#### CS 798: Multiagent Systems Multiagent Learning

Kate Larson

Cheriton School of Computer Science University of Waterloo

イロン 不同 とくほ とくほ とう

#### Outline









A B + A B +
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A

→ Ξ → < Ξ →</p>

## Introduction

- So far we have focused on computing optimal/equilibrium strategies
- Another approach: learn how to play a game
  - Play the game many times
  - Update your strategy based on experience
- Why?
  - Some aspect of the game may be unknown to you
  - Other agents may not be playing in equilibrium
  - Computing an optimal strategy is hard
  - Learning is what people do
  - . . .

くロト (過) (目) (日)

## Introduction

- So far we have focused on computing optimal/equilibrium strategies
- Another approach: learn how to play a game
  - Play the game many times
  - Update your strategy based on experience
- Why?
  - Some aspect of the game may be unknown to you
  - Other agents may not be playing in equilibrium
  - Computing an optimal strategy is hard
  - Learning is what people do

• . . .

ヘロン 人間 とくほ とくほ とう

## Introduction

- So far we have focused on computing optimal/equilibrium strategies
- Another approach: learn how to play a game
  - Play the game many times
  - Update your strategy based on experience
- Why?
  - Some aspect of the game may be unknown to you
  - Other agents may not be playing in equilibrium
  - Computing an optimal strategy is hard
  - Learning is what people do
  - . . .

ヘロン 人間 とくほ とくほ とう

#### Challenges

#### • There are other agents in the environment

- Dynamic environment (true in single agent settings)
- What others are learning depend on what our agent is learning
  - Complex global behaviour of the system
- Difficult to separate learning from teaching

	L	R
Τ	1,0	3,2
В	2,1	4,0

ヘロト 人間 ト ヘヨト ヘヨト

э

# Challenges

#### • There are other agents in the environment

- Dynamic environment (true in single agent settings)
- What others are learning depend on what our agent is learning
  - Complex global behaviour of the system
- Difficult to separate learning from teaching

	L	R
Τ	1,0	3,2
В	2,1	4,0

ヘロト 人間 ト ヘヨト ヘヨト

## Challenges

- There are other agents in the environment
  - Dynamic environment (true in single agent settings)
  - What others are learning depend on what our agent is learning
    - Complex global behaviour of the system
  - Difficult to separate learning from teaching

	L	R
Τ	1,0	3,2
В	2,1	4,0

## Challenges

- There are other agents in the environment
  - Dynamic environment (true in single agent settings)
  - What others are learning depend on what our agent is learning
    - Complex global behaviour of the system
  - Difficult to separate learning from teaching

	L	R
Τ	1,0	3,2
В	2,1	4,0

ヘロト 人間 ト ヘヨト ヘヨト

э

## Goals of Multiagent Learning

#### Or What is meant by successful learning?

- No clear answer
- Descriptive Theories
- Prescriptive Theories

くロト (過) (目) (日)

## Goals of Multiagent Learning

Or What is meant by successful learning?

- No clear answer
- Descriptive Theories
- Prescriptive Theories

( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( ) < ( )

A B A B A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A

Typically

- Agents play a normal-form game (the stage game)
- They see what happened (and get the payoffs)
- They play again
- . . .

Can be repeated finitely or infinitely

- Extensive-form game with subgame-perfect equilibrium being repetition of some NE of the stage game
- Are there other equilibria?

ヘロト ヘアト ヘビト ヘビト

Typically

- Agents play a normal-form game (the stage game)
- They see what happened (and get the payoffs)
- They play again
- . . .

#### Can be repeated finitely or infinitely

- Extensive-form game with subgame-perfect equilibrium being repetition of some NE of the stage game
- Are there other equilibria?

ヘロン 人間 とくほ とくほ とう

Typically

- Agents play a normal-form game (the stage game)
- They see what happened (and get the payoffs)
- They play again

• . . .

Can be repeated finitely or infinitely

- Extensive-form game with subgame-perfect equilibrium being repetition of some NE of the stage game
- Are there other equilibria?

ヘロン 人間 とくほ とくほ とう

Typically

- Agents play a normal-form game (the stage game)
- They see what happened (and get the payoffs)
- They play again

• . . .

Can be repeated finitely or infinitely

- Extensive-form game with subgame-perfect equilibrium being repetition of some NE of the stage game
- Are there other equilibria?

イロト イポト イヨト イヨト

-

### Finitely-repeated Prisoners' Dilemma

	С	D
С	2,2	0,3
D	3,0	1,1

- What will the agents do in the last round?
- What will the agents do in the second last round?
- . . .
- What is the equilibrium?

ヘロト ヘ戸ト ヘヨト ヘヨト

#### Finitely-repeated Prisoners' Dilemma



- What will the agents do in the last round?
- What will the agents do in the second last round?
- . . .
- What is the equilibrium?

