RL 13: POMDPs continued

Michael Herrmann

University of Edinburgh, School of Informatics

01/03/2016

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 Markovian 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)

Belief propagation

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

o observation, a action, s state, b belief (distribution over states)

Ω observation model, T state transition probability

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

Belief propagation

- Bayesian belief propagation (given action a):

$$b'(s') = \frac{\Omega(o \mid s', a) \sum_{s \in S} T(s' \mid s, a) b(s)}{\sum_{\tilde{s} \in S} \Omega(o \mid \tilde{s}, a) \sum_{s \in S} T(\tilde{s} \mid s, a) b(s)}$$

  where s are the previous states with distribution b, s' the new states with distribution b', T the state transition probabilities, and Ω the observation probabilities for the actual signals o.
- In terms of spread of the belief (variance), usually T increases uncertainty, Ω reduces uncertainty.
- In terms of the decidedness of the belief towards one state, usually T is neutral, while the effect of Ω depends on the outcome of the observation.

Value function

- Given the current belief b and the next belief b' we can compute a new iteration of the value function V_{k+1} from the current estimate V_k. Formally, we have for each action a

$$V_{k+1}^{a}(b) = r(b,a) + \gamma \int V_k(b')\, p(b'|a,b)\, db'$$

which practically is for discrete states

$$V_{k+1}^{a}(b) = r(b,a) + \gamma \sum_{s'} V_k(b')\, b'(s'|a,b)\, ds'$$

- Instead of $V{`}^{a}(b)$ we could write $Q(b,a)$

A simple POMDP Algorithm ($T = 1$)

Set time t and initial belief $b + t$

1. Choose action $a_t = \arg\max_a V_t^a(b_t)$
2. Execute action a_t and increment $t \to t + 1$
3. Read new observation o_{t+1} and reward r
4. Propagate b_t to b_{t+1} (using a_t and o_{t+1})
5. Calculate $V_{t+1}^a(b_t)$ for all a (using $V_t^a(b_{t+1})$, a_t , o_{t+1} and r)

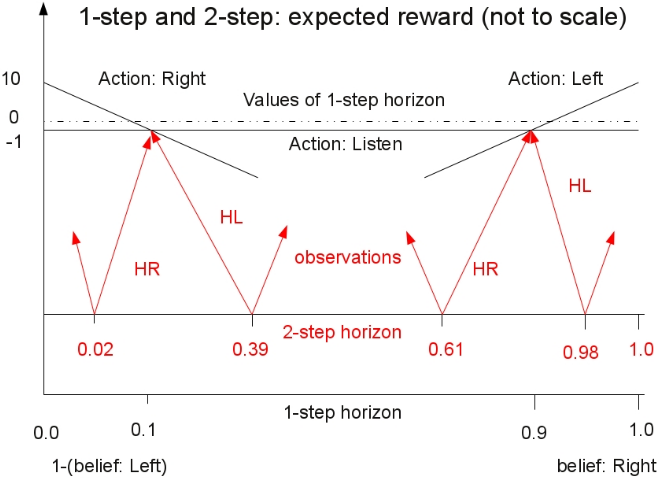
Notes: Because b is high-dimensional, it is unlikely that we have a $V_t^a(b_{t+1})$ that was recently updated, so we should calculate V_{t+1} for all b. Alternatively, we can use a set of points in the belief space.

Pruning is possible, i.e removing $V_{t+1}^a(b_t)$ if dominated by other actions

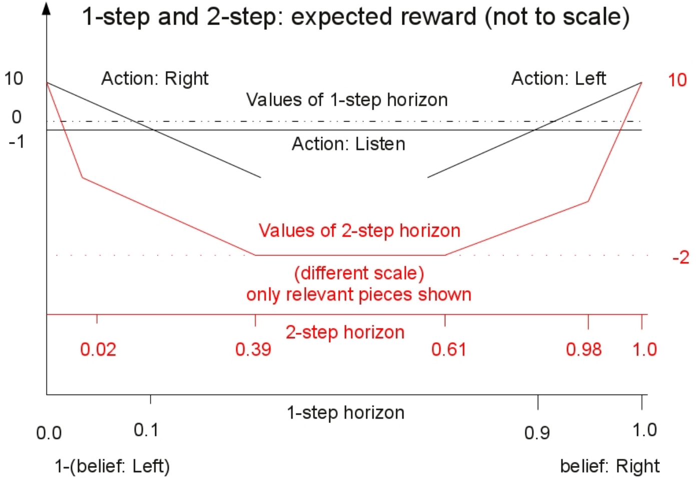
Multi-step prediction

- We do not have to interate thevalue function. We can use an update with a small learning rate instead. Then the $T = 1$ algorithm will integrate all possible futures into the value function.
- If transition and obeservation probabilities are know and using $r(b, a) = \sum_{s \in \mathcal{S}} b(s)\, r(s, a)$, steps 4. and 5. can be performed for all a and o a few (T) steps into the future (exponentially complex, but pruning helps)
- Value function over belief state is piecewise linear and convex (Sondik, 1978)

Repeated from last time: The Tiger problem



Repeated from last time: The Tiger problem



Value iteration

- Given the current belief b and the next belief b' (see previous slide) we can compute a new iteration of the value function V_{k+1} from the current estimate V_k. Formally, we have

$$V_{k+1}(b) = \max_a \left(r(b,a) + \gamma \int V_k(b')\, p(b'|a,b)\, db' \right)$$

which practically is for discrete states

$$V_{k+1}(b) = \max_a \left(r(b,a) + \gamma \sum_{s'} V_k(b')\, b'(s'|a,b)\, ds' \right)$$

- Initialisation

$$V(b) = \sum_{s \in \mathcal{S}} b(s)\, r(s)$$

- Action choice is given by the argmax

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

What is Missing in POMDPs?

- POMDPs do not describe natural metrics in environment
  - When driving, we know both global and local distances
- POMDPs do not natively recognise differences between scales
  - Uncertainty in control is entirely different from uncertainty in routing
- POMDPs conflate properties of the environment with properties of the agent
  - Roads and buildings behave differently from cars and pedestrians: we need to generalise over them differently
- POMDPs are defined in a global coordinate frame, often discrete
  - We may need many different representations in real problems

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)*
- The POMDP Page (www.pomdp.org)
- Tony's POMDP Page)cs.brown.edu/research/ai/pomdp)