

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes
Matrix Games
Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation

Simultaneous Multiagent learning in an adversarial setting

Under the supervision of:
Dr. Amitabha Mukerjee
&
Dr. K. S. Venkatesh

Sandip Gupta (Y6425)
Rohit Vaish (Y6402)

Indian Institute of Technology Kanpur

April 21, 2010

Outline

Motivation

Theoretical
Background

Markov Decision
Processes
Matrix Games
Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation

- 1 Motivation
- 2 Theoretical Background
 - Markov Decision Processes
 - Matrix Games
 - Stochastic Games
- 3 Problem Statement
- 4 Intent
- 5 Algorithm
- 6 Reward Formulation
- 7 Robot Simulation
- 8 Experiments & Results

Motivation

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes
Matrix Games
Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation

- *Multi-agent learning*: ubiquitous phenomenon
- *Adversarial Setting*: Widely studied in research beds like Robot soccer
- *Stochastic game* modeling allows us to work with results from Game Theory and Reinforcement learning disciplines.

Markov Decision Processes (MDPs)

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes

Matrix Games
Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation

MDP is a tuple (S,A,T,R) , where:

- S is the set of states
- A is the set of actions available to the agent
- T is a transition function, mapping $S \times A \times S \rightarrow [0, 1]$. It describes the mapping from the agent initially being in a given state, choosing a particular action and then attaining some other final state, to the probability space. (state-action-state space to probability)
- R is the reward function which maps the action of the agent, given a state, to a reward. Hence the mapping is $S \times A \rightarrow \mathbb{R}$

Deals with single agent interacting with (and learning) in an environment

Matrix Games

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes

Matrix Games
Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation

Matrix Games can be formally represented as a tuple $(n, A_{1\dots n}, R_{1\dots n})$, where

- n is the number of agents involved
- A_i is the set of possible actions available to i^{th} agent
- R_i is the reward of i^{th} agent, depending on the actions of all agents. Hence its a mapping from $\mathbb{A} \rightarrow \mathbb{R}$, where \mathbb{A} is the joint action space $A_1 \times \dots \times A_n$

Deals with multiple agents involved in decision making and the reward of any agent depends on the joint action of all other agents.

Stochastic Games

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes
Matrix Games
Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation

It can be described as tuple $(n, \mathbb{S}, A_{1\dots n}, T, R_{1\dots n})$, where

- n is the number of agents
- \mathbb{S} is the set of states of the environment
- A_i is the set of actions of i^{th} agent
- T is the transition function mapping the probabilities of an action in given state, and can be written as $\mathbb{S} \times \mathbb{A} \times \mathbb{S} \rightarrow [0, 1]$
- R_i is the reward function for the i^{th} agent, mapping $\mathbb{S} \times \mathbb{A} \rightarrow \mathbb{R}$

It is an extension of MDPs to multiple agent setting alternately, incorporation of non-deterministic state transitions in Matrix Games.

Keepout : The Game

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes
Matrix Games
Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation

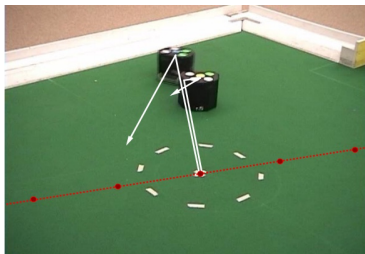
Two adversary teams - Attackers & Defenders

Attacker's Motive: To reach a target zone on the arena, within a fixed time

Defender's Motive: To prevent any attacker from reaching the target zone

Pushing is not allowed

Each such run is limited by a time frame T_0



Questions we intend to address

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes
Matrix Games
Stochastic
Games

Problem
Statement

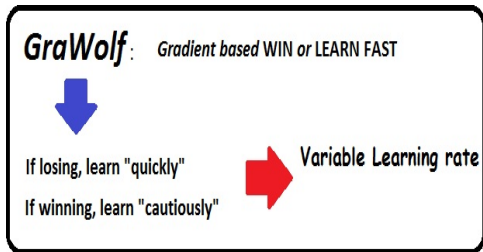
Intent

Algorithm

Reward
Formulation

- Constantly *learning* agents against their *fixed-strategy* counterparts
- Effect of various learning parameters on agents' accomplishment of the task
- Exploring various learning mechanisms, and judging their relative goodness

Implement Gra-WoLF [1] for 2 agent setting



Gra-WoLF combines two ideas from reinforcement learning :
policy gradient learning and **WoLF** variable learning rate.
Extend it to multiple agents and test on simulations

