

# Decision Making for Cooperative Agents

## Multiagent MDPs, Decentralized MDPs & POMDPs

Frans Oliehoek  
fao@csail...



6.882: Planning and Decision Making  
November 23, 2010

- 1 Decisions with Multiple Agents
- 2 Global Observations: Multiagent MDPs and POMDPs
  - Recap: Single-agent (PO)MDPs
  - Multiagent MDPs
  - Multiagent POMDPs
- 3 Local Observations: Dec-MDPs & Dec-POMDPs
  - Dec-(PO)MDPs
  - Issues When Acting on Local Observations
- 4 Solution Methods for Dec-POMDPs
  - Backwards Approach
  - Forward Approach
  - The State of the Art
- 5 Summary

# Topic: Multiple Agents

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs  
Multiagent MDPs  
Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs  
Issues

Solving  
Dec-POMDPs

Backwards Approach  
Forward Approach  
The State of the Art

Summary

References

LIS

- Course so far:
  - planning,
  - reinforcement learning (RL),
  - state uncertainty (POMDPs)
- All the above assume a **single** agent interacting with an environment.
- However, if we can build one intelligent agent, **soon we will have many!**
  - **Multiagent system (MAS)**
- Interactions between decision makers: game theory, but focus on:
  - self-interested, and often competitive agents.
  - single-shot interactions.
- This lecture:
  - teams of cooperative agents in a dynamic environment.

# Planning/Learning, on-/off-line

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs  
Multiagent MDPs  
Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs  
Issues

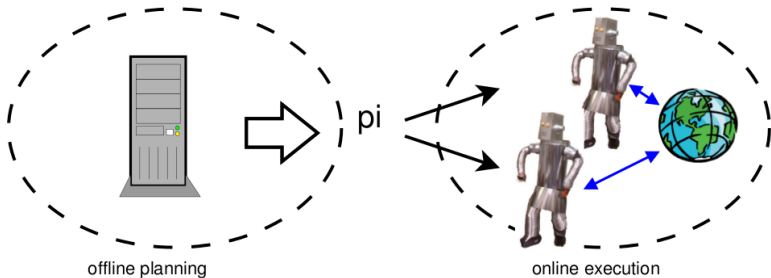
Solving  
Dec-POMDPs

Backwards Approach  
Forward Approach  
The State of the Art

Summary

References

- Focus on planning, but also provide some pointers to reinforcement learning approaches.
- Focus on the situation where
  - planning **off-line**.
  - team of agents executes the plan in an **on-line** phase.



- 1 Agents in the team receive **global observations**.
  - 1 The regular (PO)MDP model be extended to multiple agents (multiagent MDP, POMDP).
  - 2 What assumptions does that require?
  - 3 Why this still requires specialized approaches.
- 2 Agents in the team receive only **local observations**.
  - 1 No longer a reduction to a centralized model. 'Truly' decentralized. (decentralized MDP, POMDP).
  - 2 Agents do not have a Markovian signal to act on!
  - 3 Coordination vs. Exploitation of local information.
- 3 **Solving** Dec-POMDPs:
  - 1 backward approach: dynamic programming.
  - 2 forward approach: heuristic search.

- 1 Decisions with Multiple Agents
- 2 Global Observations: Multiagent MDPs and POMDPs
  - Recap: Single-agent (PO)MDPs
  - Multiagent MDPs
  - Multiagent POMDPs
- 3 Local Observations: Dec-MDPs & Dec-POMDPs
  - Dec-(PO)MDPs
  - Issues When Acting on Local Observations
- 4 Solution Methods for Dec-POMDPs
  - Backwards Approach
  - Forward Approach
  - The State of the Art
- 5 Summary

- 1 Decisions with Multiple Agents
- 2 Global Observations: Multiagent MDPs and POMDPs
  - **Recap: Single-agent (PO)MDPs**
  - Multiagent MDPs
  - Multiagent POMDPs
- 3 Local Observations: Dec-MDPs & Dec-POMDPs
  - Dec-(PO)MDPs
  - Issues When Acting on Local Observations
- 4 Solution Methods for Dec-POMDPs
  - Backwards Approach
  - Forward Approach
  - The State of the Art
- 5 Summary

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs

Multiagent MDPs  
Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs  
Issues

Solving  
Dec-POMDPs

Backwards Approach  
Forward Approach  
The State of the Art

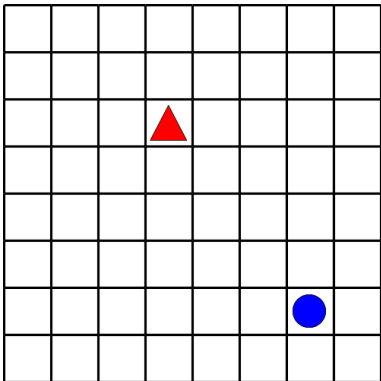
Summary

References

# A single-agent scenario

- Predator-prey

- One agent, the **predator** (blue).
- **Prey** (red) is part of environment.
- Wrap-around world (on a torus)



## Formalization

- states: relative positions.

$$s = (-3,4)$$

- actions: N,W,S,E.
- transitions:
  - probability of failure to move.
  - prey's movements.
- reward for capture.



# The Markov Decision Process

## A Markov Decision Process (MDP) $\langle \mathcal{S}, \mathcal{A}, T, R, h \rangle$

- $\mathcal{S}$  — finite set of states  $s$ .
  - $\mathcal{A}$  — finite set of actions  $a$ .
  - $T$  — transition function, specifying  $P(s'|s,a)$ .
  - $R$  — immediate reward function:  $R(s,a)$ .
  - $h$  — the **horizon** finite or infinite.
- 
- **Policy** maps states to actions  $\pi : \mathcal{S} \rightarrow \mathcal{A}$ .
  - Goal: policy that maximizes the (discounted) **return**.
  - Compute  $\pi^*$ 
    - Value, policy iteration or linear programming:  $V^*$ .
    - From  $V^*$  we can greedily construct  $\pi^*$ .
  - Finite horizon:  $V^{*,\tau+1}(s)$  for  $\tau$  time-steps-to-go.

