How to Cite:
Vimala, K., & Bharathi, S. (2022). Multi agent Markov games: Symmetric and asymmetric approach. *International Journal of Health Sciences*, *6*(S2), 2260–2267.
https://doi.org/10.53730/ijhs.v6nS2.5531

Multi agent Markov games: Symmetric and asymmetric approach

K. Vimala
Assistant Professor, Department of Mathematics Government Arts and Science College, Kadaladi-623703, Ramanathapuram-District, Tamil Nadu, India
Email: vimalamaths78@gmail.com

S. Bharathi
Assistant Professor & Research Supervisor, Department of Mathematics, Bharathiar University PG Extension and Research Center, Erode, Tamil Nadu, India
Email: bharathikamesh6@gmail.com

***Abstract*---**Modern computing systems are totally different from one that worked in the last decade. They are distributed, large and heterogeneous in structure. In the age of IOT, computers, information processing devices and humans are very tightly connected with each other to the support a combined effort. In the sense of decision processes, it would be preferable to handle their entities more as agents than stand – alone systems

***Keywords*---**modern computing, heterogeneous, IOT, computers.

Introduction

The major objective of Artificial Intelligence is to understand interactions between entities, (natural or artificial) and to suggest how to make good decisions, while taking other decision makers into account. In this thesis, the interactions between multiple agents in Markov games leads to an interdisciplinary research between two disciplines: Machine learning and Game Theory. Here we see Markov Games as an extended version of Partially Observable Markov Decision Process (POMDP), a well- known tool for modeling single agent problems such as Inventory control systems and service facility systems. Markov games can be seen as a dynamic extension to normal strategic form games, the standard models in game Theory.

Markov games, in the perspective of computer science provide a flexible and efficient method to describe different social interactions between intelligent agents. This thesis studies two alternative ways of learning in Markov Games.

Reinforcement learning model in AI is the most general model whose goal is the maximize the long – time performance of the learning agent.

In this thesis, we introduce an asymmetric learning model that is computationally efficient in multi-agent systems. Construction of hieratical order of agents is possible in this perspective. The multi-agent reinforcement learning systems, are based as Markov games. The time and space complexities of the reinforcement algorithms increase with the number of agents and the size of the problem instance. So it is necessary to use function approximates such as neural networks to model multi agent systems in real – world applications.

In this thesis, many numerical examples are provided to illustrate the Markov game models. The proposed methods are tested with small but non–trivial problems from different research domains. The list is not exhausting but includes artificial robot navigations, simplified game, and automated pricing models for intelligent agents. The thesis also contains an extensive literature survey on multi-agent reinforcement learning and various methods based as Markov Games.

Preliminaries

Introduction

This doctoral thesis deals with Multi-agent Markov games – Symmetric & Asymmetric Approach. Here multi-agent means that there are multiple number of learners present in the system. Refines decision making character of the learners are assumed is the proposed models. Actual reasoning and learning procedure take place inside each learner. It is assumed that there is a direct relation between intelligence and ability to learn. In this thesis the prime goal of agents is to learn is participate the long – time consequences of their action choice.

This is achieved through reinforcement learning method, based as POMDP. Also additional decision makes (agents) exist is the system with rational character. The relationship between agents is modeled with Markov games (MGS), directly the generalization of MDP for multi-agent systems. Markov games theory can be viewed as an extension from two different methods

- $POMDP_s$
- Classical game theory.

Single –agent POMDP were extended to multi-agent domains by associating a matrix game with each state of the POMDP. But, Markov Games can be seen as a direct extension of classical game theory having deterministic pay off to stochastic one, through matrix form of games. The following picture shows the different categories of optimization problems and the corresponding research.

	one agent	many agents
Static	Decision Theory	Static Game Theory
Dynamic	PO Markov Decision Processes	Markov Games.

The development $POMDP_s$ and MG rooted to the year 1950s. Since, then, there are two dynamic decision making models have evolved rather independently and have been studied in different fields namely applied mathematics, engineering and economics. The single agent dynamic $POMDP_s$ have been studied by many researches and efficient solving and learning methods exist for these processes. The theory of Markov Games not yet established like MDP_s. But in recent years, some numeric solution algorithms exists for these more complex processes. However, in the case of MGs, learning the game structure and optional playing methods are still very active open problems under research.

