Last Time

Offline POMDP Algorithms

- PBVI
- SARSOP

$\alpha_\pi[s]$

Index — state
Numbers — Value of executing π at state s
Correspond — π

$$\hat{V}(b) = \alpha_\pi^T b$$

This Time

Formulation Approximations
Object Oriented Programming in Julia in HW5

Disappointing Facts:

Infinite Horizon POMDPs are Undecidable
Finite Horizon POMDPs are PSPACE complete.
Among the hardest problems that can be solved using a polynomial amount of space

~~AMY~~.

POMDP algorithms (likely) have exponential complexity.

Numerical Approximation	Formulation Approximations
Offline, Online	
Solve the original problem approximately	Solve an approximate problem exactly or approximately

Name	Description	Properties	Usefulness
Certainty-Equivalence	control as if the true state is mean of belief	Optimal for LQG	★★★★★
QMDP	Full observability after 1 time step	QMDP pseudo α's are upper bound for true V	★★★★★
FIB	Takes 1 observation into account	tighter upper bound than QMDP	★★
Hindsight Opt (HOP)	Hindsight knowledge of state and outcome uncertainty	Looser Upper Bound than QMDP	★★★★
Last k observations "k Markov"	pretend that last k observations make up the state and solve that MDP	Great for Atari	★★★★
Open Loop no observations	Choose sequence of actions	Good if Alleatory is low, and epistemic is hard to reduce	★★★
Most likely obs	Plan assuming $b' = \tau(b, a, \hat{o}(b))$	No observation branching	★★★

POMDP Objective

$$\pi^* = \underset{\pi: \mathcal{B} \to \mathcal{A}}{\operatorname{argmax}} \; E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(b_t))\right]$$

Certainty Equivalence

$$\pi_{CE}(b) = \pi_{MDP}\left(\underset{s \sim b}{E}[s]\right)$$

or mode
or median

Optimal for one very special POMDP

LQG

$$s' \sim N(As + Ba, V) \qquad s_0 \sim N(\mu_0, \Sigma_0)$$
$$o \sim N(Cs' + Du, W) \qquad R(s,a) = s^T Q s + a^T R a$$

Bayesian Update — Kalman Filter

$$b_t = N(\mu_t, \Sigma_t)$$

Kalman Gain →

$$\Sigma_{t+1} = A\left(\Sigma_t - \Sigma_t C^T \left(C\Sigma_t C^T + W\right)^{-1} C\Sigma_t\right) A^T + V$$
$$L = A\Sigma_t C^T \left(\Sigma_t C^T + W\right)^{-1}$$
$$\mu_{t+1} = A\mu_t + Ba + L(o_t - C\mu_t)$$

Solution to LQR MDP $\quad a_t = -K s_t$

Solution to LQG POMDP $\quad a_t = -K \mu_t$

Certainty Equivalence Principle

Works pretty well for any problem where belief is unimodal

QMDP

$$\alpha_a[s] = Q^*_{MDP}(s,a)$$

$$\pi_{QMDP}(b) = \text{argmax}_a \; \mathop{E}_{s\sim b}\left[Q_{MDP}(s,a)\right] = \text{argmax}_a \; \alpha_a^T b$$

Breakout Rooms

$b = U(\{1,2\})$ $\quad \pi^*(b) = +10$

What is $\pi_{QMDP}(b)$

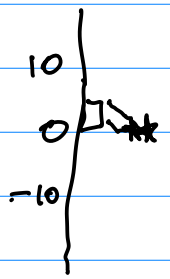
Steps

figure out $Q^*_{MDP}(s,a)$ — what would you do if you knew the state

$$\text{argmax}_a \; \mathop{E}_{s\sim b}\left[Q_{MDP}(s,a)\right]$$

$Q^*_{MDP}(1,-1) = 100-1$ $\quad Q^*_{MDP}(1, a\neq -1) =$ worse than -1

$Q^*_{MDP}(2,-1) = 100-2$ $\quad Q^*_{MDP}(2, a\neq -1) =$ worse than -1



QMDP is bad at costly information gathering

FIB

$$\pi_{FIB}(b) = \text{argmax}_a \; \alpha_a^T b$$

loop

$$\alpha_a^{(k+1)}[s] = R(s,a) + \gamma \sum_o \max_{a'} \sum_{s'} Z(o|a,s')\, T(s'|s,a)\, \alpha_{a'}^{(k)}[s']$$

HOP

$$\pi_{\text{HOP}}(b) = \underset{a^i_{t:\infty}}{\operatorname{argmax}} \frac{1}{m} \sum_{i=1}^{m} \gamma^t R(s^i_t, a^i_t)$$

subject to $s^i_{t+1} = G(s^i_t, a^i_t, w_i)$

$a^i_t = a^j_t \quad \forall\, i, j$