Explicitly Learning Policy Under Partial Observability in Multiagent Reinforcement Learning

Chen Yang[1,2], Guangkai Yang[1,2], Hao Chen[1], Junge Zhang[1,2†]
[1]*School of Artificial Intelligence, University of Chinese Academy of Sciences*
[2]*Institute of Automation, Chinese Academy of Sciences*
Beijing 100049, P.R.China
{yangchen2021, yangguangkai2019, chenhao2019}@ia.ac.cn, jgzhang@nlpr.ia.ac.cn

CONTENTS

CONTENTS

1. Introduction

University of Chinese Academy of Sciences

One key problem in MARL

Each agent in the multiagent system only observes part of the environment and has to make individual and independent decisions based only on local information, making it difficult to reach comprehensive coordination among agents. Such problem is also known as **partial observability**.

Current solutions

- Mostly based on empirical approach.
- Implicitly alleviate partial observability.
- Low learning efficiency.

Our approach

- We derive an ideal form of policy that maximizes MARL objective under partial observability.
- We propose a method that explicitly learns the optimal policy under partial observability.

2. Methodology

The Optimal Policy under Partial Observability

MARL Objective:

$$\boldsymbol{\pi}(s) = \underset{\boldsymbol{a}}{\operatorname{argmax}}\ Q_{tot}(s, \boldsymbol{a}).$$

$$\downarrow$$

$$\pi_i(s) = \underset{a}{\operatorname{argmax}}\ Q_i(s, a).$$

$$\downarrow$$

$$\pi_i(\tau_i) = \underset{a}{\operatorname{argmax}} \sum_s b_i(s) Q_i(s, a),$$

$$\downarrow$$

$$\pi_i(\tau_i) = \underset{a}{\operatorname{argmax}} \sum_s p(s|\tau_i) Q_i(s, a). \longrightarrow \text{the optimal policy}$$

IGM Condition:

$$\underset{\boldsymbol{a}}{\operatorname{argmax}}\ Q_{tot}(s, \boldsymbol{a}) = \begin{pmatrix} \operatorname{argmax}_{a_1}\ Q_1(s, a_1) \\ \vdots \\ \operatorname{argmax}_{a_n}\ Q_n(s, a_n) \end{pmatrix}.$$