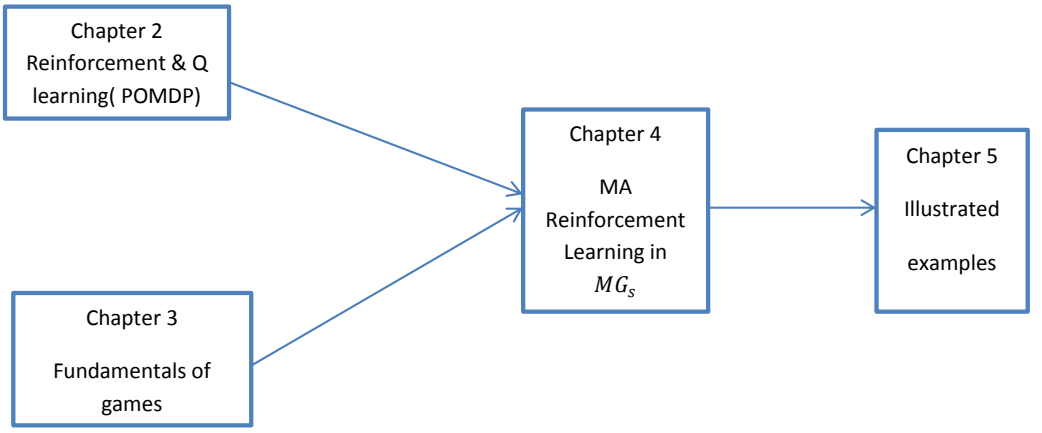
In this thesis, the focus is on efficient, both regular and numeric, learning methods is MGs. The thesis consists of an introduction part and three publications listed is section 1. 2. The publications contains most of the contributions of the Thesis and are thus referred in the appropriate positions of the text is the thesis.

The paper' primary contributions are as follows:
The following are the research' key contributions:

1. Literature Review Of The Current Status Of Markov Games, Reinforcement Learning Research.
   Tee fundamental results in the field of Markov Games and reinforcement learning, in particular multi agent reinforcement learning and partially observed Markov decision processes are covered in this thesis
2. Asymmetric Multi-Agent Reinforcement Learning Model.
   This learning model simplifies the decision making procedure in MGs by an additional requirement of the ordering among decision makers. In contrast to the symmetric learning model, the asymmetric learning model has stronger convergence properties and lower computational and space requirements.
3. Numeric Methods
   Numeric techniques for multi-agent reinforcement learning in MGs based on value functions and policy gradients. The usage of function approximates such as neural networks is required for applying multi-agent reinforcement learning approaches to genuine issues. This research proposed efficient and general gradient-based multi-agent reinforcement learning algorithms for this aim.
4. Hybrid Model For Multi-Agent Reinforcement Learning In MGS.
   In MGs, the opponents are modeled in each state of the system by using matrix games. In many problem instances this is not required and hence, in this thesis, a method is proposed for dividing the state space into two subspaces: the space of complex states and the space of simple states. The opponent is modeled only in the complex states resulting in much lower computational and space requirements in standard MGs.
5. The Research Is Organized In The Following Manner:
   The rest of the thesis is organized as depicted in Fig 1. 1. Chapter 2 contains background and fundamental of reinforcement learning based on $POMDP_s$. Chapter 3 deals with basic concepts and foundations of the mathematics of game theory. Chapter 2 and 3 together form two parallel

tracks for foundations for the next two chapters on multi-agent reinforcement learning in MG_s .

Chapter 4 is also the main chapter of the thesis and chapter 5 contains of the publications, deals with the main contributions of the author. Chapter 5 introduce two example of multi-agent reinforcement learning (Q - learning) based as Markov games.
Finally, chapter 6 concludes the thesis



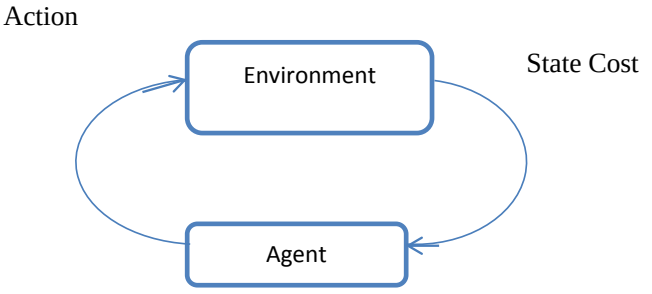


Fig 2. 1: An overview of the learning system in its environment

Implements its action selection by using the effectors and based on the action selection, the environment changes its state, perhaps stochastically. After the state transition, the agent observes the new state and an immediate reward associated with the state transition. Formally, the model of the system consists of the following parts (in this thesis it is assumed, for brevity, that the system is discrete, i. e. time, states, and actions are all discrete):

