A. Deep Multi-Agent Determinantal Q-Learning

Although our proposed Q-DPP serves as a new type of function approximator for the value function in multi-agent reinforcement learning, deep neural networks can also be seamlessly applied on Q-DPP. Specifically, one can adopt deep networks to respectively represent the quality and diversity terms in the kernels of Q-DPP, and we name such approach Deep Q-DPP. One can think of Deep Q-DPP as modeling $\boldsymbol{\mathcal{D}}$ and $\boldsymbol{\mathcal{B}}$ by neural networks rather than look-up tables. An analogy of Deep Q-DPP to Q-DPP would be Deep Q-learning (Mnih et al., 2015) to Q-learning (Watkins & Dayan, 1992). As the main motivation of introducing Q-DPP is to eliminate structural constraints and bespoke neural architecture designs in solving multi-agent cooperative tasks, we omit the study of Deep Q-DPP in the main body of this paper. Here we rather demonstrate a proof of concept for Deep Q-DPP and its effectiveness on StarCraft II micro-management tasks (Samvelyan et al., 2019b) for the review purpose.

A.1. Neural Architectures for Deep Q-DPP.

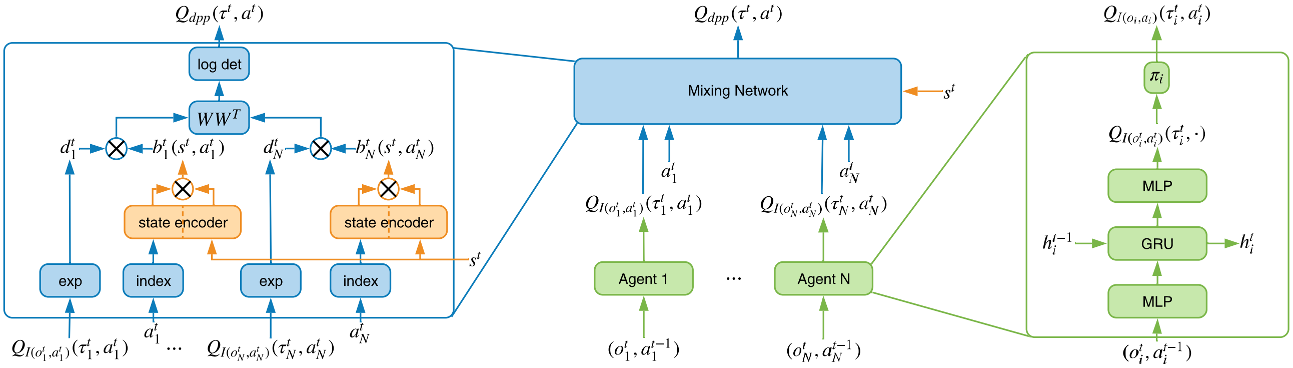

Figure 6: Neural Architecture of Deep Q-DPP.

Fig. 6 illustrates the complete setup for Deep Q-DPP. The middle part of the diagram shows the overall architecture of Q-DPP, which consists of each agent's individual Q-networks and a centralized mixing network. Details of the mixing network are presented in the left part of Fig. 6. We compute the quality term, d_i, by applying the exponential operator on the individual Q-value, and compute the diversity feature term, $\boldsymbol{b}_i$, by index the corresponding vector in $\boldsymbol{\mathcal{B}}$ through the global state s and each individual action a_i.

A critical advantage of Deep Q-DPP is that it can deal with continuous states/observations. When the input state s is continuous, we first index the raw diversity feature $\boldsymbol{b}_i'$ based on the embedding of discrete action a_i. To integrate the information of the continuous state, we use two multi-layer feed-forward neural networks f_d and f_n, which encodes the direction and norm of the diversity feature separately. f_d outputs a feature vector with same shape as $\boldsymbol{b}_i'$ indicating the **direction**, and f_n outputs a real value for computing the **norm**. In practice, we find modeling the direction and norm of the diversity features separately by two neural networks helps stabilize training, and the diversity feature vector is computed as $\boldsymbol{b_i} = f_d(\boldsymbol{b}_i', s) \times \sigma(f_n(\boldsymbol{b}_i', s))$. Finally, the centralized Q-value can then be computed from d_i and $\boldsymbol{b_i}$ following Eq. 6.

