On Cooperation in Multi-Agent Reinforcement Learning

Karush Suri

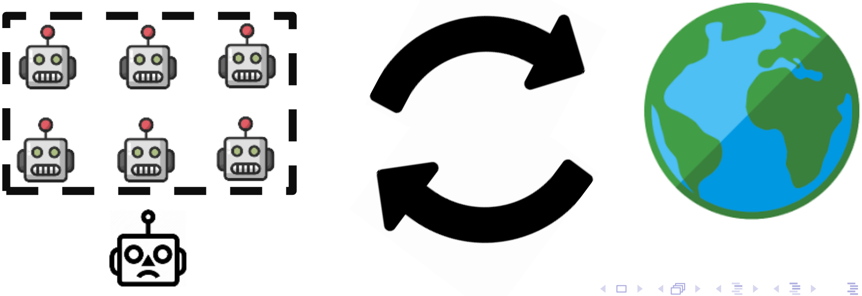
University of Toronto
Department of Electrical and Computer Engineering
ECE1657 Game Theory and Evolutionary Games
Fall 2020

Overview

- ❑ The Multi-Agent Paradigm
- ❑ Stochastic Markov Games
- ❑ Q-Learning
- ❑ Multi-Agent Reinforcement Learning (MARL)
    - Independent Q-Learning (IQL)
    - Value Decomposition Network (VDN)
    - QMIX
- ❑ Surprise Minimization
- ❑ Experiments
- ❑ Discussion

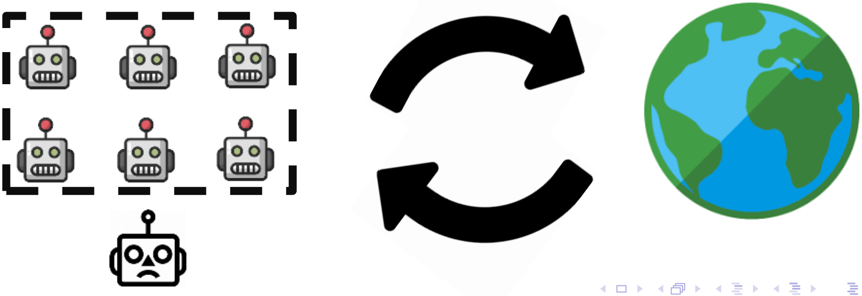
The Multi-Agent Paradigm

- More than one agent interacts with the environment to optimize strategies
- Agents may perform as a team or in selfish interests



The Multi-Agent Paradigm

- Cooperation gives rise to centralized updates with decentralized control
- Improved scalability to complex tasks which may not be not achievable for a single agent



But how can multiple agents learn to collaborate at the same time?

Stochastic Markov Games

- Setup of Stochastic Markov Games-

$$(\mathcal{S}, \mathcal{A}^1, \mathcal{A}^2 ... \mathcal{A}^n, r^1, r^2, ... r^n, N, P, \gamma)$$

$\mathcal{S}$ - State space

$\mathcal{A}^n$ - Action space of agent-n

r^n - Reward function of agent-n

N - Set of all players in the game

P - State transition probability distribution

γ - Discount factor

Stochastic Markov Games

- Also known as General Markov Games (GMGs).

$$(S, A^1, A^2 ... A^n, r^1, r^2, ... r^n, N, P, \gamma)$$

S - State space

A^n - Action space of agent-n

r^n - Reward function of agent-n

N - Set of all players in the game

P - State transition probability distribution

γ - Discount factor

Stochastic Markov Games

- Collaboration is focussed in Team Markov Games (TMGs).

$$(S, \mathcal{A}, r, N, P, \gamma)$$

S - State space

$\mathcal{A}$ - Action space of all agent

r - Reward function of all agents

N - Set of all players in the game

P - State transition probability distribution

γ - Discount factor

Lucian Buşoniu, Robert Babuška, and Bart De Schutter. Multi-agent reinforcement learning: An overview.. 2010.

Stochastic Markov Games

Team Markov Games-

$$(S, \mathcal{A}, r, N, P, \gamma)$$

- Combined strategy space for all agents.
- All agents observe the same reward.

$$\mathcal{A} : \mathcal{A}^1 \times \mathcal{A}^2 \times ...\mathcal{A}^n$$

$$r^1 = r^2 = ...r^n = r$$

But how do agents learn TMGs to maximize payoffs?