The environment is in some discrete state $s \in S$ at each time step. The learning agent has a set of action available in each state $s \in S$, denoted $A(s)$. When the system changes its state, the agent gets a real – valued reward $r \in \mathbb{R}$.

There is a function π that stipulates the behavior of the agent and thus summarizes the agent's knowledge of its environment. The main difference between supervised learning and reinforcement learning is that in reinforcement learning, the feedback from the environment does not include information about the right action choice. It only provides a punishment or a reward signal based on the current action choice of the agent. Due to the lack of correct answers, the learning system should gather information of its environment in a trial – and – error manner. This leads to the exploration vs. exploitation dilemma that will be discussed at the end of this chapter.

Suitable Performance Criteria

A rational agent always selects the action that maximizes some performance criterion that reflects design objectives of the agent. Usually it is not enough to maximize only the direct reward γ instead, it is preferable to maximize the long time performance of the agent. Three suitable long – time performance criteria are presented below. In each case, the criterion is an expected value because different sources of stochasticity can exist in the system. More-ever, all criteria depend on the policy π that stipulates the behavior of the agent.

- Expected total reward:

$E_\pi[\sum_{t=0}^{h} r_{t+1}]$. In this case, the objective function is the total, cumulative reward collected in some finite length h episode of the interaction between the agent and the environment. This error criterion is suitable for the cases in which the length of the task is known to the agent a priori. However, the difficulty is that the length of the task is known to the agent a priori. However, the difficulty is that the optimal behavior is not necessarily stationary; it could change over time.

- Expected total discounted reward:

$E_\pi[\sum_{t=0}^{\infty} \gamma^t \ r_{t+1}]$. This is almost the same as the previous criterion except that now the horizon h is infinite. With infinite horizon, the expected total reward would be unbounded and therefore the rewards are discounted by a discount factor $\gamma \in [0,1[$ that controls the balance between the significance of immediate rewards and future rewards. If γ is near zero, the agent makes its decisions almost myopically; decisions are based on immediate rewards and, on other hand, if γ is near one, the agent is willing to sacrifice immediate rewards for acquiring a large long – time expected utility value.

- Expected average reward:

$E_\pi\left[\frac{1}{h}\sum_{t=0}^{h} r_{t+1}\right]$. In this case, the goal is to maximize the total average reward over time. This criterion is complementary to the previous performance criterion and, in fact, the optimal policy maximizing the expected discounted reward criterion becomes equivalent to the policy maximizing the expected average reward criterion when γ approaches unity. The main difficulty with this criterion is that it is not possible to balance between policies that emphasize short time rewards and then policies that emphasize longtime rewards. On the other hand, the

methods that use the expected average reward criterion do not have a discount factor parameter γ and thus have one parameter less than methods based on the expected discounted reward criterion.

The second performance criterion, the expected total discounted reward criterion, is the one used most in reinforcement learning literature. This is mainly due to the fact that it is easier to handle mathematically than the other two criteria. All methods presented in this thesis utilize this performance criterion. Simple numerical examples are provided to illustrate the RL procedures in Markov games and PO MDP.

Conclusion

The equilibrium idea in MGs is dictated by the nature of the situation at hand. Several examples problems are used to test the learning methods in this thesis. In some of these experiments, the learning agents' roles are symmetric, resulting in symmetric equilibrium concepts, whereas in others, the agents' roles are asymmetric, resulting in asymmetric equilibrium concepts. The pricing challenges presented in this thesis are examples of problems in which the agents' roles are naturally varied as a result of the cost structure. In addition, the convergence speeds of both the symmetric and asymmetric learning approaches are evaluated on simple grid world issues. One justification for using the Max operator in multi agent Q-learning with team games is the asymmetric equilibrium concept. A policy gradient method is extended to multi agent domains based on this operator and tested with a simple soccer game.