イロン イボン イヨン イヨン

#### Finitely-repeated Prisoners' Dilemma



- What will the agents do in the last round?
- What will the agents do in the second last round?
- . . .
- What is the equilibrium?

イロト イポト イヨト イヨト

#### Finitely-repeated Prisoners' Dilemma



- What will the agents do in the last round?
- What will the agents do in the second last round?
- . . .
- What is the equilibrium?

ヘロト ヘアト ヘビト ヘビト

#### Finitely-repeated Prisoners' Dilemma



- What will the agents do in the last round?
- What will the agents do in the second last round?
- . . .
- What is the equilibrium?

イロト イポト イヨト イヨト

## Infinitely repeated games

- Utility?
  - If you add up the utility over infinitely many rounds, then everyone gets infinity!
- Limit of average payoff:

$$\lim_{n\to\infty}\sum_{1\leq t\leq n}\frac{u(t)}{n}$$

Discounted payoff:

$$\sum_t \delta^t u(t) \text{ for some } \delta, 0 < \delta < 1$$

イロト 不得 とくほと くほとう

## Infinitely repeated games

- Utility?
  - If you add up the utility over infinitely many rounds, then everyone gets infinity!
- Limit of average payoff:

$$\lim_{n\to\infty}\sum_{1\leq t\leq n}\frac{u(t)}{n}$$

Discounted payoff:

$$\sum_{t} \delta^{t} u(t)$$
 for some  $\delta$ , 0 <  $\delta$  < 1

ヘロト 人間 ト ヘヨト ヘヨト

# Infinitely repeated Prisoners' Dilemma

	С	D
С	2,2	0,3
D	3,0	1,1

Tit-for-tat strategy:

- Cooperate in first round
- In every later round do the same thing that the other player did in the previous round

Trigger strategy:

- Cooperate as long as everyone cooperates
- Once an agent defects, defect forever

# Infinitely repeated Prisoners' Dilemma

	С	D
С	2,2	0,3
D	3,0	1,1

Tit-for-tat strategy:

- Cooperate in first round
- In every later round do the same thing that the other player did in the previous round

Trigger strategy:

- Cooperate as long as everyone cooperates
- Once an agent defects, defect forever

# Infinitely repeated Prisoners' Dilemma

	С	D
С	2,2	0,3
D	3,0	1,1

Tit-for-tat strategy:

- Cooperate in first round
- In every later round do the same thing that the other player did in the previous round

Trigger strategy:

- Cooperate as long as everyone cooperates
- Once an agent defects, defect forever

# Infinitely repeated Prisoners' Dilemma

	С	D
С	2,2	0,3
D	3,0	1,1

Tit-for-tat strategy:

- Cooperate in first round
- In every later round do the same thing that the other player did in the previous round

Trigger strategy:

- Cooperate as long as everyone cooperates
- Once an agent defects, defect forever

## **Fictitious Play**

Early and simply learning rule

- Initialize beliefs about opponent's strategy
- Repeat
  - Play a best-response to assessed strategy of opponent
  - Observe opponent's actual play and update beliefs accordingly

Note that agent is oblivious to the other agent's utilities.

ヘロト 人間 ト ヘヨト ヘヨト

# Properties of Fictitious Play

#### Definition

An action profile a is in steady state if whenever a is played in round t then it is played in round t + 1.

#### Theorem

If a pure strategy profile is a strict NE of a stage game, then it is a steady state of fictitious play in the repeated game.

#### Theorem

If the empirical distribution of each agent's strategies converges in fictitious play then it converges to a Nash equilibrium.

イロト イポト イヨト イヨト

# Properties of Fictitious Play

#### Definition

An action profile a is in steady state if whenever a is played in round t then it is played in round t + 1.

#### Theorem

If a pure strategy profile is a strict NE of a stage game, then it is a steady state of fictitious play in the repeated game.

#### Theorem

If the empirical distribution of each agent's strategies converges in fictitious play then it converges to a Nash equilibrium.

イロト イポト イヨト イヨト

#### **Regret:**

$$R_i(a_i, t) = \frac{1}{t-1} \left[ \sum_{1 \le t' \le t-1} u_i(a_i, a_{-i,t'}) - u_i(a_{i,t'}, a_{-i,t'}) \right]$$

An algorithm has *zero-regret* if or each  $a_i$ , the regret for  $a_i$  becomes non-positive as t goes to infinity (almost surely) against any opponents

くロト (過) (目) (日)

э

**Regret:** 

$$R_i(a_i, t) = \frac{1}{t-1} \left[ \sum_{1 \le t' \le t-1} u_i(a_i, a_{-i,t'}) - u_i(a_{i,t'}, a_{-i,t'}) \right]$$

An algorithm has *zero-regret* if or each  $a_i$ , the regret for  $a_i$  becomes non-positive as t goes to infinity (almost surely) against any opponents

