Robot Learning

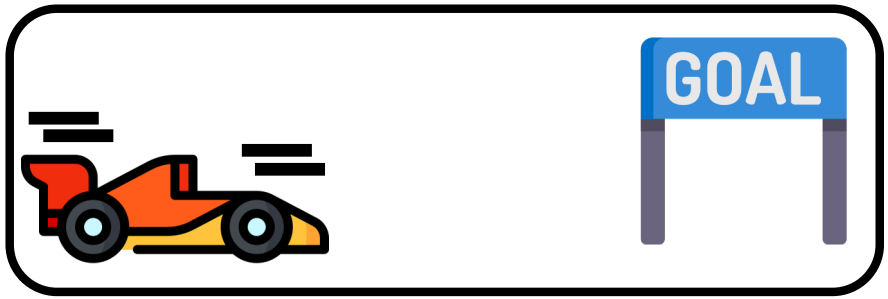
Safe Reinforcement Learning

Prof. Songhwai Oh
ECE, SNU

SAFE REINFORCEMENT LEARNING (SAFE RL)

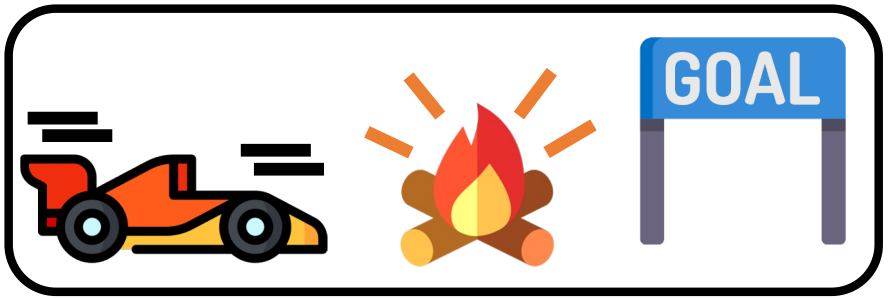
Safe Reinforcement Learning

General RL



$$\text{maximize}_\pi \sum_{t=0}^{\infty} \gamma^t R_t$$

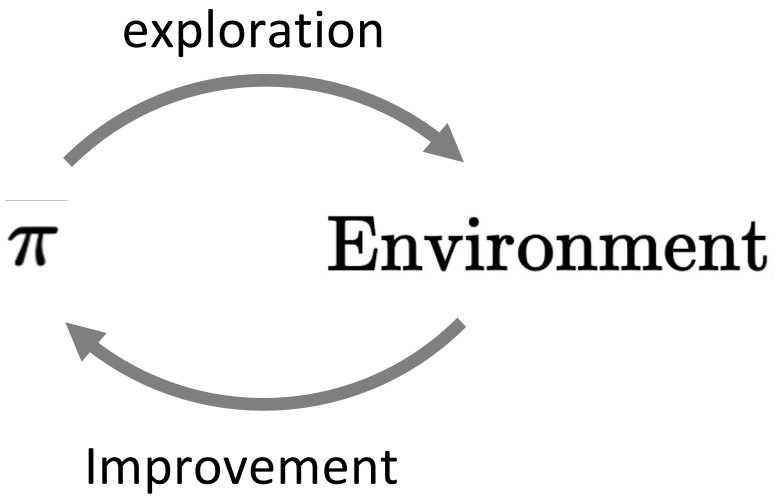
Safe RL



$$\text{maximize}_\pi \sum_{t=0}^{\infty} \gamma^t R_t \quad s.t.\ \mathbf{S}(\sum_{t=0}^{\infty} \gamma^t C_t) \leq d.$$

Approaches

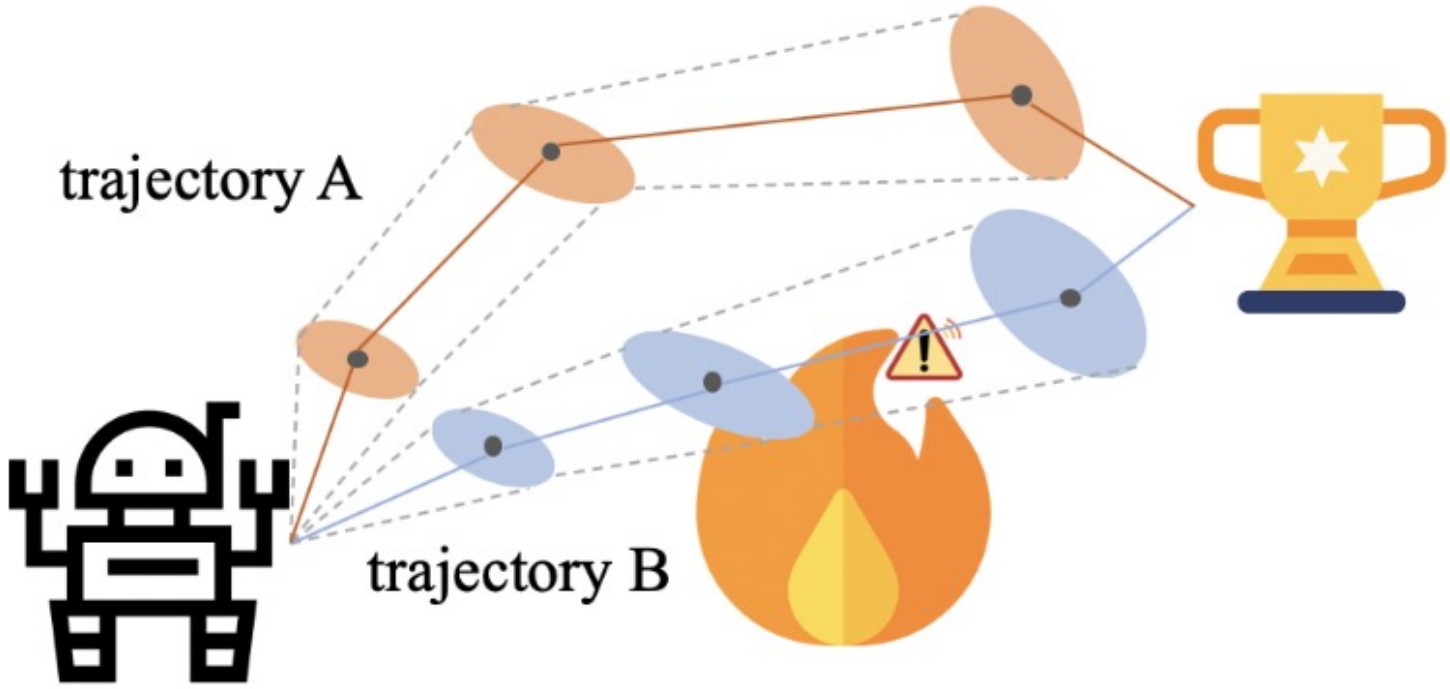
- Safe exploration
- Safe policy improvement



Objective of safe reinforcement learning (RL) :=

Maximize returns while satisfying constraints

Reinforcement Learning **Safety**



Source: Zuxin Liu, Hongyi Zhou, Baiming Chen, Sicheng Zhong, Martial Hebert, Ding Zhao, "Constrained Model-based Reinforcement Learning with Robust Cross-Entropy Method." *arXiv preprint arXiv:2010.07968* (2020).

