Constrained Reinforcement Learning

Minseok Seo [*], Kyunghwan Choi [*]

July 10, 2025
Version 0.0

Abstract

This documentation provides an overview of constrained reinforcement learning (CRL). CRL is a subfield of reinforcement learning that focuses on learning policies that satisfy certain constraints while maximizing a reward signal. The goal is to ensure that the learned policies not only achieve high rewards but also adhere to predefined constraints. In this document, we will cover the key concepts of CRL and previous research in this field.

Contents

1 Introduction

Reinforcement Learning (RL) has achieved significant advancements in recent years, particularly in the context of complex decision-making tasks. However, the depoyment of RL agents in real-world applications still have several challenges. One of the main challenges is that we cannot ensure the safety of RL agents. This is largely due to the presence of scenarios that were not encountered during training, as well as differences between the training and deployment environments. To address this issue, many approaches have been proposed to ensure the safety of RL agents. In this document, we focus on the Constrained Reinforcement Learning (CRL) framework.

1.1 Research Objectives

2 Background

2.1 Reinforcement Learning

Reinforcement Learning (RL) is a framework in which an agent interacts with an environment and learns a policy to maximize cumulative rewards. The agent observes the state of the environment, takes actions, and recieves rewards based on those actions. This process is formalized as a Markov Decision Process (MDP) [1], which provides a formal structure for modeling decision-making problems. An MDP is defined by a tuple $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, P is the state transition probability function, R is the reward function, and $\gamma \in [0, 1)$ is the discount factor. In this document, we consider a finite-horizon

[*]Minseok Seo is with the AI Graduate School, Gwangju Institute of Science and Technology (GIST), and Kyunghwan Choi is with the Cho Chun Shik Graduate School of Mobility, Korea Advanced Institute of Science and Technology (KAIST) (e-mail: seominseok@gm.gist.ac.kr and kh.choi@kaist.ac.kr).

setting and use the undiscounted return. The objective of RL is to find an optimal policy π^* that maximizes the expected cumulative reward, defined as:

$$\begin{aligned} \theta^* &= \arg\max_\theta J(\theta) \\ J(\theta) &= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} r_t\right] \end{aligned} \tag{1}$$

The policy π_θ is assumed to be differentiable function parameterized by θ, denoted as $\pi_\theta(a|s)$, which represents the probability of taking action a given state s. The expectation $\mathbb{E}_{\tau\sim\pi_\theta}$ is taken over the trajectories $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \cdots, s_T)$ generated by following the policy π_θ.

2.2 Constrained Reinforcement Learning

Constrained Reinforcement Learning (CRL) extends the standard RL framework by incorporating constraints into the learning process. CRL is formalized as a Constrained Markov Decision Process (CMDP) [2], which is defined by a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \mathcal{C}, \gamma \rangle$. Here, $\mathcal{C}$ is the cost functions associated with the constraints. The feasible policy set in a CMDP is given by:

$$\Pi_C = \{\pi : J_{c_i}(\pi) \leq d_i, \quad i = 1, \cdots, k\} \tag{2}$$

where $J_{c_i}(\pi)$ is a cost-based constraint function defined the expected cumulative cost, and d_i is the threshold for the i-th constraint. The objective of CRL is to find an optimal policy that maximizes the expected cumulative reward while satisfying the constraints. In the context of policy gradient methods, the constrained optimization problem can be formulated as follows:

$$\begin{aligned} \theta^* &= \arg\max_\theta J(\theta) \\ J(\theta) &= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} r_t\right] \text{ subject to } \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} c_t\right] \leq d \end{aligned} \tag{3}$$

2.2.1 Lagrangian Method

Constrained optimization problem defined in Eq. 3 can be solved using various methods. In this thesis, however, we consider only the Lagrangian method. By applying Lagrangian relaxation, the constrained optimization problem can be transformed into an unconstrained optimization problem, in which the constraint is incorporated into the objective function using a Lagrange multiplier λ.

$$\begin{aligned} \theta^* &= \arg\max_\theta \mathcal{L}(\theta, \lambda) \\ \mathcal{L}(\theta, \lambda) &= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} r_t\right] - \lambda\left(\mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} c_t\right] - d\right) \end{aligned} \tag{4}$$

The Lagrange multiplier λ is a non-negative scalar that adjusts the trade-off between maximizing the expected cumulative reward and satisfying the constraints.

2.3 State-wise Constrained Reinforcement Learning

State-wise Constrained Reinforcement Learning (SCRL) is a variant of CRL that imposes constraints at the state level. CRL considers the cumulative cost over the entire trajectory, while SCRL focuses on the cost at each transition. SCRL is formalized as a State-wise Constrained Markov Decision Process (SCMPD), it is quite similar to CMDP, but SCMDP enforces the constraint for every state action trasition satisfies a hard constraints. The objective of SCRL is to find an optimal policy that maximizes the expected cumulative reward while satisfying the state-wise constraints.

$$\begin{aligned} \pi^* &= \arg\max_{\pi_\theta} J(\theta) \\ J(\theta) &= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} r_t\right] \text{ subject to } \mathbb{E}_{\tau\sim\pi_\theta}[c(s,a)] \leq w, \quad \forall s \in S \end{aligned} \tag{5}$$

3 Conclusion

References

[1] M. L. Puterman, "Markov decision processes," *Handbooks in operations research and management science*, vol. 2, pp. 331–434, 1990.

[2] E. Altman, *Constrained Markov decision processes*. Routledge, 2021.