Safe Reinforcement Learning

Minseok Seo[1],

[1]Mobility Intelligence and Control Laboratory (MIC Lab)
CCS Graduate School of Mobility
Korea Advanced Institute of Science and Technology (KAIST)

Mobility Intelligence and Control Laboratory

September 5, 2025

Outline

Figure: Waymo and Tesla

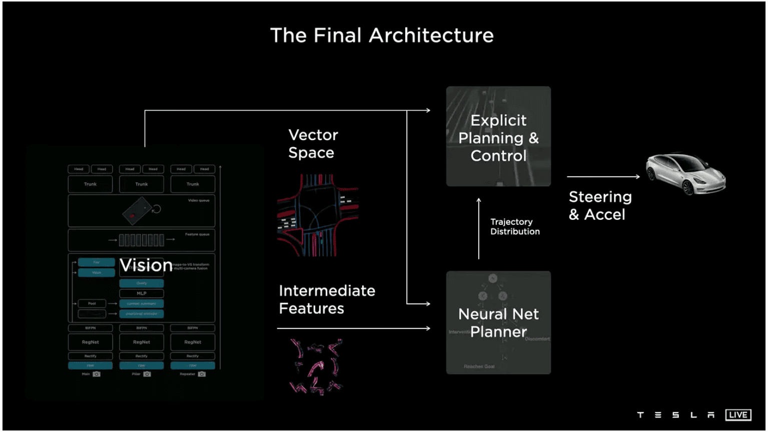


Figure: Tesla's Architecture in 2021 (source: AI Day 2021)

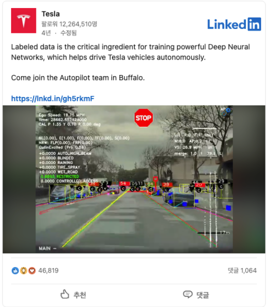


Figure: Tesla's recruitment post for data labeling positions

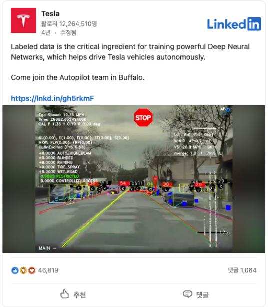


Figure: Tesla's recruitment post for data labeling positions

- Supervised learning requires a large amount of labeled data.

Motivation Supervised Learning

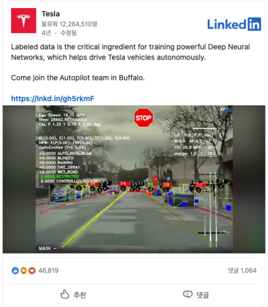


Figure: Tesla's recruitment post for data labeling positions

- Supervised learning requires a large amount of labeled data.
- Since it is created by humans, it is expensive.

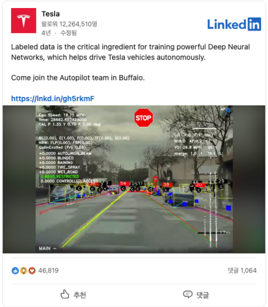


Figure: Tesla's recruitment post for data labeling positions

- Supervised learning requires a large amount of labeled data.
- Since it is created by humans, it is expensive.
- The performance of supervised learning is limited by human-labeled data.

Insert Reinforcement Learning example

Insert Reinforcement Learning framework

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

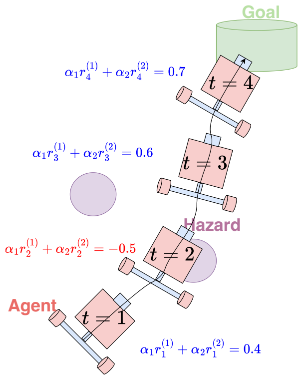


Figure: Reinforcement Learning Framework

1. Needs simulation, not labeled data

1. Needs simulation, not labeled data
2. Works in simulation, but may fail in the real world
    - Even after training, the agent may still take unsafe actions

