Individual Reward Assisted Multi-Agent Reinforcement Learning

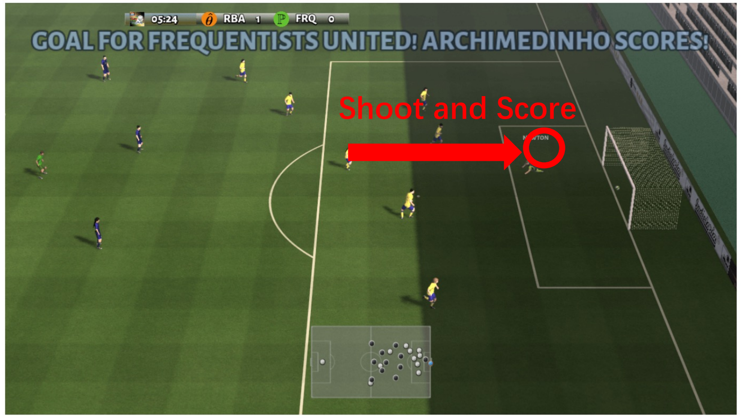
Li Wang*[1,2], Yupeng Zhang*[1,2], Yujing Hu[2], Weixun Wang[3,4], Chongjie Zhang[5], Yang Gao[1], Jianye Hao[3,4], Tangjie Lv[2], Changjie Fan[2]

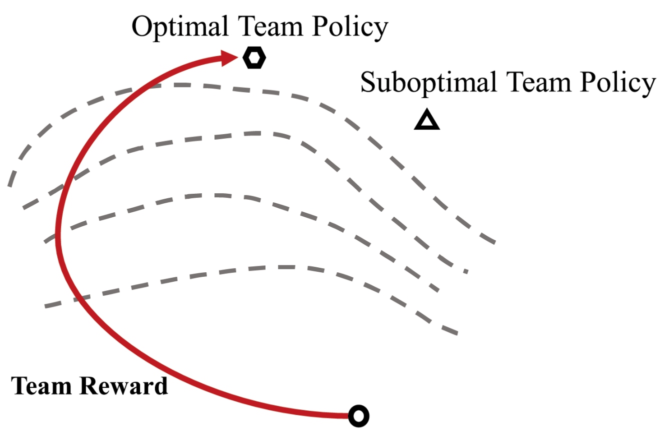
[1]Nanjing University, [2]NetEase Fuxi AI Lab, [3]Tianjin University, [4]Noah's Ark Lab, [5]Tsinghua University

July,2022

ICML
International Conference
On Machine Learning

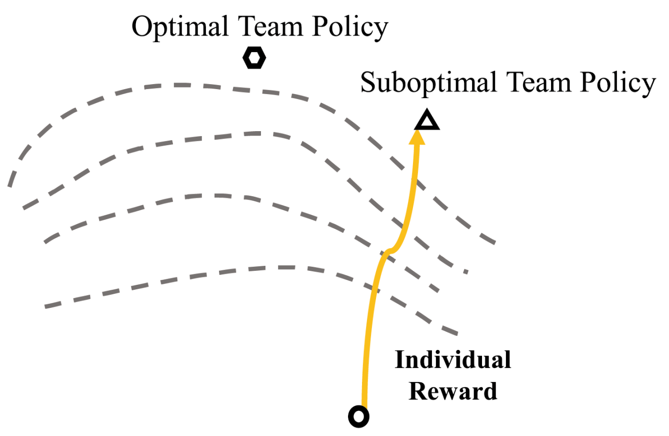
Background





Sparse team rewards are given for team goals, e.g. shot and scoring in football game.

Team rewards are **too sparse to guide an effective** cooperative policy.



Dense individual rewards are designed to assist the learning of team goals, e.g. pass in football game.

Dense individual rewards usually can lead to a **sub-optimal** cooperative policy.

