DS-GA 3001.005

NYU Center for Data Science

Reinforcement Learning

Homework 04

Exercice 1 (30 points)

Reinforcement Learning with Function Approximation:

- **1.1 (15 points)**: List some pros and cons of model-based vs. value-based vs. policy-based reinforcement learning. Provide at least 2 pros and 2 cons for each method.
- **1.2 (5 points)**: Why does parameterizing a function can help scale and generalize through larger state and/or action spaces?
- **1.3 (10 points)**: Discuss the trade-offs between using simulated experiences and real experiences in reinforcement learning. Can a combination of both be beneficial? Support your answer with a scenario where either approach alone is preferable.

Exercice 2 (30 points)

Actor-Critic Reinforcement Learning:

- **2.1(10 points)**: Provide at least 3 examples of objective measure that have been successfully used (i.e., published or discussed in the course) for policy optimization using gradient ascent
- **2.2 (5 points)**: In Actor-Critic RL, why does approximating values to define the *Critic* can help increase the *Actor*'s sampling efficiency?

- **2.3 (5 points)**: What is the impact of the regularization term added to the objective measure in the TRPO algorithm?

- **2.4 (5 points)** : Compare PPO and TRPO. Why PPO is generally preferred in practical implementations. Provide an example of a reinforcement learning scenario where PPO's advantages are particularly beneficial.

- **2.5 (5 points)** : Write down the policy gradient theorem. Explain how sampling contributes to estimating the policy gradient and why it is necessary for effective learning in reinforcement learning.

Exercice 3 (40 points)

Planning with Advanced Model-based Reinforcement Learning:

- **3.1 (5 points)**: What are the input and output of a model of the environment in a single state Bandit problem? How about in the more general case of sequential RL problem?

- **3.2 (10 points)**: What's the difference between table-lookup model, expectation model and stochastic model? Provide at least one pros and one cons for each type of model.

- **3.3 (5 points)**: Why is Experience Replay called a *non-parametric* model?

- **3.4 (5 points)**: In Monte-Carlo Tree Search, why does the tree policy become more optimal than the initial rollout policy during the search?

- **3.5 (15 points)**: Describe, in your own words, what are the differences between the AlphaGo and AlphaZero algorithms. This question is voluntarily open, and will assess your overall understanding of this technology, both in term of methodology and potential for applications

  [**Warning:**] Responses from ChatGPT & similar lack precision relative to what we discussed in the lecture. To get all the points, you need be significantly more precise than ChatGPT :)