Comparison of Safe Reinforcement Learning Algorithms in Safety Gym

Kati Moug, Sunho Jang

April 2021

Outline

Paper Outline

The main paper that we present in this talk is Benchmarking Safe Exploration in Deep Reinforcement Learning [1]. The paper:

- Argues that constrained RL provides the best standardized framework for safe RL.
- Presents a set of benchmark environments for testing constrained RL algorithms called Safety Gym, built on OpenAI Gym [2] and MuJoCo [3].
- Tests a number of state of the art algorithms in Safety Gym.

Outline

Safety in RL

- Within a simulated environment, reinforcement learning agents are free to explore without consequence.
- Complex systems, such as human-AI interaction, likely cannot be effectively trained in simulation.
- Real world training environments require safe exploration.

Safety Constraint Examples

- Self-driving cars should not injure humans
- Power grid infrastructure should not be damaged by AI system
- Conversational assistants like Siri and Alexa should not suggest harmful responses to medical queries [4]

Figure: Photo by Campbell on Unsplash

Approaches to Safe RL

- Include a penalty in the reward function for dangerous behavior
  - Difficult to tune the penalty to ensure safety specifications are met
  - No safety assurances during exploration
- Constrain the variance or risk of reward
  - Does not allow consideration of non-reward safety specifications

Constrained RL

- Separately specifies reward functions and cost functions
- Incorporates domain knowledge by separating learning tasks into learning good performance-related actions and good safety-related actions
- Increases generalizability. Cost functions for reducing harm in one problem may apply to another problem.

RL in a Constrained Markov Decision Processes (CMDP)

Settings

- Reward function $R : S \times A \times S \to \mathbb{R}$
- Cost function $C_i : S \times A \times S \to \mathbb{R} \quad (i = 1, \ldots, m)$
- d_i : The limits on the cost functions.

The reinforcement learning problem in a CMDP is defined as

$$\pi^\star = \arg\max_{\pi \in \Pi_C} J(\pi),$$

where $\Pi_C := \{\pi \in \Pi : \forall i, J_{C_i}(\pi) \leq d_i\}$.

- $J(\pi) := E_{\tau\sim\pi}[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1})]$
- $J_{C_i}(\pi) := E_{\tau\sim\pi}[\sum_{t=0}^{\infty} \gamma^t C_i(s_t, a_t, s_{t+1})]$

Outline

Safety Gym Environments

- Each benchmark has a robot that must navigate the environment and its obstacles. In line with the authors' argument to separate task performance and safety performance, the environments have separately defined cost and reward functions.
- The environments have randomized layouts, that reset at the beginning of each new episode in order to encourage the agents to learn generalized behaviors.

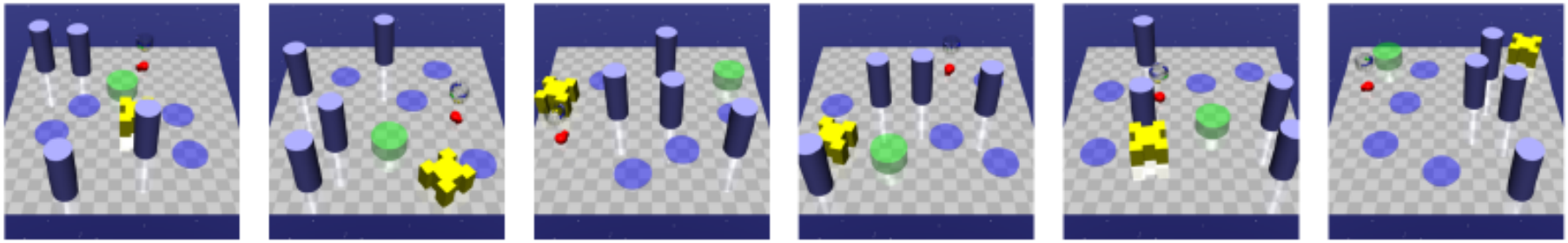
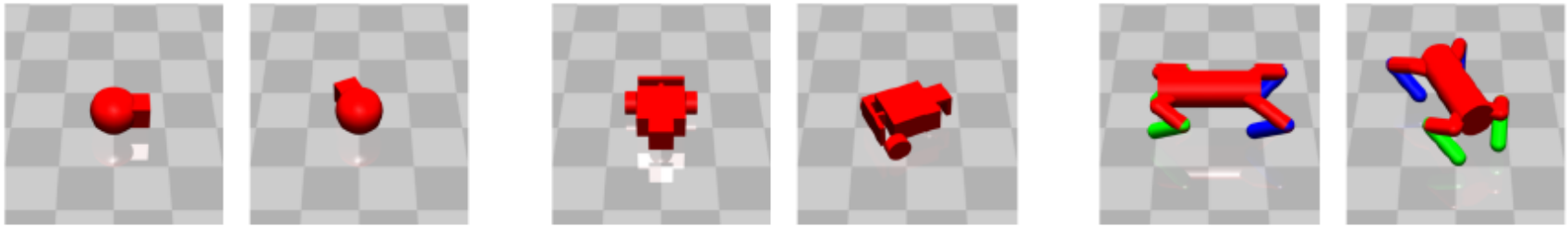


