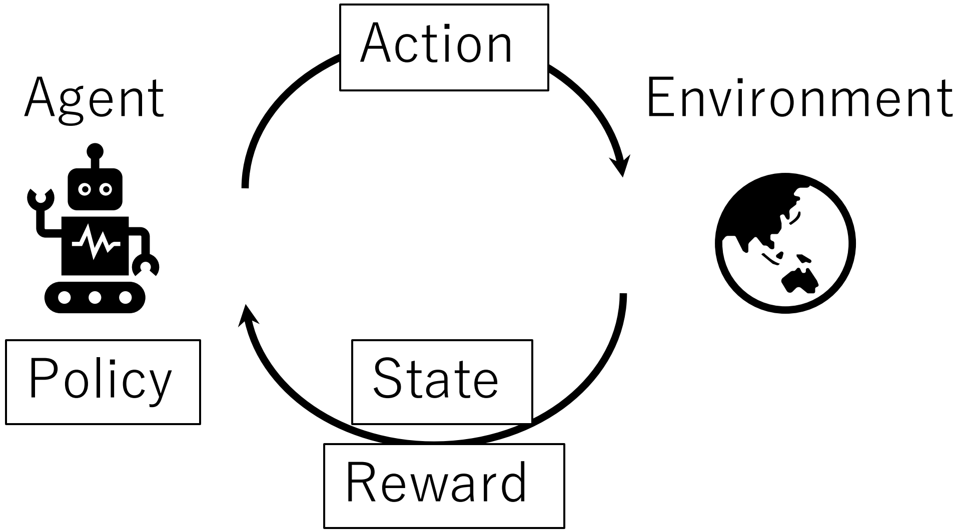
# A Survey of Constraint Formulations in Safe Reinforcement Learning

Akifumi Wachi, Xun Shen, Yanan Sui

# Reinforcement Learning (RL)





AlphaGo
(Google DeepMind)


Gran Turismo
(Sony AI)

ChatGPT

RLHF (OpenAI)

# Safety Issues in RL

Gap

**Research on RL**


AlphaGo
(Google DeepMind)


Gran Turismo
(Sony AI)

Unsafe actions → No problem ☺

**Real Applications**

Medical
Applications

Autonomous Driving

Unsafe actions → Catastrophic results ☹

⇩

Safe RL is needed!!

# Safe RL in This Talk

- Safe RL is a broad topic by definition.

- Garcia and Fernández (2015) classified optimization criteria into 4 groups:
  1. Constrained criterion
  2. Worst-case criterion
  3. Risk-sensitive criterion
  4. Others (e.g., r-squared, value-at-risk)

- This talk focuses on safe RL based on the constrained criterion.

Garcıa and Fernández. "A comprehensive survey on safe reinforcement learning." *JMLR* 16.1 (2015): 1437-1480.

# Safe RL with Constrained Criterion

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

Typical RL objective
(Expected cumulative reward)

# Potential Applications of Safe RL

Industrial Robot

Medical

Autonomous Driving


Chatbot


Space Exploration

# Diverse Required Safety Levels


Medium                Required Safety Level                High

# Diverse Constraint Formulations

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

Typical RL objective

Diverse constraint formulations depending on the applications or required safety levels

*Expectation* or *Almost surely*

*Cumulative* or *Instantaneous*

# Common Constraint Representations

**Problem Formulation 1** (Expected Cumulative Safety Constraint)

$$\max_{\pi \in \Pi} V_r^\pi(\rho)$$

subject to

$$V_c^\pi(\rho) \le \xi$$

Typical RL objective

Safety Constraint

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

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

Same Structure

$$V_{c,h}^\pi(s) := \mathbb{E}_\pi\left[\sum_{h'=h}^{H} \gamma_c^{h'} c(s_{h'}, a_{h'}) \,\middle|\, s_h = s\right]$$

$$V_c^\pi(\rho) := \mathbb{E}_{s\sim\rho}\left[V_{c,0}^\pi(s)\right]$$

# Common Constraint Representations

**<u>Problem Formulation 1</u>** (Expected Cumulative Safety Constraint)

$$\boxed{\max_{\pi \in \Pi} V_r^\pi(\rho)} \quad \text{subject to} \quad \boxed{V_c^\pi(\rho) \leq \xi}$$

Typical RL objective — Safety Constraint

- One of the most popular formulations.
  - Many well-known algorithms are based on this formulation.
  - CPO[1], {TRPO, PPO}-Lagrangian[2], RCPO[3], etc.
- Focus on the averaged performance → Relatively low required safety level

[1] Achiam+. Constrained policy optimization. In ICML, 2017.
[2] Ray+. Benchmarking safe exploration in deep reinforcement learning. arXiv preprint arXiv:1910.01708, 2019.
[3] Tessler+. Reward constrained policy optimization. In ICLR, 2019.

