Procesos de Decisión de Markov (MDPs)

Fernando Lozano

Universidad de los Andes

2 de febrero de 2023

Procesos de decisión de Markov (MDPs)

- Modelo matemático para toma de decisiones secuenciales, orientadas a una meta, bajo incertidumbre.

Procesos de decisión de Markov (MDPs)

- Modelo matemático para toma de decisiones secuenciales, orientadas a una meta, bajo incertidumbre.
- Diversos campos de aplicación (logística, control, ecología, economía, comunicaciones...).

Procesos de decisión de Markov (MDPs)

- Modelo matemático para toma de decisiones secuenciales, orientadas a una meta, bajo incertidumbre.
- Diversos campos de aplicación (logística, control, ecología, economía, comunicaciones...).
- En RL: aprendizaje a través de interacción del agente con el ambiente.

Procesos de decisión de Markov (MDPs)

- Modelo matemático para toma de decisiones secuenciales, orientadas a una meta, bajo incertidumbre.
- Diversos campos de aplicación (logística, control, ecología, economía, comunicaciones...).
- En RL: aprendizaje a través de interacción del agente con el ambiente.
- Modelo idealizado del problema de RL:

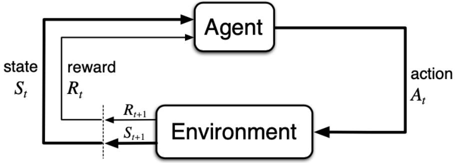
Procesos de decisión de Markov (MDPs)

- Modelo matemático para toma de decisiones secuenciales, orientadas a una meta, bajo incertidumbre.
- Diversos campos de aplicación (logística, control, ecología, economía, comunicaciones...).
- En RL: aprendizaje a través de interacción del agente con el ambiente.
- Modelo idealizado del problema de RL:
  - Permite análisis teórico, por ejemplo de convergencia de algoritmos de solución.

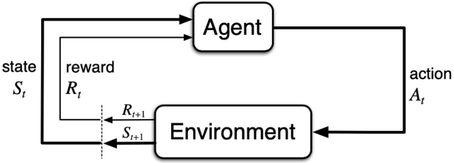
Procesos de decisión de Markov (MDPs)

- Modelo matemático para toma de decisiones secuenciales, orientadas a una meta, bajo incertidumbre.
- Diversos campos de aplicación (logística, control, ecología, economía, comunicaciones...).
- En RL: aprendizaje a través de interacción del agente con el ambiente.
- Modelo idealizado del problema de RL:
  - Permite análisis teórico, por ejemplo de convergencia de algoritmos de solución.
  - Extrapolar a situaciones reales.

Interfaz Agente-Ambiente

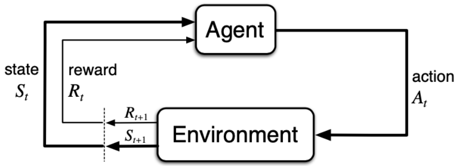


Interfaz Agente-Ambiente



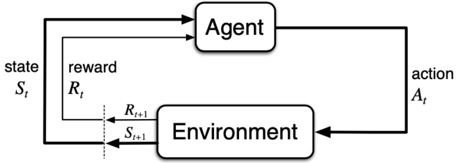
- Agente toma decisiones (acciones).

Interfaz Agente-Ambiente



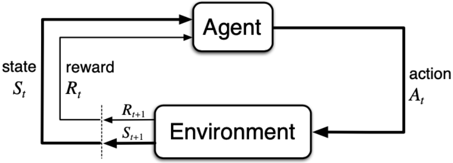
- Agente toma decisiones (acciones).
- Ambiente: todo lo que es exterior al agente.

Interfaz Agente-Ambiente



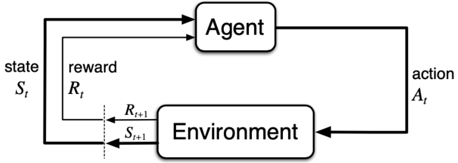
- Agente toma decisiones (acciones).
- Ambiente: todo lo que es exterior al agente.
- Ambiente y agente interactúan en pasos $t = 1, 2, \ldots$:

Interfaz Agente-Ambiente



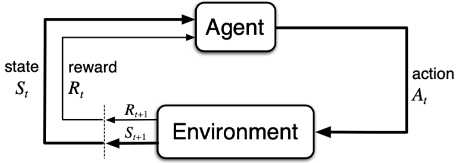
- Agente toma decisiones (acciones).
- Ambiente: todo lo que es exterior al agente.
- Ambiente y agente interactúan en pasos $t = 1, 2, \dots$:
    1. Agente recibe representación del estado del ambiente $S_t \in \mathcal{S}$

Interfaz Agente-Ambiente



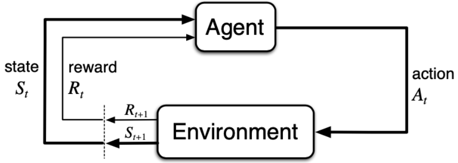
- Agente toma decisiones (acciones).
- Ambiente: todo lo que es exterior al agente.
- Ambiente y agente interactúan en pasos $t = 1, 2, \ldots$:
  1. Agente recibe representación del estado del ambiente $S_t \in \mathcal{S}$
  2. Selecciona acción $A_t \in \mathcal{A}(s_t)$.

Interfaz Agente-Ambiente



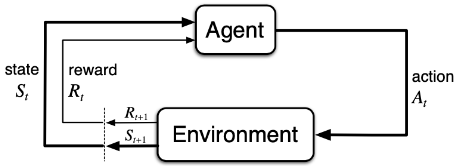
- Agente toma decisiones (acciones).
- Ambiente: todo lo que es exterior al agente.
- Ambiente y agente interactúan en pasos $t = 1, 2, \ldots$:
  1. Agente recibe representación del estado del ambiente $S_t \in \mathcal{S}$
  2. Selecciona acción $A_t \in \mathcal{A}(s_t)$.
  3. En tiempo $t + 1$ recibe recompensa $R_{t+1} \in \mathcal{R} \subset \mathbb{R}$

Interfaz Agente-Ambiente



- Agente toma decisiones (acciones).
- Ambiente: todo lo que es exterior al agente.
- Ambiente y agente interactúan en pasos $t = 1, 2, \ldots$:
  1. Agente recibe representación del estado del ambiente $S_t \in \mathcal{S}$
  2. Selecciona acción $A_t \in \mathcal{A}(s_t)$.
  3. En tiempo $t+1$ recibe recompensa $R_{t+1} \in \mathcal{R} \subset \mathbb{R}$ y pasa a estado $S_{t+1} \in \mathcal{S}$

Interfaz Agente-Ambiente



- Agente toma decisiones (acciones).
- Ambiente: todo lo que es exterior al agente.
- Ambiente y agente interactúan en pasos $t = 1, 2, \ldots$:
  1. Agente recibe representación del estado del ambiente $S_t \in \mathcal{S}$
  2. Selecciona acción $A_t \in \mathcal{A}(s_t)$.
  3. En tiempo $t + 1$ recibe recompensa $R_{t+1} \in \mathcal{R} \subset \mathbb{R}$ y pasa a estado $S_{t+1} \in \mathcal{S}$
- Trayectoria:

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, S_3, \ldots$$

Dinámica del MDP

Dinámica del MDP

- MDPs finitos: $|\mathcal{S}|, |\mathcal{A}|, |\mathcal{R}| < \infty$.

Dinámica del MDP

- 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:

Dinámica del MDP

- 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:

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

Dinámica del MDP

- 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:

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

  - $$p : \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \longrightarrow [0, 1]$$

Dinámica del MDP

- 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:

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

  - $$p \, : \, \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \longrightarrow [0, 1]$$
  - $$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r \mid s, a) = 1 \quad \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$$

- La función $p(s', r \mid s, a)$ caracteriza por completo la dinámica del MDP.

- La función $p(s', r \mid s, a)$ caracteriza por completo la dinámica del MDP.
- Valor de R_t, S_t depende únicamente de S_{t-1}, A_{t-1}, y no de valores anteriores en la trayectoria.

- La función $p(s', r \mid s, a)$ caracteriza por completo la dinámica del MDP.
- Valor de R_t, S_t depende únicamente de S_{t-1}, A_{t-1}, y no de valores anteriores en la trayectoria.
- Propiedad de Markov.

- La función $p(s', r \mid s, a)$ caracteriza por completo la dinámica del MDP.
- Valor de R_t, S_t depende únicamente de S_{t-1}, A_{t-1}, y no de valores anteriores en la trayectoria.
- Propiedad de Markov.
- Estado observado incluye toda la información relevante para el agente, en interacciones pasadas.