Figure: Randomized layouts of an environment in Safety Gym (Figure: [1])

Safety Gym Elements



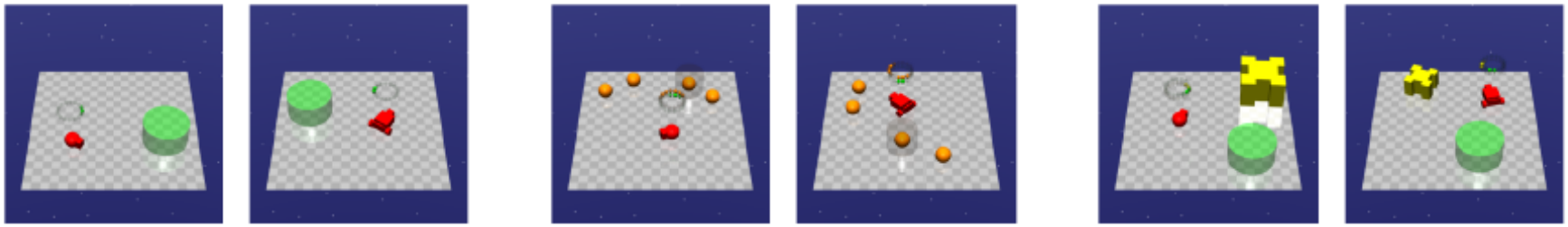
(a) Point: a simple 2D robot that can turn and move.

(b) Car: a wheeled robot with differential drive control.

(c) Doggo: a quadrupedal robot with bilateral symmetry.

Figure: Robots in Safety Gym (Figure: [1])

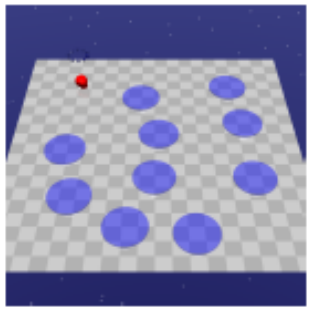
Safety Gym Elements



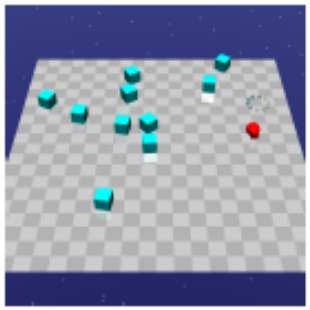
(a) Position (b) Button (c) Push

Figure: Tasks in Safety Gym (Figure: [1])

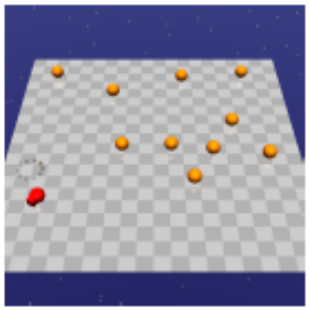
Safety Gym Elements



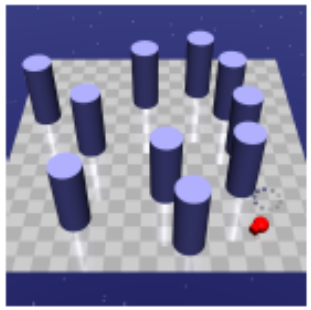
(a) Hazards, dangerous areas.



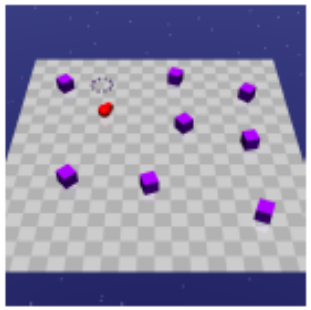
(b) Vases, fragile objects.



