

## Capítulo 3: Procesos de Decisión Finitos

de Markov.

### 1. The Agent-Environment Interface.

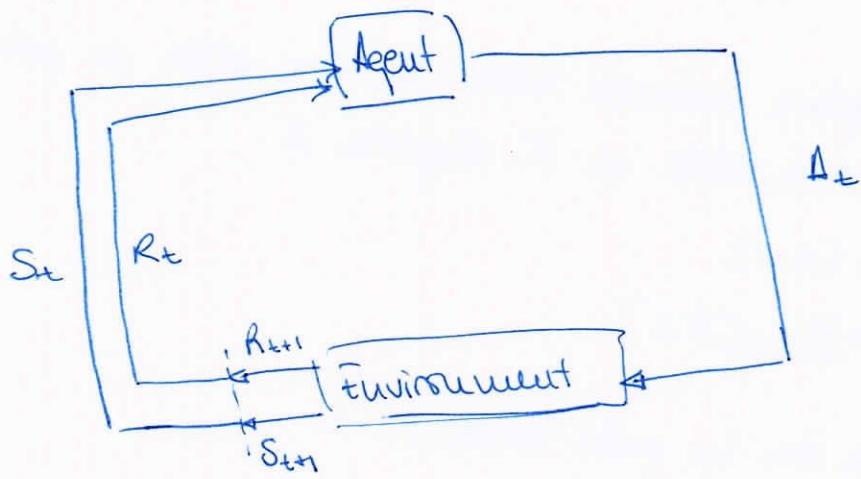
Agent = controller

Environment = Controlled System / Plant

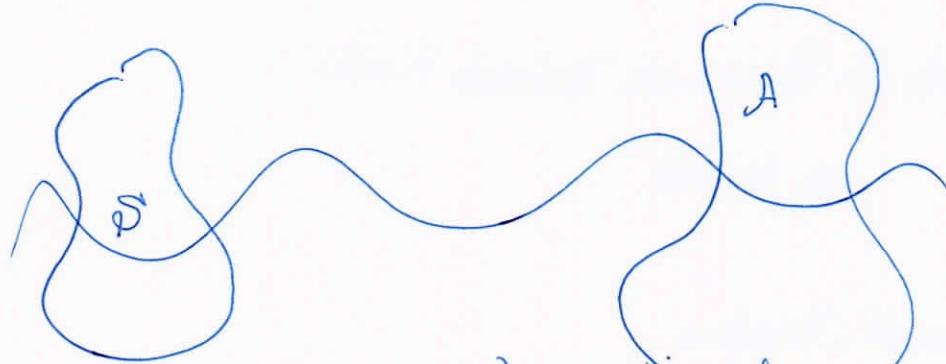
Acción = Control Signal

Cada instancia del problema RL se llama tarea.

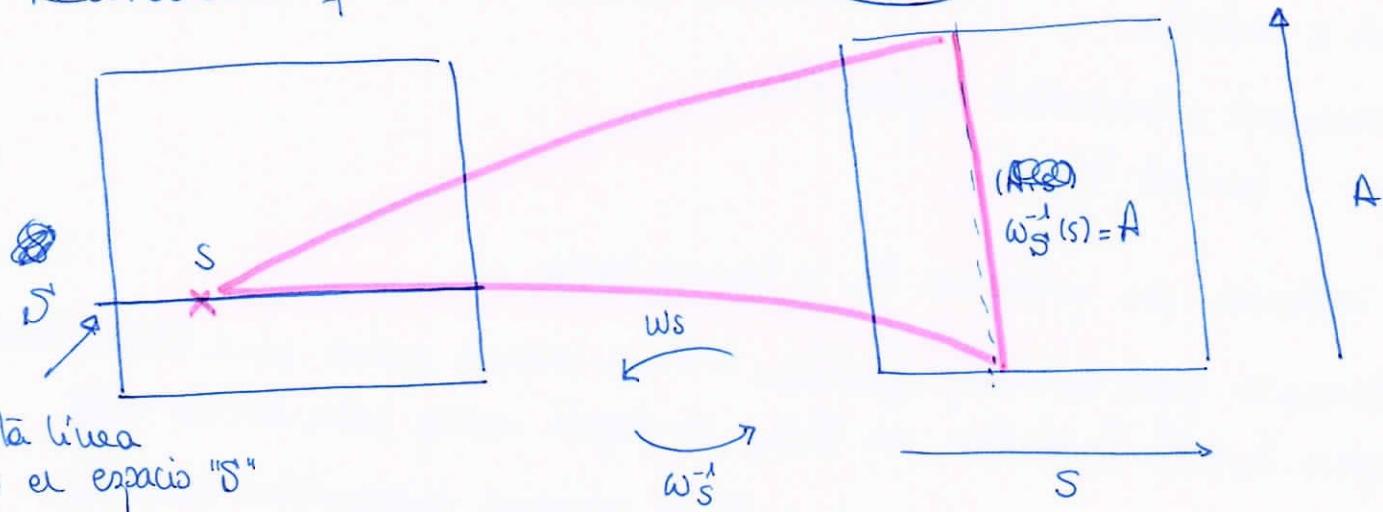
Consideremos pasos de tiempo finitos.  $\forall t$ , el agente recibe una REPRESENTACIÓN del espacio,  $S_t \in S$ . Basándose en  $S_t$ , el agente exige una acción,  $A_t$ .  $A_t \in A(S_t)$ , donde  $A(S_t)$  es el conjunto de acciones disponibles para  $S_t$ . Un paso temporal después, el agente recibe (en parte en consecuencia) una recompensa  $R_{t+1} \in \mathbb{R} \subset \mathbb{R}$ , y pasa a  $S_{t+1}$ .



En cada  $t$ , el agente genera un mapa de  $S$  a probabilidades sobre  $A(S_t)$ . Este mapa es la política o estrategia,  $\Pi_t$ .



Recordemos que es una dist. condicionada:



$$w_S : S \times A \rightarrow S$$

$$f(s) = A$$

cond.

Denotemos  $A$  al  $\sigma$ -álgebra inducida sobre  $A$ . Una dist. de prob. sobre  $A$  es una distribución sobre el  $\sigma$ -álgebra  $A$ :

$$\underline{P}_{A|S} : A \times S \rightarrow [0,1]$$

$$(A, s) \mapsto \underline{P}_{A|S}[A, s]$$

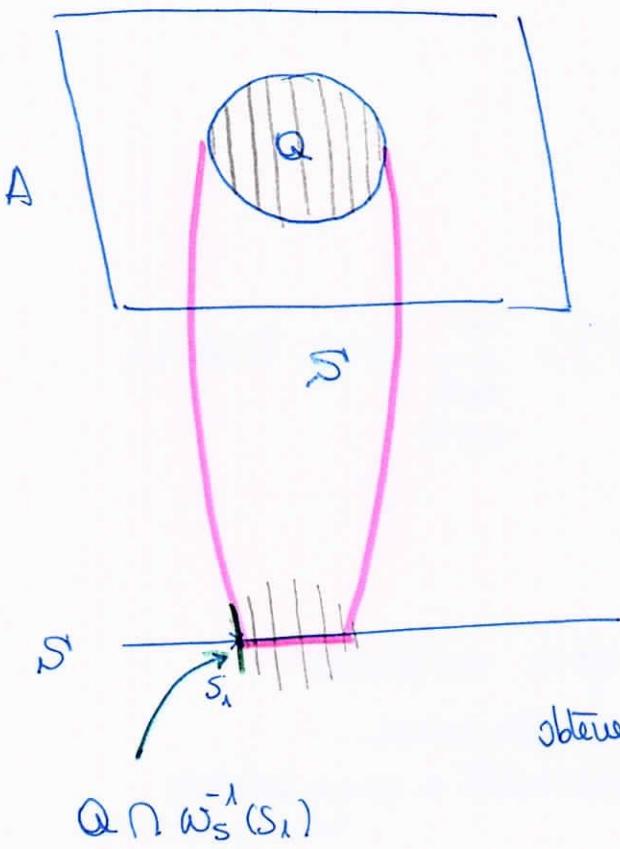
Dado un punto fijo " $s$ ", nos da la prob. sobre  $A$ .

