



INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



Minimax Value Iteration Applied to Robotic Soccer

Gonçalo Neto
Institute for Systems and Robotics
Instituto Superior Técnico
Lisbon, PORTUGAL



INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



Presentation Outline

- Framework Concepts
- Solving Two-Person Zero-Sum Stochastic Games
- Soccer as a Stochastic Game
- Results
- Conclusions and Future Work



INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



Markov Decision Processes

- Defined as a 4-tuple (S, A, T, R) where:
 - S is a set of states.
 - A is a set of actions.
 - $T: S \times A \times S \rightarrow [0,1]$ is a transition function.
 - $R: S \times A \times S \rightarrow \mathbb{R}$ is a reward function.
- Single-agent / multiple-state markovian environment.
- On an MDP a policy π is:
 - $\pi: S \times A \rightarrow [0,1]$
 - deterministic vs stochastic



INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



Optimality in MDPs

- Maximize expected reward will lead to optimal policies.
- Usual formulation: discounted reward over time.
- State values:

$$V^\pi(s) = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s, \pi \right\}$$

- Bellman Optimality Equation relates state values, for the optimal policy:
 - Optimal policy is greedy...

$$V^*(s) = \max_a \sum_{s'} T(s, a, s') \left[R(s, a, s') + \gamma V^*(s') \right]$$



INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



Dynamic Programming

- There are several Dynamic Programming algorithms.
 - They assume full knowledge of the environment.
 - Not suitable for online learning.
- A popular algorithm is **Value Iteration**.
 - Based on the Bellman Optimality Equation.
 - Iteration expression:

$$V^{k+1}(s) \leftarrow \max_a \sum_{s'} T(s, a, s') \left[R(s, a, s') + \gamma V^k(s) \right]$$



INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



Matrix Games

- Defined as a tuple $(n, A_1 \dots n, R_1 \dots n)$ where:
 - n is the number of players.
 - A_i is the set of actions for player i . A is the joint action space.
 - $R_i: A \rightarrow R$ is a reward function: the reward depends on the joint action.
- Multiple-agent / single-state environment.
- A strategy σ is a probability distribution over the actions. The joint strategy is the strategy for all the players.



INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



Matrix Games: examples

	R	P	S
R	0	-1	1
P	1	0	-1
S	-1	1	0

Player 1

	R	P	S
R	0	1	-1
P	-1	0	1
S	1	-1	0

Player 2

Rock-Paper-Scissors

	T	N
T	2	0
N	5	1

Player 1

	T	N
T	2	5
N	0	1

Player 2

Prisoner's Dilemma



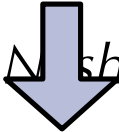
INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



Optimality in MGs

- **Best-Response Function:** set of optimal strategies given the other players current strategies.
- **Nash equilibrium:** in a game's Nash equilibrium all the players are playing a Best-Response strategy to the other players.
- Solving a MG: finding it's Nash equilibrium (or equilibria, because one game can have more than one).
- *All MGs have at least one  Nash equilibrium.*
- Types of Games: zero-sum games, team-games, general-sum games.



INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



Solving Zero-sum Games

- **Two-person Zero-sum games** (or just Zero-sum games) have the following characteristics:
 - Two opponents play against each other.
 - Their rewards are symmetrical (always sum zero).
 - Usually only one equilibrium...
 - ... If more exist they are interchangeable!!
- To find an equilibrium use *Minimax Principle*:

$$\max_{\sigma \in PD(A)} \min_{o \in O} \sum_{a \in A} \sigma(a) R(a, o)$$



INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



Stochastic Games

- Defined as a tuple $(n, S, A_1 \dots n, T, R_1 \dots n)$ where:
 - n is the number of players.
 - S is a set of states.
 - A_i is the set of actions for player i . A is the joint action.
 - $T: S \times A \times S \rightarrow [0, 1]$ is a transition function.
 - $R_i: S \times A \times S \rightarrow R$ is a reward function.
- Multiple-agent / multiple-state environment. Like an extension of MDPs and MGs.
- Markovian from the game's point of view but not from the player.
- The notion of policy can also be defined like in MDPs.



INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



Solving SGs...

- Several Reinforcement Learning and Dynamic Programming algorithms have been derived.
- Normally one type of game is solved.
 - Example: a zero-sum stochastic game is one with two players in which every state represents a zero-sum matrix game.
- A possible approach:
- **Dynamic Programming + Matrix-Game Solver**



INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



Presentation Outline

- Framework Concepts
- Solving Two-Person Zero-Sum Stochastic Games
- Soccer as a Stochastic Game
- Results
- Conclusions and Future Work



INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



Minimax Value Iteration

- Suitable for two-person zero-sum stochastic games.
- Dynamic Programming:
 - Value Iteration.
 - The state values represent Nash equilibrium values.
- Matrix Solver:
 - Minimax in each state.
- Bellman Optimality Equation:

$$V^*(s) = \max_{\pi \in PD(A)} \min_{o \in O} \sum_{a \in A} \pi(a) Q^*(s, a, o)$$

$$Q^*(s, a, o) = \sum_{s'} R(s, a, o, s') + \gamma T(s, a, o, s') V^*(s')$$



INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



If not two-person...

