ICML
International Conference
On Machine Learning

清華大學 交叉信息研究院
Institute for Interdisciplinary Information Sciences, Tsinghua University

上海
期智研究院
SHANGHAI QI ZHI INSTITUTE

清華大學电子工程系
Department of Electronic Engineering, Tsinghua University

Berkeley
UNIVERSITY OF CALIFORNIA

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

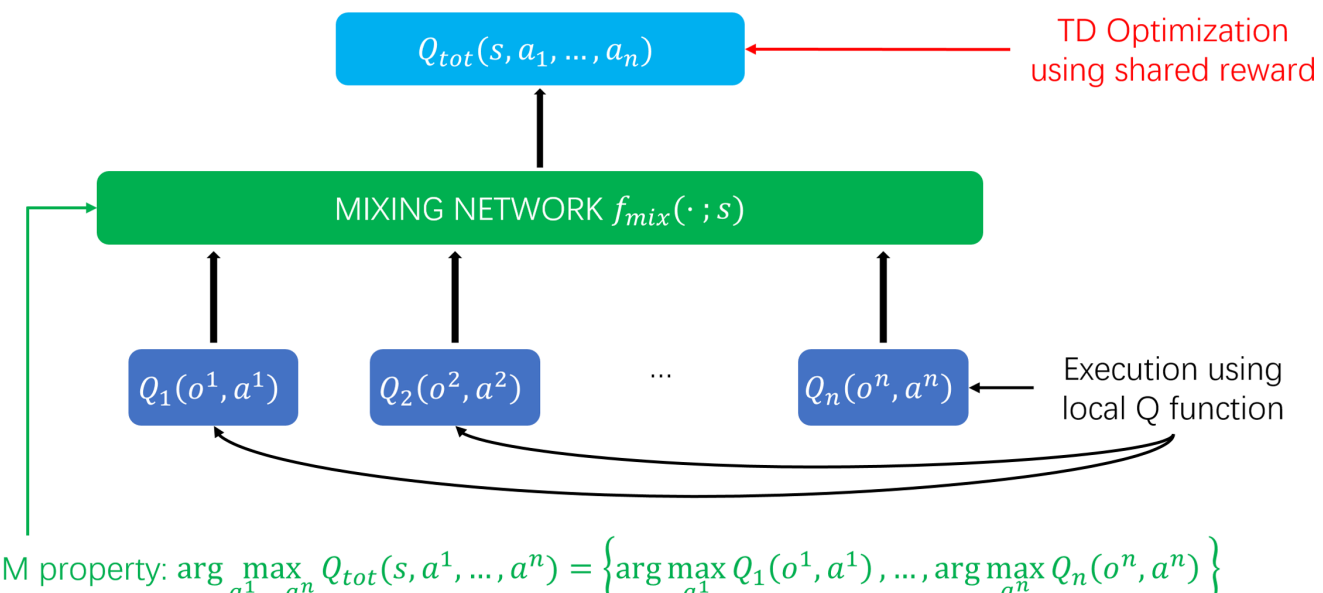
Revisiting Some Common Practices in Cooperative Multi-Agent Reinforcement Learning

Wei Fu, Chao Yu, Zelai Xu, Jiaqi Yang, Yi Wu

Correspondence to: Wei Fu <fuwth17@gmail.com>, Yi Wu <jxwuyi@gmail.com>.

Background

- Paradigms for tackling cooperative MARL
  - Value Decomposition (VD)
    - The most popular framework



Background

- Paradigms for tackling cooperative MARL
  - Value Decomposition (VD)
    - The most popular framework
    - VDN, QMIX, QPLEX, etc

Background

- Paradigms for tackling cooperative MARL
  - Value Decomposition (VD)
    - The most popular framework
    - VDN, QMIX, QPLEX, etc
  - Policy Gradient (PG)
    - COMA
    - Less popular due to the on-policy fashion

Background

- Paradigms for tackling cooperative MARL
  - Value Decomposition (VD)
    - The most popular framework
    - VDN, QMIX, QPLEX, etc
  - Policy Gradient (PG)
    - COMA
    - Less popular due to the on-policy fashion
    - Empirically, PPO can achieve strong performance

Background

- Paradigms for tackling cooperative MARL
  - Value Decomposition (VD)
    - The most popular framework
    - VDN, QMIX, QPLEX, etc
  - Policy Gradient (PG)
    - COMA
    - Less popular due to the on-policy fashion
    - Empirically, PPO can achieve strong performance ?

