西安交通大學
XI'AN JIAOTONG UNIVERSITY

Greedy-based Value Representation for Optimal Coordination in Multi-agent Reinforcement Learning

Lipeng Wan / Zeyang Liu / Xingyu Chen / Xuguang Lan* / Nannin Zheng
School of AI, Xi'an Jiaotong University
2022/6/21

Contents

Introduction

Fully-cooperative MARL

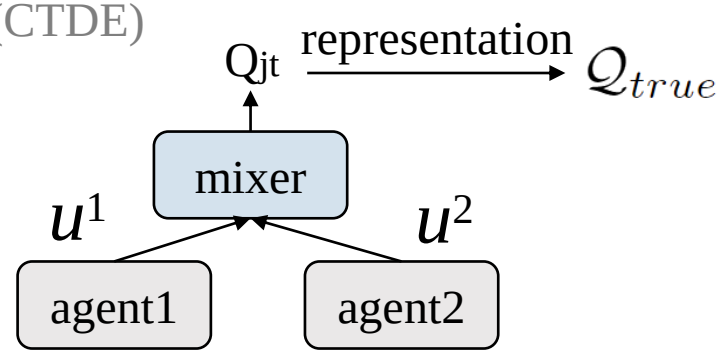
agent 1

agent 2	u_0	u_1	u_2
u_0	r_1	r_2	r_3
u_1	r_4	r_5	r_6
u_2	r_7	r_8	r_9

shared rewards

Value decomposition

Credit assignment for Centralized Training with Decentralized Execution (CTDE)



Independent Global Max (IGM)

identity of joint greedy action and the set of individual greedy actions

Qjt				u^2
	1	0	0	0.6
	0	0	0	0
	0	0	2	0.8
u^1	0.4	0	1.2	

true Q value function

Q_{true}			
	Q_{11}	**Q_{12}**	**Q_{13}**
	Q_{21}	**Q_{22}**	**Q_{23}**
	Q_{31}	**Q_{32}**	**Q_{33}**

Qjt				u^2
	1	0	0	2
	0	0	0	0
	0	0	2	1
u^1	1	0	2	

VDN:
Linear Value Decomposition (LVD)

$$Q_{jt} = \mathcal{U}^1 + \mathcal{U}^2$$

QMIX:
Monotonic Value Decomposition (MVD)

$$Q_{jt} = \omega_1 \cdot \mathcal{U}^1 + \omega_2 \cdot \mathcal{U}^2$$

* u^i: utility function
Qjt: joint Q value function

Background

Incomplete Representation Capability (IRC) of the joint Q value function

u^2 \ u^1	u_0	u_1	u_2
u_0	$u_0^1+u_0^2$	$u_0^1+u_1^2$	$u_0^1+u_2^2$
u_1	$u_1^1+u_0^2$	$u_1^1+u_1^2$	$u_1^1+u_2^2$
u_2	$u_2^1+u_0^2$	$u_2^1+u_1^2$	$u_2^1+u_2^2$

Qjt

representation →

Q_{11}	Q_{12}	Q_{13}
Q_{21}	Q_{22}	Q_{23}
Q_{31}	Q_{32}	Q_{33}

$\mathcal{Q}_{true}$

representation problem: solving an overdetermined equation system

$\{\mathcal{U}_i^1 + \mathcal{U}_j^2 = \mathcal{Q}_{ij}\}_{i,j\in[1,m]}$

number of variables: m*n (6)
number of equations: m^n (9)
equation num > variable num

no exact solution

↓

Qjt is unable to converge to $\mathcal{Q}_{true}$

↓

Qjt suffers from IRC

* m: number of options of individual action space
n: number of agents

Relative Overgeneralization (RO)

multiple potential representations

8	**-12**	**-12**
-12	**0**	**0**
-12	**0**	**6**

representation 1

7.40	**-8.33**	**-7.93**
-8.33	**-24.1**	**-23.7**
-7.93	**-23.7**	**-23.3**

representation 2

-24.4	**-14.5**	**-9.3**
-14.5	**-4.65**	**0.6**
-9.3	**0.6**	**5.8**

non-optimal coordination

Background

Existing solutions to RO

8	**-12**	**-12**
-12	**0**	**0**
-12	**0**	**6**

1. biased representation
(Weighted QMIX)
place more weight on the representation of good samples

