Paper Critique

Shuvrajeet Das, DA24D402

Course: DA7400, Fall 2024, IITM
Paper: [Curriculum Induction for Safe Reinforcement Learning]
Date: [30-10-2024]

Make sure your critique Address the following points:
1. The problem the paper is trying to address
2. Key contributions of the paper
3. Proposed algorithm/framework
4. How the proposed algorithm addressed the described problem
Note: Be concise with your explanations. Unnecessary verbosity will be penalized. Please don't exceed 2 pages.

---

1 The problem the paper is trying to address

The paper addresses the problem of safe reinforcement learning in constrained environments, framed as a constrained Markov Decision Process (CMDP). The objective is to find an optimal policy that maximizes the expected cumulative reward while keeping constraint violations under a specified threshold.

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\rho_\pi}\left[\sum_{t=0}^{T} r(s_t, a_t, s_{t+1})\right] \text{ s.t. } \mathbb{E}_{\rho_\pi}\left[\sum_{t=0}^{T} \mathbb{I}(s_t \in D)\right] \leq \kappa, \tag{1}$$

2 Key contributions of the paper

- Introduces Curriculum Induction for Safe Reinforcement Learning (CISR), a framework for safe RL that enables agents to learn without violating safety constraints.
- Proposes a curriculum policy optimization method where a teacher adapts intervention strategies over multiple learning rounds, helping the agent to safely improve performance.
- Demonstrates that CISR-trained agents achieve comparable or superior rewards to traditional RL agents while maintaining safety throughout training.
- Shows that CISR policies are transferable, allowing safe RL in new environments and with agents of varying architectures.

3 Proposed algorithm/framework

Algorithm 1 Curriculum Induction for Safe Reinforcement Learning (CISR)

Require: Interventions set $\mathcal{I}$, initial teacher policy $\pi_{T,0}$

for each round $j = 0, 1, \ldots, N_t$ **do**
  Initialize new student policy $\pi_{0,j} \leftarrow$ get_student()
  for each curriculum step $n = 0, 1, \ldots, N_s$ **do**
    Choose intervention $M_{i_n} \leftarrow \pi_{T,j}(o_{T,0}, \ldots, o_{T,n})$
    if $n > 0$ **then**
      $\pi_{n,j} \leftarrow \text{transfer}(\pi_{n-1,j})$
    end if
    Train student in CMDP M_{i_n}: $\pi_{n,j} \leftarrow \text{student.train}(M_{i_n})$
    Observe student performance: $o_{T,n} \leftarrow \phi(\pi_{n,j})$
  end for
  Update teacher policy: $\pi_{T,j+1} \leftarrow \text{teacher.train}(\{(\pi_{T,k}, V(\pi_{N_s,k}))\}_{k=1}^{j})$
end for

4 How the proposed algorithm addressed the problem

The proposed CISR algorithm addresses the safe RL problem by structuring the agent's learning process as a series of constrained Markov Decision Processes (CMDPs) with progressively reduced reliance on safety interventions. This ensures safety and maximizes reward by:

- **Curriculum-Based Learning**: The algorithm introduces safety interventions gradually, enabling the agent to explore safely and learn without incurring constraint violations.
- **Teacher Optimization of Curriculum Policy**: By evaluating each student's performance after training rounds, the teacher optimizes the curriculum policy for selecting interventions. This dynamic adaptation enables faster and safer learning by choosing interventions tailored to the agent's current performance.
- **Constraint Satisfaction during Learning**: Each CMDP in the curriculum guarantees that the agent does not exceed the safety threshold, ensuring that any policy learned is feasible and safe in the original environment.
- **Knowledge Transfer across Students**: The algorithm refines its intervention strategy by learning from the experience of multiple students, enabling it to optimize a curriculum policy that can be effectively transferred to new agents or environments.