(c) Buttons, sometimes should not be pressed.



(d) Pillars, large fixed obstacles.



(e) Gremlins, moving objects.

Figure: Obstacles in Safety Gym (Figure: [1])

Outline

Lagrangian method

Let a constrained RL problem be represented as follows

$$\max_{\pi} \; f(\pi)$$
$$\text{s.t.} \; g(\pi) \leq 0.$$

- With $f(\pi)$ the objective and $g(\pi) \leq 0$ the constraint, Lagrangian methods solve the equivalent unconstrained max-min optimization problem

$$\max_{\pi} \min_{\lambda \geq 0} \mathcal{L}(\pi, \lambda) := f(\pi) - \lambda g(\pi),$$

  by **gradient ascent on π and descent on λ.**

- This Lagrangian approach is combined with several optimization algorithm to solve problem.
  - Trust Region Policy Optimization (TRPO)
  - Proximal Policy Optimization (PPO)

Outline

RL in a Constrained Markov Decision Processes (CMDP)

Settings

- Reward function $R : S \times A \times S \to \mathbb{R}$
- Cost function $C_i : S \times A \times S \to \mathbb{R} \quad (i = 1, \ldots, m)$
- d_i : The limits on the cost functions.

The reinforcement learning problem in a CMDP is defined as

$$\pi^\star = \arg\max_{\pi \in \Pi_C} J(\pi),$$

where $\Pi_C := \{\pi \in \Pi : \forall i, J_{C_i}(\pi) \leq d_i\}$.

- $J(\pi) := E_{\tau\sim\pi}[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1})]$
- $J_{C_i}(\pi) := E_{\tau\sim\pi}[\sum_{t=0}^{\infty} \gamma^t C_i(s_t, a_t, s_{t+1})]$

Local policy search for CMDP

Local policy search for CMDPs is as following

$$\begin{aligned} \pi_{k+1} = \underset{\pi \in \Pi_\theta}{\arg\max} \; & J(\pi) \\ \text{s.t.} \; & J_{C_i}(\pi) \leq d_i \;\; i = 1, \ldots, m \\ & D(\pi, \pi_k) \leq \delta, \end{aligned}$$

where D is some distance measure, and $\delta > 0$ is a step size.

- Evaluation of the constraint functions is difficult.
- Typically requires off-policy evaluation.

Theoretical foundation

Ref. [5] proves bounds on the difference in returns of two policies as follows.

Corollary

For any policies π', π with $\epsilon^{\pi'} := \max_s |E_{a\sim\pi'}[A^\pi(s,a)]|$, the following bound holds:

$$J(\pi') - J(\pi) \geq \frac{1}{1-\gamma} E_{\substack{s\sim d^\pi \\ a\sim\pi'}} \left[A^\pi(s,a) - \frac{2\gamma\epsilon^{\pi'}}{1-\gamma} D_{TV}(\pi'||\pi)[s] \right].$$

- $A^\pi(s,a) := Q^\pi(s,a) - V^\pi(s)$
- $V^\pi(s) := E_{\tau\sim\pi}[R(\tau)|s_0 = s]$
- $Q^\pi(s,a) := E_{\tau\sim\pi}[R(\tau)|s_0 = s, a_0 = a]$
- $D_{TV}(\pi'||\pi)[s] = (1/2)\sum_a |\pi'(a|s) - \pi(a|s)|$

Theoretical foundation

From the same theorem, the bound in the difference in costs of two polices is as follows.

Corollary

For any policies π', π and any cost function C_i with $\epsilon^{\pi'} := \max_s |E_{a\sim\pi'}[A^{\pi}_{C_i}(s,a)]|$, the following bound holds:

$$J_{C_i}(\pi') - J_{C_i}(\pi) \leq \frac{1}{1-\gamma} E_{\substack{s\sim d^{\pi} \\ a\sim\pi'}} \left[A^{\pi}_{C_i}(s,a) + \frac{2\gamma\epsilon^{\pi'}_{C_i}}{1-\gamma} D_{TV}(\pi' || \pi)[s] \right].$$

