MAVEN: Multi-Agent Variational Exploration

Anuj Mahajan, Tabish Rashid, Mikayel Samvelyan, Shimon Whiteson

1.Introduction

We analyse value-based methods that are known to have superior performance in complex environments for MARL in centralised training with decentralised execution scenario. We show that the representational constraints on the joint action-values introduced by QMIX and similar methods lead to provably poor exploration and suboptimality. We show that committed exploration can be used to solve the above problem. Towards this we propose a novel approach called MAVEN that hybridises value and policy-based methods by introducing a latent space for hierarchical control. MAVEN achieves committed, temporally extended exploration, which is key to solving complex multi-agent tasks.

2.Background

- Dec-POMDP defined as a tuple $G = \langle S, U, P, r, Z, O, n, \gamma \rangle$
- S is the set of states
- U the set of available actions per agent
- agents $i \in \mathcal{A} \equiv \{1, ..., n\}$
- joint action $\mathbf{u} \in \mathbf{U} \equiv U^n$
- $P(s'|s, \mathbf{u}) : S \times \mathbf{U} \times S \to [0, 1]$ is the state transition function
- $r(s, \mathbf{u}) : S \times \mathbf{U} \to \mathbb{R}$ is the reward function
- observations $z \in Z$ according to observation function $O(s, i) : S \times \mathcal{A} \to Z$.
- γ is discount factor
- action-observation history for an agent i is $\tau^i \in T \equiv (Z \times U)^*$

3.Decentralisability

$$\arg\max_{\mathbf{u}} Q^*(s, \mathbf{u}) = \left(\arg\max_{u^1} q_1(\tau^1, u^1) \dots \arg\max_{u^n} q_n(\tau^n, u^n)\right)',$$

- QMIX uses monotonic transformations on q_i, $\frac{\partial Q_{qmix}(s,\mathbf{u})}{\partial q_i(s,u^i)} \geq 0$
- VDN uses sum of utilities $Q_{vdn}(s, \mathbf{u}) = \sum_i q_i(s, u^i)$
- QTRAN: poses the decentralisation problem as optimisation with $\mathcal{O}(|S||U|^n)$ constraints and relaxes for tractability.
- IQL approximates by treating as an independent single agent problem.

Definition (Non-monotonicity)

For any state $s \in S$ and agent $i \in \mathcal{A}$ given the actions of the other agents $u^{-i} \in U^{n-1}$, the Q-values $Q(s, (u^i, u^{-i}))$ form an ordering over the action space of agent i. Define $C(i, u^{-i}) := \{(u^i_1, ..., u^i_{|U|}) | Q(s, (u^i_j, u^{-i})) \geq Q(s, (u^i_{j+1}, u^{-i})), j \in \{1, \dots, |U|\}, u^i_j \in U, j \neq j' \implies u^i_j \neq u^i_{j'}\}$, as the set of all possible such orderings over the action-values. The joint-action value function is **non-monotonic** if $\exists i \in \mathcal{A}, u^{-i}_1 \neq u^{-i}_2$ s.t. $C(i, u^{-i}_1) \cap C(i, u^{-i}_2) = \varnothing$.

Theorem (Uniform visitation QMIX)

For n player, $k \geq 3$ action matrix games ($|\mathcal{A}| = n, |U| = k$), under uniform visitation; Q_{qmix} learns a δ-suboptimal policy for any time horizon T, for any $0 < \delta \leq R\left[\sqrt{\frac{a(b+1)}{a+b}} - 1\right]$ for the payoff matrix M (n dimensional) given by the template below, where $b = \sum_{s=1}^{k-2} \binom{n+s-1}{s}$, $a = k^n - (b+1)$, $R > 0$:

$$\begin{bmatrix} R+\delta & 0 & \dots & R \\ 0 & & \ddots & \\ \vdots & \ddots & & \vdots \\ R & \dots & & R \end{bmatrix}$$

Theorem (ϵ-greedy visitation QMIX)

For n player, $k \geq 3$ action matrix games, under ϵ-greedy visitation $\epsilon(t)$; Q_{qmix} learns a δ-suboptimal policy for any time horizon T with probability $\geq 1 - \left(\exp(-\frac{T\upsilon^2}{2}) + (k^n - 1)\exp(-\frac{T\upsilon^2}{2(k^n-1)^2})\right)$, for any $0 < \delta \leq R\left[\sqrt{a\left(\frac{\upsilon b}{2(1-\upsilon/2)(a+b)} + 1\right)} - 1\right]$ for the payoff matrix given by the template above, where $b = \sum_{s=1}^{k-2} \binom{n+s-1}{s}$, $a = k^n - (b+1)$, $R > 0$ and $\upsilon = \epsilon(T)$.

