# RL 14: POMDPs continued

Michael Herrmann

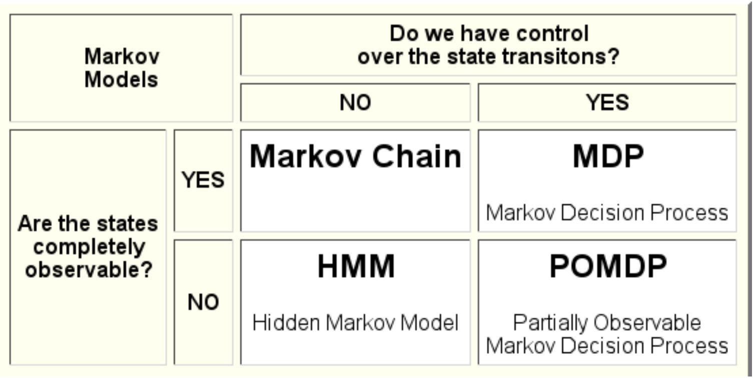
University of Edinburgh, School of Informatics

06/03/2015

# POMDPs: Points to remember

- Belief states are probability distributions over states
- Even if computationally complex, POMDPs can be useful as a modelling approach (consider simplification of the implementation in a second stage)
- POMDPs enable agents to deal with uncertainty efficiently
- POMDPs are Markov w.r.t. belief states
- Beliefs tend to blur as consequence of the state dynamics, but can refocus by incorporating observations via Bayes' rule.
- Policy trees take all possible realisations of the sequence of future observations into account, i.e. the choice of the current action depends on the average over many futures.
- This causes exponential complexity unless the time horizon is truncated (standard) or approximations are used (e.g. $\mathcal{Q}$MDP, AMPD, and sample-based methods).
- Often some states are fully observable and these may be the states where decisions are critical (e.g. a robot turning when observing a doorway)

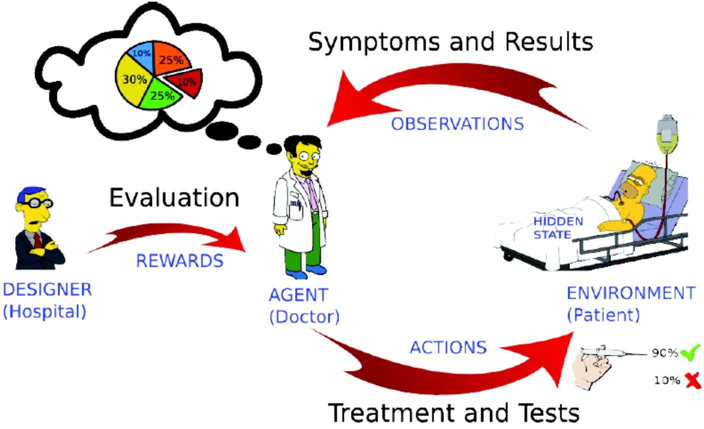
# Markov models



| Markov Models | | Do we have control over the state transitons? | |
|---|---|---|---|
| | | NO | YES |
| Are the states completely observable? | YES | Markov Chain | MDP<br>Markov Decision Process |
| | NO | HMM<br>Hidden Markov Model | POMDP<br>Partially Observable Markov Decision Process |

A. Cassandra: http://www.pomdp.org/faq.shtml

# Solving small information-gathering problems



From Mauricio Araya-Lopez, JFPDA 2013.

# POMDP: Applications

- Belief-MDP (Åström, 1965)
- Autonomous robot localisation and navigation
- Classical RL applications (test problems): elevator control, machine maintenance, structural inspection,
- Business (IT): marketing, troubleshooting, dialog systems, distributed data base queries, routing
- Operations research: Medicine, finance, fishery industry, conservation, education

Anthony Cassandra. A Survey of POMDP Applications. Presented at the AAAI Fall Symposium, 1998.

Young, S., Gasic, M., Thomson, B., & Williams, J. D. (2013). POMDP-based statistical spoken dialog systems: A review. Proceedings of the IEEE, 101(5), 1160-1179.

