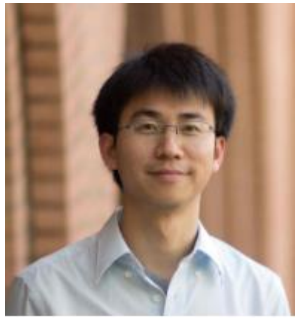
Value-Based Deep Multi-Agent Reinforcement Learning with Dynamic Sparse Training

Pihe Hu*[1], Shaolong Li*[2], Zhuoran Li[1], Ling Pan[3], Longbo Huang[1]

[1]IIIS, Tsinghua University, [2]CST, Central South University, [3]ECE, HKUST







* denotes equal contribution

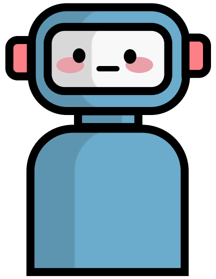
Deep MARL has been successful

StarCraft II
[Mathieu et al., 2021]

Dota 2
[Berner et al., 2019]

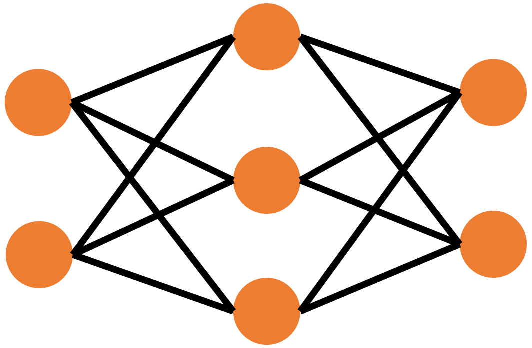
Autonomous Robots
[Chen et al., 2020]