Propiedad de Markov, otra perspectiva

Propiedad de Markov, otra perspectiva

- Secuencia de variables aleatorias

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, S_3, \ldots$$

Propiedad de Markov, otra perspectiva

- Secuencia de variables aleatorias

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, S_3, \ldots$$

- Una señal de estado tiene la Propiedad de Markov si:

Propiedad de Markov, otra perspectiva

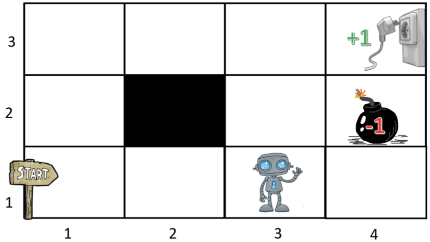
- Secuencia de variables aleatorias

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, S_3, \ldots$$

- Una señal de estado tiene la Propiedad de Markov si:

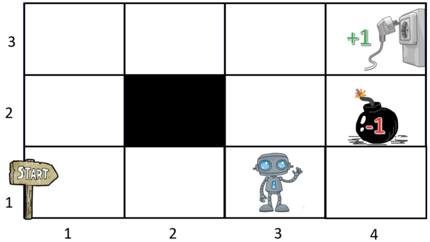
$$\begin{aligned}\mathbf{P}\left\{S_{t+1} = s', R_{t+1} = r \;\middle|\; S_t = s_t, A_t = a_t,\ R_t = r_t, S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots\right.\\ \left., R_1 = r_1, S_0 = s_0, A_0 = a_0\right\}\\ = \mathbf{P}\left\{S_{t+1} = s', R_{t+1} = r \;\middle|\; S_t = s_t, A_t = a_t\right\}\end{aligned}$$

para todo s', r, y todos los valores posibles de $s_{t+1}, r_{t+1}, s_t, a_t, r_t, s_{t-1}, a_{t-1}, \ldots, r_1, s_0, a_0$.



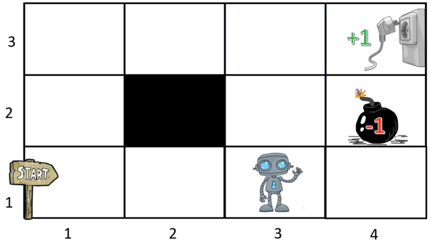
Acciones: u, d, l ,r

s	a	s'	r	$p(s', r \mid s, a)$
(1,1)	u	(1,2)	0	
(1,1)	u	(2,1)	0	
(1,1)	u	(1,2)	1	
(3,3)	r	(4,3)	1	
(3,3)	r	(4,3)	-1	



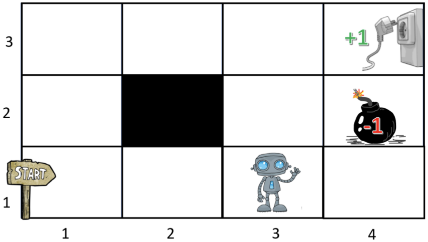
Acciones: u, d, l ,r

s	a	s'	r	$p(s', r \mid s, a)$
(1,1)	u	(1,2)	0	1
(1,1)	u	(2,1)	0	
(1,1)	u	(1,2)	1	
(3,3)	r	(4,3)	1	
(3,3)	r	(4,3)	-1	



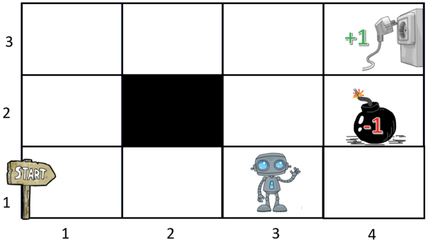
Acciones: u, d, l ,r

s	a	s'	r	$p(s', r \mid s, a)$
(1,1)	u	(1,2)	0	1
(1,1)	u	(2,1)	0	0
(1,1)	u	(1,2)	1	
(3,3)	r	(4,3)	1	
(3,3)	r	(4,3)	-1	



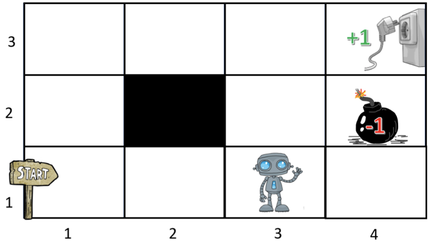
Acciones: u, d, l ,r

s	a	s'	r	$p(s', r \mid s, a)$
(1,1)	u	(1,2)	0	1
(1,1)	u	(2,1)	0	0
(1,1)	u	(1,2)	1	0
(3,3)	r	(4,3)	1	
(3,3)	r	(4,3)	-1	



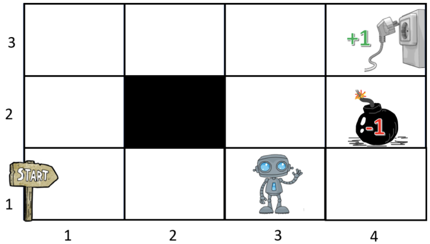
Acciones: u, d, l ,r

s	a	s'	r	$p(s', r \mid s, a)$
(1,1)	u	(1,2)	0	1
(1,1)	u	(2,1)	0	0
(1,1)	u	(1,2)	1	0
(3,3)	r	(4,3)	1	1
(3,3)	r	(4,3)	-1	



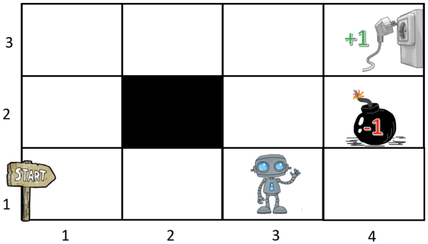
Acciones: u, d, l ,r

s	a	s'	r	$p(s', r \mid s, a)$
(1,1)	u	(1,2)	0	1
(1,1)	u	(2,1)	0	0
(1,1)	u	(1,2)	1	0
(3,3)	r	(4,3)	1	1
(3,3)	r	(4,3)	-1	0



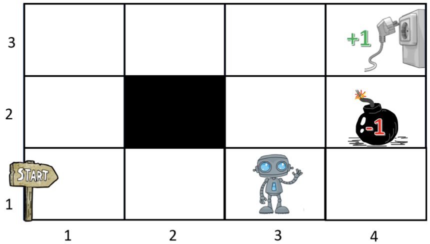
Acciones: u, d, l ,r pero con probabilidad $\frac{1}{2}$ otra dirección aleatoria.

s	a	s'	r	$p(s', r \mid s, a)$
(1,1)	u	(1,2)	0	
(1,1)	u	(2,1)	0	
(1,1)	u	(1,2)	1	
(3,1)	l	(3,2)	0	
(3,2)	u	(4,2)	-1	



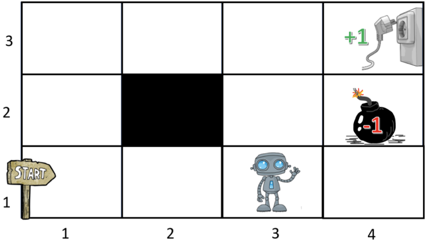
Acciones: u, d, l ,r pero con probabilidad $\frac{1}{2}$ otra dirección aleatoria.

s	a	s'	r	$p(s', r \mid s, a)$
(1,1)	u	(1,2)	0	0.5
(1,1)	u	(2,1)	0	
(1,1)	u	(1,2)	1	
(3,1)	l	(3,2)	0	
(3,2)	u	(4,2)	-1	



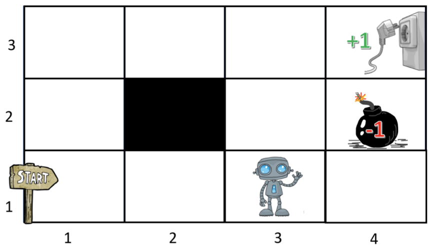
Acciones: u, d, l ,r pero con probabilidad $\frac{1}{2}$ otra dirección aleatoria.

s	a	s'	r	$p(s', r \mid s, a)$
(1,1)	u	(1,2)	0	0.5
(1,1)	u	(2,1)	0	0.5
(1,1)	u	(1,2)	1	
(3,1)	l	(3,2)	0	
(3,2)	u	(4,2)	-1	



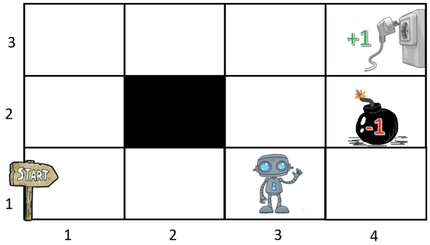
Acciones: u, d, l ,r pero con probabilidad $\frac{1}{2}$ otra dirección aleatoria.

s	a	s'	r	$p(s', r \mid s, a)$
(1,1)	u	(1,2)	0	0.5
(1,1)	u	(2,1)	0	0.5
(1,1)	u	(1,2)	1	0
(3,1)	l	(3,2)	0	
(3,2)	u	(4,2)	-1	



Acciones: u, d, l ,r pero con probabilidad $\frac{1}{2}$ otra dirección aleatoria.

s	a	s'	r	$p(s', r \mid s, a)$
(1,1)	u	(1,2)	0	0.5
(1,1)	u	(2,1)	0	0.5
(1,1)	u	(1,2)	1	0
(3,1)	l	(3,2)	0	0.25
(3,2)	u	(4,2)	-1	



Acciones: u, d, l ,r pero con probabilidad $\frac{1}{2}$ otra dirección aleatoria.

s	a	s'	r	$p(s', r \mid s, a)$
(1,1)	u	(1,2)	0	0.5
(1,1)	u	(2,1)	0	0.5
(1,1)	u	(1,2)	1	0
(3,1)	l	(3,2)	0	0.25
(3,2)	u	(4,2)	-1	0.25

Probabilidades de transición

Probabilidades de transición

- Probabilidad de que el estado resultante sea s', cuando se parte del estado s y se ejecuta la acción a.

Probabilidades de transición

- Probabilidad de que el estado resultante sea s', cuando se parte del estado s y se ejecuta la acción a.
- Función $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \longrightarrow [0, 1]$.