# Common Constraint Representations

**<u>Problem Formulation 2</u>** (Almost Surely Cumulative Safety Constraint)

$$\boxed{\max_{\pi \in \Pi} V_r^\pi(\rho)} \quad \text{subject to} \quad \boxed{\mathbb{P}_\pi\left[\sum_{h=0}^{H} \gamma_c^h c(s_h, a_h) \le \xi\right] = 1}$$

Typical RL objective

Safety Constraint

- Require the constraint satisfaction <u>with probability of 1 (i.e., almost surely)</u>.
  - $\mathbb{P}_\pi$ is used rather than $\mathbb{E}_\pi$.
  - Higher required level of safety.
- Saute RL[1] algorithm is based on this formulation.
  - Good theoretical properties + Empirical performance.

[1] Sootla+. Saute RL: Almost surely safe reinforcement learning using state augmentation. In ICML, 2022.

# Common Constraint Representations

**Problem Formulation 3** (Almost Surely Instantaneous Safety Constraint)

$$\boxed{\max_{\pi \in \Pi} V_r^\pi(\rho)} \quad \text{subject to} \quad \boxed{\mathbb{P}_\pi\left[c(s_h, a_h) \leq \xi\right] = 1, \ \forall h \in [H]}$$

Typical RL objective

Safety Constraint

- Require the constraint satisfaction with probability of 1 at every time step.
  - Very high required level of safety.
- Many algorithms are based on this formulation.
  - SMbRL[1], RL-CBF[2], SafeMDP[3], SNO-MDP[4], etc.


[1] Berkenkamp+. Safe model-based reinforcement learning with stability guarantees. In NeurIPS, 2017.
[2] Cheng+. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In AAAI, 2019
[3] Turchetta+. Safe exploration in finite Markov decision processes with Gaussian processes. In NeurIPS, 2016.
[4] Wachi and Sui. Safe reinforcement learning in constrained Markov decision processes. In ICML, 2020.

# Typical Procedure of Safe RL

**Step 1: Problem Formulation**

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

- Diverse safety constraint representations

⇩

**Step 2: Policy Optimization**

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

# Issues of Previous Safe RL Research

**Step 1: Problem Formulation**

$$\boxed{\max_{\pi \in \Pi} V_r^{\pi}(\rho)} \quad \text{subject to} \quad \boxed{\text{Safety Constraint}}$$

- Diverse safety constraint representations

$\Downarrow$

**Step 2: Policy Optimization**

- Most safe RL researches have pursued SOTA performance
- Existing survey papers have focused on <u>algorithms</u>

# Our Contributions

**Step 1: Problem Formulation**

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

- Diverse safety constraint representations

- Constraint formulation is the first step in safe RL
- Crucial to properly understand diverse constraint representations.

- **<u>Our paper provides comprehensive survey on Safe RL focusing on problem formulation in safe RL.</u>**

| 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 | – |
| | 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 | A |

Table 1: Common safe RL formulations based on the constrained criterion and associated representative work. Type indicates whether each safety RL is based on online or offline RL settings. In the Theoretical Guarantee column, **Y** indicates the (near-)optimality of the policy obtained by an algorithm. Also, **T** means that safety is guaranteed during training, and **C** means that safety is guaranteed after convergence. Note that offline algorithms are inherently safe during training since there is no interaction between the agent and the environment. In the OSS column, **A** means a public authors' implementation exists, and **SSA** is an abbreviation of the Safety Starter Agent repository (Ray *et al.* [2019], https://github.com/openai/safety-starter-agents). Also, **FSRL** (Liu *et al.* [2023a], https://github.com/liuzuxin/FSRL), **OSRL** (Liu *et al.* [2023a], https://github.com/liuzuxin/OSRL), **SafePO** (Ji *et al.* [2023], https://github.com/PKU-Alignment/Safe-Policy-Optimization), and **OmniSafe** (Ji *et al.* [2023], https://github.com/PKU-Alignment/omnisafe) are recent and actively maintained repositories for online and offline safe RL, which will lead to the ease of the process of adopting safe RL algorithms.