~~Dado un subconjunto de  $A$ ,  $B \in A$ , la distribución condicional define una función de~~

Si lo que fijamos es  $B \in A$ , entonces la dist. de prob. condicionada nos da una función  $f : S \rightarrow [0,1]$  (que no es dist. necesariamente) que nos dice ~~le~~ la probabilidad de ese set varía al cambiar de fibra.

si tiene

si tuviéramos una distribución marginal  $\Pi_A$ , podríamos entonces definir una distribución sobre el espacio total  $A \times S$ :



Aplicamos ~~W(AxS)~~  $\omega_S$   
 $w_S: S \times A \rightarrow S$

(Todos los

estados con la misma

acción cotapsan en lo mismo)

(Todas las acciones a las que corresponde el mismo  
mismo estado cotapsan en un punto)

la prob. de cada fibra se calcula con  
 $P_{A|S}$

~~que depende de s~~. Si fijamos

$Q$ , entonces  $P(S_1) = P_{A|S} [Q \cap \omega_S^{-1}(S_1), S_1]$

obtenemos una función sobre los estados.  
probabilidad de la intersección  $Q \cap \omega_S^{-1}(S_1)$  y

s\_1

Si tenemos en cuenta todos los  $s \in S$  y conocemos la probabilidad marginal  $\Pi_S(s)$ , obtenemos  $\underline{P}_A(Q) = \mathbb{E}_S [P_{A|S} (Q \cap \omega_S^{-1}(S_1), S_1)]$

En otras palabras:  $\underline{P}_{A|B}$  nos permite definir

la distribución de prob.  
sobre "A" si fijamos  
b e B.

una función  $f: B \rightarrow [0,1]$   
que nos indique cómo cambia  
la prob. de un conjunto  
de A conforme cambiamos  
los b de fibra.

\* Es decir, nos da la probabilidad de la  
fibra condicionada sobre el punto  $s$ .

la prob. de  $Q$  depende de la probabilidad  
de las fibras que <sup>intervienen</sup> en  $Q$  multiplicadas  
por la prob. de cada punto  $s \in Q$  y su medida.

$$\underline{P}_A(Q) = \sum_{\substack{s \\ \in \\ A \times S}} \Pi_S(s) \cdot \underbrace{P_{A|S} (Q \cap \omega_S^{-1}(s), s)}_A$$

y la distribución condicional  $\Pi(A|S)$  nos da la distribución de probabilidad sobre las acciones dado un estado.

Si conocemos  $\Pi(A|S)$  (o su forma) y conocemos  $\Pi(S)$ :

$$P_{AS} = \sum_{S \in S} \Pi(S) \cdot \Pi(A|S)$$

$$P_A = \sum_{A \in Q} P_{AS}$$

$$P_{AS}[Q] = \mathbb{E}_S [ P_{AS} [ Q \cap \omega^*(S), S ] ] = \sum_{\substack{S \in Q \\ a \in Q}} \Pi(S) \cdot \Pi(a|S)$$

Este planteamiento es flexible:

- los pasos temporales no tienen por qué ser uniformes.
- las acciones pueden ser de bajo o alto nivel.
- los estados pueden ser también abstractos o muy básicos

la frontera entre agente y ambiente no siempre está clara! (no siempre es la misma que la frontera física).

Regla general: "todo aquello que uno pueda cambiar arbitrariamente por el agente, se considera parte de él, y por tanto parte del ambiente".

El objetivo de la recompensa es hacer siempre favor del agente, para que no lo padezca.

### 3.2 Goals and Rewards. (y Return)

$$\text{R}_t \in \mathbb{R} \quad \forall t$$

El propósito de RL se puede resumir (informalmente) como la maximización del valor esperado de la suma acumulada de una señal exterior recibida (que llamamos recompensa).

Se una secuencia de recompensas

$$R_{t+1}, R_{t+2}, R_{t+3}, \dots$$

¿Cómo definimos una función de esa secuencia a optimizar como objetivo del RL?  $\rightarrow G_t$ , return.

En el caso más simple:

$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \dots + R_T$$

Si  $T$  no está bien determinado porque la naturaleza de la interacción no es episódica (y especialmente cuando  $T \rightarrow \infty$  potencialmente),  $G_t$  no estará bien definido así.

Para arreglarlo, usamos un descuento:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

$$\text{con } 0 \leq \gamma \leq 1$$

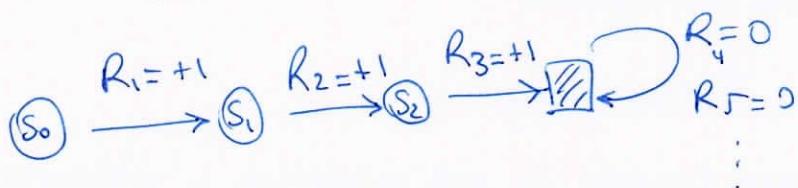
~~#~~ Cuanto  $\gamma$  esté más cerca de 1, se tiene más en cuenta el largo plazo. Un agente nubio será agresivo con  $\gamma=0$ .

Huified

Podemos definir  $G_t$  como:

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

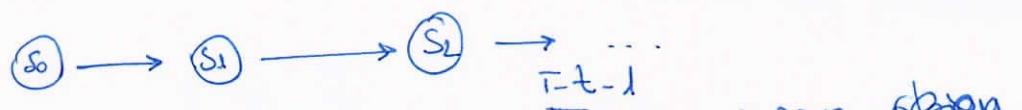
donde  $\gamma \neq 0$  pude ser  $\gamma = 1$  nle avue geda definiode porque  
admitimos un tipo de estados "absorbente":



... o bien podemos definir  $G_t$  como

$$G_t = \sum_{k=0}^{T-t-1} \gamma^k R_{t+k+1}$$

Donde  $T$  cabe la posiblidad de que sea  $\infty$ . y  $\gamma$  pude ser 1  
(pues no se dan las mitades situaciones a la vez).  
En este caso no se consideran los estados absorbentes!



Pero sabemos que si escribimos  $\sum_{k=0}^{T-t-1}$ , es porque ~~estas~~  
no se va a dar la situación de  $T \rightarrow \infty$  y  $\gamma = 1$ .

### Estados de Markov

Asumimos que el estados se prepara de alguna forma por el  
sistema.

Queremos obtener estados del ambiente (una señal de estado) que captúren de forma compacta el pasado, pero reteniendo toda la info. relevante.