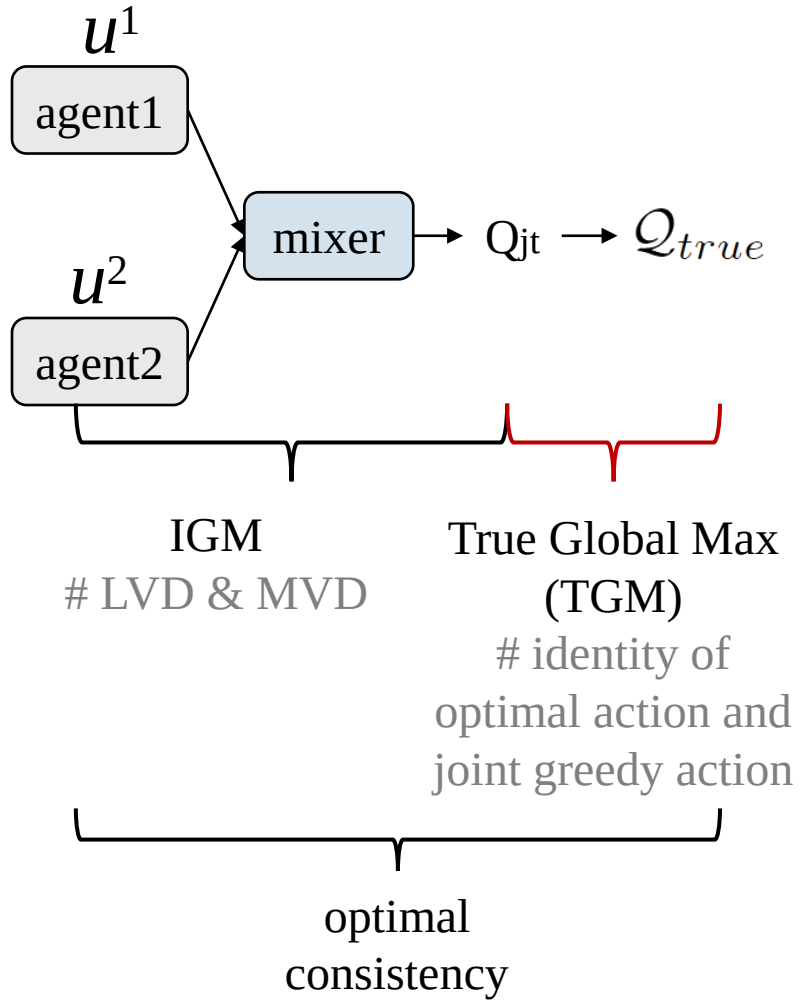
7.9	**-8.33**	**-7.93**
-8.33	**-24.1**	**-23.7**
-7.93	**-23.7**	**-23.3**

concerns:
- rely on heuristic parameters
- can not ensure to solve the RO

2. complete representation
(QPLEX, Qtran)
design joint Q value functions with complete representation capability

8	**-12**	**-12**
-12	**0**	**0**
-12	**0**	**6**

concerns：
- CRC is complex and unnecessary



Analysis

Conditions of the TGM principle for LVD and MVD

According to the equation system

$$\{\mathcal{U}_i^1 + \mathcal{U}_j^2 = \mathcal{Q}_{ij}\}_{i,j\in[1,m]}$$

Under the e-greedy visitation

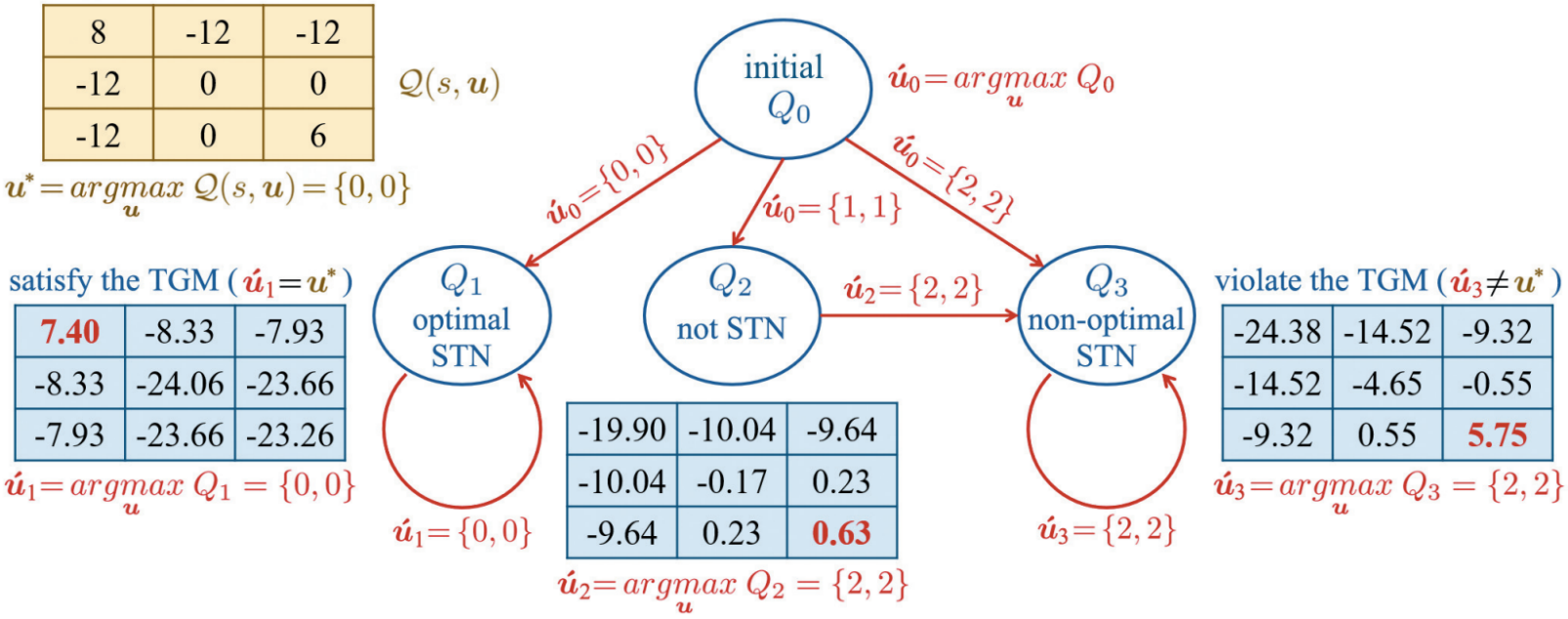
($\acute{i}$, $\acute{j}$ denotes the greedy action of agent 1, 2)

