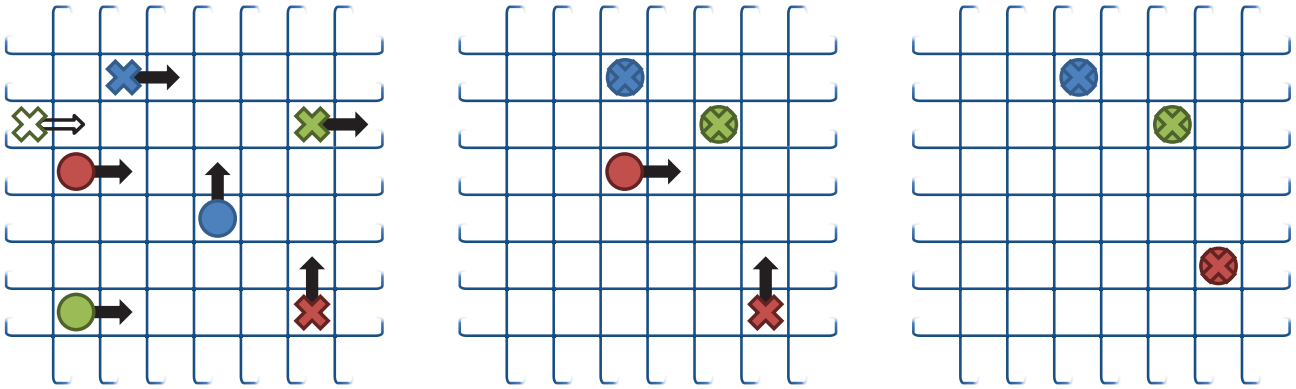
Supplemental: Deep Decentralized Multi-task Multi-agent RL under Partial Observability

The following provides additional empirical results. Some plots are reproduced from the main text to ease comparisons.

A. Multi-agent Multi-target (MAMT) Domain Overview



(a) Agents must learn inherent toroidal transition dynamics in the domain for fast target capture (e.g., see green target).

(b) In MAMT tasks, no reward is given to the team above, despite two agents successfully capturing their targets.

(c) Example successful simultaneous capture scenario, where the team is given $+1$ reward.

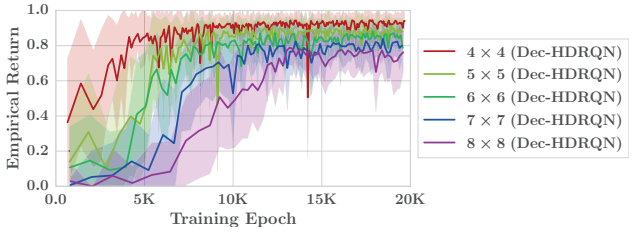
Figure 6. Visualization of MAMT domain. Agents and targets operate on a toroidal $m \times m$ gridworld. Each agent (circle) is assigned a unique target (cross) to capture, but does not observe its assigned target ID. Targets' states are fully occluded at each timestep with probability P_f. Despite the simplicity of gridworld transitions, reward sparsity makes this an especially challenging task. During both learning and execution, the team receives no reward unless all targets are captured *simultaneously* by their corresponding agents.

B. Empirical Results: Learning on Multi-agent Single-Target (MAST) Domain

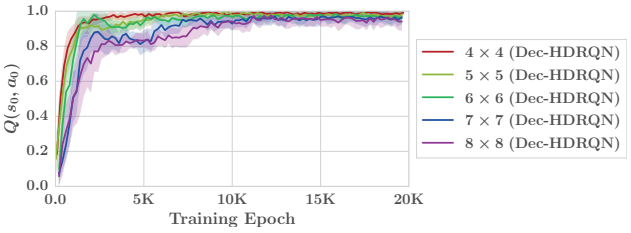
Multi-agent Single-target (MAST) domain results for Dec-DRQN and Dec-HDRQN, with 2 agents and $P_f = 0.0$ (observation flickering disabled). These results mainly illustrate that Dec-DRQN sometimes has *some* empirical success in low-noise domains with small state-space. Note that in the 8×8 task, Dec-HDRQN significantly outperforms Dec-DRQN, which converges to a sub-optimal policy despite domain simplicity.

(a) Dec-DRQN empirical returns during training.

(b) Dec-DRQN anticipated values during training.



(c) Dec-HDRQN empirical returns during training.

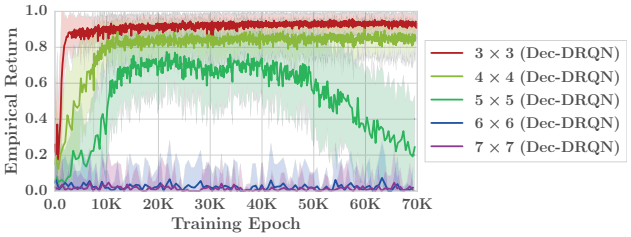


(d) Dec-HDRQN anticipated values during training.

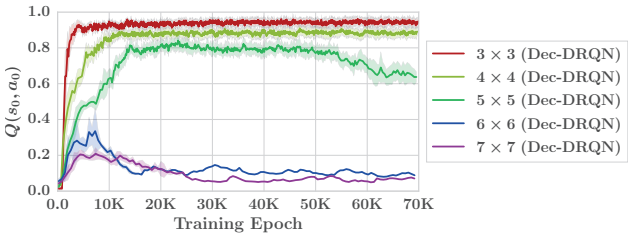
Figure 7. Multi-agent Single-target (MAST) domain results for Dec-DRQN and Dec-HDRQN, with 2 agents and $P_f = 0.0$ (observation flickering disabled). All plots conducted (at each training epoch) for a batch of 50 randomly-initialized episodes. Anticipated value plots (on right) were plotted for the exact starting states and actions undertaken for the episodes used in the plots on the left.

C. Empirical Results: Learning on MAMT Domain

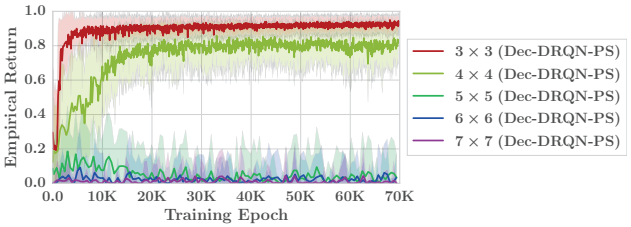
Multi-agent Single-target (MAMT) domain results, with 2 agents and $P_f = 0.3$ (observation flickering disabled). We also evaluated performance of inter-agent parameter sharing (a centralized approach) in Dec-DRQN (which we called Dec-DRQN-PS). Additionally, performance of a Double-DQN was deemed to have negligible impacts (Dec-DDRQN).



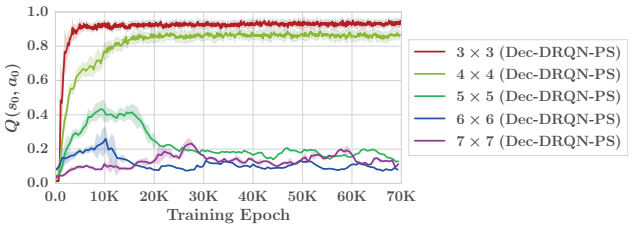
(a) Dec-DRQN empirical returns during training.



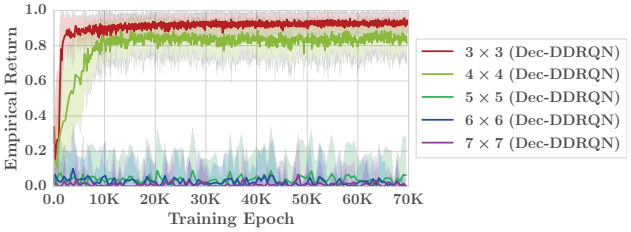
(b) Dec-DRQN anticipated values during training.



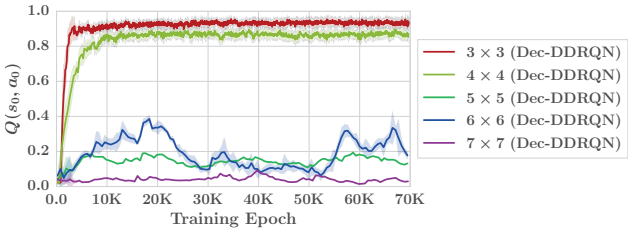
(c) Dec-DRQN-PS (with parameter sharing), empirical returns during training.



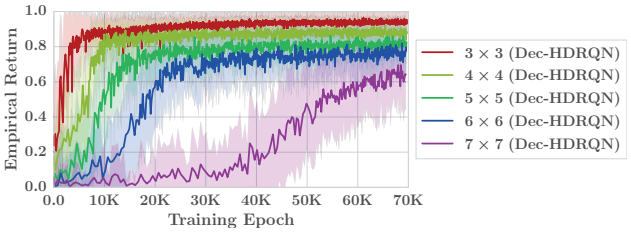
(d) Dec-DRQN-PS (with parameter sharing), anticipated values during training.



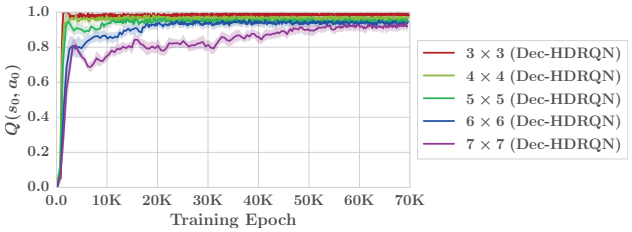
(e) Dec-DDRQN (double DRQN) empirical returns during training.



(f) Dec-DDRQN (double DRQN) anticipated values during training.



(g) Dec-HDRQN (our approach) empirical returns during training.



(h) Dec-HDRQN (our approach) anticipated values during training.

Figure 8. MAMT domain results for Dec-DRQN and Dec-HDRQN, with 2 agents and $P_f = 0.3$. All plots conducted (at each training epoch) for a batch of 50 randomly-initialized episodes. Anticipated value plots (on right) were plotted for the exact starting states and actions undertaken for the episodes used in the plots on the left.

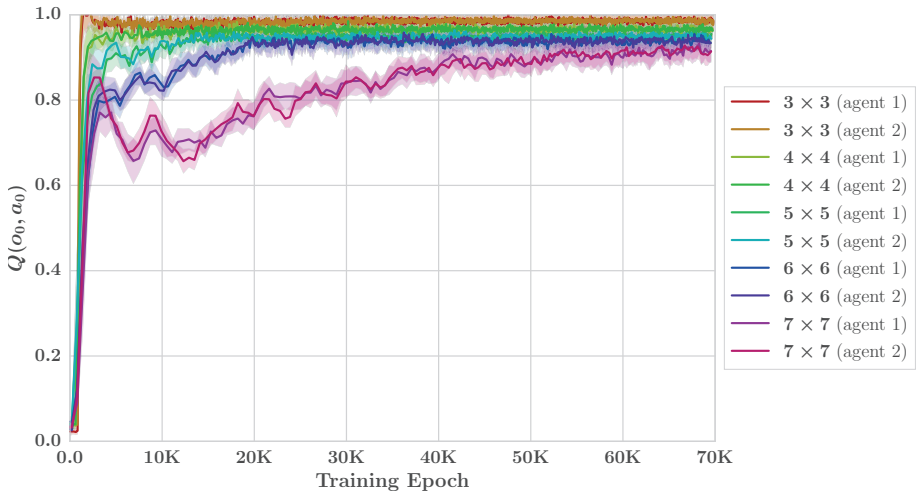
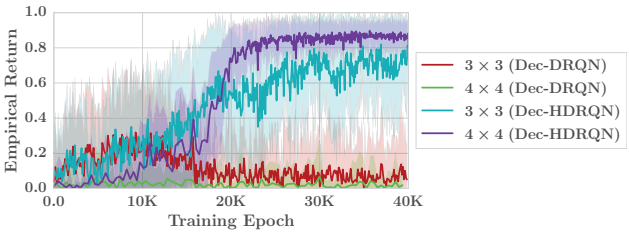
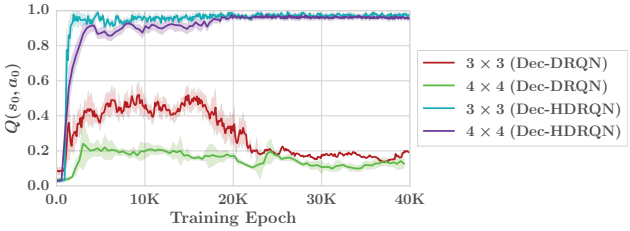


Figure 9. Comparison of agents' anticipated value plots using Dec-HDRQN during training. MAMT domain, with 2 agents and $P_f = 0.3$. All plots conducted (at each training epoch) for a batch of 50 randomly-initialized episodes. For a given task, agents have similar anticipated value convergence trends due to shared reward; differences are primarily caused by random initial states and independently sampled target occlusion events for each agent.



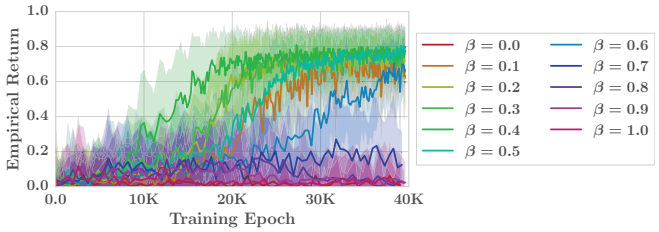
(a) Empirical returns during training. For batch of 50 randomly-initialized games.



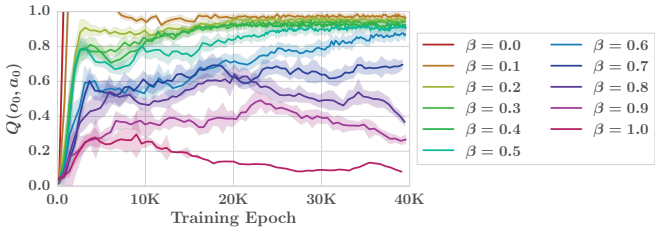
(b) Anticipated values during training. For specific starting states and actions undertaken in the same 50 randomly-initialized games as Fig. 10a.

Figure 10. MAMT domain results for Dec-DRQN and Dec-HDRQN, with $n = 3$ agents. $P_f = 0.6$ for the 3×3 task, and $P_f = 0.1$ for the 4×4 task.

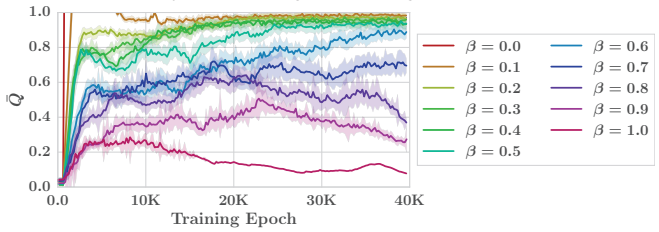
D. Empirical Results: Learning Sensitivity to Dec-HDRQN Negative Learning Rate β



(a) Sensitivity of Dec-HDRQN empirical returns to β during training. For batch of 50 randomly-initialized games.