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs

Multiagent MDPs

Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs  
Issues

Solving  
Dec-POMDPs

Backwards Approach

Forward Approach

The State of the Art

Summary

References

LIS

# Partial Observability (PO)

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs

Multiagent MDPs  
Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs  
Issues

Solving  
Dec-POMDPs

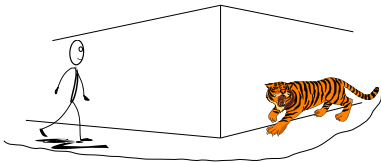
Backwards Approach  
Forward Approach  
The State of the Art

Summary

References

LIS

- **Partially observable world**: not possible to fully determine the state  $\Rightarrow$  **state uncertainty**.
- Two causes:
  - Noise — e.g., distance is approx. 1.5m.
  - Perceptual aliasing — e.g., cannot look around a corner.



# Example PO

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs

Multiagent MDPs  
Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs  
Issues

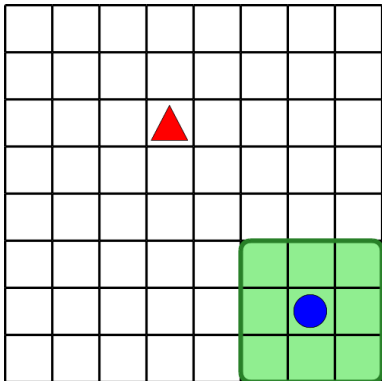
Solving  
Dec-POMDPs

Backwards Approach  
Forward Approach  
The State of the Art

Summary

References

- Single agent predator-prey, with **limited sight**.



- States same as in MDP:
  - $(-8, -8)$  up to  $(8,8)$ .
  - current  $s = (-3,4)$
- But now agent has a different observation:

$o = \text{Null}$

# Example PO

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs

Multiagent MDPs  
Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs  
Issues

Solving  
Dec-POMDPs

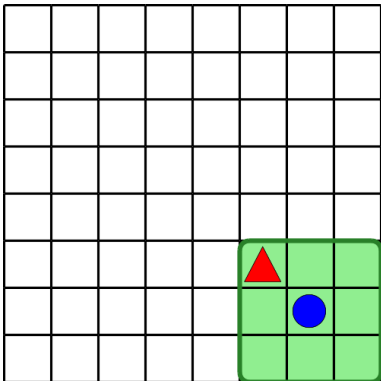
Backwards Approach  
Forward Approach  
The State of the Art

Summary

References

LIS

- Single agent predator-prey, with **limited sight**.



- States same as in MDP:
  - $(-8, -8)$  up to  $(8,8)$ .
  - current  $s = (-3,4)$
- But now agent has a different observation:

$$o = (-1,1)$$

# The POMDP model

$$\text{POMDP} = \langle S, \mathcal{A}, T, R, \mathcal{O}, O, h \rangle$$

- $\mathcal{O}$  — finite set of observations  $o$
- $O$  — observation function, providing  $P(o|a,s')$

- Observations are not a Markovian signal. . .
  - Should remember the entire history of observations?
  - No: we can maintain a **belief**.

$$b = ( \text{Pr}(-8, -8) \quad \text{Pr}(-7, -8) \quad \dots \quad \text{Pr}(7,8) \quad \text{Pr}(8,8) )^T$$

- Reduction to a 'belief-state MDP'.
- So compute  $V^*(b)$ .
  - (how to do this is more complicated, but the principle is the same.)

- 1 Decisions with Multiple Agents
- 2 Global Observations: Multiagent MDPs and POMDPs
  - Recap: Single-agent (PO)MDPs
  - **Multiagent MDPs**
  - Multiagent POMDPs
- 3 Local Observations: Dec-MDPs & Dec-POMDPs
  - Dec-(PO)MDPs
  - Issues When Acting on Local Observations
- 4 Solution Methods for Dec-POMDPs
  - Backwards Approach
  - Forward Approach
  - The State of the Art
- 5 Summary

# Multiagent planning

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs

Multiagent MDPs

Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs

Issues

Solving  
Dec-POMDPs

Backwards Approach

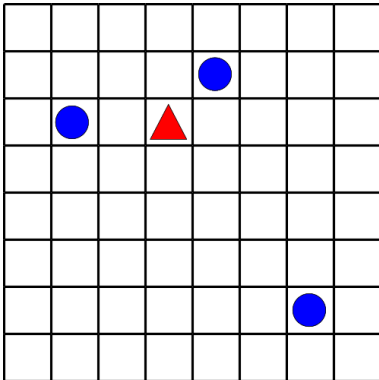
Forward Approach

The State of the Art

Summary

References

- Predator-prey with **multiple predators**.



- State:

$$s = \begin{pmatrix} (3, -4) \\ (1, 1) \\ (-2, 0) \end{pmatrix}$$

- (now with prey as point of reference)

# The Multi-agent MDP

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs

Multiagent MDPs

Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs  
Issues

Solving  
Dec-POMDPs

Backwards Approach

Forward Approach

The State of the Art

Summary

References

- Can be formalized as a **multiagent MDP (MMDP)**.
- MMDP is an MDP with multiple agents
  - $R(s, a_1, \dots, a_n)$
  - $P(s' | s, a_1, \dots, a_n)$
- It is just an MDP but with **joint actions**  $\mathbf{a} = \langle a_1, \dots, a_n \rangle$ .
- Interpretation: '**Puppeteer**' who plans with  $\mathbf{a}$ .
  - $\Rightarrow R(s, \mathbf{a})$ ,
  - $\Rightarrow P(s' | s, \mathbf{a})$



# MMDP is an MDP

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs

Multiagent MDPs

Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs

Issues

Solving  
Dec-POMDPs

Backwards Approach

Forward Approach

The State of the Art

Summary

References

- An MMDP is special case of an MDP
  - under some **assumptions** (next).
  - In practice, term 'MMDP' used when these assumptions hold.
- So MDP solution methods (value iteration, policy iteration, linear programming) apply.
- Also can consider reinforcement learning in MMDPs.
- However:
  - Number of joint actions scales **exponentially** with  $n$ .
  - Need special methods to deal with that [Guestrin et al., 2002a,b, Kok and Vlassis, 2006].