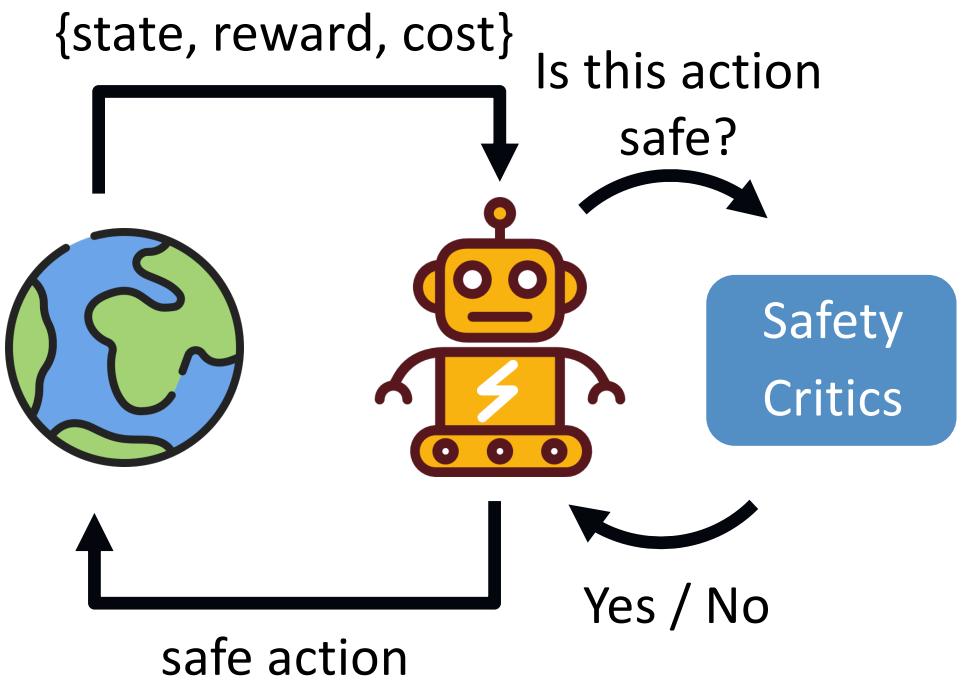
Safe RL Methods

Garcia and Fernández have classified safe RL into two categories [1].

- Exploration Process
  - Handle constraints in a way that transforms actions into a safe action during exploration
- Optimization Criteria
  - Handle constraints in a way that update policies safely using constrained optimization
  - It is divided according to what constraints are used and how the policy is updated.

[1] Garcıa, Javier, and Fernando Fernández. "A comprehensive survey on safe reinforcement learning." *Journal of Machine Learning Research* 16.1 (2015): 1437-1480.

Safe RL: Exploration Process



- Recovery-RL [1]
  - If the output of the safety critic is above a threshold, execute a recovery policy instead.
- Conservative Safety Critic [2]
  - Sample an action until the output of the safety critic is above a threshold and execute it.

Safety Critics:

$$Q_C^{\pi}(s,a) := \mathbb{E}\left[\sum_t \gamma^t C(s_t,a_t) | s_0 = s, a_0 = a\right]$$

[1] B. Thananjeyan, et. al, "Recovery RL: Safe reinforcement learning with learned recovery zones," IEEE Robot. Automat. Lett., vol. 6, no. 3, pp. 4915–4922, 2021.

[2] H. Bharadhwaj, et. al, "Conservative safety critics for exploration," in Proc. Int. Conf. Learn. Representations, 2020.

Safe RL: Exploration Process

- **Shielding method**: costs are predicted by safety monitoring modules such as safety critics, and actions are resampled or transformed to safe actions
  - Bharadhwaj, Homanga, et al. "Conservative safety critics for exploration." arXiv preprint. 2020.
  - Srinivasan, Krishnan, et al. "Learning to be safe: Deep RL with a safety critic." arXiv preprint. 2020.
  - Thananjeyan, Brijen, et al. "Recovery RL: Safe reinforcement learning with learned recovery zones." IEEE Robotics and Automation Letters. 2021.
- **Model-based method**: explicit transition models are trained and used to judge whether an action is unsafe during action planning
  - Cowen-Rivers, Alexander I., et al. "Samba: Safe model-based & active reinforcement learning." arXiv preprint. 2020.
  - Ohnishi, Motoya, et al. "Barrier-certified adaptive reinforcement learning with applications to brushbot navigation." IEEE Transactions on Robotics. 2019.
- **Uncertainty modeling using Gaussian process regression**: to safely explore, reduce the action space using uncertainty
  - Wachi, Akifumi, et al. "Safe exploration and optimization of constrained MDPs using Gaussian processes." AAAI Conference on Artificial Intelligence. 2018.
  - Kuo, Cheng-Yu, et al. "Uncertainty-aware contact-safe model-based reinforcement learning." IEEE Robotics and Automation Letters. 2021.

Safe RL: Optimization Criteria

- Policy Improvement Type
  - Lagrangian method
    - Convert the constrained problem into a dual problem using Lagrange multipliers
    - Update the policy and multipliers concurrently
  - Trust region method
    - Approximate the constrained problem within a trust region as linear-quadratic programming

- Safety Constraint Type
  - Expectation constraint
    - Expectations on stochastic variables are set as constraints
  - Risk measure constraint
    - Risk measures such as conditional value at risk (CVaR) are set as constraints.
  - Chance constraint
    - Probability of failure such as value at risk (VaR) is set as constraints.

Safe RL: Optimization Criteria

- Optimization Criterion
  - Expectation constraint
    - Expectations on a stochastic variable are set as constraints.
      - Achiam, Joshua, et al. "Constrained policy optimization." International Conference on Machine Learning. PMLR, 2017.
      - Ding, Dongsheng, et al. "Provably efficient safe exploration via primal-dual policy optimization." International Conference on Artificial Intelligence and Statistics. PMLR, 2021.
      - Liu, Yongshuai, et al. "IPO: Interior-point policy optimization under constraints." Proceedings of the AAAI Conference on Artificial Intelligence. 2020.
  - Risk measure constraint
    - Risk measures such as conditional value at risk (CVaR) are set as constraints.
      - Zhang, Jianyi, and Paul Weng. "Safe Distributional Reinforcement Learning." arXiv preprint, 2021.
      - Yang, Qisong, et al. "WCSAC: Worst-case soft actor critic for safety-constrained reinforcement learning." Proceedings of the AAAI Conference on Artificial Intelligence. 2021.
  - Chance constraint
    - Probability of failure such as value at risk (VaR) is set as constraints.
      - Chow, Yinlam, et al. "Risk-constrained reinforcement learning with percentile risk criteria." The Journal of Machine Learning Research. 2017.

Safe RL: Optimization Criteria

- Optimization Criterion
  - Expectation constraint
    - Expectations on a stochastic variable are set as constraints.
      - Achiam, Joshua, et al. "Constrained policy optimization." International Conference on Machine Learning. PMLR, 2017.
      - Ding, Dongsheng, et al. "Provably efficient safe exploration via primal-dual policy optimization." International Conference on Artificial Intelligence and Statistics. PMLR, 2021.
      - Liu, Yongshuai, et al. "IPO: Interior-point policy optimization under constraints." Proceedings of the AAAI Conference on Artificial Intelligence. 2020.
  - Risk measure constraint
    - Risk measures such as conditional value at risk (CVaR) are set as constraints.
      - Zhang, Jianyi, and Paul Weng. "Safe Distributional Reinforcement Learning." arXiv preprint, 2021.
      - Yang, Qisong, et al. "WCSAC: Worst-case soft actor critic for safety-constrained reinforcement learning." Proceedings of the AAAI Conference on Artificial Intelligence. 2021.
  - Chance constraint
    - Probability of failure such as value at risk (VaR) is set as constraints.
      - Chow, Yinlam, et al. "Risk-constrained reinforcement learning with percentile risk criteria." The Journal of Machine Learning Research. 2017.

– policy update with trust-region methods
– policy update with Lagrangian methods
– policy update with interior-point methods

Dohyeong Kim and Songhwai Oh, "**TRC: Trust Region Conditional Value at Risk for Safe Reinforcement Learning**," IEEE Robotics and Automation Letters, vol. 7, no. 2, pp. 2621-2628, Apr. 2022.

