Online and Reinforcement Learning (2025)
Home Assignment 6

Davide Marchi 777881

Contents

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

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

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

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

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.

2.2 Part 2

(a) Modification of UCB1 for Importance-Weighted Losses (Uniform Sampling).

To handle partial feedback when arms are chosen uniformly at random (probability $1/K$ for each arm), we replace the usual empirical loss estimates in UCB1 with importance-weighted estimates. Concretely, whenever an arm i is chosen, its observed loss is scaled by the factor $\frac{1}{p_i(t)} = K$, ensuring an unbiased estimator. The main changes from standard UCB1 are thus:

- **Empirical Loss Update:** Instead of adding the raw observed loss $\ell_{i,t}$, we add $K \cdot \ell_{i,t}$ to the running total for arm i.
- **Confidence Bounds:** The increased variance (due to multiplying by K) is accounted for in the confidence term, typically by a constant factor in front of the usual $\sqrt{\frac{\ln t}{N_i(t)}}$ bound.

Pseudo-Code of the Modified Algorithm:

Initialize: For each arm $i \in \{1, \ldots, K\}$,

$$\widehat{L}_i(0) = 0, \quad N_i(0) = 0.$$

For $t = 1$ **to** T:

1. Select arm
$$A_t \;=\; \arg\min_{i \in \{1,\ldots,K\}} \Big(\widehat{L}_i(t-1) \;+\; c\sqrt{\tfrac{\ln(t-1)}{N_i(t-1)}} \Big),$$
where c is a positive constant (improved parameter choice).
2. Observe the loss $\ell_{A_t,t}$ for the chosen arm A_t.

3. Update counts:

$$N_{A_t}(t) \;=\; N_{A_t}(t-1)+1, \quad N_j(t) = N_j(t-1) \;\text{ for } j \neq A_t.$$

4. Update the empirical loss estimate of arm A_t via importance weighting:

$$\widehat{L}_{A_t}(t) \;=\; \frac{N_{A_t}(t-1)\,\widehat{L}_{A_t}(t-1) \;+\; K\,\ell_{A_t,t}}{N_{A_t}(t)},$$

and keep $\widehat{L}_j(t) = \widehat{L}_j(t-1)$ for $j \neq A_t$.

Key Changes and Regret Bound:

- The use of $K \cdot \ell_{A_t,t}$ ensures an unbiased estimate of the true mean loss for arm A_t under uniform sampling ($p_i(t) = 1/K$).
- Accounting for the variance increase (due to multiplication by K) requires adjusting the confidence term by an appropriate constant factor c.
- The pseudo-regret analysis follows similarly to the standard UCB1 derivation, with an additional factor from the importance weighting. The resulting pseudo-regret remains on the order of

$$O\big(\sqrt{K\,T\ln T}\big),$$

up to constants that depend on c and problem-specific parameters.

(b) Why the Modified UCB1 Cannot Exploit the Small Variance

In the modified UCB1, each observed loss is multiplied by K (because of the importance-weighting factor $1/p_i(t)$ with $p_i(t) = 1/K$). Even if the original loss distribution has small variance, the algorithm's effective variance is dominated by the extra factor of K. UCB1 constructs confidence intervals by considering worst-case deviations of these estimates. Hence, the potential benefit of a small underlying variance cannot be directly exploited; the importance-weighted updates inflate the variance term in the confidence bounds, making the algorithm behave more conservatively (as if the variance were larger).

(c) Modifying EXP3 for Importance-Weighted Losses (Uniform Sampling)

We consider a logging policy that selects arms uniformly at random, with probability $1/K$ for each arm. We modify the EXP3 algorithm so that, whenever the logging policy selects an arm A_t, we form an unbiased estimate of its loss by scaling the observed loss by K. The pseudo-code follows closely the standard EXP3 structure but includes an importance-weighted update:

Pseudo-code for Modified EXP3:

1. **Initialization:**

$$\text{For each arm } i \in \{1, \dots, K\}, \quad w_i(1) = 1.$$

2. **For each round** $t = 1, 2, \dots, T$:

   (a) **(Logging policy)** An arm A_t is chosen *uniformly at random* with probability $\frac{1}{K}$.

   (b) **(Observe loss)** Observe the loss $\ell_{A_t,t}$ of the chosen arm.

   (c) **(Form importance-weighted loss)** For each arm i, define

$$\tilde{\ell}_{i,t} \;=\; \begin{cases} K\,\ell_{A_t,t} & \text{if } i = A_t, \\ 0 & \text{otherwise.} \end{cases}$$

   (d) **(Update weights)** For each arm i,

$$w_i(t+1) \;=\; w_i(t)\,\exp\bigl(-\eta\,\tilde{\ell}_{i,t}\bigr),$$

   where $\eta > 0$ is a learning rate to be chosen.

Expected Regret Bound (Sketch):

Let L_i be the true cumulative loss of arm i over T rounds, and let $L^* = \min_i L_i$ be the cumulative loss of the best arm in hindsight. Using standard EXP3 analysis (with partial feedback), one can show that the expected regret,

$$\mathbb{E}\bigl[\sum_{t=1}^{T} \ell_{A_t,t}\bigr] - L^*,$$

is bounded by

$$O\Bigl(\sqrt{K\,T\ln K}\Bigr),$$

up to constant factors and the choice of η. Since the logging policy is uniform, the factor K appears in the importance-weighted updates, but it is constant and known, which simplifies the analysis. With a more refined argument that accounts for small-variance assumptions (e.g., exploiting tighter concentration bounds for low-variance losses), it is possible to improve these guarantees further. In essence, if the true variance of the losses is small, one can tighten the deviation bounds on $\tilde{\ell}_{i,t}$, leading to smaller confidence terms and an improved regret rate.

(d) Anytime Modification of EXP3

For the anytime version of EXP3 (one that does not assume knowledge of a fixed horizon T), we replace the constant learning rate with one that depends on the current round t. A common choice is:

$$\eta_t \;=\; \sqrt{\frac{2\,\ln K}{K\,t}}.$$

With this time-varying learning rate, the expected regret bound remains of the same order as the fixed-horizon version, namely

$$O\Big(\sqrt{K\,T\,\ln K}\Big),$$

but with a slightly larger constant factor than if T were known in advance.