Final Project Proposal:  
MARL with Asymmetric zero-sum game

Eliav Shalelashvili  
Project supervised by Dr Yehudit Aperstein  
Afeka

ABSTRACT

In this project we are facing against a problem that described by the atmosphere of the Plants vs. Zombies game, A tower defense and strategy video game. The work we are going to discuss in this paper has arisen from and done in collaboration with one of the Israeli Aerospace Industries divisions. We study online reinforcement learning in Stochastic Games (SGs). Our goal is to achieve the highest score in a problem of two player zero sum game in a Markov environment where the transitions and one step payoffs are determined simultaneously.

INTRODUCTION

One of the primary goals of the field of artificial intelligence (AI) is to produce fully autonomous agents that interact with their environments to learn optimal behaviors, improving over time through trial and error. Crafting AI systems that are responsive and can effectively learn has been a long-standing challenge, ranging from robots, which can sense and react to the world around them, to purely software-based agents, which can interact with natural language and multimedia.  
Imagine a board of zombies approaching from some locations in the left side of the board towards the right side. Above all that, there is a light that can be positioned anywhere on the board.  
The two agents will be called 'Zombie Master' and 'Light Master' the Zombie Master is responsible of positioning the zombies in the left side and determine their initial angle and speed that will stay constant for each zombie. On the other hand, the Light Master decides where to project his light in every turn. Each zombie that leaves the left side of the board and goes under the light of the Light Master is damaged and his strength meter is lowered by some value.  
In general, the goal of the Zombie Master/Light Master is to maximize/minimize the strength of the zombies that are reaching the right side of the board.   
We will use some variants of traditional RL methods and examine their profit to our work.

LITERATURE REVIEW

The area of learning agents that master a particular game or on the other side, agents that seek the highest score over a set of games, grew to huge scales in the past few years. Since we are not facing with a studied problem nor a known game, we will divide our review into five sections:

1. Reinforcement Learning (RL) – review of some traditional and relevant RL algorithms and concepts
2. Stochastic Games
3. Nash Equilibrium in Stochastic Games
4. Learning in Stochastic Games
5. The GVG-AI competition

All along with elaboration of the potential contribution of each topic to our research due to the successes of similar problems and previous research of the domain.

1. Reinforcement Learning

Reinforcement learning is an area of Machine Learning. It is about taking suitable action to maximize reward in a particular situation. Essentially an agent (or several) is built such that it can perceive and interpret the environment in which is placed, furthermore, it can take actions and interact with it.  
Basic reinforcement learning problems are modeled as a Markov Decision process (MDP) which is a 4-tuple , where:

* is a finite set of states.
* is a finite set of actions.
* is the probability that action  in state at time will lead to state , due to action .
* is the immediate reward (or expected immediate reward) received after transitioning from state to state , due to action .

The goal is to learn a policy that maximizes the cumulative sum of discounted rewards

where are the rewards and is a discount factor, tuning parameter through which we can influence the amount of weight we give to future awards in relation to the immediate reward.

We can split the subject of RL into two main partitions: ***Model-Free*** and ***Model-Based***. In Model-Free RL,  
the agent does not have access to a model of the environment (The agent couldn’t estimate the consequences of his actions). In Model-Based RL, the agent has access to a model of the environment.  
Our focus is on the Model-Free type of learning mainly due to the advantage that it doesn’t require a model of the environment.

The Model-Free learning can be considered as two parts of ***off-policy***learning and ***on-policy***learning. an agent might be acting using one or two control policies. In *on-policy* control the agent is evaluating and simultaneously improving the exact policy that it follows. Conversely, in *off-policy* control, the agent is following one policy, but may be evaluating another – it is following a behavior policy while evaluating a target policy. In our work we will implement some *off-policy* algorithms alongside an algorithm from the tree search area called MCTS for comparison and evaluation.

1. Stochastic Games

In this paper, two-player zero-sum Stochastic Games (SGs) are considered. These games proceed like MDPs, with the exception that in each state, both players select their own actions simultaneously, which jointly determine the transition probabilities and their rewards. The zero-sum property restricts that the two players’ payoffs sum to zero.

A *Stochastic Game* (SG) is a tuple , Where:

* N is the number of the players/agents
* T: is the transition function
* is the action set for the player
* is the discount factor
* : is the reward function for player

The objective of the n agents is to find a deterministic joint policy (aka. joint strategy aka. strategy profile) (where ) so as to maximize the expected sum of their discounted payoffs. The Q-function, , is the expected sum of discounted payoffs given that the agents play joint action in state and follow policy thereafter. The optimal -function ­ is the -function for (each) optimal policy . So, captures the game structure. The agents generally do not know in advance. Sometimes, they know neither the payoff structure nor the transition probabilities.

