Partially observable MASs

Decentralized POMDPs

Frans Oliehoek
faolieho@...

MASs and distributed AI, April 1st 2008

IAS

Outline

Outline

Outline

Aspects of multiagent system

Relevant aspects:

- 1/multi-timestep?
- Communication?
- Cooperative?
- Learning or planning?
  - on-line / off-line
- Uncertainty?
  - Stochastic environment
  - Observability
  - Other agents
- etc...

Aspects of multiagent system

Relevant aspects:

- 1/multi-timestep?
- Communication?
- Cooperative?
- Learning or planning?
  - on-line / off-line
- Uncertainty?
  - Stochastic environment
  - Observability
  - Other agents
- etc...

This lecture

- Multiple timesteps
- No communication
- Cooperative
- Planning
- Various degrees of uncertainty

Markov Models

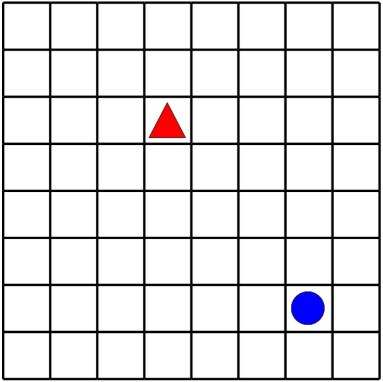
- In particular, we review different 'Markov models'.
  - MDPs, multiagent MDPs, POMDPs.
  - Generalization is the decentralized partially observable Markov decision process (Dec-POMDP).
- First part should be mostly familiar
  - Let me hear if there are questions!

Outline

IAS

Single agent

- Predator-prey with one predator
  - Prey is part of environment.



- States only contain prey position:

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

The Markov Decision Process

- A Markov Decision Process is a framework for single agent planning in a stochastic environment:
  - State of the world can change.
  - Outcome (effect) of actions is uncertain.
  - The state is fully observable.
- Formally:

A Markov Decision Process (MDP)

- is a tuple $\langle \mathcal{S}, \mathcal{A}, T, R, h \rangle$
  - $\mathcal{S}$ — finite set of states s.
  - $\mathcal{A}$ — finite set of actions a.
  - T — transition function, specifying $P(s'|s,a)$.
  - R — immediate reward function
    - $R(s,a)$ gives reward for action a in state s.
  - h — the horizon.

IAS

MDPs — 2

- Policy maps states to actions $\pi : \mathcal{S} \rightarrow \mathcal{A}$

Goal

- Find policy that maximizes the expected cumulative reward, or return:

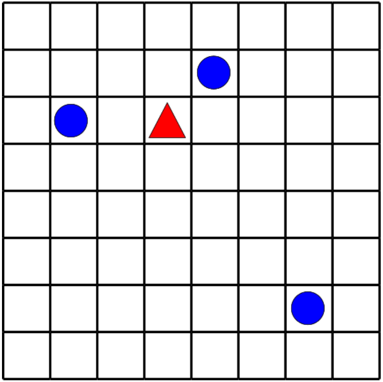
$$E(\pi) = E_{\pi}\left[\sum_{t=0}^{h} \gamma^{h} R^{t}\right]$$

- Optimal value function for τ time-steps-to-go:

$$V^{*,\tau+1}(s) = \max_{a \in \mathcal{A}} \left[R(s,a) + \sum_{s' \in \mathcal{S}} P(s'|s,a) V^{*,\tau}(s') \right]$$

- From V^{*} we can greedily construct π^{*}.

- 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

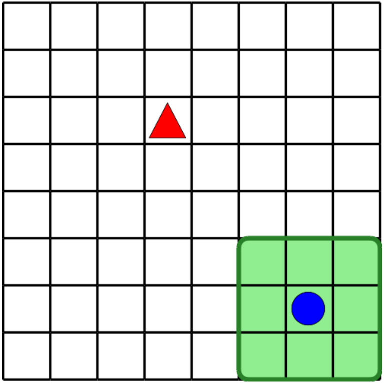
- MMDP is an MDP with multiple agents
  - Cooperative stochastic game (with identical payoffs)
    - related to "coupled learning where agents share the same reward function" [Vlassis, 2007] (But then planning!)
  - $R(s,a_1,\ldots,a_n)$
  - $P(s'|s,a_1,\ldots,a_n)$
- Because state is fully observable, all agents can perform the same reasoning.
- 'Puppeteer' who plans with joint actions, $\mathbf{a} = \langle a_1,\ldots,a_n \rangle$.
  $\Rightarrow R(s,\mathbf{a})$, $P(s'|s,\mathbf{a})$
  $\Rightarrow$Similar to regular MDP
  - Number of joint actions scales exponentially with n.