LIS

# MMDP Assumptions

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs

Multiagent MDPs

Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs

Issues

Solving  
Dec-POMDPs

Backwards Approach

Forward Approach

The State of the Art

Summary

References

When can we make the reduction to an MDP?

- Bottom line: the agents need to be able to execute the optimal MDP policy  $\pi(s) = \mathbf{a}$ .
- If  $\pi$  has been computed in an off-line stage, then each agent  $i$  has a copy. Then, **at execution**:
  - Observe  $s$
  - look up  $\pi(s) = \mathbf{a}$
  - execute  $a_i$  the individual component of  $\mathbf{a}$ .

So either

- each agent can **observe  $s$** , or
- agents can **communicate**.
  - noise-free, cost-free, instantaneous broadcast communication!

- 1 Decisions with Multiple Agents
- 2 Global Observations: Multiagent MDPs and POMDPs
  - Recap: Single-agent (PO)MDPs
  - Multiagent MDPs
  - **Multiagent POMDPs**
- 3 Local Observations: Dec-MDPs & Dec-POMDPs
  - Dec-(PO)MDPs
  - Issues When Acting on Local Observations
- 4 Solution Methods for Dec-POMDPs
  - Backwards Approach
  - Forward Approach
  - The State of the Art
- 5 Summary

# Partial observability in MASs

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs  
Multiagent MDPs  
Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs  
Issues

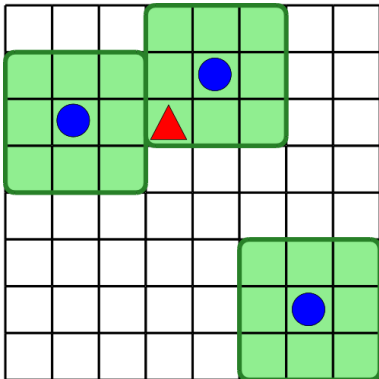
Solving  
Dec-POMDPs

Backwards Approach  
Forward Approach  
The State of the Art

Summary

References

- Predator-prey with predators with restricted sight.
- MAS where each agent gets an individual observation.



- State unchanged:

$$s = \begin{pmatrix} (3, -4) \\ (1, 1) \\ (-2, 0) \end{pmatrix}$$

- But now 3 observations
  - $o_1 = \text{Null}$
  - $o_2 = (-1, -1)$
  - $o_3 = \text{Null}$

# Multiagent POMDPs

We can formalize the problem as follows:

## Multiagent POMDP (MPOMDP)

- $n$  agents.
- $\mathcal{A} = \times_i \mathcal{A}_i$  — set of **joint actions**
  - $\mathcal{A}_i$  — actions of agent  $i$ .
  - $\mathbf{a} = \langle a_1, \dots, a_n \rangle$  one joint action
- $T — P(s'|s, \mathbf{a})$ .
- $R — R(s, \mathbf{a})$
- $\mathcal{O} = \times_i \mathcal{O}_i$  — set of **joint observations**.
  - $\mathcal{O}_i$  observations for agent  $i$ .
  - joint observation  $\mathbf{o} = \langle o_1, \dots, o_n \rangle$
- $O$  — observation function  $P(\mathbf{o}|\mathbf{a}, s')$
- $h$  — the horizon.



Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs

Multiagent MDPs

Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs

Issues

Solving  
Dec-POMDPs

Backwards Approach

Forward Approach

The State of the Art

Summary

References

LIS

# MPOMDP is a POMDP

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs

Multiagent MDPs

Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs

Issues

Solving  
Dec-POMDPs

Backwards Approach

Forward Approach

The State of the Art

Summary

References

LIS

- When agents can communicate observations freely:
  - **Reduction** to a POMDP.
  - Again, term MPOMDP typically used when these assumptions hold.
- Off-line: compute  $V^*(\mathbf{b})$ ,  $\pi^*(\mathbf{b})$ .
- On-line: At each time step,
  - synchronize local observations.
  - compute **joint** belief  $\mathbf{b}$ .
  - look up  $\pi(\mathbf{b}) = \mathbf{a}$
  - execute individual component  $a_i$ .
- Again, scales exponentially with number of agents.
  - still a very much open direction of research.

- 1 Decisions with Multiple Agents
- 2 Global Observations: Multiagent MDPs and POMDPs
  - Recap: Single-agent (PO)MDPs
  - Multiagent MDPs
  - Multiagent POMDPs
- 3 Local Observations: Dec-MDPs & Dec-POMDPs
  - Dec-(PO)MDPs
  - Issues When Acting on Local Observations
- 4 Solution Methods for Dec-POMDPs
  - Backwards Approach
  - Forward Approach
  - The State of the Art
- 5 Summary

Acting based on global information can be **impractical** for several reasons:

- Communication is not possible. . .
  - military domains, space exploration.
- . . . or has a (significant) cost
  - networks, battery power.
- . . . or is not instantaneous, or noise-free, etc.
- Moreover, the required broadcast communication does not scale with the number of agents.

The alternative: **act based on local observations.**

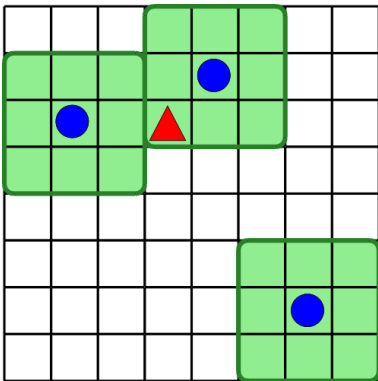
- No communication at all.



