CLASE 2: MARKOV DECISION PROCESS

C2
H1

MARKOV REWARD PROCESS $(\underbrace{S, p}_{\text{markov}}, R, \gamma)$

γ → discount factor $\gamma \in [0,1]$

$$R_s = E[R_{t+1} \mid S_t = s]$$ REWARD FUNCTION

En general ; $R_{t+1} = \Psi(S_{t+1})$

Recordar: $E[g(X)] = \sum_x g(x) p(x)$

PONER EJEMPLO EN SLIDES.

RETURN $$G_t = R_{t+1} + \gamma R_{t+2} + \ldots = \sum_{k=0}^{+\infty} \gamma^k R_{t+k+1}$$

$\gamma \simeq 0$ corto plazo

$\gamma \simeq 1$ largo plazo

Obs: $\gamma > 0$ evita ciclos en el proceso de Markov.

VALUE FUNCTION $$v(s) := E[G_t \mid S_t = s]$$ (no depende de t pues el markov es homogéneo)

$$= E[R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \ldots) \mid S_t = s]$$ RATA

$$= E[R_{t+1} + \gamma G_{t+1} \mid S_t = s] = E[E[R_{t+1} + \gamma G_{t+1} \mid S_{t+1}] \mid S_t = s]$$

$$= E[R_{t+1} + \gamma v(S_{t+1}) \mid S_t = s] = R_s + \gamma E[v(S_{t+1}) \mid S_t = s]$$

Agregar antes $E[\Psi(X) . Y \mid \cancel{X}] = \Psi(X) E[Y \mid X]$

BELLMAN EQUATION $$v(s) = R_s + \gamma \sum_{s'} v(s') \underbrace{P(S_{t+1} = s' \mid S_t = s)}_{p_{s,s'}}$$

Obs: Dados $R, p \Rightarrow$ me da ecuaciones para v

$$\begin{bmatrix} v(1) \\ \vdots \\ v(S) \end{bmatrix} = \boxed{v = R + \gamma p v}$$

. $O(|S|^3)$ complejidad

. Programación Dinámica, MC ...

MARKOV DECISION PROCESS: $(\mathcal{S}, P, R, \gamma, \mathcal{A})$ C2 H2

$\mathcal{A}$ = conjunto de acciones

Obs: LA ACCIÓN DEL AGENTE MODIFICA EL AMBIENTE

$$P^a_{ss'} := P(S_{t+1}=s' \mid S_t=s, A_t=a)$$

$$R^a_s = E[R_{t+1} \mid S_t=s, A_t=a] \qquad R_{t+1}=\Psi(S_{t+1})$$

$$\begin{tikzcd} S_t \arrow[d, dashed] \arrow[dr] & & \\ A_t \arrow[r] & S_{t+1} \arrow[r] & R_{t+1} \end{tikzcd}$$

POLÍTICA

$$\pi(a|s) := P(A_t=a \mid S_t=s)$$

NO DEPENDE DE t!!!

Decide la acción de manera aleatoria en función del ambiente (s)

$$P(S_{t+1}=s' \mid S_t=s) = \sum_a P_{S_t=s}(S_{t+1}=s' \mid A_t=a)\, P_{S_t=s}(A_t=a)$$

PROB. TOTAL PARA LA CONDICIONAL

Mencionar condicional de la condicional:

$P_B(A|C) = P(A|B,C)$

$$= \sum_a \underbrace{\pi(a|s)\, P^a_{s,s'}}_{\text{MEZCLA DE CADENAS DE MARKOV}} =: P^\pi_{s,s'}$$

$$R^\pi_s := E[R_{t+1} \mid S_t=s] = \sum_a \pi(a|s)\, R^a_s$$

STATE-VALUE $\quad v_\pi(s) = E_\pi[G_t \mid S_t=s]$

Explicar diferencia entre ambas funciones!

