Human in the Loop, Safe Reinforcement Learning for Continuous Control

DDP Final Review

Chetan Reddy N
ME19B093
Guided by Prof. Nirav Bhatt

June 13, 2024

Table of Contents

1. Introduction
2. Literature Survey
3. Problem Formulation
4. Algorithm
5. Simulation Experiments and Results
6. Conclusion
7. Future Work

Table of Contents

1. Introduction
2. Literature Survey
3. Problem Formulation
4. Algorithm
5. Simulation Experiments and Results
6. Conclusion
7. Future Work

Introduction

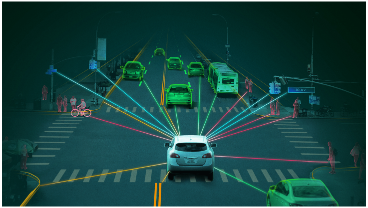
We aim to develop a framework to ensure safety during both the training and deployment phases of an RL by using human input.

Introduction

We aim to develop a framework to ensure safety during both the training and deployment phases of an RL by using human input.

Motivation:

- Safety Critical Tasks: Autonomous Driving, Healthcare Robotics etc
- As humans entrust autonomous agents with increasingly complex tasks, their involvement in the learning process becomes crucial.



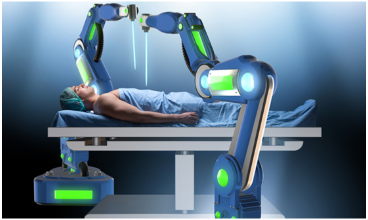


Table of Contents

1. Introduction
2. Literature Survey
3. Problem Formulation
4. Algorithm
5. Simulation Experiments and Results
6. Conclusion
7. Future Work

Literature Survey

We performed a comprehensive literature review from two directions and a taxonomy was developed:

- Safe Reinforcement Learning [1, 2, 3, 4]
- Human in the Loop RL [5, 6, 7, 8]

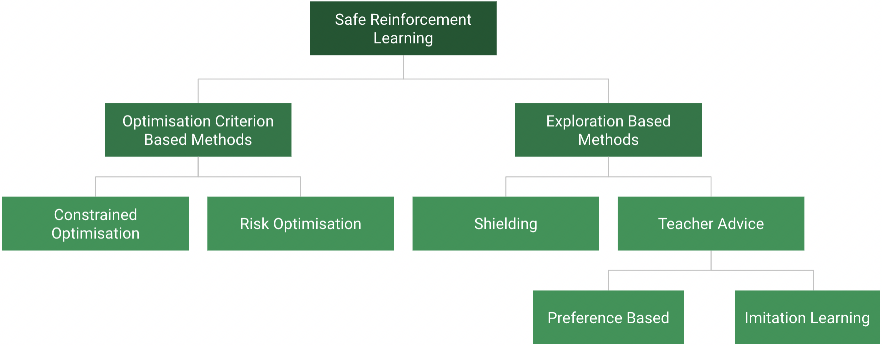


Figure: Safe RL Taxonomy

Table of Contents

1 Introduction

2 Literature Survey

3 Problem Formulation

4 Algorithm

5 Simulation Experiments and Results

6 Conclusion

7 Future Work

- Consider a Markov Decision Process $(\mathcal{S}, \mathcal{A}, f, \mathcal{R}, \gamma)$ with a State Space S, Action Space A, Transition Function f s.t $s' = f(s, a)$, Reward Structure R and discounting factor γ
- The objective is to learn a policy $\pi : \mathcal{S} \to \mathcal{A}$ which maximises the expected cumulative reward while ensuring safety in both training and deployment i.e $\forall s \in \mathcal{X}_s, f(s, \pi(s)) \in \mathcal{X}_s$

