Programación Dinámica

Fernando Lozano

Universidad de los Andes

14 de febrero de 2023

Programación Dinámica

- Bellman (1957): Control, economía, algorítmica. . . .

Programación Dinámica

- Bellman (1957): Control, economía, algorítmica. . . .
- MDPs finitos: $|\mathcal{S}|, |\mathcal{A}|, |\mathcal{R}| < \infty$.

Programación Dinámica

- Bellman (1957): Control, economía, algorítmica. . . .
- MDPs finitos: $|\mathcal{S}|, |\mathcal{A}|, |\mathcal{R}| < \infty$.
- R_t, S_t variables aleatorias discretas cuya distribución depende únicamente del estado y acción anterior:

Programación Dinámica

- Bellman (1957): Control, economía, algorítmica. . . .
- MDPs finitos: $|\mathcal{S}|, |\mathcal{A}|, |\mathcal{R}| < \infty$.
- R_t, S_t variables aleatorias discretas cuya distribución depende únicamente del estado y acción anterior:

$$p(s', r \mid s, a) \doteq \mathbf{P}\left\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\right\}, \quad \forall s', s \in \mathcal{S}, r \in \mathcal{R}, a \in \mathcal{A}(s)$$

Programación Dinámica

- Bellman (1957): Control, economía, algorítmica. . . .
- MDPs finitos: $|\mathcal{S}|, |\mathcal{A}|, |\mathcal{R}| < \infty$.
- R_t, S_t variables aleatorias discretas cuya distribución depende únicamente del estado y acción anterior:

$$p(s', r \mid s, a) \doteq \mathbf{P}\left\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\right\}, \quad \forall s', s \in \mathcal{S}, r \in \mathcal{R}, a \in \mathcal{A}(s)$$

- $p(s', r \mid s, a)$ es conocida $\forall s', s \in \mathcal{S}, r \in \mathcal{R}, a \in \mathcal{A}(s)$

Programación Dinámica

- Bellman (1957): Control, economía, algorítmica. . . .
- MDPs finitos: $|\mathcal{S}|, |\mathcal{A}|, |\mathcal{R}| < \infty$.
- R_t, S_t variables aleatorias discretas cuya distribución depende únicamente del estado y acción anterior:

$$p(s', r \mid s, a) \doteq \mathbf{P}\left\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\right\}, \quad \forall s', s \in \mathcal{S}, r \in \mathcal{R}, a \in \mathcal{A}(s)$$

- $p(s', r \mid s, a)$ es conocida $\forall s', s \in \mathcal{S}, r \in \mathcal{R}, a \in \mathcal{A}(s)$
- (no hay aprendizaje!)

- Retorno:

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

- Retorno:

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

- Retorno:

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

- Política:

$$\pi(a \mid s)$$

- Retorno:

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

- Política:

$$\pi(a \mid s) = \mathbf{P}\left\{A_t = a \mid s_t = s\right\}$$

- Retorno:

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

- Política:

$$\pi(a \mid s) = \mathbf{P}\left\{A_t = a \mid s_t = s\right\}$$

- Función de valor de estado de la política π:

$$v_\pi(s) \doteq \mathbb{E}_\pi\left\{G_t \mid S_t = s\right\}$$

- Retorno:

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

- Política:

$$\pi(a \mid s) = \mathbf{P}\left\{A_t = a \mid s_t = s\right\}$$

- Función de valor de estado de la política π:

$$v_\pi(s) \doteq \mathbb{E}_\pi\left\{G_t \mid S_t = s\right\} = \mathbb{E}_\pi\left\{\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right\}$$

- Retorno:

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

- Política:

$$\pi(a \mid s) = \mathbf{P}\left\{A_t = a \mid s_t = s\right\}$$

- Función de valor de estado de la política π:

$$v_\pi(s) \doteq \mathbb{E}_\pi\left\{G_t \mid S_t = s\right\} = \mathbb{E}_\pi\left\{\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right\}$$

- Función de valor de acción de la política π:

- Retorno:

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

- Política:

$$\pi(a \mid s) = \mathbf{P}\left\{A_t = a \mid s_t = s\right\}$$

- Función de valor de estado de la política π:

$$v_\pi(s) \doteq \mathbb{E}_\pi\left\{G_t \mid S_t = s\right\} = \mathbb{E}_\pi\left\{\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right\}$$

- Función de valor de acción de la política π:

$$q_\pi(s, a) = \mathbb{E}_\pi\left\{G_t \mid S_t = s, A_t = a\right\}$$

- Retorno:

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

- Política:

$$\pi(a \mid s) = \mathbf{P}\left\{A_t = a \mid s_t = s\right\}$$

- Función de valor de estado de la política π:

$$v_\pi(s) \doteq \mathbb{E}_\pi\left\{G_t \mid S_t = s\right\} = \mathbb{E}_\pi\left\{\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right\}$$

- Función de valor de acción de la política π:

$$q_\pi(s, a) = \mathbb{E}_\pi\left\{G_t \mid S_t = s, A_t = a\right\}$$

- Retorno:

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

- Política:

$$\pi(a \mid s) = \mathbf{P}\left\{A_t = a \mid s_t = s\right\}$$

- Función de valor de estado de la política π:

$$v_\pi(s) \doteq \mathbb{E}_\pi\left\{G_t \mid S_t = s\right\} = \mathbb{E}_\pi\left\{\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right\}$$

- Función de valor de acción de la política π:

$$\begin{aligned} q_\pi(s,a) = \mathbb{E}_\pi\left\{G_t \mid S_t = s, A_t = a\right\} \\ = \mathbb{E}_\pi\left\{\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a\right\} \end{aligned}$$

Ecuaciones de Bellman

- Para v_π:

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \left[r(s, a, s') + \gamma v_\pi(s') \right]$$

Ecuaciones de Bellman

- Para v_π:

$$
\begin{aligned}
v_\pi(s) &= \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \left[r(s, a, s') + \gamma v_\pi(s') \right] \\
&= \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi(s') \right]
\end{aligned}
$$

Ecuaciones de Bellman

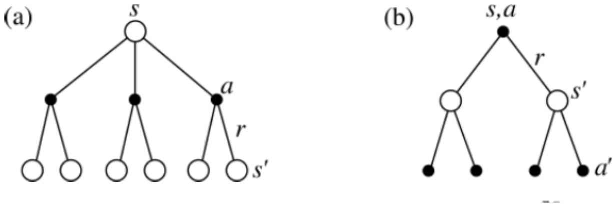
- Para v_π:

$$
\begin{aligned}
v_\pi(s) &= \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \left[r(s, a, s') + \gamma v_\pi(s') \right] \\
&= \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi(s') \right]
\end{aligned}
$$

- Para $q_\pi(s, a)$:

$$
q_\pi(s, a) = \sum_{s'} p(s' \mid s, a) \left[r(s, a, s') + \gamma \sum_{a'} q_\pi(s', a') \right]
$$

Diagramas de Backup



Funcion de valor óptima

Funcion de valor óptima

- $\pi \geq \pi' \Leftrightarrow v_\pi(s) \geq v_{\pi'}(s)$ para todo s.

Funcion de valor óptima

- $\pi \geq \pi' \Leftrightarrow v_\pi(s) \geq v_{\pi'}(s)$ para todo s.
- Política óptima: π_* tal que $\pi_* \geq \pi$ para cualquier π.

Funcion de valor óptima

- $\pi \geq \pi' \Leftrightarrow v_\pi(s) \geq v_{\pi'}(s)$ para todo s.
- Política óptima: π_* tal que $\pi_* \geq \pi$ para cualquier π.
- Políticas óptimas tienen la misma función de valor de estado óptima:

$$v_*(s) \doteq \max_\pi v_\pi(s) \quad \forall s \in \mathcal{S}$$

Funcion de valor óptima

- $\pi \geq \pi' \Leftrightarrow v_\pi(s) \geq v_{\pi'}(s)$ para todo s.
- Política óptima: π_* tal que $\pi_* \geq \pi$ para cualquier π.
- Políticas óptimas tienen la misma función de valor de estado óptima:

$$v_*(s) \doteq \max_\pi v_\pi(s) \quad \forall s \in \mathcal{S}$$

- Políticas óptimas tienen valor óptimo de la función de valor de pares estado-acción:

$$q_*(s,a) = \max_\pi q_\pi(s,a) \quad \forall s \in \mathcal{S}, a \in \mathcal{A}(s).$$

Funcion de valor óptima

- $\pi \geq \pi' \Leftrightarrow v_\pi(s) \geq v_{\pi'}(s)$ para todo s.
- Política óptima: π_* tal que $\pi_* \geq \pi$ para cualquier π.
- Políticas óptimas tienen la misma función de valor de estado óptima:

$$v_*(s) \doteq \max_\pi v_\pi(s) \quad \forall s \in \mathcal{S}$$

- Políticas óptimas tienen valor óptimo de la función de valor de pares estado-acción:

$$q_*(s,a) = \max_\pi q_\pi(s,a) \quad \forall s \in \mathcal{S}, a \in \mathcal{A}(s).$$

