Underexplored Subspace Mining for Sparse-Reward Cooperative Multi-Agent Reinforcement Learning

Yang Yu[1,2], Qiyue Yin[1,2], Junge Zhang[1,2], Hao Chen[3], Kaiqi Huang[1,2,4]
[1]*School of Artificial Intelligence, University of Chinese Academy of Sciences*
[2]*Institute of Automation, Chinese Academy of Sciences*
[3]*University of Chinese Academy of Sciences*
[4]*CAS Center for Excellence in Brain Science and Intelligence Technology*
Beijing 100049, P.R.China
yuyang2019@ia.ac.cn, chenhao915@mails.ucas.ac.cn, {qyyin, jgzhang, kqhuang}@nlpr.ia.ac.cn

CONTENTS

CONTENTS

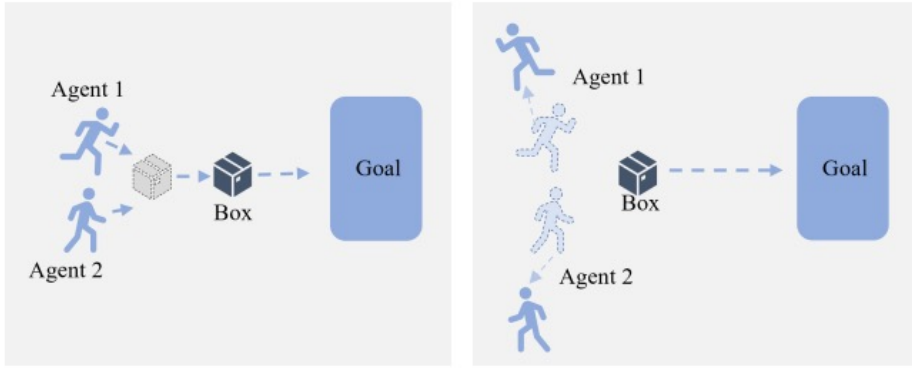
1. Introduction

One key Problem in MARL

- **sparse-reward cooperative MARL:** agents are required to cooperate over a long-time horizon to obtain a team reward.

Current Solutions

- Using influence among agents as intrinsic rewards
- Using hierarchical control to learn joint exploration
- ➢ Suffering from exploration in the joint state space



Cooperative target is often related to partial attributes, and this is no need to treat the whole state space equally. Thus, we encourage agents to do **selective exploration**, focusing on partial subspace instead of wasting time on the whole state space. Specially, we encourage agents to focus on **the underexplored subspace**, according to the principle of optimism in the face of the uncertainty.