Probabilidades de transición

- Probabilidad de que el estado resultante sea s', cuando se parte del estado s y se ejecuta la acción a.
- Función $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \longrightarrow [0,1]$.

$$p(s' \mid s, a) \doteq \mathbf{P}\left\{S_t = s' \mid S_{t-1} = s, A_{t-1} = a\right\}$$

Probabilidades de transición

- Probabilidad de que el estado resultante sea s', cuando se parte del estado s y se ejecuta la acción a.
- Función $p \,:\, \mathcal{S} \times \mathcal{S} \times \mathcal{A} \longrightarrow [0,1]$.

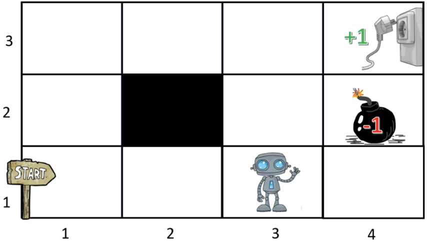
$$p(s' \mid s, a) \doteq \mathbf{P}\left\{S_t = s' \mid S_{t-1} = s, A_{t-1} = a\right\} = \sum_{r \in \mathcal{R}} p(s', r \mid s, a)$$

Probabilidades de transición

- Probabilidad de que el estado resultante sea s', cuando se parte del estado s y se ejecuta la acción a.
- Función $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \longrightarrow [0, 1]$.

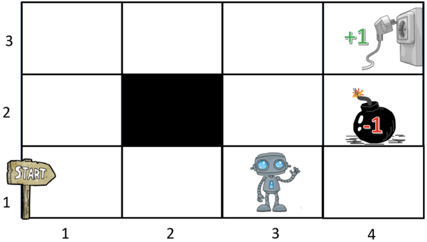
$$p(s' \mid s, a) \doteq \mathbf{P}\left\{S_t = s' \mid S_{t-1} = s, A_{t-1} = a\right\} = \sum_{r \in \mathcal{R}} p(s', r \mid s, a)$$

(sumamos probabilidades sobre los valores de recompensas posibles).



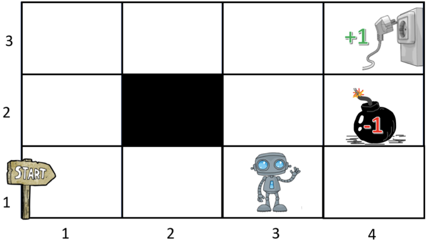
Acciones: u, d, l ,r pero con probabilidad $\frac{1}{2}$ otra dirección aleatoria.

s	a	s'	$p(s' \mid s, a)$
(1,1)	u	(1,2)	
(1,1)	u	(2,1)	
(3,1)	d	(3,2)	
(3,2)	u	(4,2)	



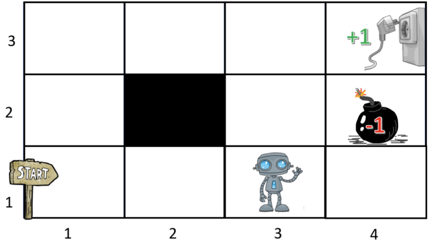
Acciones: u, d, l ,r pero con probabilidad $\frac{1}{2}$ otra dirección aleatoria.

s	a	s'	$p(s' \mid s, a)$
(1,1)	u	(1,2)	0.5
(1,1)	u	(2,1)	
(3,1)	d	(3,2)	
(3,2)	u	(4,2)	



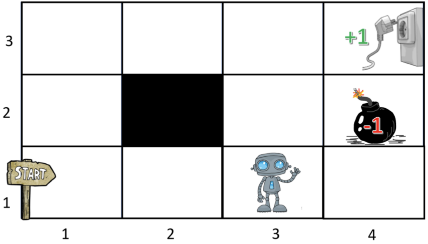
Acciones: u, d, l ,r pero con probabilidad $\frac{1}{2}$ otra dirección aleatoria.

s	a	s'	$p(s' \mid s, a)$
(1,1)	u	(1,2)	0.5
(1,1)	u	(2,1)	0.5
(3,1)	d	(3,2)	
(3,2)	u	(4,2)	



Acciones: u, d, l ,r pero con probabilidad $\frac{1}{2}$ otra dirección aleatoria.

s	a	s'	$p(s' \mid s, a)$
(1,1)	u	(1,2)	0.5
(1,1)	u	(2,1)	0.5
(3,1)	d	(3,2)	0.25
(3,2)	u	(4,2)	



Acciones: u, d, l ,r pero con probabilidad $\frac{1}{2}$ otra dirección aleatoria.

s	a	s'	$p(s' \mid s, a)$
(1,1)	u	(1,2)	0.5
(1,1)	u	(2,1)	0.5
(3,1)	d	(3,2)	0.25
(3,2)	u	(4,2)	0.25

Recompensa esperada para pares estado-acción

Recompensa esperada para pares estado-acción

- Valor esperado de la recompensa cuando está en estado s y ejecuta acción a.

Recompensa esperada para pares estado-acción

- Valor esperado de la recompensa cuando está en estado s y ejecuta acción a.
- Función $r \, : \, \mathcal{S} \times \mathcal{A} \longrightarrow \mathbb{R}$:

Recompensa esperada para pares estado-acción

- Valor esperado de la recompensa cuando está en estado s y ejecuta acción a.
- Función $r : \mathcal{S} \times \mathcal{A} \longrightarrow \mathbb{R}$:

$$r(s,a) \doteq \mathbb{E}\left[R_t \mid S_{t-1} = s, A_{t-1} = a\right]$$

Recompensa esperada para pares estado-acción

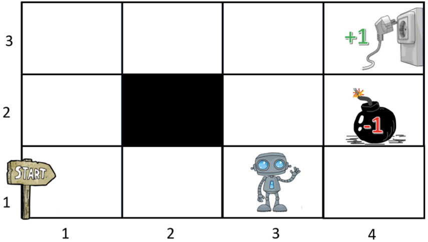
- Valor esperado de la recompensa cuando está en estado s y ejecuta acción a.
- Función $r : \mathcal{S} \times \mathcal{A} \longrightarrow \mathbb{R}$:

$$r(s,a) \doteq \mathbb{E}\left[R_t \mid S_{t-1} = s, A_{t-1} = a\right] = \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r \mid s, a)$$

Recompensa esperada para pares estado-acción

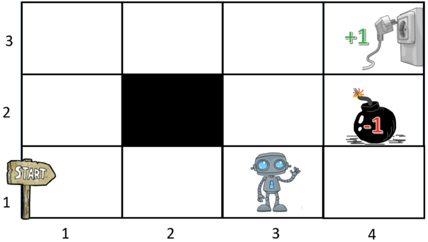
- Valor esperado de la recompensa cuando está en estado s y ejecuta acción a.
- Función $r : \mathcal{S} \times \mathcal{A} \longrightarrow \mathbb{R}$:

$$r(s,a) \doteq \mathbb{E}\left[R_t \mid S_{t-1} = s, A_{t-1} = a\right] = \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r \mid s, a)$$



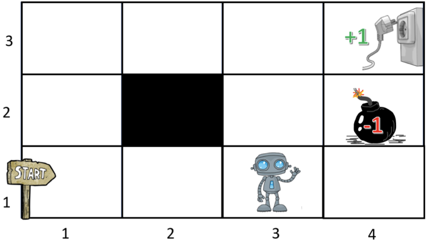
Acciones: u, d, l ,r pero con probabilidad $\frac{1}{2}$ otra dirección aleatoria.

s	a	$r(s,a)$
(1,1)	u	
(3,2)	u	
(3,2)	d	
(3,3)	l	



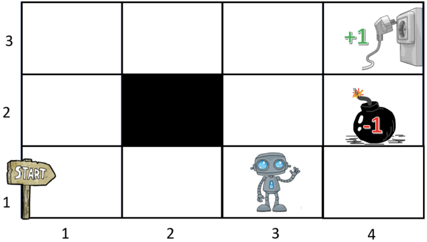
Acciones: u, d, l ,r pero con probabilidad $\frac{1}{2}$ otra dirección aleatoria.

s	a	$r(s,a)$
(1,1)	u	0
(3,2)	u	
(3,2)	d	
(3,3)	l	



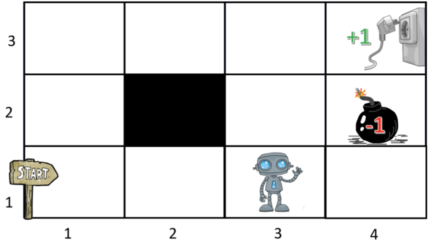
Acciones: u, d, l ,r pero con probabilidad $\frac{1}{2}$ otra dirección aleatoria.

s	a	$r(s,a)$
(1,1)	u	0
(3,2)	u	-0.25
(3,2)	d	
(3,3)	l	



Acciones: u, d, l ,r pero con probabilidad $\frac{1}{2}$ otra dirección aleatoria.

s	a	$r(s,a)$
(1,1)	u	0
(3,2)	u	-0.25
(3,2)	d	-0.5
(3,3)	l	



Acciones: u, d, l ,r pero con probabilidad $\frac{1}{2}$ otra dirección aleatoria.

s	a	$r(s,a)$
(1,1)	u	0
(3,2)	u	-0.25
(3,2)	d	-0.5
(3,3)	l	0.25

Recompensa esperada para triplas s, a, s'

Recompensa esperada para triplas s, a, s'

- Recompensa esperada para triplas estado-acción-estado siguiente.