Policy Gradient

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes

Matrix Games

Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation

θ : Policy parameters vector

ϕ_{sa} : Feature Vector for state s and action a

$$\pi_{\theta}(s, a) = \frac{e^{\theta \cdot \phi_{sa}}}{\sum_b e^{\theta \cdot \phi_{sb}}} \quad (1)$$

$$\theta_{k+1} = \theta_k + \delta_k \sum_s \left(d^{\pi_k}(s) \sum_a \frac{\partial \pi_{\theta_k}(s, a)}{\partial \theta_k} \cdot f_{w_k}(s, a) \right) \quad (2)$$

(convergence proven by Sutton *et al* [2])

$$f_{w_k}(s, a) = w_k \cdot \left[\phi_{sa} - \sum_b \pi_{\theta_k}(s, b) \cdot \phi_{sb} \right] \quad (3)$$

Policy Gradient

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes
Matrix Games
Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation

On-Line Learning:

$$\theta_{k+1} = \theta_k + \delta_k \cdot \gamma^t \cdot \sum_a \phi_{sa} \cdot \pi_{\theta_k}(s, a) \cdot f_{w_k}(s, a) \quad (4)$$

Value estimation: SARSA \rightarrow

$$e_{k+1} = \lambda \gamma^t e_k + \phi_{sa} \quad (5)$$

Weight update \rightarrow

$$w_{k+1} = w_k + e_{k+1} \alpha_k (r + \gamma^t Q_{w_k}(s', a') - Q_{w_k}(s, a)) \quad (6)$$

$$f_{w_k}(s, a) = Q_{w_k}(s, a) - \sum_a \pi_{\theta_k}(s, a) \cdot Q_{w_k}(s, a) \quad (7)$$

Variable Learning Rate (δ_k) Selection: Win or Learn Fast

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes
Matrix Games
Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation

An agent is considered “winning” if and only if:

$$\sum_a \pi_{\theta_k}(s, a) Q_{w_k}(s, a) > \sum_a \pi_{\bar{\theta}_k}(s, a) Q_{w_k}(s, a) \quad (8)$$

Selecting variable learning rate \rightarrow

$$\delta = \begin{cases} \delta^w, & \text{if winning} \\ \delta^l, & \text{otherwise} \end{cases}$$

Summary of the algorithm

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes
Matrix Games
Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation

1. Let $\alpha \in (0, 1], \delta^w < \delta^l \in (0, 1], \beta \in (0, 1]$ be learning rates. Initialize:

$$w \leftarrow \bar{0}, \theta \leftarrow \bar{0}, \bar{\theta} \leftarrow \bar{0}, e \leftarrow \bar{0}$$

2. Define:

$$\begin{aligned}\pi_{\theta}(s, a) &= \frac{e^{\theta \cdot \phi_{sa}}}{\sum_b e^{\theta \cdot \phi_{sb}}} \\ Q_w(s, a) &= w \cdot \phi_{sa} \\ f_{w_k}(s, a) &= w_k \cdot [\phi_{sa} - \sum_b \pi_{\theta_k}(s, b) \cdot \phi_{sb}]\end{aligned}$$

3. Repeat:

- (a) From state s select action a according to strategy $\pi^{\theta}(s)$
(b) Observed reward = r and next state = s' with time elapsed = t ,

$$\begin{aligned}e_{k+1} &= \lambda \gamma^t e_k + \phi_{sa} \\ w_{k+1} &= w_k + e_{k+1} \alpha_k (r + \gamma^t Q_{w_k}(s', a') - \gamma^t Q_{w_k}(s, a)) \\ \theta_{k+1} &= \theta_k + \delta_k \cdot \gamma^t \cdot \sum_a \phi_{sa} \cdot \pi_{\theta_k}(s, a) \cdot f_{w_k}(s, a)\end{aligned}$$

where,

$$\delta = \begin{cases} \delta^w, & \text{if } \sum_a \pi_{\theta_k}(s, a) Q_{w_k}(s, a) > \sum_a \pi_{\bar{\theta}_k}(s, a) Q_{w_k}(s, a) \\ \delta^l, & \text{otherwise} \end{cases}$$

- (c) Maintain an average parameter vector,

$$\bar{\theta} \leftarrow (1 - \beta) \bar{\theta} + \beta \theta$$

- (d) If s' is s_0 , or trial is over then $t \leftarrow 0$, $e \leftarrow \bar{0}$. Otherwise $t \leftarrow t+1$.