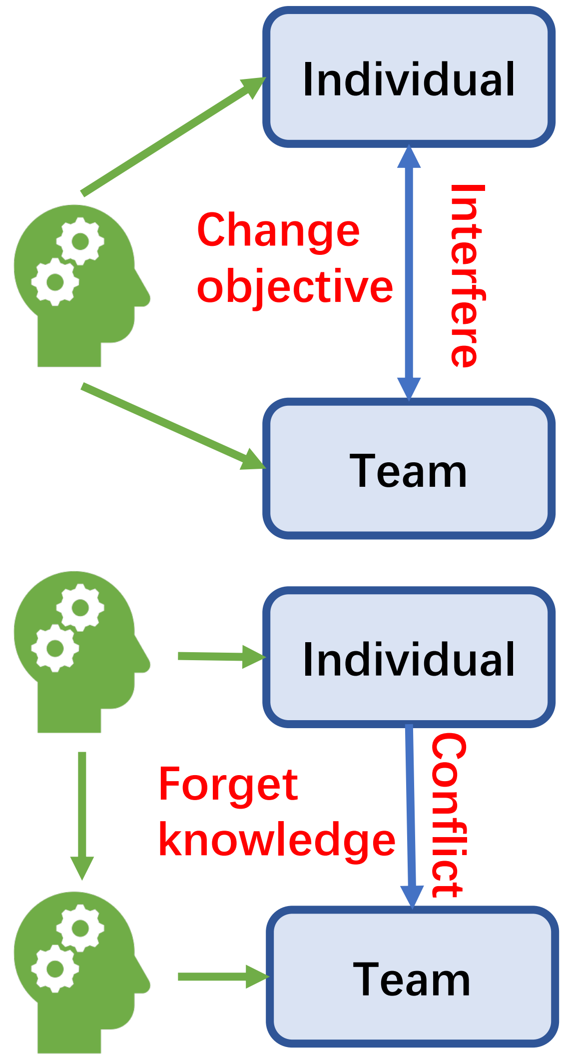
Background

Reward Shaping[Andrew Y, et al. ICML, 1999]
- sum individual rewards with team rewards as final rewards.

