Flipping-based Policy for Chance-Constrained Markov Decision Processes

Xun Shen[1], Shuo Jiang[1], Akifumi Wachi[2], Kazumune Hashimoto[1], Sebastien Gros[3]

1. Osaka University

2. LY Corporation

3. Norwegian University of Technology, NTNU

December 10 – 15, 2024

NeurIPS 2024

Chance-Constrained MDPs

Maximize $\boldsymbol{\pi} \in \boldsymbol{\Pi}$

$$\mathbb{E}\left\{\sum_{k=0}^{\infty} \gamma^k r\left(\mathbf{s}_k, \mathbf{a}_k\right) \mid \mathbf{s}_0 = \mathbf{s}\right\} \qquad \mathbf{a}_k \sim \boldsymbol{\pi}\left(\mathbf{s}_k\right)$$

Subject to

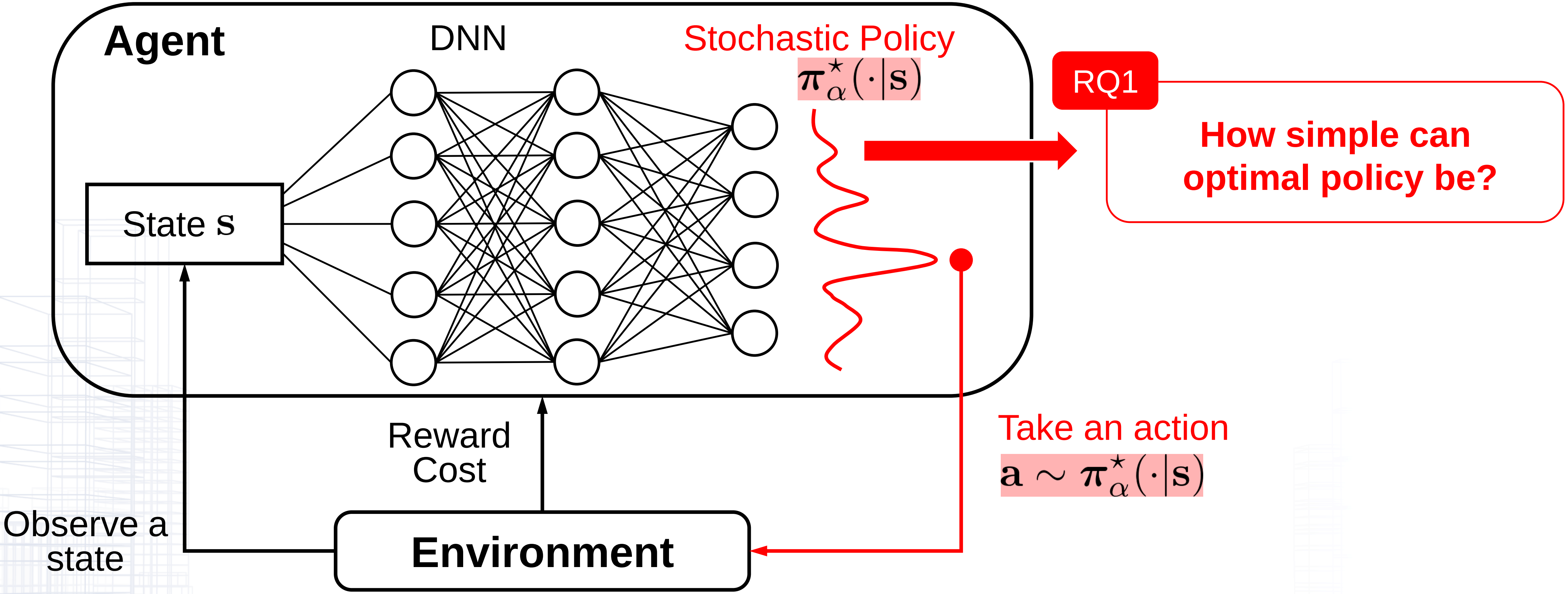
$$\mathrm{Pr}^{\boldsymbol{\pi}}_{\mathbf{s}_0,\infty}\left\{\mathbf{s}_{k+i} \in \mathbb{S}, \forall i \in [T] \mid \mathbf{s}_k \in \mathbb{S}\right\} \geq 1 - \alpha, \ \forall k$$