Q-Learning

- A form of Reinforcement Learning (RL)
- Agents select joint action a and optimize their policy $\pi(u_t|s_t)$ based on Q-values

$$Q(u, s; \theta) = \mathbb{E}_{\pi\theta}[\sum_{t=1}^{T} \gamma^t r(s, u)|s = s_t, u = u_t]$$

- Discounted returns motivate long-horizon behaviors and collaboration
- Each agent maintains its own Q-values which form the joint Q-values of all agents

Richard S. Sutton and Andrew G. Barto. Reinforcement Learning: An Introduction. 2018.

Q-Learning

- Agent policies parameterized using parameters θ.

$$\pi^* = (\pi^{1,*}, \pi^{2,*}, \dots \pi^{N,*})$$
$$s.t. \quad Q(u^a, s; \theta^*) \geq Q(u^a, s; \theta)$$
$$\forall s \in S, u \in A, a \in N$$

- Joint optimal policy is the Nash Equilibrium (NE) of the Stochastic TMG.
- But how to update policies in the long-horizon?

Richard S. Sutton and Andrew G. Barto. Reinforcement Learning: An Introduction. 2018.

Q-Learning

- Temporal Difference Learning-

$$\mathbb{L}(\theta) = \mathop{\mathbb{E}}_{b \sim R}[(y - Q(u, s; \theta))^2]$$

$$y = r + \gamma \max_{u_{t+1} \in A} Q(u_{t+1}, s_{t+1}; \theta^{-})$$

- Update θ w.r.t one step Q-value estimates
- Minimize cost using Gradient Descent
- Select best action u_{t+1} greedily using Boltzmann distribution

Richard S. Sutton and Andrew G. Barto. Reinforcement Learning: An Introduction. 2018.

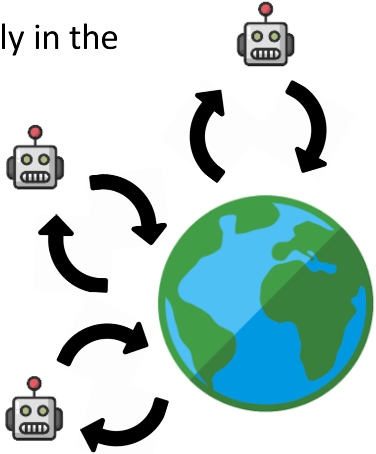
How to make use of Temporal Difference Learning in Multi-Agent settings?

- ❑ Review state-of-the-art methods in MARL.
- ❑ Improve collaboration and decentralized control.

- Independent Q-Learning (IQL)- Each agent updates its own Q-values.
- Value Decomposition Networks (VDNs)- Joint factorization of individual Q-values.
- QMIX- Monotonic $mixing$ (nonlinear factorization) of individual Q-values.

Independent Q-Learning (IQL)

- ❑ Agents interact independently in the environment.
- ❑ Each agent updates its own beliefs.
- ❑ Agents are a part of team But act in selfish interests.



Independent Q-Learning (IQL)

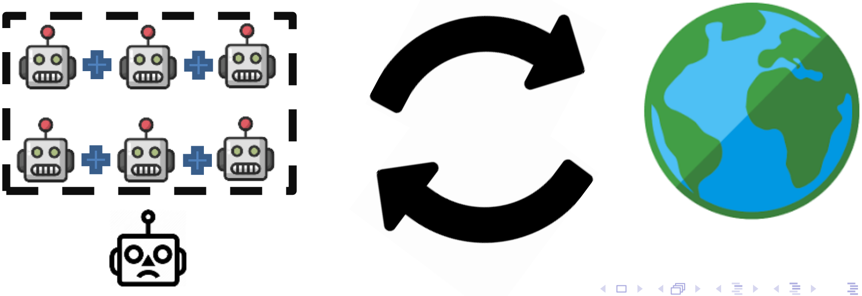
Cons-

- Each agent faces a non-stationary problem which depends on actions of other agents
- Lack of centralized information does not yield global convergence
- Computationally expensive since each agent updates its own beliefs separately.

