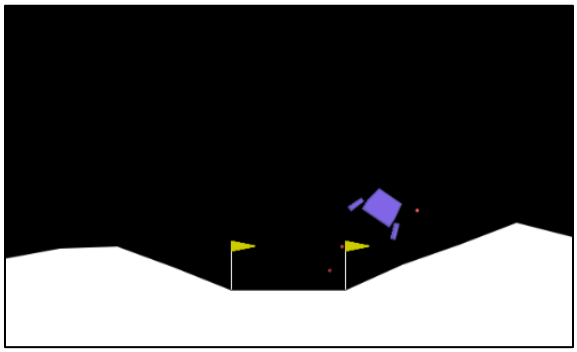
Practical Work: IQL

Aim: How does the performance of IQL depend on the dataset characteristics?

Reinforcement Learning (RL) had tremendous success, but it can take billions of environment interactions to actually learn a policy that performs well. In critical scenarios, such as autonomous driving or healthcare, this is not acceptable. Therefore, Offline RL tries to learn policies from previously recorded interactions with the environment. For the practical work, you will use the IQL offline RL algorithm on the Lunar-Lander (continuous) environment under different dataset conditions and report your findings.



Utilize an off-the-shelf online RL algorithm (e.g. PPO) to obtain 5 different datasets for the LunarLander (continuous) environment. These can range from datasets collected by a random policy to a dataset collected by a near-optimal policy, or some intermediate policy during training the online RL agent. First, analyze the different datasets. Then use those different datasets for training IQL and report the performance. As a baseline method, also implement Behavioural Cloning (BC), which is just supervised learning where you learn the action given the state on the dataset. To evaluate during training the (offline) RL agent, you need to interact with the environment and do e.g. 5-10 episodes of interactions to assess how well the policy currently performs. Also note that hyperparameter-tuning is notoriously hard in Offline RL, a good idea would be a “validation” dataset, which is only used for tuning and not for the analysis afterwards.

Note that you will not be graded on how well any of the IQL agents perform in the end, but by how well you designed and executed your experiments. You should try to come up with a sound experimental design, ask the right questions and try to conduct clean and reproducible (use seeds!) experiments to get answers. Note that one run of IQL per dataset is not sufficient, you need to run the algorithm multiple (e.g. 3-5) times to get statistically meaningful answers.

The focus is a **clean codebase and reproducibility**, you will benefit your whole (study) career from learning that. Also consider using git (and Github) if you don’t already do so. Also, it is required that you log your runs, use either tensorboard or weights&biases for that. You can use packages for the RL algorithm to create the datasets, but IQL should be re-implemented by you as part of the practical work. The code needs to be handed in to get a grade in the end.

Environment: https://gymnasium.farama.org/environments/box2d/lunar_lander/
StableBaselines for online RL: https://github.com/DLR-RM/stable-baselines3
RL dataset analysis paper: https://arxiv.org/abs/2111.04714
Statistical Analysis of RL experiments: https://arxiv.org/abs/2108.13264