OMRL - Identifiability Challenges and Effective Data Collection Strategies [Dorfman et al. 2021]

- Extension to VariBAD Algorithm
  - Off-Policy Version
- Introducing Data-Collection/Augmenting Strategy for offline data
- Identifiability Problem: MDP ambiguity
  - Special to OMRL
  - Problem of identifying MDP
  - Tackle it with **policy replaying**: online
    - Use data-policy of another MDP in current to augment data
  - Tackle it with **reward relabeling**: offline
    - Use reward function of different MDP to augment data of current one

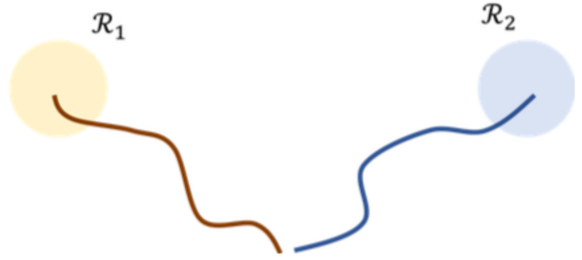
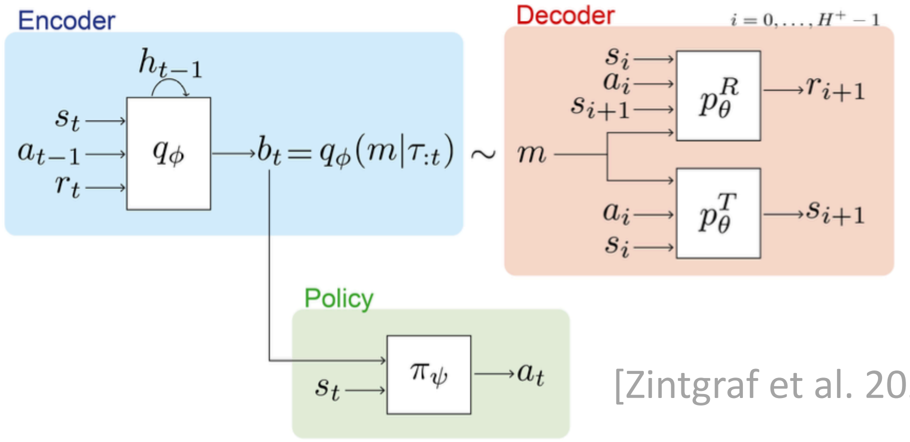


Figure 2: Reward ambiguity: from the two trajectories, it is impossible to know if there are two MDPs with different rewards (blue and yellow circles), or one MDP with rewards at both locations.



[Zintgraf et al. 2020]

OMRL - Identifiability Challenges and Effective Data Collection Strategies [Dorfman et al. 2021]

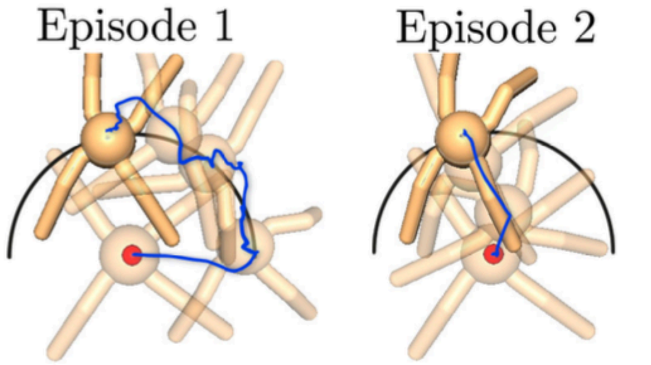


Figure 5: Ant-Semi-circle: trajectories from trained policy on a new goal. In the first episode the ant searches for the goal, and in the second one it directly moves toward the goal it has previously found. This search behavior is different from the goal-reaching behaviors that dominate the data.

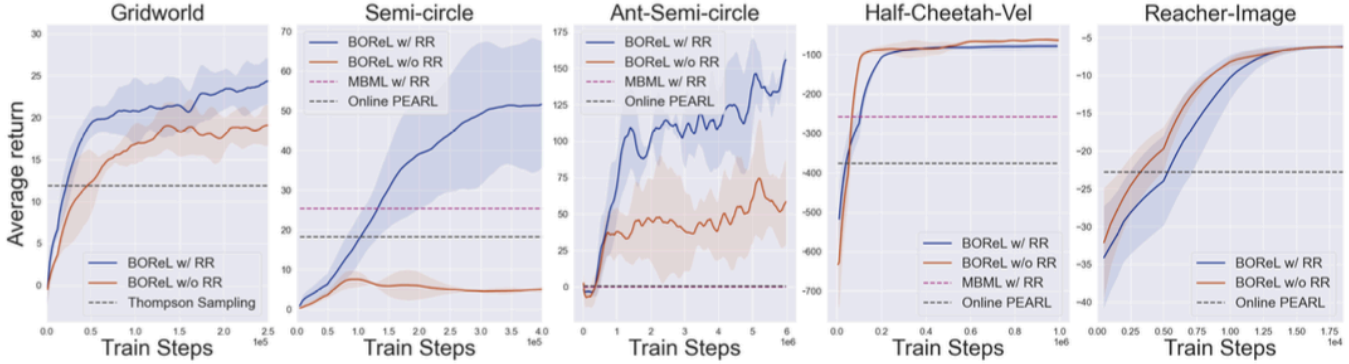


Figure 3: Offline performance on domains with varying rewards. We compare BOReL with and without reward relabeling (blue and red, respectively) with Thompson sampling baselines – calculated exactly in Gridworld, and using online PEARL and offline MBML for the other domains. Full training curves for baselines appear in the supplementary; here we plot only the best performance.