Recompensa esperada para triplas s, a, s'

- Recompensa esperada para triplas estado-acción-estado siguiente.
- Función $r\ :\ \mathcal{S} \times \mathcal{A} \times \mathcal{S} \longrightarrow \mathbb{R}$:

Recompensa esperada para triplas s, a, s'

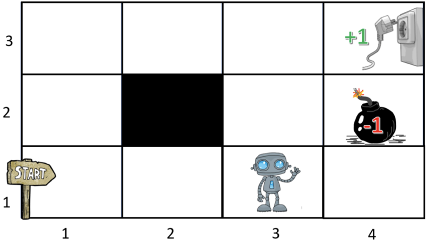
- Recompensa esperada para triplas estado-acción-estado siguiente.
- Función $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \longrightarrow \mathbb{R}$:

$$r(s, a, s') \doteq \mathbb{E}\left[R_t \mid S_{t-1} = s, A_{t-1} = a, S_t = s'\right]$$

Recompensa esperada para triplas s, a, s'

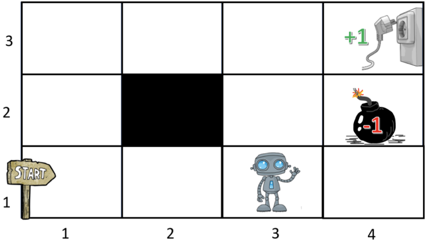
- Recompensa esperada para triplas estado-acción-estado siguiente.
- Función $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \longrightarrow \mathbb{R}$:

$$r(s,a,s') \doteq \mathbb{E}\left[R_t \mid S_{t-1} = s, A_{t-1} = a, S_t = s'\right] = \sum_{r \in \mathcal{R}} r \frac{p(s', r \mid s, a)}{p(s' \mid s, a)}$$



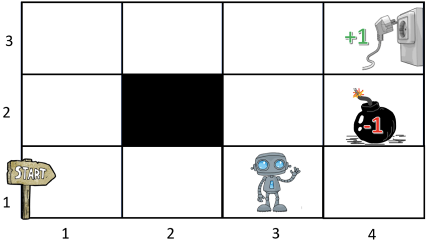
Acciones: u, d, l ,r pero con probabilidad $\frac{1}{2}$ otra dirección aleatoria.

s	a	s'	$r(s, a, s')$
(1,1)	u	(1,2)	
(3,2)	u	(4,2)	
(3,2)	d	(4,2)	
(3,3)	l	(4,3)	



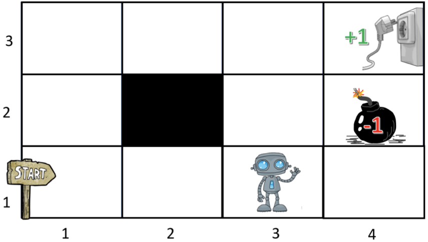
Acciones: u, d, l ,r pero con probabilidad $\frac{1}{2}$ otra dirección aleatoria.

s	a	s'	$r(s,a,s')$
(1,1)	u	(1,2)	0
(3,2)	u	(4,2)	
(3,2)	d	(4,2)	
(3,3)	l	(4,3)	



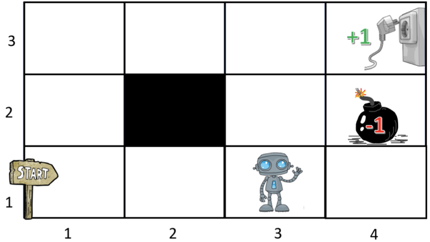
Acciones: u, d, l ,r pero con probabilidad $\frac{1}{2}$ otra dirección aleatoria.

s	a	s'	$r(s,a,s')$
(1,1)	u	(1,2)	0
(3,2)	u	(4,2)	-1
(3,2)	d	(4,2)	
(3,3)	l	(4,3)	



Acciones: u, d, l ,r pero con probabilidad $\frac{1}{2}$ otra dirección aleatoria.

s	a	s'	$r(s,a,s')$
(1,1)	u	(1,2)	0
(3,2)	u	(4,2)	-1
(3,2)	d	(4,2)	-1
(3,3)	l	(4,3)	



Acciones: u, d, l ,r pero con probabilidad $\frac{1}{2}$ otra dirección aleatoria.

s	a	s'	$r(s,a,s')$
(1,1)	u	(1,2)	0
(3,2)	u	(4,2)	-1
(3,2)	d	(4,2)	-1
(3,3)	l	(4,3)	1

Ejemplo: Robot reciclador

Ejemplo: Robot reciclador

- Robot que recolecta latas de Coca-Cola vacías en la oficina.

Ejemplo: Robot reciclador

- Robot que recolecta latas de Coca-Cola vacías en la oficina.
- Cámara, sensores, brazo...

Ejemplo: Robot reciclador

- Robot que recolecta latas de Coca-Cola vacías en la oficina.
- Cámara, sensores, brazo...
- Batería recargable.

Ejemplo: Robot reciclador

- Robot que recolecta latas de Coca-Cola vacías en la oficina.
- Cámara, sensores, brazo...
- Batería recargable.
- Meta: Recolectar máximo número de latas, sin quedarse varado por baterías.

Ejemplo: Robot reciclador

- Robot que recolecta latas de Coca-Cola vacías en la oficina.
- Cámara, sensores, brazo...
- Batería recargable.
- Meta: Recolectar máximo número de latas, sin quedarse varado por baterías.
- Elementos:

Ejemplo: Robot reciclador

- Robot que recolecta latas de Coca-Cola vacías en la oficina.
- Cámara, sensores, brazo...
- Batería recargable.
- Meta: Recolectar máximo número de latas, sin quedarse varado por baterías.
- Elementos:

  Acciones:

Ejemplo: Robot reciclador

- Robot que recolecta latas de Coca-Cola vacías en la oficina.
- Cámara, sensores, brazo...
- Batería recargable.
- Meta: Recolectar máximo número de latas, sin quedarse varado por baterías.
- Elementos:

  Acciones: Buscar lata,

Ejemplo: Robot reciclador

- Robot que recolecta latas de Coca-Cola vacías en la oficina.
- Cámara, sensores, brazo...
- Batería recargable.
- Meta: Recolectar máximo número de latas, sin quedarse varado por baterías.
- Elementos:

  Acciones: Buscar lata, esperar lata,

Ejemplo: Robot reciclador

- Robot que recolecta latas de Coca-Cola vacías en la oficina.
- Cámara, sensores, brazo...
- Batería recargable.
- Meta: Recolectar máximo número de latas, sin quedarse varado por baterías.
- Elementos:

  Acciones: Buscar lata, esperar lata, ir a recargar batería.

Ejemplo: Robot reciclador

- Robot que recolecta latas de Coca-Cola vacías en la oficina.
- Cámara, sensores, brazo...
- Batería recargable.
- Meta: Recolectar máximo número de latas, sin quedarse varado por baterías.
- Elementos:

  Acciones: Buscar lata, esperar lata, ir a recargar batería.

  Estados:

Ejemplo: Robot reciclador

- Robot que recolecta latas de Coca-Cola vacías en la oficina.
- Cámara, sensores, brazo...
- Batería recargable.
- Meta: Recolectar máximo número de latas, sin quedarse varado por baterías.
- Elementos:

  Acciones: Buscar lata, esperar lata, ir a recargar batería.

  Estados: Carga de la batería.

Ejemplo: Robot reciclador

- Robot que recolecta latas de Coca-Cola vacías en la oficina.
- Cámara, sensores, brazo...
- Batería recargable.
- Meta: Recolectar máximo número de latas, sin quedarse varado por baterías.
- Elementos:

  Acciones: Buscar lata, esperar lata, ir a recargar batería.
  Estados: Carga de la batería.
  Recompensas:

Ejemplo: Robot reciclador

- Robot que recolecta latas de Coca-Cola vacías en la oficina.
- Cámara, sensores, brazo...
- Batería recargable.
- Meta: Recolectar máximo número de latas, sin quedarse varado por baterías.
- Elementos:

  Acciones: Buscar lata, esperar lata, ir a recargar batería.
  Estados: Carga de la batería.
  Recompensas: Positiva si recoge una lata,

Ejemplo: Robot reciclador

- Robot que recolecta latas de Coca-Cola vacías en la oficina.
- Cámara, sensores, brazo...
- Batería recargable.
- Meta: Recolectar máximo número de latas, sin quedarse varado por baterías.
- Elementos:

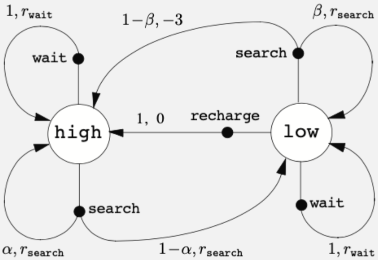
  Acciones: Buscar lata, esperar lata, ir a recargar batería.

  Estados: Carga de la batería.

  Recompensas: Positiva si recoge una lata, muy negativa si se queda sin batería.

Ejemplo: Robot reciclador

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	-



Señal de recompensa

Señal de recompensa

- Agente aprende a maximizar la recompensa acumulada a lo largo del tiempo.

Señal de recompensa

- Agente aprende a maximizar la recompensa acumulada a lo largo del tiempo.
- Recompensa acumulada debe corresponder a la meta de aprendizaje.

Señal de recompensa

- Agente aprende a maximizar la recompensa acumulada a lo largo del tiempo.
- Recompensa acumulada debe corresponder a la meta de aprendizaje.
  - Juego de tablero:

Señal de recompensa

- Agente aprende a maximizar la recompensa acumulada a lo largo del tiempo.
- Recompensa acumulada debe corresponder a la meta de aprendizaje.
  - Juego de tablero: +1, ganar, -1 perder, cero empatar.