Value Decomposition Networks (VDNs)

- Joint Q-value estimates expressed as a sum of individual estimates

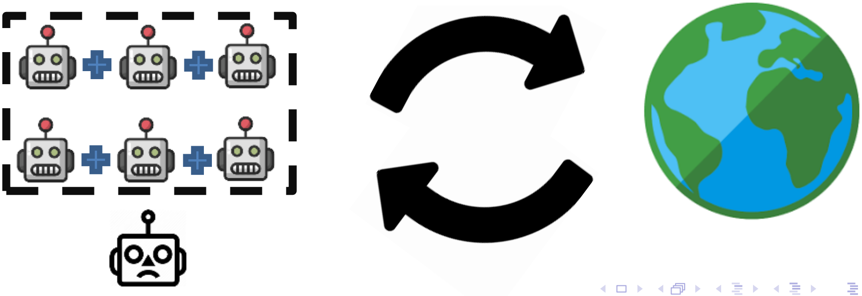
$$Q(u, s; \theta) \approx \sum_{i=1}^{n} Q_i(u^{(i)}, z^{(i)}; \theta)$$



Value Decomposition Networks (VDNs)

- ❑ Additive factorization leads to centralized information under the joint policy

$$Q(u, s, ; \theta) = \mathbb{E}_{\pi^\theta}[\sum_{t=1}^{\infty} \gamma^{t-1} r(s, u) | s = s_t, u = u_t]$$



Value Decomposition Networks (VDNs)

Pros-

- Centralized information in gradients leads to long-term collaboration
- Computationally efficient since updates require minimization of the joint cost function

Cons-

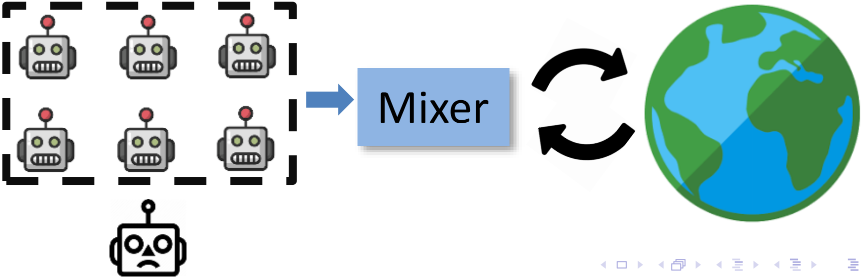
- Linear factorization leads to locally-optimal solutions
- Does not scale well in the number of agents

Multi-Agent Reinforcement Learning (MARL)

QMIX

- Nonlinear factorization of Q-value estimates using a $mixing$ component

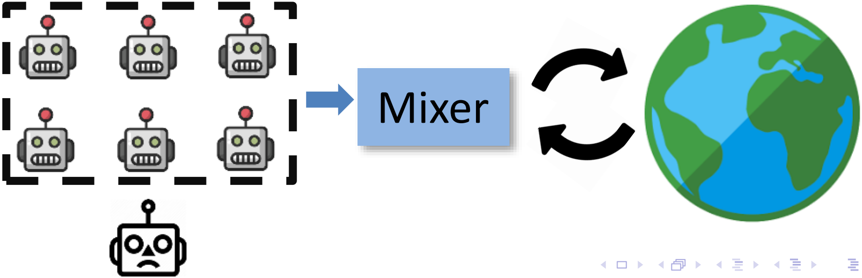
$$Q(u,s) = f(Q_1(u^{(1)}, z^{(1)}; \theta), \ldots Q_n(u^{(n)}, z^{(n)}; \theta)),$$
$$s.t.\ f : Q_i(u^{(i)}, z^{(i)}; \theta) \rightarrow Q(u, s; \theta),\ \forall i \in N$$



QMIX

- Enforce monotonicity constraints for extracting decentralized policies

$$\frac{\partial Q}{\partial Q_i} \geq 0. \quad \forall i \in N$$



QMIX

- Consistent updates require global argmax

$$\arg\max_{u} Q(u,s) = \begin{pmatrix} \arg\max_{u^{(1)}} Q_1(u^{(1)}, z^{(1)}; \theta) \\ \arg\max_{u^{(2)}} Q_2(u^{(2)}, z^{(2)}; \theta) \\ \vdots \\ \arg\max_{u^{(n)}} Q_n(u^{(n)}, z^{(n)}; \theta) \end{pmatrix}$$