- $A^{\pi}_{C_i}(s,a) := Q^{\pi}_{C_i}(s,a) - V^{\pi}_{C_i}(s)$
- $V^{\pi}_{C_i}(s) := E_{\tau\sim\pi}[C_i(\tau)|s_0 = s]$
- $Q^{\pi}_{C_i}(s,a) := E_{\tau\sim\pi}[C_i(\tau)|s_0 = s, a_0 = a]$

Constrained Policy Optimization - Strategy

- Now here is what we got

$$J(\pi') - J(\pi) \geq \frac{1}{1-\gamma} E_{\substack{s\sim d^{\pi} \\ a\sim\pi'}} \left[A^{\pi}(s,a) - \frac{2\gamma\epsilon^{\pi'}}{1-\gamma} D_{TV}(\pi' || \pi)[s] \right]$$

$$J_{C_i}(\pi') - J_{C_i}(\pi) \leq \frac{1}{1-\gamma} E_{\substack{s\sim d^{\pi} \\ a\sim\pi'}} \left[A^{\pi}_{C_i}(s,a) + \frac{2\gamma\epsilon^{\pi'}_{C_i}}{1-\gamma} D_{TV}(\pi' || \pi)[s] \right].$$

- Constrained Policy Optimization (CPO) updates the policy to π_{k+1} from π_k so that it can
  - maximize the lower bound on $J(\pi_{k+1}) - J(\pi_k)$
  - minimize the upper bound on $J_{C_i}(\pi_{k+1})$
  - Local policy search : bound on the distance between π and π_k

CPO (Trust Region Optimization)

CPO iteratively updates the policy by solving the following

$$\pi_{k+1} = \arg\max_{\pi \in \Pi_\theta} \; E_{\substack{s \sim d^{\pi_k} \\ a \sim \pi}} \left[A^{\pi_k}(s,a) \right]$$

$$\text{s.t.} \; J_{C_i}(\pi_k) + \frac{1}{1-\gamma} E_{\substack{s \sim d^{\pi_k} \\ a \sim \pi}} \left[A^{\pi_k}_{C_i}(s,a) \right] \leq d_i \;\; \forall i$$

$$\overline{D}_{KL}(\pi || \pi_k) \leq \delta.$$

Proposition (CPO Update Worst-Case Constraint Violation)

Suppose π_k, π_{k+1} are related by the above algorithm, and that Π_θ is any set of policies with $\pi_k \in \Pi_\theta$. An upper bound on the C_i-cost of π_{k+1} is

$$J_{C_i}(\pi_{k+1}) \leq d_i + \frac{\sqrt{2\delta}\gamma \epsilon^{\pi_{k+1}}_{C_i}}{(1-\gamma)^2},$$

where $\epsilon^{\pi^{k+1}}_{C_i} := \max_s |E_{a \sim \pi_{k+1}}[A^{\pi_k}_{C_i}(s,a)]|$.

Outline

Problem

The following optimization problem is solved by each algorithm :

$$\max_{\pi_\theta} \ \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{T} r_t \right]$$

$$\text{s.t.} \ \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{T} c_t \right] \leq d,$$

where c_t is the aggregate indicator cost function for the environment ($c_t = 1$ for an unsafe interaction, regardless of source) and d is a hyperparameter.

Five algorithms are compared

- Proximal Policy Optimization
- Proximal Policy Optimization-Lagrangian
- Trust Region Policy Optimization
- Trust Region Policy Optimization -Lagrangian
- Constrained Policy Optimization

Result

SG18	Return $\bar{J}_r$	Violation $\bar{M}_c$	Cost Rate $\bar{\rho}_c$
PPO	1.0	1.0	1.0
PPO-Lagrangian	0.24	0.026	**0.245**
TRPO	**1.094**	1.132	1.004
TRPO-Lagrangian	0.331	**0.018**	0.265
CPO	0.784	0.593	0.646

Figure: Normalized metrics from the conclusion of training averaged over the 18 environments and three random seeds per environment

- $\overline{J}_r(\theta)$: normalized return
- $\overline{M}_c(\theta)$: normalized constraint violation
- $\overline{\rho}_c(\theta)$: normalized cost rate