1. Needs simulation, not labeled data
2. Works in simulation, but may fail in the real world
   - Even after training, the agent may still take unsafe actions
3. Reward engineering can induce desired behaviors (multi-objective), but it is time-consuming and difficult

1. Needs simulation, not labeled data
2. Works in simulation, but may fail in the real world
    - Even after training, the agent may still take unsafe actions
3. Reward engineering can induce desired behaviors (multi-objective), but it is time-consuming and difficult

My research

Challenges 2 & 3: learning safe policies without relying on reward engineering

Outline

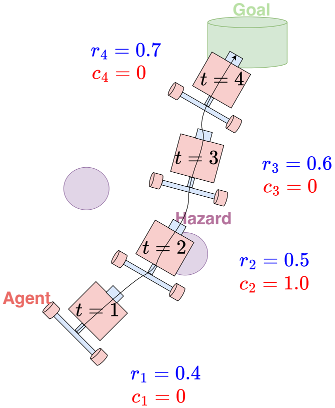


Figure: CMDP example — agent trajectory with rewards (blue) and costs (red)

Constrained Policy Optimization Problem

$$
\begin{aligned}
\theta^* &= \arg\max_{\theta} J(\theta) \\
J(\theta) &= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} r_t\right] \text{ subject to } \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} c_t\right] \leq d
\end{aligned}
\tag{2}
$$

Constrained Policy Optimization Problem

$$
\begin{aligned}
\theta^* &= \arg\max_{\theta} J(\theta) \\
J(\theta) &= \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T} r_t\right] \text{ subject to } \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T} c_t\right] \leq d
\end{aligned} \tag{2}
$$

- Find a policy that maximizes rewards while satisfying constraints.

Constrained Policy Optimization Problem

$$\theta^* = \arg\max_{\theta} J(\theta)$$
$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T} r_t\right] \text{ subject to } \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T} c_t\right] \leq d \tag{2}$$

- Find a policy that maximizes rewards while satisfying constraints.
- Directly solving constrained optimization is difficult.

Constrained Policy Optimization Problem

$$\theta^* = \arg\max_\theta J(\theta)$$
$$J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} r_t\right] \text{ subject to } \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} c_t\right] \le d \tag{2}$$

- Find a policy that maximizes rewards while satisfying constraints.
- Directly solving constrained optimization is difficult.
- By applying Lagrangian relaxation, we can convert it to an unconstrained problem.

Constrained Policy Optimization Problem

$$\theta^* = \arg\max_{\theta} \mathcal{L}(\theta, \lambda)$$
$$\mathcal{L}(\theta, \lambda) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T} r_t\right] - \lambda \left(\mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T} c_t\right] - d\right) \tag{3}$$

$$\lambda \leftarrow \left[\lambda + \beta \left(\hat{J}_c - d\right)\right]_+ \tag{4}$$

Penalty (λ) increases when constraints are violated $\rightarrow$ policy is encouraged to satisfy them.

In CMDPs, constraints are imposed on the trajectory level.

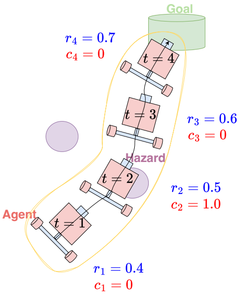


Figure: CMDP example — agent trajectory with rewards (blue) and costs (red)

In CMDPs, constraints are imposed on the trajectory level.

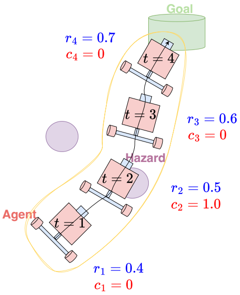


Figure: CMDP example — agent trajectory with rewards (blue) and costs (red)

Wouldn't imposing constraints at the state level allow for more precise constraint enforcement?

In state-wise constrained MDPs, constraints are imposed on each state.

