687 2017-09-12

Note Title 9/12/2017

MDPs continued

Reward discounts: $\gamma \in [0, 1]$ $\quad J(\pi) \triangleq E\left[\sum_{t=0}^{\infty} \gamma^t R_t \mid \pi\right]$

- Can be intrinsic in agent preference
- Mathematically, $\gamma < 1$ insures $J(\pi) < \infty$ $\left(J(\pi) \leq \frac{\max R_t}{1-\gamma}\right)$
  - ↳ which also guarantees policy ordering is sensible

Terminal states: A terminal state always transitions to a special state $\overset{\infty}{s}$, the "terminal absorbing state":
- cannot leave
- zero reward always

Episode: When $\overset{\infty}{s}$ is reached the current trial (an episode) ends & we start over: $t \leftarrow 0$, $s_0 \sim d_0$

MDPs can be:
- Episodic → will always reach $\overset{\infty}{s}$
- Continuous → " never " "

} → Can be mixtures, too

RL is learning from interaction with the env. (at least P is not known) → usually will not estimate it

Planning is where P & R are known, can be solved w/out env. interaction

Note: RL can still be model-based → model built through interaction.

$H_t \triangleq (S_0, A_0, R_0, S_1, A_1, R_1, \ldots, S_t, A_t, R_t)$

history

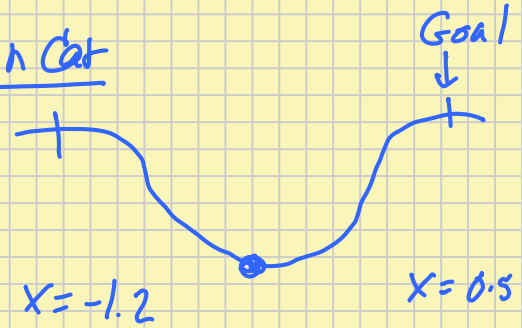
A trajectory is a history of an entire episode: H_∞

Discounted return of a trajectory $H_\infty = \sum_{t=0}^{\infty} \gamma^t R_t \triangleq G$

$J(\pi) = E[G|\pi]$

Example Problems Formulated as MDP

Mountain Car



$S = (x, v)$ ← position, velocity

A: $a \in \{\text{forward}, \text{neutral}, \text{reverse}\}$ = $\{+1, 0, -1\}$

Dynamics: $v_{t+1} = v_t + .001a - .0025 \cdot \cos(3x)$

$x_{t+1} = x_t + v_t$

If $x_{t+1} \notin [-1.2, 0.5]$ it is moved to this range, and $v_{t+1} \leftarrow 0$ (inelastic collision w/ boundary walls)

Terminal state: $x_t = 0.5$

Rewards: 1 @ goal, 0 elsewhere → works, but not informative (more later abt reward shaping)

Also: can do -1 everywhere (except terminates)

Initial state: 0.5 $\gamma = 1$ (Could model fuel + fuel usage, if we like.)

Markov Properties: $P(h,s,a,s') \triangleq Pr(S_{t+1}=s' \mid H_{t-1}=h \wedge S_t=s \wedge A_t=a)$

Markov assumption/property $Pr(S_{t+1}=s' \mid H_{t-1}=h \wedge S_t=s \wedge A_t=a) =$

$Pr(S_{t+1}=s' \mid S_t=s \wedge A_t=a)$

and rewards Markovian (likewise)

Markovian policy insures $Pr(A_t=a \mid S_t=s \wedge H_{t-1}=h) = Pr(A_t=a \mid S_t=s)$

$\pi(s,a)$ vs. $\pi(h,s,a)$

Non-Markovian → Markovian

$S_t \rightarrow (S_t, h_{t-1})$ But state space may become huge!