En el caso más general  $\mathbb{P}$ :

$$\mathbb{P}[R_{t+1} = r, S_{t+1} = s' | S_0, A_0, S_1, A_1, \dots, S_t, R_t]$$

si es la señal de estado (state signal) tiene la propiedad de Markov:

$$\mathbb{P}[R_{t+1}, S_{t+1} | R_t, S_t, A_t]$$

Ejemplo: Pole-balancing state: la representación binaria del estado del dispositivo pede un cierto grado ~~fijo~~ beneficioso para el RL aspecto, ya que le permite ignorar las diferencias finas entre estados que no habrían sido útiles al resolver la tarea.

### 3.6. Proceso de Decisión de Markov

Una tarea de RL que satisface la prop. de Markov es un proceso de decisión de Markov, MDP.

Habíamos dicho que dado un estado del entorno  $S_t$  y ejecutada una acción por el agente,  $A_t$ , el entorno responde con un nuevo estado  $S_{t+1}$  y una recompensa  $R_{t+1}$ .

También decímos que en general:

$$\mathbb{P}(R_{t+1}, S_{t+1} | H) = \mathbb{P}(R_{t+1}, S_{t+1} | S_0, A_0, R_1, \dots, R_{t-1}, S_t, A_t)$$

Y que un proceso que satisface la prop. de Markov es tal que:

$$\underline{\mathbb{P}}(R_{t+1}, S_{t+1} | t) = \underline{\mathbb{P}}(R_{t+1}, S_{t+1} | S_t, A_t)$$

Por tanto, en un MDP tenemos:

$$p(r, s' | s, a) = \underline{\mathbb{P}}(R_{t+1} = r, S_{t+1} = s' | S_t = s, A_t = a)$$

Con esto, podemos calcular:

$$* r(s, a) = \mathbb{E}[R_{t+1} | S_t = s, A_t = a] = \sum_{r \in R} \sum_{s' \in S} p(s', r | s, a)$$

$$* P(s' | s, a) = \sum_{r \in R} p(s', r | s, a) \leftarrow \text{transition probabilities}$$

$$* r(s, a, s') = \sum_{r \in R} r \cdot p(s', r | s, a) / p(s' | s, a) \leftarrow \begin{matrix} \text{Expected reward for} \\ \text{state-action-state} \\ \text{triples} \end{matrix}$$

$$\mathbb{E}[R_{t+1} | S_t = s, A_t = a, S_{t+1} = s']$$

Ej.: Robot que recarga:

Estados:  $S = \{\text{low, high}\}$ .

A (low) = {search, wait, recharge}

A (high) = {search, wait}

Si hacemos search después de low  $\Rightarrow$  ~~pasamos a~~ seguimos en low con prob.  $\beta$  y descargamos batería con prob.  $1 - \beta$ .

Si los enemigos se buscan en high  $\left\{ \begin{array}{l} \alpha \rightarrow \text{high} \\ 1-\alpha \rightarrow \text{low} \end{array} \right.$

Franch: n° de letas recolectadas durante las búsquedas.

Lata: +1 ..

$$r = \mathbb{E}[R_{t+1} | S_t = s, A_t = a, S_{t+1} = s']$$

descarga : -3

$$\mathbb{P}(S_{t+1} = \text{low} | S_t = \text{low}, A_t = \text{search}) = \beta. \quad \text{fs}$$

$$\mathbb{P}(S_{t+1} = \text{high} \mid S_t = \text{high}, A_t = \text{search}) = \alpha. \quad \checkmark$$

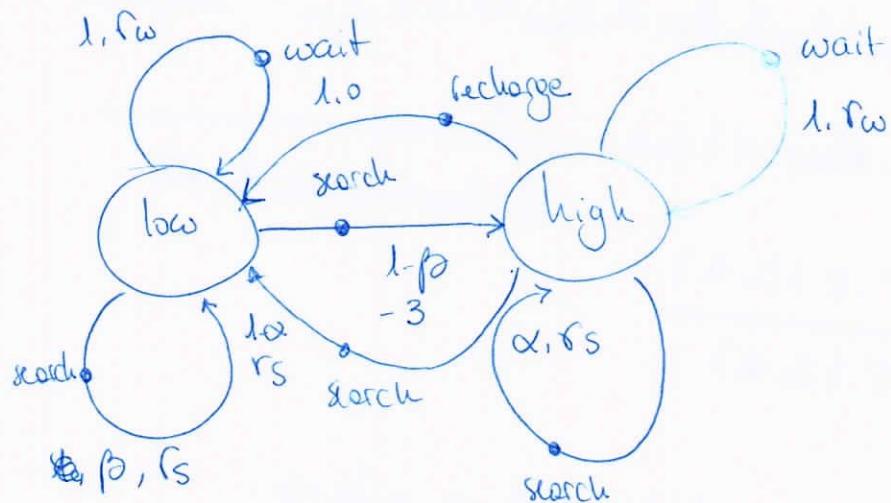
$$\mathbb{P}(S_{t+1} = \text{low} \mid S_t = \underset{\text{high}}{\cancel{\text{selected}}}, A_t = \text{search}) = 1 - \beta. \quad r = -3$$

$$\mathbb{P}(S_{t+1} = \text{high} \mid S_t = \text{low}, A_t = \text{search}) = 1 - \alpha.$$

$$\Pr(S_{t+1} = \text{low} \mid S_t = \text{low}, A_t = \text{wait}) = 1 \quad \text{fw}$$

$$\mathbb{P}(S_{t+1} = \text{high} | S_t = \text{high}, A_t = \text{wait}) = 1. \quad \text{sw}$$

$$\Pr(S_{t+1} = \text{high} \mid S_t = \text{low}, A_t = \text{recharge}) = 1.$$



## Value functions. (Función de Valor).

Hasta ahora hemos visto:

~~Probabilidad~~

$\mathbb{P}[S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a]$  determina totalmente un NDP.

A partir de aquí podemos calcular  $r(s, a) \leftarrow \begin{cases} r(s, a) \\ r(s, a; s') \\ p(s' | a, s) \end{cases}$ ,  $p(r | s, a, s')$

¿Cómo?

1)  $r(s, a) \equiv$  valor esperado de  $R_{t+1}$  sabiendo que partimos de  $S_t = s$  y  $A_t = a$ :

$$r(s, a) = \mathbb{E}[R_{t+1} | S_t = s, A_t = a] = \sum_{r \in R} r \cdot \sum_{s'} p(s' | s, a)$$

2)  $r(s, a, s') \equiv$  valor esperado de  $R_{t+1}$  tras el tripleto  $s \xrightarrow{a} s'$ :

$$\begin{aligned} r(s, a, s') &= \mathbb{E}[R_{t+1} | S_t = s, A_t = a, S_{t+1} = s'] = \sum_r r \cdot \mathbb{P}[R_{t+1} = r | S_t = s, A_t = a, S_{t+1} = s'] \\ &= \sum_r r \cdot \mathbb{P}(S_{t+1} = s' | R_{t+1} = r, S_t = s, A_t = a) \cdot \frac{1}{\mathbb{P}(S_{t+1} = s' | S_t = s, A_t = a)} \\ &= \sum_r r \cdot \frac{p(s', r | s, a)}{p(s' | s, a)} \end{aligned}$$

3)  $p(s' | a, s) \equiv$  Prob. de  $S_{t+1} = s'$  dados  $A_t = a$  y  $S_t = s$ :

$$p(s' | a, s) = \sum_{r \in R} \mathbb{P}(S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a)$$

Ahora veremos de forma más general, que los métodos de RL estiman siempre funciones de valor:

Una función de valor es una función de los estados (o del par estado-acción) que estima cuán de bueno es cada estado en términos de futuras recompensas a esperar (o más concretamente, en términos de  $E[G_t]$ ).