- 1 Decisions with Multiple Agents
- 2 Global Observations: Multiagent MDPs and POMDPs
  - Recap: Single-agent (PO)MDPs
  - Multiagent MDPs
  - Multiagent POMDPs
- 3 Local Observations: Dec-MDPs & Dec-POMDPs
  - **Dec-(PO)MDPs**
  - Issues When Acting on Local Observations
- 4 Solution Methods for Dec-POMDPs
  - Backwards Approach
  - Forward Approach
  - The State of the Art
- 5 Summary

# Decentralized MDP and POMDPs

Again, we are considering this setting.



But now:

- No communication
- Act based on local observations only!

- $s = \begin{pmatrix} (3, -4) \\ (1, 1) \\ (-2, 0) \end{pmatrix}$

- Observations

- $o_1 = \text{Null}$
- $o_2 = (-1, -1)$
- $o_3 = \text{Null}$

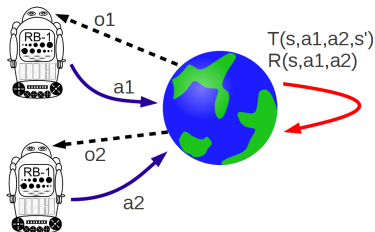
# Decentralized POMDPs

## Dec-POMDP

A **Dec-POMDP** is a **MPOMDP** without (explicit) communication.

- I.e., each agent acts only on its own observations.

- Every  $t$ :  
agent  $i$  observes  $o_i$  and  $a_i$

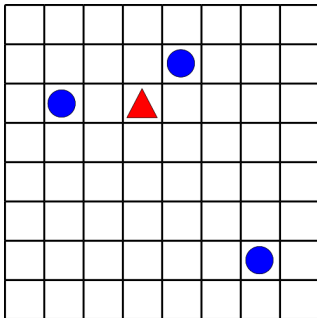


- Communication can be modeled
  - via actions and observations.
  - but is not explicit: **does not allow agents to perform a joint belief update.**

## Dec-MDP

A Dec-MDP is a Dec-POMDP that is **jointly observable**

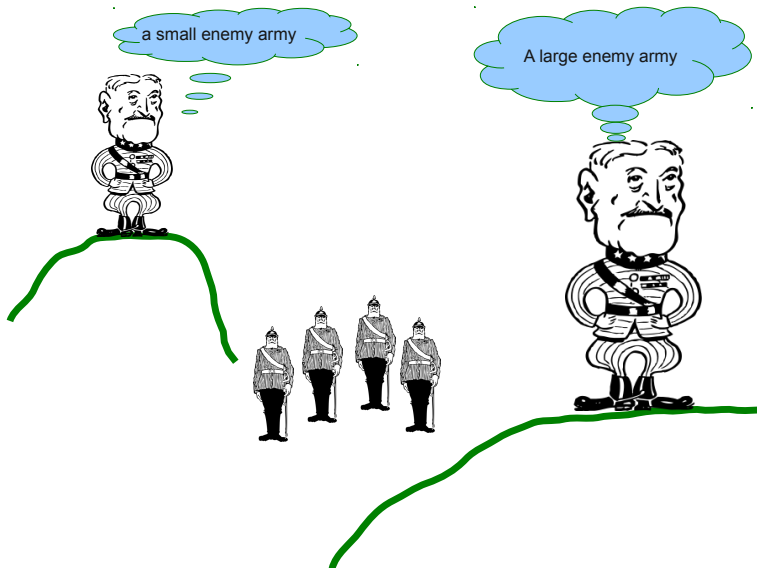
- I.e., the joint observation identifies the state.
- For instance: predator-prey where each agent only observes its own position relative to prey.



$$\begin{pmatrix} o_1 \\ o_2 \\ o_3 \end{pmatrix} = \begin{pmatrix} (3, -4) \\ (1, 1) \\ (-2, 0) \end{pmatrix}$$

Even though special case,  
**not necessarily simpler.**

- “Two generals” or “Coordinated attack” problem.



# Two Generals' problem

- “Two generals” or “Coordinated attack” problem.

## Two generals

2 states: SMALL or LARGE enemy army

2 actions: ATTACK or OBSERVE

2 observations: SMALL or LARGE

Probability of correct observation: 0.85.

Rewards:

- 1 general attacks: it loses the battle  
 $R(*, \text{ATTACK}, \text{OBSERVE}) = -10$
- Both OBSERVE: small cost  
 $R(*, \text{OBSERVE}, \text{OBSERVE}) = -1$
- Both ATTACK: depends on state  
 $R(\text{SMALL}, \text{ATTACK}, \text{ATTACK}) = +5$   
 $R(\text{LARGE}, \text{ATTACK}, \text{ATTACK}) = -20$

- Goal: find a good **joint policy**  $\pi = \langle \pi_1, \dots, \pi_n \rangle$
  - $\pi^*$  maximizes the expected return.
  - What do the policies look like?
  - Mappings from **histories of observations** to actions!
- $$\pi_i(o_i^0, o_i^1, \dots, o_i^t) = a_i$$
- $$\pi_i(\vec{o}_i^t) = a_i$$
- We will see that there is no better way (known) next.

# Goal, Histories & Policies

- Goal: find a good **joint policy**  $\pi = \langle \pi_1, \dots, \pi_n \rangle$
- $\pi^*$  maximizes the expected return.
- What do the policies look like?
- Mappings from **histories of observations** to actions!

$$\pi_i(o_i^0, o_i^1, \dots, o_i^t) = a_i$$

$$\pi_i(\vec{o}_i^t) = a_i$$

- We will see that there is no better way (known) next.



Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs  
Multiagent MDPs  
Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs  
Issues

Solving  
Dec-POMDPs

Backwards Approach  
Forward Approach  
The State of the Art

Summary

References

LIS

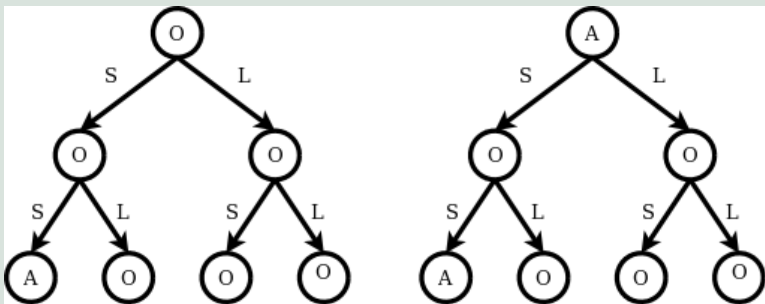


# Goal, Histories & Policies

- Goal: find a good **joint policy**  $\pi = \langle \pi_1, \dots, \pi_n \rangle$
- $\pi^*$  maximizes the expected return.
- What do the policies look like?

## A joint policy — tree representation

Individual policies can be represented as **trees**.



- 1 Decisions with Multiple Agents
- 2 Global Observations: Multiagent MDPs and POMDPs
  - Recap: Single-agent (PO)MDPs
  - Multiagent MDPs
  - Multiagent POMDPs
- 3 Local Observations: Dec-MDPs & Dec-POMDPs
  - Dec-(PO)MDPs
  - **Issues When Acting on Local Observations**
- 4 Solution Methods for Dec-POMDPs
  - Backwards Approach
  - Forward Approach
  - The State of the Art
- 5 Summary

# No Sufficient Statistic

- **No sufficient statistic during execution!**
- Reason from perspective of individual agent  $i$ .
- Assume  $\pi_j$  of other agent  $j$  is **known** and a function of its **internal state**  $l_j$ .
  - Transformation to POMDP.
  - But, need to **predict the actions**  $a_j$ 
    - E.g., to predict state transitions  $P(s'|s, a_i, a_j)$  and rewards  $R(s, a_i, a_j)$ .
  - I.e., need to track the internal state  $l_j$ .
  - Using a belief  $b_i(s, l_j)$ 
    - a sufficient statistic of future behavior of the agent  $j$ .
    - 'individual belief' over states  $b_i(s)$  is not enough.
- When  $\pi_{-i}$  not known: individual belief can not even be computed.



Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs  
Multiagent MDPs  
Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs  
Issues

Solving  
Dec-POMDPs

Backwards Approach  
Forward Approach  
The State of the Art

Summary

References

LIS

# Coordination vs. using Information

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs  
Multiagent MDPs  
Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs  
Issues

Solving  
Dec-POMDPs

Backwards Approach  
Forward Approach  
The State of the Art

Summary

References

- In these problem there is a trade-off: **coordination vs. exploiting local information**.
  - If all agents **ignore own observations**: open loop plan.
    - E.g. "ATTACK on 2nd time step"
    - maximally predictable.
    - but low quality.
  - If all agents base their action on **all their local information** (e.g. compute individual belief and execute MPOMDP policy)
    - potentially higher quality.
    - but less predictable; more likely to result in coordination failures.
- Optimal policy should balance between these.

LIS

Powerful model, but comes at a price. . .

## Dec-(PO)MDP Complexity

- finite-horizon: NEXP-complete [Bernstein et al., 2002]
  - Cast as a decision problem: Guess between EXP possibilities, then need EXP time to verify that it is a solution.
  - Most likely (if  $\text{EXP} \neq \text{NEXP}$ ) doubly exponential time in the worst case.
  - Also for  $\epsilon$ -approximate solutions!
- Infinite-horizon: undecidable
  - just like POMDPs.

# Complexity Results — 2

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs  
Multiagent MDPs  
Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs  
Issues

Solving  
Dec-POMDPs

Backwards Approach  
Forward Approach  
The State of the Art

Summary

References

- What does NEXP mean?

- Brute-force policy evaluation:

$$O \left[ \underbrace{\left( |\mathcal{A}_*| \frac{|\mathcal{O}_*|^{h-1}}{|\mathcal{O}_*|^{h-1}} \right)^n}_{\# \text{ of pure joint policies}} \cdot \underbrace{\left( |\mathcal{S}| \cdot |\mathcal{O}_*|^n \right)^h}_{\text{cost of eval. 1 pol.}} \right]$$

- '\*' denotes largest individual set.

$h$	nr. joint pols.
2	7.290e02
3	4.783e06
4	2.059e14
5	3.815e29
6	1.310e60
7	1.545e121
8	2.147e243

- Still: 1) theoretically interesting, and relevant for 2) for small problems 3) principled approximation methods.

- 1 Decisions with Multiple Agents
- 2 Global Observations: Multiagent MDPs and POMDPs
  - Recap: Single-agent (PO)MDPs
  - Multiagent MDPs
  - Multiagent POMDPs
- 3 Local Observations: Dec-MDPs & Dec-POMDPs
  - Dec-(PO)MDPs
  - Issues When Acting on Local Observations
- 4 Solution Methods for Dec-POMDPs
  - Backwards Approach
  - Forward Approach
  - The State of the Art
- 5 Summary

- We will discuss **general solution methods** for finite-horizon Dec-POMDPs:
  - Brute-force search.
  - Introduction to the two main methods.
- Remember:
  - also works for Dec-MDPs; then observations are local states.
  - general Dec-MDPs are no easier than Dec-POMDPs.
- There exist all kinds of **special cases** with specialized solution methods.
  - TOI-Dec-MDPs, ED-Dec-MDPs, Com-MTDPs, Com-Dec-POMDP, ND-POMDPs, TD-POMDPs etc.