Research Questions

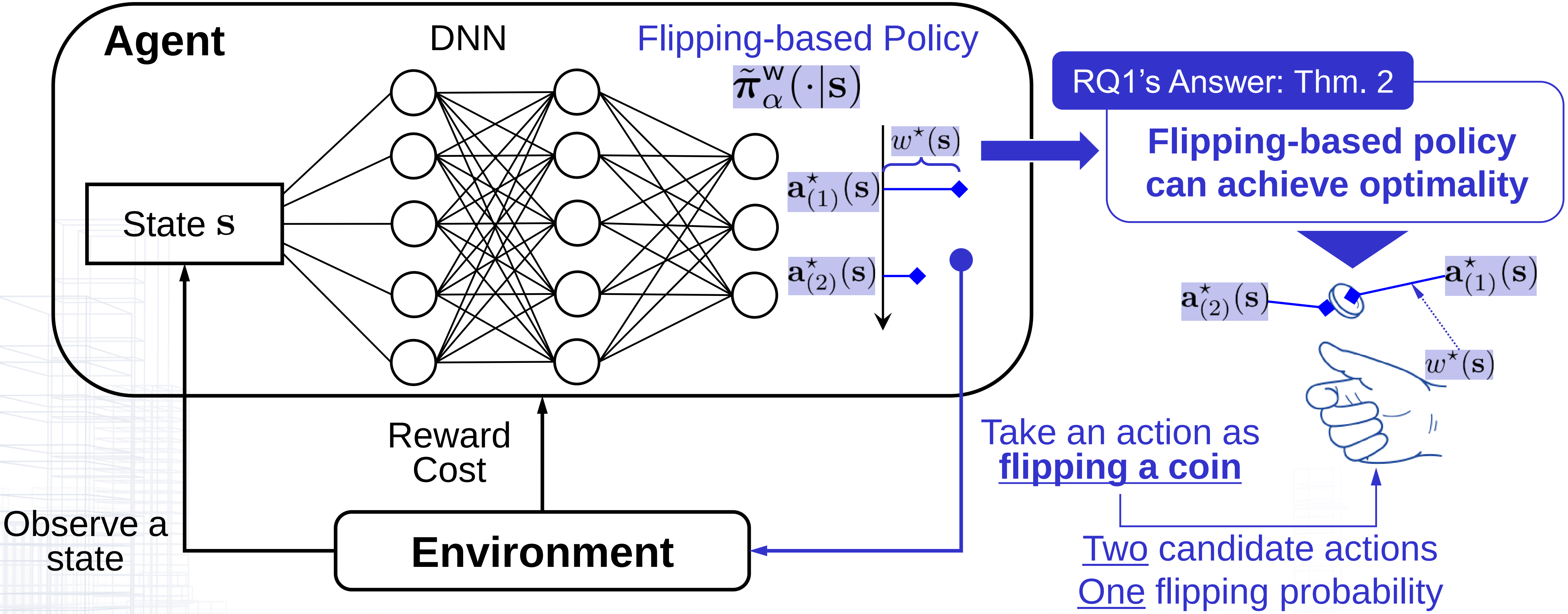
Regarding the optimal policy for CCMDPs

- RQ1: How can we **define** and **characterize** the optimal policy?
- RQ2: How can we use the existing safe RL algorithms to **effectively approximate** the optimal policy

Stochastic Policy



Flipping-based Policy



Conservative Approximation

大阪大学 OSAKA UNIVERSITY

Maximize $\boldsymbol{\pi} \in \boldsymbol{\Pi}$

$$\mathbb{E}\left\{\sum_{k=0}^{\infty} \gamma^k r\left(\mathbf{s}_k, \mathbf{a}_k\right) \mid \mathbf{s}_0 = \mathbf{s}\right\} \qquad \mathbf{a}_k \sim \boldsymbol{\pi}\left(\mathbf{s}_k\right)$$