Partial observability — single agent

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

Example

- Single agent predator-prey, with limited sight.



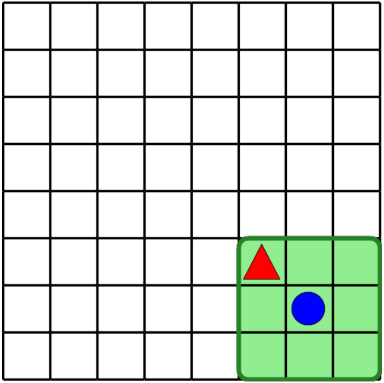
- States same as in MDP:
  - $(-8,-8)$ up to $(8,8)$.
  - current $s = (-3,4)$
- But now agent has a different observation:

$$o = \text{Null}$$

- Planning process becomes much harder!

Example

- Single agent predator-prey, with limited sight.



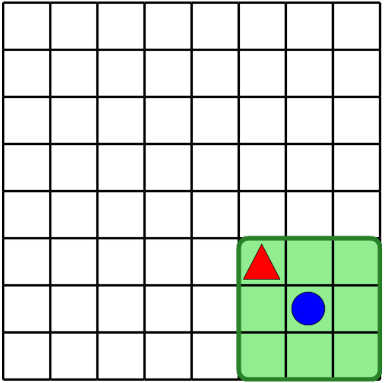
- States same as in MDP:
  - $(-8,-8)$ up to $(8,8)$.
  - current $s = (-3,4)$
- But now agent has a different observation:

$$o = (\text{-}1,1)$$

- Planning process becomes much harder!

Example

- 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:

$$o = (\text{-}1{,}1)$$

- Planning process becomes much harder!

The POMDP model

- Partially Observable MDPs (POMDPs)

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

- $\mathcal{O}$ — 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.*

Belief example

Frans Oliehoek
faolieho@...

- Thinks of a belief as a vector,

$$\begin{pmatrix} P(-8,-8) & P(-7,-8) & \ldots & P(7,8) & P(8,8) \end{pmatrix}$$

- each entry represents to probability of the corresponding state.
- E.g, uniform belief over all 289 $(= 17^2)$ states:

$$\begin{pmatrix} 0.034 & 0.034 & \ldots & 0.034 & 0.034 \end{pmatrix}$$

- Now the agent sees the prey:

$$\begin{pmatrix} 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \end{pmatrix}$$

Belief example

- Thinks of a belief as a vector,

$$\begin{pmatrix} P(-8,-8) & P(-7,-8) & \dots & P(7,8) & P(8,8) \end{pmatrix}$$

- each entry represents to probability of the corresponding state.
- E.g, uniform belief over all 289 ($= 17^2$) states:

$$\begin{pmatrix} 0.034 & 0.034 & \dots & 0.034 & 0.034 \end{pmatrix}$$

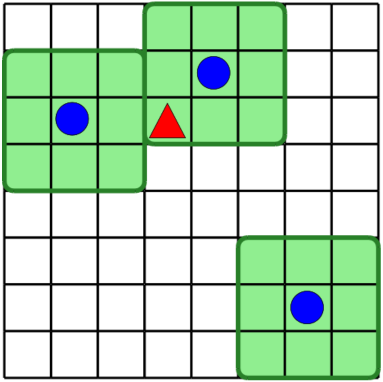
- Now the agent sees the prey:

$$\begin{pmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{pmatrix}$$

Outline

Partial observability in MASs

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



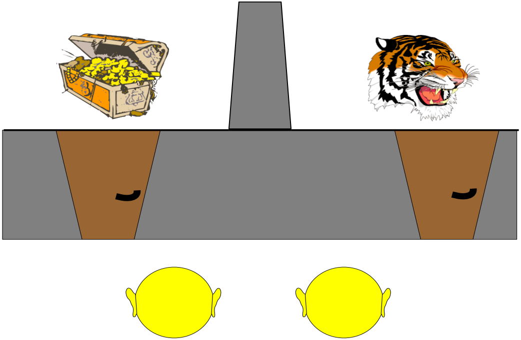
- State unchanged:

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

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

- Remember: no communication! (but still cooperative)
  - Each agent needs to select an action based its individual observation history.