Deep MARL is costly

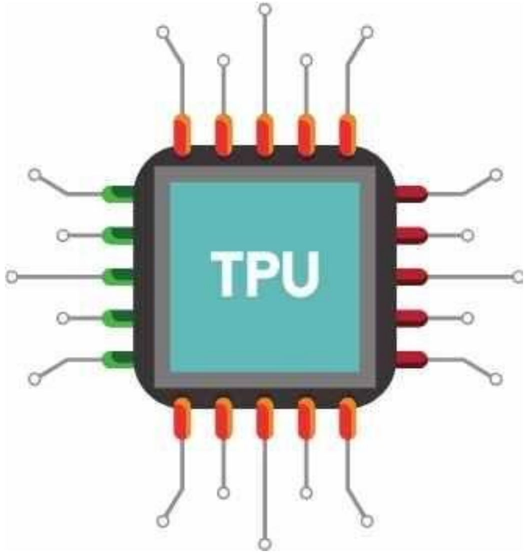


Agent

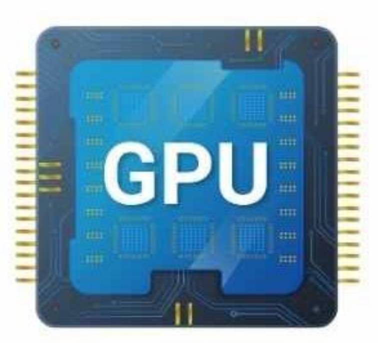
Function Approximation



Deep Neural Network
Parameters up to several Gigabytes

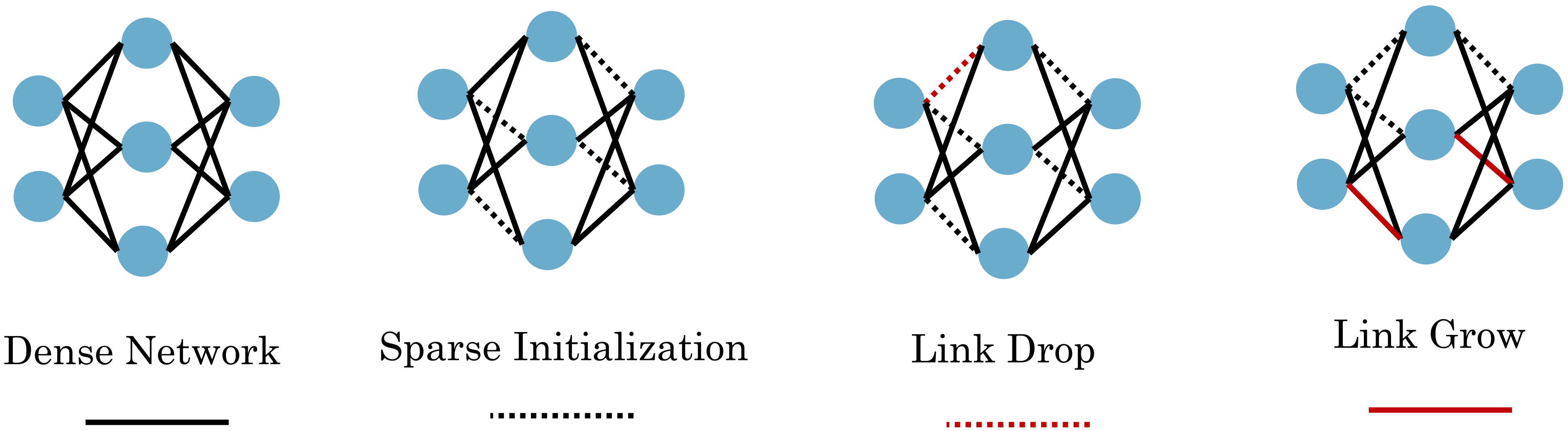


AlphaStar
[Mathieu et al., 2021]
16 TPUs
~14 days



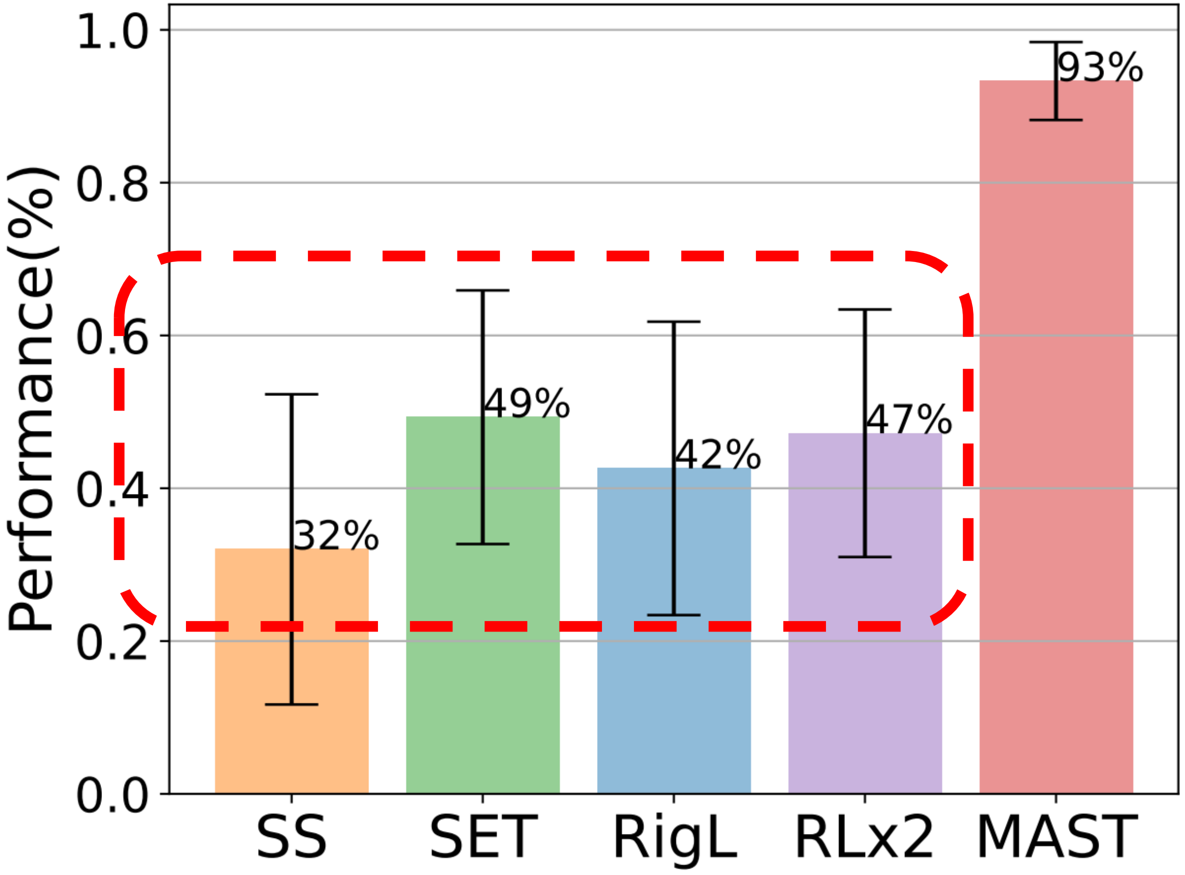
OpenAI Five
[Berner et al., 2019]
>1000 GPUs
~180 days

Dynamic sparse training



- Drop connections based on ***magnitudes***
- Explore new connections based on ***gradients*** [Evci et al., 2020]

Comparison of different sparse training methods.

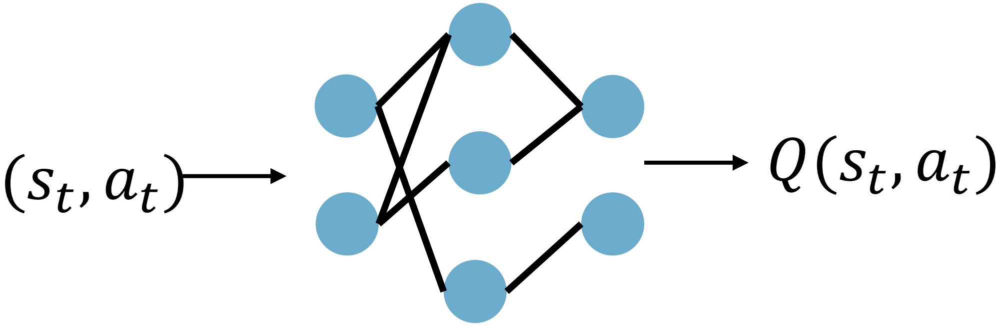


- **SS**: static sparse networks
- **SET** [Mocanu et al., 2018]
  - ➢ Existing DST method 1
- **RigL** [Evci et al., 2020]
  - ➢ Existing DST method 2
- **RLx2** [Tan et al., 2023]
  - ➢ Single-Agent DST method
- **MAST**: our proposed method

Can we train deep MARL agents effectively using ultra-sparse networks throughout?