$$\mathcal{U}_i^1 = \frac{\epsilon}{m}\sum_{k=1}^{m}(\mathcal{Q}_{ik} - \mathcal{U}_k^2) + (1-\epsilon)(\mathcal{Q}_{i\acute{j}} - \mathcal{U}_{\acute{j}}^2)$$

$$\mathcal{U}_j^2 = \frac{\epsilon}{m}\sum_{k=1}^{m}(\mathcal{Q}_{kj} - \mathcal{U}_k^1) + (1-\epsilon)(\mathcal{Q}_{\acute{i}j} - \mathcal{U}_{\acute{i}}^1)$$

The joint Q value function can be acquired

$$Q(u_i^1, u_j^2, \boldsymbol{\tau}) = \frac{\epsilon}{m}\sum_{k=1}^{m}(\mathcal{Q}_{ik} + \mathcal{Q}_{kj}) + (1-\epsilon)(\mathcal{Q}_{\acute{i}j} + \mathcal{Q}_{i\acute{j}}) - \frac{\epsilon^2}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\mathcal{Q}_{ij}$$
$$- \frac{\epsilon(1-\epsilon)}{m}\sum_{k=1}^{m}(\mathcal{Q}_{\acute{i}k} + \mathcal{Q}_{k\acute{j}}) - (1-\epsilon)^2\mathcal{Q}_{\acute{i}\acute{j}}$$

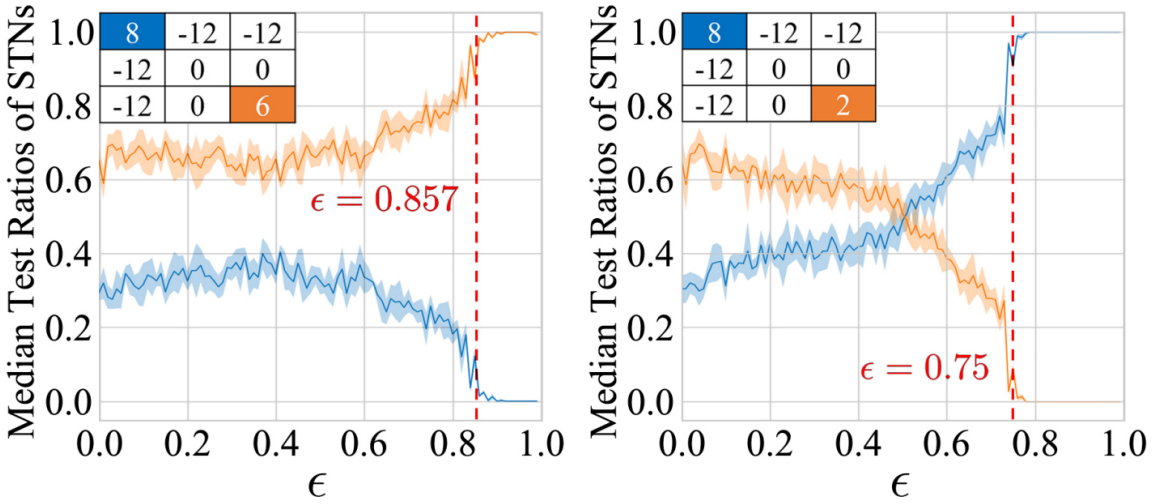


- The joint Q value function changes with the joint greedy action
- The joint Q value function is the same for LVD and MVD

