Reinforcement Learning: Tutorial 5 (week from 3. 3. 2014)

1. How can particle filters be used in the context of robot localization?

   Particle filters sample a probability distribution. The dynamics of the particles can be used to represent the change of the probability in time.

1. Algorithm **Discrete_Bayes_filter**(*Bel(x),d*)
2. η=0
3. If *d* is a perceptual data item *z* then
4. For all *x* do
5. $Bel'(x) = P(z \mid x) Bel(x)$
6. $\eta = \eta + Bel'(x)$
7. For all *x* do
8. $Bel'(x) = \eta^{-1} Bel'(x)$
9. Else if *d* is an action data item *u* then
10. For all *x* do
11. $Bel'(x) = \sum_{x'} P(x \mid u, x') \, Bel(x')$
12. Return *Bel'(x)*

2. The "art" of importance sampling: We are sampling *P*(*x*), which may be not cover the interesting aspect of the game. It is already interesting to consider sampling X/*L*(*x*) based on the distribution *P*(*x*)*L*(*x*). Why? How could it be useful in RL?
   Similarly, we can consider this scheme (see on the right ->)
   Why should one want to do this? What happens for *P*(*x*)=*W*(*x*)?

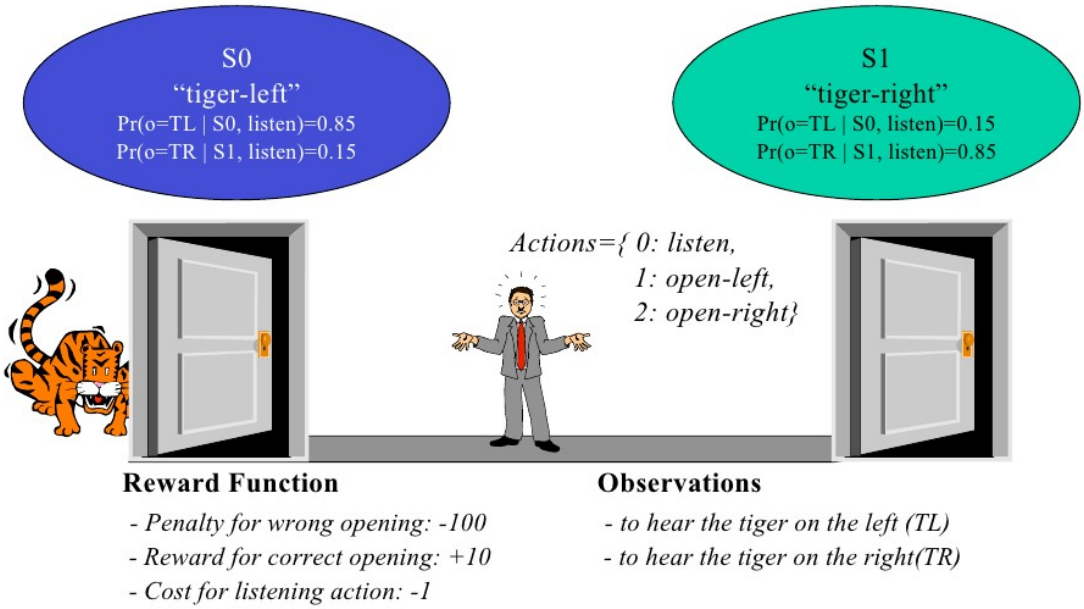
$$\overline{A} = \frac{\sum_x P(x) \, A(x) / W(x)}{\sum_x P(x) / W(x)} \; .$$

   Importance sampling is a method for variance reduction. If a state *x* is not probable (w.r.t. *P*) it will not contribute much to the average of a quantity *A* (which can be a value). As it incurs a certain cost to obtain the value of *A* in *x*, it might be better to use the information from this measurement efficiently. For *P*(*x*)=*W*(*x*) the effect of the uneven distribution cancels out and all observations of *A* eneter with an equal weight (s. http://en.wikipedia.org/wiki/Importance_sampling).

