Online and Reinforcement Learning

2024-2025

Home Assignment 6

Christian Igel **Yevgeny Seldin**
Department of Computer Science
University of Copenhagen

The deadline for this assignment is **19 March 2025, 21:00**. You must submit your *individual* solution electronically via the Absalon home page.

A solution consists of:

- A PDF file with detailed answers to the questions, which may include graphs and tables if needed. Do *not* include your full source code in the PDF file, only selected lines if you are asked to do so.
- A .zip file with all your solution source code with comments about the major steps involved in each question (see below). Source code must be submitted in the original file format, not as PDF. The programming language of the course is Python.

Important Remarks:

- IMPORTANT: Do NOT zip the PDF file, since zipped files cannot be opened in *SpeedGrader*. Zipped PDF submissions will not be graded.
- Your PDF report should be self-sufficient. I.e., it should be possible to grade it without opening the .zip file. We do not guarantee opening the .zip file when grading.
- Your code should be structured such that there is one main file (or one main file per question) that we can run to reproduce all the results presented in your report. This main file can, if you like, call other files with functions, classes, etc.
- Handwritten solutions will not be accepted.

1 PPO `(33 points) [Christian]`

Let us consider the Proximal Policy Optimization (PPO) algorithm in the variant presented in the lecture and described on the slides "Deep Reinforcement Learning".

1.1 Return expressed as advantage over another policy (7 points)

The expected return of a policy π can be expressed in terms of its advantage over another policy π_{ref} and $J(\pi_{\text{ref}})$:

$$J(\pi) = J(\pi_{\text{ref}}) + \mathbb{E}\left\{ \sum_{t=0}^{\infty} \gamma^t \underbrace{A^{\pi}_{\text{ref}}(s_t, a_t)}_{\text{advantage of following } \pi \text{ instead of } \pi_{\text{ref}}} \middle| s_0, \pi \right\}$$

Provide a proof (with intermediate steps) using the notation from the lecture.

1.2 Clipping (13 points)

PPO uses the clipping

$$\min\left(\frac{\pi(a_t,s_t)}{\pi_{\text{ref}}(a_t,s_t)}\hat{A}^{\pi}_{\text{ref}}(s_t,a_t), \text{clip}\left(\frac{\pi(a_t,s_t)}{\pi_{\text{ref}}(a_t,s_t)}, 1-\epsilon, 1+\epsilon\right)\hat{A}^{\pi}_{\text{ref}}(s_t,a_t)\right)$$

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

In the lecture, it has been stated that the gradient-based update of the PPO-Clip objective mentioned above updates the policy "if $\frac{\pi(a_t,s_t)}{\pi_{\text{ref}}(a_t,s_t)} \in [1-\epsilon, 1+\epsilon]$ or if the update leads to getting closer to this range". The first condition is trivial. Assume that the first condition does not hold. What is is meant by "if the gradient direction does not point away from the interval"? Formalize and prove this condition (assuming the first condition is not met, i.e., $\frac{\pi(a_t,s_t)}{\pi_{\text{ref}}(a_t,s_t)} \notin [1-\epsilon, 1+\epsilon]$).

1.3 Pi prime in PPO (13 points)

The policy generating the experience, see procedure "Gather experience" on the slides, uses the policy π with parameters $\boldsymbol{\theta'}$, where the probability of an action a^e_t taken in a state s^e_t (in episode e at step t) is stored in p^e_t. The PPO update, as described on the slide "PPO-Clip optimization", considers the expression $\frac{\pi(a^e_t,s^e_t)}{p^e_t}$. Why is this expression not always one?

2 Offline Evaluation of Bandit Algorithms (67 points) [Yevgeny]

Solve Exercise 5.14 Part 1 and "The theoretical part" of Part 2 in Yevgeny's lecture notes. ("The practical part" of Part 2 will be given in a later assignment.)

References