COMP0036 – Literature Review:

Multiagent Reinforcement Learning for Noised Communication in Fully Cooperative MPEs

**Jeffrey Li**, supervised by **Professor Mirco Musolesi**

BSc Computer Science, UCL

November 17, 2022

# Project Background

* TBD

# Background Knowledge

## 2.1 Markov Decision Processes

Markov Decision Processes (MDPs)[[1]](#footnote-1) are classical formulations of sequential decision-making, which seek to abstract the problem of goal-directed learning from interactions. This set of frameworks consists of two components: the agent being the decision maker and the environment in which the agent interacts with.

At each timestep , the agent would receive a representation of the environment current state and the agent selects an action where the action space depends on the current state . In response to the action, at the next time step , the agent would receive a real number reward and would be in a new state . There is a probability of those values occurring at time , given values of current state and action:

Which characterize the environments dynamics and also satisfies the law of total probability for each choice of and :

From this, we could also observe that the probability of each value of and depend entirely on the state and the action of the agent one timestep before. In other words, Markov Property states that the future depends completely on the present and not on the past, and any scenario which follows this property are defined as Markovian.

## 2.2 Reinforcement Learning

### 2.2.1 Problem Setting

Reinforcement Learning (RL) sits within the MDP framework, and the goal is the maximize the expected value of the return (cumulative sum of rewards). More precisely, it is maximizing the expected value of the discounted return with discount factor to take into account of continuous tasks which may have infinite .

This expected return is represented as value functions, which defines the expected return of an agent at a particular state, while following certain agent behaviours that are governed by what is known as policies. Policies are functions that maps from state to the probability of selecting each action , denoted as . Thus the value function at state while following policy has the following definition:

Value of a state can also be considered in conjunction with a specific action instead of an expectation over all possible actions, called Q functions, which is more commonly used in many RL algorithms and can also be used to express value functions:

Since the goal is to maximize the expected return , a RL problem can be formulated as estimating value functions given a certain policy, and search for the optimal policy that leads to the expected return (defined by its value function ) greater or equal to any other policies.

### 2.2.2 Bellman’s Equations

### 2.2.3 Well-known algorithms

Value iteration, Policy Iteration

* Known environment dynamics

Q-learning

* Model Free
* Tabular approach

## 2.2 Deep Reinforcement Learning

As shown in the Q-Learning method, the

To account for larger state spaces, RL algorithms using …

* Success in DRL
  + V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. Nature, 518(7540):529–533, 2015.
  + Mastering the game of Go with deep neural networks and tree search.
* Deep Q Networks

## Multiagent Reinforcement Learning

From the surface, it seems that Multiagent Reinforcement Learning is solving similar MDPs as RL but in multi-agent setting[[2]](#footnote-2)

More on multi-agent MDP settings…

Cooperative, Competitive, Mixed Cooperative Competitive settings

Unlike single-agent reinforcement learning, which can only interact with its environment, in the multi-agent setting, agents would be able to interconnect with other agents like what humans do naturally in our society, such as competing with each other where all agents seek to accomplish a goal over the other agents, cooperate where agents would work together to achieve a common goal or some sort of combination of both.

Cooperative AI[[3]](#footnote-3) using MARL has gained increasing attention over the years and radiates through a wide range of impactful fields, such as the AI Economist[[4]](#footnote-4) which aims to improve equality and productivity with AI-Driven Tax policies, the study of sequential social dilemmas which require multiple agents to learn policies that implement their strategic intentions[[5]](#footnote-5) and the study of reputation in cooperative systems[[6]](#footnote-6) to name a few, all of which can be integrated well into the study of strategic games and society protocols.

OpenAI agents playing hide and seek[[7]](#footnote-7)

* Multiagent bidirectionally coordinated nets for learning to play starcraft combat games.
  + Multiagent game playing

This is a very interesting problem to tackle that can be useful in practical applications such as setting up protocols for multiple robots to do security patrolling[[8]](#footnote-8),

An initial attempt to such problem is known as the Independent Q-Learning algorithm