Issue 1: Inaccurate learning target (1)

TD target	$\mathcal{T}Q_1$	$\mathcal{T}Q_2$	$\mathcal{T}Q_3$
Iter 0	0.9	0.0	0.9
Iter 1	1.7	1.7	2.0
Iter 2	2.9	2.4	3.0
⋮		⋮	



Generated by **sparse** value networks.

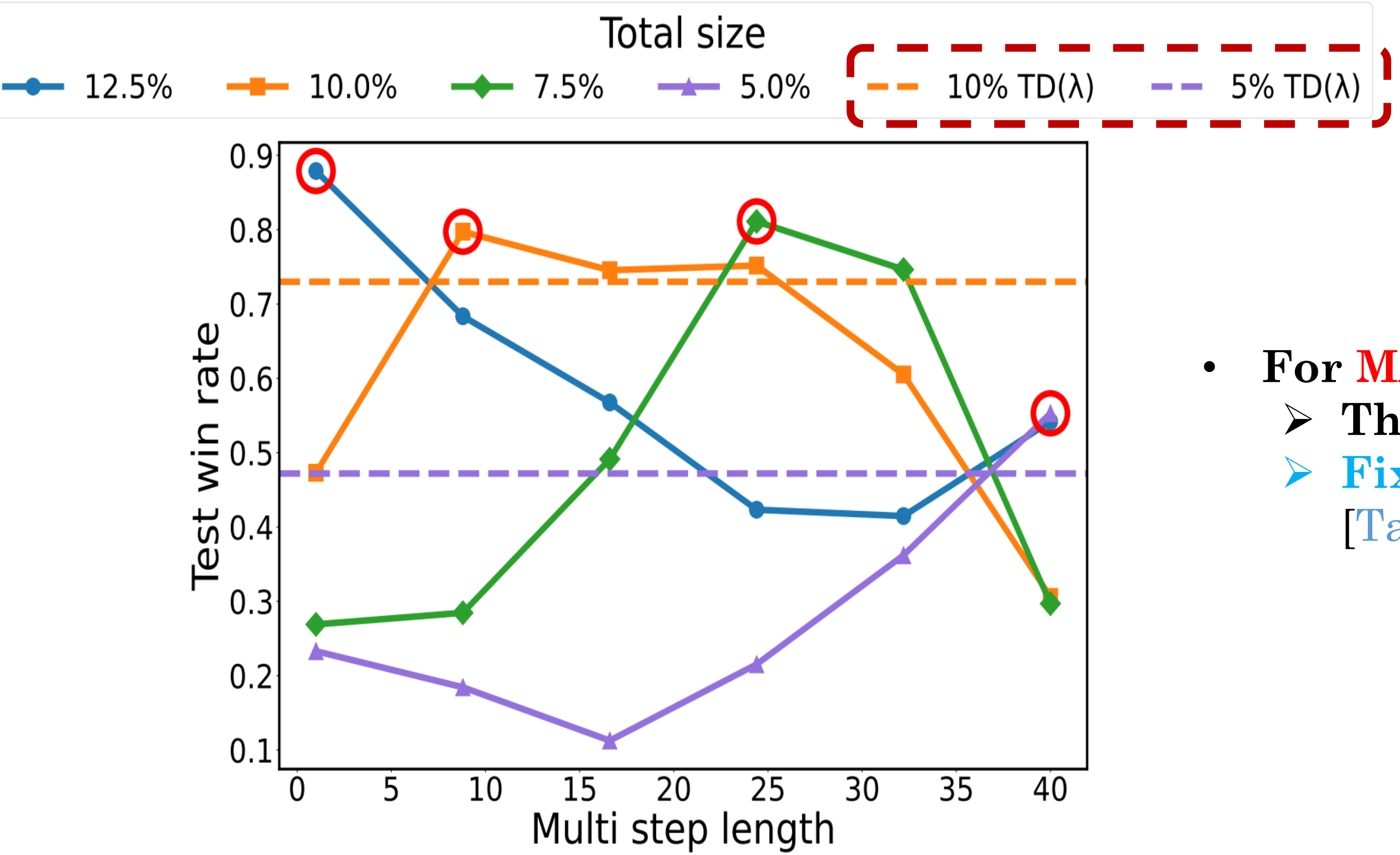
The expected TD error:

$$|\mathbb{E}_\rho[\mathcal{T}_n(s_t, \boldsymbol{u}_t)] - Q^\pi_{tot}(s_t, \boldsymbol{u}_t)| \leq \gamma^n \mathbb{E}_\rho[\underbrace{2\epsilon(s_{t+n}, \rho(s_{t+n})) + \epsilon(s_{t+n}, \pi(s_{t+n}))}_{\textit{Network fitting error}}]$$

$$+ \underbrace{|Q^\rho_{tot}(s_t, \boldsymbol{u}_t) - Q^\pi_{tot}(s_t, \boldsymbol{u}_t)|}_{\textit{Policy inconsistency error}} + \underbrace{\gamma^n \mathbb{E}_\rho[|Q^\pi_{tot}(s_{t+n}, \pi(s_{t+n})) - Q^\rho_{tot}(s_{t+n}, \rho(s_{t+n}))|]}_{\textit{Discounted policy inconsistency error}}$$

Existence of a best step length.

