MAVEN: Multi-Agent Variational Exploration

Anuj Mahajan

WhiRL, University of Oxford

Joint work with Tabish, Mika and Shimon

MARL

- Cooperative *multi-agent reinforcement learning* (MARL) is a key tool for addressing many real-world problems
- Robot swarm, autonomous cars
- Key challenges: CTDE
  - Scalability due to exponential state action space blowup
  - Decentralised execution

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 $\tau^i \in T \equiv (Z \times U)^*$

MARL problem continued

$$Q^{\pi}(s_t, \mathbf{u}_t) = \mathbb{E}_{s_{t+1:\infty}, \mathbf{u}_{t+1:\infty}} \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} | s_t, \mathbf{u}_t \right]. \quad (1)$$

The goal of the problem is to find the optimal action value function Q^* and the corresponding policy π^*.

Decentralisability

- Asserts that $\exists q_i$, such that $\forall s, \mathbf{u}$:

$$\arg\max_{\mathbf{u}} Q^*(s, \mathbf{u}) = \left(\arg\max_{u^1} q_1(s, u^1) \ldots \arg\max_{u^n} q_n(s, u^n)\right)', \tag{2}$$

Where q_i are agent utilities.

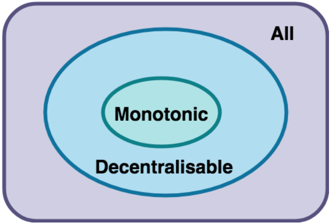


Figure 1: Classification of MARL problems.

Existing methods

- Several algorithms have been proposed which ensure decentralisability though structural constraints
- 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.

Problems with existing methods

- Existing methods do not facilitate *committed exploration*
- Imposing structural constraints on the hypothesis learnt can induce suboptimality (all existing methods suffer from this)
- Structural constraints interfere with exploration
- Use latent space to address the above problems! (MAVEN)

Analysis

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. $C(i,u^{-i}_1) \cap C(i,u^{-i}_2) = \varnothing$.

Example Non-Monotonic payoff

Table 1: (a) An example of a non-monotonic payoff matrix, (b) QMIX values under uniform visitation.

(a)

	A	*B*	*C*
A	10.4	0	10
B	0	10	10
C	10	10	10

(b)

	A	*B*	*C*
A	6.08	6.08	8.95
B	6.00	5.99	8.87
C	8.99	8.99	11.87

QMIX analysis : Uniform visitation

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 $b = \sum_{s=1}^{k-2} \binom{n+s-1}{s}$, $a = k^n - (b+1)$, $R > 0$:

$$\begin{bmatrix} R+\delta & 0 & \ldots & R \\ 0 & & \iddots & \\ \vdots & \iddots & & \vdots \\ R & \ldots & & R \end{bmatrix}$$

QMIX analysis: ϵ greedy

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 $\upsilon = \epsilon(T)$.

MAVEN: Multi-Agent Variational Exploration

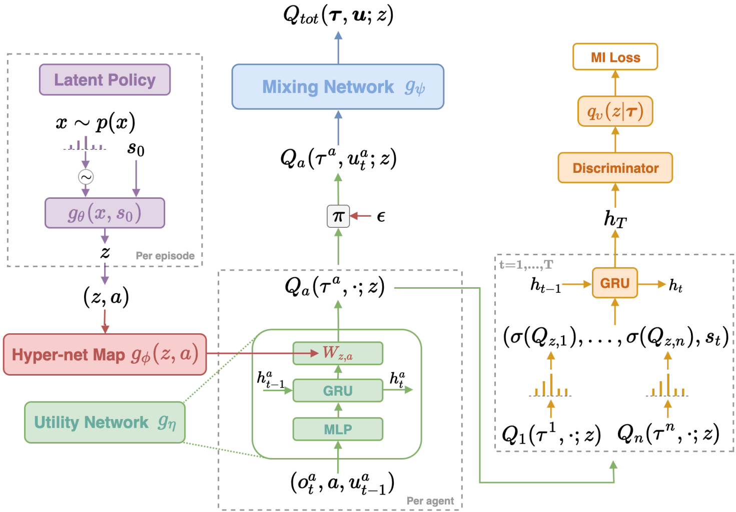


Figure 2: Architecture for MAVEN.

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:

$$\mathcal{L}_{QL}(\phi, \eta, \psi) = \mathbb{E}_{\pi_{\mathcal{A}}}[(Q(\mathbf{u}_t, s_t; z) - [r(\mathbf{u}_t, s_t) \tag{3}$$

$$+\gamma \max_{\mathbf{u}_{t+1}} Q(\mathbf{u}_{t+1}, s_{t+1}; z)])^2], \tag{4}$$

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

$$\mathcal{J}_{RL}(\theta) = \int \mathcal{R}(\tau_{\mathcal{A}}|z) p_\theta(z|s_0) \rho(s_0) dz ds_0. \tag{5}$$

Encouraging diverse behaviour with MI

- Mutual Information loss

$$\mathcal{J}_{MI} = \mathcal{H}(\sigma(\boldsymbol{\tau})) - \mathcal{H}(\sigma(\boldsymbol{\tau})|z) = \mathcal{H}(z) - \mathcal{H}(z|\sigma(\boldsymbol{\tau})), \quad (6)$$

- Tractable lower bound given by:

$$\mathcal{J}_{MI} \geq \mathcal{H}(z) + \mathbb{E}_{\sigma(\boldsymbol{\tau}),z}[\log(q_{\upsilon}(z|\sigma(\boldsymbol{\tau})))]. \quad (7)$$