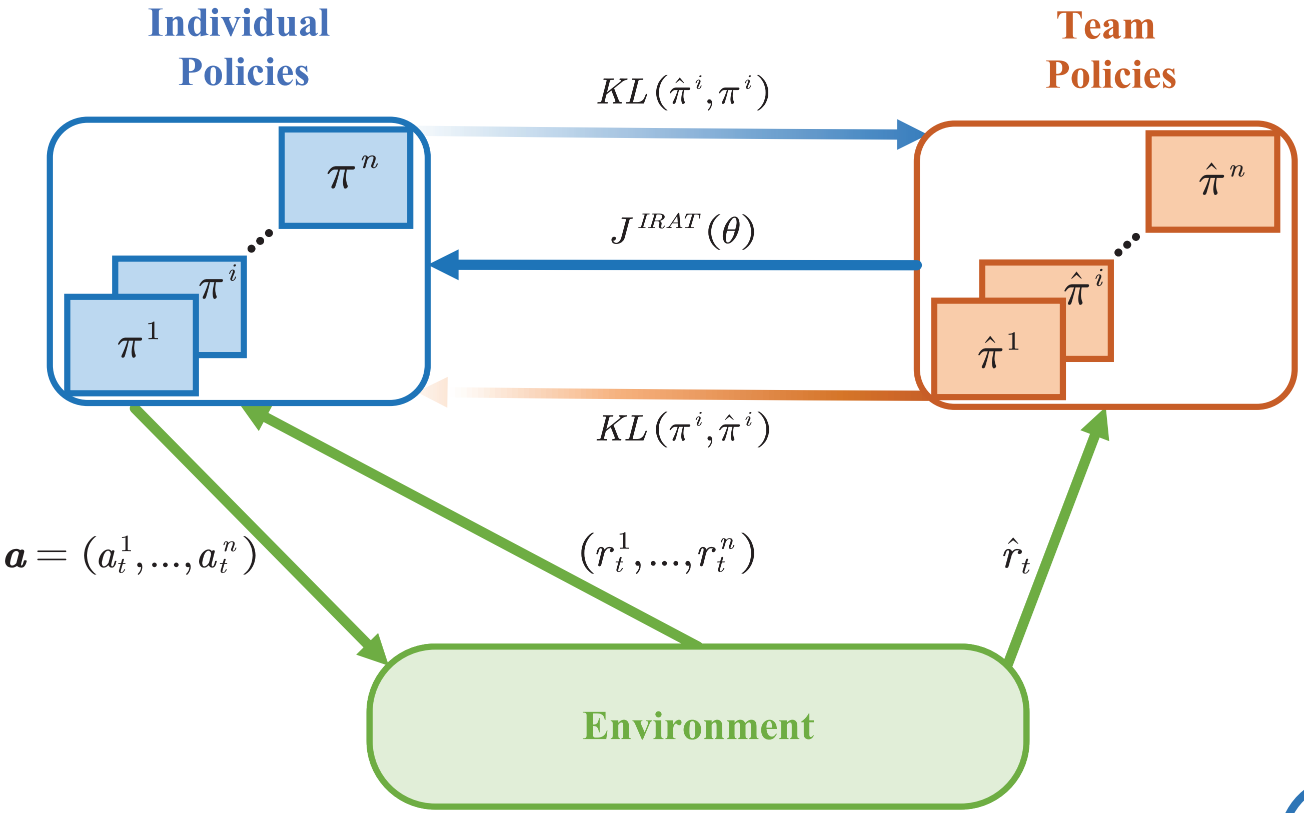
Multi-Critic[Ye D, et al. NIPS, 2020]
- maintain two critics for individual rewards and team rewards and update the policy according to the integration of them.

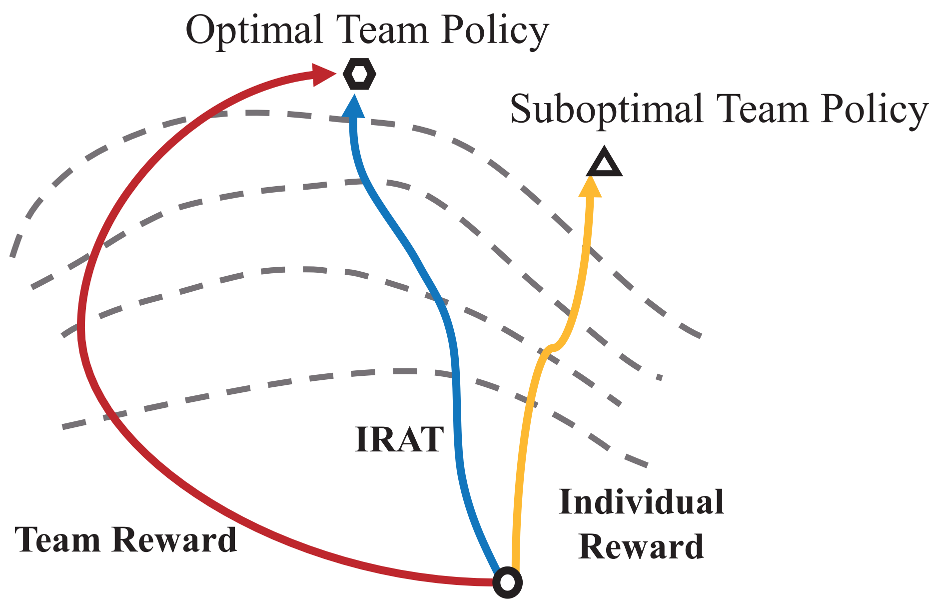
Multi-task Learning[Yu T, et al. NIPS, 2020]
- learn individual rewards and team rewards as two tasks.