For example, consider a zero-sum game with two players, one player (Player 1) wants to maximize his/her total reward, the other (Player 2) would like to minimize that amount. Similar to the case of MDPs, the reward can be discounted or undiscounted, and the game can be episodic or non-episodic.

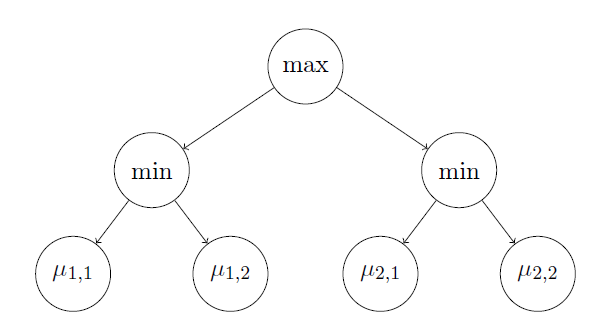


Figure 1 – Game tree when there are two actions by player

We consider a two-player two-round zero-sum game, in which player A has available actions. For each of these actions, indexed by , player B can then choose among possible actions, indexed by . , when player A chooses action and then player B chooses action j, the probability that player A wins is . We investigate the situation (see *Figure 1* for an example) from the perspective of Player A, who wants to identify a maximin action

Assuming that Player B is strategic and picks, whatever A’s action , the action minimizing , this is the best choice for A.

1. Nash Equilibrium in SGs

"In Game Theory, A Nash Equilibrium is a stable state of a system that involves several interacting participants in which no participant can gain by a change of strategy as long as all the other participants remain unchanged"   
Princeton University

A Nash equilibrium is a joint strategy where each agent’s is a best response to the others. For a stochastic game, each agent’s strategy is defined over the entire time horizon of the game.

Given a with players, a Nash Equilibrium is a tuple of strategies such that for all and ,

Where, is the set of strategies available to agent , And,

.

Is the discounted sum of rewards, with discount factor .

A Nash equilibrium is strict if the inequality above is strict. An optimal Nash equilibrium is a Nash equilibrium that gives the agents the maximal expected sum of discounted payoffs.

In the literature, SGs are typically learned under two different settings, and we will call them online and offline settings, respectively. In the offline setting, the learner controls both players in a centralized manner, and the goal is to find the equilibrium of the game [9]. This is also known as finding the worst-case optimality for each player (a.k.a. maximin or minimax policy). In this case, we care about the sample complexity, i.e., how many samples are required to estimate the worst-case optimality such that the error is below some threshold. In the online setting, the learner controls only one of the players, and plays against an arbitrary opponent [10]. In this case, we care about the learner’s regret, i.e., the difference between some benchmark measure and the learner’s total reward earned in the learning process. This benchmark can be defined as the total reward when both players play optimal policies [3], or when Player 1 plays the best stationary response to Player 2. Some of the above online-setting algorithms can find the equilibrium simply through self-playing.

1. Learning in SGs

Learning in stochastic games can be formalized as a multi-agent reinforcement learning (MARL) problem. we can say that the goal of RL is to learn equilibrium strategies through interaction with the environment.

Our work focuses on competitive settings with partially-observable MARL that has received limited attention throughout the years [2]. There were works include model-free gradient-ascent based method [3][4], simulator-supported methods to improve policies using a series of linear programs [5], Recent scalable methods use Expectation Maximization to learn finite state controller (FSC) policies [6].

The most interesting approach I've found related to our problem of competitive relation between the agents and partial observability framework is described in DEC-HDRQNS [7], that means a Decentralized Hysteretic Deep Recurrent Q-Networks model. Their approach is model-free and decentralized, learning Q-values for each agent. In contrast to policy tables or FSCs, Q-values are amenable to the multi-task distillation process as they inherently measure quality of all actions, rather than just the optimal action.

The proposed approach takes into consideration the concept of Hysteresis (lag) [8].  
Overly-optimistic MARL approaches completely ignore low returns, which are assumed to be caused by teammates’ exploratory actions. This causes severe overestimation of Q-values in stochastic domains.  
Hysteretic Q-learning, instead, uses the insight that low returns may also be caused by domain stochasticity, which should not be ignored. This approach uses two learning rates: nominal learning rate, α, is used when the TD-error is non-negative; a smaller learning rate, β, is used otherwise (where 0 < β < α < 1).