- Tenemos:

$$q_*(s,a) = \mathbb{E}\left[R_{t+1} + \gamma v_*(S_{t+1}) \,\middle|\, S_t = s, A_t = a\right]$$

Ecuaciones de optimalidad de Bellman

Ecuaciones de optimalidad de Bellman

- Para $v_*(s)$:

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

Ecuaciones de optimalidad de Bellman

- Para $v_*(s)$:

$$\begin{aligned} v_*(s) &= \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s,a) \\ &= \max_a \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma v_*(s')\right] \end{aligned}$$

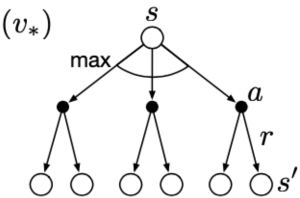
Ecuaciones de optimalidad de Bellman

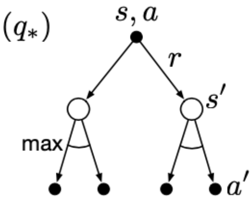
- Para $v_*(s)$:

$$\begin{aligned} v_*(s) &= \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s, a) \\ &= \max_a \sum_{s', r} p(s', r \mid s, a) \left[r + \gamma v_*(s') \right] \end{aligned}$$

- Para q_*:

$$\begin{aligned} q_*(s, a) &= \mathbb{E} \left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \mid S_t = s, A_t = a \right] \\ &= \sum_{s', r} p(s', r \mid s, a) \left[r + \gamma \max_{a'} q_*(s', a') \right] \end{aligned}$$


(v_*)
s
max
a
r
s'


(q_*)
s, a
r
s'
max
a'

Programación Dinámica

Programación Dinámica

1. Política inicial.

Programación Dinámica

1. Política inicial.
2. Evaluación de política.

Programación Dinámica

1. Política inicial.
2. Evaluación de política.
3. Mejorar política.

Cálculo de v_π

Cálculo de v_π

- Sistema de ecuaciones lineales:

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi(s') \right], \quad s \in \mathcal{S}$$

Cálculo de v_π

- Sistema de ecuaciones lineales:

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi(s') \right], \quad s \in \mathcal{S}$$

- Solución iterativa: evaluación iterativa de política

Cálculo de v_π

- Sistema de ecuaciones lineales:

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi(s') \right], \quad s \in \mathcal{S}$$

- Solución iterativa: evaluación iterativa de política
- Incializar $v_0(s)$, iterar:

$$v_{k+1}(s) \leftarrow \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_k(s') \right]$$

Cálculo de v_π

- Sistema de ecuaciones lineales:

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi(s') \right], \quad s \in \mathcal{S}$$

- Solución iterativa: evaluación iterativa de política
- Incializar $v_0(s)$, iterar:

$$v_{k+1}(s) \leftarrow \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_k(s') \right]$$

Cálculo de v_π

- Sistema de ecuaciones lineales:

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi(s') \right], \quad s \in \mathcal{S}$$

- Solución iterativa: evaluación iterativa de política
- Incializar $v_0(s)$, iterar:

$$v_{k+1}(s) \leftarrow \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_k(s') \right]$$

  - Usualmente actualización in-place.

Cálculo de v_π

- Sistema de ecuaciones lineales:

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi(s') \right], \quad s \in \mathcal{S}$$

- Solución iterativa: evaluación iterativa de política
- Incializar $v_0(s)$, iterar:

$$v_{k+1}(s) \leftarrow \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_k(s') \right]$$

  - Usualmente actualización in-place.
  - v_π es un punto fijo de este mapeo.

Cálculo de v_π

- Sistema de ecuaciones lineales:

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi(s')\right], \quad s \in \mathcal{S}$$

- Solución iterativa: evaluación iterativa de política
- Incializar $v_0(s)$, iterar:

$$v_{k+1}(s) \leftarrow \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_k(s')\right]$$

  - Usualmente actualización in-place.
  - v_π es un punto fijo de este mapeo.
  - Convergencia asimptótica: $v_k(s) \to v_\pi(s)$ cuando $k \to \infty$ si $\gamma < 1$ o episodios terminan eventualmente desde cualquier estado.

Evaluación de política

Inicialice $V(s), \pi(s)$

Evaluación de política

Incialice $V(s), \pi(s)$
repeat

Evaluación de política

Incialice $V(s), \pi(s)$

repeat

$\quad \delta \leftarrow 0$

$\quad$ **for** each $s \in \mathcal{S}$ **do**

Evaluación de política

Incialice $V(s), \pi(s)$
repeat
  $\delta \leftarrow 0$
  for each $s \in \mathcal{S}$ **do**
    $v \leftarrow V(s)$

Evaluación de política

Inicialice $V(s), \pi(s)$

repeat

$\delta \leftarrow 0$

for each $s \in \mathcal{S}$ **do**

$v \leftarrow V(s)$

$V(s) \leftarrow \sum_a \pi(s,a) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V(s')\right]$

Evaluación de política

Incialice $V(s), \pi(s)$
repeat
  $\delta \leftarrow 0$
  for each $s \in \mathcal{S}$ **do**
    $v \leftarrow V(s)$
    $V(s) \leftarrow \sum_a \pi(s,a) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V(s')\right]$
    $\delta \leftarrow \max\left(\delta, |v - V(s)|\right)$

Evaluación de política

Inicialice $V(s), \pi(s)$

repeat

$\quad \delta \leftarrow 0$

$\quad$ **for** each $s \in \mathcal{S}$ **do**

$\qquad v \leftarrow V(s)$

$\qquad V(s) \leftarrow \sum_a \pi(s,a) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V(s')\right]$

$\qquad \delta \leftarrow \max\left(\delta, |v - V(s)|\right)$

$\quad$ **end for**

Evaluación de política

Inicialice $V(s), \pi(s)$

repeat

  $\delta \leftarrow 0$

  for each $s \in \mathcal{S}$ **do**

    $v \leftarrow V(s)$

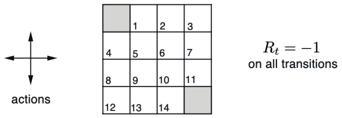
    $V(s) \leftarrow \sum_a \pi(s,a) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V(s')\right]$

    $\delta \leftarrow \max(\delta, |v - V(s)|)$

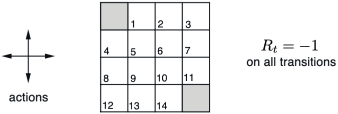
  end for

until $\delta < \epsilon$

Ejemplo

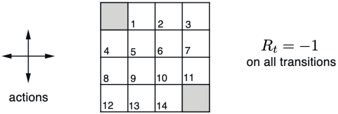


Ejemplo



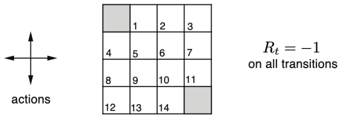
- Estados no terminales $\mathcal{S} = \{1, 3, \ldots 14\}$, estado terminal sombreado.

Ejemplo



- Estados no terminales $\mathcal{S} = \{1, 3, \dots 14\}$, estado terminal sombreado.
- Política aleatoria.

Ejemplo

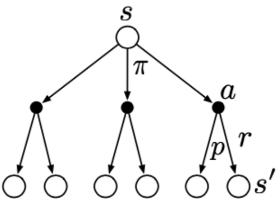


- Estados no terminales $\mathcal{S} = \{1, 3, \ldots 14\}$, estado terminal sombreado.
- Política aleatoria.
- $\gamma = 1$

0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00

0.00			
			0.00

$$V(s) \leftarrow \sum_a \pi(s,a) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V(s') \right]$$

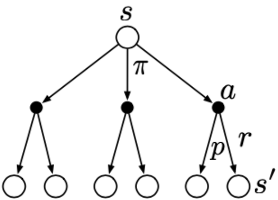


Backup diagram for v_π

0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00

0.00	-1.00		
			0.00

$$V(s) \leftarrow \sum_a \pi(s,a) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V(s') \right]$$

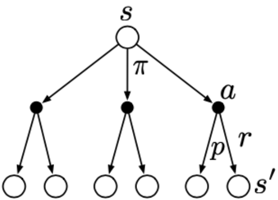


Backup diagram for v_π

0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00

0.00	-1.00	-1.00	
			0.00

$$V(s) \leftarrow \sum_a \pi(s,a) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V(s') \right]$$

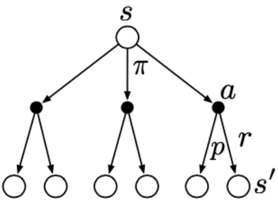


Backup diagram for v_π

0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00

0.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	0.00

$$V(s) \leftarrow \sum_{a} \pi(s,a) \sum_{s'} \sum_{r} p(s', r \mid s, a) \left[r + \gamma V(s') \right]$$