* STN: Self-Transition Node

Analysis

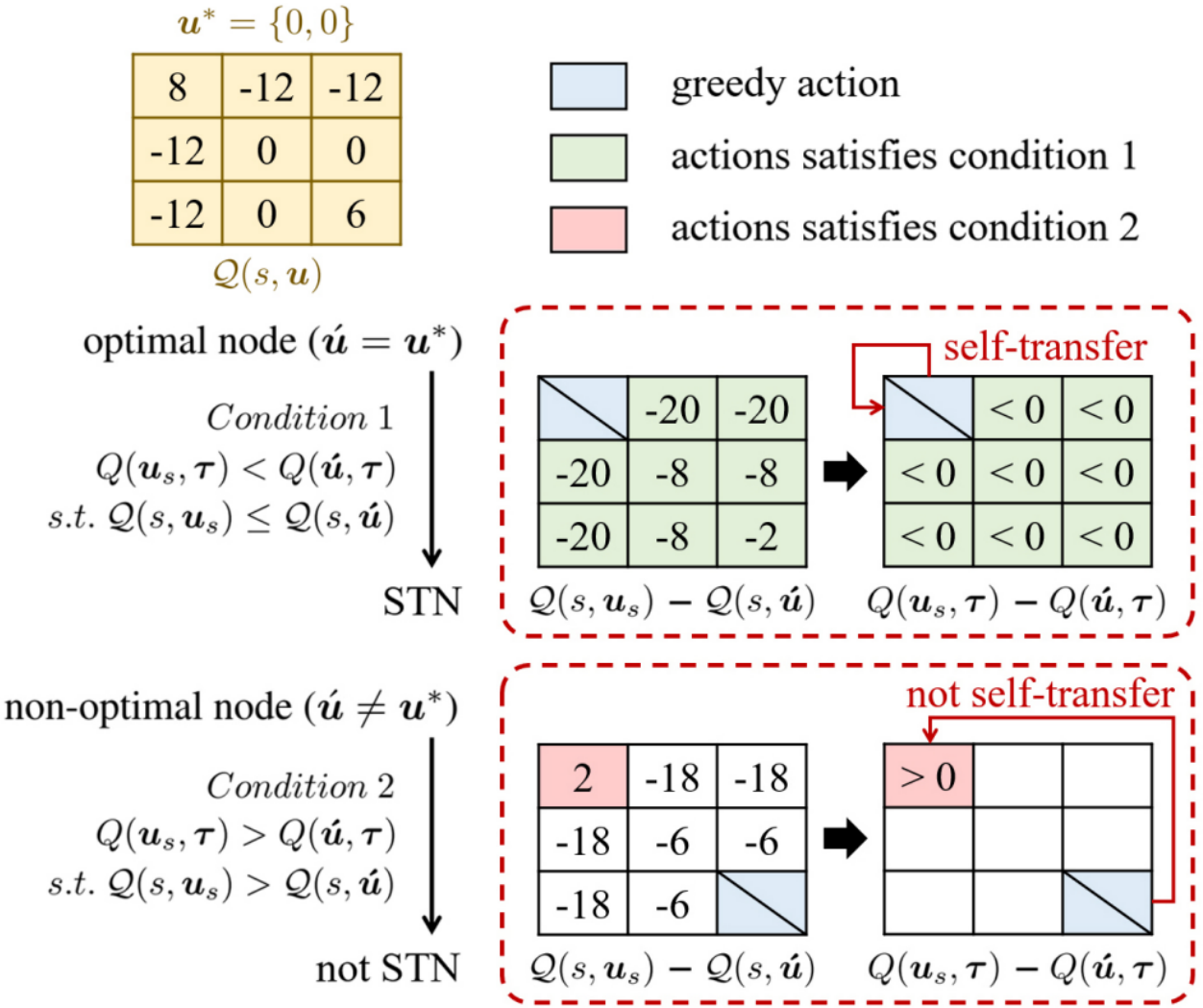
- Improving exploration helps to eliminate extra STNs



Ensuring TGM = the optimal node is the only STN

sufficient condition:

$$\#\ Condition\ 1$$
$$Q(\boldsymbol{u}_s, \boldsymbol{\tau}) < Q(\acute{\boldsymbol{u}}, \boldsymbol{\tau}) \quad s.t.\ \mathcal{Q}(s, \boldsymbol{u}_s) \leq \mathcal{Q}(s, \acute{\boldsymbol{u}})$$
$$\#\ Condition\ 2$$
$$Q(\boldsymbol{u}_s, \boldsymbol{\tau}) > Q(\acute{\boldsymbol{u}}, \boldsymbol{\tau}) \quad s.t.\ \mathcal{Q}(s, \boldsymbol{u}_s) > \mathcal{Q}(s, \acute{\boldsymbol{u}})$$



Methodology

Inferior Target Shaping (ITS)

Instead of representing the specific Q value of an inferior action, reshaping it with a Q value "no better than current greedy"

$$\mathcal{Q}_{its}(s,\boldsymbol{u}) = Q(\acute{\boldsymbol{u}},\boldsymbol{\tau}) - \alpha|Q(\acute{\boldsymbol{u}},\boldsymbol{\tau})|$$
$$s.t.\ \ \mathcal{Q}(s,\boldsymbol{u}) \le \mathcal{Q}(s,\acute{\boldsymbol{u}}) * (1+e_{Q0})\ and\ \boldsymbol{u} \ne \acute{\boldsymbol{u}}$$

8	-12	-12
-12	0	0
-12	0	6

$\mathcal{Q}(s,\boldsymbol{u})$

$\mathcal{Q}_{its}(s,\boldsymbol{u})$	(a)	(b)	(c)
$\acute{\boldsymbol{u}}$	$\{0,0\}$	$\{1,1\}$	$\{2,2\}$
$Q(\acute{\boldsymbol{u}},\boldsymbol{\tau})$	7.8	-0.2	5.8
$\mathcal{Q}_{its}(s,\boldsymbol{u})$ for inferior actions: $Q(\acute{\boldsymbol{u}},\boldsymbol{\tau}) - \alpha\|Q(\acute{\boldsymbol{u}},\boldsymbol{\tau})\|$	7.02	-0.22	5.22

greedy action
inferior action
superior action

8	7.02	7.02
7.02	7.02	7.02
7.02	7.02	7.02

(a)

8	-0.22	-0.22
-0.22	0	0
-0.22	0	6

(b)

8	5.22	5.22
5.22	5.22	5.22
5.22	5.22	6

(c)

$$Q(\boldsymbol{u}_s,\boldsymbol{\tau}) - Q(\acute{\boldsymbol{u}},\boldsymbol{\tau})$$
$$= n(\eta_1-\eta_2)\left[\mathcal{Q}(s,\acute{\boldsymbol{u}}) - (1-\alpha)Q(\acute{\boldsymbol{u}},\boldsymbol{\tau})\right] + n\eta_1 e_Q \mathcal{Q}(s,\acute{\boldsymbol{u}}) \quad (1)$$