A.2. Experiments on StarCraft II Micro-Management

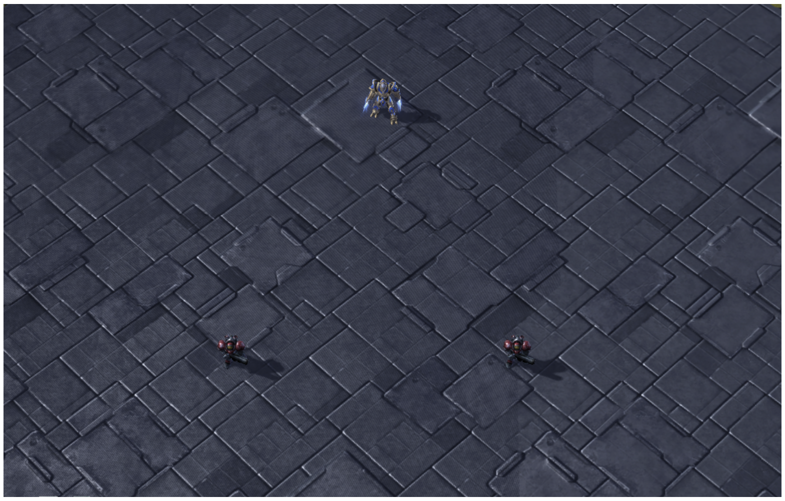

Figure 7: StarCraft II micro-management scenarios: 2 Marines vs. 1 Zealot.

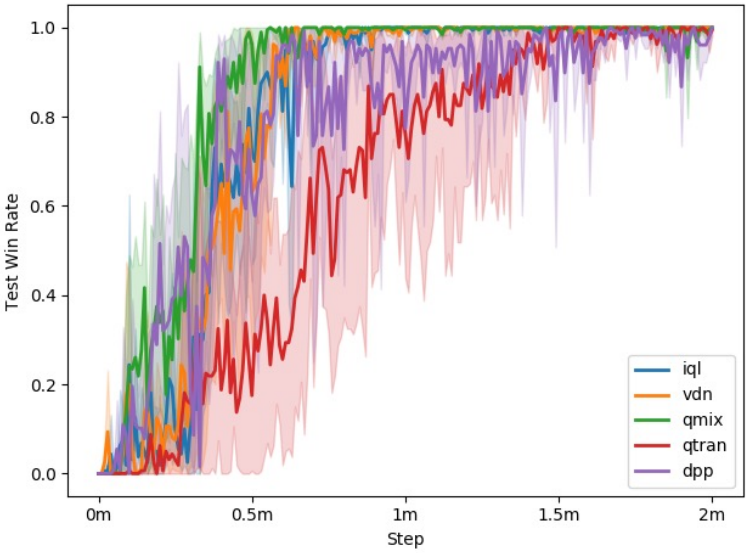

Figure 8: The performance of various algorithms on 2m_vs_1z maps.

We study a challenging micro-management game in StarCraft II in SMAC (Samvelyan et al., 2019b), i.e., **2m_vs_1z**, the screenshots of scenarios are given in Fig. 7. In the 2m_vs_1z map, we control a team of 2 Marines to fight with 1 enemy Zergling. In this task, it requires the Marine units to take advantage of their larger firing range to defeat over Zerglings which can only attack local enemies. In both cases, the agents can observe a **continuous** feature vector including the information of health, positions and weapon cooldown of other agents. In terms of reward design, we keep the default setting in (Samvelyan et al., 2019b). All agents receive a large final reward for winning a battle, at the meantime, they also receive immediate rewards that are proportional to the difference of total damages between the two teams in every timestep. We compare Q-DPP with aforementioned baseline models, i.e., IQL, VDN, QMIX, and QTRAN, and plot the results in Fig. 8. The results show that Q-DPP can perform as good as the state-of-the-art model, QMIX, even when the state feature is continuous. However, the performance is not stable and presents high variance. We will study the reason in our future works.