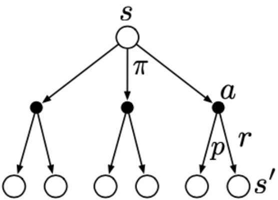


Backup diagram for v_π

0.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	0.00

0.00			
			0.00

$$V(s) \leftarrow \sum_a \pi(s,a) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V(s') \right]$$

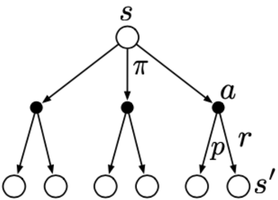


Backup diagram for v_π

0.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	0.00

0.00	-1.75		
			0.00

$$V(s) \leftarrow \sum_a \pi(s,a) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V(s') \right]$$

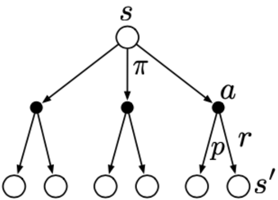


Backup diagram for v_π

0.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	0.00

0.00	-1.75	-2.00	
			0.00

$$V(s) \leftarrow \sum_a \pi(s,a) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V(s') \right]$$

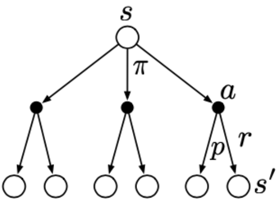


Backup diagram for v_π

0.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	0.00

0.00	-1.75	-2.00	-2.00
			0.00

$$V(s) \leftarrow \sum_a \pi(s,a) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V(s') \right]$$

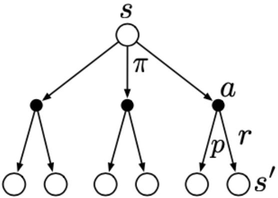


Backup diagram for v_π

0.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	0.00

0.00	-1.75	-2.00	-2.00
-1.75			
			0.00

$$V(s) \leftarrow \sum_a \pi(s,a) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V(s') \right]$$

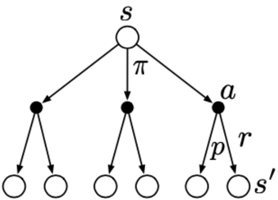


Backup diagram for v_π

0.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	0.00

0.00	-1.75	-2.00	-2.00
-1.75	-2.00	-2.00	-2.00
-2.00	-2.00	-2.00	-1.75
-2.00	-2.00	-1.75	0.00

$$V(s) \leftarrow \sum_a \pi(s,a) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V(s') \right]$$

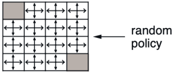


Backup diagram for v_π

v_k for the random policy | greedy policy w.r.t. v_k

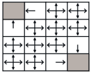
$k = 0$

0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0



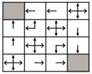
$k = 1$

0.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	0.0



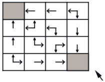
$k = 2$

0.0	-1.7	-2.0	-2.0
-1.7	-2.0	-2.0	-2.0
-2.0	-2.0	-2.0	-1.7
-2.0	-2.0	-1.7	0.0



$k = 3$

0.0	-2.4	-2.9	-3.0
-2.4	-2.9	-3.0	-2.9
-2.9	-3.0	-2.9	-2.4
-3.0	-2.9	-2.4	0.0

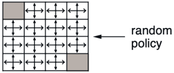


v_k for the random policy			

greedy policy w.r.t. v_k

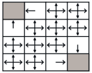
$k = 0$

0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0



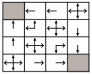
$k = 1$

0.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	0.0



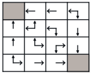
$k = 2$

0.0	-1.7	-2.0	-2.0
-1.7	-2.0	-2.0	-2.0
-2.0	-2.0	-2.0	-1.7
-2.0	-2.0	-1.7	0.0



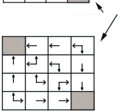
$k = 3$

0.0	-2.4	-2.9	-3.0
-2.4	-2.9	-3.0	-2.9
-2.9	-3.0	-2.9	-2.4
-3.0	-2.9	-2.4	0.0



$k = \infty$

0.0	-14.	-20.	-22.
-14.	-18.	-20.	-20.
-20.	-20.	-18.	-14.
-22.	-20.	-14.	0.0

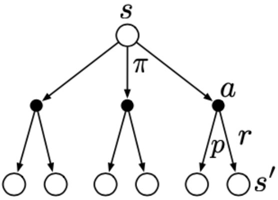


Supongamos viento empuja hacia la derecha con probabilidad 0.2

0.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	0.00

0.00	* * *		
			0.00

$$V(s) \leftarrow \sum_a \pi(s,a) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V(s')\right]$$



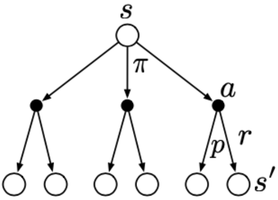
Backup diagram for v_π

Supongamos viento empuja hacia la derecha con probabilidad 0.2

0.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	0.00

0.00	-1.80		
			0.00

$$V(s) \leftarrow \sum_a \pi(s,a) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V(s') \right]$$



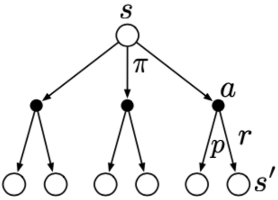
Backup diagram for v_π

Supongamos viento empuja hacia la derecha con probabilidad 0.2

0.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	0.00

0.00			
	$* * *$		0.00

$$V(s) \leftarrow \sum_a \pi(s,a) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V(s') \right]$$



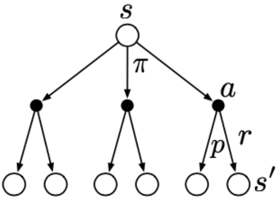
Backup diagram for v_π

Supongamos viento empuja hacia la derecha con probabilidad 0.2

0.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	0.00

0.00			
	-1.95		0.00

$$V(s) \leftarrow \sum_a \pi(s,a) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V(s')\right]$$



Backup diagram for v_π

Mejoramiento de política

Mejoramiento de política

- Suponga que conocemos $v_\pi(s)$, para $\pi(s)$ determinística y $\forall s \in \mathcal{S}$

Mejoramiento de política

- Suponga que conocemos $v_\pi(s)$, para $\pi(s)$ determinística y $\forall s \in \mathcal{S}$
- Dado $v_\pi(s)$ para un s dado, es mejor usar acción $a \neq \pi(s)$?

Mejoramiento de política

- Suponga que conocemos $v_\pi(s)$, para $\pi(s)$ determinística y $\forall s \in \mathcal{S}$
- Dado $v_\pi(s)$ para un s dado, es mejor usar acción $a \neq \pi(s)$?
- Si se selecciona a en s, y en adelante se usa π:

Mejoramiento de política

- Suponga que conocemos $v_\pi(s)$, para $\pi(s)$ determinística y $\forall s \in \mathcal{S}$
- Dado $v_\pi(s)$ para un s dado, es mejor usar acción $a \neq \pi(s)$?
- Si se selecciona a en s, y en adelante se usa π:

$$
\begin{aligned}
q_\pi(s,a) &= \mathbb{E}_\pi \left\{ R_{t+1} + \gamma v_\pi(s_{t+1}) \middle| S_t = s, A_t = a \right\} \\
&= \sum_{s'} \sum_{r} p(s', r \mid s, a) \left[r + \gamma v_\pi(s') \right]
\end{aligned}
$$

Mejoramiento de política

- Suponga que conocemos $v_\pi(s)$, para $\pi(s)$ determinística y $\forall s \in \mathcal{S}$
- Dado $v_\pi(s)$ para un s dado, es mejor usar acción $a \neq \pi(s)$?
- Si se selecciona a en s, y en adelante se usa π:

$$\begin{aligned} q_\pi(s,a) &= \mathbb{E}_\pi \left\{ R_{t+1} + \gamma v_\pi(s_{t+1}) \right| S_t = s, A_t = a \} \\ &= \sum_{s'} \sum_{r} p(s', r \mid s, a) \left[r + \gamma v_\pi(s') \right] \end{aligned}$$

- Si $q_\pi(s,a) > v_\pi(s)$ política igual a π, excepto en s, donde se reeemplaza $\pi(s)$ por a, debe ser mejor.

Teorema

(Teorema de mejoramiento de política)
Sean π, π' *dos políticas determinísticas,*

$$Si\ \forall s \in \mathcal{S}, q_\pi(s, \pi'(s)) \geq v_\pi(s) \Rightarrow v_{\pi'}(s) \geq v_\pi(s), \forall s \in \mathcal{S}$$

Teorema

(Teorema de mejoramiento de política)
Sean π, π' dos políticas determinísticas,

$$Si\ \forall s \in \mathcal{S}, q_\pi(s, \pi'(s)) \geq v_\pi(s) \Rightarrow v_{\pi'}(s) \geq v_\pi(s), \forall s \in \mathcal{S}$$

Demostración.

