Tao YANG, Xinhao SHI, Qinghan ZENG, Yulin YANG, Cheng XU, Hongzhe LIU, 2025. Optimization methods in fully cooperative scenarios: a review of multiagent reinforcement learning. *Frontiers of Information Technology & Electronic Engineering*, 26(4):479-509. https://doi.org/10.1631/FITEE.2400259

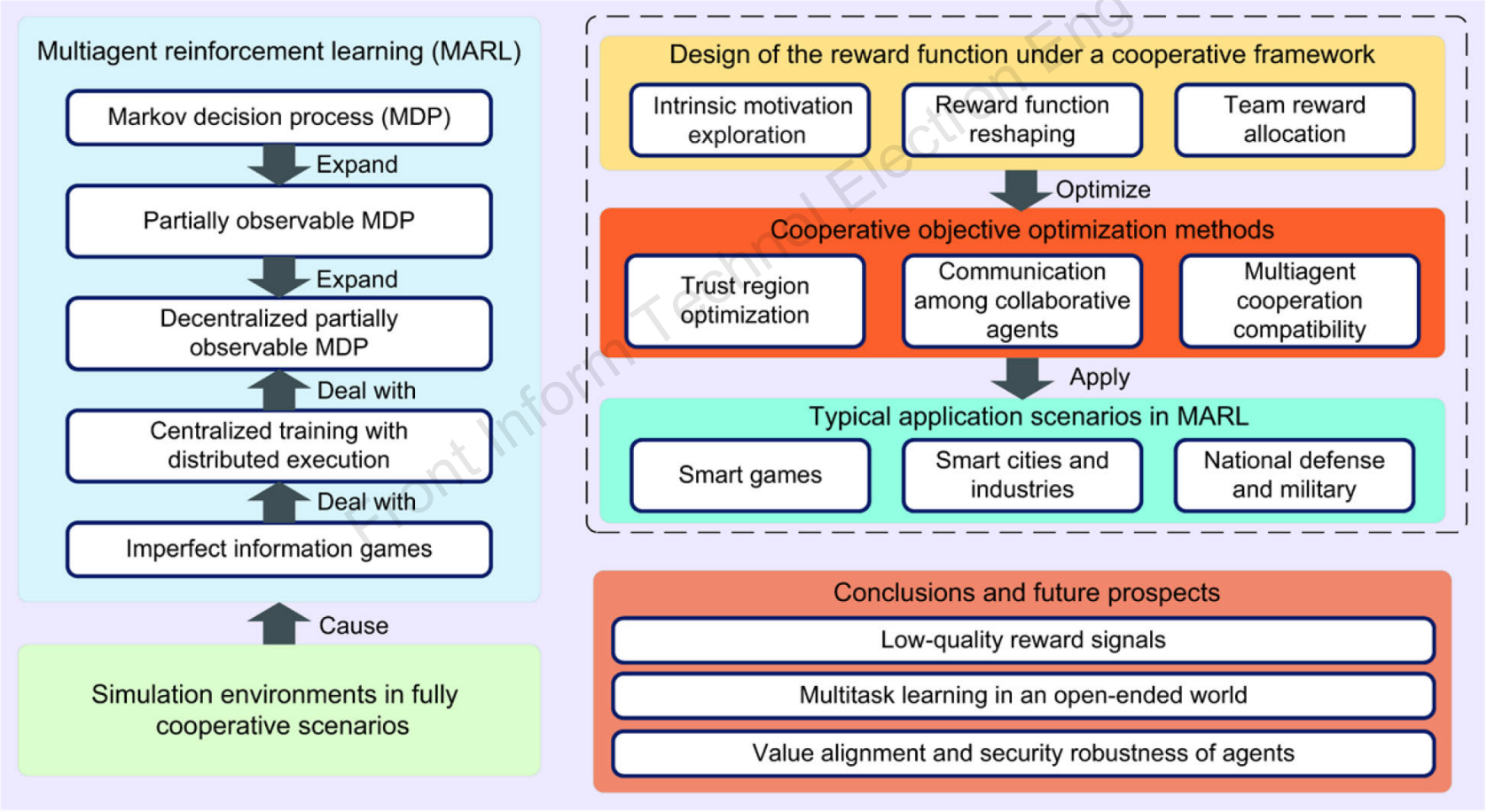
Optimization methods in fully cooperative scenarios: a review of multiagent reinforcement learning