Subject to

$$\Pr^{\boldsymbol{\pi}}_{\mathbf{s}_0,\infty}\left\{\mathbf{s}_{k+i} \in \mathbb{S}, \forall i \in [T] \mid \mathbf{s}_k \in \mathbb{S}\right\} \geq 1-\alpha, \ \forall k$$

- Conservative Approximation: Thm. 5
- Flipping-based policy can also achieve optimality: Thm. 4

Subject to

$$\mathbb{E}\left\{\sum_{i=1}^{\infty} \gamma_{\text{unsafe}}^{i} \mathbb{I}\left(\mathbf{s}_{k+i} \notin \mathbb{S}\right) \mid \mathbf{s}_k \in \mathbb{S}\right\} \leq \alpha, \ \forall k$$

Intractable ← **Joint Chance Constraint**

CPO, PCPO, CUP, P3O, …… ← **Expected Cumulative Safety Constraint**

Train Flipping-based Policy

Step 1. Construct the sample set of risk levels

$$\mathcal{Z}_S = \{\tilde{\alpha}_i\}_{i=1}^{S}, \ \tilde{\alpha}_i \sim \mathcal{U}(0,1)$$

Step 2. Optimize a policy parameter $\tilde{\boldsymbol{\theta}}_i$, by solving

$$\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} J(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\tau}_\infty \sim \boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathsf{d}}} \{R(\boldsymbol{\tau}_\infty)\} \quad \text{s.t.} \quad F^{\mathsf{d}}(\boldsymbol{\theta}) \leq \tilde{\alpha}_i.$$

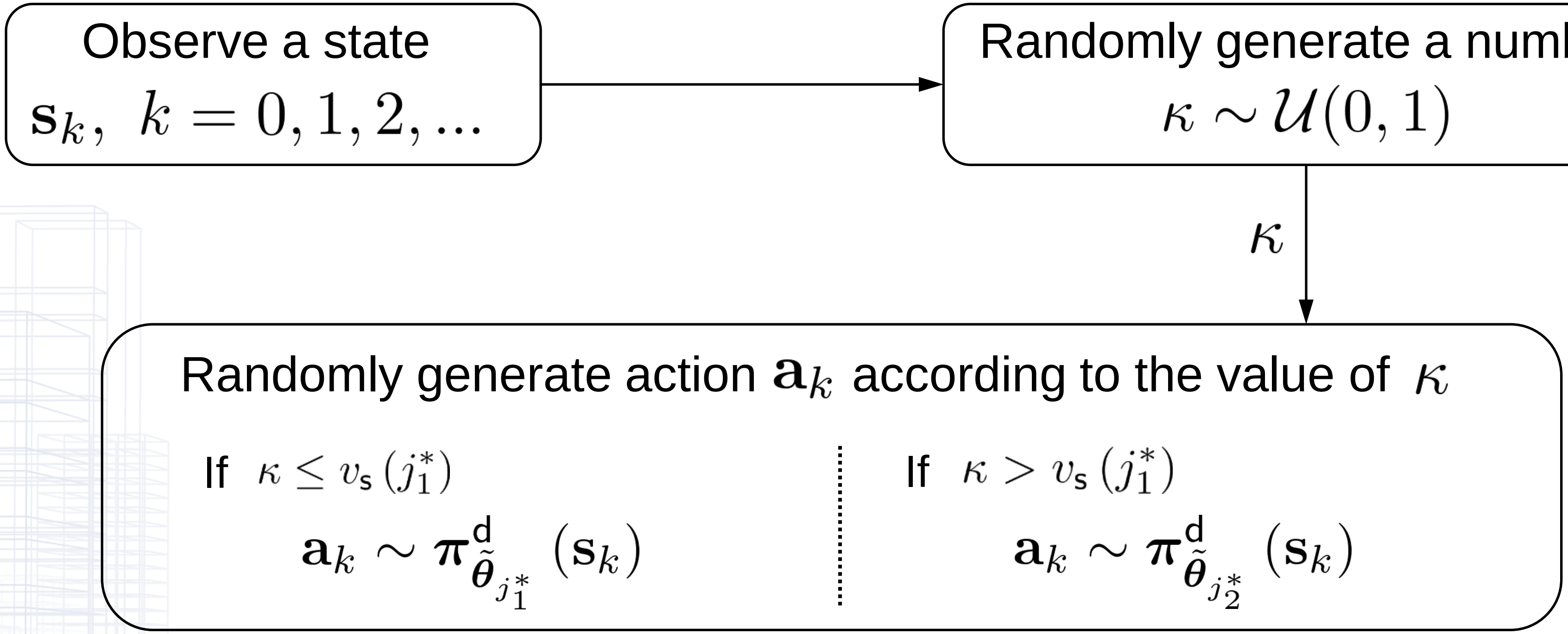
$$J(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{s} \sim \mu_0} \left\{ \mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathsf{d}}} \left\{ \sum_{k=0}^{\infty} \gamma^k r(\mathbf{s}_k, \mathbf{a}_k) \mid \mathbf{s}_0 \right\} \right\} \quad F^{\mathsf{d}}(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{s} \sim \mu_0} \left\{ \mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathsf{d}}} \left\{ \sum_{i=1}^{\infty} \gamma_{\mathsf{unsafe}}^{i} \mathbb{I}(\mathbf{s}_{k+i} \notin \mathbb{S}) \mid \mathbf{s}_0 \right\} \right\}$$