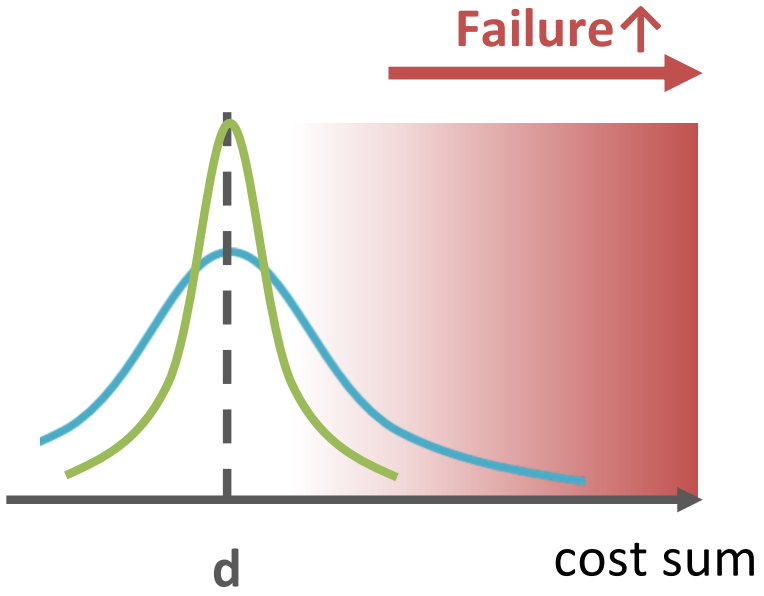
TRC: TRUST REGION CONDITIONAL VALUE AT RISK FOR SAFE REINFORCEMENT LEARNING

Safe RL Objective

- Train an RL agent while satisfying safety constraints to avoid catastrophic failure.
- Expectation-based constraints are widely used in safe RL, but it is hard to reduce the probability of the failure.

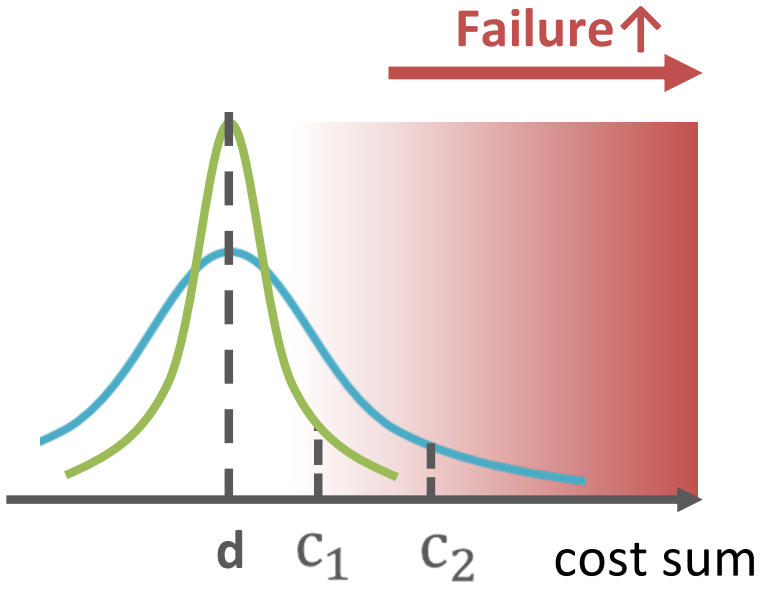
  → **Why?**

  - Expectation constraint:
    - Two distributions of cost sums have the same expectation value d.
    - However, the expectation-based constraint **cannot recognize that the green distribution is more desirable.**



How can resolve the issue?

- Use a risk measure for the safety constraints.
- Conditional value at risk (CVaR), one of the risk measures, is the conditional expectation of a stochastic variable above a certain percentile level.
  - CVaR-based constraint:
    - C_1 is the CVaR value of the green distribution and C_2 is the CVaR value of the blue distribution.
    - **CVaR can distinguish which distributions have a lower probability of failure.**



How to update a policy to satisfy the CVaR-based constraints?

- Option 1: Lagrangian method
  - The Lagrangian method relaxes a constrained problem to an unconstrained problem and is widely used in safe RL.
  - It updates the policy and the Lagrange multipliers concurrently.
  - **However, oscillations of Lagrange multipliers make training unstable.**
- Option 2: Trust-region method
  - Can handle constraints without using additional variables.
  - **Need to estimate CVaR values of any policies within the trust region.**

→ We propose to use an approximation of CVaR by deriving the **upper bound of the CVaR** within the trust region.

Constrained Markov decision process (CMDP)

- Defined as a tuple $(\mathcal{S}, \mathcal{A}, \rho, \mathcal{P}, R, C, \gamma)$.
  - State space: $\mathcal{S} \subset \mathbb{R}^n$
  - Action space: $\mathcal{A} \subset \mathbb{R}^m$
  - Initial state distribution: $\rho : \mathcal{S} \mapsto \mathbb{R}$
  - Transition model: $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto \mathbb{R}$
  - Reward function: $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto \mathbb{R}$
  - Cost function: $C : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto \mathbb{R}_{\geq 0}$
  - Discount factor: $\gamma \in [0, 1)$
- Using the cost function, define the safety constraint.

E.g., expectation-based constraint: $\mathop{\mathbb{E}}_{\rho,\pi,\mathcal{P}} \left[\sum_{t=0}^{\infty} \gamma^t C(s_t, a_t, s_{t+1}) \right] \leq d.$

Constrained Markov decision process (CMDP)

- Discounted cost sum:

$$C_\pi := \sum_{t=0}^{\infty} \gamma^t C(s_t, a_t, s_{t+1}) | \pi, \mathcal{P}, s_0 \sim \rho$$

$$C_\pi(s) := \sum_{t=0}^{\infty} \gamma^t C(s_t, a_t, s_{t+1}) | \pi, \mathcal{P}, s_0 = s,$$

$$C_\pi(s, a) := \sum_{t=0}^{\infty} \gamma^t C(s_t, a_t, s_{t+1}) | \pi, \mathcal{P}, s_0 = s, a_0 = a.$$

- Value and action value function:

$$V_C^\pi(s) := \mathbb{E}_{\pi,\mathcal{P}} \left[\sum_{t=0}^{\infty} \gamma^t C(s_t, a_t, s_{t+1}) | s_0 = s \right],$$

$$Q_C^\pi(s, a) := \mathbb{E}_{\pi,\mathcal{P}} \left[\sum_{t=0}^{\infty} \gamma^t C(s_t, a_t, s_{t+1}) | s_0 = s, a_0 = a \right],$$

$$A_C^\pi(s, a) := Q_C^\pi(s, a) - V_C^\pi(s).$$

- Discounted state distribution:

$$d^\pi(s) := (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \text{Prob}(s_t = s | \pi)$$

- **Doubly discounted state distribution*:**

$$d_2^\pi(s) := (1-\gamma^2) \sum_{t=0}^{\infty} \gamma^{2t} \text{Prob}(s_t = s | \pi)$$

- **Cost square value function*:**

$$S_C^\pi(s) := \mathbb{E}_{\pi,\mathcal{P}} \left[C_\pi(s_0)^2 | s_0 = s \right],$$

$$S_C^\pi(s, a) := \mathbb{E}_{\pi,\mathcal{P}} \left[C_\pi(s_0, a_0)^2 | s_0 = s, a_0 = a \right].$$

* These are newly introduced in the proposed method.

Conditional Value at Risk (CVaR)

$$\mathrm{CVaR}_\alpha(X) = \mathbb{E}[X | X \geq \mathrm{ICDF}(1-\alpha)],$$

where α is a risk level and ICDF is a inverse cumulative distribution function

- One of the representative risk measures used to analyze the tails of distributions in financial portfolios or management [1].