* VDN
* QMIX
* MDDPG
* MAPPO

Independent Q Learning

* Independent reinforcement learners in cooperative Markov games: a survey regarding coordination problems
  + Independent Q-Learning agents, shown in paper that they don’t perform well in multi-agent settings.
  + Policy of each agent changes during training, causing non-stationarity in the perspective of a single agent, preventing naïve approach of experience replay
    - <http://proceedings.mlr.press/v70/foerster17b/foerster17b.pdf>
      * introduces nonstationarity that makes it incompati-  
        ble with the experience replay memory on which  
        deep Q-learning relies
  + Violates Markov assumption for convergence of Q Learning
* <https://proceedings.neurips.cc/paper/2003/hash/e71e5cd119bbc5797164fb0cd7fd94a4-Abstract.html>
  + Attempts in inputting other agents’ policy parameters to Q function to overcome such non-stationarity
* <https://arxiv.org/abs/1702.08887>
  + using a multi-agent variant of importance sampling to naturally decay obsolete data
  + and conditioning each agent's value function on a fingerprint that disambiguates the age of the data sampled from the replay memory
    - Essentially indexing the experiences

Cooperative settings

* <https://ieeexplore.ieee.org/abstract/document/4399095>
  + Hysteric Q Learning

Policy Gradient

* General Policy Gradient
  + <https://proceedings.neurips.cc/paper/1999/hash/464d828b85b0bed98e80ade0a5c43b0f-Abstract.html>
  + Known to exhibit high variance gradient estimations
    - More so in multiagent context as agent’s reward depends on action of many agents, therefore reward conditioned only on agent’s own actions exhibits higher variance
* Actor Critic
  + Rather similar to GAN
  + Temporal Difference version of policy gradient
  + Contains an actor and critic
    - Actor decides which action should be taken
    - Critic inform the actor how good was the action and how it should adjust
  + Actor learns through policy gradient approach; critics evaluate the action produced by actor by computing the value function
* Extend policy gradient framework to Deterministic Policy Gradient (DPG) algorithms
  + Deep DPG is a variant of DPG where the policy and critic are approximated with deep neural networks
    - <https://arxiv.org/pdf/1509.02971.pdf?source=post_page--------------------------->
    - Also make use of experience replay as in DQN to stabilize the neural network
* <https://ojs.aaai.org/index.php/AAAI/article/view/11794> (COMA)
  + Counterfactual multi-agent policy gradients, uses centralised critic to estimate Q function and decentralised actors to optimise agents’ policies
  + Address the challenges of multi-agent credit assignment, uses a counterfactual baseline that marginalised out a single agent’s action while keeping the other agents’ actions fixed
  + Learns a single centralized critic for all agents
* <https://arxiv.org/abs/1706.02275> (MADDPG)
  + Actor critic policy gradient where the critic is augmented with extra information about the policies of other agents, while the actor only has access to local information. After training is completed, only the local actors are used at execution phase
  + Since the centralized critic function explicitly uses the decision-making policies of other agents, we additionally show that agents can learn approximate models of other agents online and effectively use them in their own policy learning procedure.
  + acting in a decentralized manner and equally applicable in cooperative and competitive settings.
  + learn a centralized critic for each agent, allowing for agents with differing reward functions including competitive scenarios
  + consider environments with explicit communication between agents
  + learns continuous policies

# Related work

This project focuses on communication in MARL context to establish efficient cooperation in coordination tasks.

* <https://link.springer.com/chapter/10.1007/978-3-540-45173-0_29>
  + Genetic algorithm to learn the languages for the predator agents in the predator-prey problem
  + Not scalable for larger problems

Learning grounded cooperative communication protocols between agents

* <https://proceedings.neurips.cc/paper/2016/file/55b1927fdafef39c48e5b73b5d61ea60-Paper.pdf>
* <https://proceedings.neurips.cc/paper/2016/file/c7635bfd99248a2cdef8249ef7bfbef4-Paper.pdf>
* usually only applicable when the communication between agents is carried out over a dedicated, differentiable communication channel.

Deciding when to communicate

* <https://arxiv.org/abs/1812.09755>
  + Extended from CommNet
  + Individualized Controlled Continuous Communication Model (IC3Net)
  + Controls communication with gating mechanism to decide when to communicate and uses individualized rewards for each agent to gain better performance and scalability
  + Applied to cooperative, semi-cooperative and competitive settings
* <https://arxiv.org/pdf/1902.01554.pdf>
  + SchedNet
  + Limited bandwidth
  + the agents share the communication medium so that only a restricted number of agents are able to simultaneously use the medium
    - to simulate state-of-the-art wireless network architectures
  + Learns to decide which agents should be entitled to  
    broadcasting their (encoded) messages, by learning the importance of each agent’s  
    partially observed information.

