INSTITUTO SUPERIOR TÉCNICO

INSTITUTO DE SISTEMAS E ROBÓTICA

Multi-agent Learning: Concepts and Methods

Gonçalo Neto
Institute for Systems and Robotics
Intelligent Systems Laboratory

INSTITUTO SUPERIOR TÉCNICO

INSTITUTO DE SISTEMAS E ROBÓTICA

Motivation

- Optimal solutions for the decision problem, particularly in multi-agent systems, are sometimes not obvious to the programmer.
- **So...** Multi-agent reinforcement learning provides a way of programming agents without the complete knowledge of the task.
- **But...** Reinforcement Learning for the single-agent domain can't always be used in a multi-agent scenario.
- **So...** there is the need to study specific reinforcement learning techniques in the presence of other agents.

INSTITUTO SUPERIOR TÉCNICO

INSTITUTO DE SISTEMAS E ROBÓTICA

Outline

- Background
- Best-Response Learners
- Equilibrium Learners
- Other Approaches
- Conclusions

Outline

- Background
- Best-Response Learners
- Equilibrium Learners
- Other Approaches
- Conclusions

INSTITUTO SUPERIOR TÉCNICO

INSTITUTO DE SISTEMAS E ROBÓTICA

INSTITUTO SUPERIOR TÉCNICO

INSTITUTO DE SISTEMAS E ROBÓTICA

Markov Decision Processes (Bellman, 57)

- Single-agent / multi-state framework with no memory: *Markov Property.*
- An Optimality Concept: maximizing expected reward.
  - Usual Formulation: discounted reward over time
  - State Values:

$$V^\pi(s) = E\left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| s_t = s, \pi \right\}$$
$$= \sum_a \pi(s,a) \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^\pi(s') \right]$$

- Reinforcement Learning to find optimal policy.
  - *Q-learning* (Watkins,89) is a possible algorithm:

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right)$$

INSTITUTO SUPERIOR TÉCNICO

INSTITUTO DE SISTEMAS E ROBÓTICA

Matrix Games (von Neumman, 47)

- Matrix Games provide a multi-agent / single-state framework.
- Optimality Concepts in Matrix Games.
  - *Best-Response Function:* set of optimal strategies given the other agents current strategies.

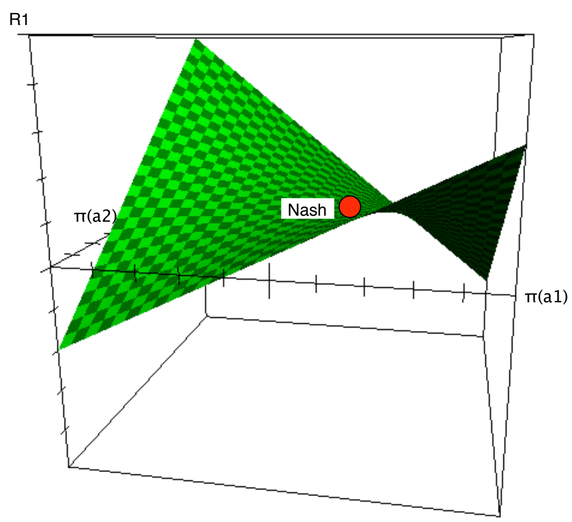
$$\forall_{\sigma_i \in PD(A_i)} \quad R_i(\langle \sigma_i^*, \sigma_{-i} \rangle) \geq R_i(\langle \sigma_i, \sigma_{-i} \rangle)$$

  - *Nash Equilibria (Nash, 50):* All agents are using best-response strategies.

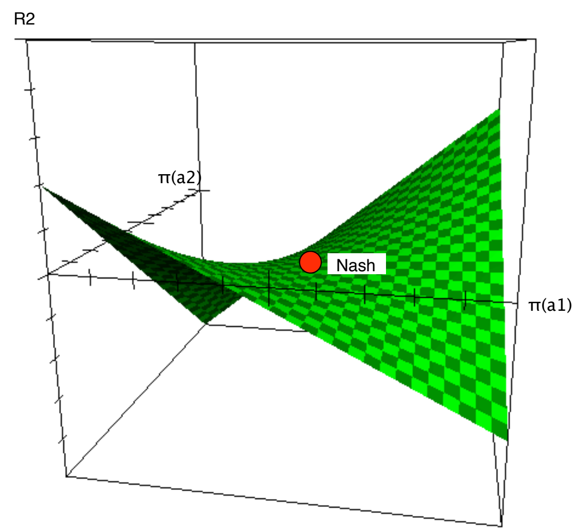
- All Matrix Games have at least one Nash Equilibrium

Game Classification: *Zero-sum*

- 2 players with opposing objectives.
- There is only one Nash equilibrium
  - Minimax to find it.



