AIMLC ZG512 -
Deep Reinforcement Learning

Session 17: Special Topics in DRL

Introducing - Safety in Reinforcement Learning

S. P. Vimal, Department of CSIS, WILP Division (vimalsp@wilp.bits-

Safety

BIG QUESTION:
How do we guarantee safety when we deploy RL in real-world applications?

S. P. Vimal, Department of CSIS, WILP Division (vimalsp@wilp.bits-pilani.ac.in)

Agenda for the session

- ➔ What do we mean by Safety in RL?
- ➔ Safe RL – Problem
- ➔ Some Applications (Autonomous Driving)

Source for the session & Recommended Reading:
A Review of Safe Reinforcement Learning: Methods, Theory and Applications, 2023, Shangding Gu,Long Yang,Yali Du, Guang Chen, Florian Walter, Jun Wang, Yaodong Yang, Alois Knoll [Available from https://arxiv.org/pdf/2205.10330.pdf]

S. P. Vimal, Department of CSIS, WILP Division (vimalsp@wilp.bits-pilani.ac.in)

Safety

BIG QUESTION:
How do we guarantee safety when we deploy RL in real-world applications?

S. P. Vimal, Department of CSIS, WILP Division (vimalsp@wilp.bits-pilani.ac.in)

Safety

Some Definitions of Safety:

(1) the condition of being protected from or unlikely to cause danger, risk, or injury [oxford]
(2) being protected from harm or other dangers
(3) controlling recognized dangers to attain an acceptable level of risk

Safety

2H3W problems:

(1) **Safety Policy.** How can we perform policy optimisation to search for a safe policy?
(2) **Safety Complexity**. How much training data is required to find a safe policy?
(3) **Safety Applications.** What is the up to date progress of safe RL applications?
(4) **Safety Benchmarks**. What benchmarks can we use to fairly and holistically examine safe RL performance?
(5) **Safety Challenges**. What are the challenges faced in future safe RL research?

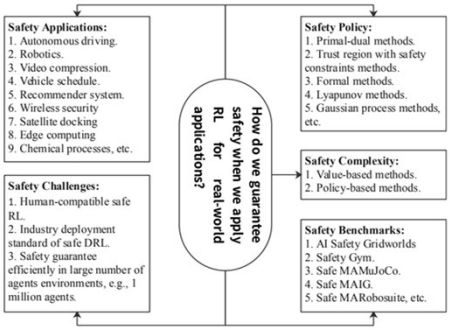
Safety

Some Definitions of Safety in the RL Sense:

(1) humans need to label environmental states as “safe” or “unsafe,” and agents are considered “safe” if “they never reach unsafe states”.
(2) agents are safe if “they act, reason, and generalize obeying human desires”

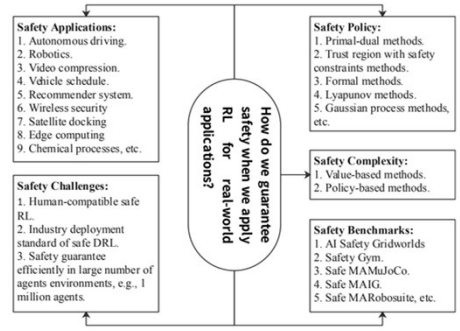
Safety

2H3W problems:



Safety

2H3W problems:



Defining SafeRL

MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \rho_0, \gamma)$

State Value Function

$$V_\pi(s) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_{t+1} | s_0 = s]$$

State-Action Value Function

$$Q_\pi(s,a) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_{t+1} | s_0 = s, a_0 = a]$$

Advantage Function

$$A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s)$$

Cost Function

$$J(\pi) = \mathbb{E}_{s\sim\rho_0(\cdot)}[V_\pi(s)]$$

Defining SafeRL

MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \rho_0, \gamma)$

discount factor

$\mathbb{P}(s'|s,a)$ Transition Model

$r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ Reward Function

$\rho_0(\cdot) : \mathcal{S} \to [0,1]$ Starting state distribution

Defining SafeRL

Constrained

CMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \rho_0, \gamma)$

MDP with safety Constraints

Defining SafeRL

Constrained

CMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \rho_0, \gamma)$

MDP with safety Constraints

$$\{(c_i, b_i)\}_{i=1}^{m}$$

Cost Value function for Constraint type i

Safety Constraint Bound.

