OmniSafe |
| | | [Zhang *et al.*, 2020] | FOCOPS | – | – | A, FSRL, SafePO, OmniSafe |
| | | [Ding *et al.*, 2020] | NPG-PD | Y | C | – |
| | | [Bharadhwaj *et al.*, 2021] | CSC | – | – | A |
| | | [Ding *et al.*, 2021] | OPDOP | Y | C | – |
| | | [Bai *et al.*, 2022] | CSPDA | Y | C | – |
| | | [As *et al.*, 2021] | LAMBDA | – | – | A |
| | | [Xu *et al.*, 2021] | CRPO | Y | C | OmniSafe |
| | | [Yu *et al.*, 2022] | SEditor | – | – | A |
| | | [Bura *et al.*, 2022] | DOPE | Y | T and C | – |
| | | [Liu *et al.*, 2022] | CVPO | Y | C | A, FSRL |
| | | [Zhang *et al.*, 2022] | P3O | – | – | A, OmniSafe |
| | Offline | [Le *et al.*, 2019] | CBPL | – | T and C | A |
| | | [Lee *et al.*, 2021] | COptiDICE | – | T | A, OSRL, OmniSafe |
| | | [Wu *et al.*, 2021] | CMOMDPs | Y | T and C | – |
| | | [Xu *et al.*, 2022] | CPQ | – | T | A, OSRL |
| | | [Liu *et al.*, 2023b] | CDT | – | T | A, OSRL |
| Problem 2 | Online | [Turchetta *et al.*, 2020] | CISR | – | – | A |
| | | [Thomas *et al.*, 2021] | SMBPO | – | C | A |
| | | [Thananjeyan *et al.*, 2021] | Recovery RL | – | – | A |
| | | [Wang *et al.*, 2023] | – | – | T and C | A |
| Problem 3 | Online | [Ono *et al.*, 2015] | CCDP | – | T and C | – |
| | | [Pfrommer *et al.*, 2022] | – | Y | T and C | – |
| | | [Mowbray *et al.*, 2022] | – | – | T and C | A |
| | | [Kordabad *et al.*, 2022] | – | – | T and C | – |
| Problem 4 | Online | [Pham *et al.*, 2018] | OptLayer | – | T and C | A |
| | | [Amani *et al.*, 2021] | SLUCB | Y | T and C | – |
| | | [Zhao *et al.*, 2023a] | SCPO | Y | C | – |
| | Offline | [Amani and Yang, 2022] | Safe-DPVI | Y | T and C | – |
| Problem 5 | Online | [Sootla *et al.*, 2022b] | Sauté RL | Y | C | A, SafePO, OmniSafe |
| | | [Sootla *et al.*, 2022a] | Simmer RL | Y | C | A, SafePO, OmniSafe |
| Problem 6 | Online | [Turchetta *et al.*, 2016] | SafeMDP | – | T and C | A |
| | | [Berkenkamp *et al.*, 2017] | SMbRL | – | T and C | A |
| | | [Fisac *et al.*, 2018] | – | – | T and C | – |
| | | [Wachi *et al.*, 2018] | SafeExpOpt-MDP | – | T and C | A |
| | | [Dalal *et al.*, 2018] | SafeLayer | – | T and C | A |
| | | [Cheng *et al.*, 2019] | RL-CBF | – | T and C | A |
| | | [Wachi and Sui, 2020] | SNO-MDP | Y | T and C | A |
| | | [Wang *et al.*, 2023] | – | – | C | – |
| Problem 7 | Online | [Shi *et al.*, 2023] | LSVI-NEW | Y | T and C | – |
| | | [Wachi *et al.*, 2023] | MASE | Y | T and C | – |

Figure 4. Common safe RL formulations based on the constrained criterion and associated representative work. Y = Yes, T = Training, C = After Convergence, and A = Authors' Implementation. Please see the paper for more details.



akifumi.wachi@lycorp.co.jp