Partially Observable Markov Decision Process

R04922023 屠政皓

Markov Decision Process (MDP)

- 4-tuple (S, A, R, T)
  - S: set of environment states
  - A: set of actions that agent can execute
  - T: stochastic transition function $T(s, a, s') = Pr(s'|s, a)$
  - R: reward function $R(s, a)$ modeling the utility of the current state and the action execution
- know completely what is the current state, and state transition determined by the state and action

Partially Observable Markov Decision Process (POMDP)

- 7-tuple (S, A, T, R, O, Ω, γ)
  - S, A, T, R are the same as MDP
  - O: the probability of observing o in state s $O(s, o) = Pr(o|s)$
  - Ω: set of all possible observations
  - γ: discounted factor indicating the rate that rewards are discounted at each step
- unsure which state we are in

Example: Baby Crying Problem

$h_0 : not\ hungry$
$h_1 : hungry$

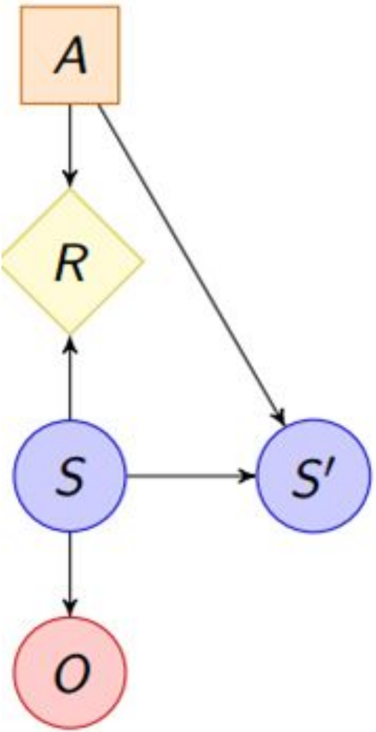
$c_0 : not\ crying$
$c_1 : crying$

$f_0 : not\ feed$
$f_1 : feed$

$R(h_0, f_1) = -5 \ R(h_1, f_1) = -15$
$R(h_0, f_0) = 0 \ R(h_1, f_0) = -10$

$Pr(c_0|h_0) = 0.9 \quad Pr(c_1|h_0) = 0.1$
$Pr(c_0|h_1) = 0.2 \quad Pr(c_1|h_1) = 0.8$

$Pr(h_0|h_0, f_0) = 0.9 \quad Pr(h_1|h_0, f_0) = 0.1$
$Pr(h_0|h_0, f_1) = 1.0 \quad Pr(h_1|h_0, f_1) = 0.0$
$Pr(h_0|h_1, f_0) = 0.0 \quad Pr(h_1|h_1, f_0) = 1.0$
$Pr(h_0|h_1, f_1) = 1.0 \quad Pr(h_1|h_1, f_1) = 0.0$



Belief Update

- consider current belief b and updated belief b', action a, observation o,

  $b = (h_0, h_1)$

  $b'(s') \propto \Sigma_{s \in S} Pr(s'|s,a) Pr(o|s') b(s)$

- Example:

  $b_0 = (0.5, 0.5)$

  $not\ \ feed,\ \ crying$

  $b_1(h_0) = b_0(h_0) Pr(h_0|h_0, f_0) Pr(c_1|h_0) + b_0(h_1) Pr(h_0|h_1, f_0) Pr(c_1|h_0)$

  $b_1(h_1) = b_0(h_0) Pr(h_1|h_0, f_0) Pr(c_1|h_1) + b_0(h_1) Pr(h_1|h_1, f_0) Pr(c_1|h_1)$

  $b_1 = (0.0928, 0.9072)$

Belief Update

$b_1 = (0.0928, 0.9072)$

$feed,\ not\ crying$

$b_2 = (1.0, 0.0)$

$not\ feed,\ not\ crying$

$b_3 = (0.9759, 0.0241)$

$not\ feed,\ not\ crying$

$b_4 = (0.9701, 0.0299)$

$not\ feed,\ crying$

$b_5 = (0.4624, 0.5376)$

......

POMDP and Belief-State MDP

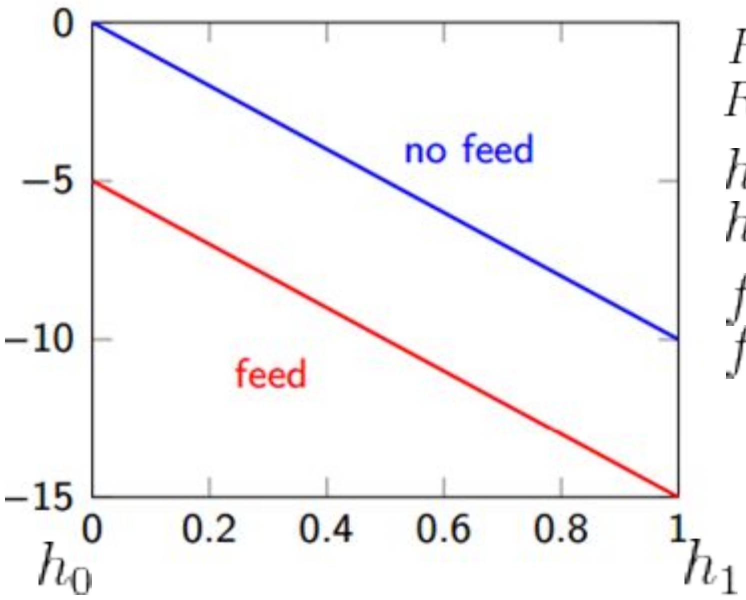
- POMDP is a MDP when states are belief states
- belief state is a probability distribution over the states of original POMDP
- transition probability is the product of actions and observations
- reward becomes the expected reward according to the belief