# Belief spaces

- In terms of belief states, POMDPs are MDPs: previous methods are applicable.
- Belief state: a line for 2 states, ..., a simplex for $N$ states
- Value function over belief state is piecewise linear and convex (Sondik, 1978)
    - Represent by points → point-based algorithms
    - Represent by vectors → $\alpha$-vector based

# Value iteration

$$
\begin{aligned}
V_t(b) &= \max_{a\in\mathcal{A}} \left( \sum_{s\in\mathcal{S}} b(s) \sum_{s'\in\mathcal{S}} T(s'|s,a) \sum_{o\in\Omega} \Omega(o|s',a) \left(R^a_{ss'o} + V_{t-1}\left(b^o_a(s')\right)\right)\right) \\
&= \max_{a\in\mathcal{A}} \left( \sum_{s\in\mathcal{S}} b(s) R(s,a) + \sum_{s\in\mathcal{S}} b(s) \sum_{s'\in\mathcal{S}} T(s'|s,a) \sum_{o\in\Omega} \Omega(o|s',a)\, V_{t-1}\left(b^o_a(s')\right)\right) \\
&= \max_{a\in\mathcal{A}} \left( \sum_{s\in\mathcal{S}} b(s) R(s,a) + \sum_{o\in\Omega} \max_k \sum_{s\in\mathcal{S}} b(s) \sum_{s'\in\mathcal{S}} T(s'|s,a)\, \Omega(o|s',a)\, \alpha^k_{t-1}(s')\right) \\
&= \max_{a\in\mathcal{A}} \left( \sum_{s\in\mathcal{S}} b(s) \left( \underbrace{R(s,a) + \sum_{o\in\Omega} \sum_{s'\in\mathcal{S}} T(s'|s,a)\, \Omega(o|s',a)\, \alpha^{l(b,a,o)}_{t-1}(s')}_{\alpha^k_t(s)} \right)\right)
\end{aligned}
$$

# Algorithm POMDP($T$) (based on a set of points $x_i$)

```
$\Upsilon = \{(0, \ldots, 0)\}$, $\mathcal{U} = \emptyset$
for $\tau = 1$ to $T$ do
    $\Upsilon' = \emptyset$
    for all $(\mathcal{U}; v_1^k, \ldots, v_N^k)$ in $\Upsilon$ do
        for all control actions $u$ do
            for all measurements $z$ do
                for $j = 1$ to $N$ do
                    $v_{j,u,z}^k = \sum_{i=1}^N v_i^k \, p\left(z|x_i\right) p\left(x_i|u, x_j\right)$
                endfor
            endfor
        endfor
    endfor
    for all control actions $u$ do
        for all $k = 1$ to $|\Upsilon|$ do
                for $i = 1$ to $N$ do
                    $v_i' = r\left(x_i, u\right) + \gamma \sum_z v_{i,u,z}^k$
                endfor
            add $u$ to $\mathcal{U}$ and $(\mathcal{U}; v_1', \ldots, v_N')$ to $\Upsilon'$
            endfor
        endfor
    optional: prune $\Upsilon$'
    $\Upsilon$=$\Upsilon$'
endfor
return $\Upsilon$
```

# Remarks on the algorithm

- Without pruning $|\Upsilon|$ increases exponentially with $T$
- The algorithm describes the determination of the value function. Value iteration, actual observations and actions are not entering.
- Further steps in algorithm
  - Find value function on policy trees up to a given $T$
  - Determine maximum over branches and perform first action
  - Recalculate policy taking into account observations and rewards
  - Update observation model, transition model and reward model
- Many variants exist.

# A systematic approach to POMDPs

POMDPs have no information about states but about observation outcomes and preformed actions.

In the following we start with a trivial expression of this fact and use this to motivate belief states.

# Finite Horizon Problems

- COMDP with horizon $T$: Optimal policy is a sequence of mappings $\pi_t^* : \mathcal{S} \to \mathcal{A}$, $t = 0, \ldots, T-1$

  $$\pi^* = (\pi_t^*)_{t=0,\ldots,T-1} = \arg\max_{\pi_0,\ldots,\pi_{T-1}} E\left[r(s_T) + \sum_{t=0}^{T-1} r(s_t, \pi_t(s_t))\right]$$

  Previously, we assumed that all $t$ share the same mapping.
- POMDP: Optimal policy $\pi^*$ depends on past actions and observations. Policies are trees

  $$\pi : (\mathcal{A} \times \mathcal{O})^* \to \mathcal{A}$$

  Optimal policy trees can be calculated by value iteration on branches of policy trees. Branches correspond to sequences of actions and observations

# Information states

In principle all previous information is needed to decide about values and actions.

This is realised by information states $I \in (\mathcal{A} \times \mathcal{O})^*$, e.g. $I_t = (a_0, o_0, ..., a_t, o_t)$.