- Optimal node (fig (a))

$$Q(\boldsymbol{u}_s,\boldsymbol{\tau}) - Q(\acute{\boldsymbol{u}},\boldsymbol{\tau})$$
$$= n(\eta_1-\eta_2)\frac{\alpha}{w}Q(\acute{\boldsymbol{u}},\boldsymbol{\tau}) + n\eta_1 e_Q \mathcal{Q}(s,\acute{\boldsymbol{u}}) < 0$$

The optimal node is always STN

- Non-optimal node (fig (b), (c)). Let Eq.1>0:

We have $\frac{\eta_1}{\eta_2} > \frac{\alpha}{\alpha + e_{Q0}}$ (2)

Improving the probability of superior actions helps removing non-optimal STNs

* $e_Q = \frac{\mathcal{Q}_{its}(s,\boldsymbol{u}_s) - \mathcal{Q}(s,\acute{\boldsymbol{u}})}{\mathcal{Q}(s,\acute{\boldsymbol{u}})}$ $\eta_1 = (\frac{\epsilon}{m})^{n-1}$ $\eta_2 = (1-\epsilon+\frac{\epsilon}{m})^{n-1}$

Methodology

Raising the probabilities of superior actions

1. Improving exploration

- Too large exploration rate

$$\frac{\eta_1}{\eta_2} > \frac{\alpha}{\alpha + e_{Q0}} \quad \text{equals} \quad \epsilon > \frac{m}{(\frac{e_{Q0}}{\alpha})^{\frac{1}{n-1}} + 1 + m - 1}$$

2. Reweighting samples

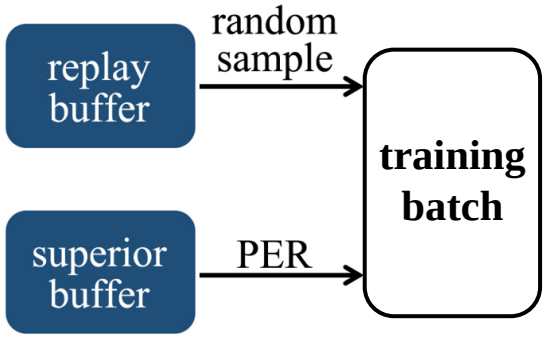
required sample weight $w > \frac{\alpha(\eta_2 - \eta_1)}{e_{Q0}\eta_1}$

m^n	3^2	3^3	3^4
w_0 (Eq.57)	3.60	50.32	**659.50**
$Q(\boldsymbol{u}_s, \boldsymbol{\tau}) - Q(\acute{\boldsymbol{u}}, \boldsymbol{\tau})$ (Eq.56)	0.01 ±0.06 (0)	-0.02 ±0.30 (0)	-0.48 ±0.75 (0)
Test $Q(\acute{\boldsymbol{u}}, \boldsymbol{\tau})$	5.95 ±0.02	5.90 ±0.06	5.93 ±0.03

- Required sample weight grows exponentially as the number of agents

Superior Experience Replay (SER)

remove the correlation of a superior action's proportion in training batch with its exploration probability