2. Methodology

Overall Framework of ELP

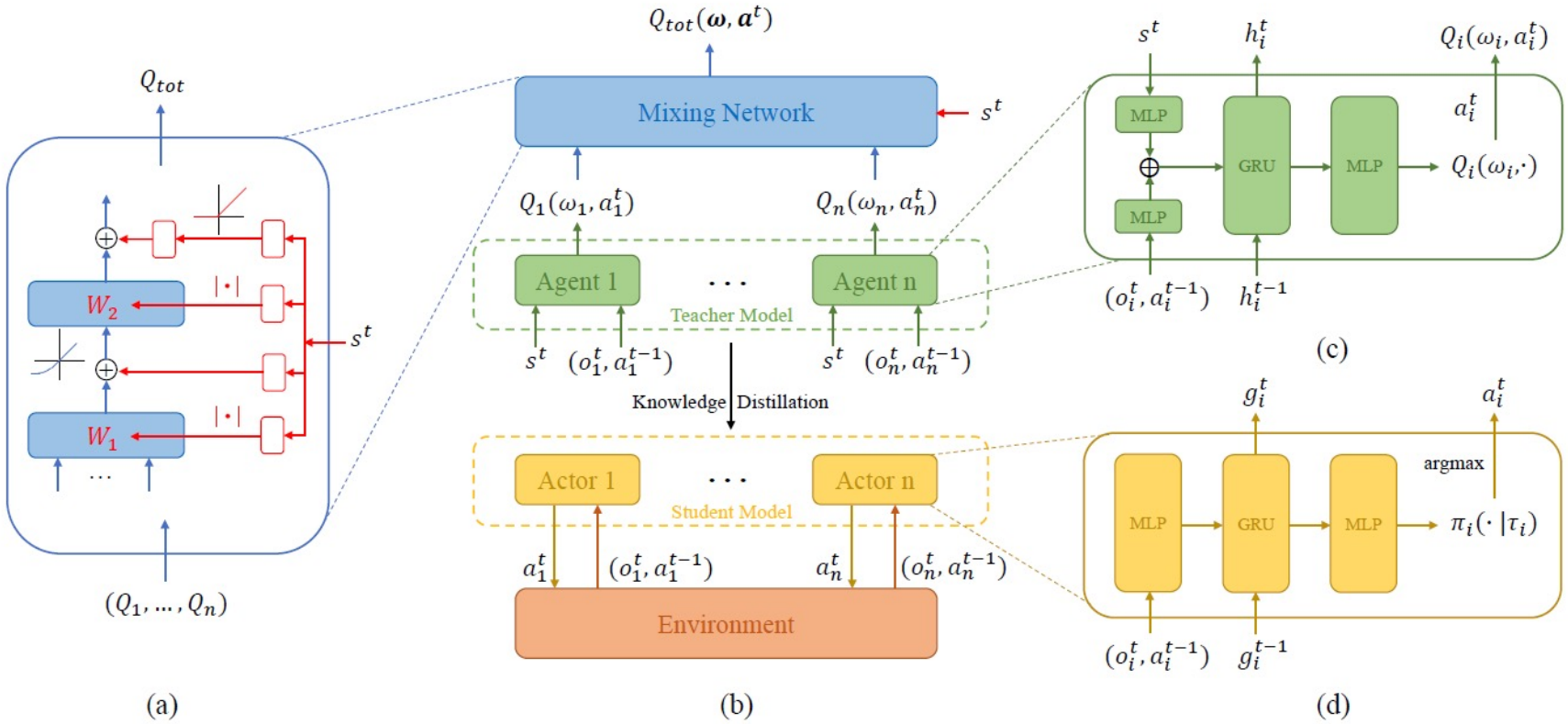

Fig. 1. The overall framework of ELP. (a) The mixing network structure. It takes Q_i as input and outputs Q_{tot}. (b) The overview of teacher-student structure. The teacher model updates the student model through knowledge distillation. (c) The global Q network structure. It estimates $Q_i(s, a)$ for agent during training. (d) The local executor network structure. It conducts deterministic policy for agent during execution.

2. Methodology

Training Procedure

To update teacher model, we minimize TD loss:

$$L_{TD}(\theta) = \sum_{b}(Q_{tot}(\boldsymbol{\omega};\theta) - r - \gamma \max_{\boldsymbol{a}'} Q_{tot}(\boldsymbol{\omega}';\theta^{-}))^2.$$

To update student model, we minimize KD loss:

$$L_{KD} = \sum_{b}\sum_{i\in\mathcal{N}}\sum_{a\in\mathcal{A}} O_i(a)(logO_i(a) - logS_i(a)).$$

The teacher model outputs Q_i, and the student model outputs O_i.We denote S_i as softmax(Q_i), and the policy of agent in execution formulates as :

$$\pi_i = \text{argmax}_a O_i(a)$$

Algorithm 1 Optimization Procedure of ELP

Initialize: hyperparameters.
Initialize: θ for teacher model and ϕ for student model.
Initialize: target parameters $\theta^{-} = \theta$.
while not terminated **do**
  Sample a batch of b trajectories from experience buffer;
  # Teacher Model Update
  Update θ by minimizing the TD loss
  # Student Model Update
  Update ϕ by minimizing the KD loss
  Update target network parameters $\theta^{-} = \theta$ in teacher model with period I;
end while

2. Methodology

Knowledge Distillation

$$L_{KD} \propto \sum_{s \in \mathcal{S}} p(s|\tau_i) O_i(a) (log O_i(a) - log S_i(a)).$$

$$\downarrow$$

$$O_i(a) \propto \exp[\sum_{s \in \mathcal{S}} p(s|\tau_i) log S_i(a)].$$

$$\downarrow$$

$$O_i(a) \propto \exp[\sum_{s \in \mathcal{S}} p(s|\tau_i) Q_i(s,a)].$$

$$\downarrow$$

$$\pi_i(\tau_i) = \operatorname*{argmax}_a \exp[\sum_s p(s|\tau_i) Q_i(s,a)]$$

$$= \operatorname*{argmax}_a \sum_s p(s|\tau_i) Q_i(s,a). \longrightarrow \text{the optimal policy}$$

3 Experiments

Experiment Setup

SMAC: StarCraft **M**ulti-**A**gent **C**hallenge

Table 1: Units configuration of selected maps in SMAC.