- With an assumption that the discounted cost sum follows a Gaussian distribution, the CVaR of discounted cost sum can be approximated as:

$$\mathrm{CVaR}_\alpha(C_\pi) \approx J_C(\pi) + \frac{\phi(\Phi^{-1}(\alpha))}{\alpha} \sqrt{J_S(\pi) - J_C(\pi)^2},$$

where $J_C(\pi) := \mathbb{E}_{s\sim\rho}\left[V_C^\pi(s)\right]$ and $J_S(\pi) := \mathbb{E}_{s\sim\rho}\left[S_C^\pi(s)\right]$.

[1] R. T. Rockafellar, S. Uryasev, et al., "Optimization of conditional value-at-risk," Journal of risk, vol. 2, pp. 21–42, 2000.

Proposed Method

CVaR-constrained safe RL problem

$$\underset{\pi}{\text{maximize}} \underset{\rho,\pi,\mathcal{P}}{\mathbb{E}} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1}) \right]$$
$$\text{s.t.} \quad \text{CVaR}_\alpha(C_\pi) \leq d/(1-\gamma)$$

- To use the trust region method for policy update, we need to estimate a CVaR of any policy within the trust region.

Cost Surrogates

- Before deriving the upper bound of CVaR, the following surrogates are defined.

$$J_C^\pi(\pi') := J_C(\pi) + \frac{1}{1-\gamma} \underset{\substack{s\sim d^\pi \\ a\sim\pi}}{\mathbb{E}} \left[\frac{\pi'(a|s)}{\pi(a|s)} A_C^\pi(s,a) \right],$$
$$J_S^\pi(\pi') := J_S(\pi) + \frac{1}{1-\gamma^2} \underset{\substack{s\sim d_2^\pi \\ a\sim\pi}}{\mathbb{E}} \left[\frac{\pi'(a|s)}{\pi(a|s)} A_S^\pi(s,a) \right],$$
$$D^\pi(\pi') := \max_s D_{\text{TV}}(\pi' || \pi)[s],$$

Proposed Method

Upper bound of CVaR

Theorem 2. *For any policies* π *and* π'*, let define* $\epsilon_C^{\pi'} := \max_s \mathbb{E}_{a\sim\pi'}[A_C^\pi(s,a)]$ *and*

$$\epsilon_{\text{CVaR}}^{\pi'} := \epsilon_S^{\pi'} + \left(J_C^\pi(\pi') - \frac{\gamma\epsilon_C^{\pi'}}{(1-\gamma)^2} D^\pi(\pi') \right) \frac{2\gamma(1+\gamma)}{1-\gamma} \epsilon_C^{\pi'}.$$

Then, the following inequality holds:

$$\text{CVaR}_\alpha(C_{\pi'}) \leq J_C^\pi(\pi') + \frac{\phi(\Phi^{-1}(\alpha))}{\alpha} \sqrt{J_S^\pi(\pi') - J_C^\pi(\pi')^2} + \frac{2}{1-\gamma} \left(\frac{\gamma\epsilon_C^{\pi'}}{1-\gamma} + \frac{\phi(\Phi^{-1}(\alpha))/\alpha}{\sqrt{J_S(\pi') - J_C(\pi')^2}} \frac{\epsilon_{\text{CVaR}}^{\pi'}}{1+\gamma} \right) D^\pi(\pi'),$$

where equality holds if $\pi = \pi'$.

- Assuming that π' exists in the trust region of π so that $D^\pi(\pi')$ is negligible, Theorem 2 gives a differentiable upper bound of CVaR.

Policy update in the trust region

- Replace the CVaR part in the CVaR-constrained safe RL problem with the upper bound derived from Theorem 2 and add a trust region constraint as in [1, 2].
- The proposed CVaR-constrained subproblem can be written as below.

$$\begin{aligned} \underset{\pi}{\text{maximize}} \quad & \mathop{\mathbb{E}}_{\substack{s\sim d^{\pi_{\text{old}}} \\ a\sim\pi_{\text{old}}}} \left[\frac{\pi(a|s)}{\pi_{\text{old}}(a|s)} A^{\pi_{\text{old}}}(s,a) \right] \\ \text{s.t.} \quad & J_C^{\pi_{\text{old}}}(\pi) + \frac{\phi(\Phi^{-1}(\alpha))}{\alpha} \sqrt{J_S^{\pi_{\text{old}}}(\pi) - J_C^{\pi_{\text{old}}}(\pi)^2} \leq \frac{d}{1-\gamma}, \\ & \mathop{\mathbb{E}}_{s\sim d_{\pi_{\text{old}}}} [D_{\text{KL}}(\pi_{\text{old}}||\pi)[s]] \leq \delta. \end{aligned} \tag{15}$$

- Find the policy update direction by approximating the objective and the CVaR constraint linearly and the KL divergence quadratically.
- Update the policy using a line search method.

[1] Schulman, John, et al. "Trust region policy optimization." International conference on machine learning. PMLR, 2015.
[2] Achiam, Joshua, et al. "Constrained policy optimization." International Conference on Machine Learning. PMLR, 2017.

Generalized advantage estimations (GAEs) for CVaR

- GAEs can control the trade-off between bias and variance of the advantage function.
- To apply the GAEs to the CVaR-constrained problem, We derive the GAEs for the k-step cost square value functions as below.

$$
\begin{aligned}
A_{S,t}^{(k+1)} - A_{S,t}^{(k)} &= \mathop{\mathbb{E}}_{\pi,\mathcal{P}} \Big[2\left(C_{t+k} + \gamma V_{C,t+k+1}^{\pi} - V_{C,t+k}^{\pi}\right) \sum_{i=t}^{t+k-1} \gamma^{k+i-t} C_i + \\
&\quad \gamma^{2k} \left(C_{t+k}^2 + 2\gamma C_{t+k} V_{C,t+k+1}^{\pi} + \gamma^2 S_{C,t+k+1}^{\pi} - S_{C,t+k}^{\pi}\right) \Big] \\
&= \gamma^{2k} \mathop{\mathbb{E}}_{\pi,\mathcal{P}} \left[C_{t+k}^2 + 2\gamma C_{t+k} V_{C,t+k+1}^{\pi} + \gamma^2 S_{C,t+k+1}^{\pi} - S_{C,t+k}^{\pi} \right]
\end{aligned}
$$

$$
\delta_t^S := C_{t+k}^2 + 2\gamma C_{t+k} V_{C,t+k+1}^{\pi} + \gamma^2 S_{C,t+k+1}^{\pi} - S_{C,t+k}^{\pi}
$$

$$
\hat{A}_{S,t}^{\mathrm{GAE}(\gamma,\lambda)} := \sum_{i=t}^{\infty} (\gamma^2\lambda)^{i-t} \delta_i^S
$$

→ GAEs for cost square value.

Proposed Method

Algorithm 1 TRC

Input: Initial policy network $\pi(a|s;\theta)$, value network $V^\pi(s;\phi)$, cost value network $V_C^\pi(s;\phi_C)$, and cost square network $S_C^\pi(s;\psi_C)$.

for epochs=1, P **do**
  Initialize trajectory memory D.
  for episodes=1, E **do**
    Get an initial state s_0 from the environment.
    for t=0, T **do**
      Sample an action $a_t \sim \pi(\cdot|s_t;\theta)$.
      Feed the action a_t to the environment.
      Get reward r_t, cost c_t, and next state. s_{t+1}, and store $(s_t, a_t, r_t, c_t, s_{t+1})$ in D.
    end for
  end for
  Calculate $\hat{A}^{\mathrm{GAE}(\gamma,\lambda)}$, $\hat{A}_C^{\mathrm{GAE}(\gamma,\lambda)}$, and $\hat{A}_S^{\mathrm{GAE}(\gamma,\lambda)}$ with $V^\pi(s;\phi)$, $V_C^\pi(s;\phi_C)$, $S_C^\pi(s;\psi_C)$, and D.
  Calculate a policy gradient from (15) using the calculated GAEs as the advantages and update $\pi(a|s;\theta)$.
  Update $V^\pi(s;\phi)$, $V_C^\pi(s;\phi_C)$, and $S_C^\pi(s;\psi_C)$ using (16), (17), (21) from D.
end for

Dohyeong Kim and Songhwai Oh, "**Efficient Off-Policy Safe Reinforcement Learning Using Trust Region Conditional Value at Risk**," IEEE Robotics and Automation Letters, vol. 7, no. 3, pp. 7644-7651, Jul. 2022.