Benchmark problem: Dec-Tiger



The Dec-Tiger problem

- Behind 1 treasure, other: a tiger. Uniform prob. s_l or s_r.
- Agents have 3 actions OpenLeft, OpenRight, Listen.
- Observations: HearLeft, HearRight (informative $\Leftrightarrow \langle Li, Li \rangle$)
- Acting jointly is always better.
- Opening door resets (s_l or s_r with 50% prob.)

Decentralized POMDPs (Dec-POMDPs)

A Dec-POMDP with n agents is a tuple $\langle \mathcal{S}, \mathcal{A}, T, R, \mathcal{O}, O, h \rangle$

- n agents.
- $\mathcal{A} = \times_i \mathcal{A}_i$ — set of joint actions
  - $\mathcal{A}_i$ — actions of agent i.
  - $\mathbf{a} = \langle a_1, \ldots, a_n \rangle$ one joint action
- T — the transition function, now giving $P(s'|s,\mathbf{a})$.
- R — now also dependent on joint actions: $R(s,\mathbf{a})$
- $\mathcal{O} = \times_i \mathcal{O}_i$ — set of joint observations.
  - $\mathcal{O}_i$ observations for agent i.
  - A joint observation $\mathbf{o} = \langle o_1, \ldots, o_n \rangle$
- O — observation function, now $P(\mathbf{o}|\mathbf{a},s')$

Dec-POMDPs—2

- Every t: 1 joint observation $\mathbf{o}$ and 1 joint action $\mathbf{a}$
  - Only observe own observation o_i and action a_i.
- So, in fact it is a cooperative POSG.
  - (partially observable stochastic game)
  - cooperative because rewards are identical.

Benchmark problem: Dec-Tiger formal—1

Dec-Tiger more formally

- $\mathcal{S} = \{s_l, s_r\}$
- $\mathcal{A}_i = \{\text{OpenLeft, OpenRight, Listen}\}$
- $\mathcal{A} = \{\langle \text{OL,OL} \rangle, \langle \text{OL,OR} \rangle, \ldots, \langle \text{Li,Li} \rangle\}$
  (9 joint actions)
- $\mathcal{O}_i = \{\text{HearLeft, HearRight}\}$
- $\mathcal{O} = \{\langle \text{HL,HL} \rangle, \langle \text{HR,HL} \rangle, \langle \text{HL,HR} \rangle, \langle \text{HR,HR} \rangle\}$
  (4 joint observations)

Benchmark problem: Dec-Tiger formal—2

Dec-Tiger more formally—2

- T — transition model. Examples:
  - $P(s_l|s_l, \langle \text{Li,Li} \rangle) = 1$ (listening doesn't change state)
  - $P(s_l|s_l,*) = 0.5$ (reset for other joint actions)
- $\mathcal{O}$ — Observation model. Examples:
  - $P(\langle HL,HL \rangle \,|\, \langle \text{Li,Li} \rangle\, ,s_l) = 0.7225$ (informative)
  - $P(\langle HL,HL \rangle \,|\, *\, ,s_l) = 0.25$ (non-informative)
- R — the reward model. Examples:
  - $R(s_l, \langle OL,Li \rangle) = -101$
  - $R(s_l, \langle OL,OL \rangle) = -50$
  - $R(s_r, \langle OL,Li \rangle) = +9$
  - $R(s_r, \langle OL,OL \rangle) = +20$
  - $R(s_l, \langle Li,Li \rangle) = -2$

Dec-POMDPs — 2

- Joint policy $\pi = \langle \pi_1, \ldots, \pi_n \rangle$ — π_i agent i's policy
  - π_i mapping from sequences of $o_i \in \mathcal{O}_i$ to actions.

Observation History for agent i

- $\vec{o}_i^{\,t} = (o_i^1, \ldots, o_i^t)$ — $\vec{\mathcal{O}}_i$

Deterministic (pure) policy for Dec-POMDPs

- $\pi_i : \vec{\mathcal{O}}_i \to \mathcal{A}_i$.