$G_t$  depende de las acciones que se ejecutén. A su vez, las acciones dependen de la política. Por tanto, denotamos los valores esperados con subíndice  $\pi$ .

De modo, una política  $\pi$  es una distr. de prob. condicionada  $\pi(a|s)$ .  
~~sobre que nos~~ Dado un estado, las de una distr. de prob. sobre las acciones  $a \in A(s)$ . Por otra parte, dado un subconjunto  $B \subseteq A$ , las de una función que predice cuánto cambia la probabilidad de un conjunto de  $B$  conforme nos cambiamos de estado.  
 El valor de un estado "s" bajo una política  $\pi$ ,  $V_\pi(s)$ , es el

retoño esperado si partimos de "s" y consideramos todos los escenarios posibles a los que podríamos desembocar siguiendo la política:

$$\begin{aligned} V_\pi(s) &= E_\pi[G_t | S_t = s] = E_\pi[R_{t+1} + \gamma R_{t+2} + \dots | S_t = s] = \\ &= E_\pi \left[ \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s \right] = \sum_{s'} p(s'|a,s) \sum_a \pi(a|s) \\ &\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} = \sum_{r,s'} p(s',r|a,s,a) \sum_a \pi(a|s) \underbrace{\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}}_{G_t} \end{aligned}$$

$$P(A, B) = P(A|B) \cdot P(B)$$

$$P(s', r | s, a) = P(r | s, s', a) \cdot p(s' | s, a)$$

$$\mathbb{E}(R_{t+1} | S_t = s, A_t = a, S_{t+1} = s') = \sum_r r \cdot p(r | s, s', a)$$

Igualmente,  $q_n(s, a)$  es el valor esperado en términos de  $n$  (en el futuro) si partimos del par  $(s, a)$ :

$$q_n(s, a) = \mathbb{E}_n [G_t | S_t = s, A_t = a] = \mathbb{E}_n \left[ \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | s, a \right] =$$

$$= \sum_{s'} p(s' | s, a) \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} = \sum_{r, s} p(r, s' | s, a) \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

$v_n(s)$ : Función Valor para la política  $\pi$ .

$q_n(s, a)$ : Función Acción-Valor para la política  $\pi$ .

Una propiedad fundamental de las funciones valor y acción-valor es que satisfacen relaciones recursivas:

i) Relación recursiva para  $V_n(s)$ :

$$V_n(s) = \mathbb{E}_n [G_t | S_t = s] = \mathbb{E}_n \left[ \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s \right] =$$

$$= \mathbb{E}_n \left[ R_{t+1} + \sum_{k=1}^{\infty} \gamma^k R_{t+k+1} | S_t = s \right] = \mathbb{E}_n \left[ R_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} | S_t = s \right] =$$

$$= \mathbb{E}_n [R_{t+1} | S_t = s] + \gamma \mathbb{E}_n \left[ \underbrace{\sum_{k=0}^{\infty} \gamma^k R_{t+k+2}}_{G_{t+1}} | S_t = s \right]$$

$$\text{Notese que } \mathbb{E}_n[G_{t+1} | S_t = s] = \sum_a \pi(a|s) \cdot \underbrace{\sum_{s'} p(s'|s,a) \mathbb{E}_n[G_{t+1}|s']}_{V_n(s')}$$

$$\text{Y que } \mathbb{E}_n[R_{t+1} | S_t = s] = \sum_a \pi(a|s) \cancel{p(s'|s,a)} r(s,a) =$$

$$= \sum_a \pi(a|s) \cdot \mathbb{E}_n(R | S, a) = \sum_a \pi(a|s) \cdot \sum_{r \in R} r \cdot p(r|s,a) =$$

$$= \sum_a \pi(a|s) \sum_{r \in R} r \cdot \sum_{s'} p(s',r|s,a)$$

Por tanto:

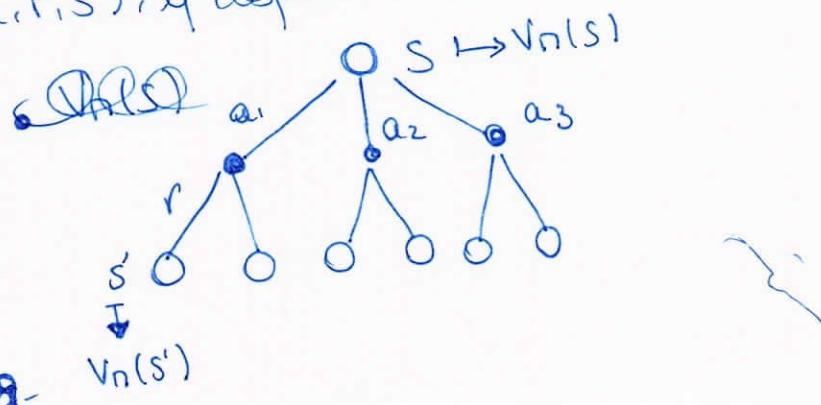
$$V_n(s) = \overbrace{\sum_a \pi(a|s) \sum_r r \sum_{s'} p(s',r|s,a)}^{(*)} + \gamma \sum_a \pi(a|s) \sum_{s'} p(s'|s,a) V_n(s')$$

$$= (*) + \gamma \sum_a \pi(a|s) \sum_{r,s'} p(s',r|s,a) V_n(s') =$$

$$= \sum_a \pi(a|s) \sum_{r,s'} p(s',r|s,a) [r + \gamma V_n(s')]$$

Donde queda implícito que:  $\begin{array}{c} a \in A(s) \\ s \in S \\ r \in R \end{array}$

Esta expresión suele ser fácilmente comprobable  
esperando! → estamos multiplicando [-] por la probabilidad de  
cada tripleta  $(a,r,s')$ , y después lo sumamos todo!



ecuación

A la relación anterior le llamamos relación de Bellman para  $V_n$ :

$$V_n(s) = \sum_a \pi(a|s) \sum_{s', r} p(s', r | s, a) [r + \gamma V_n(s')]$$

"El valor de un estado  $s$  de partida es el valor del estado que lo sucede (descuento por  $\gamma$ ) más ~~es~~ la recompensa recibida entre medias".

### 3.8. Funciones de Valor Óptimas.

Las funciones valor definen un orden parcial entre políticas: si  $V_n(s)$  es ~~mucho mayor~~ que  $V_{n'}(s)$ , entonces decimos que  $n \geq n'$ :

$$\pi \geq \pi' \iff V_n(s) \geq V_{n'}(s) \quad \forall s \in S.$$

Siempre hay una política mejor que igual que todas las demás: la óptima,  $\pi^*$ .

$$V^*(s) = \max_n V_n(s) \quad \forall s \in S.$$

~~Este~~ es

la función  $V^*$  de  ~~$V_n$~~  Todas las ~~as~~ políticas óptimas son ~~IGUALES~~ PARA TODOS LOS ESTADOS.