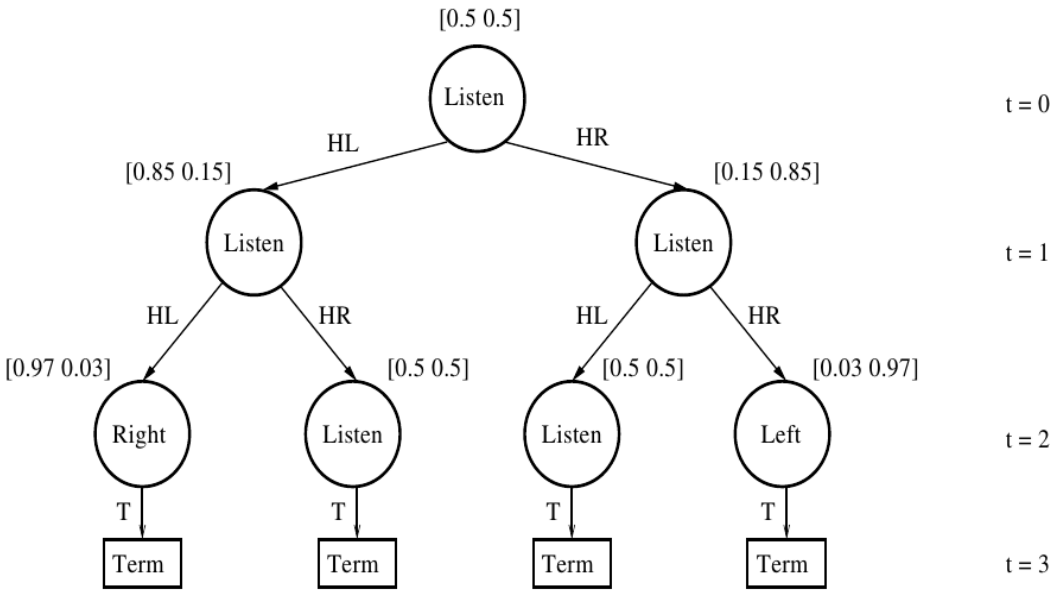
3. How are POMDPs and Hidden Markov Models (HMMs) related? Would a Viterbi algorithm be useful in POMDPs?

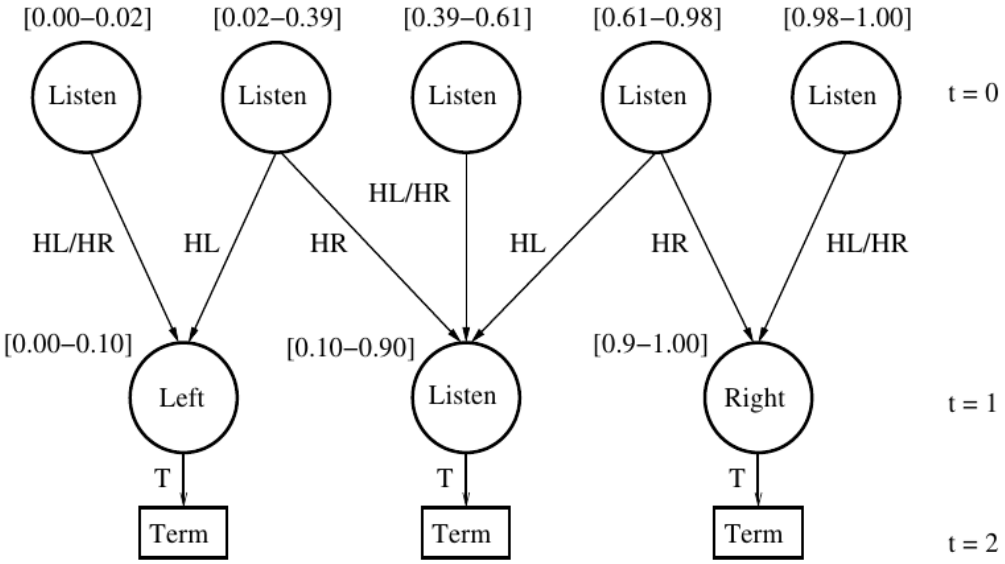
   A POMDP is an MDP. A HMM is a model of a time series that is assumed to be caused by a sequence of unobservable states, i.e. there are no actions involved such that time series cannot be influenced in any way. The HMM describes, nevertheless, the difference between POMDP and MDP in a sense, such that one could state that MPD + HMM = POMDP (which is the title of a presentation by H. Daumé).