Issue 1: Inaccurate learning target (1)



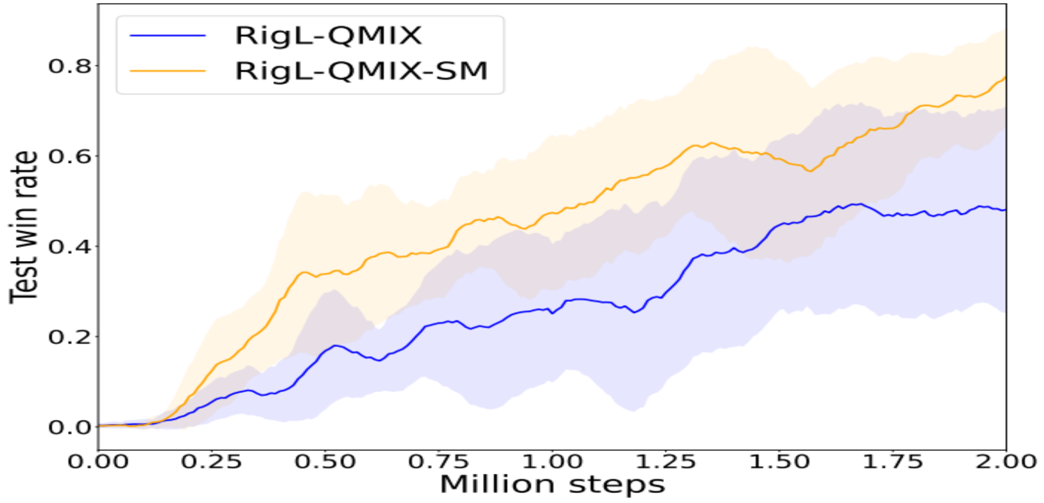
TD(λ) Target

- **For MARL**
  - ➢ **The optimal step length varies**
  - ➢ **Fixed-length multi-step target** [Tan et al., 2023] **is not feasible.**

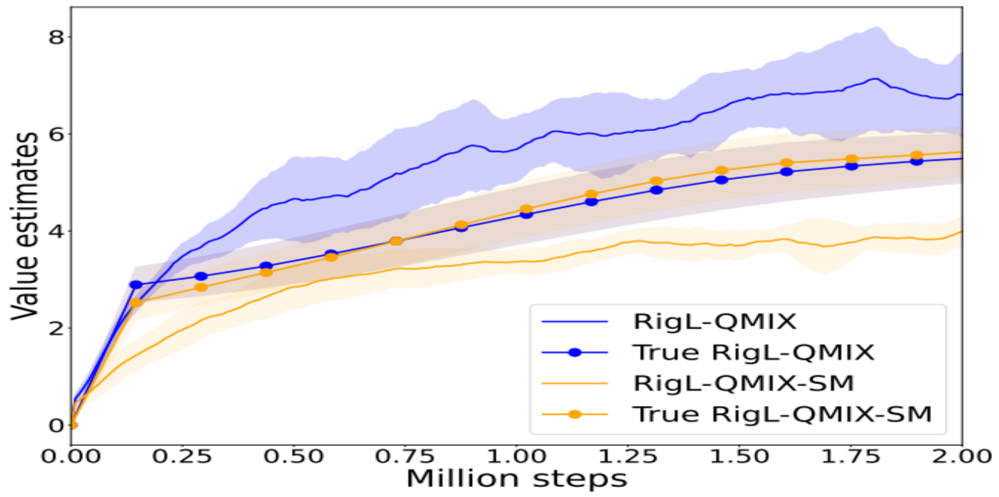
Improving the reliability of training targets.

Issue 1: Inaccurate learning target (2)

Deep **MARL** algorithms, including QMIX [Rashid et al., 2020], also grapple with the **overestimation** problem.