ヘロト 人間 ト ヘヨト ヘヨト

#### Regret matching:

$$\sigma_i^{t+1} = \frac{R^t(a_i)}{\sum_{a' \in A_i} R^t(a')}$$

- Regret matching has zero regret.
- If all players use regret matching, then play converges to the set of *weak correlated equilibria*
- Other types of regret-based learning have different properties

くロト (過) (目) (日)

Regret matching:

$$\sigma_i^{t+1} = \frac{R^t(a_i)}{\sum_{a' \in A_i} R^t(a')}$$

- Regret matching has zero regret.
- If all players use regret matching, then play converges to the set of *weak correlated equilibria*
- Other types of regret-based learning have different properties

ヘロト ヘアト ヘビト ヘビト

-

• Regret matching:

$$\sigma_i^{t+1} = \frac{R^t(a_i)}{\sum_{a' \in A_i} R^t(a')}$$

- Regret matching has zero regret.
- If all players use regret matching, then play converges to the set of *weak correlated equilibria*
- Other types of regret-based learning have different properties

ヘロン 人間 とくほ とくほ とう

• Regret matching:

$$\sigma_i^{t+1} = \frac{R^t(a_i)}{\sum_{a' \in A_i} R^t(a')}$$

- Regret matching has zero regret.
- If all players use regret matching, then play converges to the set of *weak correlated equilibria*
- Other types of regret-based learning have different properties

ヘロン 人間 とくほ とくほ とう

-

# Targeted Learning

#### Assume that there is a limited set of possible opponents

• Try to do well against these

#### **Example**: is there a learning algorithm that

- Learns to best-respond against any stationary opponent (one that always plays the same mixed strategy), and
- Converges to a Nash equilibrium when playing against a copy of itself (self-play)?

イロト イポト イヨト イヨト

# Targeted Learning

- Assume that there is a limited set of possible opponents
  - Try to do well against these

Example: is there a learning algorithm that

- Learns to best-respond against any stationary opponent (one that always plays the same mixed strategy), and
- Converges to a Nash equilibrium when playing against a copy of itself (self-play)?

イロン 不得 とくほ とくほとう

## **Stochastic Games**

- Multiple states  $S = \{S_1, \ldots, S_m\}$ 
  - Each state,  $S_i$  is a normal form game
- After a round, random transition to another state
  - Transition probabilities depend on state and action taken
- Typically discount utilities over time

Note:

- 1-state stochastic game = (infinitely) repeated game
- 1-agent stochastic game = Markov Decision Process (MDP)

・ロト ・ 同ト ・ ヨト ・ ヨト … ヨ

# **Stochastic Games**

A stochastic game is (Q, N, A, P, R) where

- Q is a finite set of states
- $A = A_1 \times \ldots \times A_n$  where  $A_i$  is a finite set of actions available to player *i*
- P: Q × A × Q → [0, 1] is the transition probability function where P(q, a, q') is the probability of transitioning from state q to state q' when joint action a is played
- *R* = *r*<sub>1</sub>,..., *r<sub>n</sub>* where *r<sub>i</sub>* : *Q* × *A* → ℝ is a real valued payoff function for player *i*

▲□▶ ▲□▶ ▲目▶ ▲目▶ 目 のへで

# Review: Q-Learning in Single Agent Settings

- Define Q(s, a) to be the value of taking action a in state s (where by value we mean expected discounted sum of future rewards)
- Optimal policy  $\pi^*(s) = \arg \max_a Q(s, a)$
- Q-Learning algorithm
  - For each s and a initialize Q(s, a)
  - Observe current state
  - Loop
    - Select a and execute it
    - Receive immediate reward r
    - Observe new state s'
    - Update

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

・ロト ・ 同ト ・ ヨト ・ ヨト … ヨ

# Minimax Q-Learning

- Zero-sum stochastic games
- Define Q<sup>π</sup><sub>i</sub>(s, a) to be the value for agent *i* when both agents follow strategy profile π starting in s and a is a joint action.
- Complications
  - Now need to initialize  $\pi$  (typically uniform distribution)
  - Update *Q* value:  $Q_{i,t+1}(s_t, a_{i,t}, a_{-i,t}) = Q_{i,t}(s_t, a_{i,t}, a_{-i,t}) + \alpha(r + \gamma \max_i \min_{a_{-i}} Q_{i,t}(s, \pi_i(s), a_{-i}) Q_{i,t}(s_t, a_{i,t}, a_{-i,t}))$

・ロ・ ・ 同・ ・ ヨ・ ・ ヨ・

• Update  $\pi$ :  $\pi(s, \cdot) = \arg \max_{\pi'(s, \cdot)} (\min_{a'_{-i}} \sum_{a'_i} (\pi'(s, a'_i)Q(s, a')))$ 

#### **Beyond Zero-Sum Stochastic Games**

イロト 不得 とくほ とくほとう