Online and Reinforcement Learning (2025)
Home Assignment 4

Davide Marchi 777881

Contents

1 Policy Gradient Methods

1.1 Baseline

We are given that the policy gradient theorem can be generalized to include an arbitrary baseline $b(s)$:

$$\nabla_\theta J(\pi) = \sum_{s \in S} \mu_\pi(s) \sum_{a \in A} \nabla_\theta \pi(s,a) \left(Q_\pi(s,a) - b(s) \right),$$

where:

- S is the state space.
- A is the action space.
- $\pi(s,a)$ is the probability of choosing action a in state s.
- $\mu_\pi(s)$ is the stationary state distribution under policy π.
- $Q_\pi(s,a)$ is the state-action value function.

The term

$$\sum_{a \in A} \nabla_\theta \pi(s,a) b(s)$$

acts as a control variate, and we must show that its expectation is zero, i.e.,

$$\mathbb{E}\left[\sum_{a \in A} \nabla_\theta \pi(s,a) b(s) \right] = 0.$$

Proof

For any state $s \in S$, note that $\pi(s, \cdot)$ is a probability distribution over A. Therefore, by definition:

$$\sum_{a \in A} \pi(s,a) = 1.$$

Differentiating both sides of the equation with respect to θ, we obtain:

$$\sum_{a \in A} \nabla_\theta \pi(s,a) = \nabla_\theta \left(\sum_{a \in A} \pi(s,a) \right) = \nabla_\theta(1) = 0.$$

Since $b(s)$ does not depend on the action a, it can be factored out of the summation:

$$\sum_{a \in A} \nabla_\theta \pi(s,a)\, b(s) = b(s) \sum_{a \in A} \nabla_\theta \pi(s,a) = b(s) \cdot 0 = 0.$$

Taking the expectation with respect to the stationary distribution $\mu_\pi(s)$, we have:

$$\mathbb{E}_{s\sim\mu_\pi}\left[\sum_{a\in A}\nabla_\theta\pi(s,a)b(s)\right] = \sum_{s\in S}\mu_\pi(s)\cdot 0 = 0.$$

Thus, we conclude that

$$\mathbb{E}\left[\sum_{a\in A}\nabla_\theta\pi(s,a)b(s)\right] = 0.$$

2 Improved Parametrization of UCB1

(Optional, but highly recommended)

3 Introduction of New Products

4 Empirical comparison of FTL and Hedge