- Goal is to maximize the expected cumulative reward.
- Cooperative $\Rightarrow$ there is an optimal pure joint policy.
  - General POSG more complicated.

Example policies

- The optimal policy for $h = 3$.

The Dec-Tiger problem policy for agent 1

- π_1 :
  - $() \rightarrow \mathit{Listen}$
  - $(\mathit{HearLeft}) \rightarrow \mathit{Listen}$
  - $(\mathit{HearRight}) \rightarrow \mathit{Listen}$
  - $(\mathit{HearLeft}, \mathit{HearRight}) \rightarrow \mathit{Listen}$
  - $(\mathit{HearRight}, \mathit{HearLeft}) \rightarrow \mathit{Listen}$
  - $(\mathit{HearLeft}, \mathit{HearLeft}) \rightarrow \mathit{OpenRight}$
  - $(\mathit{HearRight}, \mathit{HearRight}) \rightarrow \mathit{OpenLeft}$

- Optimal policy for agent 2 is identical.
  - This is not necessarily the case.

Summary of the Dec-POMDP model

- Cooperative multiagent systems.
- Partially observable environment.
- Stochastic action effects.
- Offline centralized planning.
- Online decentralized execution.
    - No communication.

Outline

IAS

Outline

Naive approach

- Simplest approach: enumerate all pure joint policies and choose the best one.
  - Remember for a POSG this is not even possible!

- Policy evaluation.
- Like MDPs (GPI).
- But also need observations:

Regular MDP

$$V_\pi^t(s) = R(s,\pi(s)) + \sum_{s'} P(s' \mid s,\pi(s)) V_\pi^{t+1}(s')$$

$$V^{t,\pi}(s^t,\vec{\mathbf{o}}^t) = R(s^t,\pi(\vec{\mathbf{o}}^t)) + \sum_{s^{t+1},\mathbf{o}^{t+1}} P(s^{t+1},\mathbf{o}^{t+1} \mid s^t,\pi(\vec{\mathbf{o}}^t)) V^{t+1,\pi}(s^{t+1},\vec{\mathbf{o}}^{t+1})$$

where

- $\vec{\mathbf{o}}^t = \langle \vec{o}_1^t,\dots,\vec{o}_n^t \rangle$ — joint observation history.
- $\pi(\vec{\mathbf{o}}^t) = \langle \pi_1(\vec{o}_1^t),\dots,\pi_n(\vec{o}_n^t) \rangle$ — joint action.
- $\vec{\mathbf{o}}^{t+1} = (\vec{\mathbf{o}}^t,\mathbf{o}^{t+1})$ — appending $\mathbf{o}^{t+1}$ to $\vec{\mathbf{o}}^t$.

Naive approach

- Simplest approach: enumerate all pure joint policies and choose the best one.
    - Remember for a POSG this is not even possible!

- Policy evaluation.
- Like MDPs (GPI).
- But also need observations:

Regular MDP

$$V^t_\pi(s) = R\left(s,\pi(s)\right) + \sum_{s'} P(s' \mid s,\pi(s))\, V^{t+1}_\pi(s')$$

$$V^{t,\pi}(s^t,\vec{\mathbf{o}}^{\,t}) = R\left(s^t,\pi(\vec{\mathbf{o}}^{\,t})\right) + \sum_{s^{t+1},\mathbf{o}^{t+1}} P(s^{t+1},\mathbf{o}^{t+1} \mid s^t,\pi(\vec{\mathbf{o}}^{\,t}))\, V^{t+1,\pi}(s^{t+1},\vec{\mathbf{o}}^{\,t+1})$$

where

- $\vec{\mathbf{o}}^{\,t} = \langle \vec{o}^{\,t}_1,\ldots,\vec{o}^{\,t}_n \rangle$ — joint observation history.
- $\pi(\vec{\mathbf{o}}^{\,t}) = \langle \pi_1(\vec{o}^{\,t}_1),\ldots,\pi_n(\vec{o}^{\,t}_n) \rangle$ — joint action.
- $\vec{\mathbf{o}}^{\,t+1} = (\vec{\mathbf{o}}^{\,t},\mathbf{o}^{t+1})$ — appending $\mathbf{o}^{t+1}$ to $\vec{\mathbf{o}}^{\,t}$.