(a) Reward function for player 1



(b) Reward function for player 2

INSTITUTO SUPERIOR TÉCNICO
INSTITUTO DE SISTEMAS E ROBÓTICA

Two-person Zero-Sum Games

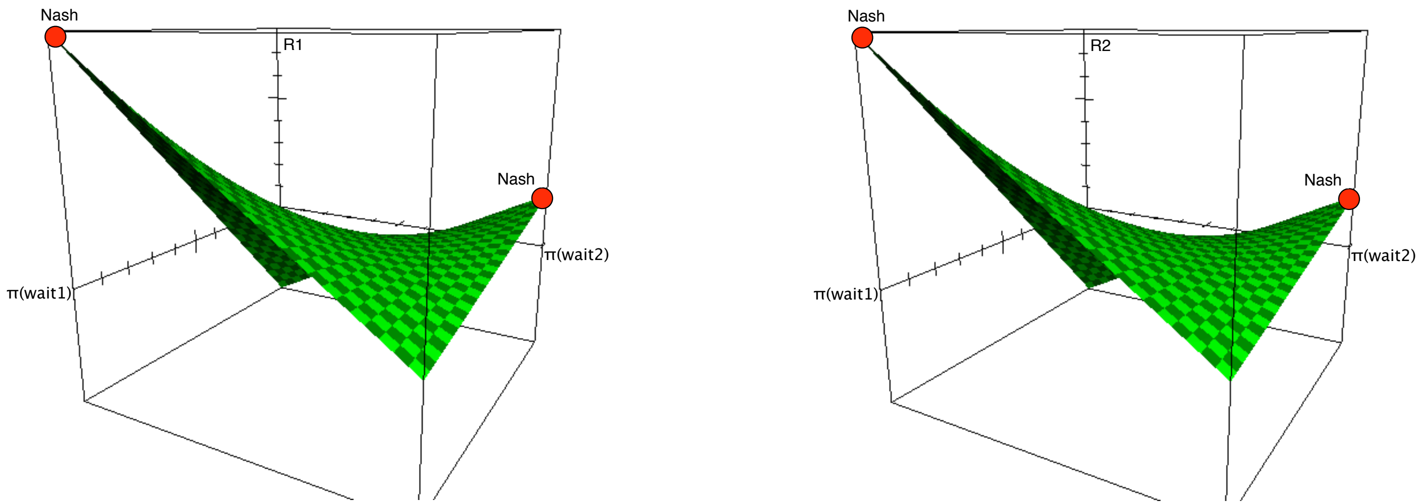
- Characteristics:
  - Two opponents play against each other.
  - Symmetrical rewards (always sum zero).
  - Usually only one equilibrium...
  - ... and if more exist they are interchangeable!
- Minimax to find an equilibrium:

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

  - Formulated as a Linear Program.
  - Solution in the strategy space: simultaneous playing invalidates deterministic strategies.

Game Classification: *Team*

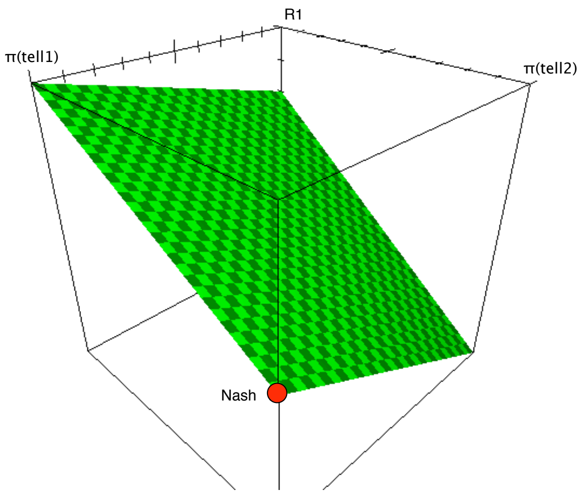
- N players with the same objective.
- Nash equilibria are deterministic.
  - Just look for higher payoffs.