- If the game is not a two-person zero-sum game but...
 - It's a **two team** game.
 - In each team, the reward is the same.
 - The rewards of both teams are symmetrical.
- ...we can consider team actions and apply the same algorithm:
 - $A = A_1 \times A_2 \times \dots \times A_n$
 - $O = O_1 \times O_2 \times \dots \times O_m$



INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



Algorithm Expression

- Based on the Bellman Optimality Equation for Two-Person Zero-Sum Stochastic Games:

$$V^{k+1}(s) \leftarrow \max_{\pi \in PD(A)} \min_{o \in O} \sum_{a \in A} \pi(a) Q^{k+1}(s, a, o)$$

$$Q^{k+1}(s, a, o) \leftarrow \sum_{s'} R(s, a, o, s') + \gamma T(s, a, o, s') V^k(s)$$



INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



Presentation Outline

- Framework Concepts
- Solving Two-Person Zero-Sum Stochastic Games
- Soccer as a Stochastic Game
- Results
- Conclusions and Future Work



INSTITUTO
SUPERIOR
TÉCNICO

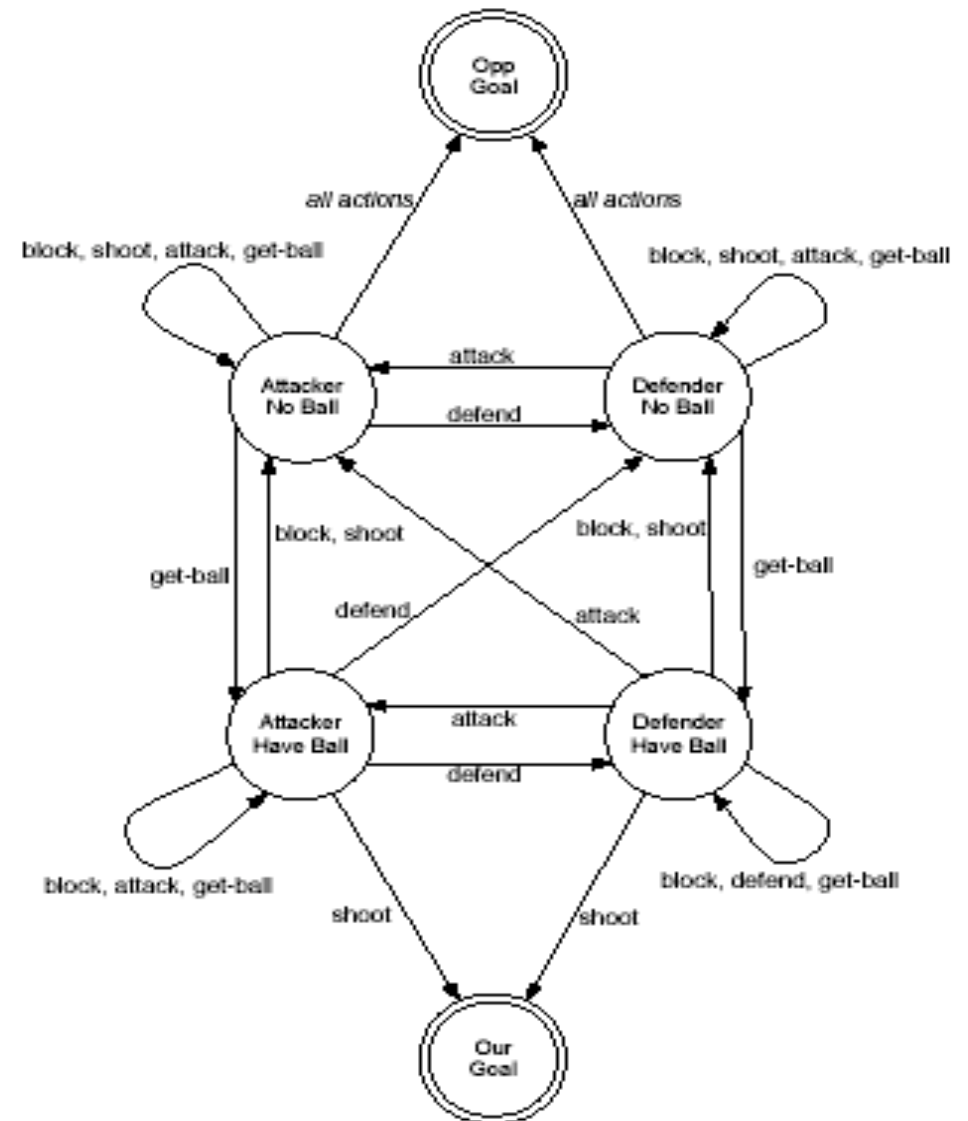


INSTITUTO DE
SISTEMAS E
ROBÓTICA



Modelling a Player

- Non-deterministic automata.
- The output of an action depends on the actions of all players...
- ... the transition probabilities are not stationary.





INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



Modeling the Game

- Symmetrical rewards for both teams.
 - Only received after a goal.
- A set of rules defines the transitions. Examples:
 - IF k players are getting-ball AND none has it ◊ One of them gets it with probability $1/k$.
 - IF a player is changing role ◊ The role is changed with probability 1 and the ball lost with probability p .
 - ...
- Used in simulation:
 - 2 teams of 2 players each
 - Different setups, with some players restricted to just one role.



INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



Presentation Outline

- Framework Concepts
- Solving Two-Person Zero-Sum Stochastic Games
- Soccer as a Stochastic Game
- Results
- Conclusions and Future Work



INSTITUTO
SUPERIOR
TÉCNICO

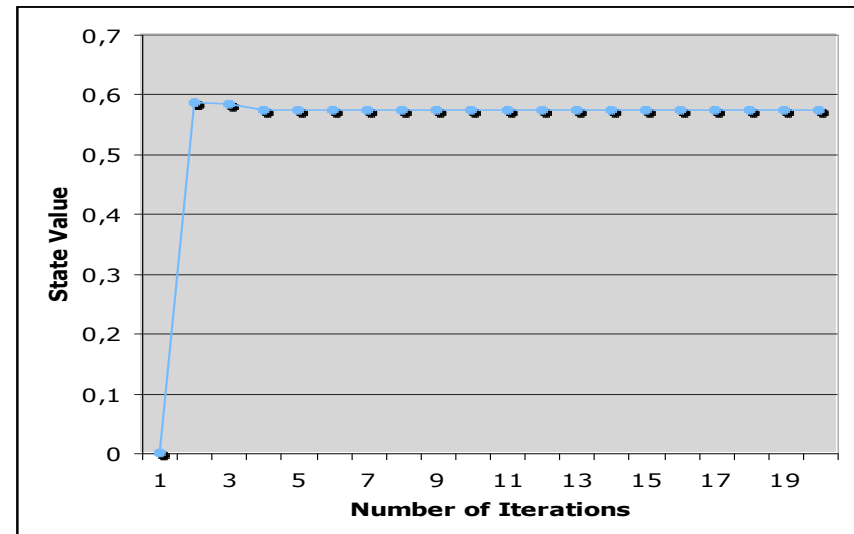


INSTITUTO DE
SISTEMAS E
ROBÓTICA



Method Convergence

- Usually converges fast but...
-for a setup with $\#S=82$ and $\#A=25$ one iteration took more than 30 minutes.
- The graphics are for $\#S=22$ and $\#A=15$.





INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



Simulation after Training

- Used a 10000 step simulation.
 - When a terminal state is reached, the game was put back in the initial state.
- Against another optimal opponent:
 - Only one game played.
 - Finished with a goalless draw.
- Against a random opponent:
 - A team with one pure Attacker scored 2974 against 326.
 - A team with one pure Defender scored 0 against 0.



INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



Presentation Outline

- Framework Concepts
- Solving Two-Person Zero-Sum Stochastic Games
- Soccer as a Stochastic Game
- Results
- Conclusions and Future Work



INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



Conclusions

- The Nash equilibrium convergence assures worst-case optimal.
 - If not possible to score more, assuming worst-case, then keep the draw.
 - Defensive teams tend to just defend
- Method suitable for offline learning.
 - Very time consuming.
 - With a large action set, linear programs slow the method
 - ◊ Efficient LP techniques needed.
- The team action approach only works for small action sets and/or small teams.



INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



Future Work and Ideas

- Observability issues.
 - Should DP assume partial observability?
 - We do we build the game model?
- Suitable learning method depends on other players type.
 - While learning / training locally, learning method could depend on the agent's beliefs about another player.
- Some actions could be discarded.
 - Example: doesn't make sense to choose get-ball while having the ball.
 - Supervisory control for enabling actions that make sense.
 - A way of incorporating knowledge.
 - Can act as a complement to reinforcement learning and dynamic programming.



INSTITUTO
SUPERIOR
TÉCNICO



INSTITUTO DE
SISTEMAS E
ROBÓTICA



Q & A