Señal de recompensa

- Agente aprende a maximizar la recompensa acumulada a lo largo del tiempo.
- Recompensa acumulada debe corresponder a la meta de aprendizaje.
  - Juego de tablero: +1, ganar, -1 perder, cero empatar.
  - Robot reciclador:

Señal de recompensa

- Agente aprende a maximizar la recompensa acumulada a lo largo del tiempo.
- Recompensa acumulada debe corresponder a la meta de aprendizaje.
  - Juego de tablero: +1, ganar, -1 perder, cero empatar.
  - Robot reciclador: positiva al recoger lata, muy negativa al descargarse

Señal de recompensa

- Agente aprende a maximizar la recompensa acumulada a lo largo del tiempo.
- Recompensa acumulada debe corresponder a la meta de aprendizaje.
  - Juego de tablero: +1, ganar, -1 perder, cero empatar.
  - Robot reciclador: positiva al recoger lata, muy negativa al descargarse
  - No se usa para indicar cómo lograr la meta de aprendizaje.

Retorno

Retorno

- Tareas episódicas:

Retorno

- Tareas episódicas: estado terminal S_T.

Retorno

- Tareas episódicas: estado terminal S_T.
- Retorno:
$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$$

Retorno

- Tareas episódicas: estado terminal S_T.
- Retorno:
$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$$
- Meta: Maximizar retorno esperado.

Retorno

- Tareas episódicas: estado terminal S_T.
- Retorno:

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$$

- Meta: Maximizar retorno esperado.
- Tareas con horizonte infinito

Retorno

- Tareas episódicas: estado terminal S_T.
- Retorno:

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$$

- Meta: Maximizar retorno esperado.
- Tareas con horizonte infinito : retorno con descuento $0 \leq \gamma \leq 1$:

Retorno

- Tareas episódicas: estado terminal S_T.
- Retorno:

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$$

- Meta: Maximizar retorno esperado.
- Tareas con horizonte infinito : retorno con descuento $0 \leq \gamma \leq 1$:

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$$

Retorno

- Tareas episódicas: estado terminal S_T.
- Retorno:

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$$

- Meta: Maximizar retorno esperado.
- Tareas con horizonte infinito : retorno con descuento $0 \leq \gamma \leq 1$:

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

Retorno

- Tareas episódicas: estado terminal S_T.
- Retorno:

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$$

- Meta: Maximizar retorno esperado.
- Tareas con horizonte infinito : retorno con descuento $0 \leq \gamma \leq 1$:

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

Retorno

- Tareas episódicas: estado terminal S_T.
- Retorno:

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$$

- Meta: Maximizar retorno esperado.
- Tareas con horizonte infinito : retorno con descuento $0 \leq \gamma \leq 1$:

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

  - Preferencia por recompensas recientes.

Retorno

- Tareas episódicas: estado terminal S_T.
- Retorno:

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$$

- Meta: Maximizar retorno esperado.
- Tareas con horizonte infinito : retorno con descuento $0 \leq \gamma \leq 1$:

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

  - Preferencia por recompensas recientes.
  - Valor típico $\gamma \approx 0{,}9$

Retorno

- Tareas episódicas: estado terminal S_T.
- Retorno:

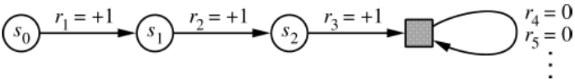
$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$$

- Meta: Maximizar retorno esperado.
- Tareas con horizonte infinito : retorno con descuento $0 \leq \gamma \leq 1$:

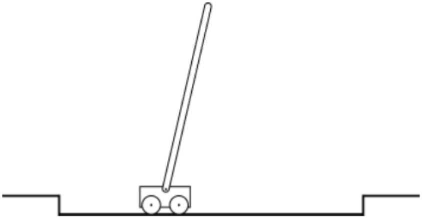
$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} = R_{t+1} + \gamma G_{t+1}$$

  - Preferencia por recompensas recientes.
  - Valor típico $\gamma \approx 0{,}9$
  - NO sirve para maximizar recompensas promedio!

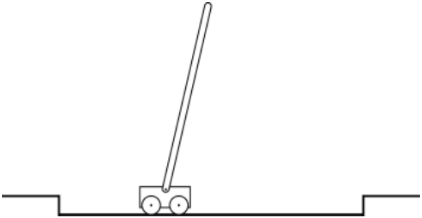
Visión unificada



Ejemplo

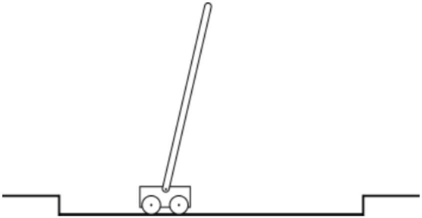


Ejemplo



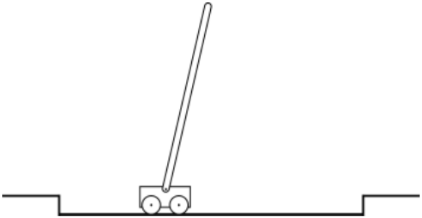
- Objetivo: Aplicar fuerzas al carro de manera que el palo no se caiga.

Ejemplo



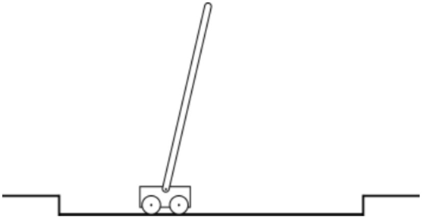
- Objetivo: Aplicar fuerzas al carro de manera que el palo no se caiga.
- Cuando se cae, el palo se devuelve a su posición vertical.

Ejemplo



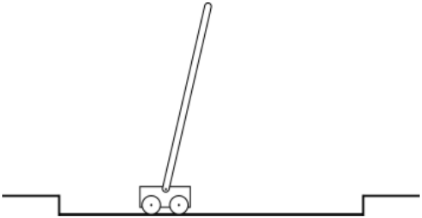
- Objetivo: Aplicar fuerzas al carro de manera que el palo no se caiga.
- Cuando se cae, el palo se devuelve a su posición vertical.
- Tarea episódica:

Ejemplo



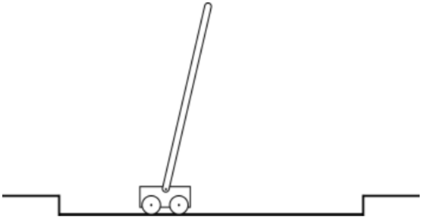
- Objetivo: Aplicar fuerzas al carro de manera que el palo no se caiga.
- Cuando se cae, el palo se devuelve a su posición vertical.
- Tarea episódica:
  - Episodio: cada intento de mantener el palo vertical.

Ejemplo



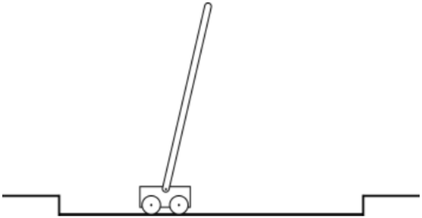
- Objetivo: Aplicar fuerzas al carro de manera que el palo no se caiga.
- Cuando se cae, el palo se devuelve a su posición vertical.
- Tarea episódica:
    - Episodio: cada intento de mantener el palo vertical.
    - Recompensa:

Ejemplo



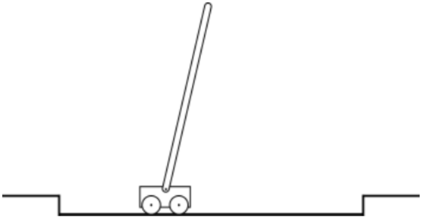
- Objetivo: Aplicar fuerzas al carro de manera que el palo no se caiga.
- Cuando se cae, el palo se devuelve a su posición vertical.
- Tarea episódica:
  - Episodio: cada intento de mantener el palo vertical.
  - Recompensa: $+1$ por cada iteración en que no se cae.

Ejemplo



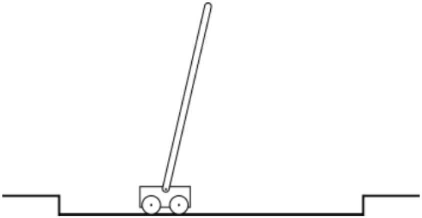
- Objetivo: Aplicar fuerzas al carro de manera que el palo no se caiga.
- Cuando se cae, el palo se devuelve a su posición vertical.
- Tarea episódica:
  - Episodio: cada intento de mantener el palo vertical.
  - Recompensa: $+1$ por cada iteración en que no se cae.
  - Retorno: tiempo total en que el palo está balanceado.

