Advanced Control for Robotics (Fall 2025)

Lecture Note 10
Policy Gradients for Reinforcement Learning

Prof. Wei Zhang
Southern University of Science and Technology

- We have learned:
  - Markov chain
  - Markov Decision Process
  - Monte Carlo Methods
- This lecture:
  - Policy gradient methods for RL

RL Problem

- Find optimal policy π

$$V^*(s) = \max_{\pi} E_{\tau \sim \pi}[R(\tau) | s_0 = s]$$

- We often parameterize policy by a parameter vector θ

- Notations:
  - $\pi_\theta := \pi_\theta(\cdot \,| s)$

  - Denote: $P_\theta(\tau)$ as the likelihood of trajectory τ under policy π_θ

- Reformulate the MDP problem as an optimization problem:
  - Assume s_0 distribution is included in trajectory likelihood $P_\theta(\tau)$
  - Utility function:

$$U(\theta) = E_{\tau \sim P_\theta(\tau)}[R(\tau)] = \sum_\tau P_\theta(\tau)\, R(\tau)$$

  - Under new notation, RL problem reduces to finding the optimal policy parameter θ

$$\max_\theta U(\theta) = \max_\theta \sum_\tau P_\theta(\tau) R(\tau)$$

  - Policy gradient type of methods use 1st order gradient ascend method to solve the above optimization problem.
    - They differ in ways of computing/estimating the gradient $\nabla_\theta U(\theta)$

- Derivation of policy gradient

- Derivation of policy gradient

- Derivation of policy gradient

Summary of Policy Gradient Derivation

- Roll out trajectories $\tau^{(i)} \sim P_\theta(\cdot),\ i = 1, \cdots, N$

- Compute the empirical mean: $\hat{g} = \frac{1}{N}\sum_i \nabla_\theta \left(\sum_t \log \pi_\theta \left(a_t^{(i)} \middle| s_t^{(i)}\right)\right) R\left(\tau^{(i)}\right)$

- By Monte Carlo: we know $E(\hat{g}) = \nabla_\theta U(\theta\ )$

- In practice, the sample mean estimate $\hat{g}$ has a high variance

- Many ways can be used to reduce the variance, and lead to different algorithms.

- REINFORCE Algorithm

1. Roll out trajectories $\{\tau_i\}_{i=1}^{N}$ from π_θ
2. Compute $\nabla_\theta \mathrm{U}(\theta) = \frac{1}{N}\sum_i \left(\sum_t \nabla_\theta \log(\pi_\theta\left(a_t^i \middle| s_t^i\right))\right) \left(\sum_t r(s_t^i, r_t^i)\right)$
3. $\theta \leftarrow \theta + \alpha \nabla_\theta U(\theta)$

- **EGLP Lemma:**

- **EGLP Corollary:**
  - $\tau_{0:t} = (s_0, a_0, \dots, s_{t-1}, a_{t-1}, s_t)$ be the partial trajectory (historical data update to t
  - Let $f(\tau_{0:t})$ be an arbitrary function of partial trajectory
  - $E_{\tau \sim \pi_\theta}[\nabla_\theta \log \pi_\theta(A_t|S_t) \cdot f(\tau_{0:t})] = 0$

- Policy gradient with temporal structure