EFFICIENT OFF-POLICY SAFE REINFORCEMENT LEARNING USING TRUST REGION CONDITIONAL VALUE AT RISK

Motivation

RLLAB
http://rllab.snu.ac.kr

In safe RL, it is important to

- Well define a constraint to prevent catastrophic failure,

  → Use risk measure-based constraints

- Satisfy the safety constraint as soon as possible.

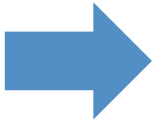
  → Improve sample efficiency

- TRC method:

1) Uses CVaR-based constraints, → Reduce catastrophic failure

2) Update policy using the trust region method → Stabilize training process

But **not sample efficient**.



We propose to extend TRC to an **off-policy** algorithm.

Off-Policy Cost Surrogates

- To use off-policy data, we propose new cost surrogates:

$$J_C^{\mu,\pi}(\pi') := J_C(\pi) + \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{d^\mu,\mu}\left[\frac{\pi'(a|s)}{\mu(a|s)}A_C^\pi(s,a)\right],$$

$$J_S^{\mu,\pi}(\pi') := J_S(\pi) + \frac{1}{1-\gamma^2}\mathop{\mathbb{E}}_{d_2^\mu,\mu}\left[\frac{\pi'(a|s)}{\mu(a|s)}A_S^\pi(s,a)\right].$$

(μ: Behavioral policy)

- Using the surrogates, we can approximate the CVaR constraint:

$$\overline{\mathrm{CVaR}}_\alpha(C_\pi) := J_C^{\mu,\pi}(\pi') + \frac{\phi(\Phi^{-1}(\alpha))}{\alpha}\sqrt{J_S^{\mu,\pi}(\pi') - J_C^{\mu,\pi}(\pi')^2} \le d.$$

Bound of Off-Policy Surrogates

- In the paper, Theorem 1 gives the bound of the CVaR approximation.

$$\left|\mathrm{CVaR}_\alpha(C_{\pi'}) - \overline{\mathrm{CVaR}}_\alpha(C_{\pi'})\right| \le \left(\frac{4\epsilon_C\gamma}{(1-\gamma)^2} + \epsilon_{\mathrm{CVaR}}\frac{\phi(\Phi^{-1}(\alpha))}{\alpha}\right) D_{\mathrm{TV}}^{\max}(\mu,\pi') D_{\mathrm{TV}}^{\max}(\pi,\pi').$$

Off-Policy TRC

- We can build a subproblem to update the policy by maximizing the lower bound of the objective and constraining the upper bound of CVaR.

$$\underset{\pi'}{\text{maximize}}\ J^{\mu,\pi}(\pi') - 4\epsilon_R\gamma D(\mu,\pi')D(\pi,\pi')/(1-\gamma)^2$$

$$\textbf{s.t.}\ \overline{\text{CVaR}}_\alpha(C_{\pi'}) + \left(\frac{4\epsilon_C\gamma}{(1-\gamma)^2} + \epsilon_{\text{CVaR}}\frac{\phi(\Phi^{-1}(\alpha))}{\alpha}\right)D(\mu,\pi')D(\pi,\pi') \le \frac{d}{1-\gamma}.$$

By approximating within the trust region,

We propose *off-policy TRC*.

$$\underset{\pi'}{\text{maximize}}\ J^{\mu,\pi}(\pi')$$

$$\textbf{s.t.}\ \overline{\text{CVaR}}_\alpha(C_{\pi'}) \le d/(1-\gamma),$$

$$D_{\text{KL}}(\pi||\pi') + \delta_{\text{old}} \le \delta.$$

- The solution can be obtained by a linear and quadratic constrained linear programming (LQCLP) solver.
- δ_{old} is for the adaptive trust region, which will be covered in the next slide.

Adaptive Trust Region

- The trust region can be defined as below in the off-policy setting.

$$D(\mu, \pi')D(\pi, \pi') \leq \delta.$$

- Using the triangular inequality, → $D(\mu, \pi') \leq D(\mu, \pi) + D(\pi, \pi')$
  we can simplify the trust region.

$$D_{\mathrm{KL}}(\pi||\pi') \leq \delta + D_{\mathrm{KL}}(\mu||\pi)/2$$

$$- \sqrt{D_{\mathrm{KL}}(\mu||\pi)\left(\delta + D_{\mathrm{KL}}(\mu||\pi)/4\right)} = \delta - \delta_{\mathrm{old}}.$$

Retrace Estimator

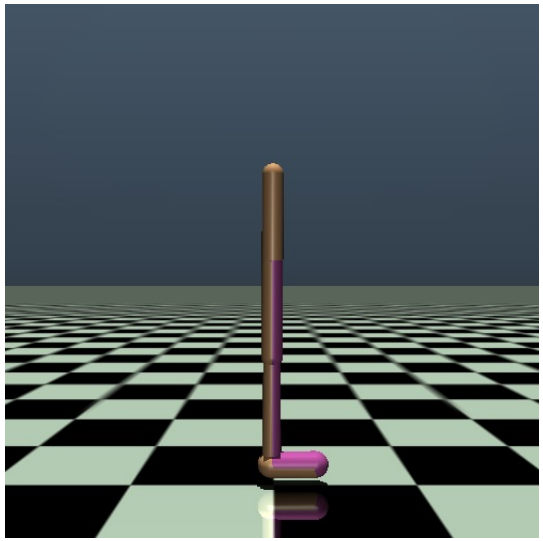
- Like GAEs, the retrace estimator can control the bias-variance trade-off in the **off-policy** setting.
- We derive the retrace estimator for the cost square value:

$$\overline{V}_{S,t} = c_t^2 + 2\gamma c_t V_C^{\pi}(s_{t+1}) + \gamma^2 V_S^{\pi}(s_{t+1}) + \gamma^2 \lambda \overline{\rho}_{t+1}(\overline{V}_{S,t+1} - V_S^{\pi}(s_{t+1})).$$

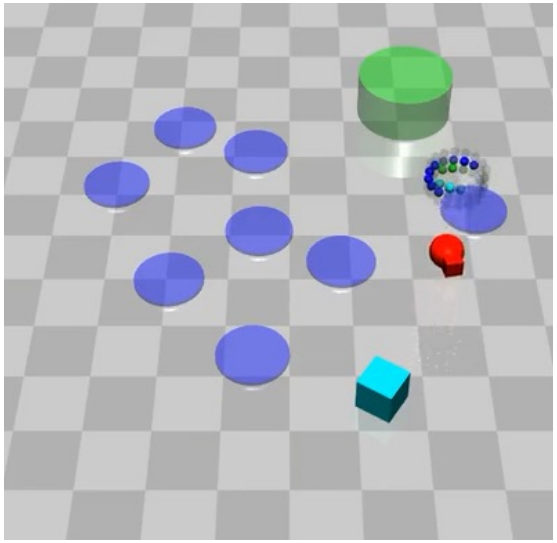
, where $\overline{\rho}_t = \min(1, \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)})$.

Environments

MuJoCo [2]



Safety Gym [1]



Jackal Robot [3]

[1] A. Ray, J. Achiam, and D. Amodei, “Benchmarking safe exploration in deep reinforcement learning,” arXiv preprint arXiv:1910.01708, vol. 7, 2019.

[2] Todorov, Emanuel, Tom Erez, and Yuval Tassa. "Mujoco: A physics engine for model-based control." *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012.