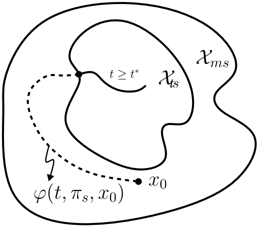


Figure: Depiction of Safe Sets

Notation	Meaning
$\mathcal{X}_{ts}$	Truly Safe State Space
$\mathcal{X}_{ms}$	Marginally Safe State Space
$\mathcal{X}_s$	Safe State Space $\mathcal{X}_{ts}$
CSP or π_s	Conservative Safe Policy

Figure: Notations

Table of Contents

1 Introduction

2 Literature Survey

3 Problem Formulation

4 Algorithm

5 Simulation Experiments and Results

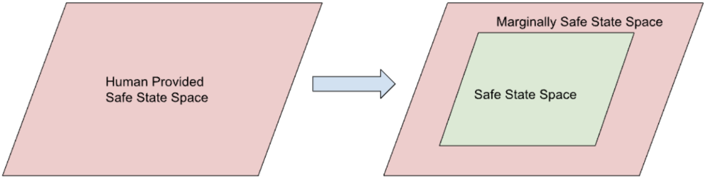
6 Conclusion

7 Future Work

Algorithm

Pseudo Code

- Initialise Policy and Critic Parameters (θ and ϕ) of the DDPG Model
- Repeat
  - If $s \in S_{safe}$, $a = \pi_\theta(s)$
  - Elif $s \in S_{marginally\ safe}$, $a = argmax_{a' \in A_{safe}(s)} Q_\phi(s, a')$
  - The Replay Buffer is populated with the tuples (s, a, s', r)
  - $\nabla_\phi L(\phi) = \nabla_\phi \frac{1}{|B|} \sum \left(r + \gamma Q_{\phi_{targ}}(s', \mu_{\theta_{targ}}(s')) - Q_\phi(s, a)\right)^2$
  - $\nabla_\theta L(\theta) = \nabla_\theta \frac{1}{|B|} \sum_{s \in B}(Q_\phi(s, \mu_\theta(s)) + (a - \mu_\theta(s))^2)$
  - The weights are updated using the gradients as defined above
- Until Convergence



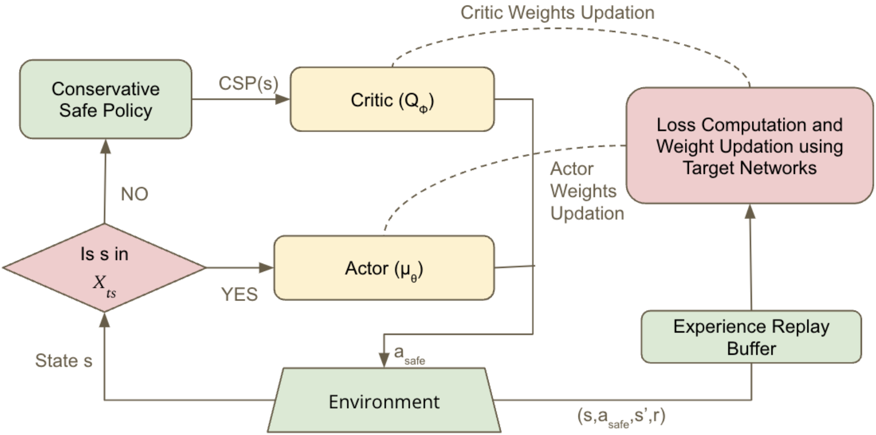


Figure: Model Pipeline

Table of Contents

1 Introduction

2 Literature Survey

3 Problem Formulation

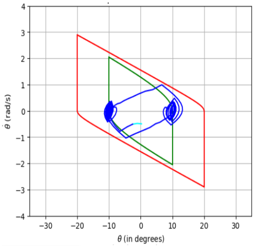
4 Algorithm

5 Simulation Experiments and Results

6 Conclusion

7 Future Work

Simulation Experiments and Results



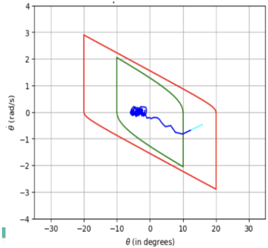


Figure: Trajectories of Episodes during Training and Testing

- During training, we see that the trajectory often enters the marginally safe region and is immediately pushed back into the safe region using the human provided conservative actions
- During testing, we see that the trained agent learns to stay inside the safe region