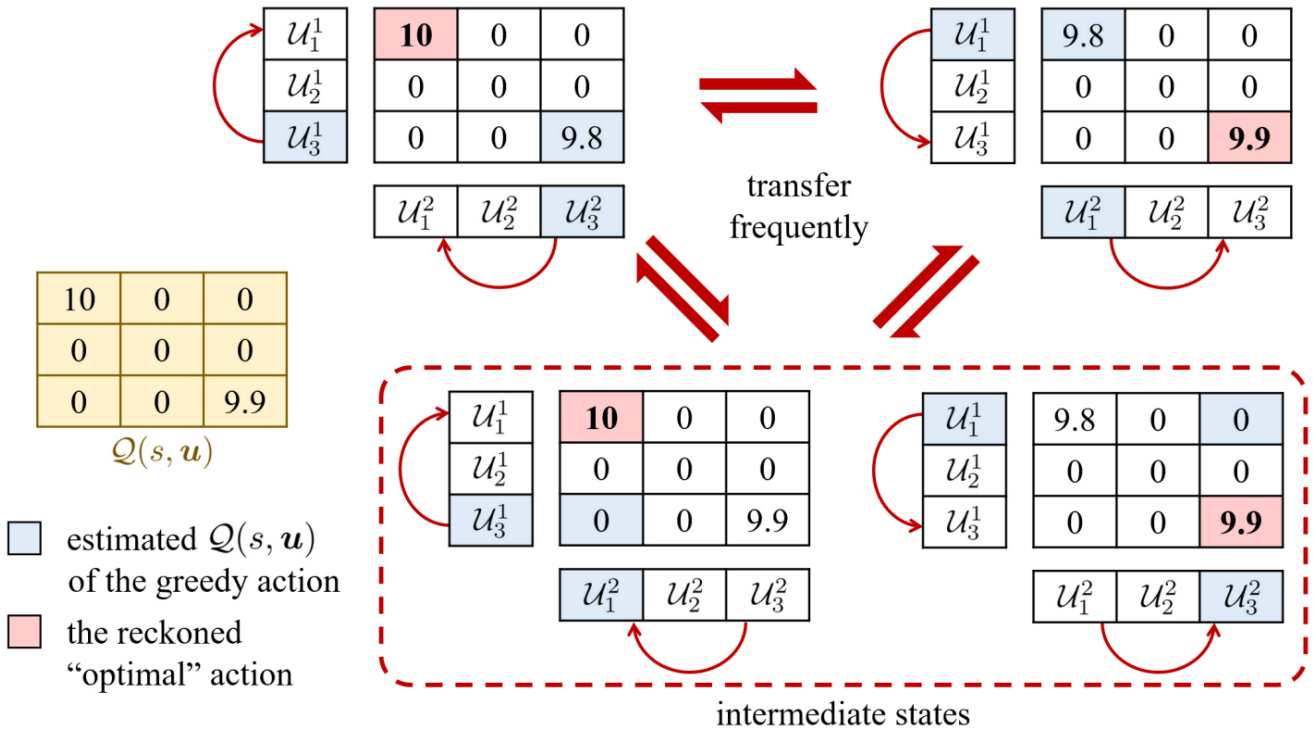


Relative weight for samples from superior buffer

$$w_{ser} > \frac{\alpha}{e_{Q0}}(\eta_2 - \eta_1)\eta_s - \eta_1\eta_s$$

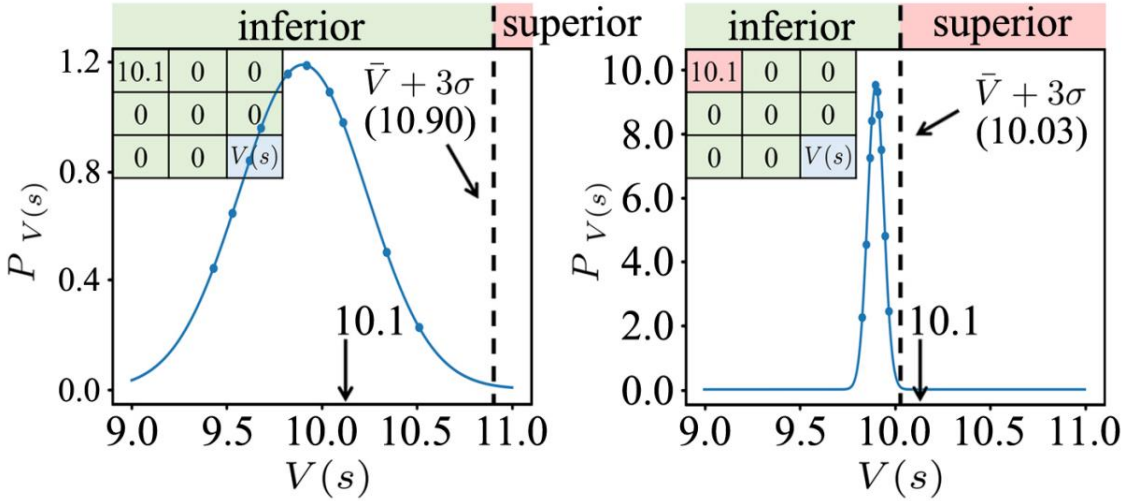
Methodology

optimality-stability paradox



adaptive trade-off between optimality and stability

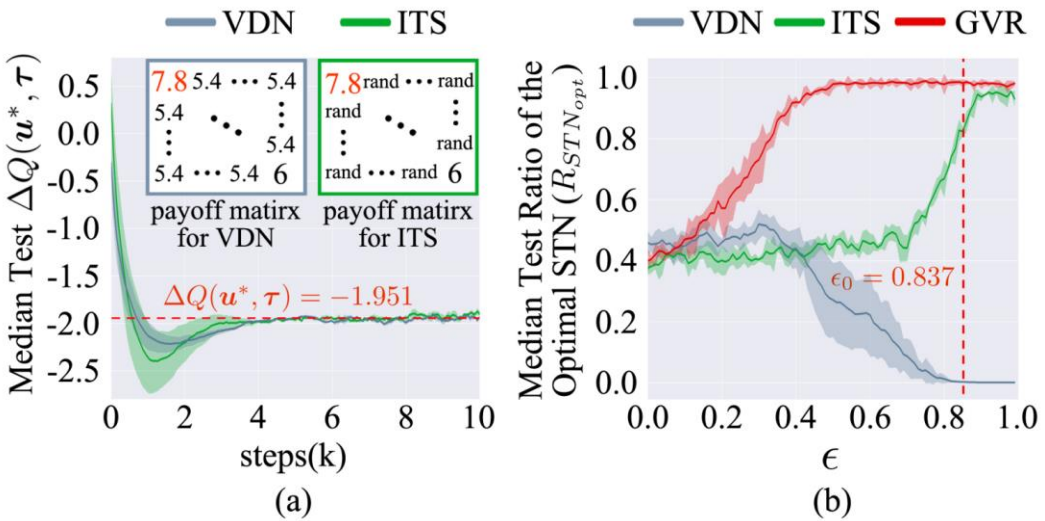
filter out superior actions with large estimation uncertainty



