Online and Reinforcement Learning (2025)
Home Assignment 6

Davide Marchi 777881

Contents

1 PPO **2**

1.1 Return expressed as advantage over another policy 2

1.2 Clipping 3

1.3 Pi prime in PPO 4

2 Offline Evaluation of Bandit Algorithms **4**

2.1 Part 1 4

1 PPO

1.1 Return expressed as advantage over another policy

We wish to prove that the expected return of a policy π can be written as

$$J(\pi) = J(\pi_{\text{ref}}) + \mathbb{E}_{s_0 \sim p_0, \pi}\left[\sum_{t=0}^{\infty} \gamma^t \, A^{\pi_{\text{ref}}}(s_t, a_t)\right], \tag{1}$$

where the advantage function is defined by

$$A^{\pi_{\text{ref}}}(s, a) = Q^{\pi_{\text{ref}}}(s, a) - V^{\pi_{\text{ref}}}(s).$$

Proof: Recall that for any state s the state-value function of π_{ref} is given by

$$V^{\pi_{\text{ref}}}(s) = \mathbb{E}_{a \sim \pi_{\text{ref}}}\left[Q^{\pi_{\text{ref}}}(s, a)\right].$$

Consider a trajectory $\tau = (s_0, a_0, s_1, a_1, \dots)$ generated by following π. For any finite horizon T, we can write

$$\begin{aligned}\sum_{t=0}^{T} \gamma^t A^{\pi_{\text{ref}}}(s_t, a_t) &= \sum_{t=0}^{T} \gamma^t \Big(Q^{\pi_{\text{ref}}}(s_t, a_t) - V^{\pi_{\text{ref}}}(s_t)\Big) \\ &= \sum_{t=0}^{T} \gamma^t \left(r(s_t, a_t) + \gamma \, V^{\pi_{\text{ref}}}(s_{t+1}) - V^{\pi_{\text{ref}}}(s_t)\right).\end{aligned}$$

Notice that the sum telescopes. To see this, rearrange the terms:

$$\sum_{t=0}^{T} \gamma^t r(s_t, a_t) + \sum_{t=0}^{T} \left(\gamma^{t+1} V^{\pi_{\text{ref}}}(s_{t+1}) - \gamma^t V^{\pi_{\text{ref}}}(s_t)\right).$$

The second sum is telescopic:

$$\sum_{t=0}^{T} \left(\gamma^{t+1} V^{\pi_{\text{ref}}}(s_{t+1}) - \gamma^t V^{\pi_{\text{ref}}}(s_t)\right) = -V^{\pi_{\text{ref}}}(s_0) + \gamma^{T+1} V^{\pi_{\text{ref}}}(s_{T+1}).$$

Assuming that $V^{\pi_{\text{ref}}}(s)$ is bounded and $\gamma \in (0, 1)$, as $T \to \infty$ we have $\gamma^{T+1} V^{\pi_{\text{ref}}}(s_{T+1}) \to 0$. Therefore,

$$\sum_{t=0}^{\infty} \gamma^t A^{\pi_{\text{ref}}}(s_t, a_t) = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) - V^{\pi_{\text{ref}}}(s_0).$$

Taking the expectation over trajectories starting from $s_0 \sim p_0$ (and using the definition $J(\pi) = \mathbb{E}[\sum_{t \geq 0} \gamma^t r(s_t, a_t)]$) gives

$$J(\pi) = \mathbb{E}_{s_0 \sim p_0, \pi}\left[\sum_{t=0}^{\infty} \gamma^t A^{\pi_{\text{ref}}}(s_t, a_t)\right] + \mathbb{E}_{s_0 \sim p_0}\Big[V^{\pi_{\text{ref}}}(s_0)\Big].$$

Since by definition $J(\pi_{\text{ref}}) = \mathbb{E}_{s_0 \sim p_0}\left[V^{\pi_{\text{ref}}}(s_0)\right]$, we obtain the desired result (1).

1.2 Clipping

In PPO the surrogate objective for a given state-action pair is defined as

$$L^{\text{CLIP}}(\theta) = \min\Bigg(r(\theta)\, \hat{A}^{\pi_{\text{ref}}}(s,a),\ \text{clip}\Big(r(\theta),\, 1-\epsilon,\, 1+\epsilon\Big)\, \hat{A}^{\pi_{\text{ref}}}(s,a) \Bigg),$$

where

$$r(\theta) = \frac{\pi_\theta(a|s)}{\pi_{\text{ref}}(a|s)}$$

and the clipping function is defined by

$$\text{clip}(x,\, l,\, u) = \min\{\max(x,l), u\}.$$

Intuitive discussion: The gradient update will change $\pi_\theta(a|s)$ (and hence $r(\theta)$) if either

(i) $r(\theta) \in [1-\epsilon,\, 1+\epsilon]$, or

(ii) when $r(\theta) \notin [1-\epsilon,\, 1+\epsilon]$, the *unclipped* term $r(\theta)\, \hat{A}^{\pi_{\text{ref}}}(s,a)$ has a gradient that *points toward* the interval $[1-\epsilon,\, 1+\epsilon]$.