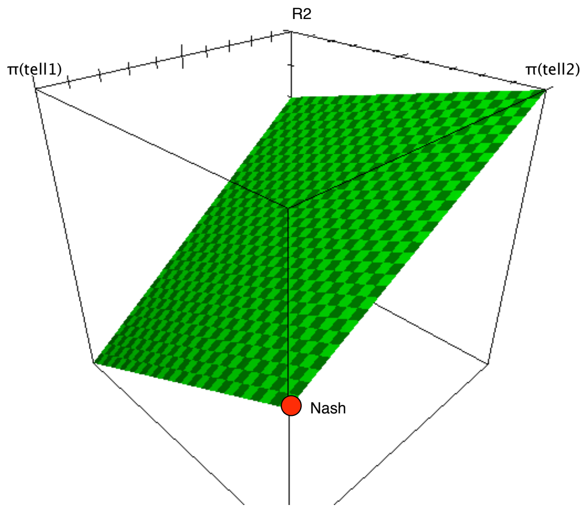
(a) Reward function for player 1 (b) Reward function for player 2

Game Classification: *General-sum*

- All kinds of games.
- Several Nash equilibria requiring complex solutions.
  - With 2 players it is possible to use quadratic programming.



(a) Reward function for player 1



(b) Reward function for player 1

Stochastic Games (Shapley, 53)

INSTITUTO SUPERIOR TÉCNICO

INSTITUTO DE SISTEMAS E ROBÓTICA

- Multiple-state / Multiple-agent environment. Like an extension of MDPs and MGs.
- Markovian but not from each player's point of view.
- Optimality concepts in Stochastic Games:
  - The discounted reward over time is usually considered, as in Markov Decision Processes.

$$V_i^\pi(s) = \sum_a \pi(s,a) \sum_{s'} T(s,a,s') \left[R_i(s,a,s') + \gamma V_i^\pi(s') \right]$$

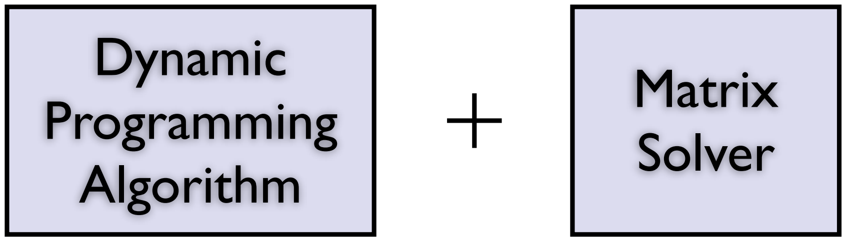
  - *Best-response function:* defined for policies with the state values as reference.

$$\forall_{\pi_i \in S \times PD(A_i)}, \forall_{s \in S} \qquad V_i^{\langle \pi_i^*, \pi_{-i} \rangle}(s) \geq V_i^{\langle \pi_i, \pi_{-i} \rangle}(s)$$

  - *Nash equilibria:* All players are using best-response policies.

Solving Stochastic Games

- Usually, each algorithm solves one type of game.
- A common approach:



INSTITUTO SUPERIOR TÉCNICO

INSTITUTO DE SISTEMAS E ROBÓTICA

INSTITUTO SUPERIOR TÉCNICO

INSTITUTO DE SISTEMAS E ROBÓTICA

Minimax Value Iteration

- Suitable for (two-person) zero-sum stochastic games.

Value Iteration Algorithm	+	Minimax Solver in each State

- Algorithm expression (based on the Bellman optimality equation for the zero-sum SG)

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

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

  - See [Owen, 95] for a convergence proof.
  - Same as Minimax-Q learning algorithm [Littman, 94].

Stationary Opponents

- The game reduces to an MDP with:

$$S^{MDP} = S^{SG}$$

$$A^{MDP} = A_i^{SG}$$

$$T^{MDP}(s, a_i, s') = \sum_{a_{-i} \in A_{-i}^{SG}} \pi_{-i}(s, a_{-i})\, T^{SG}(s, \langle a_i, a_{-i} \rangle, s')$$

$$R^{MDP}(s, a_i, s') = \sum_{a_{-i} \in A_{-i}^{SG}} \pi_{-i}(s, a_{-i})\, T^{SG}(s, \langle a_i, a_{-i} \rangle, s')\, R^{SG}(s, \langle a_i, a_{-i} \rangle, s')$$

INSTITUTO SUPERIOR TÉCNICO

INSTITUTO DE SISTEMAS E ROBÓTICA

Outline

- Background
- Best-Response Learners
- Equilibrium Learners
- Other Approaches
- Conclusions

INSTITUTO SUPERIOR TÉCNICO

INSTITUTO DE SISTEMAS E ROBÓTICA

Best-response learners

- Not specifically concerned with Nash equilibria.
- This methods adapt to the other players trying taking advantage of their weaknesses.
- Three popular approaches:
  - MDP methods.
  - Joint-action learners (JALs) (Claus and Boutilier, 98) and Opponent Modelling (Uther and Veloso, 97).
  - WoLF Policy Hill Climber (Bowling and Veloso, 01).