Transfer Learning [Liu Y, et al. IJCAI, 2019]
- pre-train the policies with the individual rewards and then fine-tuned with team rewards.



Method – Individual Reward Assisted Team Policy Learning (IRAT)





Exploration and sample generation for the learning of the team policy

The key idea:

- learn an **individual policy** from individual reward and a **team policy** from team reward
- put **discrepancy constraints** on the two policies

IRAT learns two policies from two rewards simultaneously but respectively:

- **utilize the knowledge of individual rewards** to assist the learning of team policy
- **avoid the interference** when using one policy to learn two reward objectives

Method – Individual Policy Learning

- The individual policy need to adjust its sampling behavior based on the current learning of the team policy for producing samples with higher team reward:
  - When two policies are **consistent**, the individual policy should learn **quickly.**
  - When two policies **conflict** too much, the individual policy should update **carefully.**

- Similarity between π_θ and $\hat{\pi}_{\hat{\theta}}$ is defined as:

$$\sigma_t^i(\theta^i) = \frac{\pi_{\theta^i}(a_t^i \mid \tau_t^i)}{\hat{\pi}_{\hat{\theta}^i}(a_t^i \mid \tau_t^i)}$$

- A new cooperation-oriented objective is:

$$J^{IRAT}(\theta^i) = \mathbb{E}\left[\mathrm{clip}\left(\sigma_t^i(\theta^i), 1-\xi, 1+\xi\right)A_t^i\right]$$

- Combine $J^{IRAT}(\theta^i)$ with its original optimization objective $J^{CLIP}(\theta^i)$:

$$J^{TC}(\theta^i) = \mathbb{E}\left[\mathbb{I}_{\sigma_t^i \le 1}\max\left(J^{CLIP}(\theta^i), J^{IRAT}(\theta^i)\right) + \mathbb{I}_{\sigma_t^i > 1}\min\left(J^{CLIP}(\theta^i), J^{IRAT}(\theta^i)\right)\right]$$

- An increasing-effect **KL regularizer** is introduced to distill team policy knowledge:

$$J(\theta^i) = \mathbb{E}\left[J^{TC}(\theta^i) - \alpha KL(\hat{\pi}^i, \pi^i)\right]$$

Method – Team Policy Learning

- Team policy uses learning objective corrected by importance sampling:

$$\hat{\sigma}_t^i(\hat{\theta}^i) = \frac{\hat{\pi}_{\hat{\theta}^i}(a_t^i \mid \tau_t^i)}{\pi_{\theta_{old}^i}(a_t^i \mid \tau_t^i)}$$

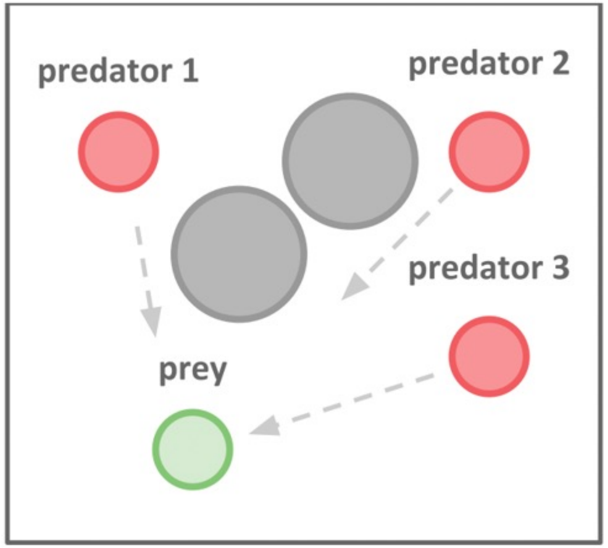
- A decreasing-effect KL regularizer to ensure effective update.