(a) Win rates

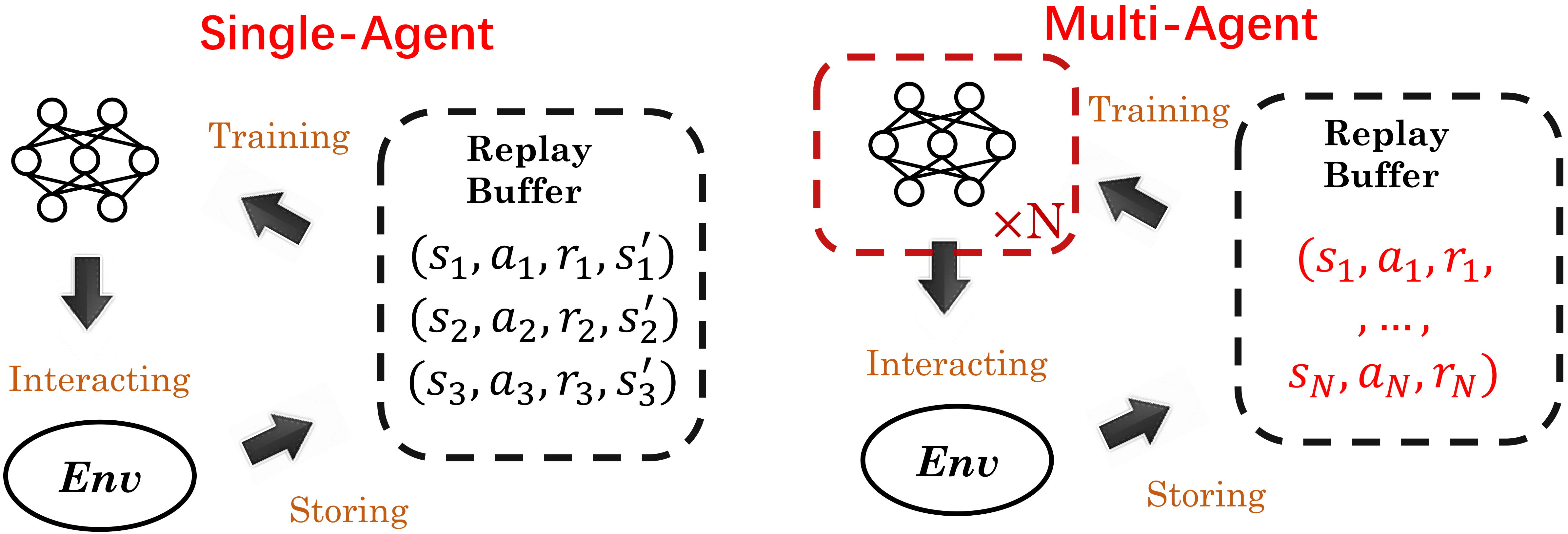


(b) Estimated values

- Soft Mellowmax Operator: $\mathrm{sm}_\omega(Q_i(\tau,\cdot)) = \frac{1}{\omega}\log\left[\sum_{u\in\mathcal{U}} \frac{\exp\left(\alpha Q_i\left(\tau,u\right)\right)}{\sum_{u'\in\mathcal{U}}\exp\left(\alpha Q_i\left(\tau,u'\right)\right)} \exp\left(\omega Q_i\left(\tau,u\right)\right)\right]$

Reducing overestimation of training targets.