In our work we will take the proposed methods under consideration and will cover some RL algorithms in stochastic games, like:

* Minimax-Q learning [1] (based on linear programing duality)
* Double Deep-Q network
* Monte Carlo Tree Search (MCTS)
* Combination of RL with MCTS
  + MaastCTS2
  + OLMCTS
  + SARSA-UCT
* Decentralized Hysteretic Deep Recurrent Q-Networks model

1. The GVG-AI competition

The General Video Game AI competition (GVGAI) was created in order to test these general agents on a multitude of real-time games (both stochastic and deterministic) under the same conditions and constraints. It has received significant international attention in the seven years it has been running and has allowed for many interesting algorithms to be tested on the large number of problems.

The GVG-AI Competition explores the problem of creating controllers for general video game playing, in such platform, researches have an opportunity to test their agents via participating in the competitions.

The past few years have led to some great RL algorithms like 'MaastCTS' and 'OLMCTS' [11], both based the MCTS [12] algorithm. All of these have proven themselves in the competitions, therefor, might be useful and beneficial to our research.

The platform of GVG-AI letting the competitors test their algorithms on some environments built specially for them (by DeepMind) to challenge and push them to their limits. By going over all the proposed environments, I found some games with a lot of resemblance to us, like in our research, there were two players in a zero-sum stochastic game alongside the fact that the action space of the agents is much like ours, for the illustration in this paper I want to elaborate on two games that might be of our interest:

1. Ghostbusters – a version of the known Atari game with improvements to satisfy the competition demands.
   1. Players:
      1. One player is the ghost
      2. The other is the hunter
   2. Actions:
      1. The ghost can pass through walls and wraps around the level
      2. The hunter shoots missiles and moves faster than the ghost
   3. Goal:
      1. The aim of the ghost is to either avoid dying or catch the hunter
      2. The goal of the hunter is to avoid the ghost that can hurt him and shoot the ghost
   4. Algorithms with best results:
      1. MCTS
      2. [MaastCTS2](https://github.com/DennisSoemers/MaastCTS2/tree/master/Two-Player/src/MaastCTS2)

The game Ghostbusters briefly described above, reminds our game in many manners, the similarity of the actions the agents take, the ghost that is capable of moving around the grid while targeting to catch the hunter, much like our Light Master (described later). And on the other hand, the hunter that shoot missiles in many directions to try and catch the ghost, just like our Zombie Master (described later).

1. Upgrade-X
   1. Environment:
      1. Two players (agents)
      2. Both of the players are located in a two-dimensional square which they can't leave
   2. Actions:
      1. Both of the players have some laser cannons, which they can move around
      2. If they run into a laser, they lose health points
   3. Goal:
      1. The winner is the player that survives or the one with most points at the end of the game
   4. Algorithms with best results
      1. OLMCTS algorithm
      2. SARSA-UCT algorithm
      3. [MaastCTS2](https://github.com/DennisSoemers/MaastCTS2/tree/master/Two-Player/src/MaastCTS2)

Like in the previous game, the Upgrade-X games have a bunch of similarities with our game like, the two agents moving around with laser cannons are responsible of the canon direction and their goal is to maximize their health/strength points until the end of the game, a process that reminds a lot our Zombie Agent (described later) that is responsible to his zombies direction and velocities alongside the goal of maximizing their strength throughout the game.

OUR MODEL - Zombie invasion problem

1. Assumptions:
   1. The time and space are discrete
   2. The system operates in discrete time over a horizon T
   3. The system area is represented by N-by-M grid with integer coordinates
   4. A zombie and marking might take coordinates on the integer grid (cells)
   5. At each time moment a zombie can move one cell in right direction
   6. The mark (light) is represented by a square area A-by-A (A is odd)

Figure 2: Positions example

1. Stochastic two players game where
   1. is the number of players
   2. is the transition function
   3. is the action set for the player
   4. is the discount factor (for now )
   5. is the reward function for player *i*

We deal with stochastic two players zero-sum game, i.e.

with limited information on one side (asymmetric information).

1. Player 1: Zombie master
   1. Objectives: maximize average lifetime of zombies
      1. Sum lifetime of all zombies and average over all game rounds
   2. Action:
      1. Decide coordinate y, where the next zombie should start. Action is an integer number from 0 to N (N size of the board). Therefore
   3. Available information:
      1. A zombie location matrix N-by-M
      2. Each cell in the matrix is 0 (no zombie) or 1 (occupied by zombie)

Denote the collection of variables (i.e. observations, actions) available to player 1 at time t by , where .