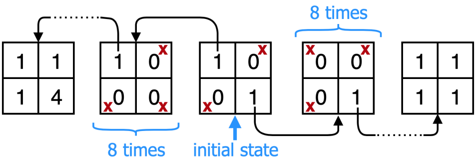
- The variational approximation can also be seen as a discriminator/critic that induces an auxiliary reward field $r^{z}_{aux}(\boldsymbol{\tau}) = \log(q_{\upsilon}(z|\sigma(\boldsymbol{\tau}))) - \log(p(z))$ on the trajectory space.
- Overall objective becomes:

$$\max_{\upsilon,\phi,\eta,\psi,\theta} \mathcal{J}_{RL}(\theta) + \lambda_{MI}\mathcal{J}_{V}(\upsilon,\phi,\eta,\psi) - \lambda_{QL}\mathcal{L}_{QL}(\phi,\eta,\psi), \quad (8)$$

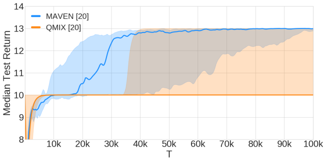
Experiments

- Toy domain: Matrix games
- StarCraft-2

m-step matrix games



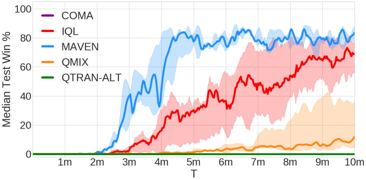
(a)



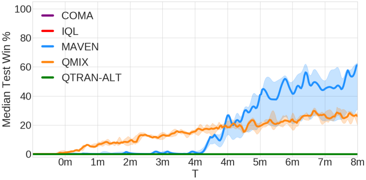
(b)

Figure 3: (a) m-step matrix game for $m = 10$ case (b) median return of MAVEN and QMIX method on 10-step matrix game for 100k training steps, averaged over 20 random initializations (2nd and 3rd quartile is shaded).

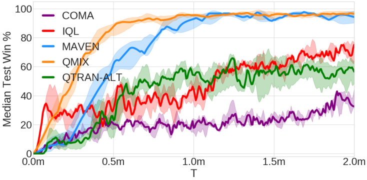
StarCraft-2 SMAC



(a) `corridor` **Super Hard**



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

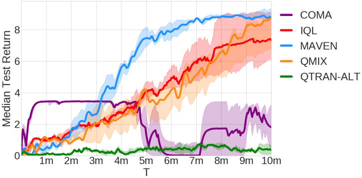


(c) `2s3z` **Easy**

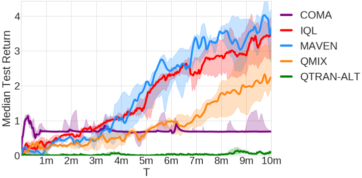
Figure 4: The performance of various algorithms on three SMAC maps.

StarCraft-2 Exploration experiments

(a) `zealot_cave`



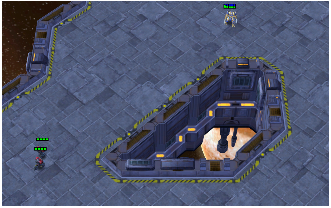
(b) `zealot_cave depth 3`



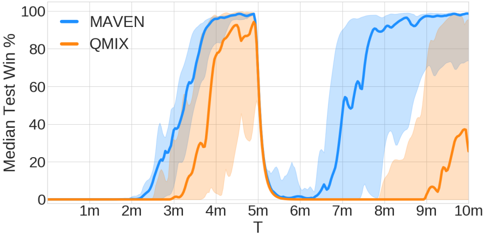
(c) `zealot_cave depth 4`

Figure 5: State exploration and policy robustness

StarCraft-2 Robustness experiments



(a) 2_corridors



(b) Shorter corridor closed at 5mil steps

Figure 6: State exploration and policy robustness

Representation capacity

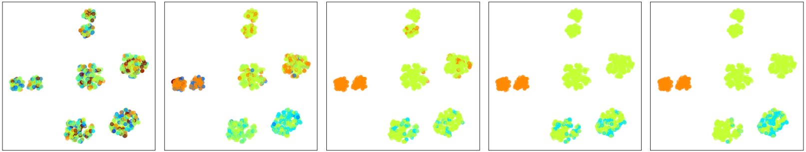


Figure 7: tsne plot for s_0 labelled with z, 16 categories, `3s5z` initial (left) to final (right)

Ablations

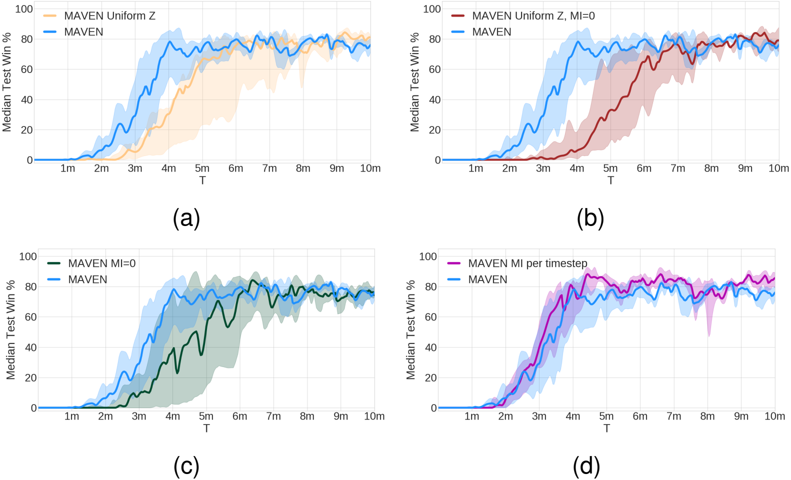


Figure 8: Figs. 8(a) and 8(b) investigate uniform hierarchical policy. Figs. 8(c) and 8(d) investigate effects of MI loss.

Thanks!

Questions?