Key words: Multiagent reinforcement learning (MARL); Cooperative framework; Reward function; Cooperative objective optimization

Tao YANG; Xinhao SHI
E-mail: 2023108351092 3@buu.edu.cn; 2022108351092 7@buu.edu.cn
These two authors contributed equally to this work.
ORCID: Tao YANG, https://orcid.org/0009-0006-7873-2959
Xinhao SHI, https://orcid.org/0009-0007-7240-7458

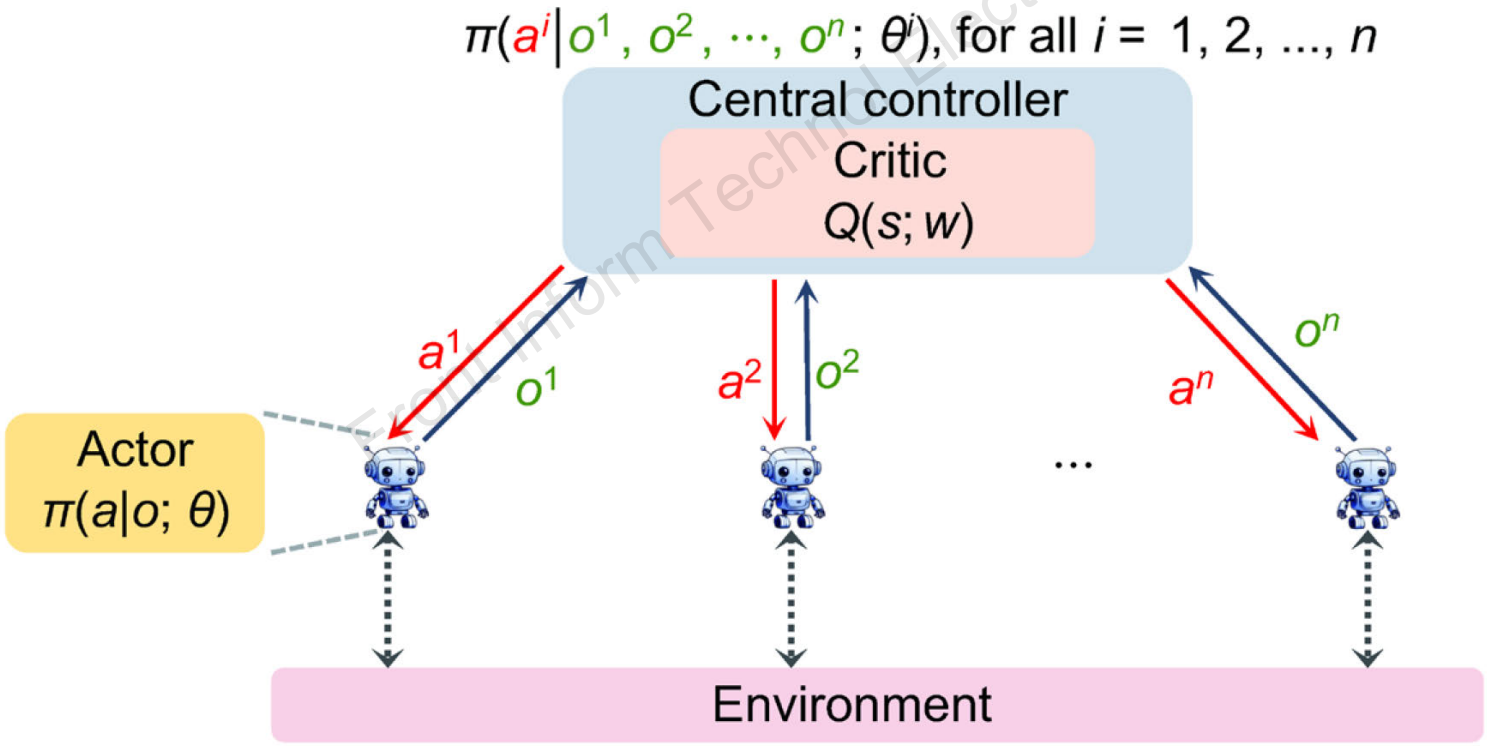
Overall structure of the review

- This review systematically explores MARL's core theories, reward mechanisms, cooperative optimization, real-world applications, evaluation tools, and future directions.