2. Methodology

Bonus-Based Method

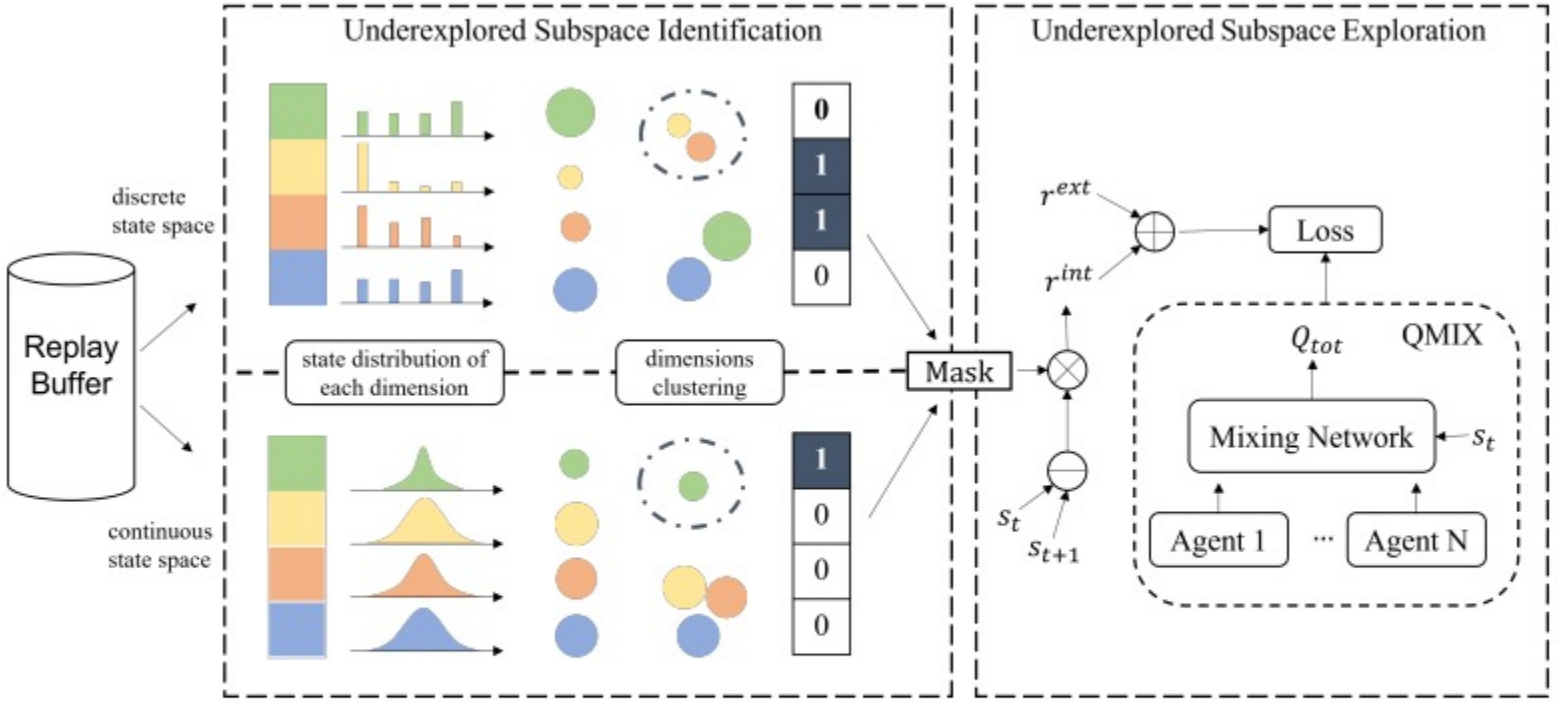


Fig. 2. The framework of USM, which identifies the underexplored subspace and encourages the exploration in this subspace.

$$L(\theta) = \mathbb{E}_{(\boldsymbol{\tau},\boldsymbol{u},r,s)\sim D}\left[\left(y^{tot} - Q_{tot}(\boldsymbol{\tau},\boldsymbol{u},s;\theta)\right)^2\right] \qquad y^{tot} = r^{ext} + \omega r^{int} + \gamma \max_{\boldsymbol{u'}} Q_{tot}(\boldsymbol{\tau'},\boldsymbol{u'},s';\theta^-)$$