IAS

Complexity of Dec-POMDPs

- Optimally solving Dec-POMDPs is NEXP-complete [Bernstein et al., 2002]
  - Most likely (if EXP$\neq$NEXP) doubly exponential in h

- Brute-force policy evaluation:

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

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

Outline

Dec-POMDPs as series of BGs—1

- Emery-Montemerlo et al. [2004] introduced an approximation for Dec-POMDPs using series of Bayesian Games (BGs).
  - 1 BG for each time step.
- Solving the BGs for stage $0,\ldots,h-1$ gives an (approximate) solution.
  - I.e., find $\beta^{0,*},\beta^{1,*},\ldots,\beta^{h-1,*}$
- 'Forward-sweep policy computation' (FSPC).

Dec-POMDPs as series of BGs—2

- What do these BGs look like?

BG is $\langle n, \mathcal{A}, \Theta, P(\cdot), Q \rangle$

- $\boldsymbol{\theta} = \langle \theta_1, \ldots, \theta_n \rangle$
- $P(\boldsymbol{\theta})$ prob. distr. over joint types.
- A payoff function $Q(\vec{\boldsymbol{\theta}}^t, \mathbf{a})$.

Action-observation history

- $\vec{\theta}_i^t = \left(a_i^0, o_i^1, a_i^1, \ldots, a_i^{t-1}, o_i^t \right) — \vec{\Theta}_i$

BG for time step t of a Dec-POMDP

- Types are action-observation histories: $\theta_i \equiv \vec{\theta}_i^t$.
- Given the past joint policy $\varphi^t = (\beta^{0,*}, \beta^{1,*}, \ldots, \beta^{t-1,*})$, probabilities $P(\vec{\boldsymbol{\theta}}^t)$ are known.

Dec-POMDPs as series of BGs—2

- What do these BGs look like?

BG is $\langle n, \mathcal{A}, \Theta, P(\cdot), Q \rangle$

- $\boldsymbol{\theta} = \langle \theta_1, \ldots, \theta_n \rangle$
- $P(\boldsymbol{\theta})$ prob. distr. over joint types.
- A payoff function $Q(\vec{\boldsymbol{\theta}}^t, \mathbf{a})$.

Action-observation history

- $\vec{\theta}_i^t = \left(a_i^0, o_i^1, a_i^1, \ldots, a_i^{t-1}, o_i^t \right)$ — $\vec{\Theta}_i$

BG for time step t of a Dec-POMDP

- Types are action-observation histories: $\theta_i \equiv \vec{\theta}_i^t$.
- Given the past joint policy $\varphi^t = (\beta^{0,*}, \beta^{1,*}, \ldots, \beta^{t-1,*})$, probabilities $P(\vec{\boldsymbol{\theta}}^t)$ are known.

Dec-POMDPs as series of BGs—2

- What do these BGs look like?

BG is $\langle n,\mathcal{A},\boldsymbol{\Theta},P(\cdot),Q\rangle$

- $\boldsymbol{\theta} = \langle \theta_1, \ldots, \theta_n \rangle$
- $P(\boldsymbol{\theta})$ prob. distr. over joint types.
- A payoff function $Q(\vec{\boldsymbol{\theta}}^t,\mathbf{a})$.

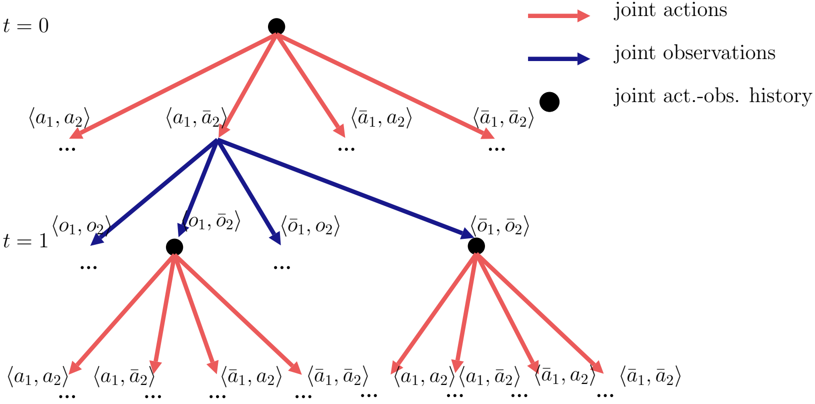
Action-observation history

- $\vec{\theta}_i^t = \left(a_i^0, o_i^1, a_i^1, \ldots, a_i^{t-1}, o_i^t\right)$ — $\vec{\Theta}_i$