Experiments

Matrix Games

- Verification of optimal consistency under sufficient exploration

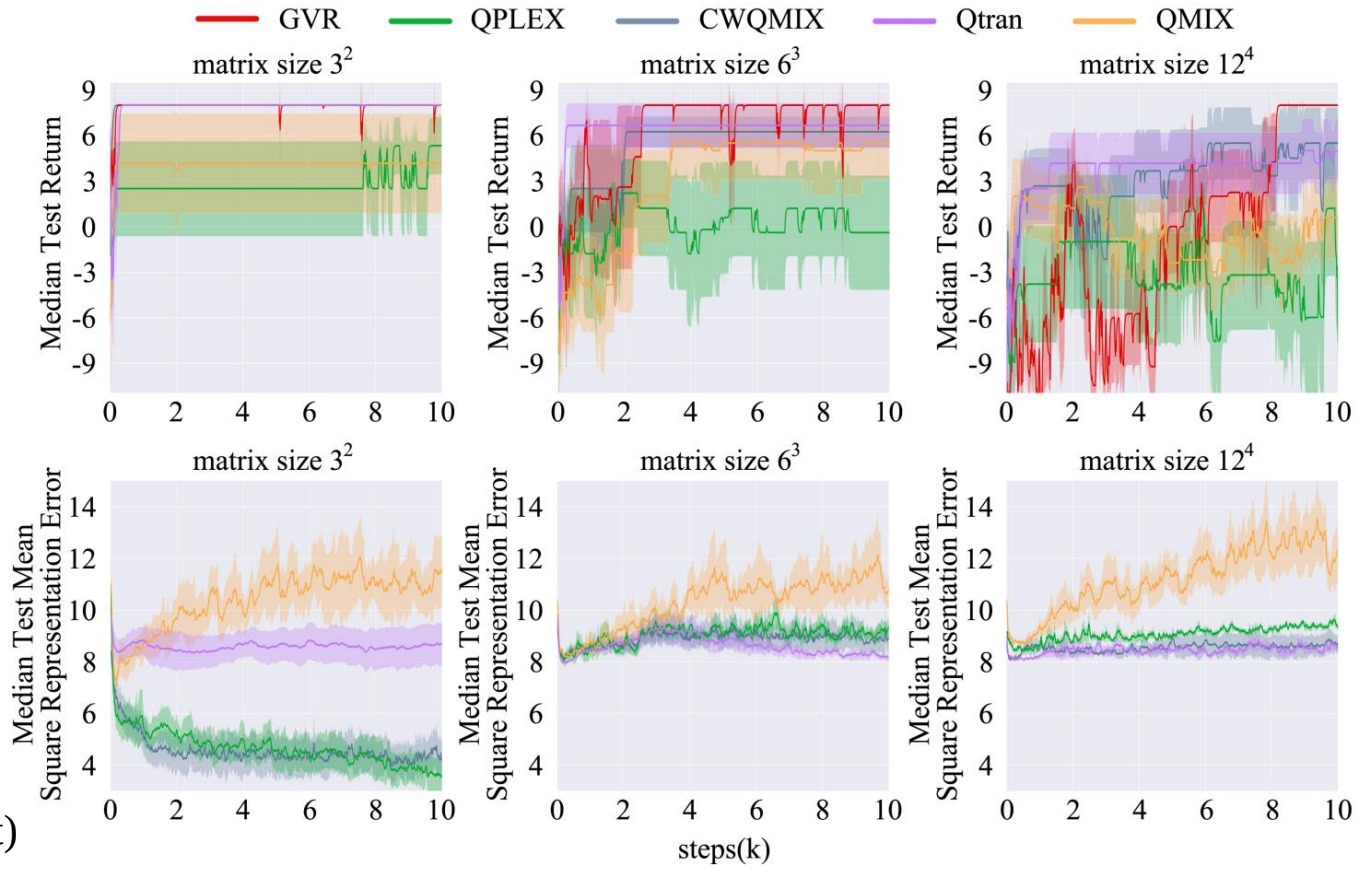


Experimental settings

$$\mathcal{Q}(s, \boldsymbol{u}) = \begin{cases} 6(1 + e_Q) & \boldsymbol{u} = \{0, 0\} \\ 6 & \boldsymbol{u} = \{m, m\} \\ random(-20, 6) & others \end{cases}$$

$\alpha = 0.1, e_Q = 0.3$ (left) $\alpha = 0.1, \epsilon = 0.2, e_Q = 1/3$ (right)

