Part 3 - Collaboration, Assistance, Coordination

Today: MDP / POMDP formulations

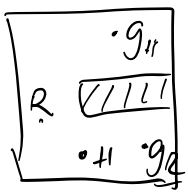
Collaboration:
- shared objective — shared workspace
- H-R team
- objective typically known by both

Assistance:
- shared objective — shared workspace or shared autonomy
- only H knows objective
- R's job is to help, despite uncertainty

Coordination:
- different objectives — shared workspace
- H's objective not necessarily known to R and vice versa

↳ simplified: avoidance: H is obj unknown, R needs to avoid H

1) Avoid as an MDP



$S = (S_H, S_R)$

$a = a_R$

$R(S,a)$: $S_H \neq S_R$, $S_R \neq$ ass, efficiency ⋮

$P(S' \mid S, a)$: $P(S_R' \mid S_R, a_R)$ ✓

$$P(S_H' \mid S, a_R) = \sum_{a_H} P(S_H', a_H \mid S, a_R)$$

$$= \sum_{a_H} \underbrace{P(S_H' \mid S_H, a_H)}_{\text{\# motion model}} \underbrace{P(a_H \mid S, a_R)}_{\text{\# policy}}$$

↓ avoid simplifies this to $P(a_H \mid S_H)$

1) learn a H policy from H action data

learn $\hat{\pi}_H = S_H \mapsto a_H$

compute $\pi_R^* = BR(\hat{\pi}_H)$ (i.e. solve MDP above)

predict $\hat{\xi}_H$ from env. params

$\xi_R^* = \arg\max_{\xi_R} U(\xi_R, \hat{\xi}_H)$

2) model H as (approx.) rational

assume/learn via IRL $R_H(S_H, a_H)$

↳ MDP_H: $S_H, a_H, P(S_H' \mid S_H, a_H), R_H$

↳ compute $\hat{\pi}_H = \pi_H^*$

OR $\hat{\pi}_H(a_H \mid S_H) \propto \exp(\beta Q_H(S_H, a_H))$

compute $\pi_R^* = BR(\hat{\pi}_H)$

$\hat{\xi}_H = \arg\max_{\xi_H} U_H(\xi_H)$

OR $P(\xi_H) \propto \exp(\beta U_H(\xi_H))$

$\xi_R^* = \arg\max_{\xi_R} U(\xi_R, \hat{\xi}_H)$

OR expectation.

Note: Same (static) goal/reward for H as in training
assumes H ignores R's existance

II Collaborate as an MDP
H no longer ignores R
1) learn a H policy from H-H collaboration data
(same as before, except H-H data, and R's reward)
2) solve for joint collaborative policies
$S=(S_R, S_H)$ $a=(a_R, a_H)$ $P(s'_R, s'_H | s_R, s_H, a_R, a_H)$ $R(s,a)$ — achieve task
a) compute $\hat{\pi}: S_R, S_H \mapsto a_R, a_H \hookrightarrow \pi_H$
$(s^*_R, s^*_H) = \underset{s_R, s_H}{\text{argmin}}\ H(s_R, s_H)$
note: $s^*_R = \underset{s_R}{\text{argmin}}\ H(s_R, s^*_H)$

III Assist as a POMDP
$S=(S_R, S_H, \theta)$ ← human's goal/intent/preferences
$a=a_R$
$R(s,a)$ - achieve θ
$o=a_H$
$P(s'|s,a)$ - depends on π_H as before
$P(o|s')$ - given by π_H
- keep belief on θ
- info gather (when valuable)
- do things that are useful now, knowing more info comes l