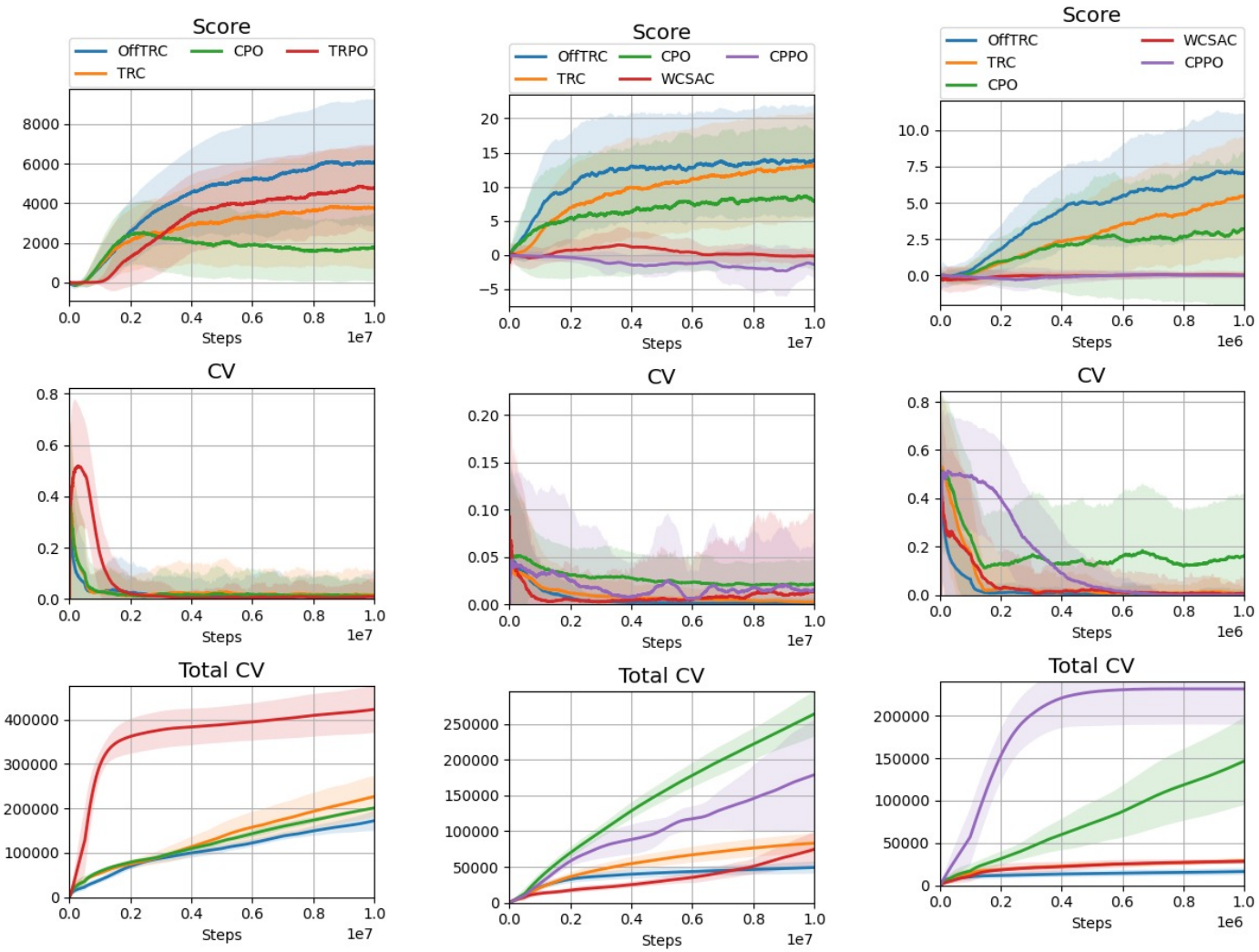
[3] C. R. Inc., “Jackal UGV - small weatherproof robot,” Sep. 2015. [On- line]. Available: https://clearpathrobotics.com/jackal-small-unmanned- ground-vehicle/

Experiments

Results

- Score:
  reward sum divided by the number of constraint violations.
- Total CV:
  the total number of constraint violations during the training.

Half-Cheetah (MuJoCo) | Point-Goal (Safety Gym) | Jackal Robot



RLLAB
Robot Learning Laboratory

Efficient Off-Policy Safe Reinforcement Learning Using Trust Region Conditional Value at Risk

Dohyeong Kim and Songhwai Oh

Robot Learning Laboratory
Department of Electrical and Computer Engineering,
Seoul National University

Dohyeong Kim, Kyungjae Lee, and Songhwai Oh, "**Trust Region-Based Safe Distributional Reinforcement Learning for Multiple Constraints**," in Proc. of Neural Information Processing Systems (NeurIPS), Dec. 2023.

TRUST REGION-BASED SAFE DISTRIBUTIONAL REINFORCEMENT LEARNING FOR MULTIPLE CONSTRAINTS

RLLAB
Robot Learning Laboratory

Trust Region-Based Safe Distributional Reinforcement Learning for Multiple Constraints

Dohyeong Kim[1], Kyungjae Lee[2], and Songhwai Oh[1]

[1]Electrical and Computer Engineering, Seoul National University

[2]Artificial Intelligence Graduate School, Chung-Ang University

** The voice is generated by OpenAI API.*

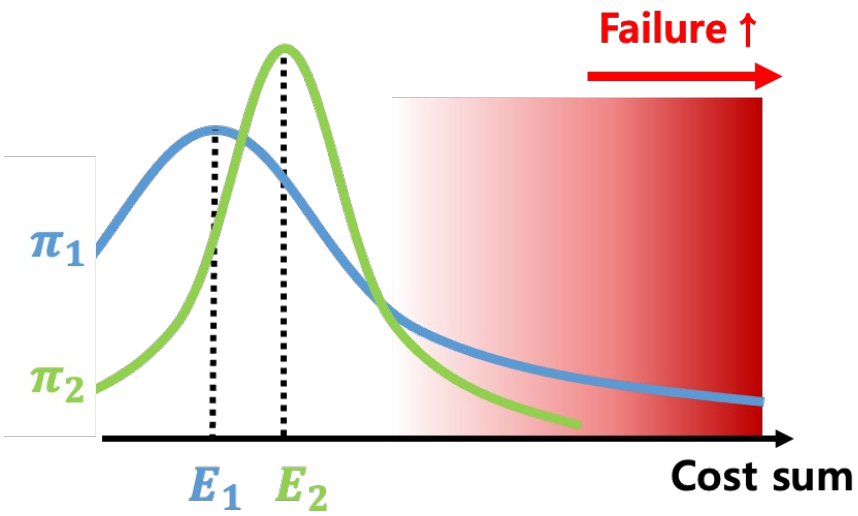
Dohyeong Kim, Kyungjae Lee, and Songhwai Oh, "**Trust Region-Based Safe Distributional Reinforcement Learning for Multiple Constraints**," in Proc. of Neural Information Processing Systems (NeurIPS), Dec. 2023.

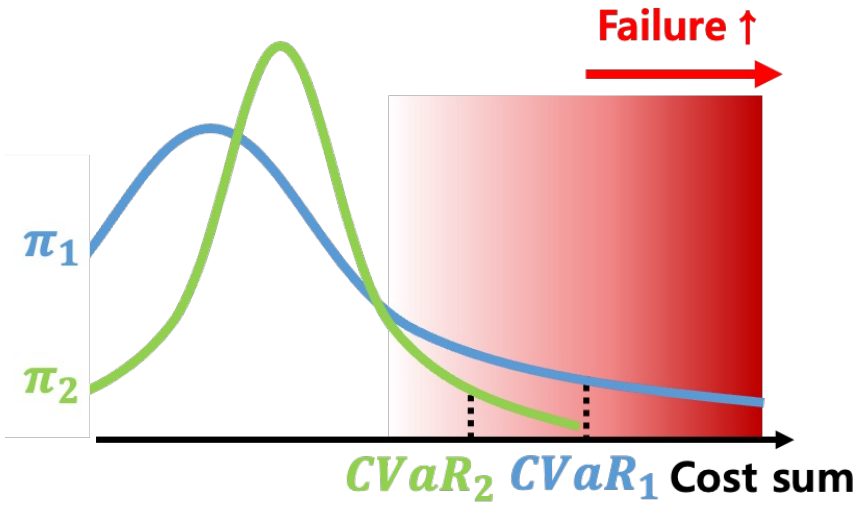
Dohyeong Kim, Taehyun Cho, Seungyub Han, Hojun Chung, Kyungjae Lee, and Songhwai Oh, "**Spectral-Risk Safe Reinforcement Learning with Convergence Guarantees**," in Proc. of Neural Information Processing Systems (NeurIPS), Dec. 2024.

