Iteración de Valor

Dado s_0

a_0

$s_1 \sim P_{s_0 a_0}$

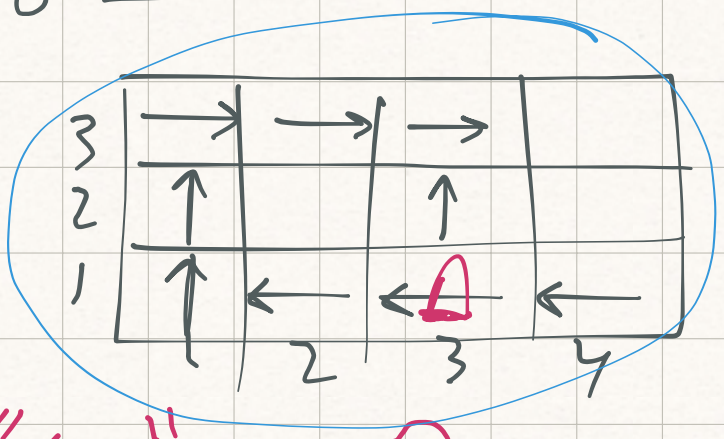
a_1

$s_2 \sim P_{s_1 a_1}$

$$R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \dots\dots$$

$\pi : S \longrightarrow A$

Política Inicial



$\pi((3,1)) = "\leftarrow"$

$a_0 = \pi(s_0)$

$P_{s_{\pi(s)}}(s')$

$P_{sa}(s')$

En la búsqueda de la mejor política

- V^{π}, V^{*}, π^{*}

la fn. de Valor para política π.

$$V^{\pi} : S \to R$$

$V^{\pi}(s)$ es la ganancia total de recompensas esperada, empezando en el estado s y ejecutando π dado.

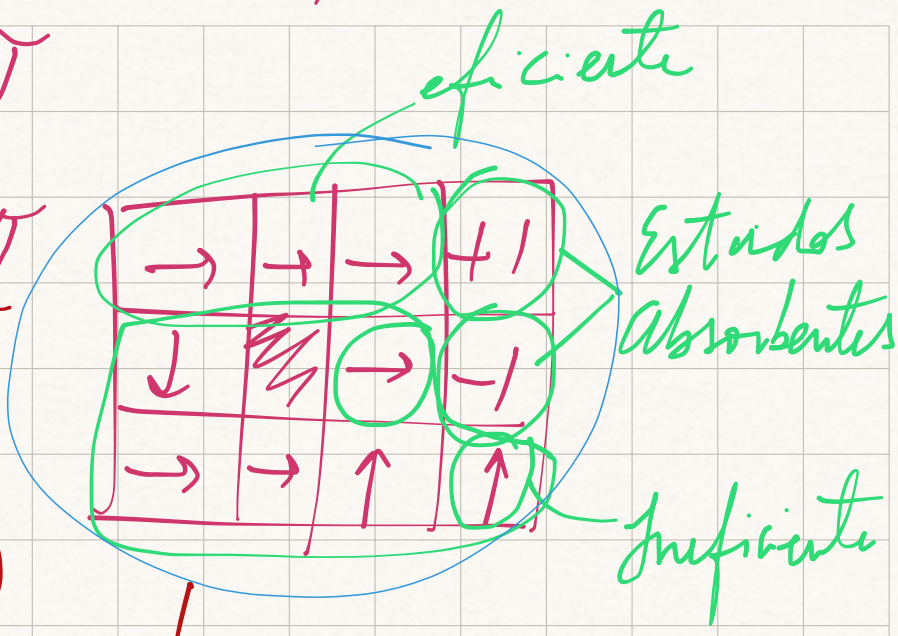
$$V^{\pi}(s) = \underline{\underline{E}}\left[R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \dots \mid \pi, s_0 = s\right]$$

→ Función de Valor para la

política π

Ejemplo

π =



$P = 0.8\ (0.1, 0.1)$

$\gamma = 0.99$

V^{π}

0.52	0.73	0.77	+ 1
-0.9		-0.82	- 1
-0.88	-0.87	-0.85	- 1

Ecuación de Bellman, que gobierna la función de Valor.

$$V^{\pi}(s) = R(s_0) + \gamma \sum_{s'} P_{s\pi(s)}(s')\, V^{\pi}(s')$$

(Sistema de Ec. lineales 11 incog y 11 ecuaciones

$$V^{\pi}(s) = E\left[R(s_0) + \gamma\left(R(s_1) + R(s_2)\right)\right]$$

$\pi, s_0 : s$

recompensa inmediata

$V^{\pi}(s_1)$

$$V^{\pi}(s) = E\left[R(s_0) + \gamma\, V^{\pi}(s_1)\right]$$

$$V^{\pi}(s) = R(s) + \gamma \sum_{s'} P_{s\pi(s)}(s')\, V^{\pi}(s')$$

$s' \sim P_{s\pi(s)}$

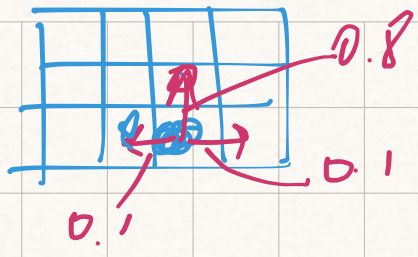
$s \longrightarrow a = \pi(s)$

$s' \sim P_{sa}, \quad a = \pi(s)$

Sist. Ec. lineal

$\pi(3,1) = \uparrow$

tomamos $V^{\pi}(5,1)$



$$V^{\pi}(3,1) = R(3,1) + \gamma(0.8\,V^{\pi}(3,2) + 0.1\,V^{\pi}(2,1) + 0.1\,V^{\pi}(4,1))$$

Cada $\boxed{}$ → estado.

V^{π} $\boxed{r}$ → incógnita (1 linea)

$$\underbrace{\begin{bmatrix} V^{\pi}(1,1) \\ V^{\pi}(1,2) \\ \vdots \\ V^{\pi}(4,3) \end{bmatrix}}_{V^{*}} \in \mathbb{R}^{11}$$

$$V^*(s) = \max_{\pi} V^{\pi}(s)$$

Optimal Value $\begin{cases} V^{\pi} - \text{policy} \\ V^* - \underline{\text{optimal}} \end{cases}$