Centralized training with decentralized execution

- During training, agents collaboratively learn with access to global state information using a centralized method; during execution, agents independently make decisions using only their local information.



Design of the reward function

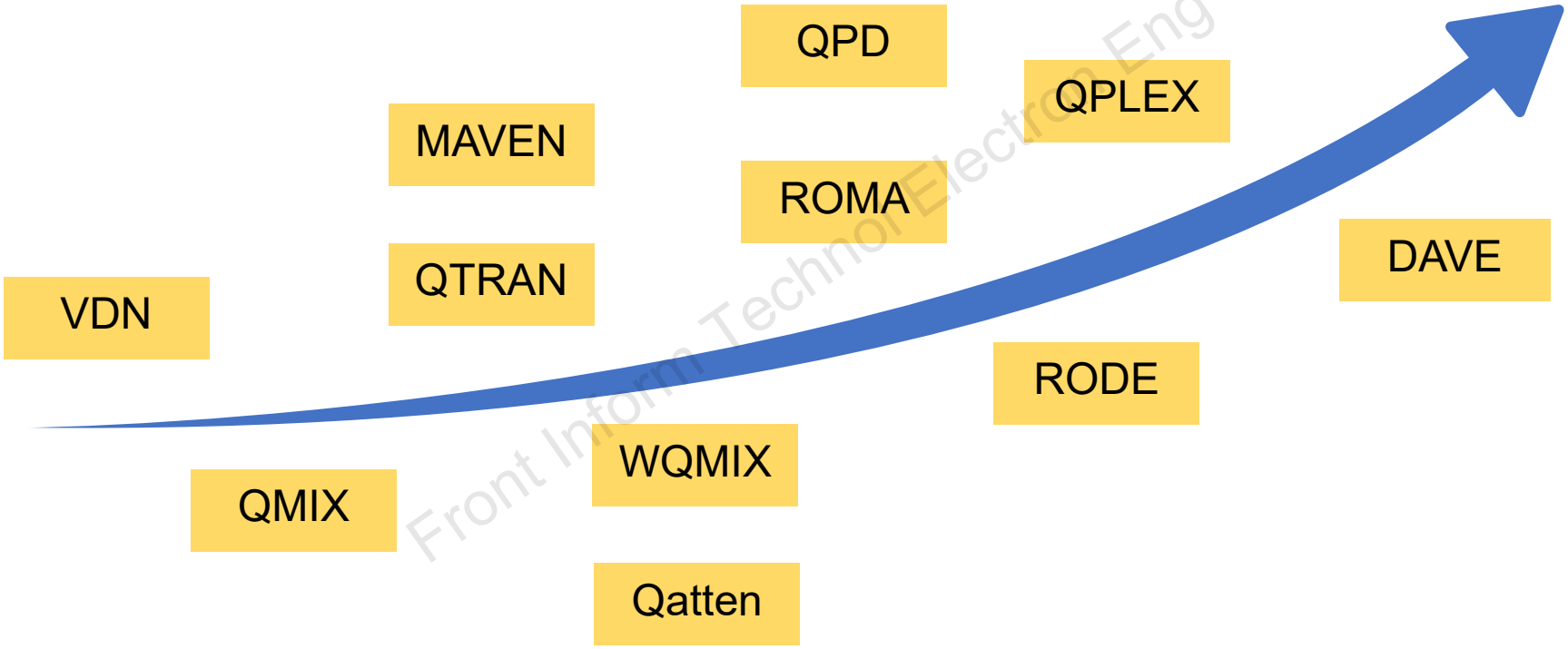
- Intrinsic motivation exploration
  - **Counting**
    - Density-based pseudo-count
    - Indirect pseudo-count
    - State abstraction
  - **Predictive model**
    - Prediction error
    - Prediction outcome discrepancy
    - Improvement of prediction accuracy
  - **Information theory**
    - Information gain
    - Maximum entropy
    - Mutual information