- The total learning objective of team policy is:

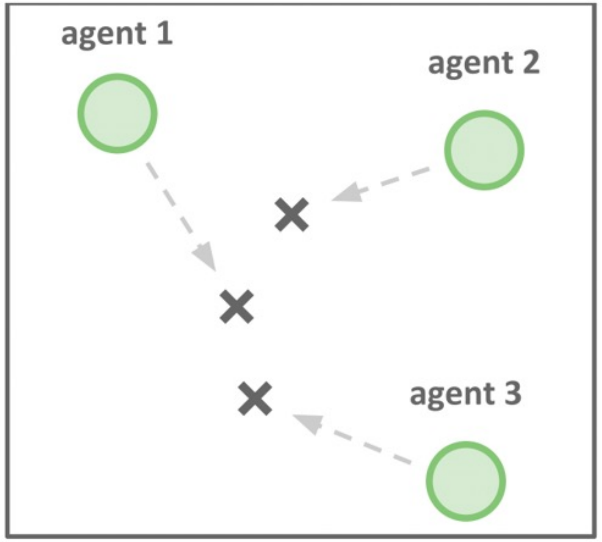
$$\hat{J}(\hat{\theta}^i) = \mathbb{E}\left[\min\left(\hat{\sigma}_t^i(\hat{\theta}^i)\hat{A}_t, \mathrm{clip}\left(\hat{\sigma}_t^i(\hat{\theta}^i), 1-\zeta, 1+\zeta\right)\hat{A}_t\right) - \beta KL\left(\pi^i, \hat{\pi}^i\right)\right]$$

Where β is a decreasing coefficient.

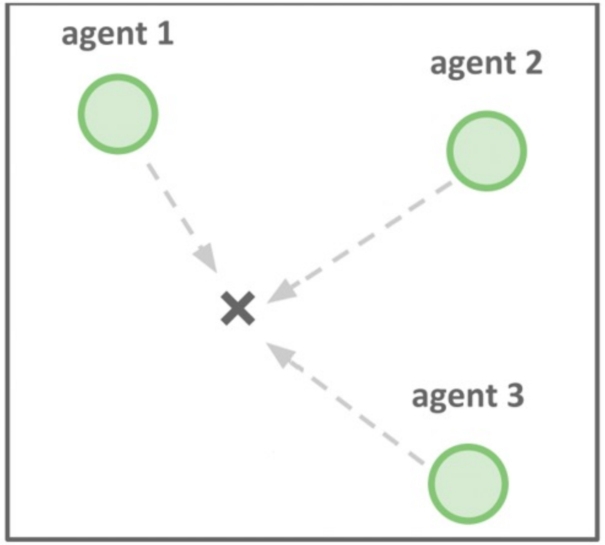
Experiments



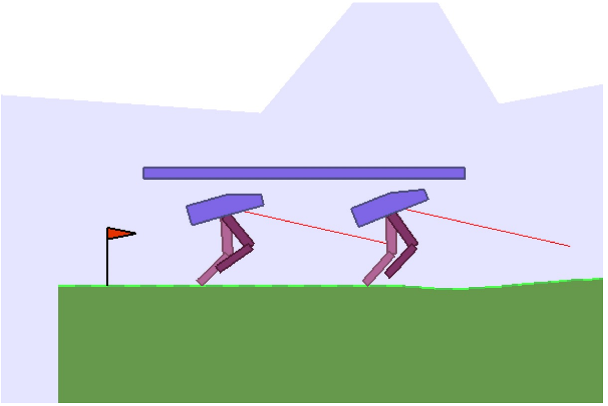
Predator-Prey



Spread



Attack



Multi-Walker

Multi-Agent Particle Environment
[Lowe R, et al. NIPS, 2017]

Team reward:

- Positive num when archive team goal
- 0 in other cases.