$v_\pi(s) \leq q_\pi(s, \pi'(s))$

$\square$

Teorema

(Teorema de mejoramiento de política)
Sean π, π' dos políticas determinísticas,

$$\text{Si } \forall s \in \mathcal{S}, q_\pi(s, \pi'(s)) \geq v_\pi(s) \Rightarrow v_{\pi'}(s) \geq v_\pi(s), \forall s \in \mathcal{S}$$

Demostración.

$$\begin{aligned} v_\pi(s) &\leq q_\pi(s, \pi'(s)) \\ &= \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma v_\pi(s_{t+1}) \middle| S_t = s \right\} \end{aligned}$$

Teorema

(Teorema de mejoramiento de política)
Sean π, π' dos políticas determinísticas,

$$Si\ \forall s \in \mathcal{S}, q_\pi(s, \pi'(s)) \geq v_\pi(s) \Rightarrow v_{\pi'}(s) \geq v_\pi(s), \forall s \in \mathcal{S}$$

Demostración.

$$\begin{aligned} v_\pi(s) &\leq q_\pi(s, \pi'(s)) \\ &= \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma v_\pi(s_{t+1}) \middle| S_t = s \right\} \\ &\leq \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma q_\pi(s_{t+1}, \pi'(s_{t+1})) \middle| S_t = s \right\} \end{aligned}$$

Teorema

(Teorema de mejoramiento de política)
Sean π, π' dos políticas determinísticas,

$$Si\ \forall s \in \mathcal{S}, q_\pi(s, \pi'(s)) \geq v_\pi(s) \Rightarrow v_{\pi'}(s) \geq v_\pi(s), \forall s \in \mathcal{S}$$

Demostración.

$$\begin{aligned} v_\pi(s) &\leq q_\pi(s, \pi'(s)) \\ &= \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma v_\pi(s_{t+1}) \middle| S_t = s \right\} \\ &\leq \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma q_\pi(s_{t+1}, \pi'(s_{t+1})) \middle| S_t = s \right\} \\ &= \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma \mathbb{E}_{\pi'}\left\{ R_{t+2} + \gamma v_\pi(s_{t+2}) \right\} \middle| S_t = s \right\} \end{aligned}$$

Teorema

(Teorema de mejoramiento de política)
Sean π, π' dos políticas determinísticas,

$$Si\ \forall s \in \mathcal{S}, q_\pi(s, \pi'(s)) \geq v_\pi(s) \Rightarrow v_{\pi'}(s) \geq v_\pi(s), \forall s \in \mathcal{S}$$

Demostración.

$$\begin{aligned}
v_\pi(s) &\leq q_\pi(s, \pi'(s)) \\
&= \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma v_\pi(s_{t+1}) \middle| S_t = s \right\} \\
&\leq \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma q_\pi(s_{t+1}, \pi'(s_{t+1})) \middle| S_t = s \right\} \\
&= \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma \mathbb{E}_{\pi'}\left\{ R_{t+2} + \gamma v_\pi(s_{t+2}) \right\} \middle| S_t = s \right\} \\
&= \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma R_{t+2} + \gamma^2 v_\pi(s_{t+2}) \middle| S_t = s \right\}
\end{aligned}$$

Teorema

(Teorema de mejoramiento de política)
Sean π, π' dos políticas determinísticas,

$$Si\ \forall s \in \mathcal{S}, q_\pi(s, \pi'(s)) \geq v_\pi(s) \Rightarrow v_{\pi'}(s) \geq v_\pi(s), \forall s \in \mathcal{S}$$

Demostración.

$$\begin{aligned}
v_\pi(s) &\leq q_\pi(s, \pi'(s)) \\
&= \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma v_\pi(s_{t+1}) \middle| S_t = s \right\} \\
&\leq \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma q_\pi(s_{t+1}, \pi'(s_{t+1})) \middle| S_t = s \right\} \\
&= \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma \mathbb{E}_{\pi'}\{R_{t+2} + \gamma v_\pi(s_{t+2})\} \middle| S_t = s \right\} \\
&= \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma R_{t+2} + \gamma^2 v_\pi(s_{t+2}) \middle| S_t = s \right\} \\
&\ \vdots
\end{aligned}$$

Teorema

(Teorema de mejoramiento de política)
Sean π, π' dos políticas determinísticas,

$$\textit{Si } \forall s \in \mathcal{S}, q_\pi(s, \pi'(s)) \geq v_\pi(s) \Rightarrow v_{\pi'}(s) \geq v_\pi(s), \forall s \in \mathcal{S}$$

Demostración.

$$\begin{aligned}
v_\pi(s) &\leq q_\pi(s, \pi'(s)) \\
&= \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma v_\pi(s_{t+1}) \middle| S_t = s \right\} \\
&\leq \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma q_\pi(s_{t+1}, \pi'(s_{t+1})) \middle| S_t = s \right\} \\
&= \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma \mathbb{E}_{\pi'}\{R_{t+2} + \gamma v_\pi(s_{t+2})\} \middle| S_t = s \right\} \\
&= \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma R_{t+2} + \gamma^2 v_\pi(s_{t+2}) \middle| S_t = s \right\} \\
&\vdots \\
&\leq \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \dots \middle| S_t = s \right\}
\end{aligned}$$

$\square$

Teorema

(Teorema de mejoramiento de política)
Sean π, π' dos políticas determinísticas,

$$Si\ \forall s \in \mathcal{S}, q_\pi(s, \pi'(s)) \geq v_\pi(s) \Rightarrow v_{\pi'}(s) \geq v_\pi(s), \forall s \in \mathcal{S}$$

Demostración.

$$\begin{aligned}
v_\pi(s) &\leq q_\pi(s, \pi'(s)) \\
&= \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma v_\pi(s_{t+1}) \,\middle|\, S_t = s \right\} \\
&\leq \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma q_\pi(s_{t+1}, \pi'(s_{t+1})) \,\middle|\, S_t = s \right\} \\
&= \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma \mathbb{E}_{\pi'}\left\{ R_{t+2} + \gamma v_\pi(s_{t+2}) \right\} \,\middle|\, S_t = s \right\} \\
&= \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma R_{t+2} + \gamma^2 v_\pi(s_{t+2}) \,\middle|\, S_t = s \right\} \\
&\ \vdots \\
&\leq \mathbb{E}_{\pi'}\left\{ R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \dots \,\middle|\, S_t = s \right\} = v_{\pi'}(s)
\end{aligned}$$

□

- Nueva política es greedy con respecto a q_π:

- Nueva política es greedy con respecto a q_π:

$$\pi'(s) = \arg\max_a q_\pi(s, a)$$

- Nueva política es greedy con respecto a q_π:

$$
\begin{aligned}
\pi'(s) &= \arg\max_a q_\pi(s,a) \\
&= \arg\max_a \mathbb{E}_\pi \left\{ R_{t+1} + \gamma v_\pi(S_{t+1}) \middle| S_t = s, A_t = a \right\}
\end{aligned}
$$

- Nueva política es greedy con respecto a q_π:

$$\begin{aligned}\pi'(s) &= \arg\max_a q_\pi(s,a)\\ &= \arg\max_a \mathbb{E}_\pi\left\{ R_{t+1} + \gamma v_\pi(S_{t+1}) \middle| S_t = s, A_t = a\right\}\\ &= \arg\max_a \sum_{s'}\sum_r p(s', r \mid s, a)\left[r + \gamma v_\pi(s')\right]\end{aligned}$$

- Nueva política es greedy con respecto a q_π:

$$\begin{aligned}\pi'(s) &= \arg\max_a q_\pi(s,a)\\ &= \arg\max_a \mathbb{E}_\pi\left\{ R_{t+1} + \gamma v_\pi(S_{t+1}) \middle| S_t = s, A_t = a\right\}\\ &= \arg\max_a \sum_{s'}\sum_r p(s', r \mid s, a)\left[r + \gamma v_\pi(s')\right]\end{aligned}$$

- π' satisface teorema de mejoramiento,

- Nueva política es greedy con respecto a q_π:

$$\begin{aligned}\pi'(s) &= \arg\max_a q_\pi(s,a)\\ &= \arg\max_a \mathbb{E}_\pi\left\{ R_{t+1} + \gamma v_\pi(S_{t+1}) \middle| S_t = s, A_t = a\right\}\\ &= \arg\max_a \sum_{s'}\sum_r p(s', r \mid s, a)\left[r + \gamma v_\pi(s')\right]\end{aligned}$$

- π' satisface teorema de mejoramiento, luego $v_{\pi'}(s) \geq v_\pi(s)\ \forall s \in \mathcal{S}$

.