Lo mismo ocurre con  $q_n(s, a)$ :

$$q^*(s, a) = \max_n q_n(s, a) \quad \forall (s, a) \in S \times A(S)$$

$V^*$  es la función valor para una política óptima. Por tanto, satisface la ecuación

$$V^*(s) = \sum_a \pi(a|s) \sum_{s', r} p(s', r|s, a) [r + \gamma V^*(s')]$$

Sabemos que  $\pi^*$  esce la acción que maximiza el ~~síntesis~~ retorno dado esa acción:

$$V^*(s) = \max_{a \in A(s)} q^*(s, a) = \max_{a \in A(s)} \mathbb{E}_{\pi^*}[G_t | S_t=s, A_t=a] =$$

$$= \max_{a \in A(s)} \mathbb{E}_{\pi^*} \left[ \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t=s, A_t=a \right] = \max_{a \in A(s)} \mathbb{E}_{\pi^*} \left[ R_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} | S_t=s, A_t=a \right] =$$

$$= \max_{a \in A(s)} \mathbb{E}_{\pi^*} \left[ R_{t+1} \underbrace{\dots}_{| S_t=s, A_t=a} + \gamma \max_{a' \in A(s')} \mathbb{E}_{\pi^*} \left[ \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} | S_t=s, A_t=a \right] \right] =$$

$$\Leftrightarrow = \max_{a \in A(s)} \sum_{r, s'} r \cdot \mathbb{P}[s', r | s, a] + \gamma \max_{a' \in A(s')} \left[ \sum_{r, s'} \mathbb{P}(s', r | s, a) \sum_{a''} \pi(a'|s') \mathbb{E}_{\pi^*} [G_{t+1} | S_{t+1}=s', A_{t+1}=a''] \right]$$

Valores terminos o finales:

$$\bullet \quad \mathbb{E}_{\pi^*} [R_{t+1} | S_t=s, A_t=a] = \sum_{r, s'} \mathbb{P}[s', r | s, a]$$

$$\max_a \left( \mathbb{E}_{\pi^*} \left[ \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} | S_t=s, A_t=a \right] \right) \left( \mathbb{E}_{\pi^*} [G_{t+1} | S_t=s, A_t=a] \right) =$$

$$\max_a \left( \sum_{r, s'} \mathbb{P}[s', r | s, a] \sum_{a'} \mathbb{E}_{\pi^*} [G_{t+1} | S_{t+1}=s', A_{t+1}=a'] \right)$$

Con lo que queda:

$$V_*(s) = \max_{a \in A(s)} \sum_{s' \in S} p(s'|s,a) [r + \gamma \max_{a' \in A(s')} q_*(s',a')]$$

$V_*(s')$

~~q<sub>n</sub>(s',a')~~

~~q<sub>n</sub>(s',a')~~

~~q<sub>n</sub>(s',a')~~

2) Ecación recursiva para  $q$ :

$$\begin{aligned} q_n(s,a) &= \mathbb{E}_n[G_t | S_t=s, A_t=a] = \mathbb{E}_n\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t=s, A_t=a\right] = \\ &= \mathbb{E}_n[R_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} | S_t=s, A_t=a] = \mathbb{E}_n[R_{t+1} | S_t=s, A_t=a] + \\ &+ \gamma \mathbb{E}_n[G_{t+1} | S_t=s, A_t=a] = \sum_{s',r} r \cdot p(s',r|s,a) + \gamma \sum_{s',a'} p(s',a'|s,a) \cdot \\ &\cdot \sum_{a'} \pi(a'|s') \mathbb{E}_n[G_{t+1} | S_{t+1}=s', A_{t+1}=a'] = \sum_{s',r} p(s',r|s,a) [r + \gamma \sum_{a'} \pi(a'|s') q_n(s',a')] \end{aligned}$$

Es decir:

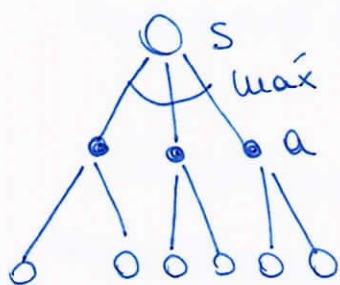
$$q_n(s,a) = \sum_{s',r} p(s',r|s,a) [r + \gamma \sum_{a'} \pi(a'|s') q_n(s',a')]$$

y la ecación óptima para  $q_*$ :

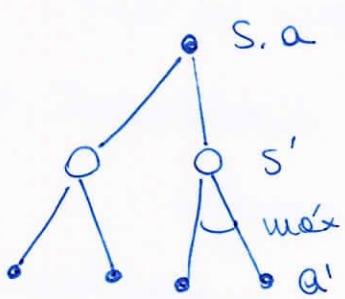
"El valor de un estado y una acción bajo una política óptima debe reflejar esperar la recompensa esperada de esa acción y ese estado más el valor ~~esperado~~ del ~~para~~ acción - ~~a~~ estado  $(s',a')$  al que llegamos ni excepcion la mejor acción siguiente,  $a'$ :"

$$q_*(s,a) = \sum_{s',r} p(s',r|s,a) [r + \gamma \max_{a' \in A(s')} q_*(s',a')]$$

Diagramas:



Para  $V^*(s)$



para  $q^*(s,a)$

Tcs. de bellman óptimas para el robot:

$$s = \{l, h\}$$

$$\omega = \{\text{search, wait, recharge}\} = \{s, w, r\}$$

$$V^*(s) = \max \left\{ \begin{array}{l} p(h|h,s) \cdot [r(h,s,h) + \gamma \cdot V^*(h)] + p(l|h,s) \dots , \\ p(h|h,w) [ \dots ] + p(l|h,w) [ \dots ] \end{array} \right.$$

Idem para  $V^*(l)$

Ej. 3.12: Resolviendo el mundo cuadrado

Estado en A:  $r(A) = +10, S' = A'$

" " B:  $r(B) = +5, S' = B'$

$$A = \{N, E, W, S\}$$

off the grid: -1.

Resol de acciones:  $\emptyset$

¿Cómo encontrar  $\nabla(s)$ ?

Si resolvemos directamente, estaremos usando 3 asunciones

que para ver se cumplen:

- 1) Conocemos bien las dinámicas del ambiente
- 2) Podemos calcular la solución (poder computacional alto).
- 3) Prop. de Markov.

### Ejercicios Capítulo 3.

Ej. 3.2.:

## Ejercicios del Capítulo 3.

Ej. 3.2 Is the MDP framework adequate to represent all goal-directed learning tasks?

Un MDP es apropiado cuando se verifica la propiedad de Markov ( $\omega$  se puede aproximar como un proceso de Markov):

$$p = f(s', r, s, a).$$

(Algunas ~~importantes~~ ~~relevantes~~ ~~comúnmente~~ ~~homólogas~~ derivadas de procesos psicológicos complejos no se puede representar de este modo?)

Ej. 3.3 Consider the problem of driving (...)

La lnea agente - ~~en~~ ambiente distingue los elementos sobre los que el agente tiene control, los conocimientos del entorno.

El agente conductor tiene control total (con excepción de desórdenes psicomotorias) sobre los pedales que presiona, el volante que dirige y el freno que acciona.

Ej. 3.4 Table del ejemplo 3.3 pero para  $p(s', r | s, a)$ ?

s	a	s'	r	$p(s', r   s, a)$
h	s	h*	Search	$\alpha$
h	s	l	$r_s$	$1-\alpha$
l	s	h	-3	$1-\beta$
l	s	l	$r_s$	$\beta$
h	w	h	$r_w$	1
l	w	l	$r_w$	1
l	r	h	0	1

Ej. 3.5. Modificar la ec. 3.3 para el caso epizódico.