- Pseudo-argmax similar to pseudo-gradient except that constraints are enforced

$$L(\theta) = \mathop{\mathbb{E}}_{b\sim R}[(r + \gamma \max_{u_{t+1}\in A} Q(u_{t+1}, s_{t+1}; \theta^{-}) - Q(u_t, s_t; \theta))^2]$$

QMIX

Pros-

- ❑ Consistency in centralized information
- ❑ Nonlinearity results in globally-optimal solutions
- ❑ Scalable in the number of agents

Cons-

- ❑ QMIX's argmax is not always correct*
- ❑ Prone to high variance under stochastic dynamics

* Rashid, Tabish, et al. "Weighted QMIX: Expanding Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning." (2020).

How to tackle stochastic dynamics in MARL?

Surprise Minimization

- Stochastic states arise from anomalous transitions termed as surprise.
- Minimizing surprise is essential for allowing the retrieval of best response.
- However, estimating surprise is challenging due to high stochasticity under fast-paced dynamics.
- We propose an energy-based formulation of surprise which allows agents to improve performance

Surprise Minimization

- Define an energy operator which allows estimation of surprise-

$$\mathcal{T}V^{a}_{surp}(s,u,\sigma) = \log \sum_{a=1}^{N} \exp\left(V^{a}_{surp}(s,u,\sigma)\right)$$

σ - standard deviation of state distributions

$V^{a}_{surp}(s,u,\sigma)$ - surprise value function

- Surprise value function assigns a value to each surprising state which is encoded by the energy operator

Surprise Minimization

- Choice of energy operator is suitable as it forms a contraction on surprise value function
- Formulation of the energy-based cost function-

$$L(\theta) = \mathop{\mathbb{E}}_{b\sim R}[\frac{1}{2}(y - (Q(u, s; \theta) + \beta \log \sum_{a=1}^{N} \exp (V_{surp}^{a}(s, u, \sigma))))^2]$$

$$y = r + \gamma \max_{u'} Q(u', s'; \theta^-) + \beta \log \sum_{a=1}^{N} \exp (V_{surp}^{a}(s', u', \sigma'))$$

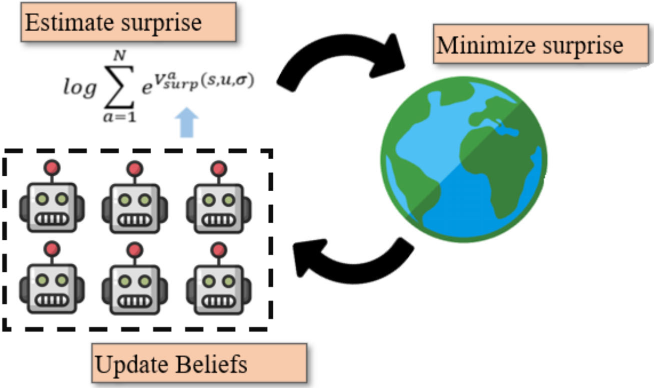
$$L(\theta) = \mathop{\mathbb{E}}_{b\sim R}[\frac{1}{2}(r + \gamma \max_{u'} Q(u', s'; \theta^-) + \beta E - Q(u, s; \theta))^2]$$

Surprise Minimization

- Surprise formulation serves as intrinsic motivation or reward regularization for agents
- Temperature parameter to balance noisy estimates with actual reward
- We further show that an optimal policy π^* consists of minimum surprise upon convergence

$$L(\theta) = \mathop{\mathbb{E}}_{b \sim R}[\frac{1}{2}(r + \gamma \max_{u'} Q(u', s'; \theta^-) + \beta E - Q(u, s; \theta))^2]$$