Map Name	Ally Units	Enemy Units
8m	8 Marines	8 Marines
3s5z	3 Stalkers & 5 Zealots	3 Stalkers & 5 Zealots
5m_vs_6m	5 Marines	6 Marines
2c_vs_64zg	2 Colossi	64 Zerglings
MMM2	1 Medivac, 2 Marauders & 7 Marines	1 Medivac, 3 Marauders & 8 Marines
27m_vs_30m	27 Marines	30 Marines

Baseline methods: QMIX, WQMIX(CW-QMIX & OW-QMIX), QTRAN, QPLEX, Qatten.

3 Experiments

Main Results

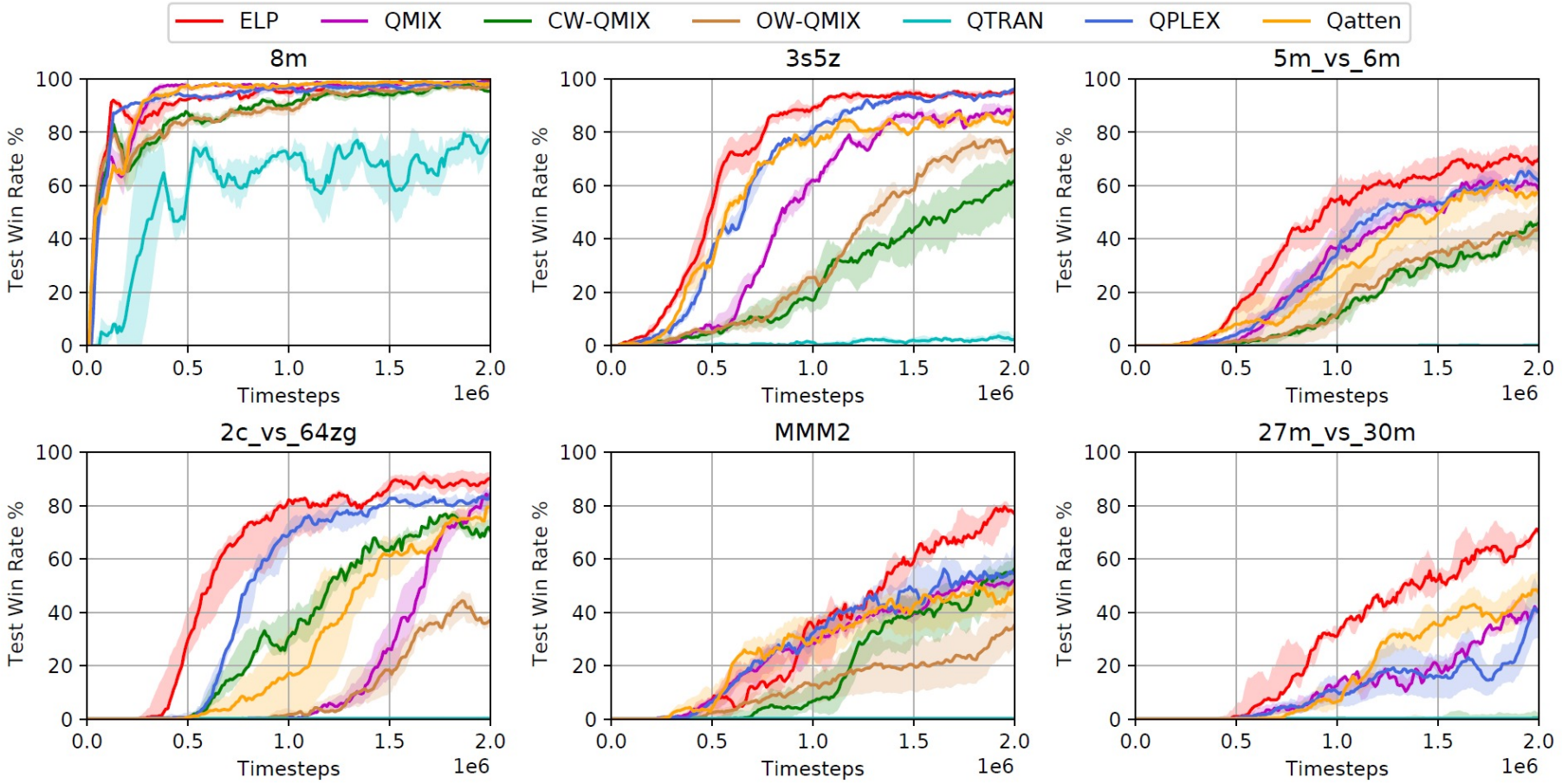

Fig. 2. Learning curves of SMAC over all 6 maps.