- Nueva política es greedy con respecto a q_π:

$$\begin{aligned}
\pi'(s) &= \arg\max_a q_\pi(s,a) \\
&= \arg\max_a \mathbb{E}_\pi \left\{ R_{t+1} + \gamma v_\pi(S_{t+1}) \right| S_t = s, A_t = a\} \\
&= \arg\max_a \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi(s') \right]
\end{aligned}$$

- π' satisface teorema de mejoramiento, luego $v_{\pi'}(s) \geq v_\pi(s)\ \forall s \in \mathcal{S}$
- si $v_{\pi'}(s) = v_\pi(s)\ \forall s \in \mathcal{S}$:

- Nueva política es greedy con respecto a q_π:

$$\begin{aligned}\pi'(s) &= \arg\max_a q_\pi(s,a)\\ &= \arg\max_a \mathbb{E}_\pi\left\{ R_{t+1} + \gamma v_\pi(S_{t+1}) \middle| S_t = s, A_t = a\right\}\\ &= \arg\max_a \sum_{s'}\sum_r p(s', r \mid s, a)\left[r + \gamma v_\pi(s')\right]\end{aligned}$$

- π' satisface teorema de mejoramiento, luego $v_{\pi'}(s) \geq v_\pi(s)\ \forall s \in \mathcal{S}$
- si $v_{\pi'}(s) = v_\pi(s)\ \forall s \in \mathcal{S}$:

$$v_{\pi'}(s) = \max_a \sum_{s'}\sum_r p(s', r \mid s, a)\left[r + \gamma v_\pi(s')\right]$$

.

- Nueva política es greedy con respecto a q_π:

$$
\begin{aligned}
\pi'(s) &= \arg\max_a q_\pi(s,a) \\
&= \arg\max_a \mathbb{E}_\pi \left\{ R_{t+1} + \gamma v_\pi(S_{t+1}) \middle| S_t = s, A_t = a \right\} \\
&= \arg\max_a \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi(s') \right]
\end{aligned}
$$

- π' satisface teorema de mejoramiento, luego $v_{\pi'}(s) \geq v_\pi(s)\ \forall s \in \mathcal{S}$
- si $v_{\pi'}(s) = v_\pi(s)\ \forall s \in \mathcal{S}$:

$$
v_{\pi'}(s) = \max_a \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi(s') \right]
$$

π' satisface ecuación de optimalidad de Bellman.

Iteración de políticas

Iteración de políticas

π_0

Iteración de políticas

$$\pi_0 \xrightarrow{\mathbf{E}} v_{\pi_0}$$

Iteración de políticas

$$\pi_0 \xrightarrow{\mathbf{E}} v_{\pi_0} \xrightarrow{\mathbf{I}} \pi_1$$

Iteración de políticas

$$\pi_0 \xrightarrow{\mathbf{E}} v_{\pi_0} \xrightarrow{\mathbf{I}} \pi_1 \xrightarrow{\mathbf{E}} v_{\pi_1}$$

Iteración de políticas

$$\pi_0 \xrightarrow{\mathbf{E}} v_{\pi_0} \xrightarrow{\mathbf{I}} \pi_1 \xrightarrow{\mathbf{E}} v_{\pi_1} \xrightarrow{\mathbf{I}} \pi_2$$

Iteración de políticas

$$\pi_0 \xrightarrow{\mathbf{E}} v_{\pi_0} \xrightarrow{\mathbf{I}} \pi_1 \xrightarrow{\mathbf{E}} v_{\pi_1} \xrightarrow{\mathbf{I}} \pi_2 \xrightarrow{\mathbf{E}} \dots$$

Iteración de políticas

$$\pi_0 \xrightarrow{\mathbf{E}} v_{\pi_0} \xrightarrow{\mathbf{I}} \pi_1 \xrightarrow{\mathbf{E}} v_{\pi_1} \xrightarrow{\mathbf{I}} \pi_2 \xrightarrow{\mathbf{E}} \dots \xrightarrow{\mathbf{I}} \pi_*$$

Iteración de políticas

$$\pi_0 \xrightarrow{\mathbf{E}} v_{\pi_0} \xrightarrow{\mathbf{I}} \pi_1 \xrightarrow{\mathbf{E}} v_{\pi_1} \xrightarrow{\mathbf{I}} \pi_2 \xrightarrow{\mathbf{E}} \dots \xrightarrow{\mathbf{I}} \pi_* \xrightarrow{\mathbf{E}} v_{\pi^*}$$

Iteración de políticas

$$\pi_0 \xrightarrow{\mathbf{E}} v_{\pi_0} \xrightarrow{\mathbf{I}} \pi_1 \xrightarrow{\mathbf{E}} v_{\pi_1} \xrightarrow{\mathbf{I}} \pi_2 \xrightarrow{\mathbf{E}} \dots \xrightarrow{\mathbf{I}} \pi_* \xrightarrow{\mathbf{E}} v_{\pi^*}$$

- $\pi_{i+1} \geq \pi_i$

Iteración de políticas

$$\pi_0 \xrightarrow{\mathbf{E}} v_{\pi_0} \xrightarrow{\mathbf{I}} \pi_1 \xrightarrow{\mathbf{E}} v_{\pi_1} \xrightarrow{\mathbf{I}} \pi_2 \xrightarrow{\mathbf{E}} \dots \xrightarrow{\mathbf{I}} \pi_* \xrightarrow{\mathbf{E}} v_{\pi^*}$$

- $\pi_{i+1} \geq \pi_i$
- MDP finito tiene número finito de políticas $\Rightarrow$ converge a π_* en un número finito de iteraciones.

Iteración de políticas

$$\pi_0 \xrightarrow{\mathbf{E}} v_{\pi_0} \xrightarrow{\mathbf{I}} \pi_1 \xrightarrow{\mathbf{E}} v_{\pi_1} \xrightarrow{\mathbf{I}} \pi_2 \xrightarrow{\mathbf{E}} \dots \xrightarrow{\mathbf{I}} \pi_* \xrightarrow{\mathbf{E}} v_{\pi^*}$$

- $\pi_{i+1} \geq \pi_i$
- MDP finito tiene número finito de políticas $\Rightarrow$ converge a π_* en un número finito de iteraciones.
- Usualmente converge en pocas iteraciones.

v_k for the random policy

$k = 0$

0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0

$k = 1$

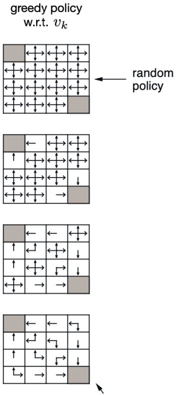
0.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	0.0

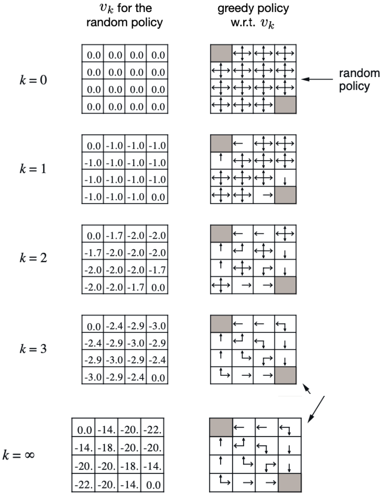
$k = 2$

0.0	-1.7	-2.0	-2.0
-1.7	-2.0	-2.0	-2.0
-2.0	-2.0	-2.0	-1.7
-2.0	-2.0	-1.7	0.0

$k = 3$

0.0	-2.4	-2.9	-3.0
-2.4	-2.9	-3.0	-2.9
-2.9	-3.0	-2.9	-2.4
-3.0	-2.9	-2.4	0.0




v_k for the random policy
greedy policy w.r.t. v_k
random policy
$k = 0$
0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0
$k = 1$
0.0 -1.0 -1.0 -1.0
-1.0 -1.0 -1.0 -1.0
-1.0 -1.0 -1.0 -1.0
-1.0 -1.0 -1.0 0.0
$k = 2$
0.0 -1.7 -2.0 -2.0
-1.7 -2.0 -2.0 -2.0
-2.0 -2.0 -2.0 -1.7
-2.0 -2.0 -1.7 0.0
$k = 3$
0.0 -2.4 -2.9 -3.0
-2.4 -2.9 -3.0 -2.9
-2.9 -3.0 -2.9 -2.4
-3.0 -2.9 -2.4 0.0
$k = \infty$
0.0 -14. -20. -22.
-14. -18. -20. -20.
-20. -20. -18. -14.
-22. -20. -14. 0.0

Iteración de Valor

- Evaluación de v_π, es costosa computacionalmente.

Iteración de Valor

- Evaluación de v_π, es costosa computacionalmente. Convergencia exacta sólo en el límite.

Iteración de Valor

- Evaluación de v_π, es costosa computacionalmente. Convergencia exacta sólo en el límite.
- Se puede obtener una política óptima, aun con valores de v_π que no son exactos.

Iteración de Valor

- Evaluación de v_π, es costosa computacionalmente. Convergencia exacta sólo en el límite.
- Se puede obtener una política óptima, aun con valores de v_π que no son exactos.
- Idea: truncar evaluación después de cierto número de iteraciones.