Surprise Minimization



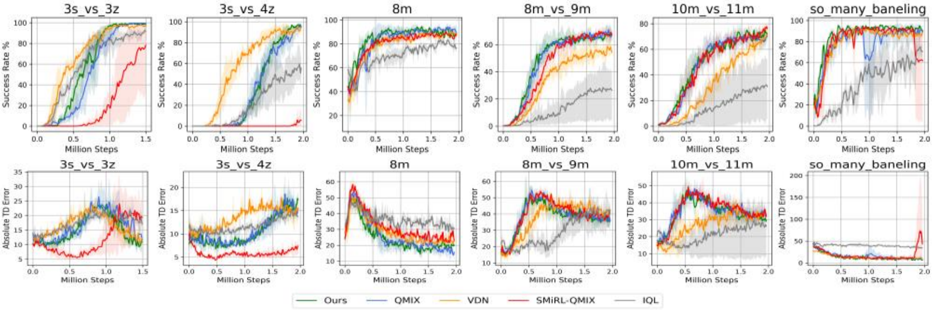
$$L(\theta) = \mathop{\mathbb{E}}_{b \sim R}[\frac{1}{2}(r + \gamma \max_{u'} Q(u', s'; \theta^-) + \beta E - Q(u, s; \theta))^2]$$

But how do we validate the suitability of surprise minimization?

Experiments

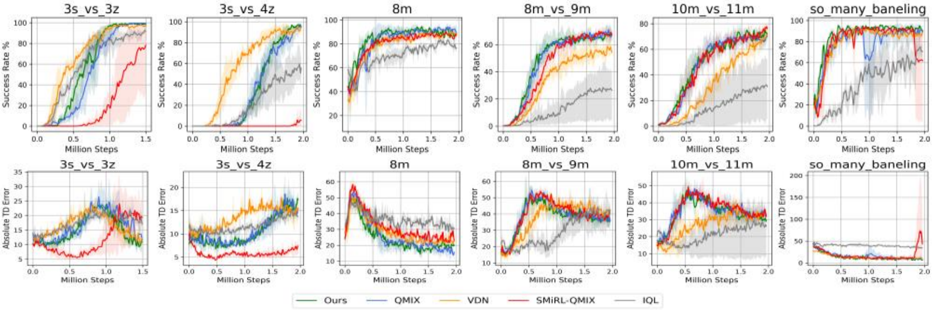
- Compare iterative performance of agents on a MARL benchmark
- Evaluate agents on large-scale StarCraft II tasks consisting of combat scenarios
- Agents need to strategically collaborate in order to defeat the opponent team
- Experiments consist of homogenous agents (teams formed for same set of agents)
- Results averaged over 5 random runs for 2 million iteration steps

Experiments



- **Top**- Surprise minimization scheme demonstrates improved win rate on 4 out of 6 tasks
- **Bottom**- Minimization of TD error
- Note the initial rise in error due to exploration

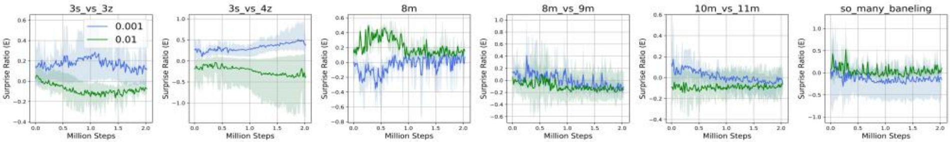
Experiments



- **Top**- collaboration-based schemes demonstrate improved performance in comparison to IQL
- **Bottom**- High variance and bias in IQL cost
- Note the initial rise in error due to exploration

Surprise Minimization

Dependence of surprise-based scheme on temperature parameter



Surprise-robust behaviors during collaboration

Discussion

Conclusions-

- ❑ Partial observability in TMGs a hindrance towards optimal strategy execution
- ❑ Cooperation facilitates scalability and convergence in MARL
- ❑ Requirement of surprise-robust schemes in stochastic dynamics
- ❑ Theoretical and empirical evaluation depict suitability of energy-based scheme

Limitations and Future Work-

- ❑ Improve performance gains in the case of large number of agents
- ❑ Alternatives to surprise value function in case of noisy estimates
- ❑ Robust behavior a consequence of counterfactual states, i.e.- transfer strategies across agents to facilitate local communication

Discussion

Thank You!