Reinforcement Learning

Algorithms

Felipe Costa

Related material at:
https://github.com/fe-lipe-c/rl_algorithms

Reinforcement Learning Algorithms

Policy Gradient Methods

In policy-based RL, the policy is a function π_θ, where θ are the parameters.[1]

Advantage Actor-Critic (A2C)

A2C is a policy gradient algorithm that combines the actor-critic architecture with advantage functions to improve sample efficiency and stability in reinforcement learning. A2C belongs to the family of actor-critic methods, which maintain two components:

- An **actor** (policy network) that decides which actions to take
- A **critic** (value network) that evaluates how good those actions are

The key innovation in A2C is using the "advantage function" to reduce variance in policy gradient updates while maintaining an unbiased estimate of the gradient.

The foundation of A2C is the policy gradient theorem. For a policy π_θ parameterized by θ, the gradient of the expected return $J(\theta)$ is:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \cdot R_t \right] \tag{1}$$

Where: - τ is a trajectory $(s_0, a_0, r_0, s_1, a_1, r_1, \dots)$ - R_t is the return (discounted sum of rewards) from time t

References

[1] Csaba Szepesvári. *Algorithms of Reinforcement Learning*. 1st ed. Morgan and Claypool, 2019.