Ejemplo



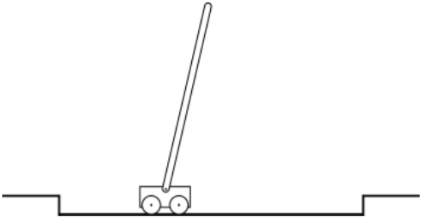
- Objetivo: Aplicar fuerzas al carro de manera que el palo no se caiga.
- Cuando se cae, el palo se devuelve a su posición vertical.
- Tarea episódica:
  - Episodio: cada intento de mantener el palo vertical.
  - Recompensa: $+1$ por cada iteración en que no se cae.
  - Retorno: tiempo total en que el palo está balanceado.
- Tarea contínua:

Ejemplo



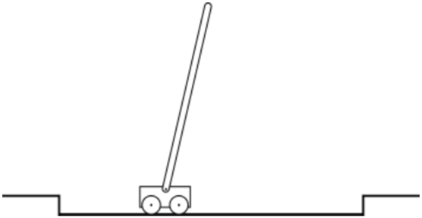
- Objetivo: Aplicar fuerzas al carro de manera que el palo no se caiga.
- Cuando se cae, el palo se devuelve a su posición vertical.
- Tarea episódica:
  - Episodio: cada intento de mantener el palo vertical.
  - Recompensa: +1 por cada iteración en que no se cae.
  - Retorno: tiempo total en que el palo está balanceado.
- Tarea contínua:
  - Recompensa:

Ejemplo



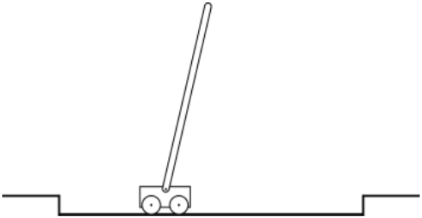
- Objetivo: Aplicar fuerzas al carro de manera que el palo no se caiga.
- Cuando se cae, el palo se devuelve a su posición vertical.
- Tarea episódica:
  - Episodio: cada intento de mantener el palo vertical.
  - Recompensa: +1 por cada iteración en que no se cae.
  - Retorno: tiempo total en que el palo está balanceado.
- Tarea contínua:
  - Recompensa: -1 si se cae, 0 si no.

Ejemplo



- Objetivo: Aplicar fuerzas al carro de manera que el palo no se caiga.
- Cuando se cae, el palo se devuelve a su posición vertical.
- Tarea episódica:
  - Episodio: cada intento de mantener el palo vertical.
  - Recompensa: +1 por cada iteración en que no se cae.
  - Retorno: tiempo total en que el palo está balanceado.
- Tarea contínua:
  - Recompensa: -1 si se cae, 0 si no.
  - Decuento γ

Ejemplo



- Objetivo: Aplicar fuerzas al carro de manera que el palo no se caiga.
- Cuando se cae, el palo se devuelve a su posición vertical.
- Tarea episódica:
  - Episodio: cada intento de mantener el palo vertical.
  - Recompensa: $+1$ por cada iteración en que no se cae.
  - Retorno: tiempo total en que el palo está balanceado.
- Tarea contínua:
  - Recompensa: -1 si se cae, 0 si no.
  - Decuento γ, retorno: $-\gamma^k$

Política (policy)

- Define comportamiento del agente.

Política (policy)

- Define comportamiento del agente.
- Mapeo estado $\rightarrow$ probabilidad de seleccionar acción en ese estado:

Política (policy)

- Define comportamiento del agente.
- Mapeo estado $\rightarrow$ probabilidad de seleccionar acción en ese estado:

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

Política (policy)

- Define comportamiento del agente.
- Mapeo estado $\rightarrow$ probabilidad de seleccionar acción en ese estado:

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

Política (policy)

- Define comportamiento del agente.
- Mapeo estado $\rightarrow$ probabilidad de seleccionar acción en ese estado:

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

  - Políticas determinísticas: $\pi(s) = a$

Política (policy)

- Define comportamiento del agente.
- Mapeo estado $\rightarrow$ probabilidad de seleccionar acción en ese estado:

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

  - Políticas determinísticas: $\pi(s) = a$
  - Políticas soft: $\pi(a \mid s) > 0 \ \forall a \in \mathcal{A}$

Política (policy)

- Define comportamiento del agente.
- Mapeo estado $\rightarrow$ probabilidad de seleccionar acción en ese estado:

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

  - Políticas determinísticas: $\pi(s) = a$
  - Políticas soft: $\pi(a \mid s) > 0 \ \forall a \in \mathcal{A} \rightarrow$ exploración.

Política (policy)

- Define comportamiento del agente.
- Mapeo estado $\rightarrow$ probabilidad de seleccionar acción en ese estado:

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

  - Políticas determinísticas: $\pi(s) = a$
  - Políticas soft: $\pi(a \mid s) > 0 \ \forall a \in \mathcal{A} \rightarrow$ exploración.
- Aprendizaje: Encontrar buenas políticas.

Funciones de valor

Funciones de valor

- Valor de un estado:

Funciones de valor

- Valor de un estado:
  - Indica qué tan bueno es estar en un estado dado, en términos de el retorno esperado .

Funciones de valor

- Valor de un estado:
  - Indica qué tan bueno es estar en un estado dado, en términos de el retorno esperado .
  - Asociado a una política $\pi(a \mid s)$.

Funciones de valor

- Valor de un estado:
  - Indica qué tan bueno es estar en un estado dado, en términos de el retorno esperado .
  - Asociado a una política $\pi(a \mid s)$.
  - Valor esperado del retorno, comenzando en s y siguiendo π de ahí en adelante:

Funciones de valor

- Valor de un estado:
  - Indica qué tan bueno es estar en un estado dado, en términos de el retorno esperado .
  - Asociado a una política $\pi(a \mid s)$.
  - Valor esperado del retorno, comenzando en s y siguiendo π de ahí en adelante:

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

Funciones de valor

- Valor de un estado:
  - Indica qué tan bueno es estar en un estado dado, en términos de el retorno esperado .
  - Asociado a una política $\pi(a \mid s)$.
  - Valor esperado del retorno, comenzando en s y siguiendo π de ahí en adelante:

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

es la función de valor de estado de la política π

Funciones de valor

- Valor de un estado:
  - Indica qué tan bueno es estar en un estado dado, en términos de el retorno esperado .
  - Asociado a una política $\pi(a \mid s)$.
  - Valor esperado del retorno, comenzando en s y siguiendo π de ahí en adelante:

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

    es la función de valor de estado de la política π

- Similarmente la función de valor de acción de la política π:

Funciones de valor

- Valor de un estado:
  - Indica qué tan bueno es estar en un estado dado, en términos de el retorno esperado .
  - Asociado a una política $\pi(a \mid s)$.
  - Valor esperado del retorno, comenzando en s y siguiendo π de ahí en adelante:

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

    es la función de valor de estado de la política π

- Similarmente la función de valor de acción de la política π:

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

Funciones de valor

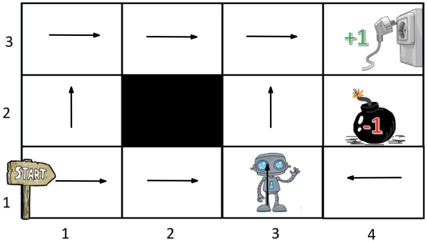
- Valor de un estado:
  - Indica qué tan bueno es estar en un estado dado, en términos de el retorno esperado .
  - Asociado a una política $\pi(a \mid s)$.
  - Valor esperado del retorno, comenzando en s y siguiendo π de ahí en adelante:

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

    es la función de valor de estado de la política π

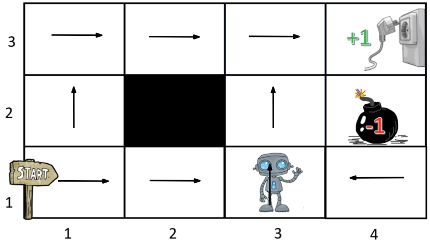
- Similarmente la función de valor de acción de la política π:

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



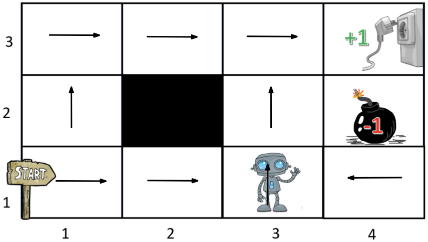
$$\gamma = 1$$

s	$v_\pi(s)$
(4,3)	
(3,3)	
(2,3)	
(1,1)	
(4,2)	



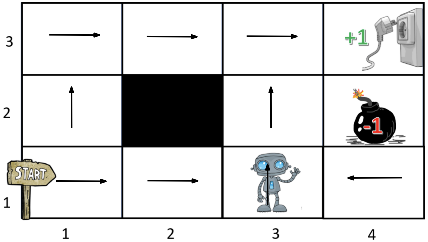
$$\gamma = 1$$

s	$v_\pi(s)$
(4,3)	1
(3,3)	
(2,3)	
(1,1)	
(4,2)	



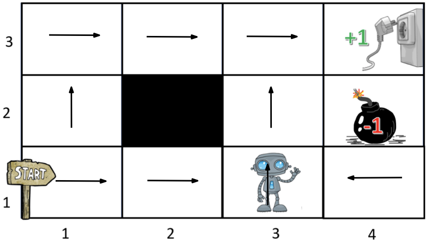
$$\gamma = 1$$

s	$v_\pi(s)$
(4,3)	1
(3,3)	1
(2,3)	
(1,1)	
(4,2)	



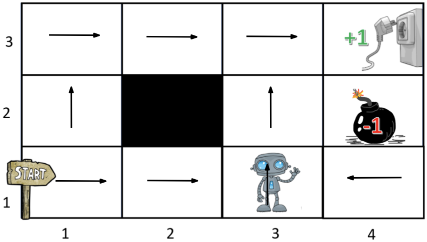
$$\gamma = 1$$

s	$v_\pi(s)$
(4,3)	1
(3,3)	1
(2,3)	
(1,1)	
(4,2)	