Step 3. Solve a linear program to obtain parameters $\left(\nu_{\mathsf{s}}(j_1^*), \nu_{\mathsf{s}}(j_2^*), \tilde{\boldsymbol{\theta}}_{j_1^*}, \tilde{\boldsymbol{\theta}}_{j_2^*}\right)$ for flipping-based policy

$$\max_{\nu_{\mathsf{s}}(1),\dots,\nu_{\mathsf{s}}(S) \in [0,1]^S} \sum_{i=1}^{S} J(\tilde{\boldsymbol{\theta}}_i)\nu_{\mathsf{s}}(i) \quad \text{s.t.} \quad \sum_{i=1}^{S} \nu_{\mathsf{s}}(i) F^{\mathsf{d}}(\tilde{\boldsymbol{\theta}}_i) \geq 1 - \alpha, \ \sum_{i=1}^{S} \nu_{\mathsf{s}}(i) = 1.$$

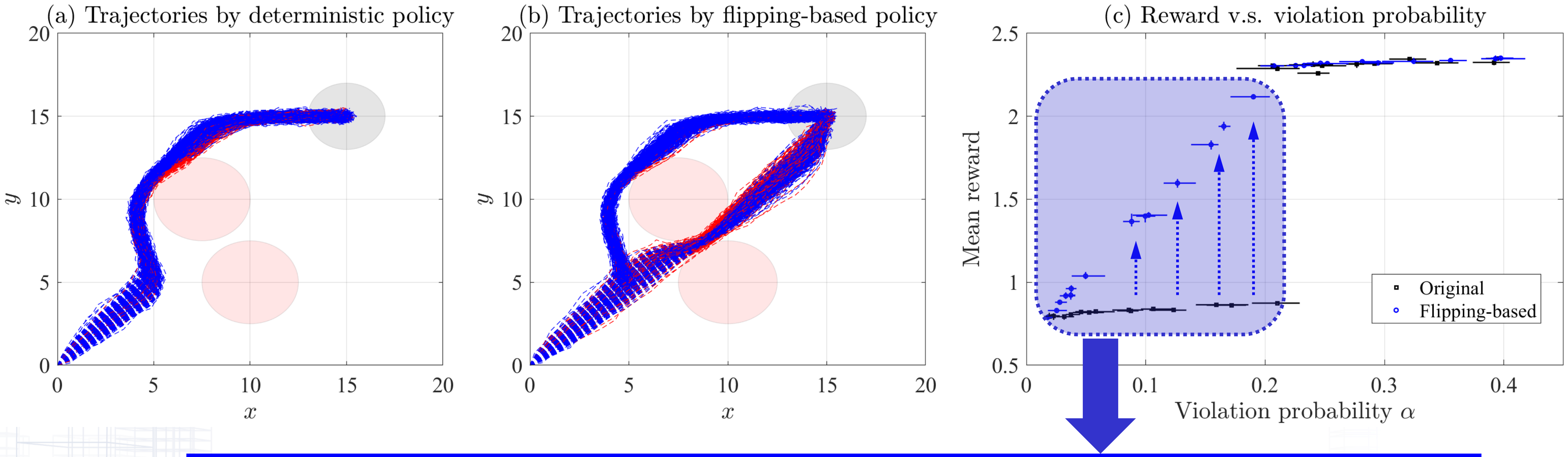
Optimal solution has two non-zero element

