Recoverability and Safe Policies

Let T be the recall time

s_0 is the desired home state

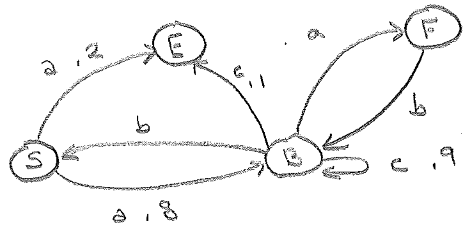
def: π_0 is δ-safe wrt s_0 & stopping time T iff

$$\exists \pi_r \text{ s.t. } E_\beta E^{P}_{s_0,\pi_0}\left[E^{P}_{s_T,\pi_r}[B_0]\right] \geq \delta$$

in general NP hard

→ Alternative ideas of safety?

1) safe states or actions?



$\delta = .8$ safe π: a a b a b ...

unsafe a c c c

2) 1 return π for each possible potential sample from belief

2 MDPs, prob of each is 0.5

see examples in paper

if total expected reward under any MDP drawn from the belief is bounded $\in [0,1]$

then $(P_{sas'} - E_\beta[P_{sas'}]) E_{s',\pi,P}[V] \geq P_{sas'} - E_\beta[P_{sas'}]$ if $P_{sas'} - E_\beta[P_{sas'}] < 0$

and if cap at 0 then always an underestimate of term

thm 2, let β be a belief s.t. for any policy π and any starting state s, the total expected reward in any MDP drawn from the belief is between 0 an 1; i.e. $0 \leq E^P_{s,\pi}[V] \leq 1$, β-almost surely. Then the following bound holds for any policy π & starting state s $E_\beta E^P_{s,\pi} \sum_{t=0}^{\infty} R_{S_t,A_t} \geq E^P_{s,\pi} \sum_{t=0}^{\infty} (E_\beta[R_{S_t,A_t}] + \sigma^\beta_{S_t,A_t})$ where $\sigma^\beta_{s,a} \equiv \sum_{s'} E_\beta[\min(0, P_{sas'} - E_\beta[P_{sas'}])]$

lemma 3: given belief β & policy π, there exists a policy dependent correction $\sigma^{\beta,\pi}$ such that the MDP w/transition measure $p \equiv E_\beta P$ and rewards $r + \sigma^{\beta,\pi}$ where $r \equiv E_\beta R$ has the same expected total return as the belief for any initial distribution. Formally

$$\forall \rho \quad E_\beta E^P_{\rho,\pi} \sum_{t=0}^{\infty} R_{S_t,A_t} = E^p_{\rho,\pi}\left(r_{s,a} + \sigma^{\beta,\pi}_{s,a}\right)$$

$$\sigma^{\beta,\pi}_{s,a} \equiv \sum_{s'} E_\beta\left[\left(P_{s,a,s'} - E_\beta\left[P_{s,a,s'}\right]\right) E_{s',\pi,P}[V]\right]$$

proof: Markov property under belief β reads

$$(1A)\quad E_\beta E^P_{s,\pi}[V] = \sum_a \pi_{s,a} \underbrace{E_\beta[R_{s,a}]}_{\equiv r_{s,a}} + \sum_a \pi_{s,a} \sum_{s'} E_\beta\left[P_{sas'} E^P_{s',\pi}[V]\right]$$

Markov property assuming expected transition frequencies & expected rewards with corrections is

$$(1B)\quad E^p_{s,\pi}[\bar V] = \sum_a \pi_{s,a}\left(r_{s,a} + \sigma^{\beta,\pi}_{s,a}\right) + \sum_a \pi_{s,a} \sum_{s'} p_{s,a,s'} E^p_{s',\pi}[\bar V]$$

let $\Delta_s \equiv E_\beta E^P_{s,\pi}[V] - E^p_{s,\pi}[\bar V]$

subtract, 1A − 1B

$$\Delta_s = -\sum_a \pi_{s,a}\sigma^{\beta,\pi}_{s,a} + \sum_a \pi_{s,a} \sum_{s'}\left(E_\beta\left[P_{sas'} E^P_{s'\pi}[V]\right] - p_{sas'} E^p_{s'\pi}[\bar V]\right)$$