# Off-line Planning

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs  
Multiagent MDPs  
Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs  
Issues

Solving  
Dec-POMDPs

Backwards Approach  
Forward Approach  
The State of the Art

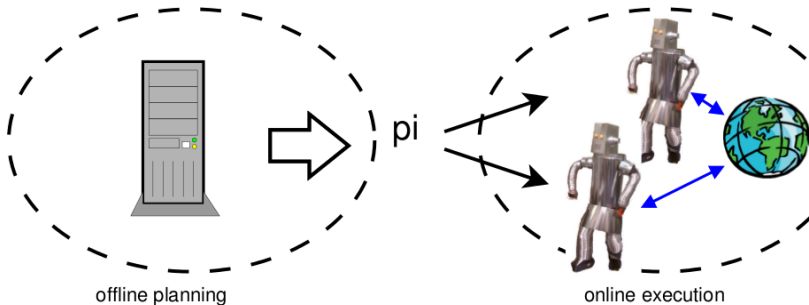
Summary

References

LIS

## Remember:

- planning off-line.
- team of agents executes the plan in an on-line phase.



# Brute Force Search

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs  
Multiagent MDPs  
Multiagent POMDPs

Local  
Observations  
Dec-(PO)MDPs  
Issues

Solving  
Dec-POMDPs

Backwards Approach  
Forward Approach  
The State of the Art

Summary

References

LIS

- The **stupidest algorithm** possible: Brute force search.
  - We only need to consider deterministic joint policies.
  - There are finitely many.
  - So evaluate them all and pick the best.

$$V_{\pi}^t(s^t, \vec{o}^t) = R(s^t, \pi(\vec{o}^t)) + \sum_{s^{t+1} \in \mathcal{S}} \sum_{\mathbf{o}^{t+1} \in \mathcal{O}} \Pr(s^{t+1}, \mathbf{o}^{t+1} | s^t, \pi(\vec{o}^t)) V_{\pi}^{t+1}(s^{t+1}, \vec{o}^{t+1}). \quad (1)$$

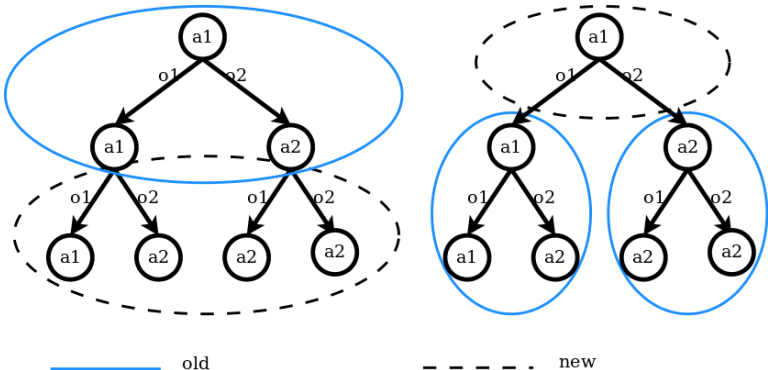
$$V(\pi) = \sum_{s^0 \in \mathcal{S}} V_{\pi}(s^0, \vec{\theta}_{\emptyset}) \mathbf{b}^0(s^0). \quad (2)$$

# Two Main Approaches

Decision Making for Cooperative Agents

Frans Oliehoek  
fao@csail...

- 2 Main approaches: 'forward' and 'backward'



Intro MASs

Global Observations

(PO)MDPs  
Multiagent MDPs  
Multiagent POMDPs

Local Observations  
Dec-(PO)MDPs  
Issues

Solving Dec-POMDPs

Backwards Approach  
Forward Approach  
The State of the Art

Summary

References

LIS

- 1 Decisions with Multiple Agents
- 2 Global Observations: Multiagent MDPs and POMDPs
  - Recap: Single-agent (PO)MDPs
  - Multiagent MDPs
  - Multiagent POMDPs
- 3 Local Observations: Dec-MDPs & Dec-POMDPs
  - Dec-(PO)MDPs
  - Issues When Acting on Local Observations
- 4 Solution Methods for Dec-POMDPs
  - **Backwards Approach**
  - Forward Approach
  - The State of the Art
- 5 Summary

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs  
Multiagent MDPs  
Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs  
Issues

Solving  
Dec-POMDPs

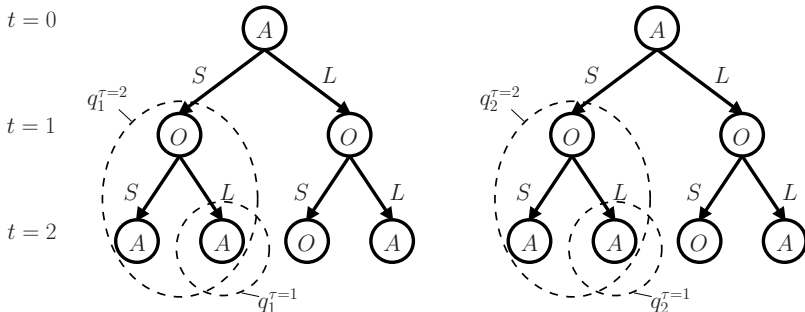
Backwards Approach  
Forward Approach  
The State of the Art

Summary

References

# Backward Approach

- Backward approach: dynamic programming for Dec-POMDPs [Hansen et al., 2004].



- Works on **sub-tree policies**  $q_i^{\tau=k}$ .
- $\tau$  denotes time-to-go.
- Given a policy  $\pi_i$ , and a history  $\vec{o}_i^t$   
 $\Rightarrow$  can find the **implicated sub-tree**  $q_i^{\tau=h-t}$ .

# Determining the Value

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs  
Multiagent MDPs  
Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs  
Issues

Solving  
Dec-POMDPs

Backwards Approach  
Forward Approach  
The State of the Art

Summary

References

Determining **values** for joint sub-tree policies.