2. Methodology

USM in Discrete State Space

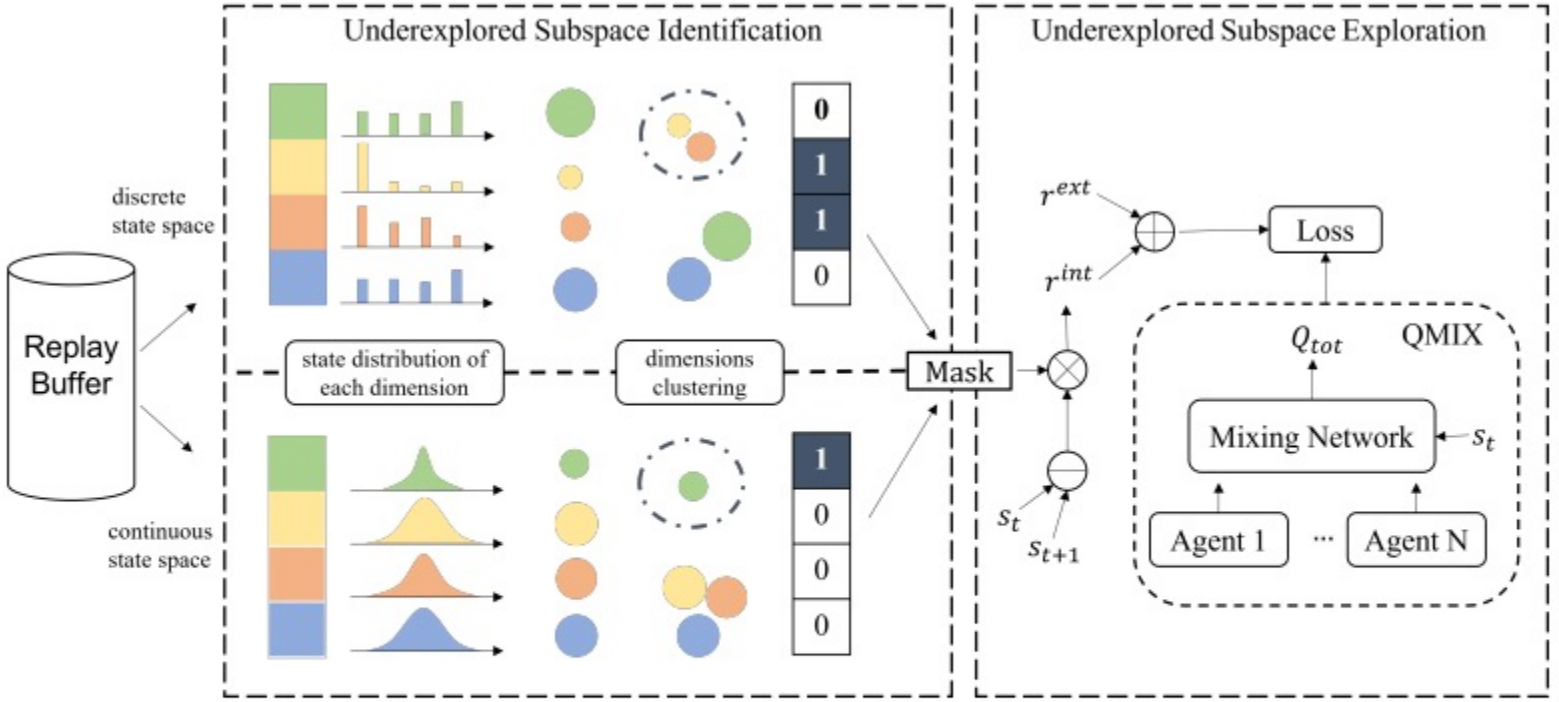


Fig. 2. The framework of USM, which identifies the underexplored subspace and encourages the exploration in this subspace.

$$p_i(*) = \frac{c_i(*)}{\sum_{v \in V_i} c_i(v)} \qquad e_i = \frac{H_i}{H_{max,i}} = \frac{-\sum_{v \in V_i} p_i(v) \log p_i(v)}{log(|V_i|)} \qquad r_t^{int}(s_t, s_{t+1}) = \frac{\sum_{i=1}^{N} \mathbb{I}\left[s_t^i \neq s_{t+1}^i\right] m_i}{(\sum_{i=1}^{N} m_i)\sqrt{N(s_{t+1})}}$$