Iteración de Valor

- Evaluación de v_π, es costosa computacionalmente. Convergencia exacta sólo en el límite.
- Se puede obtener una política óptima, aun con valores de v_π que no son exactos.
- Idea: truncar evaluación después de cierto número de iteraciones.
- Una sóla iteración: algoritmo de iteración de valor:

Iteración de Valor

- Evaluación de v_π, es costosa computacionalmente. Convergencia exacta sólo en el límite.
- Se puede obtener una política óptima, aun con valores de v_π que no son exactos.
- Idea: truncar evaluación después de cierto número de iteraciones.
- Una sóla iteración: algoritmo de iteración de valor:

$$v_{k+1}(s) \doteq \max_a \mathbb{E}\left[R_{t+1} + \gamma v_k(S_{t+1}) \,\middle|\, S_t = s, A_t = a\right]$$

Iteración de Valor

- Evaluación de v_π, es costosa computacionalmente. Convergencia exacta sólo en el límite.
- Se puede obtener una política óptima, aun con valores de v_π que no son exactos.
- Idea: truncar evaluación después de cierto número de iteraciones.
- Una sóla iteración: algoritmo de iteración de valor:

$$\begin{aligned} v_{k+1}(s) &\doteq \max_a \mathbb{E}\left[R_{t+1} + \gamma v_k(S_{t+1}) \,\middle|\, S_t = s, A_t = a\right] \\ &= \max_a \sum_{s'} \sum_{r} p(s', r \mid s, a) \left[r + \gamma v_k(s')\right] \end{aligned}$$

Iteración de valor

Inicialice $V(s)$

Iteración de valor

Incialice $V(s)$
repeat

Iteración de valor

InciaIice $V(s)$

repeat

$\delta \leftarrow 0$

Iteración de valor

Inicialice $V(s)$
repeat
    $\delta \leftarrow 0$
    for each $s \in \mathcal{S}$ **do**

Iteración de valor

Inicialice $V(s)$
repeat
  $\delta \leftarrow 0$
  for each $s \in \mathcal{S}$ **do**
    $v \leftarrow V(s)$

Iteración de valor

Inicialice $V(s)$

repeat

$\delta \leftarrow 0$

for each $s \in \mathcal{S}$ **do**

$v \leftarrow V(s)$

$V(s) \leftarrow \max_a \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V_k(s')\right]$

Iteración de valor

Inicialice $V(s)$

repeat

  $\delta \leftarrow 0$

  for each $s \in \mathcal{S}$ **do**

    $v \leftarrow V(s)$

    $V(s) \leftarrow \max_a \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V_k(s')\right]$

    $\delta \leftarrow \max\left(\delta, |v - V(s)|\right)$

Iteración de valor

Inicialice $V(s)$
repeat
  $\delta \leftarrow 0$
  for each $s \in \mathcal{S}$ **do**
    $v \leftarrow V(s)$
    $V(s) \leftarrow \max_a \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V_k(s')\right]$
    $\delta \leftarrow \max\left(\delta, |v - V(s)|\right)$
  end for

Iteración de valor

Inicialice $V(s)$
repeat
  $\delta \leftarrow 0$
  for each $s \in \mathcal{S}$ **do**
    $v \leftarrow V(s)$
    $V(s) \leftarrow \max_a \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V_k(s')\right]$
    $\delta \leftarrow \max\left(\delta, |v - V(s)|\right)$
  end for
until $\delta < \epsilon$

Iteración de valor

Inicialice $V(s)$

repeat

  $\delta \leftarrow 0$

  for each $s \in \mathcal{S}$ **do**

    $v \leftarrow V(s)$

    $V(s) \leftarrow \max_a \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V_k(s')\right]$

    $\delta \leftarrow \max\left(\delta, |v - V(s)|\right)$

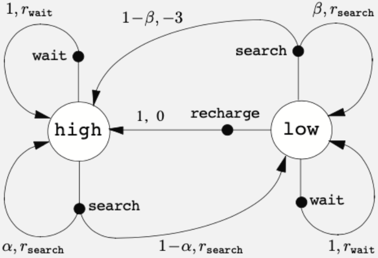
  end for

until $\delta < \epsilon$

Output π tal que $\pi(s) = \max_a \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma V(s')\right]$

Ejemplo

s	a	s'	$p(s' \mid s,a)$	$r(s,a,s')$
high	search	high	α	$r_{\texttt{search}}$
high	search	low	$1-\alpha$	$r_{\texttt{search}}$
low	search	high	$1-\beta$	-3
low	search	low	β	$r_{\texttt{search}}$
high	wait	high	1	$r_{\texttt{wait}}$
high	wait	low	0	-
low	wait	high	0	-
low	wait	low	1	$r_{\texttt{wait}}$
low	recharge	high	1	0
low	recharge	low	0	-

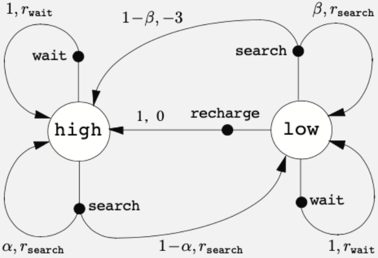


$$\alpha = \frac{3}{4},\ \beta = \frac{1}{4},\ r_{\texttt{search}} = 2,\ r_{\texttt{wait}} = 1,\ \gamma = 0{,}9$$

- Iteración de Valor a partir de $V(\texttt{high}) = 15$ y $V(\texttt{low}) = 12$.

Ejemplo

s	a	s'	$p(s' \mid s, a)$	$r(s, a, s')$
high	search	high	α	$r_{\texttt{search}}$
high	search	low	$1-\alpha$	$r_{\texttt{search}}$
low	search	high	$1-\beta$	-3
low	search	low	β	$r_{\texttt{search}}$
high	wait	high	1	$r_{\texttt{wait}}$
high	wait	low	0	-
low	wait	high	0	-
low	wait	low	1	$r_{\texttt{wait}}$
low	recharge	high	1	0
low	recharge	low	0	-

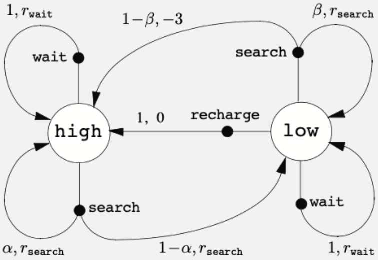


$$\alpha = \frac{3}{4},\ \beta = \frac{1}{4},\ r_{\texttt{search}} = 2,\ r_{\texttt{wait}} = 1,\ \gamma = 0{,}9$$

- Iteración de Valor a partir de $V(\texttt{high}) = 15$ y $V(\texttt{low}) = 12$.
  - Para V_1:

Ejemplo

s	a	s'	$p(s' \mid s, a)$	$r(s, a, s')$
high	search	high	α	$r_{\texttt{search}}$
high	search	low	$1-\alpha$	$r_{\texttt{search}}$
low	search	high	$1-\beta$	-3
low	search	low	β	$r_{\texttt{search}}$
high	wait	high	1	$r_{\texttt{wait}}$
high	wait	low	0	-
low	wait	high	0	-
low	wait	low	1	$r_{\texttt{wait}}$
low	recharge	high	1	0
low	recharge	low	0	-

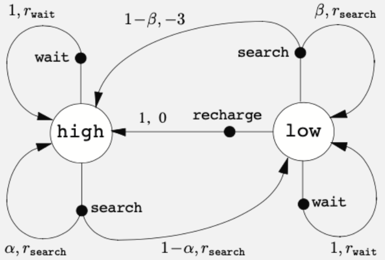


$$\alpha = \frac{3}{4},\ \beta = \frac{1}{4},\ r_{\texttt{search}} = 2,\ r_{\texttt{wait}} = 1,\ \gamma = 0{,}9$$

- Iteración de Valor a partir de $V(\texttt{high}) = 15$ y $V(\texttt{low}) = 12$.
  - Para V_1:
    `wait`: $1 + 0{,}9 \times 15 = 14{,}5$

Ejemplo

s	a	s'	$p(s' \mid s, a)$	$r(s, a, s')$
high	search	high	α	$r_{\texttt{search}}$
high	search	low	$1-\alpha$	$r_{\texttt{search}}$
low	search	high	$1-\beta$	-3
low	search	low	β	$r_{\texttt{search}}$
high	wait	high	1	$r_{\texttt{wait}}$
high	wait	low	0	-
low	wait	high	0	-
low	wait	low	1	$r_{\texttt{wait}}$
low	recharge	high	1	0
low	recharge	low	0	-



$$\alpha = \frac{3}{4},\ \beta = \frac{1}{4},\ r_{\texttt{search}} = 2,\ r_{\texttt{wait}} = 1,\ \gamma = 0{,}9$$