INSTITUTO SUPERIOR TÉCNICO

INSTITUTO DE SISTEMAS E ROBÓTICA

MDP Methods

- Use reinforcement learning methods for Markov Decision Processes to learn in Stochastic Games: *Q-learning, Sarsa, Actor-critic, ...*
- Some success with this approach (Tan, 93; Sen *et al*, 94).
- **Pros:**
  - Simple implementation.
- **Cons:**
  - Cannot learn stochastic policies (MDP optimal is deterministic).
  - Environment is not stationary from the agent's point of view (MDP methods assume stationarity).

INSTITUTO SUPERIOR TÉCNICO

INSTITUTO DE SISTEMAS E ROBÓTICA

JALs and Opponent Modelling

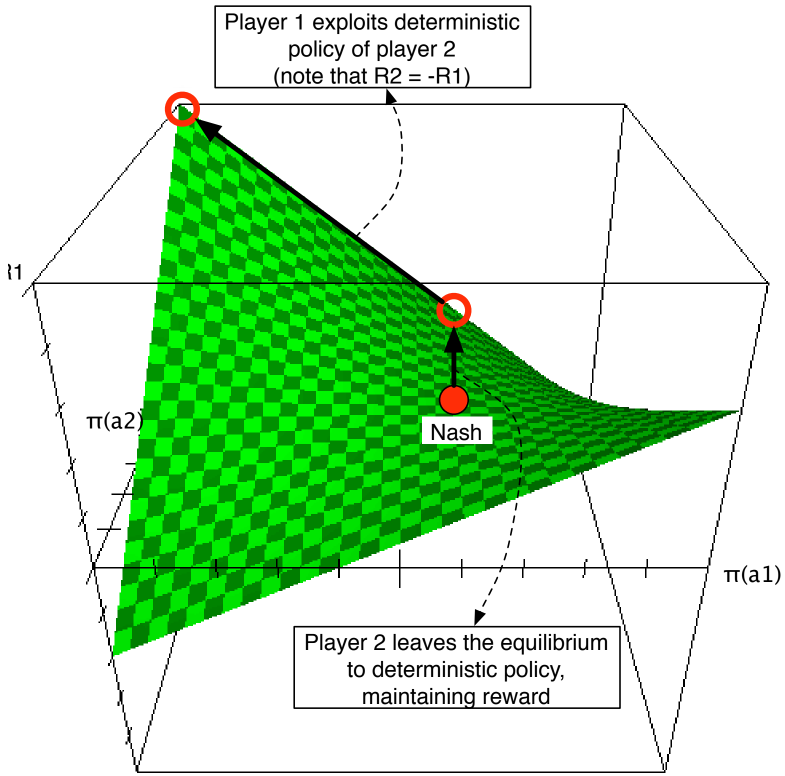
- Learn Q-values based on joint actions.
- Maintain statistics of the opponents actions to compute joint policies.
  - In JALs when deciding, Q-values are replaced by:

$$EV(a_i) = \sum_{a_{-i} \in A_{-i}} Q(\langle a_i, a_{-i} \rangle) \prod_{j \neq i} \widehat{\pi}_j(a_{-i}[j])$$

- **Pros:**
  - Use information of the other players.
- **Cons:**
  - Also learn deterministic policies (max operator).

Deterministic vs Stochastic

- Deterministic policies can be exploited.
  - Most Nash equilibria are stochastic...
- An example:



WoLF Policy Hill Climber

- Modifies the policy directly (Hill Climbing procedure)
- WoLF stands for *Win or Learn Fast*, meaning that the learning rate changes when the agent is winning/ loosing.

$$\sum_{a'} \pi(s, a')Q(s, a') > \sum_{a'} \tilde{\pi}(s, a')Q(s, a')$$

- **Pros:**
  - Can learn stochastic policies.
  - Variable learning rate controls exploration.
  - Converges to Nash when all are playing best-response.
- **Cons:**
  - Assumes convergence to stationary policies of the other agents.