Individual reward ：

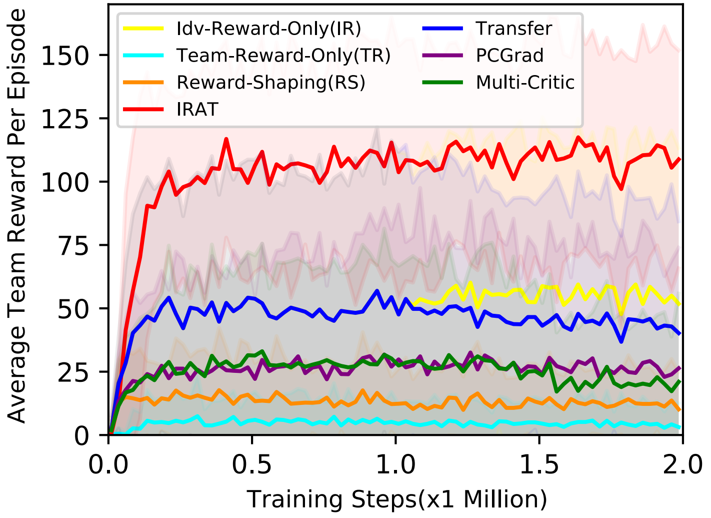
- Predator-Prey:Useful Individual Reward
- Spread: Misleading Individual Reward
- Attack: Conflicting Individual Reward

Multi-Walker [Gupta, J. K, et al. AAMAS, 2017]

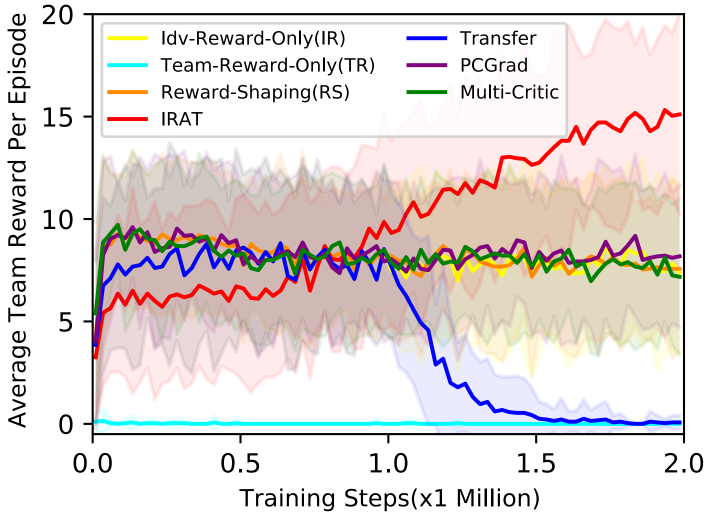
Team reward:

- Not sparse but hard to learn

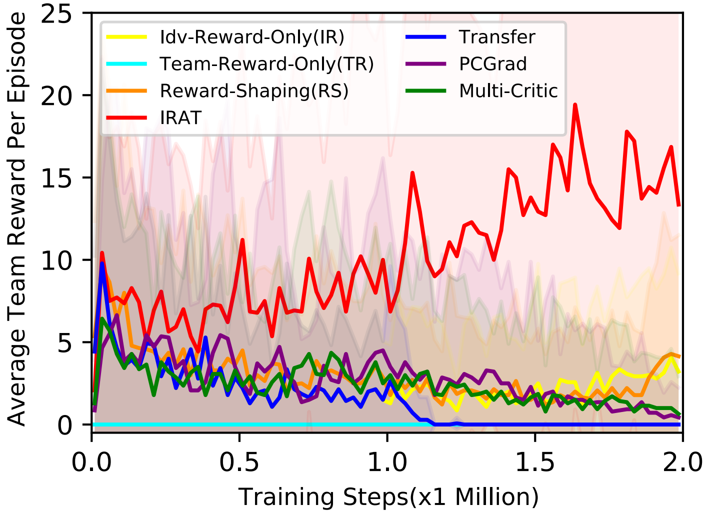
Experiments



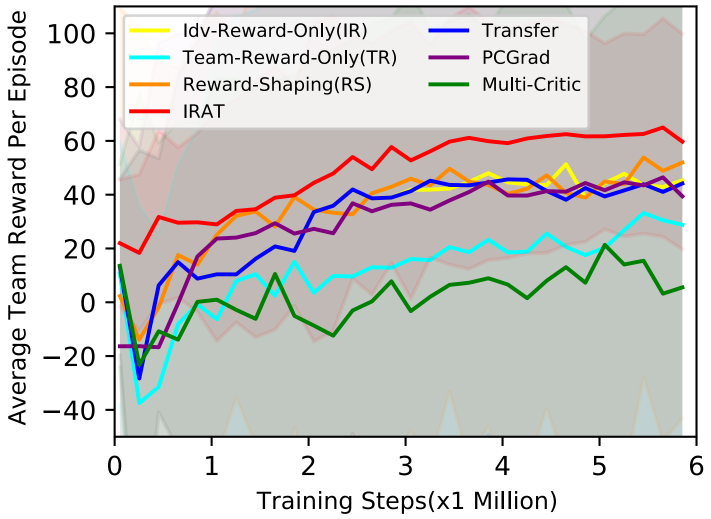
Predator-Prey: Useful Individual Reward



Spread: Misleading Individual Reward



Attack: Conflicting Individual Reward



Multiwalker: Not sparse Team Reward

IRAT outperforms other methods, even when the individual rewards sometimes mislead or conflict with the team rewards.

Experiments

Google Research Football (GRF)

[Kurach K, et al. AAAI, 2020]