3 Experiments

Ablation Study 1

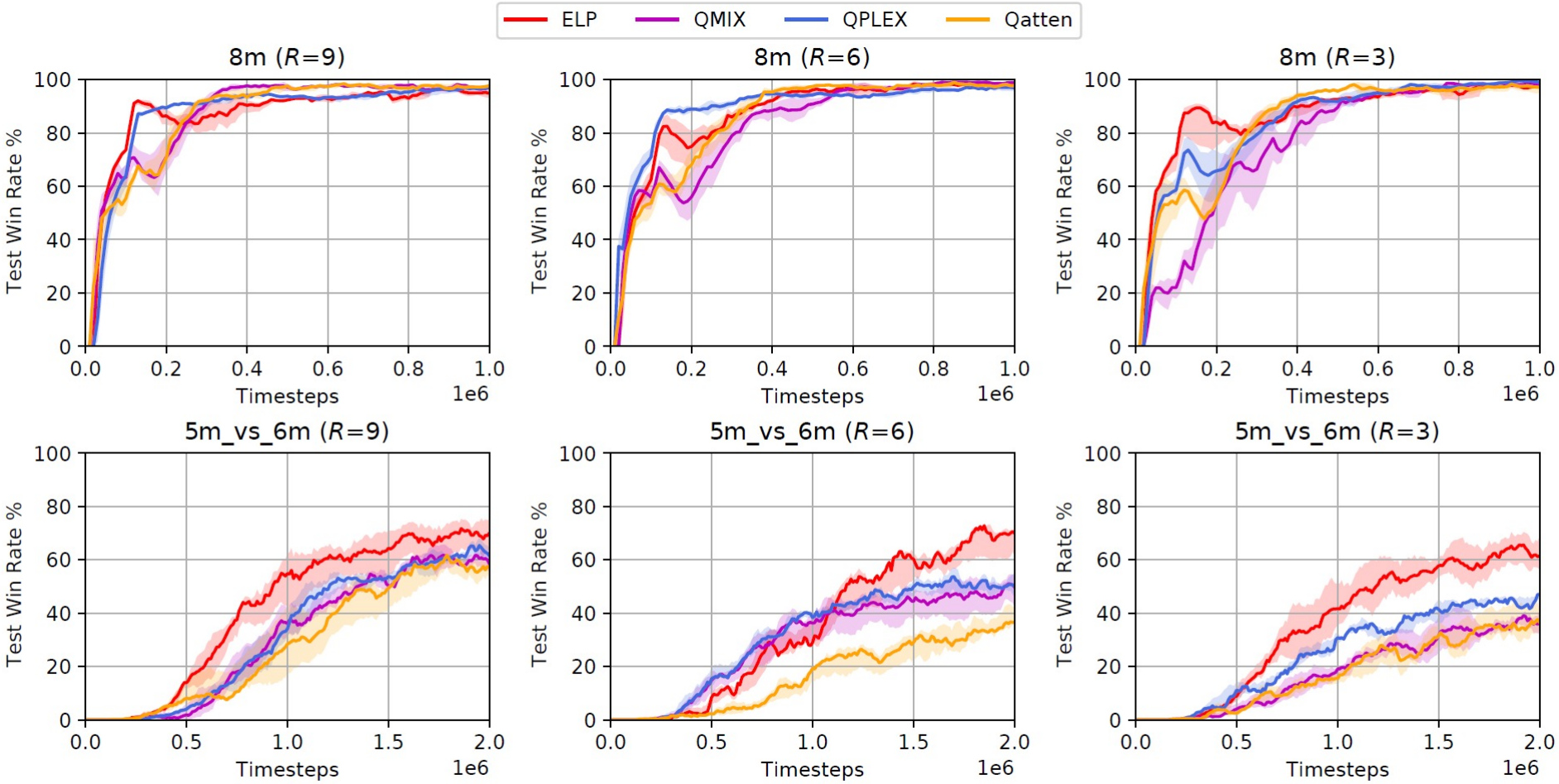

Fig. 4. Learning curves of SMAC under different sight ranges.

