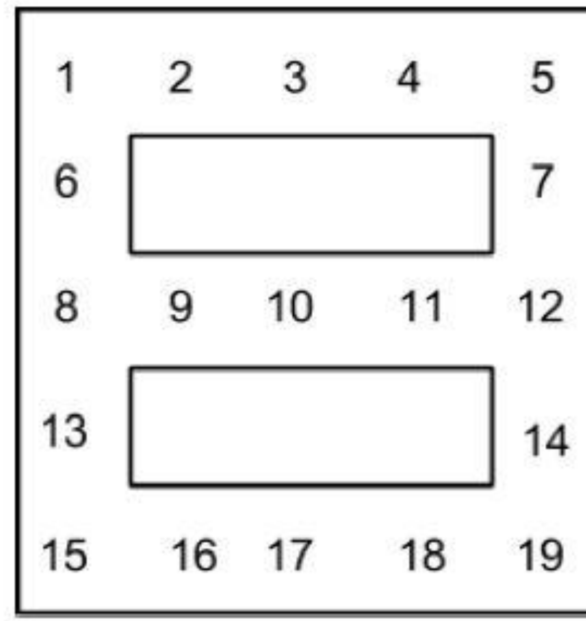
Partial Observability

B. Ravindran

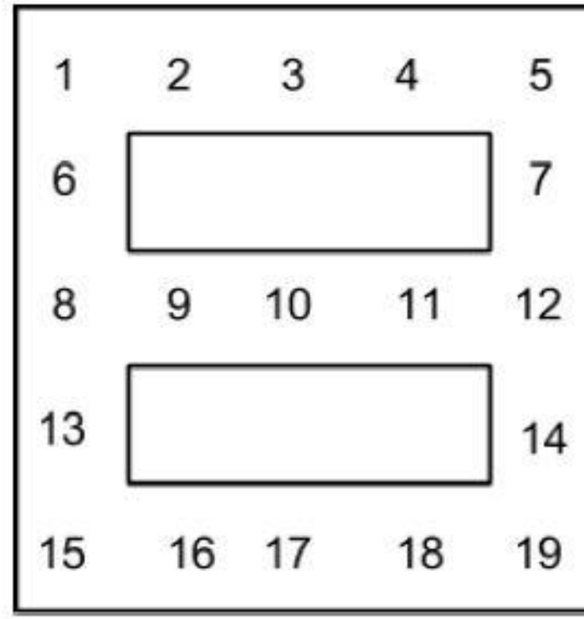
Partial Observability

- Agent receives information in the form
  (b_N, b_S, b_E, b_W)
  where the subscript indicates a direction
  $b_x = 1$ if corridor is blocked in the x direction
- Given observation (1,1,0,0)
  this may refer to any of the following states
  {2,3,4,9,10,11,16,17,18}



Partial Observability

Which among the 19 states shown in the figure can be unambiguously identified using the 4-direction blocked information?



Approaches to Handle Partial Observability

- Ostrich Approach!
- POMDPs
- History Based Methods
  - U-Tree (McCallum, 1996)
    - used to learn a task-specific state representation that makes perceptual and memory distinctions only where needed for the task at hand
  - MC-AIXI (Veness, Ng, Hutter, Uther & Silver, 2011)
- Predictive State Representations (PSRs)
  - Representation based on predictions of future observations

POMDP – Partially Observable MDP

- In a partially observable MDP (POMDP) the system dynamics are determined by an MDP but the agent cannot directly observe the underlying state.

POMDP – Partially Observable MDP

- Formally a POMDP is a seven-tuple (S,A,P,R,Ω,O,γ), where
  - S is a set of states
  - A is a set of actions
  - P is a set of conditional transition probabilities between states
  - R is the reward function
  - Ω is a set of observations
  - O is a set of conditional observation probabilities
  - $\gamma \in [0,1]$ is the discount factor

- On taking action $a \in A$, the environment transitions from current state s to next state s' with probability $P(s'|s,a)$
- At the same time, the agent receives an observation $o \in \Omega$ which depends on the new state of the environment with probability $O(a, s', o)$.

POMDP – Partially Observable MDP

POMDPs offer a versatile model that allows for:

- Uncertainty in knowledge of state
- Noisy observations
- Uncertainty in effects of actions

Potential Applications:

- Maintenance scheduling, Quality control
- Robot Navigation
- Treatment Planning, Medical Diagnosis

and many others..