Tile Coding

Tile Coding [2] is popular technique for handling intractable feature spaces by creating a set of boolean feature from a set of continuous features.

An example of multiple, overlapping grid tilings is shown below:

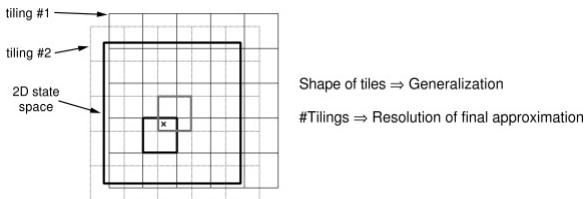


Figure: Tile Coding (Image source [2])

Accuracy v/s Approximation tradeoff!

Proposed Reward function formulation

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes
Matrix Games
Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation

Attacker reward function:

$$R_A = \pm R_0 + \mathbf{f}(\mathbf{r}) - \mathbf{g}(\mathbf{t}) \quad (9)$$

Defender reward function:

$$R_D = \mp R_0 - \mathbf{f}(\mathbf{r}) + \mathbf{g}(\mathbf{t}) \quad (10)$$

Function forms experimented with:

- $f(r) = \frac{(\text{iniAttackerDist} - \text{currAttackerDist})}{\text{iniAttackerDist}}$
- $g(t) = \frac{t}{\text{maxTrials}}$

ZERO SUM GAME \Rightarrow unique Nash Equilibrium

Simulation Snapshot

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes

Matrix Games

Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation

Condition	Hit Count
91 (no condition)	break always (currently 0)
93 (no condition)	break always (currently 0)
97 (no condition)	break always (currently 0)
78 (no condition)	break always (currently 0)
966 (no condition)	break always (currently 0)
99 (no condition)	break always (currently 0)
27 (no condition)	break always (currently 0)
352 (no condition)	break always (currently 0)

Implementation Details

- State (s): The seven dimensional state vector is described as:

- 1 Radial distance of attacker
- 2 Radial distance of defender
- 3 Angle between them
- 4 Attacker velocity magnitude
- 5 Attacker Velocity angle wrt radial line
- 6 Defender Velocity magnitude
- 7 Defender Velocity angle wrt radial line

- Action

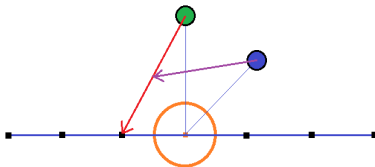


Figure: Action Space

Implementation Details

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes

Matrix Games

Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation

- State-Action (s, a): The state-action space is eight dimensional (Action point is eighth dimension).
- Tiling: Used CMACS tile coding [2] for our simulation to construct the feature vector. 16 tilings and 64 hash bins.
- Reward: A reward of ± 1 is used for attacker/defender depending on who wins the task. (in addition to the proposed formulation)

Methodology contd.

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes

Matrix Games

Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation

Libraries used:

- 1 OpenCV - Image processing module
- 2 NXT++ - interfacing with the Lego robots, converting actuation module output to bluetooth signals
- 3 SDL library - graphical output of the simulation

Experiment 1: Random v/s Learning agents

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes
Matrix Games
Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation

Value estimation learning rates:

$$\alpha_{attacker} = \frac{0.002}{1 + \frac{t}{200000}}$$

$$\alpha_{defender} = \frac{0.1}{1 + \frac{t}{200000}}$$

We obtained the following wins-losses table for the attacker and defender over 150 runs of the game:

Attacker-Defender	Attacker Wins	Defender Wins
L v/s L	0.42	0.58
L v/s R	0.48	0.52
R v/s L	0.39	0.61

Table: Keepout Performance Table

Conclusion

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes

Matrix Games

Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation

- Learning agents far outperform those that make random action selection.
- We also realize that the formulation of the proposed reward function tends to favor the defender in this particular case.

Experiment 2: Effect of Value estimation learning rate on performance

The following graph was obtained from this experiment:

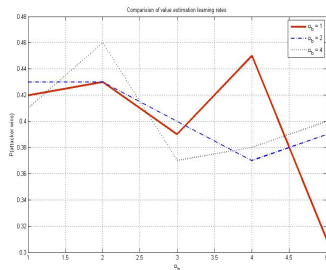
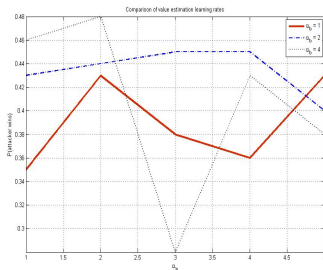


Figure: Comparing Value estimation learning rates α for (a) rewards = R_0 and (b) our reward formulation

other parameters:

attacker speed = 15, defender speed = 10 units. Each time a sequence of 150 trials is conducted.