Tu caso. no sólo interemos trabajando con S, sino con S<sup>+</sup> (el conjunto de estados, incluidos los terminales):

$$\sum_{s' \in S^+} \sum_{r \in R} p(s', r | s, a) = 1 \quad \forall s \in S, \forall a \in A$$

Ej. 3.6. Pole-balancing extendido como episódico ( $T$ ) pero con desenfado:

$$G_t = \sum_{k=0}^T \gamma^k R_{t+k+1}$$

En este caso, si el error se produce en tiempo  $k$ , (ándremos):

$$\text{Ge} \approx \gamma^{\underline{K}} \quad \text{Ge} \quad G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots + \gamma^{K-1} R_{t+K} = \gamma^{K-1} \cdot (-1) =$$

$$\underline{t} + \underline{Q} + \underline{Q} - \underline{Q}$$

$$= -\gamma^{k-1} \quad \text{A mayor } \underline{k}, \text{ mayor } \underline{b_k}.$$

Aquí finalizará el episodio y únicamente se recibirá un  
correo dependiente de cuándo se produce esa caída.  
En el caso descrito en el libro, tendríamos una zona <sup>dpendiente</sup> de toda  
las veces que caemos lo que estamos penalizando la

3.7 En este caso, no se está guardando información de cuánto  
está teniendo el agente en salir del laberinto, con lo que  
para el robot ~~se~~ de igual tener muchos ~~tener~~ muchos  
o poco.

Ej. 3.8.  $\gamma = 0.5$ ,  $R_1 = -1$ ,  $R_2 = 2$ ,  $R_3 = 6$ ,  $R_4 = 3$  y  $R_5 = 2$ , con  $T=5$

¿  $G_0, \dots, G_5$  ?

$$\begin{aligned} & \cancel{\text{G}_0 + \cancel{\text{R}_0}} \\ & \cancel{\gamma R_0 + \gamma G_0} = 0 \\ & G_5 = G_0 + \gamma R_0 \quad \infty \end{aligned}$$

$$\cancel{\text{G}_0} \\ G_4 = \cancel{\gamma G_3} + R_5 = 2$$

$$G_3 = R_4 + \gamma G_4 = 3 + 0.5 \cdot 2 = 4$$

$$G_2 = R_3 + \gamma G_3 = 6 + 0.5 \cdot 4 = 8$$

$$G_1 = R_2 + \gamma G_2 = 2 + 0.5 \cdot 8 = 6$$

$$G_0 = R_1 + \gamma G_1 = -1 + 0.5 \cdot 6 = 2$$

Ej. 3.9.  $\gamma = 0.9$ ,  $R_1 = 2$ ,  $R_{>1} = 7$ .  $G_1$ ?  $G_0$ ?

$$G_0 = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} = R_1 + \gamma \cdot 7 + \sum_{k=0}^{\infty} \gamma^k \cancel{R_{t+k+1}} = 2 + \gamma \cdot 7 + \underbrace{\sum_{k=0}^{\infty} \gamma^k}_{\gamma < 1} =$$

$$= 2 + \frac{7 \cdot \gamma}{1 - \gamma}$$

$$G_1 = \frac{G_0 - R_1}{\gamma} = \frac{G_0 - 2}{\gamma} = \frac{7}{1 - \gamma} = 70.$$

Ej. 3.11.  $\pi$  es localística.  $E[R_{t+1}]$ ?

$$E[R_{t+1}] = \sum_r p(r|s) \cdot r = \sum_a \pi(a|s) \cdot \sum_{s'} p(s', r|s, a) \cdot r$$

Ej. 3.12. Equation for  $v_n$  in terms of  $q_n$  and  $\pi$ ?

Hay 2:

- $v_n(s) = \sum_a \pi(a|s) q_n(s, a)$  (entornos del presente)

$$\begin{aligned} v_n(s) &= \sum_a \pi(a|s) \sum_{s', r} p(s', r|s, a) [r + \gamma v_n(s')] = \\ &= \sum_a \pi(a|s) \sum_{s', r} p(s', r|s, a) [r + \sum_{a'} \pi(a'|s') q(s'|a')] \end{aligned}$$

Ej. 3.13.  $q_n$  entornos de  $v_n$  y  $\pi$ ?

$$q_n(s, a) = \frac{v_n(s)}{\sum_a \pi(a|s)}$$

bien:

$$\begin{aligned} q_n(s, a) &= E_n[G_t | S_t = s, A_t = a] = E_n[R_{t+1} + \gamma \sum_0^\infty \gamma^k R_{t+k+2} | s, a] = \\ &= E_n[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a] = \\ &= E_n[R_{t+1} | S_t = s, A_t = a] + \gamma E_n[G_{t+1} | S_t = s, A_t = a] = \\ &= \sum_{r, s'} p(s', r | s, a) \underbrace{\pi}_{r, s'} + \gamma \sum_{r, s'} \sum_a \pi(a | s') q_n(s', a') \end{aligned}$$

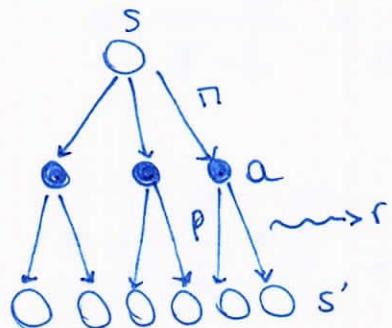
Y sustituyendo  $v_n(s) = \sum_a \pi(a|s) q_n(s, a)$ :

$$q_n(s, a) = \sum_{r, s'} p(r, s'|s, a) [r + \gamma v_n(s')]$$

Eg. 3.14 Bellman eq. must hold for each state for the value function  $V_n$  shown in fig. 3.2 of example 3.5...

Eg. Bellman:

$$V_n(s) = \sum_a \pi(a|s) \sum_{s'} p(s', r|s, a) [r + \gamma V_n(s')]$$



$$\gamma = 0.9$$

$$s = (0, 0); V_n(s) = 0.7 \quad \downarrow r = 0; a = e; p(a|s) = 0.25$$

$$s' = (1, 0); V_n(s') = 0.4$$

$$\pi(u|s) = \pi(s|s) = \pi(w|s) = \pi(e|s) = 0.25.$$

$\pi$ : deterministic

~~$$0.7 = \pi(e|s) [r + \gamma V_n(s')] = 0.25 \cdot [0 + 0.9 \cdot 0.4]$$~~

~~$$= 0.25 + 0.36 = 0.61 \pm 0.1.$$~~

$$0.7 = \pi(e|s) [0 + 0.9 \cdot 0.4] + \pi(u|s) [0 + 0.9 \cdot 2.3] +$$

$$0.7 = \pi(e|s) [0 + 0.9 \cdot 0.7] + \pi(s|s) [0 + 0.9 \cdot (-0.4)] + \pi(w|s) [0 + 0.9 \cdot 0.7] =$$

$$0.36 + 2.07 + 0.63 - 0.36 = \frac{1}{4} 2.7 = 0.675$$

Ej. 3.15. Rewards are positive for goals, negative for running into edge of the world, and zero the rest of the time.