2. Methodology

USM in Continuous State Space

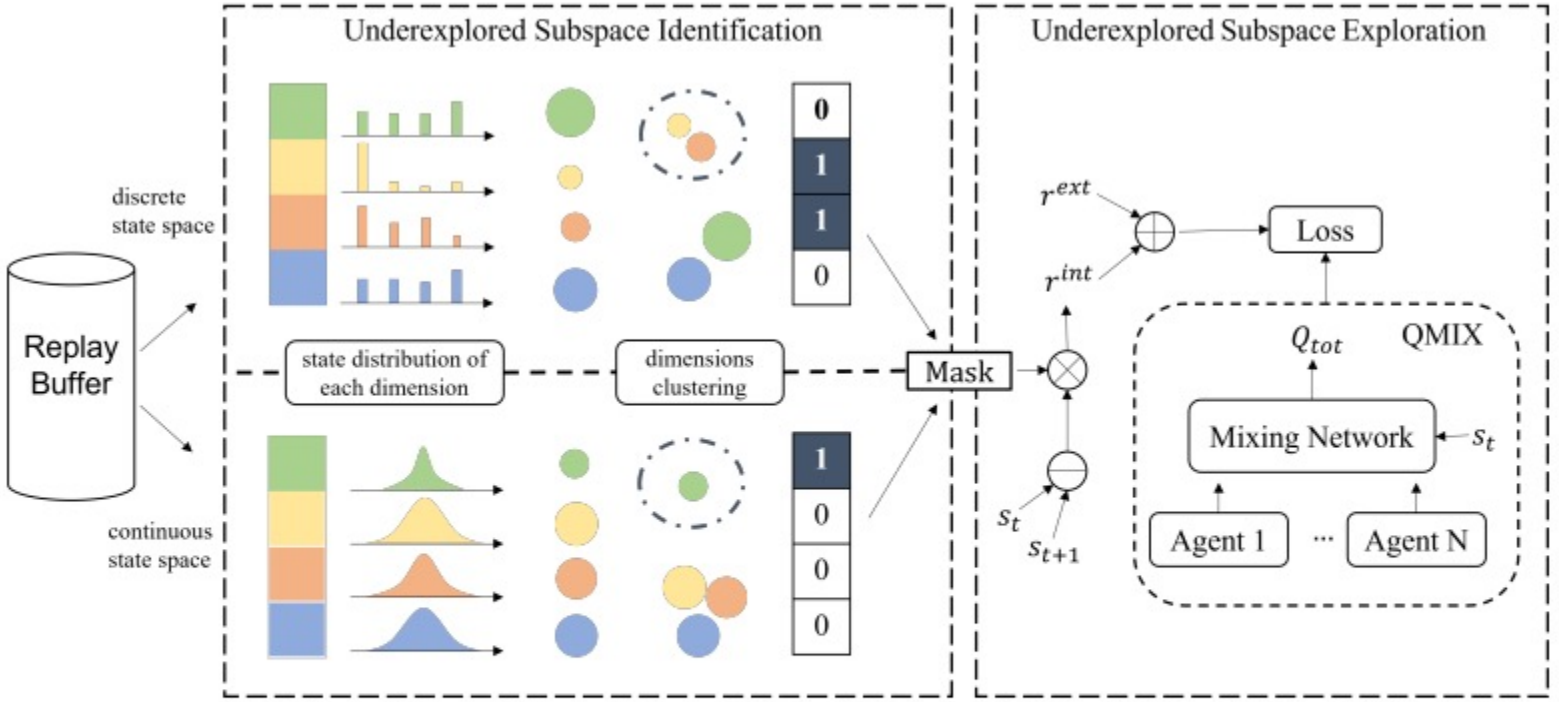


Fig. 2. The framework of USM, which identifies the underexplored subspace and encourages the exploration in this subspace.

$$var_i = \sigma_i^2 = \frac{1}{n}\sum_{d \in D_i}(d - \overline{D_i})^2$$

$$r_t^{int}(s_t, s_{t+1}) = \frac{\sum_{i=1}^{N} |s_t^i - s_{t+1}^i| m_i}{\sum_{i=1}^{N} m_i}$$