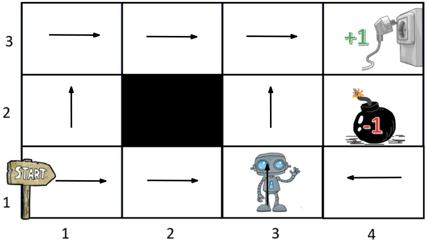
$$\gamma = 1$$

s	$v_\pi(s)$
(4,3)	1
(3,3)	1
(2,3)	1
(1,1)	1
(4,2)	



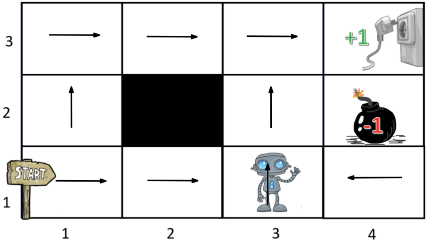
$$\gamma = 1$$

s	$v_\pi(s)$
(4,3)	1
(3,3)	1
(2,3)	1
(1,1)	1
(4,2)	-1



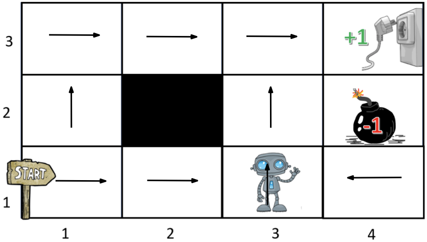
$$\gamma = 0{,}9$$

s	$v_\pi(s)$
(4,3)	
(3,3)	
(2,3)	
(1,1)	



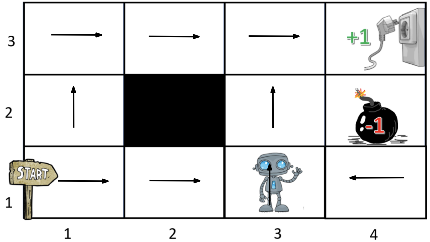
$$\gamma = 0{,}9$$

s	$v_\pi(s)$
(4,3)	1
(3,3)	
(2,3)	
(1,1)	



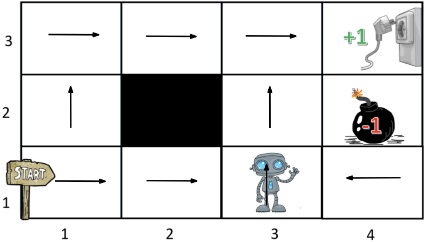
$$\gamma = 0{,}9$$

s	$v_\pi(s)$
(4,3)	1
(3,3)	0.9
(2,3)	
(1,1)	



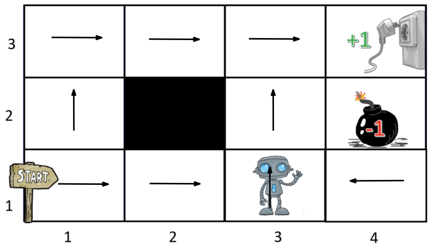
$$\gamma = 0{,}9$$

s	$v_\pi(s)$
(4,3)	1
(3,3)	0.9
(2,3)	0.81
(1,1)	



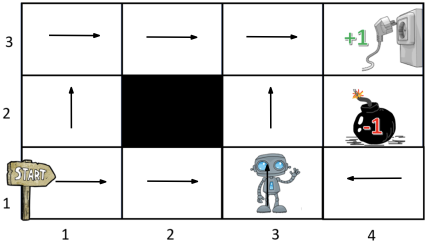
$$\gamma = 0{,}9$$

s	$v_\pi(s)$
(4,3)	1
(3,3)	0.9
(2,3)	0.81
(1,1)	0.59



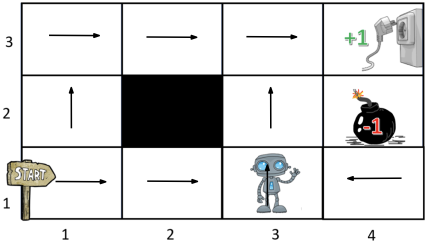
$$\gamma = 0{,}9$$

s	a	$q_\pi(s,a)$
(3,3)	r	
(3,3)	l	
(3,1)	l	



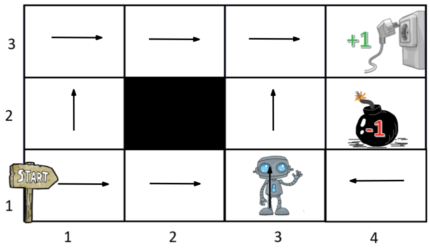
$$\gamma = 0{,}9$$

s	a	$q_\pi(s,a)$
(3,3)	r	0.9
(3,3)	l	
(3,1)	l	



$$\gamma = 0{,}9$$

s	a	$q_\pi(s,a)$
(3,3)	r	0.9
(3,3)	l	0.73
(3,1)	l	



$$\gamma = 0{,}9$$

s	a	$q_\pi(s,a)$
(3,3)	r	0.9
(3,3)	l	0.73
(3,1)	l	0.59

Ecuación de Bellman para v_π

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

Ecuación de Bellman para v_π

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s \right\} \\
&= \mathbb{E}_\pi \left\{ R_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \,\middle|\, S_t = s \right\}
\end{aligned}
$$

Ecuación de Bellman para v_π

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s \right\} \\
&= \mathbb{E}_\pi \left\{ R_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \,\middle|\, S_t = s \right\} \\
&= \sum_a \pi(a \mid s)
\end{aligned}
$$

Ecuación de Bellman para v_π

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s \right\} \\
&= \mathbb{E}_\pi \left\{ R_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \,\middle|\, S_t = s \right\} \\
&= \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s, a)
\end{aligned}
$$

Ecuación de Bellman para v_π

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s \right\} \\
&= \mathbb{E}_\pi \left\{ R_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \,\middle|\, S_t = s \right\} \\
&= \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \Big[r(s, a, s')
\end{aligned}
$$

Ecuación de Bellman para v_π

$$\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s \right\} \\
&= \mathbb{E}_\pi \left\{ R_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \,\middle|\, S_t = s \right\} \\
&= \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \left[r(s, a, s') + \gamma \mathbb{E}_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \,\middle|\, S_{t+1} = s' \right\} \right]
\end{aligned}$$

Ecuación de Bellman para v_π

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s \right\} \\
&= \mathbb{E}_\pi \left\{ R_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \,\middle|\, S_t = s \right\} \\
&= \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \left[r(s, a, s') + \gamma \mathbb{E}_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \,\middle|\, S_{t+1} = s' \right\} \right]
\end{aligned}
$$

Ecuación de Bellman para v_π

$$\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi\left\{\sum_{k=0}^{\infty}\gamma^k R_{t+k+1} \mid S_t = s\right\} \\
&= \mathbb{E}_\pi\left\{R_{t+1} + \gamma\sum_{k=0}^{\infty}\gamma^k R_{t+k+2} \,\middle|\, S_t = s\right\} \\
&= \sum_a \pi(a \mid s)\sum_{s'} p(s' \mid s, a)\left[r(s,a,s') + \gamma\mathbb{E}_\pi\left\{\sum_{k=0}^{\infty}\gamma^k R_{t+k+2} \,\middle|\, S_{t+1} = s'\right\}\right] \\
&= \sum_a \pi(a \mid s)\sum_{s'} p(s' \mid s, a)\left[r(s,a,s') + \gamma v_\pi(s')\right]
\end{aligned}$$

Ecuación de Bellman para v_π

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s \right\} \\
&= \mathbb{E}_\pi \left\{ R_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \,\middle|\, S_t = s \right\} \\
&= \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \left[r(s, a, s') + \gamma \mathbb{E}_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \,\middle|\, S_{t+1} = s' \right\} \right] \\
&= \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \left[r(s, a, s') + \gamma v_\pi(s') \right] \\
&= \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi(s') \right]
\end{aligned}
$$

Ecuación de Bellman para v_π

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s \right\} \\
&= \mathbb{E}_\pi \left\{ R_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \,\middle|\, S_t = s \right\} \\
&= \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \left[r(s, a, s') + \gamma \mathbb{E}_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \,\middle|\, S_{t+1} = s' \right\} \right] \\
&= \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \left[r(s, a, s') + \gamma v_\pi(s') \right] \\
&= \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi(s') \right]
\end{aligned}
$$

- Relación entre el valor de un estado y el valor de sus sucesores.