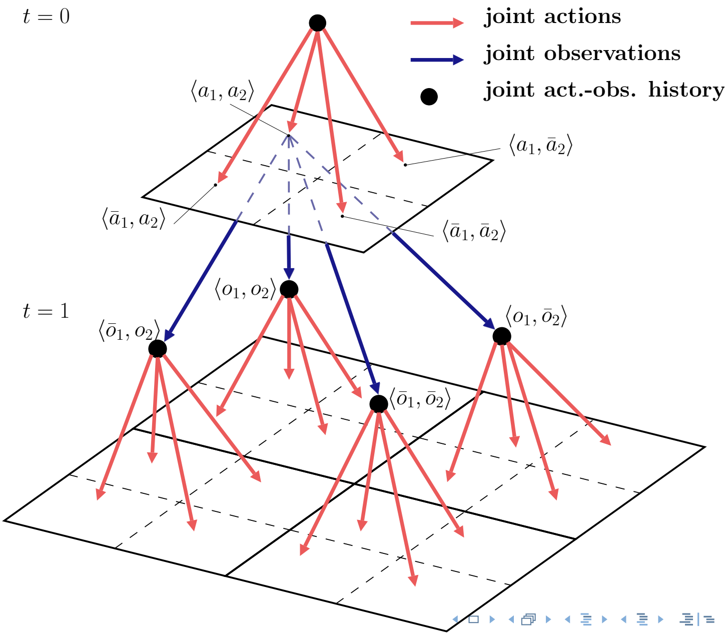
BG for time step t of a Dec-POMDP

- Types are action-observation histories: $\theta_i \equiv \vec{\theta}_i^t$.
- Given the past joint policy $\varphi^t = (\beta^{0,*}, \beta^{1,*}, \ldots, \beta^{t-1,*})$, probabilities $P(\vec{\boldsymbol{\theta}}^t)$ are known.

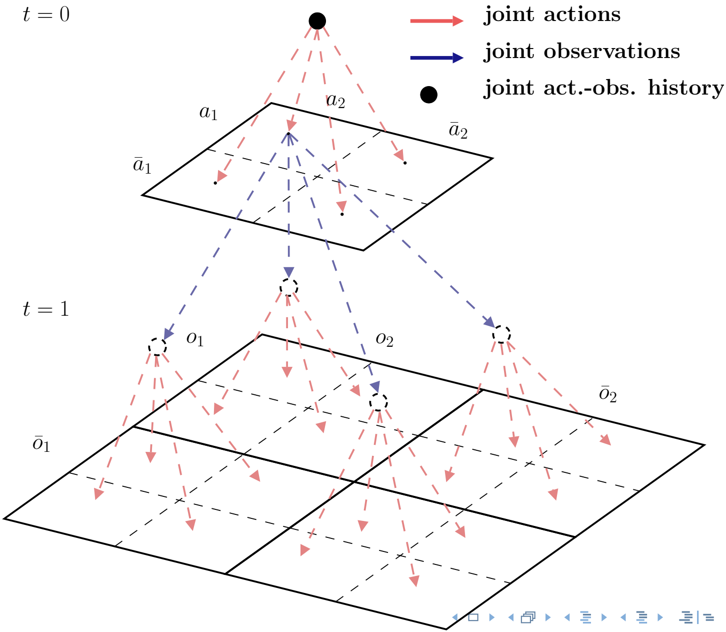
Visualization: Dec-POMDPs are trees



Visual: Dec-POMDPs as series of BGs—1



Visual: Dec-POMDPs as series of BGs—2



Visual: Dec-POMDPs as series of BGs—3

$\vec{\theta}_1^{t=0}$ \ $\vec{\theta}_2^{t=0}$		()	
		a_2	$\bar{a}_2$
()	a_1	$+2.75$	-4.1
	$\bar{a}_1$	-0.9	$+0.3$