Conclusion

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes
Matrix Games
Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation

- Neither agent's performance is directly related to increase in its value estimation learning rate.
- This points to the fact that plainly increasing this rate might not essentially improve the performance of the agent over an extended sequence of trials.

Experiment 3: Effect of speeds on performance

attackerDot.speed = A, defenderDot.speed = 5 * B

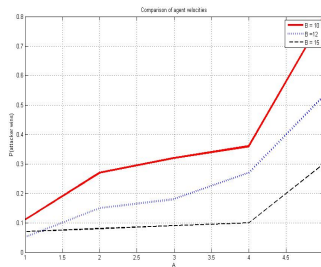
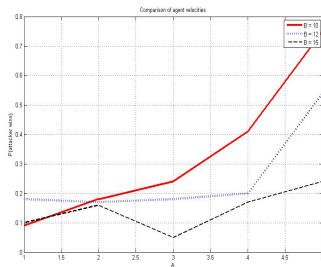


Figure: Comparing attacker and defender speeds for (a) rewards = R_0 and (b) our reward formulation

Value estimation rates:

$$\alpha_a = 4, \alpha_b = 2$$

Conclusion

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes
Matrix Games
Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation

This section provides a straightforward and much logical explanation of the fact that increase in attacker's speed **MUST** be reciprocated by a concomitant increase in defender's speed. Obviously, to chase a fast moving agent, another quick agent is required.

Experiment 4: Effect of initial angle on performance

Simultaneous Multiagent learning in an adversarial setting

Sandip Gupta (Y6425)
Rohit Vaish (Y6402)

Outline

Motivation

Theoretical Background

Markov Decision Processes
Matrix Games
Stochastic Games

Problem Statement

Intent

Algorithm

Reward Formulation

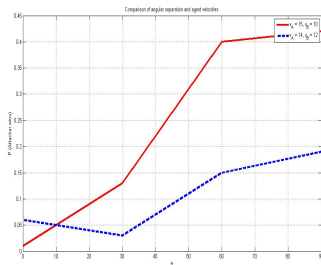
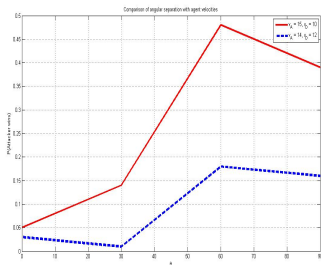


Figure: Comparing Angular separation and agent velocities for (a) rewards = R_0 and (b) our reward formulation

Parameters:

- attackerDot.speed= 15, defenderDot.speed= 10
- $\alpha_a = 2, \alpha_b = 4$
- Number of episodes = 100

Conclusion

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes

Matrix Games

Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation

This experiment provides two important insights:

- 1 A decrease in the relative speeds of the agents tends to favor the defender over a wide range of initial angular separations.
- 2 For a larger value of initial angular separation, the attacker tends to gain from our reward function formulation in comparison to that of [1].

Conclusion

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes

Matrix Games
Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation

We can conclude that

- Learning agents are far superior in task realization as compared to random agents
- Experiments with different learning parameters and initial settings provide us some very basic and logical conclusions, thereby stressing on the genuinity of our reward function formulation and the learning scheme as a whole

References

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes
Matrix Games
Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation



M. Bowling and M. Veloso.

Simultaneous adversarial multi-robot learning.

In *IJCAI'03: Proceedings of the 18th international joint conference on Artificial intelligence*, pages 699–704, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc.



R. S. Sutton and A. G. Barto.

Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning).

The MIT Press, March 1998.

Simultaneous
Multiagent
learning in an
adversarial
setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical
Background

Markov Decision
Processes

Matrix Games

Stochastic
Games

Problem
Statement

Intent

Algorithm

Reward
Formulation

Thank You

Simultaneous Multiagent learning in an adversarial setting

Sandip Gupta
(Y6425)
Rohit Vaish
(Y6402)

Outline

Motivation

Theoretical Background

Markov Decision Processes

Matrix Games

Stochastic Games

Problem Statement

Intent

Algorithm

Reward Formulation