History based methods

- Underlying dynamics of a POMDP are Markovian.
- No direct access to the current state.
- Takes decisions by keeping track of (possibly) the entire history of the process:
- t = 1 : $< \mathbf{O_0} >$
- t = 2 : $< \mathbf{O_0, A_0, O_1} >$
- t = 3 : $< \mathbf{O_0, A_0, O_1, A_1, O_2} >$
- t = 10: $< \mathbf{O_0, A_0, O_1, \ldots., O_8, A_8, O_9} >$

History based methods

- We can build more "memory" or history into our states by using a higher order Markov system.
- eg. k-th order Markov system:

$$\mathbb{P}(X_t = x_t | X_0 = x_0, \ldots, X_{t-1} = x_{t-1})$$
$$= \mathbb{P}(X_t = x_t | X_{t-k} = x_{t-k}, \ldots, X_{t-1} = x_{t-1})$$

- Additional history **may** have predictive value.
- Higher the order, greater the computation needed.

History based methods

Issues with history-based approaches:

- Large and growing state spaces.
- Difficult to get parameterized representations for variable length states.
- Possible wastage of computation.

Belief States

- History-based policy grows exponentially with horizon: Not suitable for infinite-horizon POMDPs.
- Solution: Use a “belief state” that sufficiently summarizes history.
- Belief state: A probability distribution over states.
- Belief space: Set of all possible probability distributions.
- Update belief state every time we take an action and see a new observation.

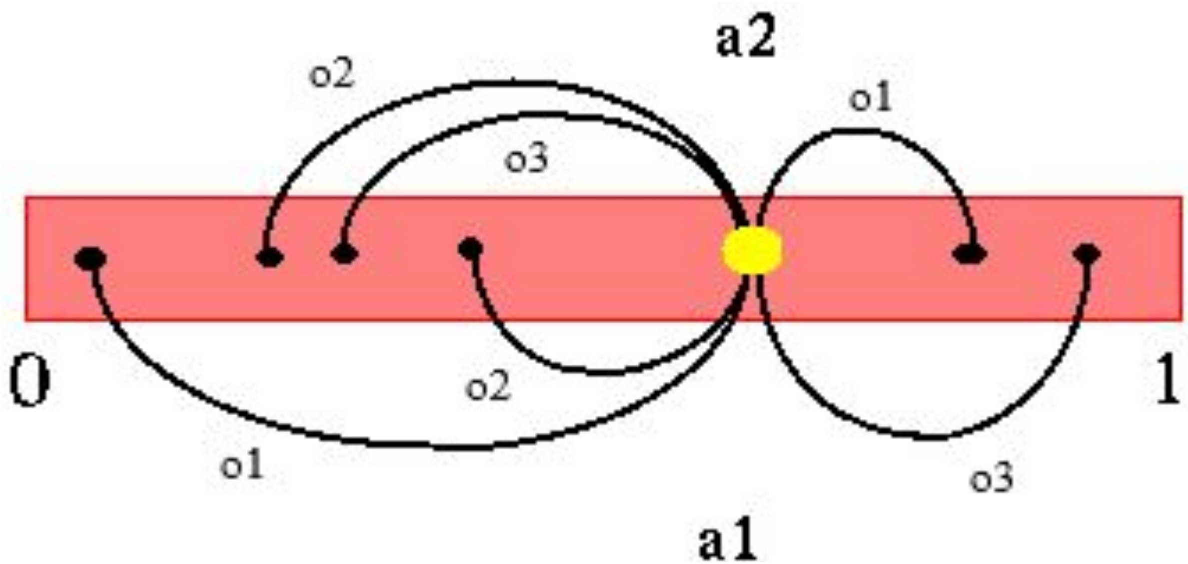
Belief State

- Consider a 2-state MDP with 2 actions a_1 and a_2.
- Probability for being in s_1: p
- Therefore, probability of being in s_2: 1-p
- We can represent belief space as [0, 1]:

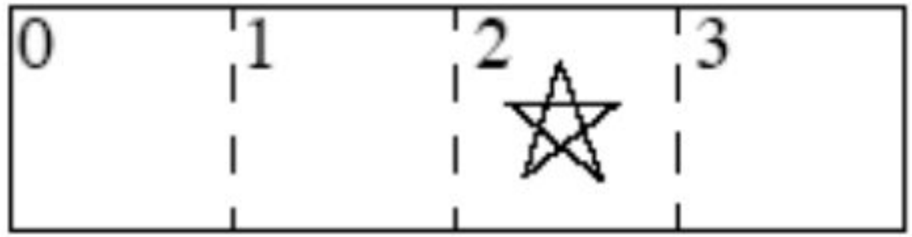
- Far right: In s_1 with probability 1.