$\vec{\theta}_1^{t=1}$ \ $\vec{\theta}_2^{t=1}$		(a_2,o_2)		$(a_2,\bar{o}_2)$		...
		a_2	$\bar{a}_2$	a_2	$\bar{a}_2$	
(a_1,o_1)	a_1	-0.3	$+0.6$	-0.6	$+4.0$	...
	$\bar{a}_1$	-0.6	$+2.0$	-1.3	$+3.6$	...
$(a_1,\bar{o}_1)$	a_1	$+3.1$	$+4.4$	-1.9	$+1.0$	...
	$\bar{a}_1$	$+1.1$	-2.9	$+2.0$	-0.4	
$(\bar{a}_1,o_1)$	a_1	-0.4	-0.9	-0.5	-1.0	...
	$\bar{a}_1$	-0.9	-4.5	-1.0	$+3.5$	...
$(\bar{a}_1,\bar{o}_1)$	...	...	...	...	...	...

- Dark entries not realized given $\langle a_1, a_2 \rangle$ at $t = 0$.
- Entries $Q(\vec{\theta}^t, \mathbf{a})$ represent expected reward.

IAS

BG payoff functions

BG payoff functions

- Policies will be good when using an appropriate payoff function.
- Unclear how to compute...
- Using the 'underlying MDP'.
- Further reading [Oliehoek and Vlassis, 2007]

Summarizing...

- Dec-POMDP can be modeled using BGs.
- FSPC can deal with (somewhat) larger problems
  - Size of BGs still grows exponentially!

Outline

Summary

- Dec-POMDP
  - MASs under partial observability
  - No communication
  - off-line (centralized) planning for on-line (decentralized) execution.
- Planning for a Dec-POMDP is hard.
  - BFS intractable for all but the smallest problems.
  - Approximation through BGs allows for somewhat larger problems.

References

- 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. Approximate solutions for partially observable stochastic games with common payoffs. In *Proc. of the International Joint Conference on Autonomous Agents and Multi Agent Systems*, pages 136–143, 2004.

F. A. Oliehoek and N. Vlassis. Q-value functions for decentralized POMDPs. In *Proc. of the International Joint Conference on Autonomous Agents and Multi Agent Systems*, pages 833–840, May 2007.

N. Vlassis. *A Concise Introduction to Multiagent Systems and Distributed Artificial Intelligence*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2007.

Outline

5 Appendix

scratch space

Intuition behind complexity — 1

- For MDPs, the solution can be efficiently found by dynamic programming.
  - Simple calculation of optimal Q-value function:

$$Q^{*,t}(s,a) = R(s,a) + \sum_{s'} P(s'|s,a) \max_{a'} Q^{*,t+1}(s',a')$$

  - Greedy extraction of optimal policy:

$$\forall_t \forall_s \quad \pi^{*,t}(s) = \arg\max_{a \in \mathcal{A}} Q^{*,t}(s,a)$$

- For POMDPs: states are replaced by beliefs b (probability distributions over states).
  - $b_{a,o}$ can be calculated from preceding belief b by Bayes' rule.
  - I.e., for each action-observation history, there is 1 belief.

Intuition behind complexity — 2

Action-observation history

- individual — $\vec{\theta}_i^t = \left(a_i^0, o_i^1, a_i^1, ..., a_i^{t-1}, o_i^t\right)$
- joint — $\vec{\boldsymbol{\theta}}^t = \left\langle \vec{\theta}_1^t, ..., \vec{\theta}_n^t \right\rangle$

Similar to POMDP, in a Dec-POMDP:

- 'joint belief' — prob. distr. over states $b^{\vec{\boldsymbol{\theta}}^t}$.

However...

- The agents can not observe $\vec{\boldsymbol{\theta}}^t$.
  - Can't calculate $b^{\vec{\boldsymbol{\theta}}^t}$ during execution.
  - Can't condition their actions on $b^{\vec{\boldsymbol{\theta}}^t}$.
  - Can't calculate Q^* as easy.

Instead each agent i will have to reason about $\vec{\boldsymbol{\theta}}^t$

- I.e., given $\vec{\theta}_i^t$, what is $P(\vec{\theta}_{\neq i}^t | \vec{\theta}_i^t)$?