3 Experiments

Ablation Study 2

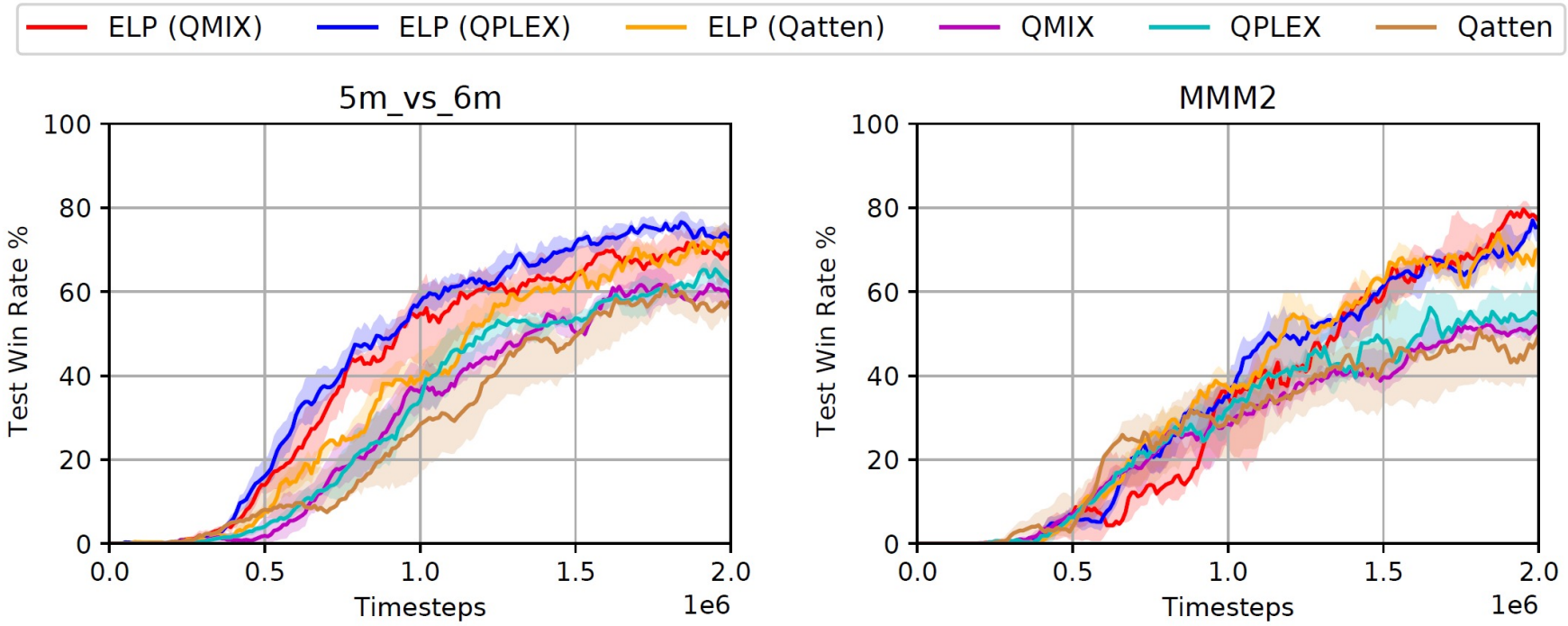

Fig. 5. Ablations for mixing networks in teacher model.