Some Training Curves - Safety Gymnasium Environment

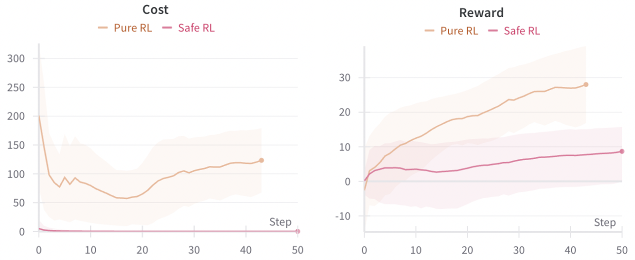


Figure: Metrics for Implementations with and without safety layer

Algorithm	Reward	Cost
Pure RL	47.6	203
Pure RL (Modified Reward)	2.85	0
Pure CSP	4.70	0
Safe RL	22.10	0

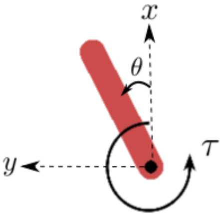
Table: Reward and Cost during Evaluation/Deployment

Table of Contents

1. Introduction
2. Literature Survey
3. Problem Formulation
4. Algorithm
5. Simulation Experiments and Results
6. Conclusion
7. Future Work

Conclusion

- A comprehensive literature survey was performed and a taxonomy was developed.
- A modified version of the Deep Deterministic Policy Gradient algorithm was implemented with a safety layer for continuous control.
- The algorithm was tested in the inverted pendulum environment and the safety-gymnasium environment which is a benchmark library for safe RL
- **Won 3rd Place for the Best Poster Award** with during the WSAI Annual Research Showcase 2024.



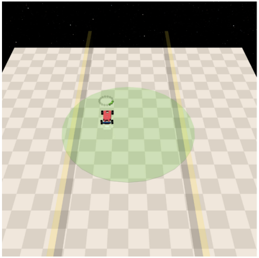


Table of Contents

1 Introduction

2 Literature Survey

3 Problem Formulation

4 Algorithm

5 Simulation Experiments and Results

6 Conclusion

7 Future Work

What Next?

- We are aiming to submit our work to the International Conference on Control, Automation, Robotics and Vision (ICARCV 2024) which has a deadline of June 30, 2024.
- There are multiple directions this work can be continued in by the future students
  - The provided CSP can be formulated as a distribution and probabilistic guarantees of safety can be established.
  - The work presently assumes that CSP and SSS are readily provided by the human. Model dynamics perhaps can be used to automate this to some extent.
  - Experiments can be performed on Real Life Mobile Robots

Thank you!
Questions?

J. Garcıa and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig, "Safe learning in robotics: From learning-based control to safe reinforcement learning," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, pp. 411–444, 2022.

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

G. Dalal, K. Dvijotham, M. Vecerik, T. Hester, C. Paduraru, and Y. Tassa, "Safe exploration in continuous action spaces," *arXiv preprint arXiv:1801.08757*, 2018.

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *Advances in neural information processing systems*, vol. 30, 2017.

W. Saunders, G. Sastry, A. Stuhlmueller, and O. Evans, "Trial without error: Towards safe reinforcement learning via human intervention," *arXiv preprint arXiv:1707.05173*, 2017.

C. Frye and I. Feige, "Parenting: Safe reinforcement learning from human input," *arXiv preprint arXiv:1902.06766*, 2019.

H. Hoang, T. Mai, and P. Varakantham, "Imitate the good and avoid the bad: An incremental approach to safe reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 12439–12447, 2024.