SPECTRAL-RISK SAFE REINFORCEMENT LEARNING WITH CONVERGENCE GUARANTEES

Motivation

- Safe RL: $max_\pi \mathbb{E}[\Sigma_{t=0}^{\infty} \gamma^t R_t]$ s.t. $\mathbb{E}\left[\Sigma_{t=0}^{\infty} \gamma^t C_{i,t}\right] \leq d_i/(1-\gamma)$.





Motivation

- A risk-constrained RL (RCRL) problem:

$$\max_\pi J_R(\pi) \text{ s.t. } \mathcal{R}_i(C_i^\pi) \leq d_i \ \forall_i, \text{ where } \mathcal{R}_i \text{ is a risk measure.}$$

- Due to the **nonlinearity of risk measures**, it is challenging to develop a safe RL algorithm that guarantees **convergence to an optimal policy**.

$\Longrightarrow$ Propose a bilevel optimization framework for risk-constrained RL using the duality of spectral risk and show convergence guarantees in tabular settings.

Spectral Risk Measure

- Definition:

$$\mathcal{R}_\sigma(X) \coloneqq \int_0^1 F_X^{-1}(u)\sigma(u)du,$$

where σ (spectrum) is an increasing function, $\sigma \geq 0$, and $\int_0^1 \sigma(u)du = 1$.

- Example:

Conditional value at risk (CVaR): $\sigma(u) = \mathbf{1}_{u \geq \alpha}/(1-\alpha)$.

Spectral Risk Measure

- Definition:

$$\mathcal{R}_\sigma(X) \coloneqq \int_0^1 F_X^{-1}(u)\sigma(u)du,$$

where σ (spectrum) is an increasing function, $\sigma \geq 0$, and $\int_0^1 \sigma(u)du = 1$.

- Dual form expression:

$$\mathcal{R}_\sigma(X) = \inf_g \mathbb{E}[g(X)] + \int_0^1 g^*(\sigma(u))du,$$

where

1. g is an increasing convex function
2. $g^*(y) = \inf_x xy - g(x)$ is the convex conjugate of g
3. $\mathcal{R}_\sigma^g(X) \coloneqq \mathbb{E}[g(X)] + \int_0^1 g^*(\sigma(u))du$ is a sub-risk measure.

Bilevel Optimization Framework

- Reformulation of the RCRL problem:

$$\max_\pi J_R(\pi) \text{ s.t. } \mathcal{R}_{\sigma_i}(C_i^\pi) \leq d_i \ \forall_i.$$

$$\Longrightarrow \sup_{g_1,\ldots,g_N} \underbrace{\max_\pi J_R(\pi) \text{ s.t. } \mathcal{R}_{\sigma_i}^{g_i}(C_i^\pi) \leq d_i \ \forall_i.}_{\text{Inner problem}}$$

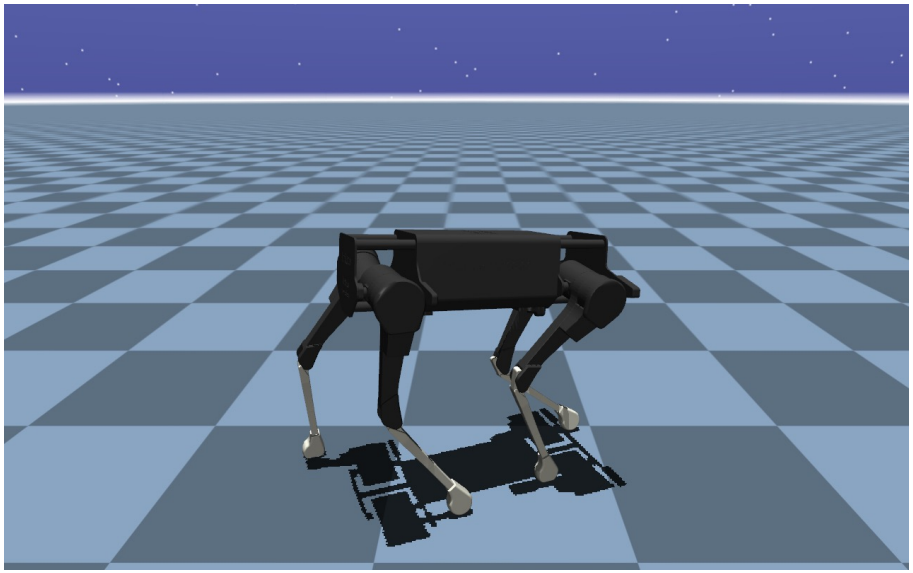
Outer problem

Experimental Results

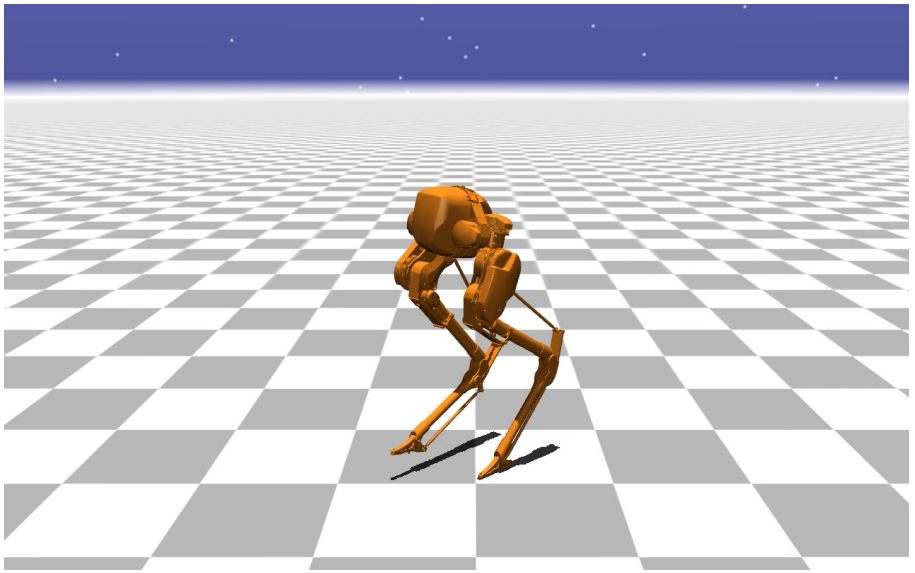
RLLAB
http://rllab.snu.ac.kr

- Legged robot locomotion tasks:

Quadrupedal (Laikago)



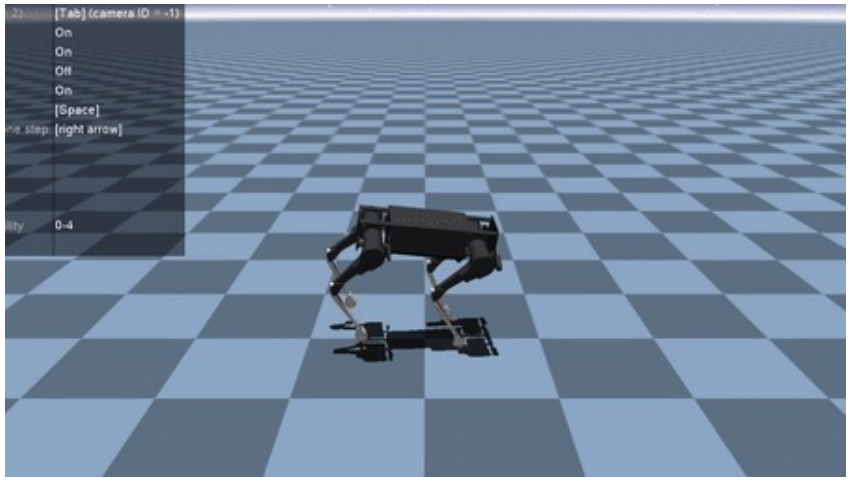
Bipedal (Cassie)



