Second Paper Summary, EE245 Spring 2025

Kunyi Yu
Department of Computer Science and Engineering
University of California, Riverside
Riverside, CA 92521, USA
kyu135@ucr.edu

Abstract

The second paper summary due on Monday, May 19, 2025. The author chose the paper titled "Benchmarking Safe Exploration in Deep Reinforcement Learning" by Alex Ray, Joshua Achiam, and Dario Amodei from OpenAI.

1 Summary of Major Contributions

When training a reinforcement learning (RL) agent interacting with human, it is crucial to ensure a safe exploration. Otherwise, the agent may take actions that are harmful to human or the environment, such as autonomous vehicles accidents, AI systems causing power grid failures, or question-answering systems generating misleading medical advice. As a result, the paper [1] advance the study of it by proposing three major contributions:

First, the authors propose to standardize constrained RL to be the major formalism for safe exploration. They argue that safety specifications are indipendent of the performance specifications, and constrained RL is a natural way to formalize this. The authors also note that the constrained RL is scalable to high-dimensional methods.

Second, the paper provides the Safety Gym benchmark suite consisting of 18 high-dimensional continuous control tasks for safe exploration, 9 tasks for debugging, and tools for building extensive environments. They claim each environment express task performance and safety via a reward function and a set of additional cost functions. Moreover, Safety Gym has different levels of difficulty, randomized initial states, and highly extensive environments.

Finally, the authors establish a set of baselines for both unconstrained RL algorithms such as Trust Region Policy Optimization (TRPO) [2] and Proximal Policy Optimization (PPO) [3], and constrained RL algorithms such as Lagrangian methods, Constrained Policy Optimization (CPO) [4], and a constrained version of TRPO. They found that the performance of CPO is poor compared to Lagrangian methods.

2 Relation to Prior Work

Three aspects of prior work are reviewed in Section 2 of the paper:

Safety Overviews: Amodei et al. [5] and Leike et al. [6] provide two taxonomies and examples of safety problems in AI, including safe exploration. Two additional surveys discuss non-learning approaches to safety, which are omitted here because the paper focuses on modern deep RL methods.

Safety Definitions and Algorithms: The paper mentions 13 safety definitions or their variations, including: labeling states of environments as “safe” or “unsafe” [7], which is often related to constraints [8]; considering agents to be safe if they act, reason, and generalize within human preferences [9–12]; and so on.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

The mentioned RL algorithms include: using ensembles to learn generalized safe policies [13]; preventing unsafe actions by learning action-time interventions [14, 15]; using human interventions to avoid unsafe actions [16]; and so on.

Benchmarking RL and RL Safety: The paper mentions several general RL benchmarks, including: ALE [17], OpenAI Gym [18], DeepMind Control Suite [19], MuJoCo simulator [20], and CoinRun [21]. Unlike these general RL benchmarks, Leike et al. [6] provide grid worlds demonstrating AI safety problems by using dueling reward functions evaluating both performance and safety.

Section 3 of the paper introduces several basic concepts. An optimal policy in **Constrained RL** is given by $\pi^* = \arg\max_{\pi \in \Pi_C} J_r(\pi)$, where Π_C is the feasible set of policies that satisfy the constraints. The feasible set in a constrained MDP is given by $\Pi_C = \{\pi : J_{c_i}(\pi) \leq d_i, \forall i\}$, where d_i is the threshold for the i-th constraint. Furthermore, the **choice of using constrained RL** for safety can address two practical problems. The **agent alignment** problem is universally applicable to all RL problems, and constrained RL is suitable for many methods to alleviate it. Lastly, the authors discuss the drawbacks of other safety approaches and the differences between constrained RL and multi-objective RL.

3 Summary of the Method

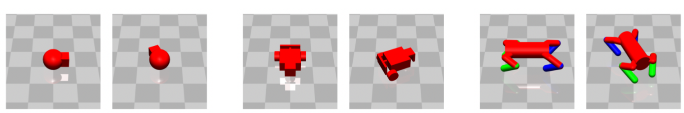
Due to the fact that the paper introduces a RL benchmark envrionment instead of a new algorithm, this part will highlight the characteristics and some details of Safety Gym.

Safety Gym has two major components: (1) an environment-builder that permits creating new environments with varing physics elements, performance requirements, and safety requirements; and (2) a set of pre-defined environments as benchmarks to standardize the evaluation of algorithms on the safe exploration problem.