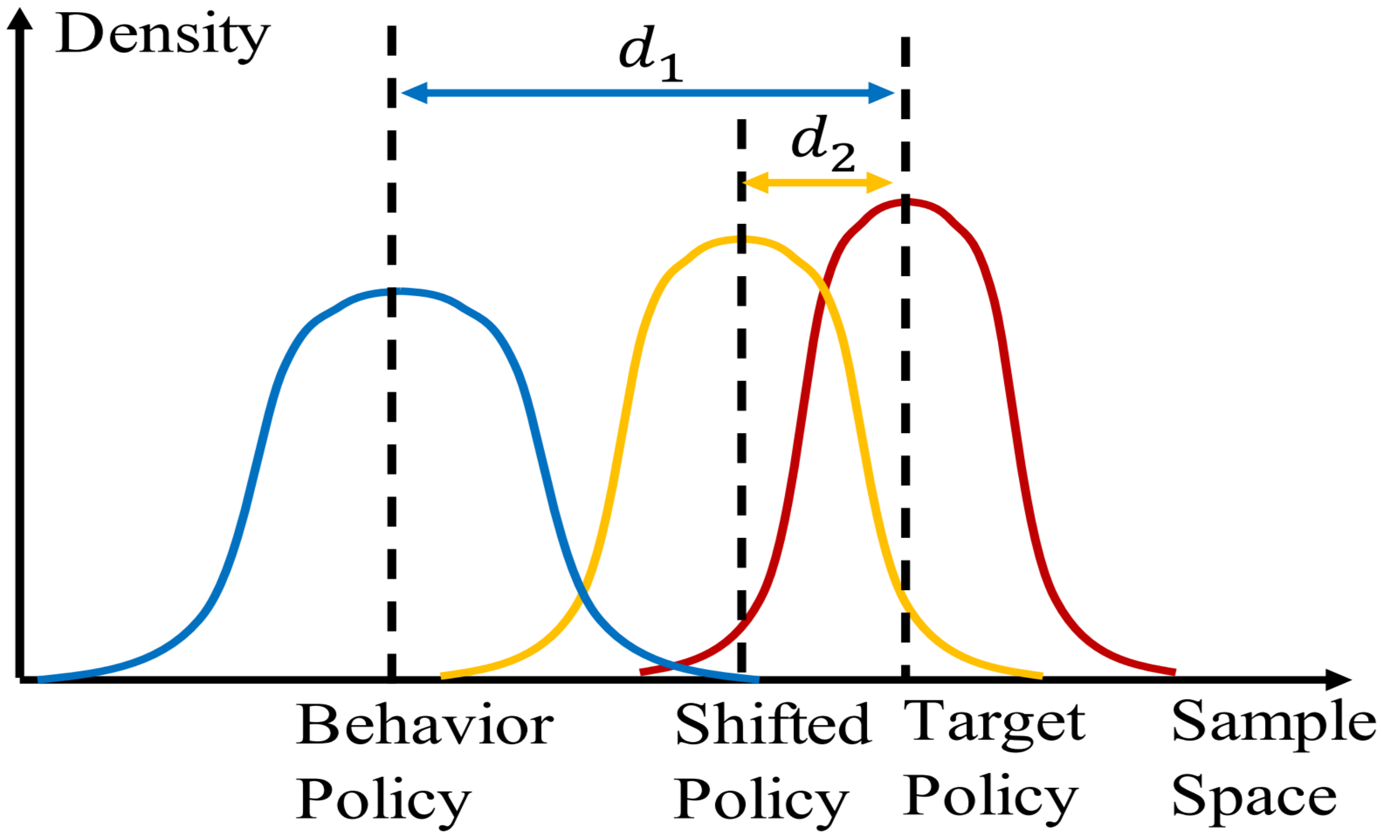
Issue 2: Unstationary data distribution



Transition-level buffer tricks [Banerjee et al., 2022; Tan et al., 2023] are not feasible in **MARL** settings, as transitions are in episode form.

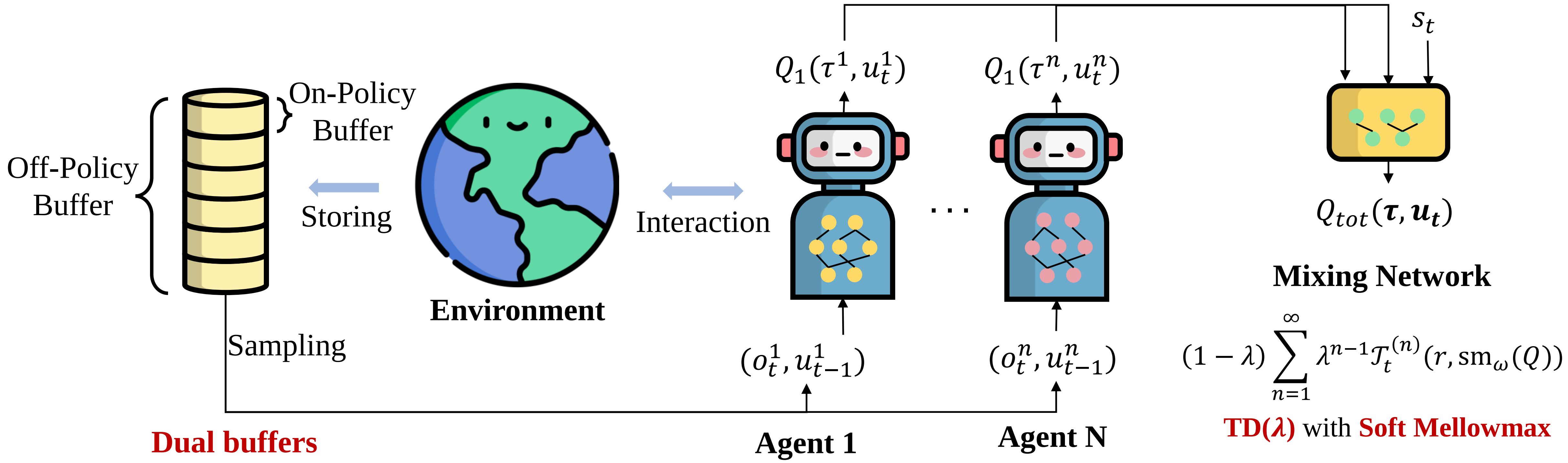
Issue 2: Unstationary data distribution

A **dual buffer mechanism** utilizing two First-in-First-Out (FIFO) replay buffe



Improving the rationality of sample distribution.