Results -Figures

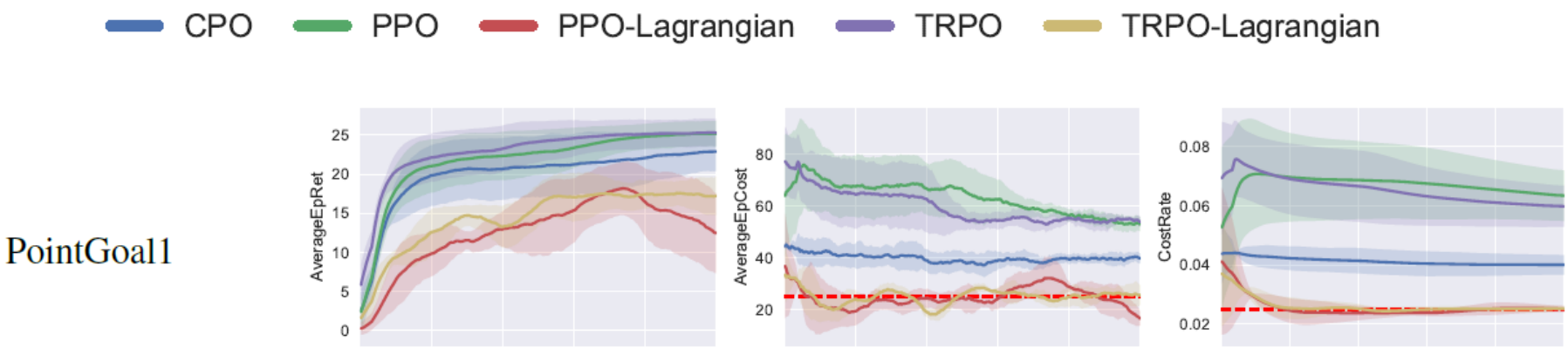


Figure: Result of all algorithms

The x-axis of these graphs is the number of interactions with environment and the y-axis is the following in the order from the left to the right

- $J_r(\theta)$: The average episodic return
- $J_c(\theta)$: The quantity we aim to constrain
- ρ_c : The average cost over the entirety of training

Conclusion

- There is a trade-off between costs and rewards.
- While the Lagrangian methods with TRPO, PPO reliably enforce constraints, Constrained Policy Optimization has pretty large costs. It is expected to be from the approximation errors in CPO.

Our interest - CVaR

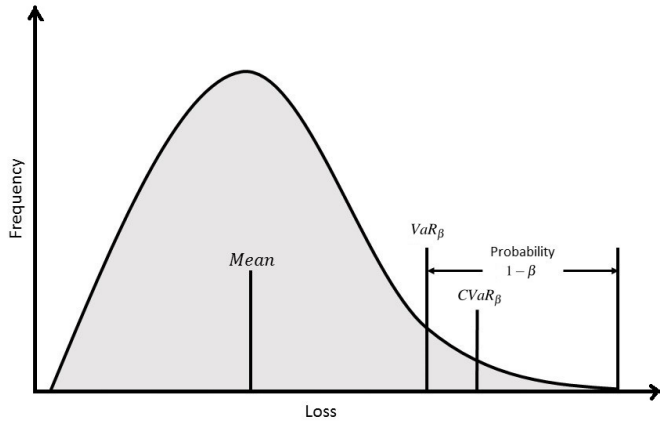


Figure: Description of Conditional Value at Risk (CVaR) [6]

Our goal is to develop algorithms for the constrained RL problem with CVaR as the risk measure which is defined as

$$\text{CVaR}_\alpha(X) := \min_{\nu \in \mathbb{R}} \left\{ \nu + \frac{1}{1-\alpha} \mathbb{E}\left[(X-\nu)^+\right] \right\}$$

which corresponds to the expectation value of the costs in *risky* range.

References I

- A. Ray, J. Achiam, and D. Amodei, "Benchmarking safe exploration in deep reinforcement learning," *arXiv preprint arXiv:1910.01708*, 2019.
- G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "OpenAI gym," *arXiv preprint arXiv:1606.01540*, 2016.
- E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 5026–5033.
- T. W. Bickmore, H. Trinh, S. Olafsson, T. K. O'Leary, R. Asadi, N. M. Rickles, and R. Cruz, "Patient and consumer safety risks when using conversational assistants for medical information: An observational study of Siri, Alexa, and Google Assistant," *Journal of Medical Internet research*, vol. 20, no. 9, p. e11510, 2018.

References II

J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *International Conference on Machine Learning*. PMLR, 2017, pp. 22–31.

J. McAllister, Z. Li, J. Liu, and U. Simonsmeier, "Epo dosage optimization for anemia management: Stochastic control under uncertainty using conditional value at risk," *Processes*, vol. 6, no. 5, p. 60, 2018.