4.MAVEN

- Fixing z gives a joint action-value function $Q(\mathbf{u}, s; z, \phi, \eta, \psi)$ which implicitly defines a greedy deterministic policy $\pi_{\mathcal{A}}(\mathbf{u}|s; z, \phi, \eta, \psi)$. This gives the corresponding Q-learning loss:

$$\mathcal{L}_{QL}(\phi, \eta, \psi) = \mathbb{E}_{\pi_{\mathcal{A}}}[(Q(\mathbf{u}_t, s_t; z) - [r(\mathbf{u}_t, s_t) + \gamma \max_{\mathbf{u}_{t+1}} Q(\mathbf{u}_{t+1}, s_{t+1}; z)])^2],$$

- The hierarchical policy objective for z, freezing the parameters ψ, η, ϕ is given by:

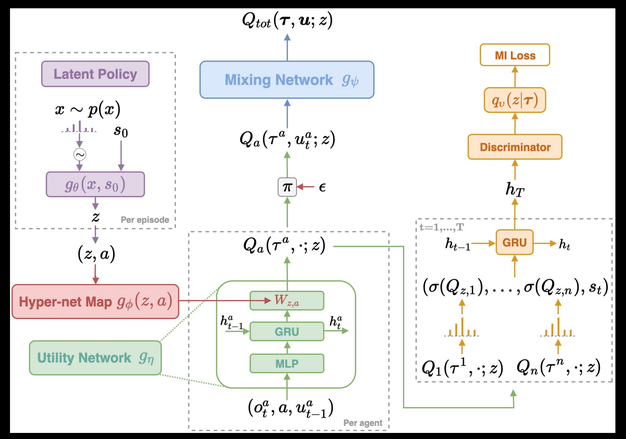
$$\mathcal{J}_{RL}(\theta) = \int \mathcal{R}(\tau_{\mathcal{A}}|z) p_\theta(z|s_0)\rho(s_0)dzds_0.$$

- Tractable lower bound on mutual information $\mathcal{J}_{MI}$ given by:

$$\mathcal{J}_{MI} \geq \mathcal{H}(z) + \mathbb{E}_{\sigma(\tau),z}[\log(q_\upsilon(z|\sigma(\tau)))].$$

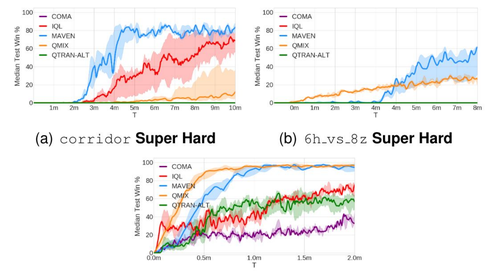
- Overall objective becomes:

$$\max_{\upsilon,\phi,\eta,\psi,\theta} \mathcal{J}_{RL}(\theta) + \lambda_{MI}\mathcal{J}_V(\upsilon, \phi, \eta, \psi) - \lambda_{QL}\mathcal{L}_{QL}(\phi, \eta, \psi),$$



5.Experiments

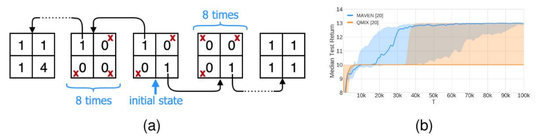
StarCraft-2 SMAC



(a) `corridor` **Super Hard** (b) `6h_vs_8z` **Super Hard**

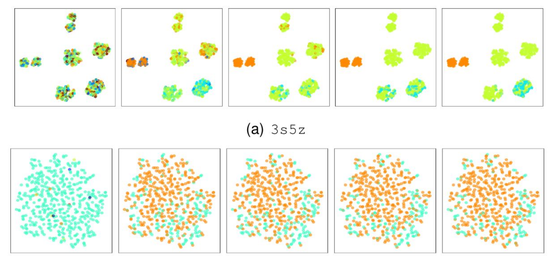
(c) `2s3z` **Easy**

m-step matrix games



(a) (b)

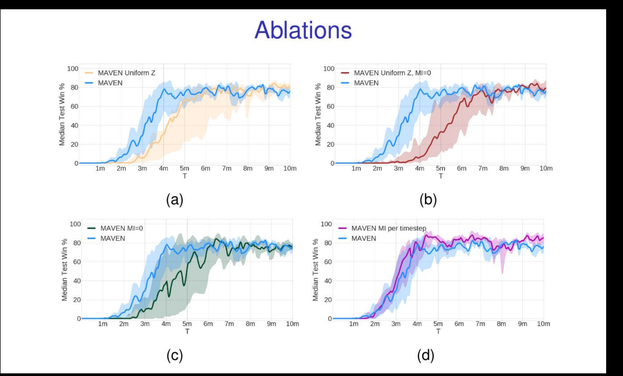
Representation capacity



(a) `3s5z`

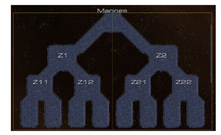
(b) `micro_corridor`

Ablations

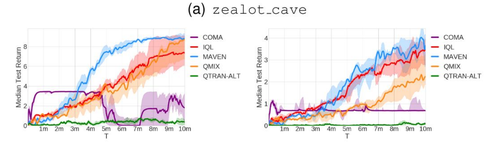


(a) (b) (c) (d)

StarCraft-2 Exploration experiments



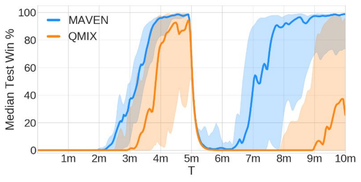
(a) `zealot_cave`



(b) `zealot_cave depth 3` (c) `zealot_cave depth 4`

StarCraft-2 Robustness experiments

(a) `2_corridors`



(b) `Shorter corridor closed at 5mil steps`

Contact

`anuj.mahajan@cs.ox.ac.uk`