Updating Belief State

- Assume we start from belief state ***b*** (yellow dot).
- If we take an action a_1 and observe o_1, the next belief state is fully determined.
- With finite actions and observations, there are only finite possible next belief states.

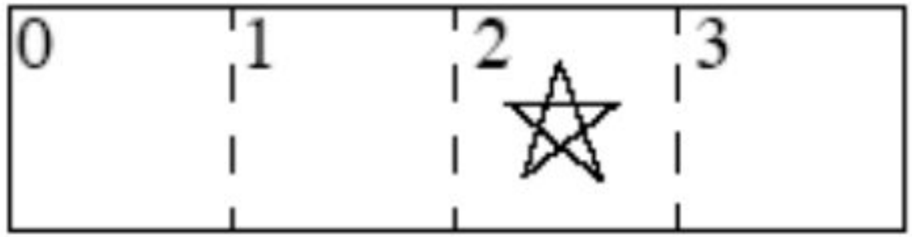


Another Example



- Two actions: left, right; deterministic
- If agent moves into a wall, stays in current state
- If agent reaches the goal state (star), moves randomly to state 0, 1, or 3, and receives reward 1
- Agent can only observe whether or not it is in the goal state

Another Example



- **b**: belief state
- **b(s)** = prob agent is in state s
- After goal: (1/3, 1/3, 0, 1/3)
- After action right and not observing the goal: (0, 1/2, 0, 1/2)
- After moving right again and still not observing the goal: (0, 0, 0, 1)

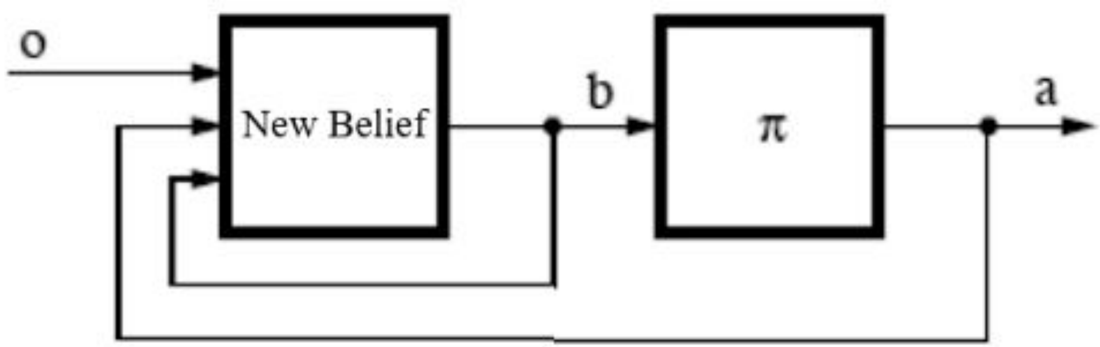
Belief State MDPs

- POMDPs can be viewed as belief state MDPs:
  - States: B (beliefs)
  - Actions: A
  - Transitions:
  - Rewards:

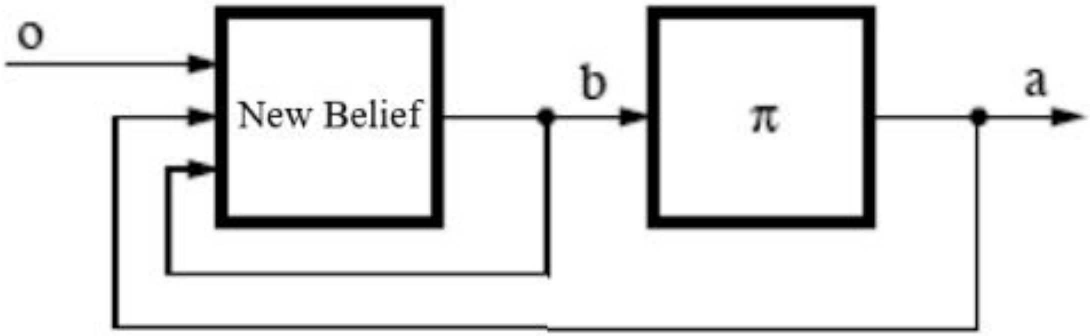
- Belief state MDPs can be considered MDPs.
- The belief space is continuous.