Then it holds (trivially) that

$$P\left(I_{t+1} = (a_0, o_1, ..., a_{t+1}, o_{t+1})\,|a_{t+1}, I_t\right) = P(o_{t+1}|a_{t+1}, I_t)$$

A POMDP is an information state MDP (with a huge state space)

Rewards

$$r_I\left(I_t, a_t\right) = \sum_{s\in\mathcal{S}} P\left(s_t = s|I_t\right) r\left(s, a\right)$$

# Value iteration on information states

Initialisation

$$V_T(I_T) = \sum_{s \in \mathcal{S}} P(s_T = s | I_T)\, r(s)$$

Bellman optimality equation

$$V^*(I_t) = \max_{a \in \mathcal{A}} \Bigg( \sum_{s \in \mathcal{S}} P(s_t = s | I_t)\, r(s, a) + \gamma \sum_{o \in O} P(o | I_t, a)\, V^*(I_{t+1} = (a_0, \ldots, a = a_t, o = o_{t+1})) \Bigg)$$

State space grows exponentially with $T$ (episode length)

Belief states summarise information states by distributions of states.

# Equivalence of Information States and Belief States

For an underlying Markov process, information states and belief states are equivalent in the sense that the lead to the same optimal value functions.

Belief states summarise information states by distributions of states

Optimal value functions are defined by the Bellman optimality equation.

Equivalency in terms of the value functions follows from the compatibility of the belief update with the Bellman equation

# Belief update

$$\begin{aligned} b'(s') &= P(s'|z,a,b) \\ &= \frac{P(z|s',a,b)\,P(s'|a,b)}{P(z|a,b)} \\ &= \frac{P(z|s',a)\sum_{s\in\mathcal{S}} P(s'|a,b,s)\,P(s|a,b)}{P(z|a,b)} \\ &= \frac{O(z,s',a)\sum_{s\in\mathcal{S}} T(s',a,s)\,b(s)}{P(z|a,b)} \end{aligned}$$

$z$ observation, $u$ action, $s$ state, $b$ belief (distribution of states)

$O$ observation model, $T$ state transition probability

Rewards on belief states: $\rho(b,u) = \sum_{s\in\mathcal{S}} b(s)\,R(s,a)$

# Value iteration on belief states

Initialisation

$$V_T\left(b\right) = \sum_{s\in\mathcal{S}} b\left(s\right) r\left(s\right)$$

Bellman equation

$$V\left(b\right) = \max_{a\in\mathcal{A}} \left( \sum_{s\in\mathcal{S}} b\left(s\right) r\left(s, a\right) \right. \\ \left. + \gamma \sum_{o\in O} \sum_{s', s''\in\mathcal{S}} P\left(o|s'', a\right) P\left(s''|s', a\right) b'\left(s'\right) V\left(b'\left(o, a\right)\right) \right)$$

# Policy iteration

Backup in belief space:

$$V(b) = \max_u \left( r(b,u) + \gamma \int V(b')\, p(b'|u,b)\, db' \right)$$

$$p(b'|u,b) = \int p(b'|u,b,z)\, p(z|u,b)\, dz$$

$$V(b) = \max_u \left( r(b,u) + \gamma \int \left[ \int V(b')\, p(b'|u,b,z)\, db' \right] p(z|u,b) dz \right)$$

# Recent and current research

- Solution of Gridworld POMDPs (M. Hausknecht, 2000)
- Point-based value iteration (J. Pineau, 2003)
- Large problems: Heuristic Search Value Iteration (T. Smith & R. Simmons, 2004): 12545 states, considering bounds for the value function over belief states
- Learning POMDPs from data (Learning a model of the dynamics)
    - compressed predictive state representation
    - Bayes-adaptive POMDPs (tracking the dynamics of belief states)
- Policy search, hierarchical POMDPs, decentralised POMDPs, ...

Joelle Pineau (2013) A POMDP Tutorial. *European Workshop on Reinforcement Learning.*

# Summary on POMDPs

- POMDPs compute the optimal action in partially observable, stochastic domains.
- For finite horizon problems, the resulting value functions are piece-wise linear and convex, but very complicated
- A number of heuristic and stochastic approaches are available to reduce the complexity.
- Combinations with other RL approaches possible
- POMDPs have been applied successfully to realistic problem is robotics

# References

Thrun, S., Burgard, W., & Fox, D. (2005). Probabilistic robotics. MIT press. Chapters 15 and 16. (text book)

Milos Hausknecht (2000) Value-function approximations for partially observable Markov decision processes. *Journal of Artificial Intelligence Research* **13**, 33-94. (detailled paper)

Joelle Pineau (2013) A POMDP Tutorial. *European Workshop on Reinforcement Learning. (review on recent research)*