4. Discuss the tiger problem (from: Dr. Stephan Timmer "Introduction to POMDPs")




```
# This is the tiger problem of AAAI paper fame in the new POMDP
# format. This format is still experimental and subject to change

discount: 0.75
values: reward
states: tiger-left tiger-right
actions: listen, open-left, open-right
observations: tiger-left, tiger-right

Transitions:
listen -> identity
open-left -> uniform
open-right -> uniform

Observations
listen (in either state):     0.85 0.15
                              0.15 0.85
open-left: uniform
open-right: uniform

Rewards:
R:listen : * : * : * -1
R:open-left : tiger-left : * : * -100
R:open-left : tiger-right : * : * 10
R:open-right : tiger-left : * : * 10
R:open-right : tiger-right : * : * -100
```





5. Consider the application to a POMDP to the problem of controlling several elevators problem. For what definition of states does any uncertainty arise? Discuss the advantage of a POMDP over a state abstraction (that does not distinguish between states that can be confused). Compare to the original Barto&Crites approach (see final slides of the lecture RL09). Can the design of the elevator operation be changed such that this uncertainty is removed/reduced?

   For a low complexity (few elevators, few floors, single persons) the state space can be described completely using a discrete set of states. Already for the Appleton Tower this will not easily work. The state of the system is determined by the positions of the cars, the number of people inside the cars and their destinations and the people waiting for a car and their destinations. Uncertainty arises because the latter is not known (apart from "up" and "down" from the relevant flow). This can be ignored, by as shown in the literature, there are preferred destinations from each floor. These may depend on building and time of day, but can in principle be learned

by an appropriately designed POMDP algorithm. Nevertheless, a POMDP may require substantially more computation than the neural-network-based approach in Barto&Crites. Consider also that there is noise in the system (some people may press a wrong button), that some states are improbably and can perhaps, instead of being learned, be dealt with by a fallback option based on a simple standard algorithm without reducing the total performance of the system and that people have a subjective view (do not enjoy being passed by an elevator car while waiting even if this is good for the global performance).

6. Recall the discussion of "afterstates" from a previous tutorial. Afterstates are an option to include the reaction of an opponent into the own policy. Under what conditions would it make sense to reformulate the problem as POMDPs?

   An afterstate can be used to provide the learner information about the reaction of an opponent. The combined "state plus afterstate" is available when making the next move such that there is no state uncertainty. An POMDP can anticipate these states probabilistically and would thus in principle be able to use these anticipation for learning policy trees that extend a few steps into the future. A standard algorithm (such as SARSA) would simply learn the average which may be enough if the opponent is actually stochastic (it can still have more or less probable reactions). If the opponent follows a strategy a POMDP may be more appropriate. Consider, however, that in order to be Markovian the POMDP should operate in strategy space (a strategy could become clear a few states later), which is already very complex such that compromises must be taken (e.g. with Markovianity) in order to realise such an approach.

7. Consider a robot moving down a hallway as a 1D problem with states being sections of the track of a length of 1m. The robot's speed is 1m/s +/- 0.1m/s (assume a uniform distribution of deviations). Discuss the belief propagation in standard POMDP vs. the corresponding effects in an augmented MDP or in QMDP. Think of a navigation task which is then to be solved by either of these methods.

   The robot receives input about every second when it enters a new section of the track. Due to the uncertainty in the speed there is a chance of staying in the same section for one more step or to skip a section. We don't know the distribution of the errors, but can assume e.g. that a standard deviation of 0.1 m/s would lead to entering the next state in 0.68 of the cases and to remaining in the same state or to skipping a state in 0.16 of the cases (A Gaussian distribution gives similar values). We can ignore transitions to other states. See problem 1 above for the further procedure. An augmented MDP takes the width of the resulting state distribution into account in terms of an entropy (and can prefer thus e.g. states that disambiguate the spatial uncertainty by information from walls or doors etc.). An QMPD is taking into account the first step explicitly but assumes certainty for the values of later states in analogy to Q-learning where it is assumed that the best policy is followed after the next state.

8. Assume a robot moving in a dark environment where information is available only from touch sensors. The robot learns to move successfully using a POMDP. Now the lights are switched on and the robot can use again its excellent visual system. How can it use the information from POMDP for initialising a simpler RL method?

   The idea is that now full state information is available, i.e. the observation probabilities become trivial (while the state transitions can still be stochastic). The belief states (which are probability distributions over the state space) become now concentrated to a unique state such that the relevant equations can be reformulated in terms of the states. E.g. for the Bellmann optimality equation

$$V^*(b) = \max_{a \in A} \left[r(b,a) + \gamma \sum_{o \in O} \Omega(o \mid b,a) V^*(\tau(b,a,o)) \right]$$

   will now become

$$V^*(s) = R(s) + \max_a \gamma \sum_{s'} P(s'|s,a) V^*(s')$$

9. Have a look at a review paper such as Anthony R. Cassandra (1998) A Survey of POMDP Applications. Discuss set-up, advantages and limitations of POMDP in the mentioned application problems (or do this simply for any of the examples above).