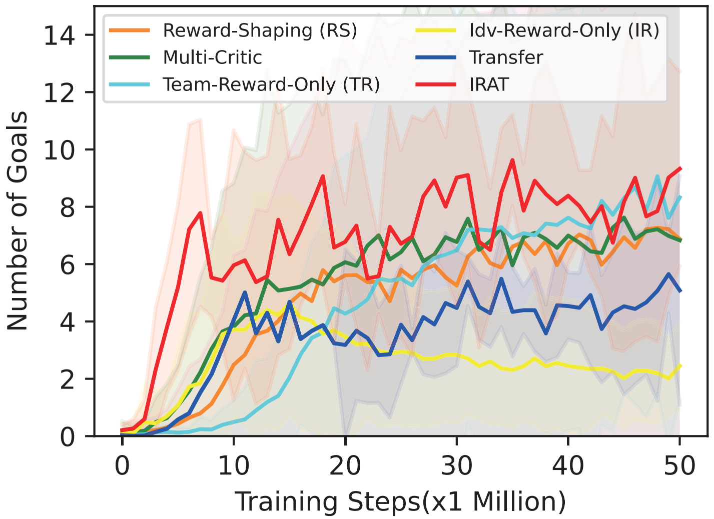
5-vs-5 half-court offense

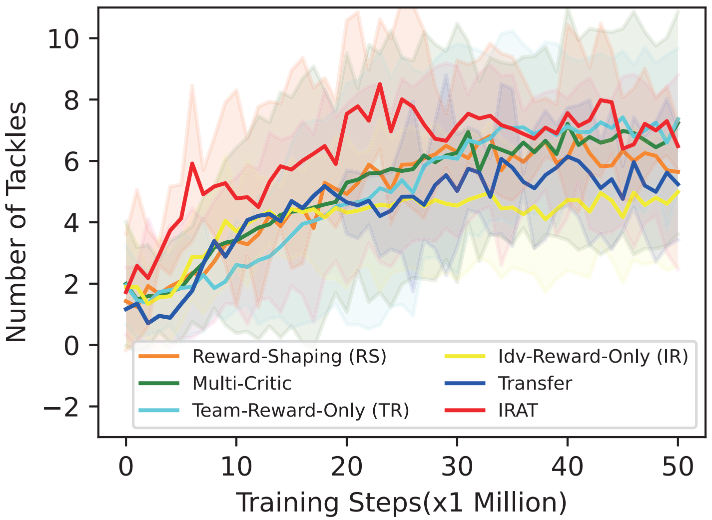
Team reward

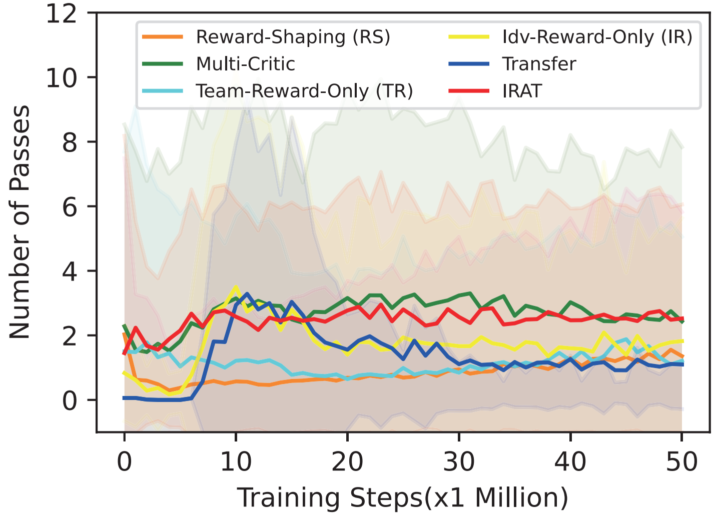
- 1 for team scores a goal
- 0 in other cases.

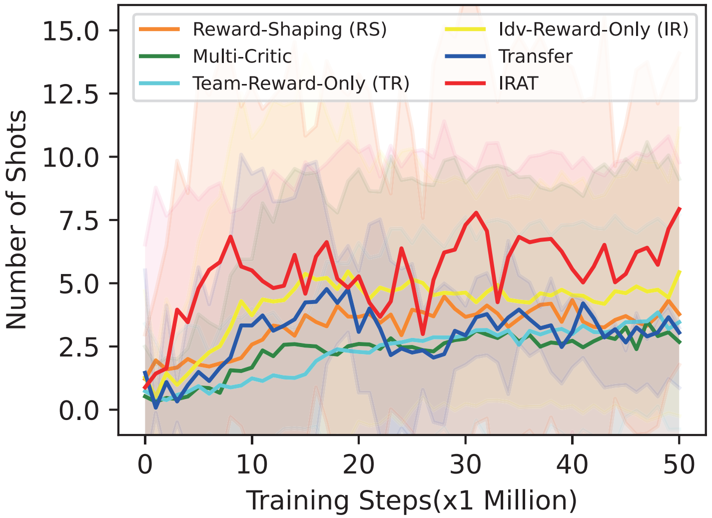
Individual rewards

- position rewards,
- shooting rewards,
- ball-passing rewards,
- ball-possession rewards.









IRAT significantly outperforms the other methods with higher goal scores and much faster convergence.

Reference

1. Andrew Y. Ng, Daishi Harada, Stuart Russell: Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. ICML 1999: 278-287.
2. Ye D, Chen G, Zhang W, et al. Towards playing full moba games with deep reinforcement learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 621-632.
3. Yu T, Kumars S, Gupta A, et al. Gradient surgery for multi-task learning[J].Advances in Neural Information Processing Systems, 2020, 33: 5824-5836.
4. Liu Y, Hu Y, Gao Y, et al. Value function transfer for deep multi-agent reinforcement learning based on N-step returns[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. 2019: 457-463.
5. Lowe R, Wu Y I, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[J]. Advances in neural information processing systems, 2017, 30.
6. Gupta, J. K., Egorov, M., and Kochenderfer, M. J. Cooperative multi-agent control using deep reinforcement learning. In *Autonomous Agents and Multiagent Systems*, pp. 66–83. Springer, 2017.
7. Kurach K, Raichuk A, et al. Google research football: A novel reinforcement learning environment[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 34: 04. 2020: 4501-4510.