Implement Flipping-based Policy



Numerical Example

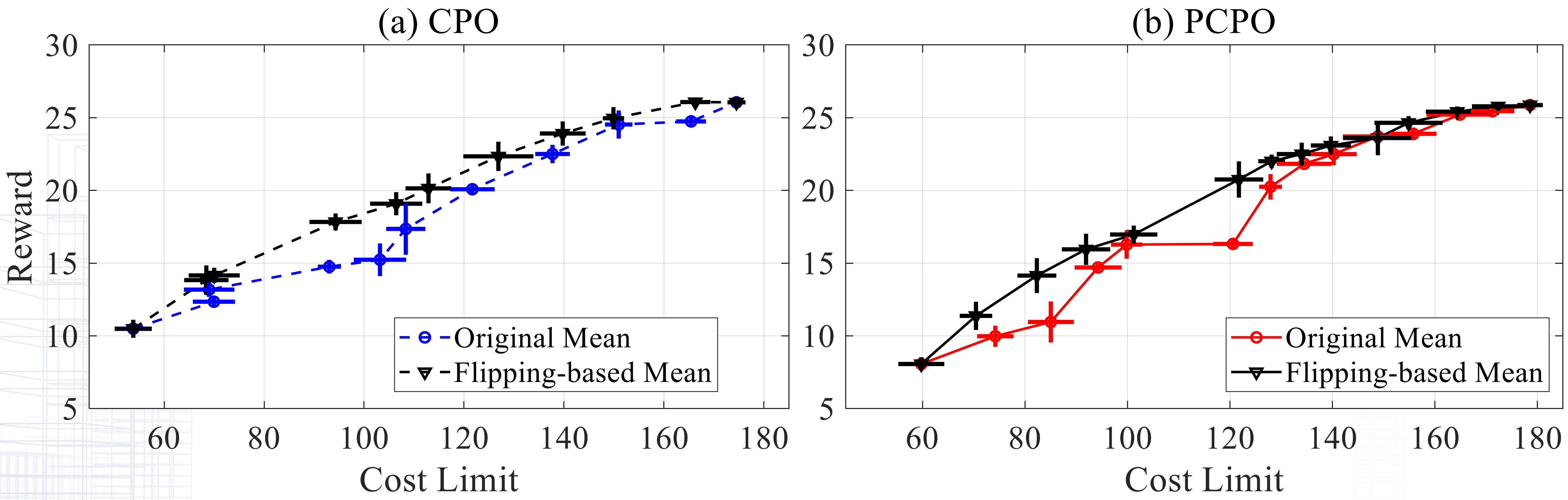
- Intuitive example of trajectory planning and control