Solving POMDP

- B: set of belief states
- policy π : B $\rightarrow$ A
- find a policy that maximizes $E[\Sigma_t \gamma^t R(b_t, a_t)|\pi]$

Alpha Vector

- a vector with |S| dimensions
- first consider doing an action in a initial belief state and get expected reward
-



$$R(h_0, f_1) = -5 \quad R(h_1, f_1) = -15$$
$$R(h_0, f_0) = 0 \quad R(h_1, f_0) = -10$$

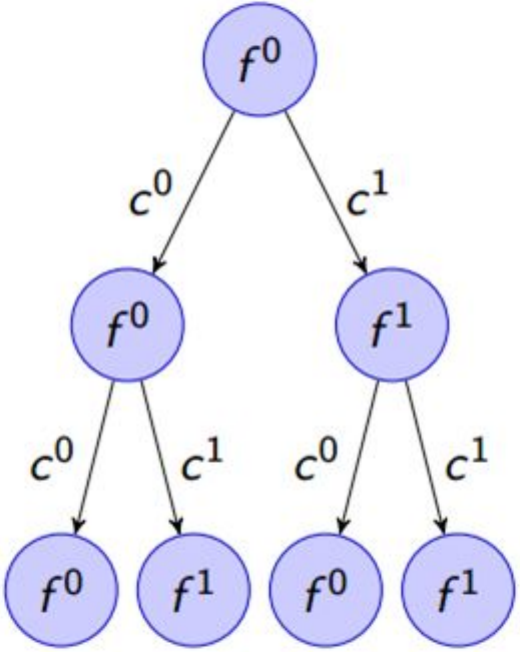
$$h_0 : not\ hungry$$
$$h_1 : hungry$$

$$f_0 : not\ feed$$
$$f_1 : feed$$

$$\alpha_{f0} = (0, -10)$$
$$\alpha_{f1} = (-5, -15)$$

Conditional Plans

- specifies what to do from a initial belief state after each possible observations up to a certain horizon
- 3-step conditional plan

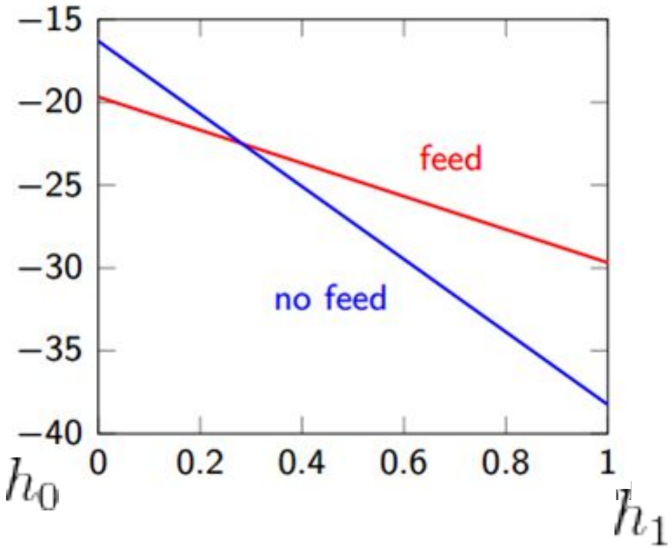


Value Iteration

- $U^*(s) = max_{a \in A}[R(s,a) + \gamma \Sigma_{s' \in S} T(s'|s,a) U^*(s')]$

  $U_1^*(s) = max_{a \in A} R(s,a)$
  $U_2^*(s) = max_{a \in A}[R(s,a) + \gamma \Sigma_{s' \in S} T(s'|s,a) U_1^*(s')]$
  ......

- input: A POMDP
- output: a set of alpha vectors
- for a belief state b, the action is $argmax_{a \in A} b \cdot \alpha_a$
- number of alpha can grow up exponentially



Point-Based Value Iteration and Optimization

- may not need to consider all the belief states
- Point-Based Value Iteration (PBVI)
  - approximate the solution by only consider a finite set of belief
  - the approximation error can be bounded
- compile the output of the PBVI to an finite state machine and can do further optimization on the size of the FSM

Reference

- Pineau, J., Gordon, G., & Thrun, S. (2003, August). Point-based value iteration: An anytime algorithm for POMDPs. In IJCAI (Vol. 3, pp. 1025-1032).
- Kochenderfer, M. J., Amato, C., Chowdhary, G., How, J. P., Reynolds, H. J. D., Thornton, J. R., ... & Vian, J. (2015). Decision making under uncertainty: theory and application. MIT press.
- Grzes, M., Poupart, P., Yang, X., & Hoey, J. (2014). Energy Efficient Execution of POMDP Policies.