MAST: Multi-Agent Sparse Training



Empirical Results

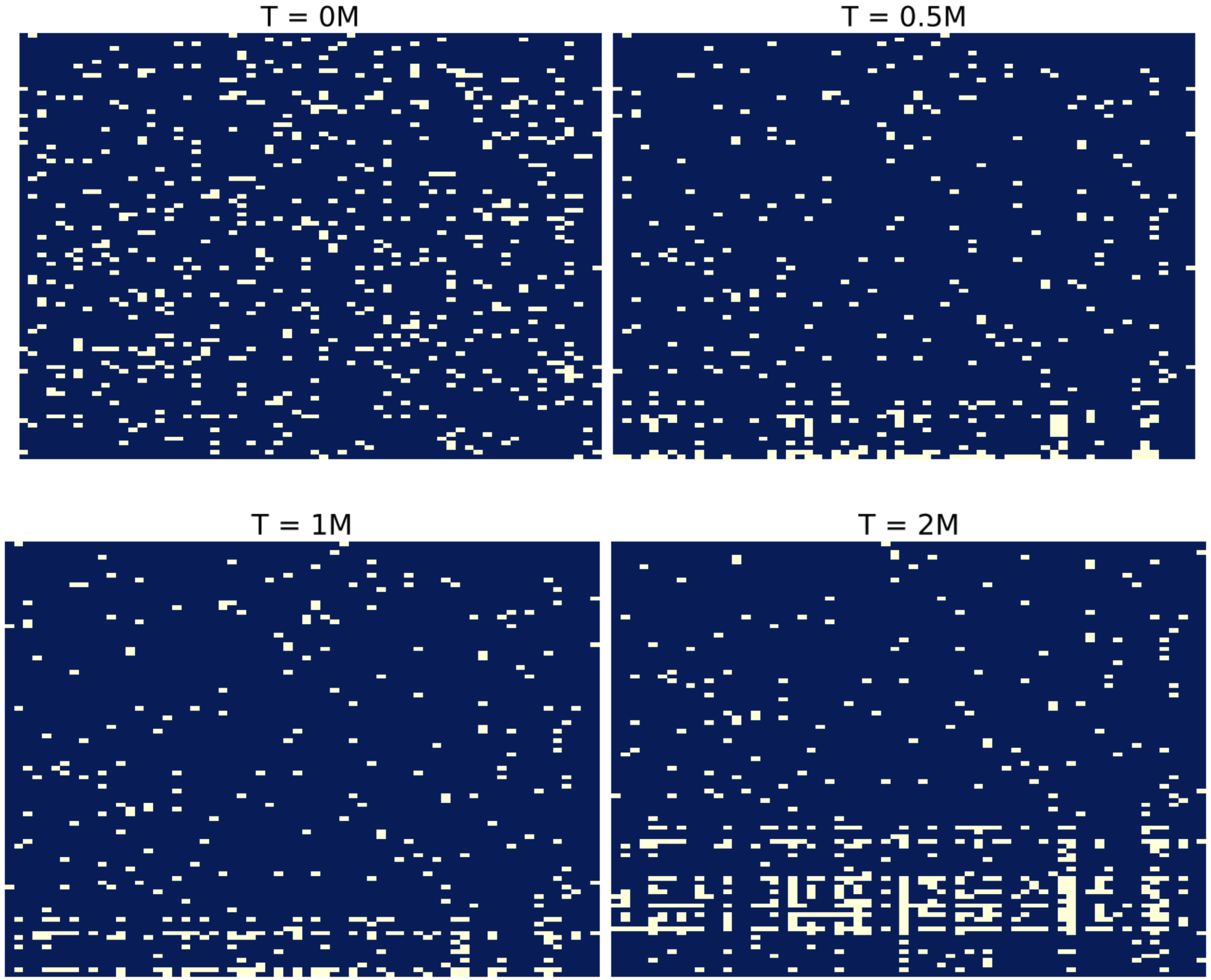
Alg.	Env.	Sp.	Total Size	FLOPs (Train)	FLOPs (Test)	Tiny (%)	SS (%)	SET (%)	RigL (%)	RLx2 (%)	MAST (%)
Q-MIX	3m	95%	0.066x	0.051x	0.050x	98.3	91.6	96.0	95.3	12.1	**100.9**
	2s3z	95%	0.062x	0.051x	0.050x	83.7	73.0	77.6	69.4	45.8	**98.0**
	3s5z	90%	0.109x	0.101x	0.100x	68.2	34.0	52.3	45.2	50.1	**99.0**
	64*	90%	0.106x	0.100x	0.100x	58.2	40.2	67.1	48.7	9.9	**97.6**
	Avg.	92%	0.086x	0.076x	0.075x	77.1	59.7	73.2	64.6	29.8	**98.9**
WQ-MIX	3m	90%	0.108x	0.100x	0.100x	98.3	96.9	97.8	97.8	98.0	**98.6**
	2s3z	90%	0.106x	0.100x	0.100x	89.6	75.4	85.9	86.8	87.3	**100.2**
	3s5z	90%	0.105x	0.100x	0.100x	70.7	62.5	56.0	50.4	60.7	**96.1**
	64*	90%	0.104x	0.100x	0.100x	51.0	29.6	44.1	41.0	52.8	**98.4**
	Avg.	90%	0.106x	0.100x	0.100x	77.4	66.1	70.9	69.0	74.7	**98.1**
RES	3m	95%	0.066x	0.055x	0.050x	97.8	95.6	97.3	91.1	97.9	**99.8**
	2s3z	90%	0.111x	0.104x	0.100x	96.5	92.8	92.8	94.7	94.0	**98.4**
	3s5z	85%	0.158x	0.154x	0.150x	95.1	89.0	90.3	92.8	86.2	**99.4**
	64*	85%	0.155x	0.151x	0.150x	83.3	39.1	44.1	35.3	72.7	**104.9**
	Avg.	89%	0.122x	0.116x	0.112x	93.2	79.1	81.1	78.5	87.7	**100.6**

- Up to 20x FLOPs reduction for both training and inference
- Less than 3% performance degradation