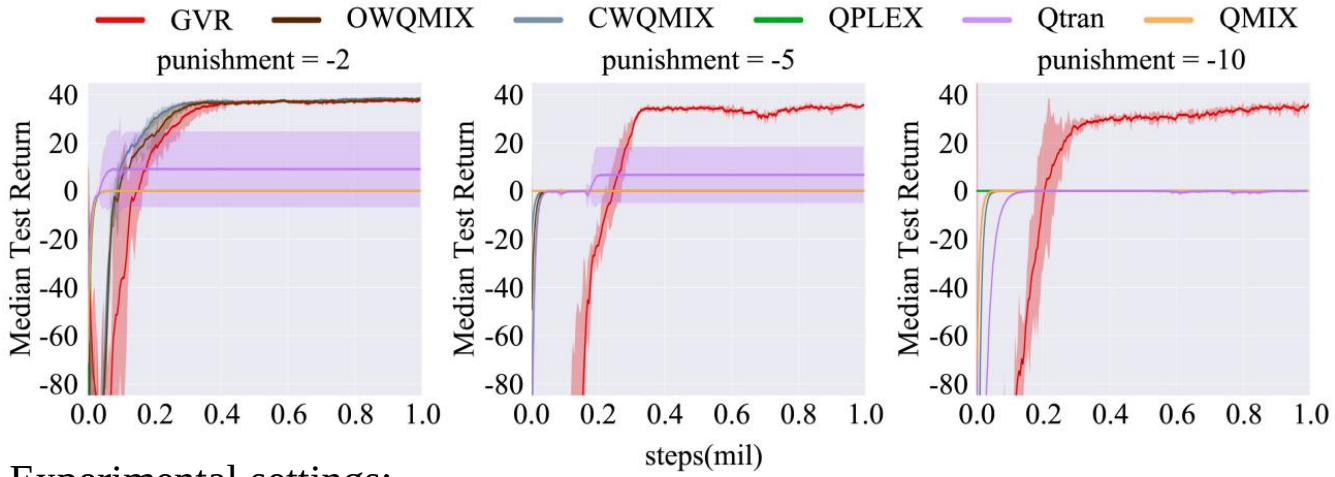
- Comparison with baselines



Experiments

Predator-Prey

- Preys are assigned with random policies
- suppose k agents capture the same prey at the same time-step
  - reward=0 if k=0
  - reward<0 (=punishment) if k=1
  - reward>0 if k>1



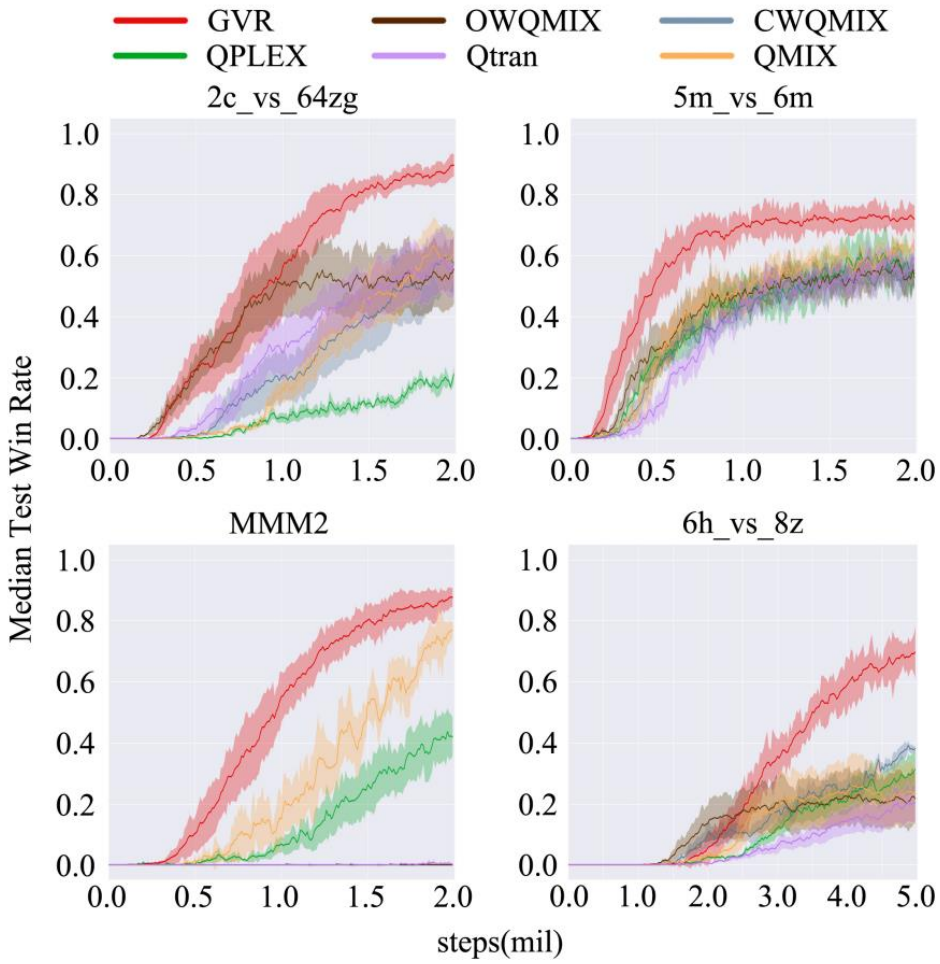
Experimental settings:

RNN shared by all agents as the utility functions

MVD for GVR, α=0.2

exploration rate damps from 1 to 0.05 during 1m steps (for 6h_vs_8z) or 50k steps (for the others)

Starcraft Multi-agent Challenge (SMAC)



Conclusions

1. Relative over generalization arises from different representations under different action distributions

- There is no difference between LVD and MVD from the perspective of value representation.
- When ε grows to 1, the action distributions under different greedy actions converge to the uniform distribution, where the representations also converge.

2. Solving the RO is equivalent to ensure the optimal consistency

- A sufficient condition to ensure the optimal consistency is: compared to the greedy action, the joint Q value of an inferior action is smaller; the joint Q value of a superior action is larger.
- Inferior target shaping ensures the optimal node is an STN, but removing the non-optimal STN rely on further raising the probability of superior actions.

3. The paradox in optimality and stability

- A trade-off between the optimality and stability can be achieve by moderate tolerance on sub-optimality.

Thanks !