Design of the reward function

- Reward function reshaping
  - **Potential-based reward shaping** supplements rewards with state potential differences.
  - **Roadmap of reward shaping** accelerates learning with state potential differences without altering optimal policies.
  - **Dynamic potential** adapts rewards to environmental and behavioral changes via variable potential functions.
  - **Relative entropy inverse reinforcement learning** infers rewards by aligning agent behaviors with experts via entropy minimization.
  - **Meta-learning reward shaping** learns adaptable shaping functions through meta-learning for higher efficiency and better generalizability.

Design of the reward function

- Team reward allocation—credit assignment



Cooperative objective optimization methods

- Trust region optimization—policy update stability

Table 4 Summary of monotonic boundary constraint representation methods

Reference	Update	Sample efficiency	Monotonic bound
Wang XH et al., 2023	Sequential	Low	$4\varepsilon \sum_{i=1}^{n} \alpha_i \left(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\alpha_i)}\right)$
Yu et al., 2022	Simultaneous	High	$4\varepsilon \sum_{i=1}^{n} \frac{\alpha_i}{1-\gamma}$
Wu ZF et al., 2021	Simultaneous	High	$4\varepsilon \sum_{i=1}^{n} \alpha_i \left(\frac{1}{1-\gamma} - \frac{1}{1-\gamma\left(1-\sum_{j=1}^{n} \alpha_j\right)}\right)$
Kuba et al., 2021	Sequential	High	$4\varepsilon \sum_{i=1}^{n} \alpha_i \left(\frac{1}{1-\gamma} - \frac{1}{1-\gamma\left(1-\sum_{j=1}^{n} \alpha_j\right)}\right)$
Wang XH et al., 2023	Sequential	High	$4\varepsilon \sum_{i=1}^{n} \alpha_i \left(\frac{1}{1-\gamma} - \frac{1}{1-\gamma\left(1-\sum_{j\in e_i \cup \{i\}} \alpha_j\right)}\right) + \sum_{i=1}^{n} \frac{\xi_i}{1-\gamma}$
Fu et al., 2022	Sequential	High	$4\varepsilon \sum_{i=1}^{n} \frac{\alpha_i}{1-\gamma}$
Zhuang et al., 2023	Simultaneous	High	$4\varepsilon \frac{\alpha_i}{1-\gamma}$
Yang TP et al., 2021	Simultaneous	High	$4\varepsilon \sum_{i=1}^{n} \frac{\alpha_i}{1-\gamma}$
Ye et al., 2023	Sequential	High	$4\varepsilon \sum_{i=1}^{n} \alpha_i \left(\frac{1}{1-\gamma} - \frac{1}{1-\gamma\left(1-\sum_{j=1}^{n} \alpha_j\right)}\right)$

ε represents the maximum absolute value of the advantage function under specific policies, γ is a discount factor, α is the maximum total variation distance between the original policy and the updated policy for each agent, and ξ_i represents the error compensation term for the i^{th} agent