Belief State MDPs

- Perfect memory controller (Cassandra et. al.)
- SE (State Estimator):
  - Computes the agent's new belief state as a function of the old belief state, the last action and the current observation.
- Policy: Learning as in an MDP, except with *beliefs* instead of *states*.



Computing Next Belief



$$
\begin{aligned}
SE_{s'}(b,a,o) &= Pr(s'|a,o,b) \\
&= \frac{Pr(o|s',a,b)Pr(s'|a,b)}{Pr(o|a,b)} \\
&= \frac{O(a,s',o)\sum_s \boldsymbol{P}(s,a,s')b(s)}{Pr(o|a,b)}
\end{aligned}
$$

Where $Pr(o|a,b)$ is a normalising factor defined as

$$Pr(o|a,b) = \sum_{s'} O(a,s',o) \sum_{s \in S} \boldsymbol{P}(s,a,s')b(s)$$

Q-MDP

- Assume knowledge of the underlying MDP.
- However, current state not known.
- Make the (usually false) assumption that one step of control leads to full observability.
- Steps:
  - Evaluate value function using MDP knowledge.
  - Compute expected return for each action.

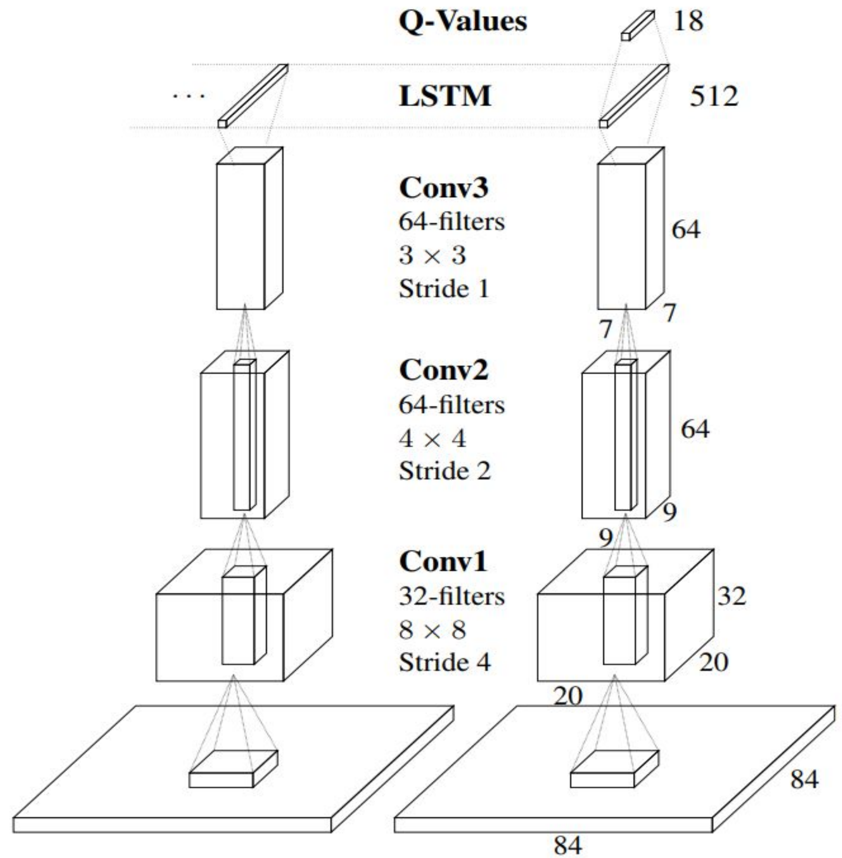
$$Q_t(s_i, a) = \sum_i b_t(s_i) Q_t(s_i, a)$$

  - Select action that yields the highest value.

Recent methods: DQN + LSTMs

- Deep Recurrent Q-Learning for Partially Observable MDPs (Hausknecht et. al., 2017).

- Track history using hidden states of LSTM.



Predictive State Representations

- Can we just look at how well we can predict the future, rather than history?
- Predictive State Representations: A New Theory for Modeling Dynamical Systems (Singh et. al.).
- PSRs rely solely on observable quantities; unlike POMDPs.
- Tests: Future observation-action sequences.
- PSR: Set of tests + Probabilities that tests are true
- Potentially more reliable with strong representational power.

Predictive State Representations

- Tests: Future observation-action sequences.
- PSR: Set of tests + Probabilities that tests are true
- No notion of underlying state space -> No need to keep history.