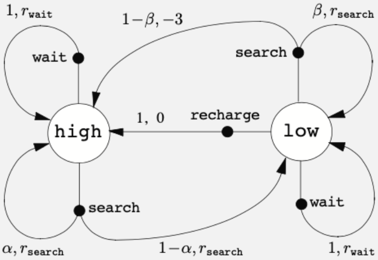
- Iteración de Valor a partir de $V(\texttt{high}) = 15$ y $V(\texttt{low}) = 12$.
  - Para V_1:

    `wait`: $1 + 0{,}9 \times 15 = 14{,}5$

    `search`: $0{,}75(2 + 0{,}9 \times 15) + 0{,}25(2 + 0{,}9 \times 12) = 14{,}825$

Ejemplo

s	a	s'	$p(s' \mid s, a)$	$r(s, a, s')$
`high`	`search`	`high`	α	$r_{\texttt{search}}$
`high`	`search`	`low`	$1-\alpha$	$r_{\texttt{search}}$
`low`	`search`	`high`	$1-\beta$	-3
`low`	`search`	`low`	β	$r_{\texttt{search}}$
`high`	`wait`	`high`	1	$r_{\texttt{wait}}$
`high`	`wait`	`low`	0	-
`low`	`wait`	`high`	0	-
`low`	`wait`	`low`	1	$r_{\texttt{wait}}$
`low`	`recharge`	`high`	1	0
`low`	`recharge`	`low`	0	-



$$\alpha = \frac{3}{4},\ \beta = \frac{1}{4},\ r_{\texttt{search}} = 2,\ r_{\texttt{wait}} = 1,\ \gamma = 0{,}9$$

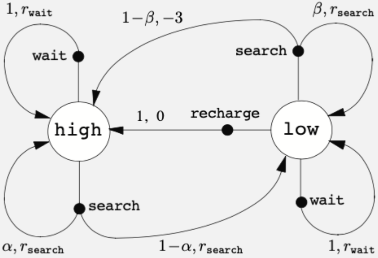
- Iteración de Valor a partir de $V(\texttt{high}) = 15$ y $V(\texttt{low}) = 12$.
  - Para V_1:

    `wait`: $1 + 0{,}9 \times 15 = 14{,}5$

    `search`: $0{,}75(2 + 0{,}9 \times 15) + 0{,}25(2 + 0{,}9 \times 12) = 14{,}825$
  - Para V_2:

Ejemplo

s	a	s'	$p(s' \mid s, a)$	$r(s, a, s')$
high	search	high	α	$r_{\texttt{search}}$
high	search	low	$1-\alpha$	$r_{\texttt{search}}$
low	search	high	$1-\beta$	-3
low	search	low	β	$r_{\texttt{search}}$
high	wait	high	1	$r_{\texttt{wait}}$
high	wait	low	0	-
low	wait	high	0	-
low	wait	low	1	$r_{\texttt{wait}}$
low	recharge	high	1	0
low	recharge	low	0	-

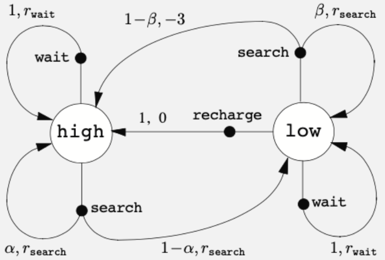


$$\alpha = \frac{3}{4},\ \beta = \frac{1}{4},\ r_{\texttt{search}} = 2,\ r_{\texttt{wait}} = 1,\ \gamma = 0{,}9$$

- Iteración de Valor a partir de $V(\texttt{high}) = 15$ y $V(\texttt{low}) = 12$.
  - Para V_1:
    `wait`: $1 + 0{,}9 \times 15 = 14{,}5$
    `search`: $0{,}75(2 + 0{,}9 \times 15) + 0{,}25(2 + 0{,}9 \times 12) = 14{,}825$
  - Para V_2:
    `wait`: $1 + 0{,}9 \times 12 = 11{,}8$

Ejemplo

s	a	s'	$p(s' \mid s, a)$	$r(s, a, s')$
`high`	`search`	`high`	α	$r_{\texttt{search}}$
`high`	`search`	`low`	$1-\alpha$	$r_{\texttt{search}}$
`low`	`search`	`high`	$1-\beta$	-3
`low`	`search`	`low`	β	$r_{\texttt{search}}$
`high`	`wait`	`high`	1	$r_{\texttt{wait}}$
`high`	`wait`	`low`	0	-
`low`	`wait`	`high`	0	-
`low`	`wait`	`low`	1	$r_{\texttt{wait}}$
`low`	`recharge`	`high`	1	0
`low`	`recharge`	`low`	0	-



$$\alpha = \frac{3}{4},\ \beta = \frac{1}{4},\ r_{\texttt{search}} = 2,\ r_{\texttt{wait}} = 1,\ \gamma = 0{,}9$$

- Iteración de Valor a partir de $V(\texttt{high}) = 15$ y $V(\texttt{low}) = 12$.
  - Para V_1:

    `wait`: $1 + 0{,}9 \times 15 = 14{,}5$

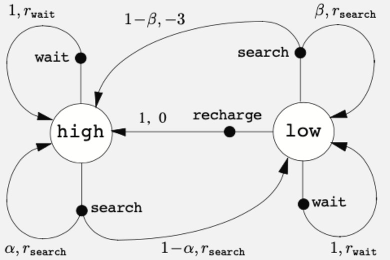
    `search`: $0{,}75(2 + 0{,}9 \times 15) + 0{,}25(2 + 0{,}9 \times 12) = 14{,}825$
  - Para V_2:

    `wait`: $1 + 0{,}9 \times 12 = 11{,}8$

    `search`: $0{,}75(-3 + 0{,}9 \times 15) + 0{,}25(2 + 0{,}9 \times 12) = 11{,}075$

Ejemplo

s	a	s'	$p(s' \mid s, a)$	$r(s, a, s')$
high	search	high	α	$r_{\texttt{search}}$
high	search	low	$1-\alpha$	$r_{\texttt{search}}$
low	search	high	$1-\beta$	-3
low	search	low	β	$r_{\texttt{search}}$
high	wait	high	1	$r_{\texttt{wait}}$
high	wait	low	0	-
low	wait	high	0	-
low	wait	low	1	$r_{\texttt{wait}}$
low	recharge	high	1	0
low	recharge	low	0	-

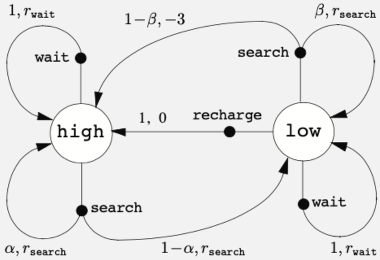


$$\alpha = \frac{3}{4},\ \beta = \frac{1}{4},\ r_{\texttt{search}} = 2,\ r_{\texttt{wait}} = 1,\ \gamma = 0{,}9$$

- Iteración de Valor a partir de $V(\texttt{high}) = 15$ y $V(\texttt{low}) = 12$.
  - Para V_1:
    - `wait`: $1 + 0{,}9 \times 15 = 14{,}5$
    - `search`: $0{,}75(2 + 0{,}9 \times 15) + 0{,}25(2 + 0{,}9 \times 12) = 14{,}825$
  - Para V_2:
    - `wait`: $1 + 0{,}9 \times 12 = 11{,}8$
    - `search`: $0{,}75(-3 + 0{,}9 \times 15) + 0{,}25(2 + 0{,}9 \times 12) = 11{,}075$
    - `recharge`: $0{,}9 \times 15 = 13{,}5$

Ejemplo

s	a	s'	$p(s' \mid s, a)$	$r(s, a, s')$
high	search	high	α	$r_{\texttt{search}}$
high	search	low	$1-\alpha$	$r_{\texttt{search}}$
low	search	high	$1-\beta$	-3
low	search	low	β	$r_{\texttt{search}}$
high	wait	high	1	$r_{\texttt{wait}}$
high	wait	low	0	-
low	wait	high	0	-
low	wait	low	1	$r_{\texttt{wait}}$
low	recharge	high	1	0
low	recharge	low	0	-



$$\alpha = \frac{3}{4},\ \beta = \frac{1}{4},\ r_{\texttt{search}} = 2,\ r_{\texttt{wait}} = 1,\ \gamma = 0{,}9$$

- Iteración de Valor a partir de $V(\texttt{high}) = 15$ y $V(\texttt{low}) = 12$.
  - Para V_1:
    - wait: $1 + 0{,}9 \times 15 = 14{,}5$
    - search: $0{,}75(2 + 0{,}9 \times 15) + 0{,}25(2 + 0{,}9 \times 12) = 14{,}825$
  - Para V_2:
    - wait: $1 + 0{,}9 \times 12 = 11{,}8$
    - search: $0{,}75(-3 + 0{,}9 \times 15) + 0{,}25(2 + 0{,}9 \times 12) = 11{,}075$
    - recharge: $0{,}9 \times 15 = 13{,}5$

- Política greedy?