3 Experiments

Experiment Setup

TABLE I
STATE SPACE OF PWE AND SMAC

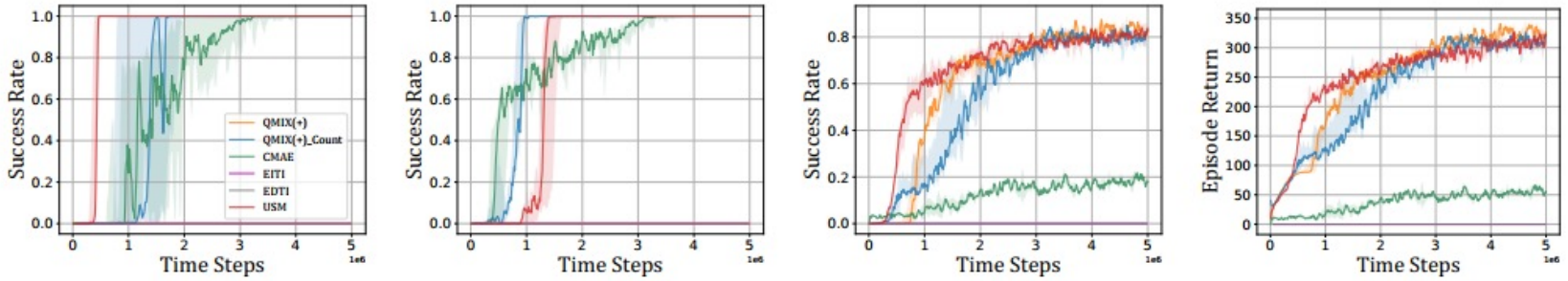
Name	Agents	Dimensions	State Space
PushBox	2	6, discrete	$\approx 1.1 \times 10^7$
Pass	2	5, discrete	$\approx 1.6 \times 10^6$
Island	2	10, discrete	$\approx 1.1 \times 10^{10}$
3m	3_vs_3	21, continuous	-
2m_vs_1z	2_vs_1	12, continuous	-
3s_vs_5z	3_vs_5	35, continuous	-
25m	25_vs_25	175, continuous	-

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

Fig. 3. The descriptions of tasks in PWE and SMAC. (a) PushBox. (b) Pass. (c) Island. (d) $5m$. A task of SMAC. In these tasks, the extrinsic reward or the largest extrinsic reward will not be given unless the cooperation target is achieved, which encourages agents to explore long-term cooperative strategies.

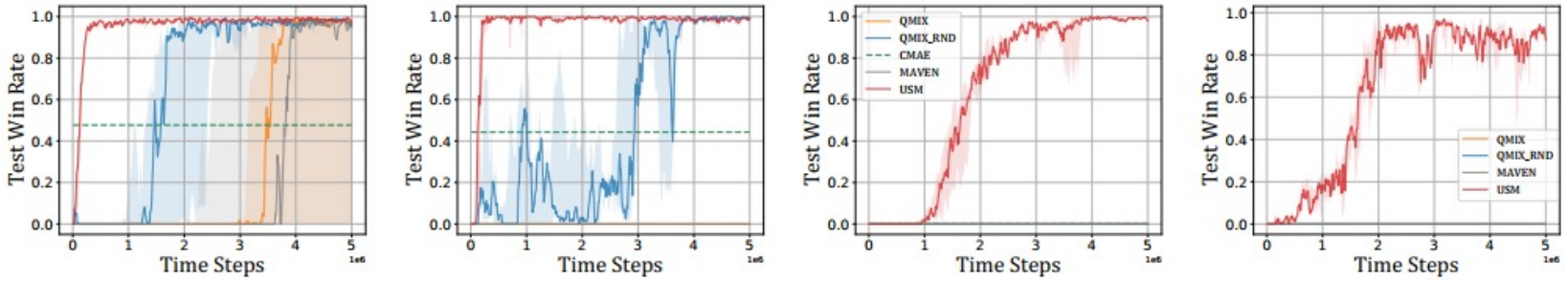
3 Experiments

Main Results



(a) PushBox (b) Pass (c) Island (d) Island

Fig. 4. Learning curves of different methods during training in PWE tasks. (c) and (d) demonstrate the success rate and the episode return of different methods in Island respectively.