Background

- Paradigms for tackling cooperative MARL
  - Value Decomposition (VD)
    - The most popular framework
    - VDN, QMIX, QPLEX, etc
  - Policy Gradient (PG)
    - COMA
    - Less popular due to the on-policy fashion
    - Empirically, PPO can achieve strong performance ?
  - Common practice: policy sharing among agents

Overview

ICML
International Conference
On Machine Learning

- We attempt to partially explain PPO’s effectiveness

Overview

- We attempt to partially explain PPO's effectiveness
- In environments with a multi-modal reward function

Overview

- We attempt to partially explain PPO's effectiveness

- In environments with a multi-modal reward function
  - VD and policy sharing can **fundamentally fail**

Overview

- We attempt to partially explain PPO's effectiveness

- In environments with a multi-modal reward function
  - VD and policy sharing can **fundamentally fail**
  - agent-specific PG can **provably converge to one mode**

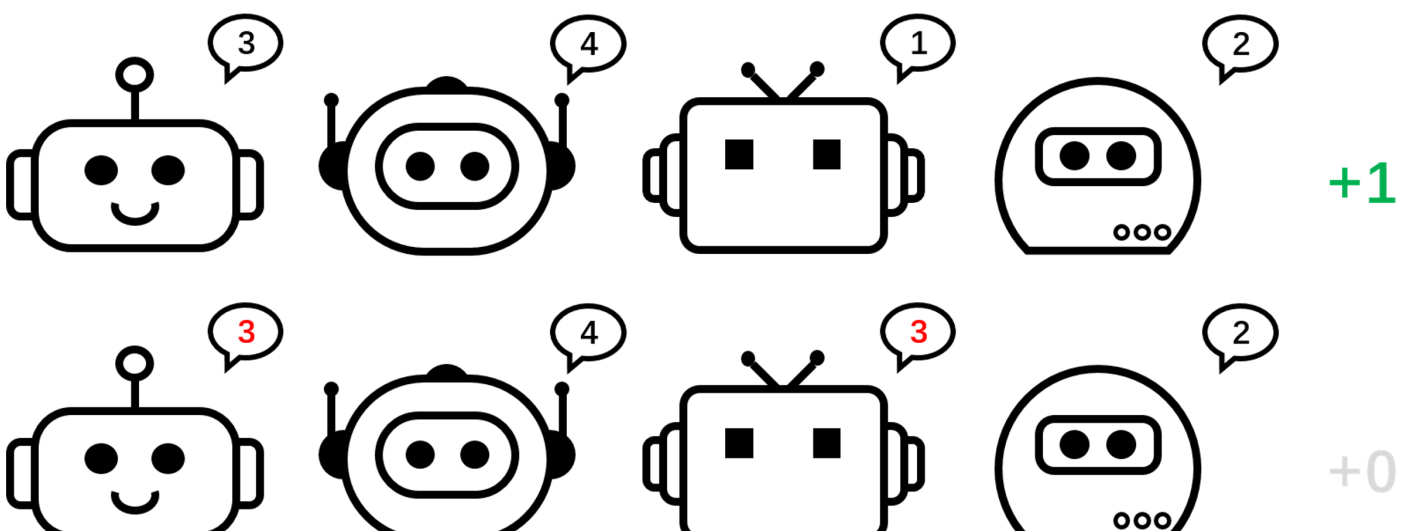
Overview

- We attempt to partially explain PPO's effectiveness

- In environments with a multi-modal reward function
  - VD and policy sharing can **fundamentally fail**
  - agent-specific PG can **provably converge to one mode**
  - **an auto-regressive policy** can cover *all* modes

Permutation Game

- N-player permutation game
  - Action space {1,…, N}
  - Reward=1 if and only if the joint action is a permutation over N