Empirical Results

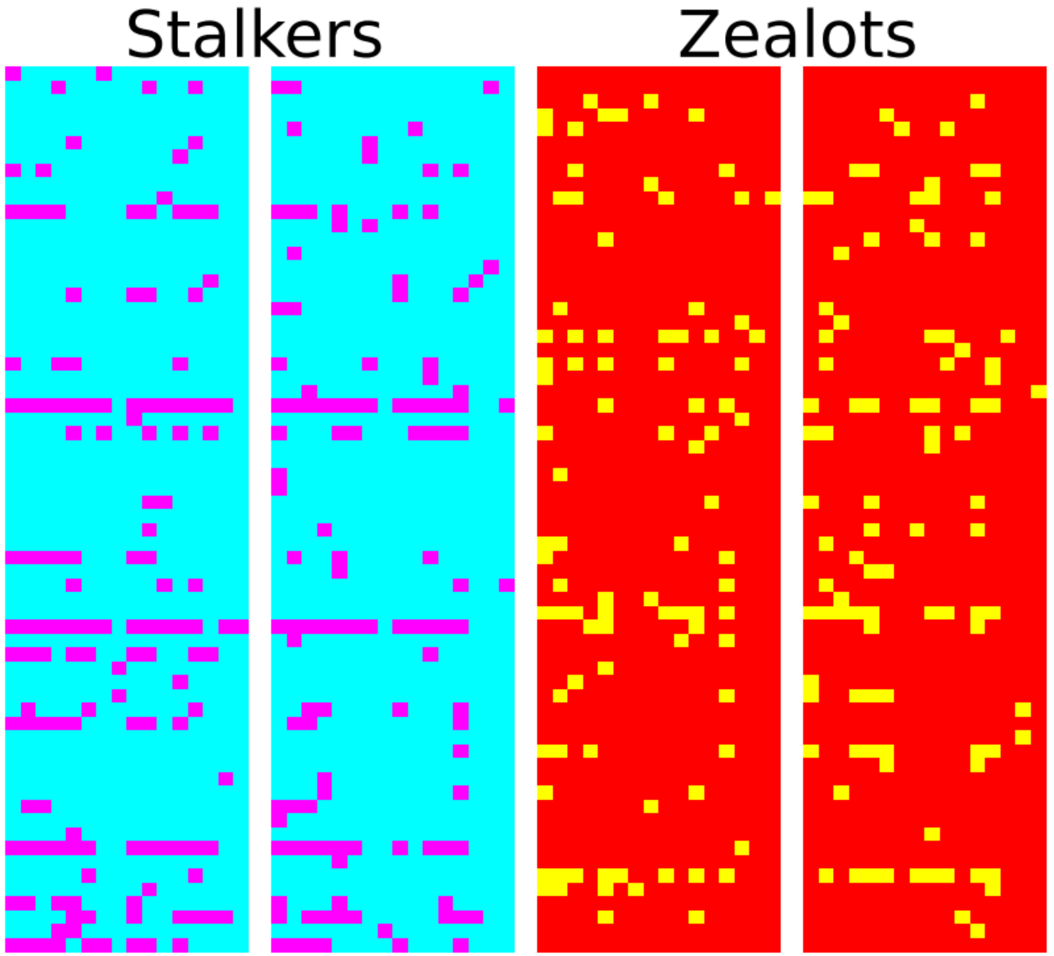
SMAC Benchmark
[Samvelyan et al., 2019]

Agent mask visualization



Empirical Results

StarCraft II
2s3z Map



Agent mask visualization

Conclusion

MAST

- Hybrid TD(λ) Target
- Soft Mellowmax Operator
- Dual Buffers

Sparse Training for Deep MARL Models

- Enhanced Training Stability
- Precise Value Estimation

Experiment Results

- 20x FLOPs reduction
- 3% Performance Loss

References

[Banerjee et al., 2022] Chayan Banerjee, Zhiyong Chen, and Nasimul Noman. Improved soft actor-critic: Mixing prioritized off-policy samples with on-policy experiences. IEEE Transactions on Neural Networks and Learning Systems, 2022.

[Berner et al., 2019] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. arXiv preprint arXiv:1912.06680, 2019.

[Chen et al., 2020] Yu-Jia Chen, Deng-Kai Chang, and Cheng Zhang. Autonomous tracking using a swarm of uavs: A constrained multi-agent reinforcement learning approach. IEEE Transactions on Vehicular Technology, 69(11):13702–13717, 2020b.

[Evci et al., 2020] Utku Evci, Trevor Gale, Jacob Menick, Pablo Samuel Castro, and Erich Elsen. Rigging the lottery: Making all tickets winners. In International Conference on Machine Learning, pages 2943–2952. PMLR, 2020.

[Mathieu et al., 2021] Michael Mathieu, Sherjil Ozair, Srivatsan Srinivasan, Caglar Gulcehre, Shangtong Zhang, Ray Jiang, Tom Le Paine, Konrad Zolna, Richard Powell, Julian Schrittwieser, et al. Starcraft ii unplugged: Large scale offline reinforcement learning. In Deep RL Workshop NeurIPS 2021, 2021.

[Mocanu et al., 2018] Decebal Constantin Mocanu, Elena Mocanu, Peter Stone, Phuong H Nguyen, Madeleine Gibescu, and Antonio Liotta. Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. Nature communications, 9(1):2383, 2018.

[Rashid et al., 2020] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. The Journal of Machine Learning Research, 21(1):7234–7284, 2020.

[Samvelyan et al., 2019] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar,