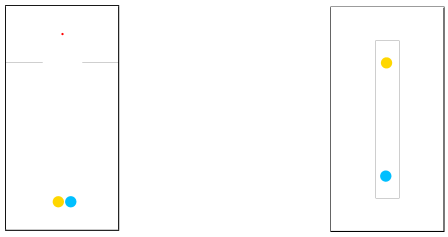

(a) Marching domain. Agents need to march through the exit. **(b) *Narrow* domain. Valid moving area is within the gray box.**

Figure 1: Sequential Decision Making Tasks

Definition 2.1. A Q-function is relative overgeneralization free if for any agent i and any of its two actions a^i and b^i, we always have $\mathbb{E}_{a_t^{-i}\sim\pi^{-i}(a_t^{-i}|o_t^{-i})}[Q(s,a^i,a^{-i})] > \mathbb{E}_{a_t^{-i}\sim\pi^{-i}(a_t^{-i}|o_t^{-i})}[Q(s,b^i,a^{-i})]$ given that $Q(s,a^i,a^{*-i}) > Q(s,b^i,a^{*-i})$, where $\pi^{-i}(a_t^{-i}|o_t^{-i})$ is some arbitrary policy of all the other agents, and a^{*-i} is the best response action of all the other agents.

A Q-function can be decomposed into an advantage function and a state value function, and the state value function does not have the action involved. Thus, we are really looking for an advantage function that is relative overgeneralization free. A natural choice would be a quadratic function with a block diagonal matrix.

Proposition 2.2. *Suppose* $A(s,\vec{a}) = -\frac{1}{2}(\vec{a}-\mu(s))^T M(s)(\vec{a}-\mu(s))$ *is the advantage function for some state* s *and joint actions* $\vec{a}$ *of all* n *agents, where* $M(s)$ *is a positive definite block diagonal matrix with* n *blocks with each block of size* d_i *by* d_i*, namely, the action dimension of agent* i*. Then the corresponding state-action value function* $Q(s,\vec{a})$ *is relative overgeneralization free.* (See [7] for proof).

3 EXPERIMENTS

We applied three problems to test our methods: *Max of Two Quadratics* (described in [6]); and *Marching* and *Narrow*, sequential decision making tasks which pose difficulties in cooperation in simple games.

In the Marching and Narrow games there are two agents, each with a radius of 0.05. Both games terminate after 200 steps. In Marching the agents must march towards a red dot. A shaped reward is provided for the two agents based on the distance between the agent's center point and the red dot. The agents receive a large penalty (-10) if they collide or are too far from one another (the distance between them is ≥0.11), and receive a large reward (10) if they reach the red dot. The purpose of this game is to test whether the agents can coordinate their moving speeds.

Narrow requires the two agents to swap positions. They start at opposite ends in an aisle, and must pass one another to reach their goals. The aisle is narrow and requires coordination between two agents, because when they collide they receive a large penalty (-10). If both agents reach their goal positions, they both receive a large reward (10). We considered several variations of this game, with different aisle width and shaped reward functions.

Table 1 shows the results of the proposed algorithms in different domains compared against other algorithms. In the repeated game Max of Two Quadratics, MADDPG generally converged to a sub-optimal Nash Equilibrium, while MAIRL and MGAIL made use of the demonstration and converged to the optimal Nash Equilibrium. In Marching, MGAIL performed the best. Although MADDPG and

	MAIRL	*MGAIL*	*MADDPG*	*TRPO*
Max of Two Quadratics	**9.78 ±0.07**	**8.98 ±0.35**	-0.03 ±0.02	
Marching	**-152.88 ±129.76**	**-80.88 ±52.08**	**-178.36 ±89.46**	-324.58 ±111.88
Narrow (u, 0.2)	**5.99 ±0.84**	**5.49 ±1.79**	-2.00 ±0.00	-2.00 ±0.00
Narrow (u, 0.205)	**4.53 ±2.51**	**5.54 ±2.15**	-2.00 ±0.00	-2.08 ±0.16
Narrow (sh, 0.2)	**-4.40 ±0.85**	**-4.94 ±1.80**	-23.46 ±6.10	-79.21 ±6.96
Narrow (sh, 0.205)	**-6.76 ±3.78**	**-5.23 ±2.13**	**-39.98 ±37.06**	-79.34 ±4.54
Narrow (s, 0.2)	**5.95 ±0.83**	**5.44 ±1.78**	-2.80 ±1.32	-2.63 ±0.14
Narrow (s, 0.205)	**4.50 ±2.51**	**5.50 ±2.15**	-3.68 ±1.04	-2.66 ±0.17

u=*unshaped reward* sh=*shaped reward* s=*scaled shaped reward*
Number in the parentheses marks the width of the field.

Table 1: Convergence results of algorithms by domain, showing the mean ± standard error over 5 trials. The best-performing algorithms on each task are shown in boldface.

TRPO learned to march to the exit and to avoid relative overgeneralization, they had to make many collisions during movement.

In all cases of Narrow, MGAIL and MAIRL agents successfully reached the target. However, in the unshaped reward and scaled shaped reward setting, both MADDPG and TRPO agents failed to learn to pass one another. To understand the difference, we noticed that our imitation learning agents used an advantage function as reward, and $A_{soft}(s,a) = \mathbb{E}_{s'}[R(s,a) + V_{soft}(s') - V_{soft}(s)]$, which can be thought of as an original reward function shaped by a soft value function. Thus, our imitation learners were receiving a better reward signal for exploration.

4 CONCLUSION AND FUTURE WORK

In this paper we proposed two methods to achieve better coordination in cooperative continuous games based on imitation learning. We showed that the proposed combined with coordination samples can avoid relative overgeneralization in cooperative games. A drawback of our approach is that the input space of our discriminator can grow linearly with the number of agents. We will investigate how to solve this issue in future work.

REFERENCES

[1] Justin Fu, Katie Luo, and Sergey Levine. 2018. Learning Robust Rewards with Adverserial Inverse Reinforcement Learning. In *International Conference on Learning Representations*.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[3] Jonathan Ho and Stefano Ermon. 2016. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*. 4565–4573.

[4] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*. 6382–6393.

[5] Ermo Wei and Sean Luke. 2016. Lenient Learning in Independent-Learner Stochastic Cooperative Games. *Journal of Machine Learning Research* 17, 84 (2016), 1–42.

[6] Ermo Wei, Drew Wicke, David Freelan, and Sean Luke. 2018. Multiagent Soft Q-Learning. In *AAAI Fall Symposium on Data Efficient Reinforcement Learning*.

[7] Ermo Wei, Drew Wicke, and Sean Luke. 2019. *Multiagent Adversarial Inverse Reinforcement Learning*. Technical Report GMU-CS-TR-2019-2. Department of Computer Science, George Mason University, 4400 University Drive MSN 4A5, Fairfax, VA 22030-4444 USA.

[8] Rudolph Paul Wiegand. 2004. *An Analysis of Cooperative Coevolutionary Algorithms*. Ph.D. Dissertation. Department of Computer Science, George Mason University.