Recoverability and Safe Policies

Let $T$ be the recall time
So is the desired home state
def: $\pi_0$ is $\delta$-safe wrt $s_0$ at stopping time $T$ iff
\[ \exists \pi \text{ s.t. } E_{\pi_0} E_{s_0, \pi_0}^P [E_{s_0, \pi_0}^P [B_0]] \geq \delta \]

in general NP hard

→ Alternative ideas of safety?

1) Safe states or actions?

2) I return $\pi$ 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 \( (p_{ss'} - E_B [p_{ss'}]) E_s \pi \mathbb{E} [V] \geq p_{ss'} - E_B [p_{ss'}] \) if \( p_{ss'} - E_B [p_{ss'}] \leq 0 \)

and if cap at 0 then always an underestimate of term.

**thm 2.** Let \( B \) be a belief set. For any policy \( \pi \) and any starting state \( s \), the total expected reward in any MDP drawn from the belief is between 0 and 1; i.e., \( 0 \leq E^B_s \pi \mathbb{E} [V] \leq 1 \) almost surely.

Then the following bound holds for any policy \( \pi \) and starting state \( s \):

\[
E^B_s \mathbb{E} \sum_{t=0}^{\infty} R_{st} \pi_t \geq E^B_s \pi \sum_{t=0}^{\infty} (E_B [R_{st} \pi_t] + \sigma^{B}_{st} \pi_t) \quad \text{where} \quad \sigma^{B}_{st} \equiv \sum_{s'} E_B [\min(0, p_{ss'} - E_B [p_{ss'}])] \]
Lemma 3: given belief $\beta$ and policy $\pi$, there exists a policy dependent correction $\sigma^{B,\pi}$ such that the MDP $w/transition$ measure $p = \beta \rho_p$ and rewards $r = \sigma^{B,\pi}$ where $\bar{r} = \beta \rho_r$ has the same expected total return as the belief for any initial distribution. Formally

$$V_p = \beta \rho_p, \sigma^{B,\pi} \sum_{t=0}^{\infty} R s_{t+1} = \beta \rho_p, (r_{s,a} + \sigma^{B,\pi}_{s,a})$$

$$\sigma^{B,\pi}_{s,a} = \sum_{s'} \beta \rho_s \left[ (p_{s,a,s'} - \beta \rho_s \rho_{s',\pi}) \rho_{s',\pi} \rho_p \left[V\right] \right]$$

Proof: Markov property under belief $\beta$ reads

$$(16) \; \beta \rho_s \rho_p \left[V\right] = \sum_{s'} \pi_{s,s'} \beta \rho_s \rho_p \left[R_{s,s'}\right] + \sum_{s'} \pi_{s,s'} \sum_{s''} \rho_{s'',\pi} \beta \rho_s \rho_p \left[P_{s,a,s''} \beta \rho_{s'',\pi} \rho_p \left[V\right] \right]$$

Markov property assuming expected transition frequencies $\rho_s$ and expected rewards with corrections $\sigma^{B,\pi}$ produces

$$(18) \; \beta \rho_s \rho_p \left[V\right] = \sum_{s'} \pi_{s,s'} \left(r_{s,a} + \sigma^{B,\pi}_{s,a}\right) + \sum_{s'} \pi_{s,s'} \sum_{s''} \rho_{s''} \rho_{s',\pi} \beta \rho_s \rho_p \left[P_{s,a,s''} \beta \rho_{s'',\pi} \rho_p \left[V\right] \right]$$

let $\Delta_s = \beta \rho_s \rho_p \left[V\right] - \beta \rho_s \rho_p \left[V\right]$

Subtract, 1A - 1B

$$\Delta_s = -\sum_{s'} \pi_{s,s'} \sigma^{B,\pi}_{s,a} + \sum_{s'} \pi_{s,s'} \sum_{s''} \left( \beta \rho_s \rho_p \left[P_{s,a,s''} \beta \rho_s \rho_p \left[V\right] \right] - p_{s,a,s''} \beta \rho_p \left[V\right] \right)$$

where used defn of $p_{s,a,s''}$

$$= \sum_{s'} \pi_{s,s'} \sum_{s''} \beta \rho_s \rho_p \left[P_{s,a,s''} \beta \rho_s \rho_p \left[V\right] \right] \rho_{s',\pi} \rho_p \left[V\right]$$

$$(4) \quad \text{sub in defn of } \sigma$$

$$+ \sum_{s'} \pi_{s,s'} \sum_{s''} \left( \beta \rho_s \rho_p \left[P_{s,a,s''} \beta \rho_s \rho_p \left[V\right] \right] - p_{s,a,s''} \beta \rho_p \left[V\right] \right)$$

$$= \sum_{s'} \pi_{s,s'} \sum_{s''} \rho_{s',\pi} \rho_p \left[V\right]$$

$$- \sum_{s'} \pi_{s,s'} \sum_{s''} \rho_{s',\pi} \rho_p \left[V\right]$$

$$= \sum_{s'} \pi_{s,s'} \sum_{s''} \rho_{s',\pi} \rho_p \left[V\right]$$

$\Delta_s$ 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 idenitically 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.
\[
\max_{\pi_t, \pi_r} \mathbb{E}_{s_0, \pi_0} \sum_{t=1}^{\infty} r_{st, at} + \delta \sigma_{st, at} \tag{5a}
\]
subject to
\[
\mathbb{E}_{s_0, \pi_0} \sum_{t=0}^{\infty} (1-\gamma) V_{st} + \gamma \sigma_{st, at} \geq \delta \tag{5b}
\]
and
\[
V_r = \mathbb{E}_{s_t, \pi_t} \sum_{t=0}^{\infty} (1.s_{st} = s_0 + (1-1.s_{st} = s_0) \sigma_{st, at}) \tag{5c}
\]

value function of MDP with transition measure \( p (1.1.s_{st} = s_0) \) and reward \( 1.s_{st} = s_0 + (1.s_{st} = s_0) \sigma_{st, at} \)

under policy \( \pi_r \)

The return policy \( \pi_r \) only appears in Equation 5c. Therefore the above optimization can be done in 2 steps.

1. Solve for optimal return policy \( \pi_r \) given expected transition measure \( p \) and also compute optimal value function \( V_r \):
\[
V_r = \max_{\pi_r} \mathbb{E}_{s, \pi_r} \sum_{t=0}^{\infty} (1.s_{st} = s_0 + (1-1.s_{st} = s_0) \sigma_{st, at})
\]

2. Compute optimal exploration policy \( \pi_e \) under strict safety constraint by solving a constrained MDP:
\[
\max_{\pi_0} \mathbb{E}_{s_0, \pi_0} \sum_{t=1}^{\infty} r_{st, at} + \delta \sigma_{st, at}
\]
subject to
\[
\mathbb{E}_{s_0, \pi_0} \sum_{t=1}^{\infty} ((1-\gamma)V_{st} + \gamma \sigma_{st, at}) \geq \delta
\]

They solved as a linear program.

Grid world experiment
Exploration: adapted R-max
Safety costs:
\[
\sigma_{st, at} = -2 \mathbb{E}_B \left[ P_{s, at} \right] (1-\mathbb{E}_B \left[ P_{s, at} \right]) = -2 \mathbb{V}_{2B} \left[ P_{s, at} \right]
\]