Exercise 7: Policy Gradient

GitHub classroom: `https://classroom.github.com/a/i0PTF2Hw`
The purpose of this exercise is to get you accustomed to implementing policy gradient methods. For this, you will be implementing the REINFORCE algorithm to solve the `CartPole-v1` environment. Your tasks are the following:

1. **Policy Gradient Implementation**

   - Complete the `Policy` class in the code with 2 Linear units to map the states to probabilities over actions.
   - Implement `compute_returns` method to compute the discounted returns G_t for each state in a trajectory.
   - Implement the `policy_improvement` step to update the policy given the rewards and probabilities from the last trajectory.
   - Use the policy in the `act` method to sample action and return its log probability.

2. **Questions**

   - How does the length of the trajectories affect the training?
   - How could a baseline be implemented to stabilize the training?
   - Does the same network architecture and learning rate work for `LunarLander-v2`?
   - How is the sample complexity (how many steps it takes to solve the environment) of this algorithm related to the DQN from the last exercise?

Please write your answers in `answers.txt`