ACTION-VALUE FUNCTION $\quad q_\pi(s,a) = E_\pi[G_t \mid S_t=s, A_t=a]$

DESARROLLAR PARA C/DEF EL EJEMPLO DE SLIDES

C2
H3

Bellman Expectation Equation

$$v_\pi(s) = E_\pi\left[R_{t+1} + \gamma\, v_\pi(S_{t+1}) \mid S_t = s\right]$$

$$q_\pi(s,a) = E_\pi\left[R_{t+1} + \gamma\, q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a\right]$$

$$\boxed{v_\pi(s)} = E_\pi\left[E\left[G_t \mid A_t\right] \mid S_t = s\right]$$

$$E\left[\quad \mid A_t = a\right] = \sum_\omega \omega\, P_{S_t=s}(\omega \mid A_t = a) = \sum_\omega \omega\, P\left(\omega \,\middle|\, \begin{matrix} S_t = s \\ A_t = a \end{matrix}\right)$$

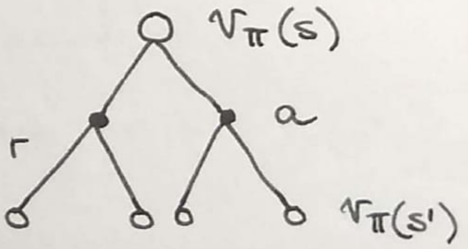
$$= \sum_a \left\{ \sum_\omega \omega\, P(\omega \mid S_t = s, A_t = a) \right\} P(A_t = a \mid S_t = s)$$

$$= \sum_a E_\pi\left[G_t \mid S_t = s, A_t = a\right] \pi(a \mid s) = \boxed{\sum_a q_\pi(s|a)\, \pi(a|s)}$$

$$\boxed{q_\pi(s,a)} = R_s^a + \gamma \sum_{s'} v_\pi(s')\, p_{ss'}^a \qquad \left(\begin{matrix}\text{by conditioning} \\ \text{on } S_{t+1}\end{matrix}\right)$$

Mezclamos $\boxed{v_\pi(s) = \sum_a \left[R_s^a + \sum_{s'} v_\pi(s')\, p_{ss'}^a\right] \pi(a|s)}$

$$\begin{tikzcd} S_t \arrow[d] \arrow[dr] & & \\ A_t \arrow[r] & S_{t+1} \arrow[r] & R_{t+1} \end{tikzcd}$$



$$q_\pi(s,a) = R_s^a + \gamma \sum_{s'} p_{ss'}^a \sum_{a'} \pi(a'|s')\, q_\pi(s',a')$$

$$\boxed{v_\pi = R^\pi + \gamma\, p^\pi v_\pi}$$

¿Para qué fórmulas recursivas?

FUNCIÓN DE VALOR ÓPTIMA:

$$v_*(s) = \max_{\pi} v_\pi(s)$$

$$q_*(s,a) = \max_{\pi} q_\pi(s,a)$$

DEFINICIÓN: $\pi \geq \pi'$ si $v_\pi(s) \geq v_{\pi'}(s) \quad \forall s$

TEOREMA: Para cualquier MDP: (Con Dynamic Programming diría que porque se entiende ~~porque~~ función)

- $\exists\ \pi_*$ óptima / $\pi_* \geq \pi \quad \forall \pi$
- $v_*(s) = v_{\pi^*}(s)$; $q_*(s,a) = q_{\pi^*}(s,a) \quad \forall s,a$

Obs: Si tenemos $q_* \Rightarrow \pi_*(a|s) := \begin{cases} 1 & \text{si } a = \underset{a}{\text{argmax}}\ q_*(s,a) \\ 0 & \text{cc} \end{cases}$

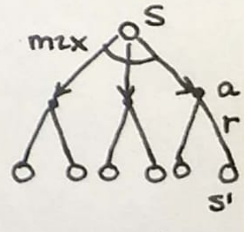
π_* es óptima y determinística.

ECUACIONES DE OPTIMALIDAD DE BELLMAN

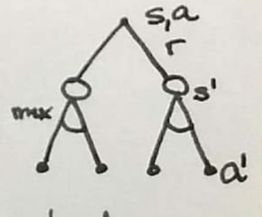
1) $$\boxed{v_*(s) = \max_{a \in A(s)} q_{\pi_*}(s,a)}$$

Recordar: $v_{\pi_*}(s) = \sum_a q_{\pi_*}(s,a)\, \pi_*(a|s)$

2) $$v_*(s) \overset{\cdots}{=} \max_a \sum_{s',r} p(s',r|s,a)\left[r + \gamma v_*(s')\right]$$



3) $$q_*(s,a) \overset{\cdots}{=} \sum_{s',r} p(s',r|s,a)\left[r + \gamma \max_{a'} q_*(s',a')\right]$$



NO LINEALES ∴ no hay fórmula cerrada → Métodos iterativos!!!