FPT Education
FPT UNIVERSITY

Policy Gradient for Continuing Tasks

Objectives

- ❑ Describe the objective for policy gradient algorithms.
- ❑ Describe the policy gradient theorem

FPT Education
FPT UNIVERSITY

Objective for Learning Policies

- What is the goal of Agent?
- The goal of Reinforcement Learning is maximizing the reward in the long run.

$$R_t, R_{t+1}, R_{t+2}, \cdots$$

FPT Education
FPT UNIVERSITY

Objective for Learning Policies

- Formalizing the gold as an Objective
  - Episodic: the sum of rewards over a whole episode

$$G_t = \sum_{t=0}^{T} R_t$$

  - Continuing:
    - The discounted return

$$G_t = \sum_{t=0}^{\infty} \gamma^t R_t$$

    - Sum of the differences between the immediate reward and its avera

$$G_t = \sum_{t=0}^{\infty} R_t - r(\pi)$$

FPT Education
FPT UNIVERSITY

Objective for Learning Policies

- The average reward objective

$$r(\pi) = \sum_s \mu(s) \sum_a \pi(a \mid s, \boldsymbol{\theta}) \sum_{s',r} p(s', r \mid s, a) r$$

- The overall average reward by considering the fraction of time we spend in state S under policy Pi.
- The expected reward is a sum over S of the expected reward in a state weighted by Mu of S, r Pi is our average reward learning objective.

FPT Education
FPT UNIVERSITY

Objective for Learning Policies

- Optimizing the average reward objective
  - Our goal of policy optimization will be to find a policy which maximizes the average reward.

$$\nabla r(\pi) = \nabla \sum_{s} \mu(s) \sum_{a} \pi(a \mid s, \boldsymbol{\theta}) \sum_{s',r} p(s', r \mid s, a) r$$

Objective for Learning Policies

- The challenge of the policy gradient methods
  - The main difficulty is that modifying our policy changes the distribution Mu.

$$\nabla_{\boldsymbol{\theta}} r(\pi) = \nabla_{\boldsymbol{\theta}} \sum_s \mu(s) \sum_a \pi(a \mid s, \boldsymbol{\theta}) \sum_{s',r} p(s', r \mid s, a) r$$

Depends on $\boldsymbol{\theta}$

  - It does not change as the weights and the parameterized value function chains

$$\begin{aligned}\nabla_{\mathbf{w}} \overline{VE} &= \nabla_{\mathbf{w}} \sum_s \mu(s)[v_\pi(s) - \hat{v}(s, \mathbf{w})]^2 \\ &= \sum_s \mu(s) \nabla_{\mathbf{w}} [v_\pi(s) - \hat{v}(s, \mathbf{w})]^2\end{aligned}$$

FPT Education
FPT UNIVERSITY

Objective for Learning Policies

- The objective of policy gradient algorithms:
  - It is to directly optimize the parameters of a parameterized policy in order to maximize the expected cumulative rewards obtained by an agent in an environment.
  - It aim to directly learn a policy that selects actions based on the observed states.
  - It compute an estimate of the gradient of the expected return with respect to the policy parameters.

The Policy Gradient Theorem

- It provides a theoretical foundation for optimizing parameterized policies using gradient-based methods.

- The theorem establishes a relationship between the expected return of a policy and the gradient of the policy parameters with respect to this expected return.

The Policy Gradient Theorem

- The policy gradient theorem:

$$\nabla r(\pi) = \sum_s \mu(s) \sum_a \nabla \pi(a \mid s, \boldsymbol{\theta}) q_\pi(s, a)$$

The Policy Gradient Theorem

- The gradient of the policy parameters is estimated using samples obtained through interactions with the environment.

- By directly optimizing the policy parameters along the direction of the gradient, these methods aim to improve the policy's performance and maximize the expected return over time.

FPT Education
FPT UNIVERSITY

Summary

- Describe the objective for policy gradient algorithms.
- Describe the policy gradient theorem

Q & A