Are the signs of these rewards important, or only the intervals between them?

$$G_t = \sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + c) = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + \frac{c}{1-\gamma} \Rightarrow$$

$$\Rightarrow V_n = \mathbb{E}_n [G_t | S_t = s] = \mathbb{E} \left[ \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \right] + \frac{c}{1-\gamma} = \\ = \sum_a \pi(a|s) \sum_{s', r} p(s', r | s, a) [r + \gamma V_n(s')] + V_c$$

$$\text{Donde } V_c = \frac{c}{1-\gamma}$$

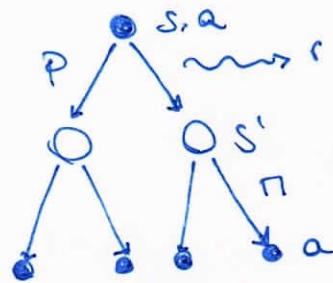
Ej. 3.16. Ahora considera añadir una constante  $c$  en una tasa episódica.

$$G_t = \sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + c)$$

En el caso episódico, una vez alcanza  $T$ , bajaría  $c$   
 $R_{t+T} = 0$ , y el  $1^{\text{er}}$  mundo desaparece, mientras que  
 $c$  multiplicaría a  $\sum_{k=0}^T \gamma^k$ , con  $T$  diferente al episodio

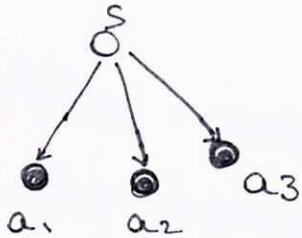
Por lo tanto, tendrán más valor aquellos estados en los que se espere es permanecer más tiempo en el tablero.

Ej. 3.17. ¿Ec. Bellman para  $q_n(s, a)$ ?



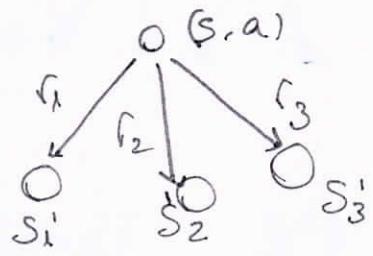
$$\begin{aligned}
 q_n(s, a) &= \mathbb{E}_n[G_t | S_t = s, A_t = a] = \mathbb{E}_n \left[ \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s, A_t = a \right] \\
 &= \mathbb{E}_n [R_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} | S_t = s, A_t = a] = \\
 &= \mathbb{E}_n [R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a] = \\
 &= \gamma \sum_{r, s'} p(s', r | s, a) r + \gamma \mathbb{E}_n [G_{t+1} | S_t = s, A_t = a] = \\
 &= \sum_{r, s'} p(s', r | s, a) \left[ r + \gamma \sum_{a'} \pi(a' | s') \mathbb{E}_n [G_{t+1} | S_{t+1} = s', A_{t+1} = a'] \right] \\
 &= \sum_{r, s'} p(s', r | s, a) \left[ r + \gamma \sum_{a'} \pi(a' | s') q_n(s', a') \right]
 \end{aligned}$$

Ej. 3.18



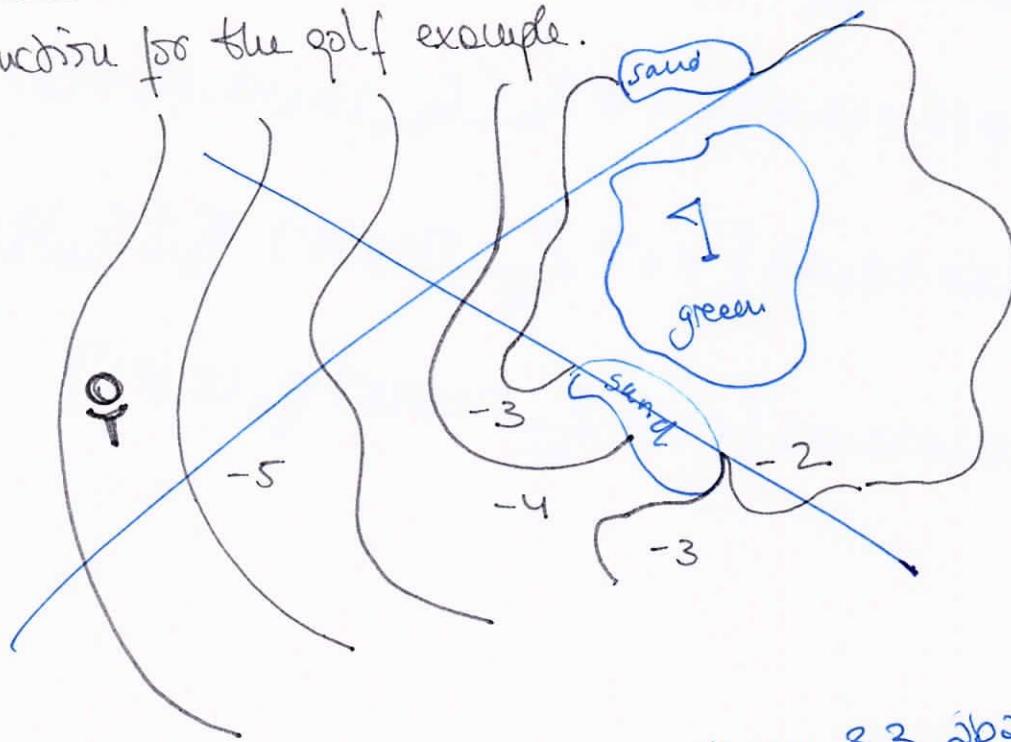
$$V_n(s) = \sum_a \pi(a | s) q(s, a).$$

Ej. 3.19.



$$q_n(s, a) = \sum_{r, s'} p(s', r | s, a) \cancel{[r + \gamma v_n(s')]} \quad [r + \gamma v_n(s')]$$

Ej. 3.20. Draw or describe the optimal state-value ( $v_n(s)$ ) function for the golf example.



¿Optimo? → El mínimo de figura 3.3 abajo: lo óptimo es empujar con driver, seguir con driver y después poner.

3.21

Si primero usemos poner, nos quedamos en la región de -6.  
Después usariamos diver + diver + poner. (-6 + (-3) + (-2) + (-1))

3.22.

$$V_{n_L}(S_0) = \mathbb{E}_{\pi_e} [G_t | S_t = S] = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} = \sum_a \pi(a|S) \sum_{S'_L} p(S'_L | S, a)$$

$$\cdot [r + \gamma V_n(S'_L)] = 1.0 [1 + V_n(S'_L)] = 1 + V_n(S'_L)$$

$$V_n(S'_L) = \gamma V_n(S_0)$$

$$\Rightarrow V_n(S_0) = 1 + V_n(S'_L) \Rightarrow V_n(S_0) = 1 + \gamma^2 V_n(S_0)$$

$$\gamma^2 x - x + 1 = 0 \Rightarrow x = \frac{1 \pm \sqrt{1 - 4}}{2}$$

$$\text{Solución} \quad (\gamma^2 - 1)V_n(S_0) + 1 = 0.$$

$$V_n(S_0) = \frac{-1}{\gamma^2 - 1} = \frac{1}{1 - \gamma^2} = \frac{1}{1 - 0.9^2} = 5.26$$

$$\left. \begin{array}{l} V_{n_L}(S_0) = \gamma V_{n_L}(S'_R) \\ V_{n_L}(S'_R) = 2 + \gamma V_{n_L}(S_0) \end{array} \right\} \begin{aligned} V_{n_L}(S_0) &= \gamma(2 + \gamma V_{n_L}(S_0)) = \\ &= 2\gamma + \gamma^2 V_{n_L}(S_0) \end{aligned}$$

$$\Rightarrow (\gamma^2 - 1)V_{n_L}(S_0) + 2\gamma = 0.$$

$$V_{n_L}(S_0) = \frac{-2\gamma}{\gamma^2 - 1} \Rightarrow 48 \quad \frac{2 \cdot 0.9}{0.19} = 9.47$$

Por lo tanto el resultado es 9.47.