(a) $3m$ (b) $2m\_vs\_1z$ (c) $3s\_vs\_5z$ (d) $25m$

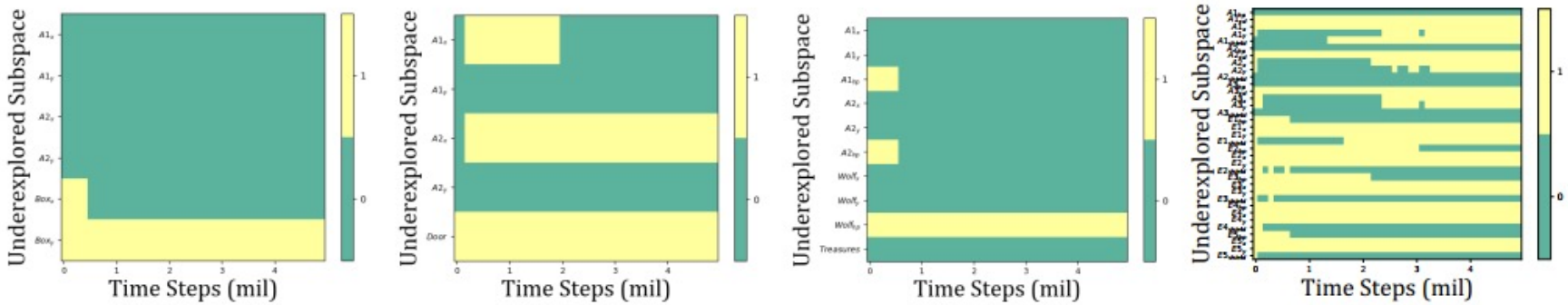
Fig. 5. Test win rate of different methods in SMAC tasks. USM becomes the only method succeeding in games with larger state space like $25m$ or complicated cooperation dynamics like $3s\_vs\_5z$

3 Experiments

Visualization

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

Fig. 3. The descriptions of tasks in PWE and SMAC. (a) PushBox. (b) Pass. (c) Island. (d) $5m$. A task of SMAC. In these tasks, the extrinsic reward or the largest extrinsic reward will not be given unless the cooperation target is achieved, which encourages agents to explore long-term cooperative strategies.



(a) PushBox (b) Pass (c) Island (d) 3s_vs_5z

Fig. 6. Identified underexplored subspace (shown in yellow) of USM in different PWE tasks and SMAC task. In SMAC tasks, the state space describes the health, cooldown, x, y locations, shield of each ally unit and the health, x, y locations, shield of each enemy unit.

3 Experiments

Ablation Studies

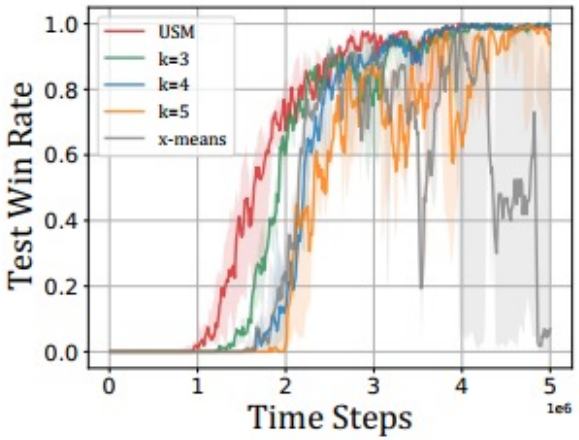
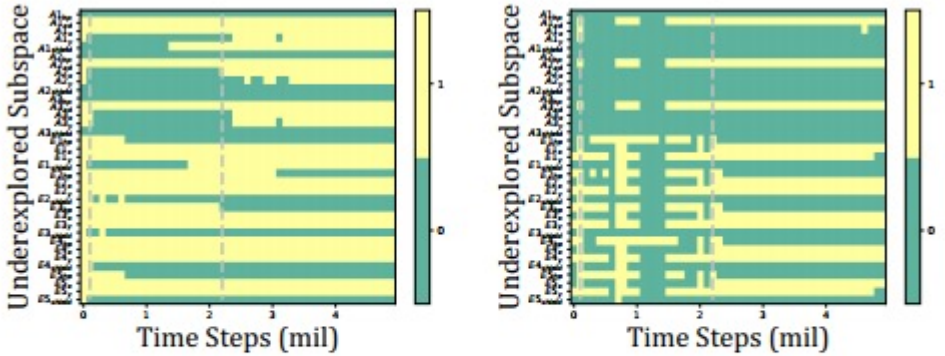


Fig. 7. Ablation study about the clustering method used in USM. Compared with USM, the learning of other numbers of cluster centers is slow and unstable.



(a) USM's focus (b) x-means' focus

Fig. 8. The focused subspace of USM and x-means. There is a difference of the focused area between these two method in different stages divided by gray dotted lines.

4 Conclusion

A novel type of intrinsic reward that encourages agents to do selective exploration

- alleviating the inefficient exploration problem in large state space
- achieving a significant performance especially in challenging sparse-reward cooperative games

Future work

- Apply USM to pixel-input tasks


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