Permutation Game

Value decomposition (VD) methods can fail

- **Theorem 4.1.** *Value decomposition* (VD) *cannot represent the underlying global Q-function in the 2-player Permutation game.*

$$Q_{tot}(a^1, a^2) = f_{mix}\big(Q_1(a^1), Q_2(a^2)\big)$$

Agent 1 \ Agent 2		
	0	1
1	1	0

Permutation Game

Value decomposition (VD) methods can fail

- **Theorem 4.1.** *Value decomposition* (VD) *cannot represent the underlying global Q-function in the 2-player Permutation game.*

$$Q_{tot}(a^1, a^2) = f_{mix}\big(Q_1(a^1), Q_2(a^2)\big)$$

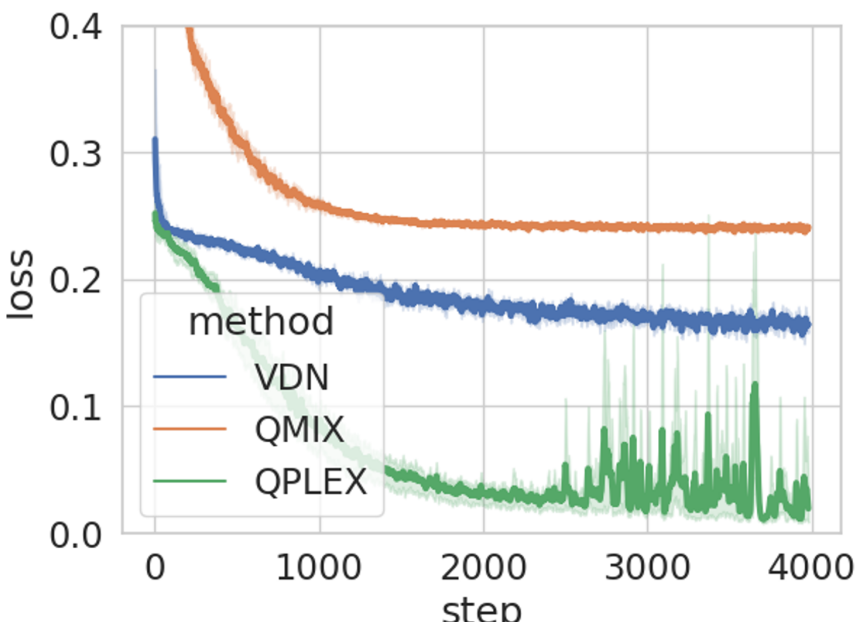
Agent 1 \ Agent 2		
	0	1
1	0	

- Intuition
  - to satisfy the Individual Global Max (IGM) property
  - Q_{tot} cannot represent two symmetric and equally optimal modes

Permutation Game

Value decomposition (VD) methods can fail

- **Theorem 4.1.** *Value decomposition* (VD) *cannot represent the underlying global Q-function in the 2-player Permutation game.*



Permutation Game

Agent-specific policy gradient (PG) methods provably converges

- **Theorem 4.2.** *Shared policy learning* (PG-sh) *cannot learn an optimal policy for the 2-player Permutation game.*

Permutation Game

Agent-specific policy gradient (PG) methods provably converges

- **Theorem 4.2.** *Shared policy learning* (PG-sh) *cannot learn an optimal policy for the 2-player Permutation game.*

- **Theorem 4.3 & 4.4.** *Agent-specific policy learning via stochastic policy gradient can represent and learn an optimal policy in the 2-player Permutation game.*

Permutation Game

An auto-regressive policy can cover all modes

- Agent-specific policy learning converges to a random mode

Can we learn a single policy that cover all the optimal modes?

Permutation Game

An auto-regressive policy can cover all modes

- Agent-specific policy learning converges to a random mode

  Can we learn a single policy that cover all the optimal modes?

- Factorized representation of agent-specific learning

$$\pi(a^1, \ldots, a^n \mid o^1, \ldots, o^n) \approx \prod_{i=1}^{n} \pi_i(a^i \mid o^i)$$

Limited Expressiveness

Permutation Game

An auto-regressive policy can cover all modes

- Agent-specific policy learning converges to a random mode

  Can we learn a single policy that cover all the optimal modes?