Via implied sub-trees  $V_{\pi}^t(s^t, \vec{o}^t)$  translates to

$$V(s^t, q_1^{\tau=h-t}, q_2^{\tau=h-t}) = R(s^t, \mathbf{a}^t) + \sum_{s^{t+1} \in \mathcal{S}} \sum_{\mathbf{o}^{t+1} \in \mathcal{O}} \Pr(s^{t+1}, \mathbf{o}^{t+1} | s^t, \mathbf{a}^t) V(s^{t+1}, q_1^{\tau=h-t}(o_1^{t+1}), q_2^{\tau=h-t}(o_2^{t+1})).$$

where

- $\mathbf{a}^t$  is specified by the roots of  $q_1^{\tau=h-t}, q_2^{\tau=h-t}$ .
- $q_i^{\tau=h-t}(o_i^{t+1})$  is the sub-tree  $q_i^{\tau=h-t-1}$  for  $o_i^{t+1}$ .

# DP for Dec-POMDPs

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs  
Multiagent MDPs  
Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs  
Issues

Solving  
Dec-POMDPs

Backwards Approach  
Forward Approach  
The State of the Art

Summary

References

LIS

- Now have a method for valuating joint sub-tree policies of increasing length.
- If some sub-tree policies are bad: than they will likely result in bad values.
- So BFS may be improved:

DP avoids evaluation of policies with bad sub-tree policies!

- $Q_i^{\tau=k}$  is **set** of  $\tau = k$  sub-tree policies for agent  $i$ .

Initialize:  $\forall_i Q_i^{\tau=1} = \mathcal{A}_i$ .

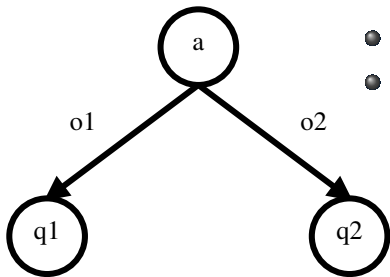
For  $k = 2 : h$

$\forall_i Q_i^{\tau=k} = \text{ExhaustiveBackup}(Q_i^{\tau=k-1})$

Prune dominated  $q_i^{\tau=k} \in Q_i^{\tau=k}$ .

# Exhaustive Backup

- Exhaustive backup: generate  $Q_i^{\tau=k+1}$  from  $Q_i^{\tau=k}$  :



- select an action  $a_i$
- for each observation  $o_i$ , select a  $q_i^{\tau=k} \in Q_i^{\tau=k}$ .

- The number of such policies
  - is exponential in the number of observations.
  - grows doubly exponentially with  $\tau$  (unless a lot can be pruned).

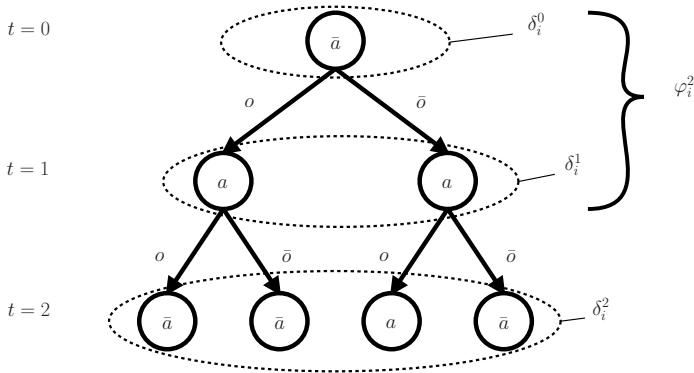


- Many **improvements** have been proposed for the DP algorithm:
  - policy compression
  - point-based DP (PBDP)
    - samples belief points and computes maximizing (non-dominated) sub-tree policies for those
  - memory bounded DP
    - PBDP that maintains a maximum of sub-trees for each agent:  $|Q_i^{\tau=k}| = \text{maxtrees}$
  - improvements to exhaustive backup.

- 1 Decisions with Multiple Agents
- 2 Global Observations: Multiagent MDPs and POMDPs
  - Recap: Single-agent (PO)MDPs
  - Multiagent MDPs
  - Multiagent POMDPs
- 3 Local Observations: Dec-MDPs & Dec-POMDPs
  - Dec-(PO)MDPs
  - Issues When Acting on Local Observations
- 4 Solution Methods for Dec-POMDPs
  - Backwards Approach
  - **Forward Approach**
  - The State of the Art
- 5 Summary

# Forward Approach

- Forward approach: **heuristic search** over partially specified joint policies  $\varphi^t = (\delta^0, \delta^1, \dots, \delta^{t-1})$ .



- Remember: we search over **joint** partial policies. (figure shows an individual policy)

## Multiagent A\* (MAA\*) [Szer et al., 2005]

- Really, it is A\*.
- Search **nodes** correspond to past joint policies

$$\varphi^t = (\delta^0, \delta^1, \dots, \delta^{t-1})$$

- **Heuristic value**

$$F(\varphi^t) = G(\varphi^t) + H(\varphi^t)$$

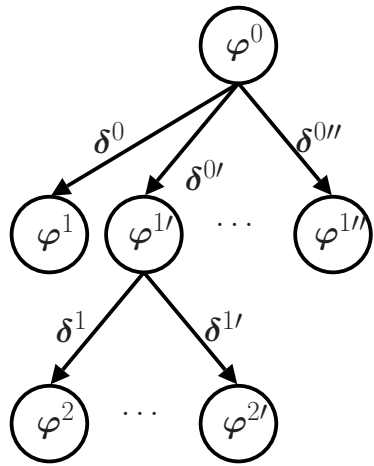
where

- $G(\varphi^t)$  is the true expected reward over the first  $t$  stages.
- $H(\varphi^t)$  is an *admissible* heuristic: optimistic estimate of the reward for the last  $\tau = h - t$  stages.
- **Expanding** a node  $\varphi^t$ , means creating all possible children:

$$\{\varphi^{t+1} = (\varphi^t, \delta^t)\}$$

- **Select** node to expand next based on  $F$ -value.

## MAA\* search tree:



## Improvements to the MAA\*:

- Lossless clustering of histories.
- Do not fully expand the nodes, but only as required.

Both are based on insights gained by interpreting search tree nodes as a 'Bayesian game'.