Defining SafeRL

CMDP

Now define $V_\pi^{c_i}$, $Q_\pi^{c_i}$, $A_\pi^{c_i}$ for each c_i

$$V_\pi^{c_i}(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t c_i(s_t, a_t) | s_0 = s \right]$$

$$\{(c_i, b_i)\}_{i=1}^{m}$$

Cost Value function for Constraint type i

Safety Constraint Bound.

Defining SafeRL

CMDP

Now define $V_\pi^{c_i}$, $Q_\pi^{c_i}$, $A_\pi^{c_i}$ for each c_i

$$\{(c_i, b_i)\}_{i=1}^{m}$$

Cost Value function for Constraint type i

Safety Constraint Bound.

Defining SafeRL

CMDP

Now define $V_\pi^{c_i}$, $Q_\pi^{c_i}$, $A_\pi^{c_i}$ for each c_i

$$V_\pi^{c_i}(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t c_i(s_t, a_t) | s_0 = s \right]$$

$$\{(c_i, b_i)\}_{i=1}^{m}$$

Compare!

$$V_\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} | s_0 = s \right]$$

Defining SafeRL

CMDP

Now define $V_\pi^{c_i}$, $Q_\pi^{c_i}$, $A_\pi^{c_i}$ for each c_i

$$V_\pi^{c_i}(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t c_i(s_t, a_t) | s_0 = s\right]$$

$$\{(c_i, b_i)\}_{i=1}^m$$

Expected Cost function

$$C_i(\pi) = \mathbb{E}_{s\sim\rho_0(\cdot)}[V_\pi^{c_i}(s)]$$

Defining SafeRL

CMDP

Constraint Bound

$$V_\pi^{c_i}(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t c_i(s_t, a_t) | s_0 = s\right]$$

$$\Pi_C = \cap_{i=1}^m \{\pi \in \Pi_S \text{ and } C_i(\pi) \leq b_i\}$$

Feasible Policy Set

Expected Cost function

$$C_i(\pi) = \mathbb{E}_{s\sim\rho_0(\cdot)}[V_\pi^{c_i}(s)]$$

Defining SafeRL

CMDP

$$V_\pi^{c_i}(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t c_i(s_t, a_t) | s_0 = s\right]$$

Expected Cost function

$$C_i(\pi) = \mathbb{E}_{s\sim\rho_0(\cdot)}[V_\pi^{c_i}(s)]$$

Defining SafeRL

CMDP

$$\Pi_C = \cap_{i=1}^m \{\pi \in \Pi_S \text{ and } C_i(\pi) \leq b_i\}$$

Defining SafeRL

CMDP

$$\max_{\pi \in \Pi_S} J(\pi), \text{ such that } \mathbf{c}(\pi) \leq \mathbf{b}$$

Maximize reward performance; Subject to Cost constraint; Vector of C_i's; Vector of bounds

$$\Pi_C = \cap_{i=1}^{m} \{\pi \in \Pi_S \text{ and } C_i(\pi) \leq b_i\}$$

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

SafeRL – Categorization

Model Based SafeRL

Model-free SafeRL

Defining SafeRL

CMDP

$$\max_{\pi \in \Pi_S} J(\pi), \text{ such that } \mathbf{c}(\pi) \leq \mathbf{b}$$

Maximize reward performance; Subject to Cost constraint; Vector of C_i's; Vector of bounds

$$\Pi_C = \cap_{i=1}^{m} \{\pi \in \Pi_S \text{ and } C_i(\pi) \leq b_i\}$$

CMDP Optimization Function

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

Applications | Autonomous Driving

Safe Reinforcement Learning on Autonomous Vehicles

David Isele, Alireza Nakhaei, and Kikuo Fujimura
Honda Research Institute USA
{disele, anakhaei, kfujimura}@honda-ri.com

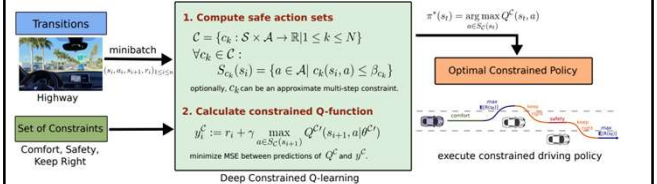
Abstract— There have been numerous advances in reinforcement learning, but the typically unconstrained exploration of the learning process prevents the adoption of these methods in many safety critical applications. Recent work in safe reinforcement learning uses idealized models to achieve their guarantees, but these models do not easily accommodate the stochasticity or high-dimensionality of real world systems. We investigate how prediction provides a general and intuitive framework to constraint exploration, and show how it can be used to safely learn intersection handling behaviors on an autonomous vehicle.