There are two paradigms to solving

Communication treated as differentiable process (continuous) optimized through backprop

* Tend to converge quickly to higher-quality policies compared to traditional RL framework
* <https://proceedings.neurips.cc/paper/2016/file/c7635bfd99248a2cdef8249ef7bfbef4-Paper.pdf>
  + Centralised training, decentralised execution
  + DIAL approach, allows real valued messages to pass between agents during centralised learning allowing gradients to be pushed through the communication channel, and allows real-valued messages to pass between agents during centralised learning, thereby treating communication actions as bottleneck connections between agents. As a result, gradients can be pushed through the communication channel, yielding a system that is end-to-end trainable even across agents. During decentralised execution, real-valued messages are discretised and mapped to the discrete set of communication actions allowed by the task. Because DIAL passes gradients from agent to agent, it is an inherently deep learning approach.
* <https://arxiv.org/pdf/1605.07736.pdf>
  + CommNet
  + Learns a shared Deep Neural Net that is shared across agents
  + Shared reward

Message treated as an extension to action space

* Communication behaviour learned and optimized via standard reinforcement learning
* <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5045926>
  + Make use of tabular Q Learning to solve the predator-prey task with communication
  + Similarly, in <https://scholar.google.co.uk/scholar?q=Efficient+Distributed+Reinforcement+Learning+through+Agreement,&hl=zh-CN&as_sdt=0&as_vis=1&oi=scholart>
* <https://proceedings.neurips.cc/paper/2016/file/c7635bfd99248a2cdef8249ef7bfbef4-Paper.pdf>
  + Same paper as DIAL
  + In case that communicated messages are discrete values and gradient are not able to be calculated
  + RIAL method makes use of deep Q Learning for better scalability to learn content of the message
* <https://arxiv.org/pdf/1602.02672.pdf>
  + DDRQN
  + Discover communication protocols to solve multiagent learning problems based on well-known riddles (partially observable tasks)
  + Not really solving coordination problems

The approach most similar to my project:

<https://ojs.aaai.org/index.php/AAAI/article/view/6205>

* + Communication Learning via Backpropagation in Discrete Channels with Unknown Noise
    - Stochastic message encoding/decoding procedure that makes a discrete communication channel mathematically equivalent to an analogue channel with additive noise
    - Then which gradients can be backpropagated

1. Howard, R. A. (1960). Dynamic programming and Markov processes. John Wiley. [↑](#footnote-ref-1)
2. <https://courses.cs.duke.edu/spring07/cps296.3/littman94markov.pdf> [↑](#footnote-ref-2)
3. A. Dafoe, Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson, and T. Graepel, “Cooperative AI: Machines must learn to find common ground,” *Nature News*, 04-May-2021. [Online]. Available: https://www.nature.com/articles/d41586-021-01170-0. [Accessed: 27-Mar-2022]. [↑](#footnote-ref-3)
4. S. Zheng, “The AI economist: Improving equality and productivity with AI-driven tax policies,” *Salesforce Research*, 30-Jun-2020. [Online]. Available: https://blog.salesforceairesearch.com/the-ai-economist/. [Accessed: 30-Mar-2022]. [↑](#footnote-ref-4)
5. “Understanding agent cooperation,” *RSS*. [Online]. Available: https://www.deepmind.com/blog/understanding-agent-cooperation. [Accessed: 30-Mar-2022]. [↑](#footnote-ref-5)
6. Nicolas Anastassacos, Julian García, Stephen Hailes, Mirco Musolesi: “Cooperation and Reputation Dynamics with Reinforcement Learning”, 2021; [http://arxiv.org/abs/2102.07523 arXiv:2102.07523]. [Accessed: 30-Mar-2022]. [↑](#footnote-ref-6)
7. https://arxiv.org/abs/1909.07528 [↑](#footnote-ref-7)
8. “Collaborative security robots use multi-agent Reinforcement Learning,” *YouTube*, 15-Apr-2021. [Online]. Available: https://www.youtube.com/watch?v=7QDr\_zxQitU. [Accessed: 30-Mar-2022]. [↑](#footnote-ref-8)