- It is also possible to compute an **approximate solution** by not doing any backtracking (just expanding 1 child). [Emery-Montemerlo et al., 2005]

- 1 Decisions with Multiple Agents
- 2 Global Observations: Multiagent MDPs and POMDPs
  - Recap: Single-agent (PO)MDPs
  - Multiagent MDPs
  - Multiagent POMDPs
- 3 Local Observations: Dec-MDPs & Dec-POMDPs
  - Dec-(PO)MDPs
  - Issues When Acting on Local Observations
- 4 Solution Methods for Dec-POMDPs
  - Backwards Approach
  - Forward Approach
  - **The State of the Art**
- 5 Summary

# Optimal Solutions

To give an idea, taken from [Oliehoek et al., 2009].

DEC-TIGER 2 states, 3 actions, 2 observations

$h$	$V^*$	$T_{GMAA^*}(s)$	$T_{cluster}(s)$
2	-4.0000	$\leq 0.01$	$\leq 0.01$
3	5.1908	0.02	$\leq 0.01$
4	4.8028	3,069.4	1.50
5	7.0265	-	130.82

BROADCASTCHANNEL 4 states, 3 actions, 2 observations

$h$	$V^*$	$T_{GMAA^*}(s)$	$T_{cluster}(s)$
2	2.0000	$\leq 0.01$	$\leq 0.01$
3	2.9900	$\leq 0.01$	$\leq 0.01$
5	4.7900	-	$\leq 0.01$
25	22.8815	-	1.67

FIREFIGHTING  $\langle n_h = 3, n_f = 3 \rangle$  27 states, 3 actions, 2 observations

$h$	$V^*$	$T_{GMAA^*}(s)$	$T_{cluster}(s)$
2	-4.3825	0.03	0.03
3	-5.7370	0.91	0.70
4	-6.5789	5605.3	5823.5

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs  
Multiagent MDPs  
Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs  
Issues

Solving  
Dec-POMDPs

Backwards Approach  
Forward Approach

The State of the Art

Summary

References

LIS



# Approximate Solutions

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs  
Multiagent MDPs  
Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs  
Issues

Solving  
Dec-POMDPs

Backwards Approach  
Forward Approach

The State of the Art

Summary

References

LIS

- Memory bounded DP (MBDP) is **linear in the horizon**, so scales to arbitrary horizons.
  - solution quality is generally good, but that may be because the problems are too simple.
- Similar for approximate Bayesian game approach by Emery-Montemerlo et al. [2005].
- Recently MBDP was modified into a sampling based approach
  - Works with a simulator.
  - Algorithm (but perhaps not simulator itself) scales linear with the number of agents.
  - results up to **20 agents** and long horizons.
- In my PhD: approximations on value function for factored firefighting: scales to **1000 agents**.

- 1 Decisions with Multiple Agents
- 2 Global Observations: Multiagent MDPs and POMDPs
  - Recap: Single-agent (PO)MDPs
  - Multiagent MDPs
  - Multiagent POMDPs
- 3 Local Observations: Dec-MDPs & Dec-POMDPs
  - Dec-(PO)MDPs
  - Issues When Acting on Local Observations
- 4 Solution Methods for Dec-POMDPs
  - Backwards Approach
  - Forward Approach
  - The State of the Art
- 5 Summary

# Summary

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs  
Multiagent MDPs  
Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs  
Issues

Solving  
Dec-POMDPs

Backwards Approach  
Forward Approach  
The State of the Art

Summary

References

- 1 Agents in the team receive **global observations**.
  - 1 multiagent (PO)MDP is a special cases of normal (PO)MDP.
  - 2 But, requires strong assumptions on observability or communication.
  - 3 Specialized approaches to deal with number of joint actions.
- 2 Agents in the team receive only **local observations**.
  - 1 decentralized MDP, POMDP: 'Truly' decentralized.
  - 2 Several issues: no Markovian signal, Coordination vs. Exploitation of local information, Complexity.
- 3 Solving Dec-POMDPs
  - 1 backward approach: dynamic programming.
  - 2 forward approach: heuristic search.

LIS

# References

- D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein. The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, 27(4):819–840, 2002.
- R. Emery-Montemerlo, G. Gordon, J. Schneider, and S. Thrun. Game theoretic control for robot teams. In *Proc. of the IEEE International Conference on Robotics and Automation*, pages 1175–1181, 2005.
- C. Guestrin, D. Koller, and R. Parr. Multiagent planning with factored MDPs. In *Advances in Neural Information Processing Systems 14*, pages 1523–1530, 2002a.
- C. Guestrin, M. Lagoudakis, and R. Parr. Coordinated reinforcement learning. In *Proc. of the International Conference on Machine Learning*, pages 227–234, 2002b.
- E. A. Hansen, D. S. Bernstein, and S. Zilberstein. Dynamic programming for partially observable stochastic games. In *Proc. of the National Conference on Artificial Intelligence*, pages 709–715, 2004.
- J. R. Kok and N. Vlassis. Collaborative multiagent reinforcement learning by payoff propagation. *Journal of Machine Learning Research*, 7:1789–1828, 2006.
- F. A. Oliehoek, S. Whiteson, and M. T. J. Spaan. Lossless clustering of histories in decentralized POMDPs. In *Proc. of The International Joint Conference on Autonomous Agents and Multi Agent Systems*, pages 577–584, May 2009.
- D. Szer, F. Charpillet, and S. Zilberstein. MAA\*: A heuristic search algorithm for solving decentralized POMDPs. In *Proc. of Uncertainty in Artificial Intelligence*, pages 576–583, 2005.

Decision  
Making for  
Cooperative  
Agents

Frans Oliehoek  
fao@csail...

Intro MASs

Global  
Observations

(PO)MDPs  
Multiagent MDPs  
Multiagent POMDPs

Local  
Observations

Dec-(PO)MDPs  
Issues

Solving  
Dec-POMDPs

Backwards Approach  
Forward Approach  
The State of the Art

Summary

References

LIS