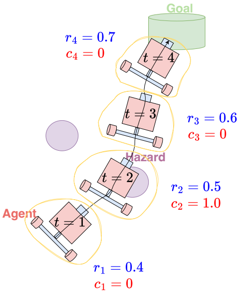


Figure: State-wise constrained MDP example — agent trajectory with rewards (blue) and costs (red)

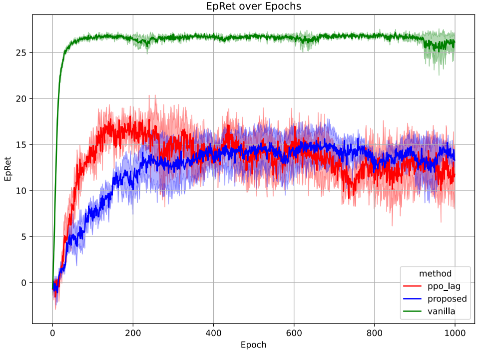
$$\pi^* = \arg\max_{\pi_\theta} J(\theta)$$
$$J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} r_t\right] \text{ subject to } \mathbb{E}_{\tau\sim\pi_\theta}[c(s,a)] \le w, \quad \forall s \in S \tag{5}$$

$$\lambda(s) \leftarrow \lambda(s) + \beta(\hat{J}_c - w) \tag{6}$$

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

$$\lambda(s) \leftarrow \lambda(s) + \beta(\hat{J}_c - w) \tag{6}$$

The Lagrange multiplier is replaced by a neural network's output instead of a scalar.



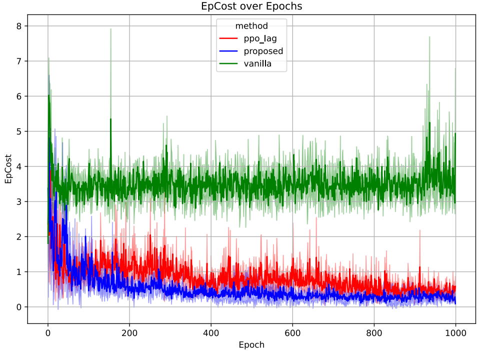


Figure: Performance comparison on Safety Gym tasks

Insert video

Limitations of the Proposed Approach

- Sensitivity to Lagrange multiplier (init., learning rate)

Limitations of the Proposed Approach

- Sensitivity to Lagrange multiplier (init., learning rate)
- Constraint threshold setting is difficult
  - Too strict $\rightarrow$ overly conservative policy
  - Too lenient $\rightarrow$ constraints not enforced

Limitations of the Proposed Approach

- Sensitivity to Lagrange multiplier (init., learning rate)
- Constraint threshold setting is difficult
  - Too strict $\rightarrow$ overly conservative policy
  - Too lenient $\rightarrow$ constraints not enforced
- Needs sufficient violations during training

Summary

Summary

- Major companies focus on supervised learning for autonomous driving

Summary

- Major companies focus on supervised learning for autonomous driving
- Supervised learning is costly and data-dependent

Summary

- Major companies focus on supervised learning for autonomous driving
- Supervised learning is costly and data-dependent
- Reinforcement learning can learn policies in simulation and transfer to the real world

Summary

- Major companies focus on supervised learning for autonomous driving
- Supervised learning is costly and data-dependent
- Reinforcement learning can learn policies in simulation and transfer to the real world
- Risk of unsafe behavior due to the Sim2Real gap

Summary

- Major companies focus on supervised learning for autonomous driving
- Supervised learning is costly and data-dependent
- Reinforcement learning can learn policies in simulation and transfer to the real world
- Risk of unsafe behavior due to the Sim2Real gap
- Research direction: state-wise constrained policy learning

Summary

- Major companies focus on supervised learning for autonomous driving
- Supervised learning is costly and data-dependent
- Reinforcement learning can learn policies in simulation and transfer to the real world
- Risk of unsafe behavior due to the Sim2Real gap
- Research direction: state-wise constrained policy learning

But there are still challenges to solve..

Future Work

Future Work

- Extension to Model-Based RL

Thank you for your attention!

References I

If you need! or remove this frame.