- Política greedy?
  - Para V_1:

- Política greedy?
  - Para V_1:
    `wait`: $$1 + 0{,}9 \times 14{,}825 = 14{,}34$$

- Política greedy?
  - Para V_1:

$$\begin{array}{ll} \texttt{wait}: & 1 + 0{,}9 \times 14{,}825 = 14{,}34 \\ \texttt{search}: & 0{,}75(2 + 0{,}9 \times 14{,}825) + 0{,}25(2 + 0{,}9 \times 13{,}5) = 15{,}04 \end{array}$$

- Política greedy?
  - Para V_1:
    `wait`: $1 + 0{,}9 \times 14{,}825 = 14{,}34$
    `search`: $0{,}75(2 + 0{,}9 \times 14{,}825) + 0{,}25(2 + 0{,}9 \times 13{,}5) = 15{,}04$
  - Para V_2:

- Política greedy?
  - Para V_1:

    `wait`: $$1 + 0{,}9 \times 14{,}825 = 14{,}34$$

    `search`: $$0{,}75(2 + 0{,}9 \times 14{,}825) + 0{,}25(2 + 0{,}9 \times 13{,}5) = 15{,}04$$
  - Para V_2:

    `wait`: $$1 + 0{,}9 \times 13{,}5 = 13{,}15$$

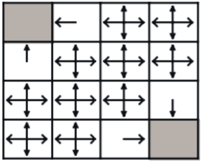
- Política greedy?
  - Para V_1:
    `wait`: $$1 + 0{,}9 \times 14{,}825 = 14{,}34$$
    `search`: $$0{,}75(2 + 0{,}9 \times 14{,}825) + 0{,}25(2 + 0{,}9 \times 13{,}5) = 15{,}04$$
  - Para V_2:
    `wait`: $$1 + 0{,}9 \times 13{,}5 = 13{,}15$$
    `search`: $$0{,}75(-3 + 0{,}9 \times 14{,}825) + 0{,}25(2 + 0{,}9 \times 13{,}5) = 11{,}29$$

- Política greedy?
  - Para V_1:
    `wait`: $1 + 0{,}9 \times 14{,}825 = 14{,}34$
    `search`: $0{,}75(2 + 0{,}9 \times 14{,}825) + 0{,}25(2 + 0{,}9 \times 13{,}5) = 15{,}04$
  - Para V_2:
    `wait`: $1 + 0{,}9 \times 13{,}5 = 13{,}15$
    `search`: $0{,}75(-3 + 0{,}9 \times 14{,}825) + 0{,}25(2 + 0{,}9 \times 13{,}5) = 11{,}29$
    `recharge`: $0{,}9 \times 14{,}825 = 13{,}34$

- Política greedy?
  - Para V_1:
    `wait`: $1 + 0{,}9 \times 14{,}825 = 14{,}34$
    `search`: $0{,}75(2 + 0{,}9 \times 14{,}825) + 0{,}25(2 + 0{,}9 \times 13{,}5) = 15{,}04$
  - Para V_2:
    `wait`: $1 + 0{,}9 \times 13{,}5 = 13{,}15$
    `search`: $0{,}75(-3 + 0{,}9 \times 14{,}825) + 0{,}25(2 + 0{,}9 \times 13{,}5) = 11{,}29$
    `recharge`: $0{,}9 \times 14{,}825 = 13{,}34$

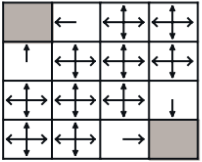
0.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	0.00

0.00			
			0.00



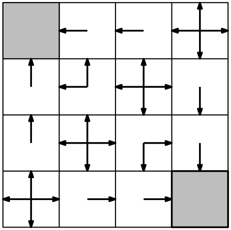
0.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	-1.00
-1.00	-1.00	-1.00	0.00

0.00	-1.00	-2.00	-2.00
-1.00	-2.00	-2.00	-2.00
-2.00	-2.00	-2.00	-1.00
-2.00	-2.00	-1.00	0.00



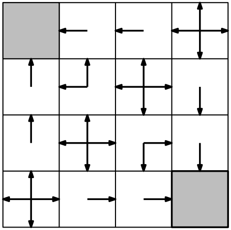
0.00	-1.00	-2.00	-2.00
-1.00	-2.00	-2.00	-2.00
-2.00	-2.00	-2.00	-1.00
-2.00	-2.00	-1.00	0.00

0.00			
			0.00



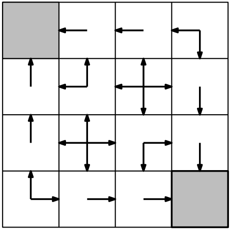
0.00	-1.00	-2.00	-2.00
-1.00	-2.00	-2.00	-2.00
-2.00	-2.00	-2.00	-1.00
-2.00	-2.00	-1.00	0.00

0.00	-1.00	-2.00	-3.00
-1.00	-2.00	-3.00	-2.00
-2.00	-2.00	-2.00	-1.00
-3.00	-2.00	-1.00	0.00



0.00	-1.00	-2.00	-2.00
-1.00	-2.00	-2.00	-2.00
-2.00	-2.00	-2.00	-1.00
-2.00	-2.00	-1.00	0.00

0.00	-1.00	-2.00	-3.00
-1.00	-2.00	-3.00	-2.00
-2.00	-3.00	-2.00	-1.00
-3.00	-2.00	-1.00	0.00



Ejemplo: Gambler's problem

Ejemplo: Gambler's problem

- Apostador apuesta $\$M$ dinero a que al lanzar una moneda caerá en cara.

Ejemplo: Gambler's problem

- Apostador apuesta $\$M$ dinero a que al lanzar una moneda caerá en cara.
  - Si cae cara, gana $\$M$, si cae sello pierde $\$M$.

Ejemplo: Gambler's problem

- Apostador apuesta $\$M$ dinero a que al lanzar una moneda caerá en cara.
  - Si cae cara, gana $\$M$, si cae sello pierde $\$M$.
  - Gana el juego si completa \$100, pierde si se queda sin dinero.

Ejemplo: Gambler's problem

- Apostador apuesta $\$M$ dinero a que al lanzar una moneda caerá en cara.
  - Si cae cara, gana $\$M$, si cae sello pierde $\$M$.
  - Gana el juego si completa \$100, pierde si se queda sin dinero.
  - MDP:

Ejemplo: Gambler's problem

- Apostador apuesta $\$M$ dinero a que al lanzar una moneda caerá en cara.
  - Si cae cara, gana $\$M$, si cae sello pierde $\$M$.
  - Gana el juego si completa \$100, pierde si se queda sin dinero.
  - MDP:
    - Estados $s \in \{1, 2, \dots, 99\}$.

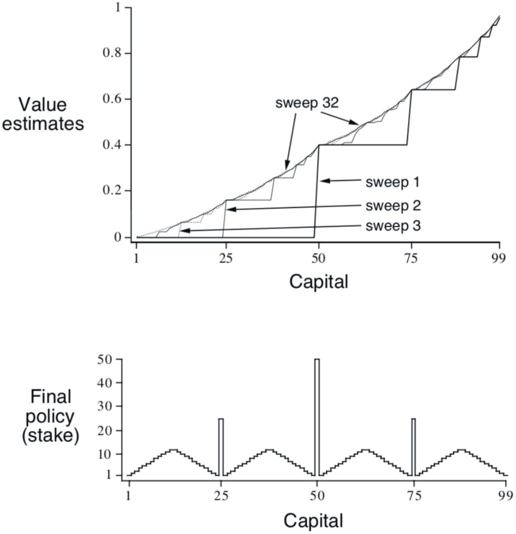
Ejemplo: Gambler's problem

- Apostador apuesta $\$M$ dinero a que al lanzar una moneda caerá en cara.
  - Si cae cara, gana $\$M$, si cae sello pierde $\$M$.
  - Gana el juego si completa \$100, pierde si se queda sin dinero.
  - MDP:
    - Estados $s \in \{1, 2, \ldots, 99\}$.
    - Acciones $a \in \{1, 2, \ldots, \text{mín}(s, 100 - s)\}$

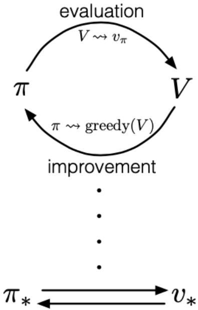
Ejemplo: Gambler's problem

- Apostador apuesta $\$M$ dinero a que al lanzar una moneda caerá en cara.
  - Si cae cara, gana $\$M$, si cae sello pierde $\$M$.
  - Gana el juego si completa \$100, pierde si se queda sin dinero.
  - MDP:
    - Estados $s \in \{1, 2, \ldots, 99\}$.
    - Acciones $a \in \{1, 2, \ldots, \text{mín}(s, 100 - s)\}$
    - Recompensa $+1$ cuando alcanza \$100, 0 en otro caso.

Solución para $p_h = 0{,}4$



Iteración de política generalizada



Iteración de política generalizada