where used defn of $r_{s,a}$

$$= \sum_a \pi_{s,a} \sum_{s'} E_\beta\left[\left(\underbrace{E_\beta[P_{s,a,s'}]}_{\equiv p_{s,a,s'}}{}^{(1)} - P_{s,a,s'}{}^{(2)}\right) E_{s',\pi,P}[V]\right]$$

sub in defn of σ

$$+ \sum_a \pi_{s,a} \sum_{s'}\left(E_\beta\left[P_{s,a,s'} E_{s',\pi,P}[V]\right]^{(3)} - p_{sas'} E^p_{s',\pi}[\bar V]^{(4)}\right)$$

$$= \sum_a \pi_{s,a} \sum_{s'} p_{s,a,s'} E_\beta E_{s',\pi,P}[V]$$

$p_{s,a,s'}$ is indep of β and (2) & (3) cancel

$$- \sum_a \pi_{s,a} \sum_{s'} p_{sas'} E^p_{s',\pi}[\bar V]$$

$$= \sum_a \pi_{s,a} \sum_{s'} p_{s,a,s'} \Delta_{s'}$$

defn of $\Delta_{s'}$

Δ satisfies the same eqn as a value function in a MDP w/transition measure p and zero rewards.
Since the value func in such an MDP is uniquely defined & identically 0, $\Delta_s = 0$. □

but in general don't know V & transformation requires V
if have a bound on V, can use to get a lower bound

$$\max_{\pi_0, \pi_r} E^{rp}_{s_0 \pi_0} \sum_t r_{s_t, A_t} + \xi^{\beta}_{s_t A_t} \qquad (5A)$$

$$\text{subject to} \quad E^{rp}_{s_0 \pi_0} \sum_{t=0}^{\infty} (1-\gamma) V_{s_t} + \gamma \sigma^{\beta}_{s_t, A_t} \geq \delta \qquad (5B)$$

$$\text{and} \quad V_s = \underbrace{E^{p \cdot (1-1_{s=s_0})}_{s \pi_r} \sum_{t=0}^{\infty} \left(1_{s_t = s_0} + (1 - 1_{s_t = s_0}) \sigma^{\beta}_{s_t, A_t}\right)} \qquad (5c)$$

value func of MDP with transition measure $p(1 - 1_{s=s_0})$ and reward $1_{s=s_0} + (1 - 1_{s=s_0}) \sigma^{\beta}_{s, a}$ under policy π_r

The return policy π_r only appears in Equation 5c.
Therefore the above optimization can be done in 2 steps

Step 1: solve for optimal return policy π_r^* given expected transition measure p & also compute optimal value function V^*_s,

$$V^*_s = \max_{\pi_r} E^{p(1-1_{s=s_0})}_{s, \pi_r} \sum_{t=0}^{\infty} \left(1_{s_t = s_0} + (1 - 1_{s_t = s_0}) \sigma^{\beta}_{s_t, A_t}\right)$$

Step 2: compute optimal exploration policy π^*_0 under strict safety constraint by solving a constrained MDP

$$\max_{\pi_0} E^{rp}_{s_0, \pi_0} \sum_t r_{s_t, a_t} + \xi^{\beta}_{s_t, A_t}$$

$$\text{s.t.} \quad E^{rp}_{s_0, \pi_0} \sum_t \left((1-\gamma) V^*_{s_t} + \gamma \sigma^{\beta}_{s_t, A_t}\right) \geq \delta$$

they solved as a linear program

grid world experiment
exploration: adapted R-max
safety costs

$$\sigma^{\beta}_{s, a} = -2 E_{\beta}[p_{s,a}](1 - E_{\beta}[p_{s,a}]) = -2 \mathrm{Var}_{\beta}[p_{s,a}]$$