List of representative algorithms associated with each formulation

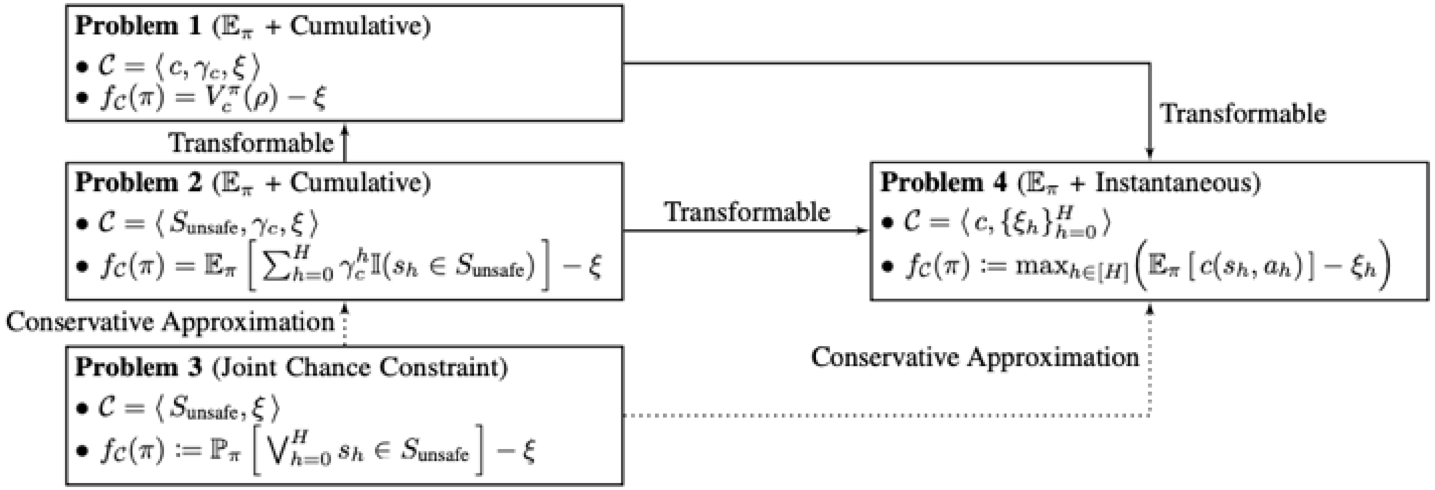



Figure 2: Relations among common safe RL formulations based on $\mathbb{E}_\pi$ and the one with chance constraints.

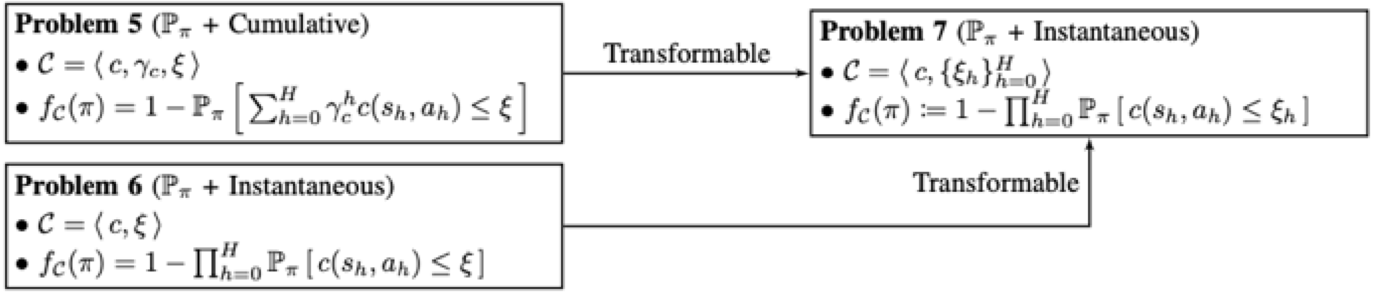



Figure 3: Relations among common safe RL formulations based on $\mathbb{P}_\pi$ (i.e., almost-surely constraints).

Theoretical relations between each constraint representation

# Conclusion (Take Home Messages)

- Safety is an important issue in RL
  - Diverse problem settings → Diverse constraint representations
- Our paper provides
  - Comprehensive survey of safe RL literature from the perspective of constraint formulations
  - Theoretical analysis on interrelations between each formulation

**Thank you!!**

**Contact: wachi.akifumi [at] gmail.com**

**Paper:** https://arxiv.org/abs/2402.02025