Improve average reward through the linear combination of risks

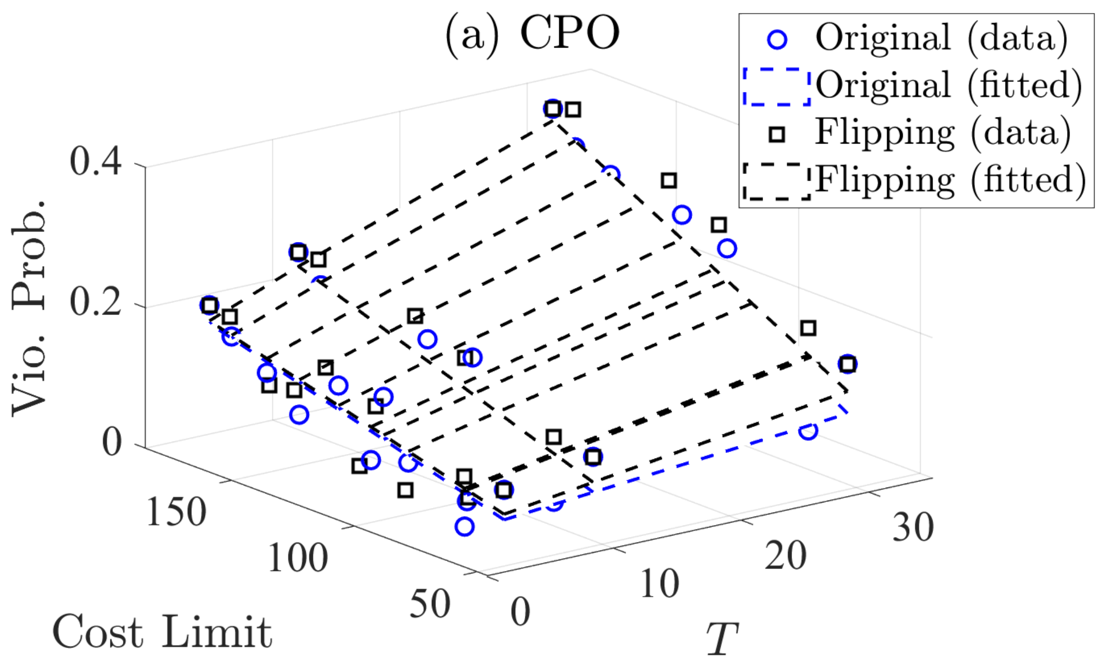
Experimental Validation

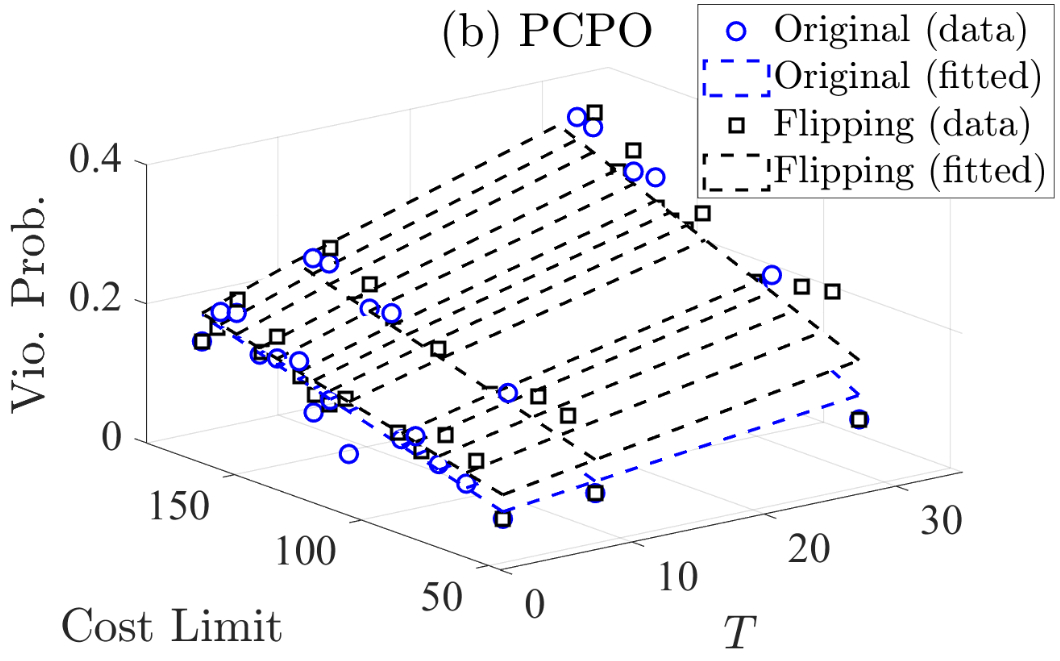
- Enhance existing safe RL algorithms (e.g., CPO, PCPO)
- Increase the expected reward under the required level of risk



Experimental Validation

- Expected cumulative safety (average cost) v.s. violation probability





Thanks for your kind attention!