The first condition is trivial. Now assume that $r(\theta) \notin [1-\epsilon, 1+\epsilon]$. There are two cases:

Case 1: $r(\theta) > 1+\epsilon$. Then

$$\text{clip}(r(\theta), 1-\epsilon, 1+\epsilon) = 1+\epsilon.$$

Now, consider the sign of $\hat{A}^{\pi_{\text{ref}}}(s,a)$:

- If $\hat{A}^{\pi_{\text{ref}}}(s,a) > 0$, then

  $$r(\theta)\, \hat{A}^{\pi_{\text{ref}}}(s,a) > (1+\epsilon)\, \hat{A}^{\pi_{\text{ref}}}(s,a).$$

  Thus, the minimum in the objective is the clipped term, which is constant with respect to θ (i.e., its gradient is zero).

- If $\hat{A}^{\pi_{\text{ref}}}(s,a) < 0$, then

  $$r(\theta)\, \hat{A}^{\pi_{\text{ref}}}(s,a) < (1+\epsilon)\, \hat{A}^{\pi_{\text{ref}}}(s,a),$$

  so the unclipped term is active. Its gradient with respect to θ is proportional to

  $$\nabla_\theta \Big[r(\theta)\, \hat{A}^{\pi_{\text{ref}}}(s,a) \Big] = \hat{A}^{\pi_{\text{ref}}}(s,a) \nabla_\theta r(\theta).$$

  Since $\hat{A}^{\pi_{\text{ref}}}(s,a) < 0$, the gradient will be *negative* (assuming $\nabla_\theta r(\theta) > 0$ for an increase in $r(\theta)$), which implies that the update will *decrease* $r(\theta)$ — that is, it pushes $r(\theta)$ *toward* the boundary $1+\epsilon$ rather than away from it.

Case 2: $r(\theta) < 1 - \epsilon$. By a similar argument, if $\hat{A}^{\pi_{\text{ref}}}(s,a) > 0$, then the gradient of the unclipped term (which is active in this case) is positive and pushes $r(\theta)$ upward toward $1 - \epsilon$.

Formalization: Assume that $r(\theta) \notin [1-\epsilon, 1+\epsilon]$. Then:

$$\begin{cases} r(\theta) > 1+\epsilon & \text{and} \quad \hat{A}^{\pi_{\text{ref}}}(s,a) < 0, \\ r(\theta) < 1-\epsilon & \text{and} \quad \hat{A}^{\pi_{\text{ref}}}(s,a) > 0. \end{cases}$$

In these cases, the derivative of the unclipped objective is

$$\nabla_\theta \Big[r(\theta)\, \hat{A}^{\pi_{\text{ref}}}(s,a) \Big] = \hat{A}^{\pi_{\text{ref}}}(s,a)\, \nabla_\theta r(\theta).$$

Thus, when $\hat{A}^{\pi_{\text{ref}}}(s,a)$ has the sign that makes this derivative point toward the interval, the gradient-based update will reduce the deviation of $r(\theta)$ from $[1-\epsilon, 1+\epsilon]$. In other words, even if the current ratio is outside the interval, the gradient direction (if it does not point away from the interval) will drive it closer to the interval, ensuring that the policy update is conservative.

1.3 Pi prime in PPO

In the *Gather experience* phase, the policy that generates the data is π with parameters θ' (denoted as the behavior policy), and the probability of taking action a_t^e in state s_t^e is stored as

$$p_t^e = \pi_{\theta'}(a_t^e | s_t^e).$$

Later, during the PPO update the current policy π_θ (with updated parameters θ) is used. Thus, the ratio

$$\frac{\pi_\theta(a_t^e|s_t^e)}{p_t^e} = \frac{\pi_\theta(a_t^e|s_t^e)}{\pi_{\theta'}(a_t^e|s_t^e)}$$

generally differs from 1 because $\theta \neq \theta'$ in general. In other words, since the policy is updated over time, the probability of taking the same action in the same state under the current policy is typically not equal to the probability under the behavior policy that generated the data. This ratio is used as an importance sampling correction to account for the discrepancy between the two policies.

2 Offline Evaluation of Bandit Algorithms

2.1 Part 1

Evaluating algorithms for online learning with limited feedback in real-life scenarios is challenging. While the most direct approach is to implement an algorithm and measure its performance live, this is often not feasible due to:

- **Potential risks, costs, and time delays.** Deploying a potentially suboptimal algorithm can lead to high financial or reputational costs. Moreover, once the algorithm is running, it takes time to collect sufficient data for analysis, which delays insights and can be inefficient if the algorithm underperforms.

- **Difficulty of controlled experimentation.** Once data is collected based on a particular algorithm's actions, it is nearly impossible to "replay" the same conditions to test different algorithms under identical circumstances. This lack of controlled repetition makes fair comparisons difficult.