The framework of Safety Gym is based on OpenAI Gym [18] interface and MuJoCo physics engine [20]. Each pre-defined environment contains a robot agent aiming to navigate in a cluttered environment to reach a goal, while following safety constraints like how to interact with objects and areas. Task objectives are defined by a reward function, and safety constraints are defined by a set of cost functions. The generalization of the environment is achieved by randomizing the initial state but not explicitly partitioning environments into training and test sets.

The environment-builder includes the following components:

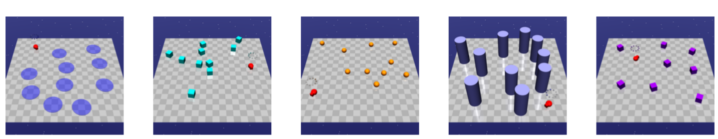
Pre-made robots include a 2D Point, a 2D Car, and a 3D Doggo. They can perceive and interact with the environment by sensors and actuators. The action space is continuous. A demo figure from the paper [1] is shown below:



(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 1: Pre-made robots in Safety Gym. These robots are used in our benchmark environments.

Tasks include Goal (move to a series of goal locations), Button (press a series of buttons), and Push (move a series of blocks). Reward functions could be sparse or dense. A demo figure from the paper [1] is shown below:



(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 3: Constraint elements used in our environments.

Constraint options include: Hazards (dangerous areas), Vases (objects to avoid), Pillars (immobile obstacles), Buttons (incorrectly goals), and Gremlins (moving objects).

Observations space options are highly configurable, including: standard robot sensors, joint position and velocity sensors, compasses for pointing to goals, and lidar.

Lastly, users may enable **layout randomization**. Additionally, the **Safety Gym Benchmark Suite** serves as a zero-shot evaluation tool to assess the generalization performance of RL algorithms.

4 Summary of Major Results

The experiments evaluate several RL algorithms (mentioned before) on Safety Gym environments. Key results are summarized below:

- Unconstrained RL gets high returns but violates safety.
- Constrained RL lowers returns to stay safe.
- Level 2 tasks are harder with more hazards.
- CPO fails to satisfy constraints; Lagrangian works better.
- Doggo learns with standard RL, but constrained RL fails.

5 Summary of Strengths

The paper provides a comprehensive but brief overview of the safety problem in AI, including the definitions, algorithms, and benchmarks. The authors also provide a clear motivation for using constrained RL for safe exploration. Moreover, the Safety Gym benchmark suite is well-designed and following the popular OpenAI Gym format, making it easy to use.

6 Summary of Weaknesses

The Safety Gym community is barely active. While Gym is among the top three most popular OpenAI repositories with 36k stars, Safety Gym is much less popular, ranking 87th with just over 500 stars. More importantly, whereas Gym has over 1,700 commits and 300 contributors, Safety Gym has only 1 commit and is already archived. This indicates that Safety Gym is not widely used or researched.

However, there is a successor of Safety Gym called Safety Gymnasium by Ji et al. [22] published in 2023 with similar stars and more commits.

Acknowledgments

This paper summary was independently conducted by the author Kunyi Yu. Meanwhile, the author would like to thank Professor Jiachen Li and TAs for their patient guidance and help.

References

[1] Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 7(1):2, 2019.

[2] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[3] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[4] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pages 22–31. PMLR, 2017.

[5] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[6] Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. Ai safety gridworlds. *arXiv preprint arXiv:1711.09883*, 2017.

[7] Alexander Hans, Daniel Schneegaß, Anton Maximilian Schäfer, and Steffen Udluft. Safe exploration for reinforcement learning. In *ESANN*, pages 143–148, 2008.

[8] Moshe Haviv. On constrained markov decision processes. *Operations research letters*, 19(1): 25–28, 1996.

[9] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

[10] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[11] Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

[12] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

[13] Zachary Kenton, Angelos Filos, Owain Evans, and Yarin Gal. Generalizing from a few environments in safety-critical reinforcement learning. *arXiv preprint arXiv:1907.01475*, 2019.

[14] Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*, 2018.

[15] Yinlam Chow, Ofir Nachum, Aleksandra Faust, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. Lyapunov-based safe policy optimization for continuous control. *arXiv preprint arXiv:1901.10031*, 2019.

[16] William Saunders, Girish Sastry, Andreas Stuhlmueller, and Owain Evans. Trial without error: Towards safe reinforcement learning via human intervention. *arXiv preprint arXiv:1707.05173*, 2017.

[17] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of artificial intelligence research*, 47:253–279, 2013.

[18] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[19] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

[20] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

[21] Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *International conference on machine learning*, pages 1282–1289. PMLR, 2019.

[22] Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. Safety gymnasium: A unified safe reinforcement learning benchmark. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL `https://openreview.net/forum?id=WZmlxIuIGR`.