SaferRL

Release 0.1

Rong

May 18, 2022

USER DOCUMENTATION

SaferRL is a Python library that makes it easier to learn about safe reinforcement learning.

Note: This project is under active development.

CHAPTER

ONE

INTRODUCTION

Table of Contents

1.1 What This Is

Welcome to SaferRL!

This is a resource that makes it easier to learn about safe reinforcement learning.

This module contains a variety of helpful resources, including:

- a short [introduction](https://) to Safe RL terminology, kinds of algorithms, and basic theory,
- a [curated list](https://) of important papers organized by topic,
- and a well-documented [code repo](https://) of short, standalone implementations of key algorithms.

CHAPTER

TWO

PART 1: KEY CONCEPTS IN SAFE RL

Table of Contents

Welcome to our introduction to safe reinforcement learning! Here, we aim to acquaint you with

- the language and notation used to discuss the subject,
- a high-level explanation of what Safe RL algorithms do,
- and a little bit of the core math that underlies the algorithms.

2.1 Why Safe RL?

2.2 Key Concepts and Terminology

2.2.1 States and Observations

2.2.2 Policies

CHAPTER

THREE

KEY PAPERS IN SAFE RL

What follows is a list of papers in Safe RL that are worth reading. This is *far* from comprehensive, but should provide a useful starting point for someone looking to do research in the field.

Table of Contents

3.1 1. General Review

1. Unsolved Problems in ML Safety, Hendrycks et al, 2022.
2. Concrete Problems in AI Safety, Amodei et al, 2016.
3. A Comprehensive Survey on Safe Reinforcement Learning, García et al, 2015.

3.2 2. Model-free RL

1. Constrained Policy Optimization, Achiam et al, 2017.
2. Lyapunov-based Safe Policy Optimization for Continuous Control, Chow et al, 2019
3. Batch Policy Learning under Constraints, Le et al, 2019
4. Reward Constrained Policy Optimization, Tessler et all, 2019
5. Responsive Safety in Reinforcement Learning by PID Lagrangian Methods, Stooke et al, 2020
6. Projection-based Constrained Policy Optimization, Yang et al, 2020.

3.3 3. Model-based RL

1. Safe Model-based Reinforcement Learning with Stability Guarantees, Berkenkamp et al, 2017
2. Constrained model predictive control: Stability and optimality, Mayne et al, 2000
3. Constrained Policy Optimization via Bayesian World Models, As et al, 2022

3.4 4. Transfer Learning

1. Learning to be Safe: Deep RL with a Safety Critic, Srinivasan et al, 2020

3.5 5. Ensemble Learning

1. Gerneralzieing from a Few Environments in Safety-critical Reinforcement Learning, Kenton et al, 2019
2. Leave No Trace: Learning to Reset for Safe and Autonomous Reinforcement Learning, Eysenbach et al, 2018

3.6 6. Human in The Loop

1. Trial without Error: Towards Safe Reinforcement Learning via Human Intervention, Saunders et al, 2017

3.7 7. Curriculum Learning

1. Safe Reinforcement Learning via Curriculum Induction, Turchetta et al, 2020

3.8 8. Risk-sensitive RL

1. Risk-Sensitive Reinforcement Learning Applied to Control under Constraints, Geibel et al, 2005
2. Risk-Aware Transfer in Reinforcement Learning using Successor Features, Gimelfarb et al, 2021
3. Exponential Bellman Equation and Improved Regret Bounds for Risk-Sensitive Reinforcement Learning, Fei et al, 2021

3.9 9. Formal Methods

1. Verifiably Safe Exploration for End-to-End Reinforcement Learning, Hunt et al, 2021. See also Guarantee Safety in Training and Testing for related work.

CHAPTER

FOUR

BENCHMARKS AND METRICS IN SAFE RL

Table of Contents

We're evaluating the SaferRL algorithm implementations in a set of standard benchmarks and metrics.

4.1 1. Benchmarks

1. Open AI Safety Gym.
2. Open AI CoinRun.
3. DeepMind AI Safety Gridworlds.

4.2 2. Metrics

CHAPTER

FIVE

CONSTRAINED POLICY OPTIMIZATION

Table of Contents

5.1 Background

5.1.1 Quick Facts

5.1.2 Key Equations

5.1.3 Pseudocode

5.2 References

5.2.1 Relevant Papers

- *Constrained Policy Optimization*, Joshua et al. 2017

5.2.2 Why These Papers?

CHAPTER

SIX

ACKNOWLEDGEMENTS

We gratefully acknowledge the contributions of the many people who helped get this project off of the ground

CHAPTER

SEVEN

ABOUT THE AUTHOR

SaferRL was primarily developed by Rong, a Postdoc at TU Berlin working on topics related to reinforcement learning and AI safety.

CHAPTER

EIGHT

INDICES AND TABLES

- genindex
- modindex
- search