Fig. 1. An autonomous vehicle navigating an intersection. Prediction is used to shield the vehicle from making dangerous decisions, while allowing it to learn policies that are both efficient and not disruptive to other vehicles.

Applications | Autonomous Driving

Deep Constrained Q-learning

Gabriel Kalweit[1,*], Maria Huegle[1,*], Moritz Werling[2], Joschka Boedecker[1,3]
[1]Neurorobotics Lab, University of Freiburg, Germany
[2]BMWGroup, Germany
[3]BrainLinks-BrainTools, University of Freiburg, Germany



Applications | Autonomous Driving

Deep Constrained Q-learning

Gabriel Kalweit[1,*], Maria Huegle[1,*], Moritz Werling[2], Joschka Boedecker[1,3]

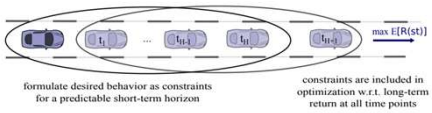
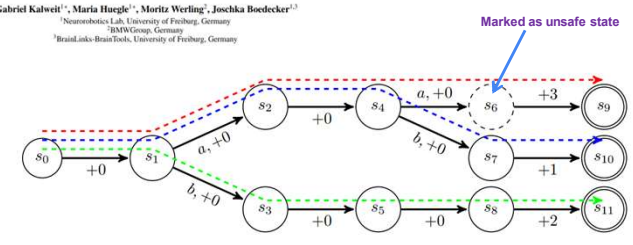


Figure 2: Methodology of Constrained Q-learning, exemplary for the application of autonomous driving. Traffic rules are ensured in predictable short-term horizon. Long-term goals are optimized by optimization of long-term return.

Applications | Autonomous Driving

Deep Constrained Q-learning

Gabriel Kalweit[1,*], Maria Huegle[1,*], Moritz Werling[2], Joschka Boedecker[1,3]
[1]Neurorobotics Lab, University of Freiburg, Germany
[2]BMWGroup, Germany
[3]BrainLinks-BrainTools, University of Freiburg, Germany

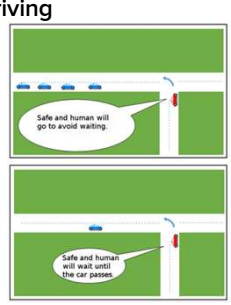


Applications | Autonomous Driving

Safe Reinforcement Learning on Autonomous Vehicles

David Isele, Alireza Nakhaei, and Kikuo Fujimura
Honda Research Institute USA
{disele, anakhaei, kfujimura}@honda-ri.com

The specific application we investigate is making a turn at an unsigned intersection. This problem was recently explored as a non-safety constrained RL domain [12] where it was noted that the learned policy, which optimized efficiency, might be disruptive to traffic vehicles in practice. The primary concerns of these maneuvers are safety and efficiency, but balancing the two is a dynamic task. In dense traffic we may wish to seize an opportunity that leaves only several meters of a safety margin, as we might have to wait an unacceptable amount of time for the next opportunity. However in sparse traffic, adding an increased margin will only add a negligible delay and will likely be preferable to the passengers and other traffic vehicles. Figure 2 demonstrates the scenarios in dense and sparse traffic.

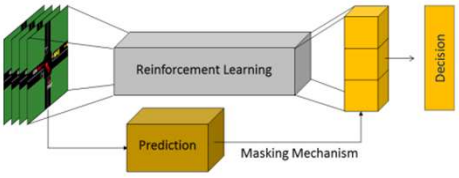


Applications | Autonomous Driving

Safe Reinforcement Learning on Autonomous Vehicles

David Isele, Alireza Nakhaei, and Kikuo Fujimura
Honda Research Institute USA
{disele, anakhaei, kfujimura}@honda-ri.com

https://arxiv.org/pdf/1910.00399.pdf



BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Thank you

Required Readings and references

1. Deep Constrained Q-learning Gabriel Kalweit, Maria Huegle, Moritz Werling,Joschka Boedecker
2. Safe Reinforcement Learning on Autonomous Vehicles David Isele, Alireza Nakhaei, and Kikuo Fujimura Honda Research Institute USA
3. A Review of Safe Reinforcement Learning: Methods, Theory and Applications, 2023, https://arxiv.org/pdf/2205.10330.pdf