Con estos pasos, hemos calculado  $V_n(s_0)$  para  $\Pi_L$  y  $\Pi_R$ .

Habrá que repetir los pasos con  $\Pi_S$  y  $\Pi_R$ . Dado que  $\Pi \cdot \Pi > \Pi' \iff V_n > V_{n'}$   $\forall s$  y  $V_{n_R}(s_0) > V_{n_L}(s_0)$  para el estado  $s_0$ , entonces también se cumplirá para  $S_L$  y  $S_R$ .

Por tanto,  $\Pi_R > \Pi_L$  y  $\Pi_R = \Pi^*$  si  $\gamma = 0.9$ .

Ej. 3.23: Ec. Bellman para  $q^*$  en el caso del robot.

Ec. Bellman:

$$\text{Diagrama: } \begin{array}{c} s, a \\ \swarrow \quad \searrow \\ s' \quad r \end{array} \quad q_n(s, a) = \sum_{r, s'} p(s', r | s, a) [r + \gamma \max_{a'} q_n(s', a')]$$

$$q_n(s', a')$$

la ec. óptima nos basará alternativamente entre todas las acciones, más que seleccionará  $\arg \max_{a'} q_n(s', a')$

~~$$q^*(s, a) = \sum_{r, s'} p(s', r | s, a) [r + \gamma \max_{a'} q^*(s', a)]$$~~

En el caso del robot

$$q^*(h, s) = \alpha [r_h + \gamma \max_{a'} q^*(h, a')] + (1-\alpha) [r_s +$$

~~$$+ \gamma \max_{a'} q^*(l, a')]$$~~

$$q^*(l, s) = (1-\beta) [-3 + \gamma \max_{a'} q^*(h, a')] + \beta [r_s +$$

$$+ \gamma \max_{a'} q^*(l, a')]$$

... 4 lo mismo para  $\begin{cases} (h, \omega) \\ (l, \omega) \\ (l, rch) \end{cases}$

Ej. 3.24. Usar los conocimientos de  $\pi^*$  y de  $V_{\text{MSP}}$  (ej. 3.8) para expresar este valor simbólicamente y después aproxímalos a 3 decimales.

Ec. óptima Bellman:

$$V_*(s) = \max_a \sum_{r, s'} p(s'|s, a) [r + \gamma V_*(s')]$$

Ec. 3.8:

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

$$V_*(s) = \max_a [0.25 \cdot \gamma V_*(\rightarrow) + 0.25 \cdot \gamma V_*(\downarrow) + 0.25 \cdot \gamma V_*(\leftarrow) + 0.25 \cdot (-1 + \gamma V_*(\uparrow))]$$

~~$0.25 \cdot 0.9 \cdot 22.0 = 4.5$~~

~~$0.25 \cdot (-1 + 0.9 \cdot 24.4) = 20.96 ?$~~

El estado de mayor  $V$  es un estado especial, de modo que lleguemos a él, nos lleva a  $A'$ . Siguiendo  $\pi^*$ , que sea lo que lleguemos en él, nos lleva a  $A'$ . Siguiendo  $\pi^*$ , que sea en  $A'$  lo mejor que podemos hacer es subir hasta  $A$ , con lo cual entraremos en un bucle donde recibimos  $R \neq 0$  cuando  $k$  es múltiplo de 5:

$$\begin{aligned} \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} &= 10 \cdot \gamma^0 + 0 \cdot \gamma^1 + 0 \cdot \gamma^2 + \dots + 10 \gamma^5 + 0 \cdot \gamma^6 + \dots = \\ &= 10 \cdot \gamma^0 + 10 \gamma^5 + 10 \gamma^{10} + 10 \gamma^{15} = \\ &= 10 \sum_{k=0}^{\infty} \gamma^{5k} = \frac{10}{1 - \gamma^5} \approx 24.419. \end{aligned}$$

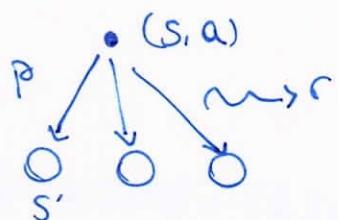
Ej. 3.25 Give an equation for  $V^*$  in terms of  $q^*$

$$V_n(s) = \sum_a \pi(a|s) q_n(s, a) \Rightarrow \arg \max_a q^*(s, a)$$

Ej. 3.26 Give an eq. for  $q^*$  in terms of  $V^*$  and the last 4-argument p.

$$q^*(s, a) = \sum_{r, s'} p(s', r|s, a) [r + \gamma V^*(s')]$$

Ej. 3.27: Give an eq. for  $\pi^*$  in terms of  $q^*$ .



$$\pi^*(a|s) = \arg \max_a q^*(s, a)$$

Ej. 3.28: Eq. for  $\pi^*$  in terms of  $V^*$ :

$$\pi^*(a|s) = \arg \max_a \sum_{r, s'} p(s', r|s, a) [r + \gamma V^*(s')]$$

Ej. 3.29. Reescribir  $V_n, V^*, q_n$  y  $q^*$  en términos de  $p(s'|s, a)$  y  $r(s, a)$

$$p(s'|s, a) = \sum_r p(r|s', r|s, a)$$

$$r(s, a) = \mathbb{E}_n[R_{t+1}|S_t=s, A_t=a] = \sum_{r, s'} r p(s', r|s, a)$$

$$V_n(s) = \sum_a \pi(a|s) \sum_{r, s'} p(r, s'|s, a) [r + \gamma V_n(s')] =$$

$$= \sum_a \pi(a|s) r(s, a) + \sum_a \pi(a|s) \sum_{s'} p(s'|s, a) V_n(s')$$

o bien

$$V_n(s) \rightarrow \sum_a \pi(a|s) \sum_{s'} p(s'|s, a)$$

(Ej. 3.29)

... Por tanto:

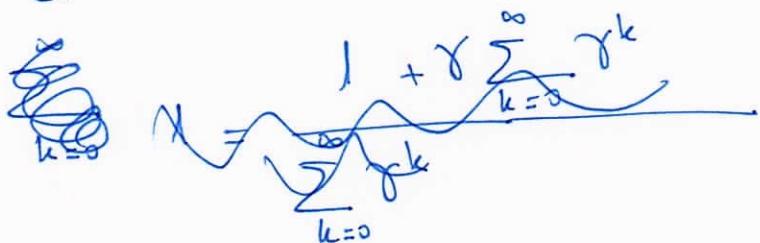
$$\begin{aligned} V_n(s) &= \sum_a \pi(a|s) \sum_{r, s'} p(s', r|s, a) [r + \gamma V_n(s')] = \\ &= \sum_a \pi(a|s) r(s, a) + \gamma \sum_a \pi(a|s) \sum_{s'} p(s'|s, a) V_n(s') \end{aligned}$$

y similar para las otras ecs.

Ej. 3.10:

Demoststrar  $\sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$  cuando  $|\gamma| < 1$

$$\sum_{k=0}^{\infty} \gamma^k = 1 + \gamma \sum_{k=0}^{\infty} \gamma^k$$



④

$$\sum_{k=0}^{\infty} \gamma^k - \gamma \sum_{k=0}^{\infty} \gamma^k = 1$$

Como  $\gamma \neq 1$ :

$$\sum_{k=0}^{\infty} \gamma^k (\lambda - \gamma) = 1$$

$$\sum_{k=0}^{\infty} \gamma^k = \frac{1}{\lambda - \gamma}$$