- Factorized representation of agent-specific learning

$$\pi(a^1, \dots, a^n \mid o^1, \dots, o^n) \approx \prod_{i=1}^{n} \pi_i(a^i \mid o^i)$$

Limited Expressiveness

- Auto-regressive (AR) modeling

$$\pi(a^1, \dots, a^n \mid o^1, \dots, o^n) \approx \prod_{i=1}^{n} \pi_i(a^i \mid o^i, a^1, \dots, a^{i-1})$$

Minimal Assumption

Stronger Expressiveness

Permutation Game

An auto-regressive policy can cover all modes

- Agent-specific policy learning converges to a random mode

Can we learn a single policy that cover all the optimal modes?

- Mode coverage in the 4-player permutation game

Individual Auto-Regressive Payoff

Learning for Higher Rewards

StarCraft Multi-Agent Challenge and Google Research Football

ICML
International Conference
On Machine Learning

- Intuitively multi-modal
- SMAC
  - **Agent-specific PG outperforms the SOTA VD method (RODE) on 9/10 maps**
- GRF
  - **Agent-specific PG outperform the SOTA VD method (CDS) in 5/6 scenarios**

Map	PG-ID	PG-sh.	PG-Ind.	RODE
1c3s5z	**100.0(0.0)**	97.4(1.0)	99.1(0.7)	**100.0(0.0)**
2s3z	**100.0(0.7)**	99.0(0.5)	99.1(0.9)	**100.0(0.0)**
3s_vs_5z	**100.0(0.6)**	96.7(1.8)	93.8(1.8)	78.9(4.2)
3s5z	**96.9(0.7)**	95.2(1.5)	80.4(3.3)	93.8(2.0)
3s5z_vs_3s6z	84.4(34.0)	42.3(4.0)	37.8(5.6)	**96.8(25.1)**
5m_vs_6m	**89.1(2.5)**	35.3(2.1)	44.4(2.9)	71.1(9.2)
6h_vs_8z	**88.3(3.7)**	79.9(4.8)	11.4(2.5)	78.1(37.0)
10m_vs_11m	**96.9(4.8)**	86.5(2.3)	78.4(2.7)	95.3(2.2)
corridor	**100.0(1.2)**	92.6(2.4)	82.2(1.8)	65.6(32.1)
MMM2	90.6(2.8)	**92.3(1.9)**	13.0(3.7)	89.8(6.7)

Scenario	PG-ID	PG-Ind.	CDS(QMIX)	CDS(QPLEX)
3v1	**90.7(1.5)**	**90.2(1.7)**	73.7(3.4)	83.6(4.0)
CA(Easy)	79.5(6.7)	**92.4(2.6)**	43.0(5.6)	40.4(4.8)
CA(Hard)	**68.2(2.0)**	**67.8(3.4)**	35.4(2.9)	34.8(3.4)
Corner	**27.3(1.3)**	21.6(2.4)	1.8(0.3)	20.8(1.7)
PS	43.4(8.8)	53.3(3.5)	83.5(4.0)	**86.8(1.9)**
RPS	66.6(3.1)	**78.8(1.5)**	65.5(7.0)	75.1(2.4)

Learning for Multi-Modal Behavior

StarCraft Multi-Agent Challenge and Google Research Football

Marines **keep standing** and perform attacks alternately while ensuring there is **only one attacking marine at each timestep**.

"Tiki-Taka" style behavior: **each player keeps passing the ball to their teammates**.

Conclusion & Takeaways

ICML
International Conference
On Machine Learning

- In environments with a multi-modal reward function
  - VD and policy sharing can lead to **unsatisfactory behaviors**
  - **Agent-specific PG can be preferred!**
    - learning higher rewards
    - learning multi-modal behavior

- Regarding PG implementation in MARL, we suggest …
  - including the **agent-specific information**
  - learning an **AR policy for emergent behavior**

清華大學 交叉信息研究院
Institute for Interdisciplinary Information Sciences, Tsinghua University

清華大學电子工程系
Department of Electronic Engineering, Tsinghua University

Thank you for your attention!

Check our paper website at

https://sites.google.com/view/revisiting-marl.

Correspondence to: Wei Fu <fuwth17@gmail.com>, Yi Wu <jxwuyi@gmail.com>.