INSTITUTO SUPERIOR TÉCNICO

INSTITUTO DE SISTEMAS E ROBÓTICA

Outline

- Background
- Best-Response Learners
- Equilibrium Learners
- Other Approaches
- Conclusions

Equilibrium Learners

- Specifically try to learn Nash equilibrium policies.
- Basic Idea: each policy is a collection of MG strategies (one for each state) where the reward matrix is defined has:

$$R_i(a) = Q_i^{\pi^*}(s, a)$$

- The Q-values can be computed like Q-learning:

$$\forall_{i=1\ldots n} \quad Q_i(s,a) \rightarrow Q_i(s,a) + \alpha\Big(r_i + \gamma V_i(s') - Q_i(s,a)\Big)$$

  - where the state value V is computes as the Nash equilibrium value for agent i.
- Problem: Several Nash equilibria!!
  - which one to choose??

INSTITUTO SUPERIOR TÉCNICO

INSTITUTO DE SISTEMAS E ROBÓTICA

Minimax-Q (Littman, 94)

- Find Nash equilibria in zero-sum games.
- Nash state values can be found with minimax:

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

  - Can be formulated as a linear program.
- **Pros:**
  - Lower bound for agent performance.
  - Convergence has been prooved – very solid for its domain.
- **Cons:**
  - Large actions spaces lead to big linear programs.

Nash-Q (Hu and Wellman, 98)

- Addresses the problem of learning in 2-player general-sum games.
- Quadratic programming to find Nash state values.
- Several equilibria (which one to choose??). Solved by strict conditions.
- **Pros:**
  - Applicable to a wider range of problems.
- **Cons:**
  - Convergence conditions are too strict and unrealistic
    - All intermediate games must have one equilibrium AND
    - It must be either a saddle point (like zero-sum games) or a global maximum (like team games).

INSTITUTO SUPERIOR TÉCNICO

INSTITUTO DE SISTEMAS E ROBÓTICA

Friend-or-Foe-Q (Littman, 01)

- Motivated by the assumptions of Nash-Q, it is restricted to a class of problems:
  - The agent is either playing against a Foe or with a Friend, and is informed by an external oracle.
- Two different solutions:
  - *Friend:* the game is cooperative and has a Nash at a global maximum – found using **max** operator (like MDPs).
  - *Foe:* the game is adversarial and has a Nash at a saddle point – use **minimax** operator (like Minimax-Q).
- **Pros:**
  - Solid in its domain (no strange convergence conditions).
- **Cons:**
  - When playing *Friend*, might need an oracle too coordinate equilibrium choice (all with the same payoff) – does learning make sense in this situation?

INSTITUTO SUPERIOR TÉCNICO

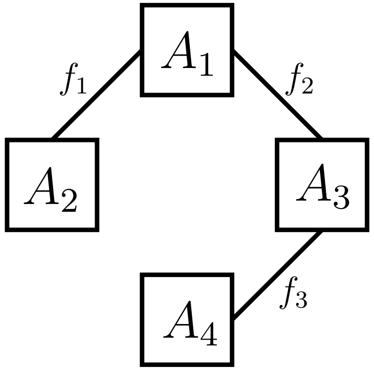
INSTITUTO DE SISTEMAS E ROBÓTICA

Outline

- Background
- Best-Response Learners
- Equilibrium Learners
- Other Approaches
- Conclusions

Other Approaches

- Local Context-Specific Coordination (for team games) (Kok and Vlassis, 04).
  - Coordination graphs for decoupling coordination.
  - Decompose global reward function into sum of functions.



- Beliefs about other agents (Tang and Kaelbling, 03).
  - Agent maintains beliefs about other agent's policies.
  - Converges to a cyclic solution that does better than best-response in average – another solution concept!!

INSTITUTO SUPERIOR TÉCNICO

INSTITUTO DE SISTEMAS E ROBÓTICA

Outline

- Background
- Best-Response Learners
- Equilibrium Learners
- Other Approaches
- Conclusions

INSTITUTO SUPERIOR TÉCNICO

INSTITUTO DE SISTEMAS E ROBÓTICA

Conclusions

- *Best-response learners:*
  - Exploit the environment (including other agents) in the best way they can.
  - Might end up with higher payoffs then Nash equilibria.
  - WoLF-PHC is best suited for learning in the multi-agent domain.
- *Equilibrium Learners:*
  - Provide a lower bound to the performance.
  - Problem with computing and choosing equilibria.
  - Minimax-Q is the most solid in its domain, although FFQ also does well. Nash-Q imposes too strict conditions, although it provides a nice general formulation.
  - There has been some criticism to the equilibrium approach (Shoham *et al,* 04)

INSTITUTO SUPERIOR TÉCNICO

INSTITUTO DE SISTEMAS E ROBÓTICA

Q&A