

## Exercise 10: RL - Solution

Name: .....

UTID:.....

Today we'll explore a simple example to highlight the advantages of different reinforcement learning algorithms. Consider learning to drive a car (continuous states, continuous actions). You already spent days cooking up a good reward function, you penalize traffic accidents by their severity, you encourage staying on the road, following a navigation system, and even signaling to other drivers what you're going to do. However, you have not yet found a good policy to drive. You try three algorithms: imitation learning, policy gradient, and Q-learning. For each algorithm highlight what its strengths and weaknesses are

### a) Imitation learning

- does not generalize well to scenes not seen in the training environment
- automatically explores the training environment
- is exponential in the possible trajectories
- is exponential in the possible states
- needs to crash the car a few times to learn
- works well with existing data augmentation
- requires human (ground truth) trajectories

### b) Deep Q-Learning

- does not generalize well to scenes not seen in the training environment
- automatically explores the training environment
- is exponential in the possible trajectories
- is exponential in the possible states
- needs to crash the car a few times to learn
- works well with existing data augmentation
- requires human (ground truth) trajectories

### c) Policy gradient

- does not generalize well to scenes not seen in the training environment
- automatically explores the training environment
- is exponential in the possible trajectories
- is exponential in the possible states
- needs to crash the car a few times to learn
- works well with existing data augmentation
- requires human (ground truth) trajectories

**d)** Given the strength and weaknesses, which algorithms (if any) would you use to train your self-driving car?

*If restricted to those four, a good option would be to initialize the network using imitation and fine-tune using policy gradient. Note that both algorithms use the same network (a policy), just different loss functions. Deep Q-Learning on the other hand uses a very different network (predict action-value pairs).*

**e)** Both Deep Q-Learning and Policy gradient are exponential, however in different quantities. Briefly explain how the two algorithms are exponential, and which one you'd prefer for our driving example.

*Policy gradient might work better, as it is exponential in trajectories, while Q-learning is exponential in a (continuous) state-action pair. Even-though PG is exponential in trajectories the gradient can still be approximated with a small set of trajectories.*

In class today, we'll look at gradient free optimization, a third algorithm to optimize the RL objective. It is also exponential, but not in the states or trajectories, instead in the parameters of the learned policy.