Because MGs have a large number of free parameters to learn and multiple possible solution concepts with various properties, scholars in the field of reinforcement learning have been mostly interested in theorising about them. However, there are equilibrium notions, such as the Stackelberg equilibrium concept, that allow you to evaluate games fast and hence speed up the learning process significantly. However, this approach necessitates the learning agents' acceptance of their roles, and general-sum games do not allow for general convergence proofs. As a result, finding a learning approach that is guaranteed to converge to an equilibrium policy for general-sum problems remains an open question. After learning the stage games in an MG, it is feasible to classify and investigate the properties of these games in order to learn more about the underlying process. This classification could be done both manually or automatically using classification and clustering techniques. The study of game encoding, or how stage games might be provided to classification techniques, is an exciting future research field.

References

1. C.J. Watkins, Learning from Delayed Rewards, PhD thesis, Cambridge University, Cambridge, England, 1989
2. Cherian, Jacob; Jacob, Jolly; Qureshi, Rubina; Gaikar, Vilas. 2020. ‘Relationship between Entry Grades and Attrition Trends in the Context of Higher Education: Implication for Open Innovation of Education

Policy' MDPI, Switzerland, Journal of Open Innovation Technology, Market and Complexity, Vol- 6, Issue- 4: 199

3. Gavin A. Rummery and Mahesan Niranjan. On-line Q-learning using connectionist systems. Technical Report CUED/FINFENG/TR166, Cambridge University, Engineering Department, 1994.
4. John F. Nash. Equilibrium points in N-person games. Proceedings of National Academy of Sciences of the United States of America, 36, 1950.
5. Jonathan F. Bard. Practical Bilevel Optimization—Algorithms and Applications. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.
6. Leemon Baird and Andrew Moore. Gradient descent for general reinforcement learning. In Advances in Neural Information Processing Systems, volume 11, Cambridge, MA, 1999. MIT Press.
7. Leonid Peshkin, Kee-Eung Kim, Nicolas Meuleau, and Leslie P. Kaelbling. Learning to cooperate via policy-search. In Proceedings of the Sixteenth Conference on Uncertainty in Artifical Intelligence (UAI-2000), Stanford, CA, 2000. Morgan Kaufmann Publishers.
8. Michael L. Littman. Friend-or-Foe Q-learning in general-sum games. In Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williamstown, MA, 2001. Morgan Kaufmann Publishers.
9. Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In Proceedings of the Eleventh International Conference on Machine Learning (ICML-1994), New Brunswick, NJ, 1994. Morgan Kaufmann Publishers.
10. Sameer A., Gaikar V. (2019), 'Study of Demographic Variables on Financial Goal of Urban Individuals' in International Journal of Research, Vol. 9 (1), July-December 2019, Pp. 24 – 34.
11. Satinder Singh, Michael Kearns, and Yishay Mansour. Nash convergence.
12. Tamer Basar and Geert J. Olsder. Dynamic Noncooperative Game Theory, volume 160 of Mathematics in Science and Engineering.Academic Press Inc. (London) Ltd., London, UK, 1982.
13. Tom M. Mitchell. Machine Learning. McGraw-Hill, New York, NY, 1997.
14. *Vilas Gaikar, Sameer Aziz Lakhani. (2020). 'Demographic Variables Influencing Financial Investment Of Urban Individuals: A Case Study Of Selected Districts Of Maharashtra State', International Journal of Advanced Science and Technology, 29(05), Pp.962 – 974.*
15. Ville J. K on open. Asymmetric multi agent reinforcement learning. In Proceedings of the 2003 WIC International Conference on Intelligent Agent Technology (IAT-2003), Halifax, Canada, 2003. IEEE Press.
16. Ville J. K ¨ on ¨ onen. Gradient based method for symmetric and asymmetric multi agent reinforcement learning. In Proceedings of the Fourth International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2003), Hong Kong, China, 2003. Springer-Verlag.
17. X. Wang and T. Sandholm, Learning near-Pareto-optimal conventions in polynomial time, in: Advances in Neural Information Processing Systems, (Vol. 16), Cambridge, MA, 2003.MIT Press.

18. X. Wang and T. Sandholm, Reinforcement learning to play an optimal Nash equilibrium in team Markov games, in Advances in Neural Information Processing Systems, volume 15, Cambridge, MA, 2002. MIT Press.