Multi-Task Offline Reinforcement Learning

Experimental Evaluation of the Generalizability of the Soft Actor-Critic + Behavioral Cloning Algorithm

TU Delft

1 Introduction / Background

Offline RL:

- No environment interaction [1].
- Learns optimal policy from a dataset [1].
- Important when environment interaction is too costly or expensive.

Multi-Task RL:

- Agent learns from multiple tasks.
- Single-task methods often fail in multi-task settings [2].

2 Research Questions

- Can SAC combined with BC effectively generalize to new tasks within a multi-task RL environment?
- What characteristics of the offline dataset are critical for the success or failure of SAC+BC in such settings?

3 Environment

- Discrete Action Space.
- 40 Tasks per Configuration.
- Task Characteristics:
  1. Agent Location ▼
  2. Goal Location ■
  3. Topology

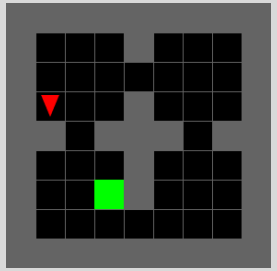


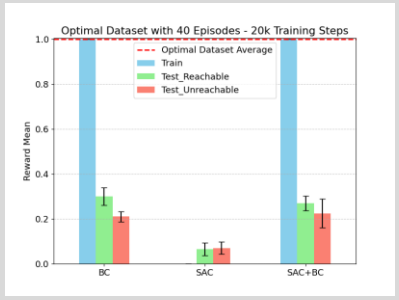
Figure: 4-Room Grid [3]

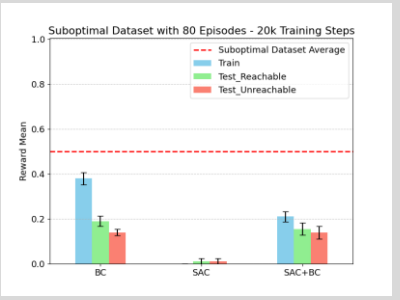
4 Experimental Setup

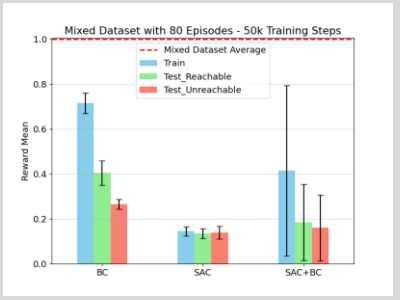
1. Create Datasets
   - Quality: Optimal, Suboptimal, Mixed.
   - Size: 40, 80, 200, 400 Episodes/Tasks.
2. Implement Algorithms
   - Behavioral Cloning (BC).
   - Soft Actor-Critic (SAC).
   - SAC+BC → Add BC term to SAC.
3. Hyperparameter Tuning with Optuna using a Tree-Structured Parzen Estimator (TPE) to sample more of the promising parameters.
4. Training & Evaluation of BC, SAC, and SAC+BC with up to 50k training steps.

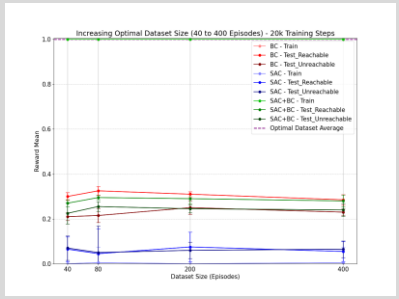
5 Results

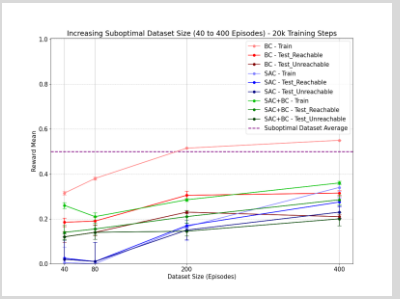
- SAC+BC generalizes comparable to BC on optimal and suboptimal data.
- SAC+BC generalizes best when trained on higher quality (optimal) data.
- SAC+BC performs with high volatility when trained on mixed data.
- SAC mostly achieves mean rewards below 0.25.
- Increasing suboptimal and mixed data size enhances performance of SAC+BC.
- Increasing optimal data size had no impact on SAC+BC.

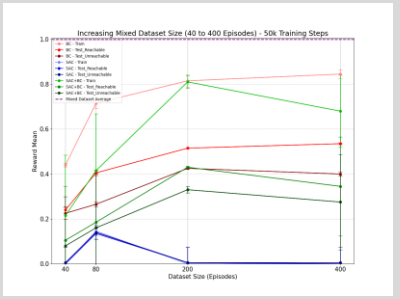












6 Conclusion

- Marginal generalization gap between BC and SAC+BC on optimal and suboptimal data but gap is not clear on mixed data.
- Study suggest increasing data size improves generalization of SAC+BC if and only if this increases diversity.
- SAC underperforms in an offline setting.

7 Limitations

- Only one environment tested, so results may not hold for other environments.
- Hardware limitations have restricted hyperparameter tuning.
- Only five seeds were used, so the results of SAC+BC are not stable with mixed data.

8 References

[1] S. Levine *et al.*, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," arXiv preprint arXiv:2005.01643, 2020.

[2] I. Mediratta *et al.*, "The Generalization Gap in Offline Reinforcement Learning," arXiv preprint arXiv:2312.05742, 2023.

[3] M. Weltevrede, "Four Room: A simple 4-room grid world environment to test. generalisation behaviour of RL agents," 2022. *Online*. Available: https://github.com/MWeltevrede/four_room. Accessed: Apr. 28, 2023.

Author: Axel Geist | A.O.Geist@student.tudelft.nl Supervisor: Max Weltevrede Professor: Matthijs Spaan Examiner: Elena Congeduti