Cooperative objective optimization methods

- Communication among collaborative agents

Table 5 Summary of communication-based MARL algorithms

Communication type	Algorithm name	Algorithm summary
Based on value function approximation	FedQMIX (Cao SH et al., 2024)	Adding regularization penalties to punish the use of additional communication rounds, thereby improving the communication efficiency of agents
	C2E (Du XQ et al., 2024)	Using a set of critic networks that communicate with each other to estimate action values more accurately
	DDRQN (Foerster et al., 2016)	Using deep recurrent Q-networks to evaluate agent actions and communication strategies
Based on policy gradient methods	DACOM (Yuan TT et al., 2023)	Introducing the TimeNet component, which adjusts the waiting time for agents to receive messages from others, addressing delay-related uncertainties
	BiCNet (Peng P et al., 2017)	Introducing bidirectional coordination networks to facilitate effective communication among multiple agents
	MD-MADDPG (Pesce and Montana, 2020)	Using shared memory as a communication channel, where agents read and provide information before action execution
	Intrinsic A3C (Jaques et al., 2019)	Providing additional incentives for collaborative actions with high mutual information
	MACC (Vanneste et al., 2020)	Using counterfactual reasoning to train both action and communication strategies of agents
Improving communication flexibility and learning efficiency	ATOC (Jiang JC and Lu, 2018)	Using attention models to guide agent communication timing and information integration
	MAC (Miuccio et al., 2024)	Agents learn to control communication protocols to communicate effectively while considering communication overhead
	NASA (Abdel-Aziz et al., 2024)	Agents jointly learn adaptive communication protocols within a dynamic state space
	RTS (Canese et al., 2024)	Optimizing communication protocols to reduce data loss between agents; exhibiting strong robustness to the number of agents
	I2C (Ding et al., 2020)	Proposing an independent inference communication mechanism, allowing agents to learn the behaviors of others without explicit communication
Based on attention	MACRL (Xiao J et al., 2023)	Employing graph attention mechanisms to generate agent aggregation vectors based on the calculated interagent distance relevancy
	AERL (Pu et al., 2023)	Using spatiotemporal attention mechanisms to filter communication information and expand the communication range of agents
	TarMAC (Das et al., 2019)	Allowing agents to actively select the recipients of messages through a signature-based soft attention mechanism

Cooperative objective optimization methods

- Multiagent cooperation compatibility—strategy alignment
  - Definition
    - Cooperation incompatibility arises from conflicts in strategies or behaviors among agents.
    - Causes include divergent goals, inconsistent action choices, and strategy misalignments.
  - Challenge
    - Traditional MARL often focuses on individual agent performance rather than cooperative skill development, which leads to difficulties in achieving harmonious multiagent collaboration.

Typical application scenarios in MARL

- Smart games
- Smart cities and industries
- National defense and military
- Simulation environments in fully cooperative scenarios

Table 6 Introduction to typical multiagent test environments

Environment name	Action space	Original learning mode	Reward	Observation
SMAC (Samvelyan et al., 2019)	Discrete	Cooperative	Mixed	Partial
MPE (Lowe et al., 2017)	Hybrid	Mixed	Dense	Full
MAMuJoCo (de Witt et al., 2021)	Continuous	Cooperative	Dense	Partial
GRF (Kurach et al., 2020)	Discrete	Mixed	Sparse	Full
SISL (Gupta et al., 2017)	Hybrid	Cooperative	Dense	Full
LBF (Papoudakis et al., 2022)	Discrete	Mixed	Dense	Partial
RWARE (Papoudakis et al., 2022)	Discrete	Cooperative	Sparse	Partial
MAgent (Zheng LM et al., 2018)	Discrete	Mixed	Dense	Partial
Pommerman (Resnick et al., 2018)	Discrete	Mixed	Sparse	Full
MetaDrive (Li QY et al., 2023)	Continuous	Collaborative	Dense	Partial
MATE (Pan et al., 2022)	Hybrid	Mixed	Dense	Partial
GoBigger (Zhang M et al., 2023)	Continuous	Mixed	Dense	Mixed
Overcooked (Carroll et al., 2019)	Discrete	Cooperative	Dense	Full
MAPDN (Wang JH et al., 2021a)	Continuous	Cooperative	Dense	Partial
Hide Seek (Baker et al., 2020)	Discrete	Mixed	Dense	Partial

Conclusions and future prospects

- Low-quality reward signals
  - Sparse, team-level, deceptive, and delayed reward signals limit learning effectiveness.
- Multitask learning in an open-ended world
  - Dynamic, evolving environments require agents to learn diverse, interdependent tasks simultaneously.
  - Challenges include imperfect information, large state/action spaces, complex planning, and long-term reasoning.
- Value alignment and security robustness of agents
  - Ensuring that agent goals align with human values and ethical norms
  - Risks of reward hacking and power-seeking behaviors by agents