MDPs with discounted cost as MDPs with finite random duration

“Death and Discouting” by Adam Shwartz

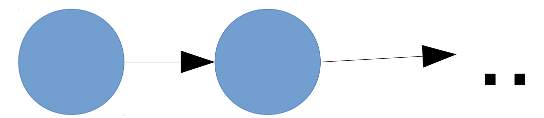
COMP 767 - Reinforcement Learning
January 20th

~Nicolas Angelard-Gontier

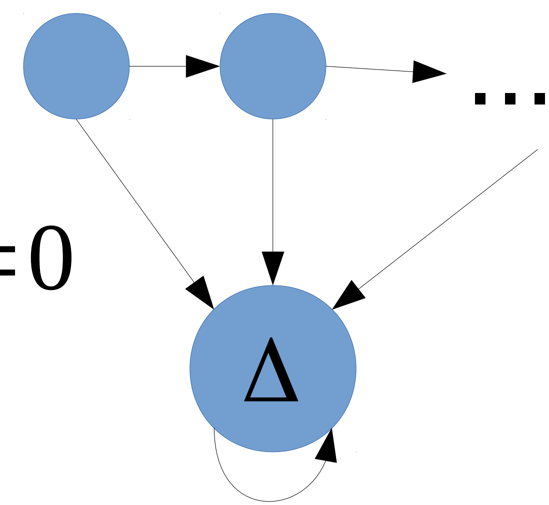
- They show that:
  - 1) MDPs with discounted cost ~ MDPs with finite random duration. Discout factor ~ life-span
  - 2) “an objective function which is a linear combination of several discounted costs does NOT, in general, model processes with several time scales”
- Focus on 1)

- M1: Classical discounted MDP with

$$V(x;\pi) = E_x^{\pi} \sum_{n=0}^{\infty} \beta^n c(x_n, a_n)$$



- M2: Now add an absorbing sate Δ ('cemetery') with:
  - $p(\Delta | x, a) = 1 - \beta$
  - $p(y | x, a) = \beta\, p(y | x, a)$
  - $p(\Delta | \Delta, a) = 1$ and $c_\Delta(\Delta, a) = 0$



$$V_\Delta(x;\pi_2) = \tilde{E}_x^{\pi_2} \sum_{n=0}^{\infty} c_\Delta(x_n, a_n)$$

- Let $\{\zeta_n\}$ be a sequence of {0,1} i.i.d. rnd.var. with mean β so that $x_n = \Delta \Leftrightarrow \zeta_n = 0$.
  T = first time step t when $\zeta_t = 0$

- Let $\{\zeta_n\}$ be a sequence of {0,1} i.i.d. rnd.var. with mean β so that $x_n = \Delta \Leftrightarrow \zeta_n = 0$.
  T = first time step t when $\zeta_t = 0$
- Lemma: Can replace discount factor of M1 by geometric time-horizon:

$$V(x;\pi) = E_x^\pi \sum_{n=0}^{\infty} \beta^n c(x_n, a_n)$$

$$= E_x^\pi \sum_{n=0}^{\infty} \left(\prod_{t=0}^{n-1} \zeta_t\right) c(x_n, a_n)$$

$$= E_x^\pi \sum_{n=0}^{T-1} c(x_n, a_n)$$

- Policies in M2 rely on richer info:

  $$\tilde{h}_n = x_0 \zeta_0 a_0 \dots x_{n-1} \zeta_{n-1} a_{n-1} x_n$$

- Given a policy π_2 in M2, we can define π in M1:
  - $\pi(.|x_0 a_0 \dots x_n) = \pi_2(.|x_0 1 a_0 \dots 1 x_n)$
- Conversely, given π, we can define π_2 :
  - if $\zeta_t = 1 \, \forall t < n$ then $\pi_2(.|\tilde{h}_n) = \pi(.|h_n)$
  - else $\pi_2(.|\tilde{h}_n) = a_\Delta$
- Thm: M1 and M2 are equivalent in the following sense:
  - $E_x^\pi \beta^n c(x_n, a_n) = E_x^\pi \left(\prod_{t=0}^{n-1} \zeta_t\right) c(x_n, a_n)$

    $$= \tilde{E}_x^{\pi_2} \left(\prod_{t=0}^{n-1} \zeta_t\right) c_\Delta(x_n, a_n) = \tilde{E}_x^{\pi_2} c_\Delta(x_n, a_n)¿$$

so: $V(x, \pi) = V_\Delta(x, \pi_2)$