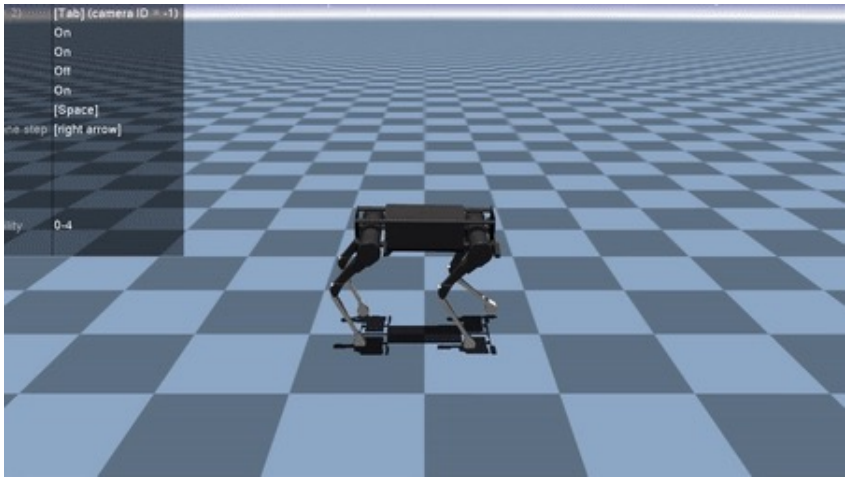
Experimental Results

RLLAB
http://rllab.snu.ac.kr

SRCPO
(Proposed)

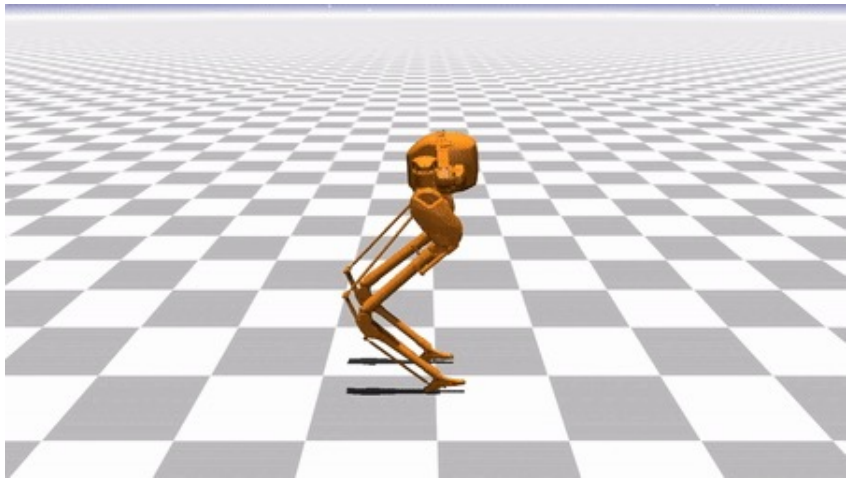


WCSAC-D

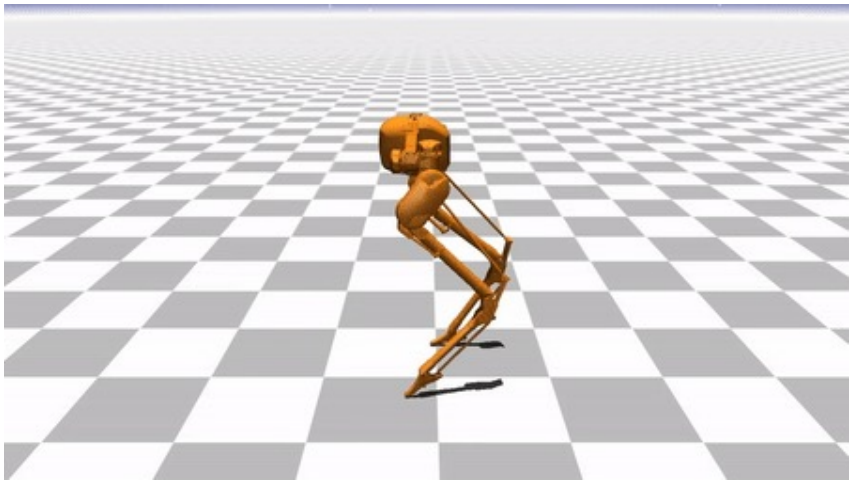


Experimental Results

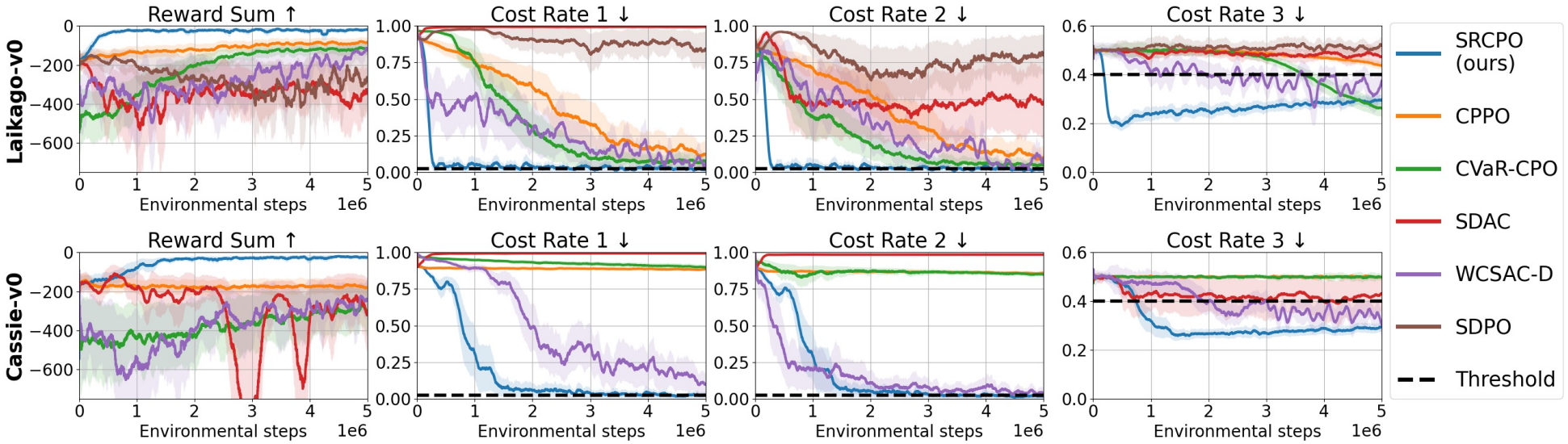
SRCPO
(Proposed)



WCSAC-D



Experimental Results



Acrobat Robots

APPLICATION

Stage-Wise Reward Shaping for Acrobatic Robots: A Constrained Multi-Objective Reinforcement Learning Approach

RLLAB
http://rllab.snu.ac.kr

Dohyeong Kim, Hyeokjin Kwon, Junseok Kim, Gunmin Lee, and Songhwai Oh, "**Stage-Wise Reward Shaping for Acrobatic Robots: A Constrained Multi-Objective Reinforcement Learning Approach**," in Proc of the IEEE International Conference on Robotics and Automation (ICRA), May 2025.