Denote a subset of all observations until time t and actions until time t-1 by . Therefore contains a set of t N-by-M matrices and a history of choices from the set

* 1. Player 2: Light master
     1. Objectives: minimize average lifetime of zombies
        1. Sum lifetime of all zombies and average over all game rounds
     2. Actions
        1. (x, y coordinates) where to put the center of the light

Thus

* + 1. Available information:
       1. A tensor (two 2-d matrix)
          1. A zombie location matrix, such that
          2. A zombie strength matrix, with non-empty cells at zombie location, such that

* + - 1. The mark (light) at time t, , i.e. the player’s action
      2. Therefore, the available information at time t is

Denote a subset of all observations until time t and actions until time t-1 by

* 1. Game rules
     1. Played on the board
     2. Discrete clock
     3. Each clock tic
        1. Zombie decides where a new zombie will appear
           1. A new zombie’s hit points is equal to 0:
        2. All previous existing zombies are moved right 1 cell, i.e. if for

In general,

* + - 1. Zombies that go over the right boundary disappear
      2. Each zombie that inside marked region (light) got additional hit of the amount c. Meaning the hit points are increased by c (default value c=1)
      3. Each remaining zombie heal itself by multiplying hit point by (1-epsilon) factor. Thus
      4. Once all hit points are calculated and falling zombies are removed from the board, there is a “kill process” that might remove some zombies from the board
      5. For each zombie a utility function U is calculated based on the hit point. U produce values from [0,1] such that zombie with no hit point get 0 and zombies with large value of hit points get utility close to 1
      6. For each zombie a biased coin is tossed with the probability equal to the value of utility function
      7. If the outcome is positive, zombie is removed from the board
      8. The reward (for zombie master) for the round is computed and its equal to the number of zombies that are still in play
      9. The round ends by plants selecting a new place for the light

Next steps

1. Build the entire environment of the game – June 2020
   1. will contain two agents' possible interactions
   2. will be compatible with the openAI gym framework to enable potential of wider research.
2. Examine the proposed algorithms – July 2020
   1. DQNs – basic and successful
   2. MCTS – from the tree search area – will use for reference
   3. *Combining RL with MCTS – GVG-AI competition (alphaGo\*, MaastCTS2, OLMCTS, SARSA-UCT etc.)*
   4. DEC-HDRQNS – proven successful
3. Test the results over a different utility and reward functions – August 2020
4. Increase degree of simulation precision – September 2020
   1. Consider round markings, finer resolution, continuous coordinates etc.

**REFERENCES**

[1] Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994* (pp. 157-163). Morgan Kaufmann.

[2] Omidshafiei, S., Pazis, J., Amato, C., How, J. P., & Vian, J. (2017, August). Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (pp. 2681-2690). JMLR. org.

[3] Peshkin, L., Kim, K. E., Meuleau, N., & Kaelbling, L. P. (2001). Learning to cooperate via policy search. *arXiv preprint cs/0105032*.

[4] Dutech, A., Buffet, O., & Charpillet, F. (2001, August). Multi-agent systems by incremental gradient reinforcement learning. In *International Joint Conference on Artificial Intelligence* (Vol. 17, No. 1, pp. 833-838). LAWRENCE ERLBAUM ASSOCIATES LTD.

[5] Wu, F., Zilberstein, S., & Chen, X. (2012). Rollout sampling policy iteration for decentralized POMDPs. *arXiv preprint arXiv:1203.3528*.

[6] Liu, M., Amato, C., Anesta, E. P., Griffith, J. D., & How, J. P. (2016, March). Learning for decentralized control of multiagent systems in large, partially-observable stochastic environments. In *Thirtieth AAAI Conference on Artificial Intelligence*.

[7] Omidshafiei, S., Pazis, J., Amato, C., How, J. P., & Vian, J. (2017, August). Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (pp. 2681-2690). JMLR. org.

[8] Matignon, L., Laurent, G. J., & Le Fort-Piat, N. (2007, October). Hysteretic q-learning: an algorithm for decentralized reinforcement learning in cooperative multi-agent teams. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 64-69). IEEE.

[9] Matignon, L., Laurent, G. J., & Le Fort-Piat, N. (2007, October). Hysteretic q-learning: an algorithm for decentralized reinforcement learning in cooperative multi-agent teams. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 64-69). IEEE.

[10] Conitzer, V., & Sandholm, T. (2007). AWESOME: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. *Machine Learning*, *67*(1-2), 23-43.

[11] Gaina, R. D., Couëtoux, A., Soemers, D. J., Winands, M. H., Vodopivec, T., Kirchgeßner, F., ... & Perez-Liebana, D. (2017). The 2016 two-player gvgai competition. *IEEE Transactions on Games*, *10*(2), 209-220.

[12] Vodopivec, T., Samothrakis,dz S., & Ster, B. (2017). On Monte Carlo tree search and reinforcement learning. *Journal of Artificial Intelligence Research*, *60*, 881-936.