3 Experiments

Ablation Study 3

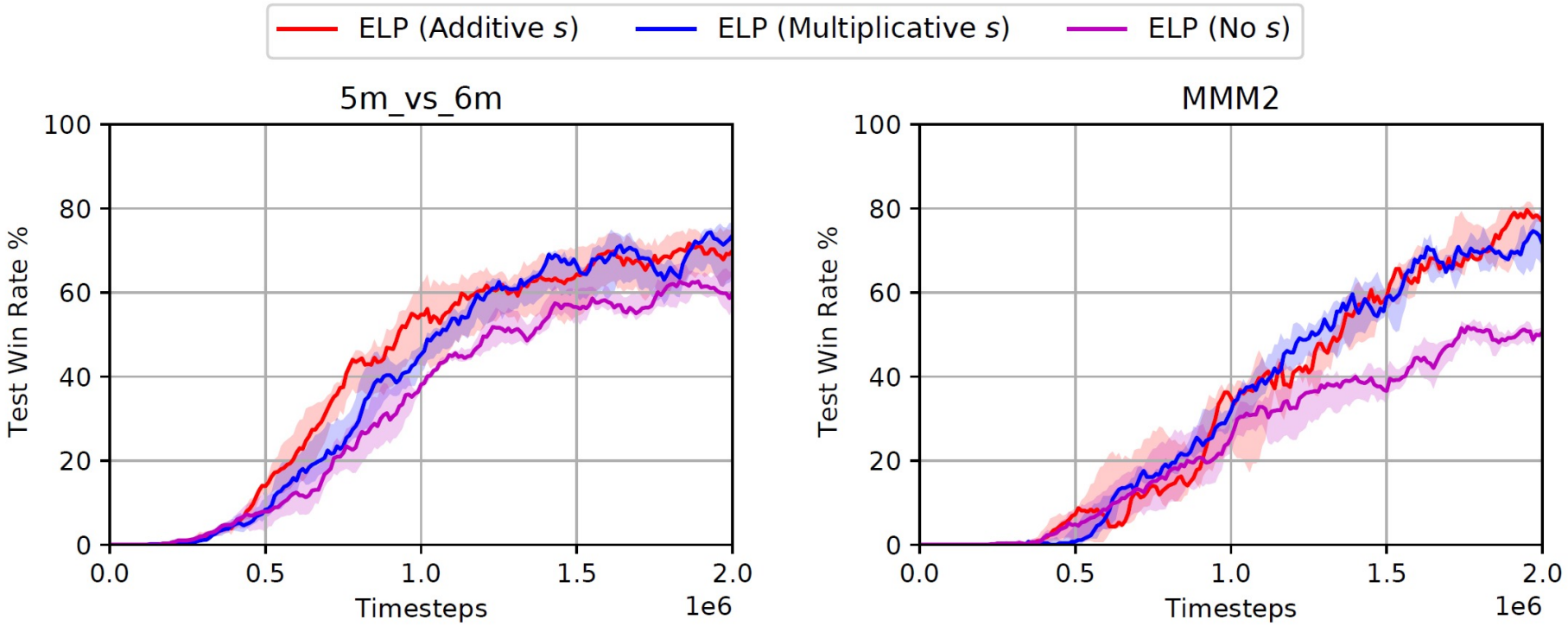

Fig. 6. Ablations for the inclusion of s in global Q networks.

4 Conclusion

Our contribution

- We formally analyze and derive the optimal policy under partial observability in MARL.
- We propose a knowledge distillation approach to explicitly learn towards optimal policies.
- The procedure in our method can easily extend to existing methods.

Future work

- We plan to extend our approach to policy-based methods.


中国科学院大学
University of Chinese Academy of Sciences
CASIA
Thank you.