Artificial Neural Networks (Gerstner). Exercises for week 11

Deep Reinforcement Learning 2

Exercise 1. Uncorrelated mini-batches in A2C.

In the lecture you have seen a simple example of a single weight w that changes with temporally correlated updates Δw (red dots and red curve on slide 6). Reshuffling the updates in time led to a more stable learning dynamics (blue dots and blue curve). This example illustrates the effect of sampling iid from the replay buffer for off-policy methods like DQN. For on-policy methods, the proposed solution is to run multiple actors in parallel.

a. Sketch a figure similar to the one on slide 6 for 4 parallel actors and 2 to 3 episodes per actor. Mark the starts of new episodes for each actor with vertical lines. Hint: the episodes can have different lengths.

b. Draw in the same figure approximately the values $\frac{1}{4}\sum_{k=1}^{4}\Delta w^{(k)}$ for all time points.

c. Write a caption to the figure that explains, why the proposed solution of A2C helps to stabilize learning.

Exercise 2. Proximal Policy Optimization.

a. In the derivation of Proximal Policy Optimization methods the ratio $r_{\theta'}(s_t, a_t) = \frac{\pi_{\theta'}(a_t;s_t)}{\pi_\theta(a_t;s_t)}$ appeared on the last line on slide 15. Convince yourself that the equality on the last line of slide 15 is correct by explicitly writing out the expectations in the same way as we did on slide 14.
Hint: write something like $E_{s_t,a_t\sim p_{\theta'},\pi_{\theta'}}\left[\sum_{t=0}^{\infty}\gamma^t A_\theta(s_t,a_t)\right] = \sum \ldots = \sum \ldots = E_{s_t,a_t\sim p_{\theta'},\pi_\theta}\cdots$

b. Show that the summands in the loss function of PPO-CLIP can also be written in the form

$$\ell(r_{\theta'}) = \min(r_{\theta'}\gamma^t A_\theta, g(\epsilon, \gamma^t A_\theta)), \quad \text{with} \quad g(\epsilon, A) = \begin{cases} (1+\epsilon)A & A \geq 0 \\ (1-\epsilon)A & A < 0 \end{cases}$$

c. Sketch $\ell(r)$ as a function of r in two figures: one where A_θ is positive and one where A_θ is negative.

d. Write a caption to your figures that explains, why one can safely run a few steps of gradient ascent on $\hat{L}^{\text{CLIP}}(\theta')$ without risking that $\pi_{\theta'}$ would move too far away from π_θ.

Exercise 3. Deep Deterministic Policy Gradient.

a. How many input and output neurons does the Q-network of DQN have, if the input consists of 100-dimensional vectors and there are 10 possible actions?

b. How many input and output neurons does the Q-network of DDPG have, if the input consists of 100-dimensional vectors and the action space is 10 dimensional?

c. Explain, why it would not be a good idea to use $\hat{Q}(s_{j+1}, a_{j+1})$ in line 7 of the DDPG algorithm (slide 21).

d. Explain, why it would not be a good idea to use $\hat{Q}(s_{j+1}, \hat{\pi}(s_{j+1}) + \epsilon)$ in line 7 of the DDPG algorithm.

Exercise 4. Background Planning.

In this exercise we look again at the simple map of Europe on slide 26. You will run value iteration with goal Vienna. This means that we will keep $V(\mathsf{V}) = 0$ all the time.

a. Initialize all V- and Q-values to zero.

b. Apply the update rule of value iteration (equation 1 on slide 26) to all cities in parallel. Hint 1: keep $V(\mathsf{V}) = 0$ and don't worry, if you find e.g. $V(\mathsf{Z}) = -2$. Hint 2: "In parallel" means, that you should assume $V(s') = 0$ when running this step for the first time and take the value that you obtained in the previous iteration otherwise.

c. Repeat step b. until convergence.

d. Convince yourself that value iteration found the optimal solution. Write down, how you convinced yourself that the optimal solution was found.

Exercise 5. AlphaZero.

In this exercise you will manually compute some steps in one iteration of AlphaZero's Monte Carlo Tree Search. Assume that from the previous Monte Carlo Tree Search you have already a tree that starts at state s_0 with $N(s_0, a_1) = 8, W(s_0, a_1) = 4, P(s_0, a_1) = 0.2$, $N(s_0, a_2) = 24, W(s_0, a_2) = 16, P(s_0, a_1) = 0.8$. We fix $C(s) = 1$ and $\tau = 1$.

a. Determine the action (a_1 or a_2) that AlphaZero would take in the selection step of MCTS in state s_0.

b. Is it a greedy or an exploratory action that was taken in a?

c. Update N, P, W and Q under the assumption that the expansion step led to $v = 0.7$.

d. Compute the probability that AlphaZero would take the actual action a_1 now.