Exercise: Exploration

This exercise showcases the impact of different exploration strategies. In this assignment you will implement a policy using no exploration, a policy using ϵ-greedy and one using ϵz-greedy [Dabney et al., 2020: .`https://arxiv.org/pdf/2006.01782.pdf`] The ϵz-greedy policy samples not only a random action but also a duration for which the action will be played. You can find the algorithm in Appendix B. We will use grid environments. Find the assignment here: `https://classroom.github.com/a/PlqGM3vD`.

1. **Implement ϵ(z)-greedy**
   Your task is to implement the (non)-ϵ(z) policy in `Policy.__call__`. Use the member variable `disable_exploration` to enable complete greedy behavior. Hint: You can switch from ϵz-greedy to ϵ-greedy by setting `duration_max`.

2. **Implement Sampling of the Duration**
   Implement the sampling of the duration in `Policy.sample_duration`. Hint: Check the paper for the hyperparameter μ.

3. **Configure Policies**
   Add the hyperparameters to `policy_classes` to create a greedy, ϵ-greedy and ϵz-greedy policy.

4. **Run and Observe**
   Run `exploration.py` and note the differences in the results in `answers.txt`. Upload the figures to `plots`. Is the current algorithm well suited for the problem? What could be a way to improve it (think of the previous lectures)? You can also play with the hyperparameters (e.g., γ and ϵ) and try different environments (e.g., bigger grid).