(b) Sensitivity of Dec-HDRQN predicted action-values to β during training. For specific starting states and actions undertaken in the same 50 randomly-initialized games of Fig. 11a.

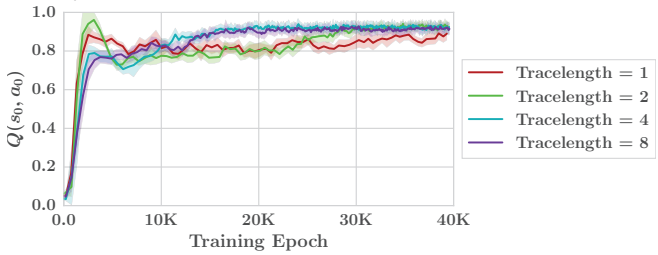


(c) Sensitivity of Dec-HDRQN average Q values to β during training. For random minibatch of 32 experienced observation inputs.

Figure 11. Learning sensitivity to β for 6×6, 2 agent MAMT domain with $P_f = 0.25$. All plots for agent $i = 0$. $\beta = 1$ corresponds to Decentralized Q-learning, $\beta = 0$ corresponds to Distributed Q-learning (not including the distributed policy update step).

E. Empirical Results: Learning Sensitivity to Dec-HDRQN Recurrent Training Tracelength Parameter τ

(a) Dec-HDRQN sensitivity to tracelength τ. 6x6 MAMT domain with $P_f = 0.25$.

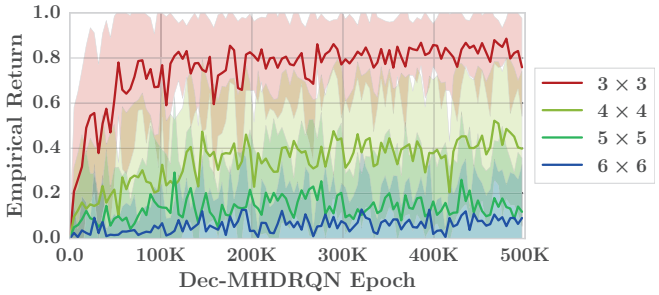


(b) Sensitivity of Dec-HDRQN predicted action-values to recurrent training tracelength parameter. For specific starting states and actions undertaken in the same 50 randomly-initialized games of Fig. 12a.

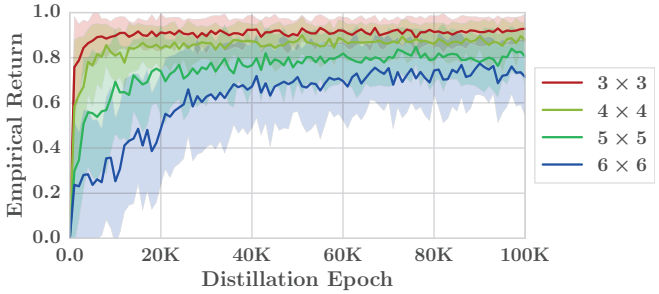
Figure 12. Learning sensitivity to recurrent training tracelength parameter for 6×6, 2 agent MAMT domain with $P_f = 0.25$. All plots for agent $i = 0$. $\beta = 1$ corresponds to Decentralized Q-learning, $\beta = 0$ corresponds to Distributed Q-learning (not including the distributed policy update step).

F. Empirical Results: Multi-tasking Performance Comparison

The below plots show multi-tasking performance of both the distillation and Multi-HDRQN approaches. Both approaches were trained on the 3×3 through 6×6 MAMT tasks. Multi-DRQN failed to achieve specialized-level performance on all tasks, despite 500K training epochs. By contrast, the proposed MT-MARL distillation approach achieves nominal performance after 100K epochs.



(a) MT-MARL via Multi-HDQRN.



(b) MT-MARL via specialized and distilled Dec-HDRQN.

Figure 13. Multi-task performance on MAMT domain, $n = 2$ agents and $P_f = 0.3$.