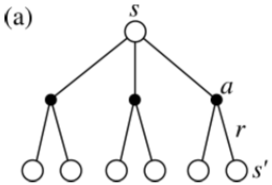
Ecuación de Bellman para v_π

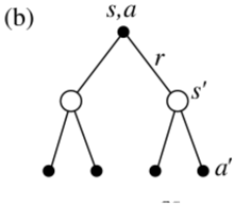
$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s \right\} \\
&= \mathbb{E}_\pi \left\{ R_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \,\middle|\, S_t = s \right\} \\
&= \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \left[r(s,a,s') + \gamma \mathbb{E}_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \,\middle|\, S_{t+1} = s' \right\} \right] \\
&= \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \left[r(s,a,s') + \gamma v_\pi(s') \right] \\
&= \sum_a \pi(a \mid s) \sum_{s'} \sum_r p(s', r \mid s, a) \left[r + \gamma v_\pi(s') \right]
\end{aligned}
$$

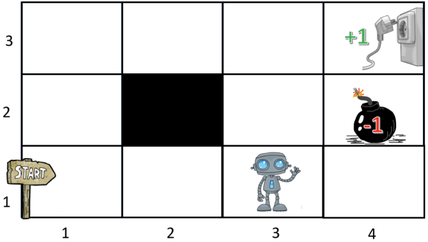
- Relación entre el valor de un estado y el valor de sus sucesores.
- Similarmente:

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

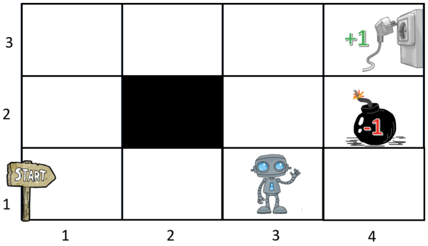
Diagramas de Backup





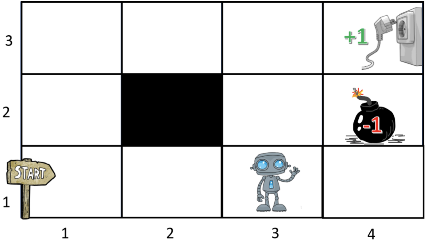


Acciones: u, d, l ,r. Política $\pi(a \mid s)$ aleatoria, $\gamma = 1$.



Acciones: u, d, l ,r. Política $\pi(a \mid s)$ aleatoria, $\gamma = 1$.

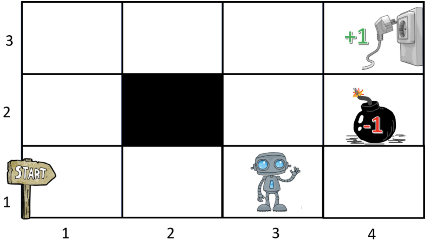
$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \left[r(s, a, s') + \gamma {\color{red} v_\pi(s')} \right]$$



Acciones: u, d, l ,r. Política $\pi(a \mid s)$ aleatoria, $\gamma = 1$.

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

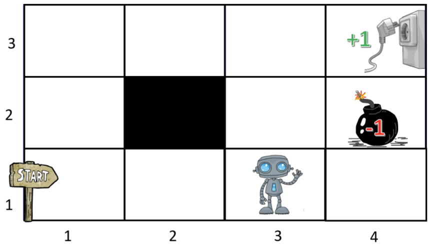
$$v_\pi(3, 1) =$$



Acciones: u, d, l ,r. Política $\pi(a \mid s)$ aleatoria, $\gamma = 1$.

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

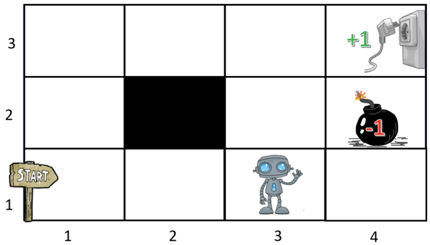
$$v_\pi(3, 1) = \frac{1}{3} v_\pi(2, 1) +$$



Acciones: u, d, l ,r. Política $\pi(a \mid s)$ aleatoria, $\gamma = 1$.

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

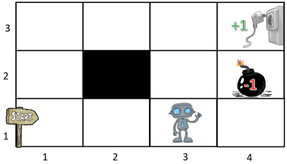
$$v_\pi(3,1) = \frac{1}{3} v_\pi(2,1) + \frac{1}{3} v_\pi(3,2) +$$



Acciones: u, d, l ,r. Política $\pi(a \mid s)$ aleatoria, $\gamma = 1$.

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

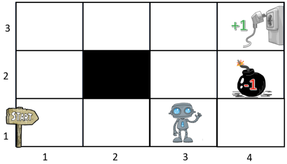
$$v_\pi(3,1) = \frac{1}{3} v_\pi(2,1) + \frac{1}{3} v_\pi(3,2) + \frac{1}{3} v_\pi(4,1)$$



Acciones: u, d, l ,r pero con probabilidad $\frac{1}{3}$ otra dirección aleatoria.

Política $\pi(a \mid s)$ aleatoria, $\gamma = 1$.

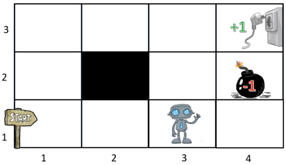
$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s, a) \left[r(s, a, s') + \gamma v_\pi(s') \right]$$



Acciones: u, d, l ,r pero con probabilidad $\frac{1}{3}$ otra dirección aleatoria.
Política $\pi(a \mid s)$ aleatoria, $\gamma = 1$.

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

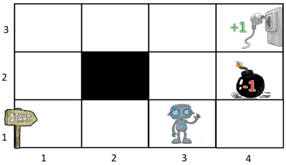
$$v_\pi(3, 1) =$$



Acciones: u, d, l ,r pero con probabilidad $\frac{1}{3}$ otra dirección aleatoria.
Política $\pi(a \mid s)$ aleatoria, $\gamma = 1$.

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

$$v_\pi(3,1) = \underbrace{\frac{1}{3}\left(\frac{2}{3} v_\pi(3,2) + \frac{1}{6} v_\pi(2,1) + \frac{1}{6} v_\pi(4,1) \right)}_{a=u}$$

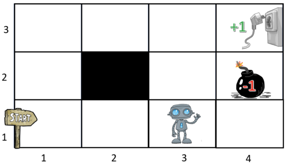


Acciones: u, d, l ,r pero con probabilidad $\frac{1}{3}$ otra dirección aleatoria.
Política $\pi(a \mid s)$ aleatoria, $\gamma = 1$.

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

$$v_\pi(3,1) = \underbrace{\frac{1}{3}\left(\frac{2}{3}v_\pi(3,2) + \frac{1}{6}v_\pi(2,1) + \frac{1}{6}v_\pi(4,1)\right)}_{a=u}$$

$$+ \underbrace{\frac{1}{3}\left(\frac{1}{6}v_\pi(3,2) + \frac{2}{3}v_\pi(2,1) + \frac{1}{6}v_\pi(4,1)\right)}_{a=l}$$

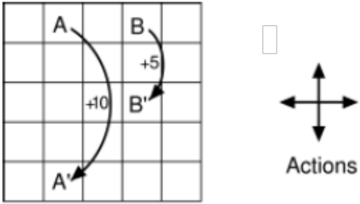


Acciones: u, d, l ,r pero con probabilidad $\frac{1}{3}$ otra dirección aleatoria.
Política $\pi(a \mid s)$ aleatoria, $\gamma = 1$.

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

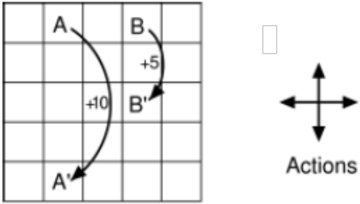
$$v_\pi(3,1) = \underbrace{\frac{1}{3}\left(\frac{2}{3}v_\pi(3,2) + \frac{1}{6}v_\pi(2,1) + \frac{1}{6}v_\pi(4,1)\right)}_{a=u}$$
$$+ \underbrace{\frac{1}{3}\left(\frac{1}{6}v_\pi(3,2) + \frac{2}{3}v_\pi(2,1) + \frac{1}{6}v_\pi(4,1)\right)}_{a=l}$$
$$+ \underbrace{\frac{1}{3}\left(\frac{1}{6}v_\pi(3,2) + \frac{1}{6}v_\pi(2,1) + \frac{2}{3}v_\pi(4,1)\right)}_{a=d}$$

Ejemplo: Gridworld



3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0

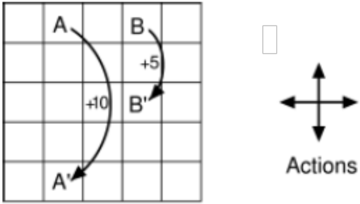
Ejemplo: Gridworld



3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0

- Intentar salirse del cuadro: recompensa -1.

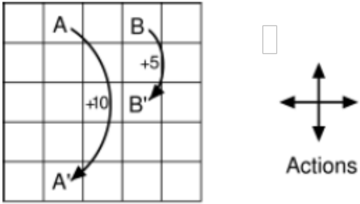
Ejemplo: Gridworld



3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0

- Intentar salirse del cuadro: recompensa -1.
- A: Recompensa +10, acciones llevan a A'.
- B: Recompensa +5, acciones llevan a B'.

Ejemplo: Gridworld



3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0

- Intentar salirse del cuadro: recompensa -1.
- A: Recompensa $+10$, acciones llevan a A'.
- B: Recompensa $+5$, acciones llevan a B'.
- $\pi(a \mid s)$ aleatoria, $\gamma = 0{,}9$

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.
- En MDP finitos existe por lo menos una política óptima.

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.
- En MDP finitos existe por lo menos una política óptima.
- 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.
- En MDP finitos existe por lo menos una política óptima.
- 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.
- En MDP finitos existe por lo menos una política óptima.
- 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

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

Ecuaciones de optimalidad de Bellman

$$
\begin{aligned}
v_*(s) &= \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s, a) \\
&= \max_{a} \mathbb{E}_{\pi_*} \{ G_t \mid S_t = s, A_t = a \}
\end{aligned}
$$

Ecuaciones de optimalidad de Bellman

$$
\begin{aligned}
v_*(s) &= \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s,a) \\
&= \max_{a} \mathbb{E}_{\pi_*} \{ G_t \mid S_t = s, A_t = a \}
\end{aligned}
$$

Ecuaciones de optimalidad de Bellman

$$
\begin{aligned}
v_*(s) &= \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s, a) \\
&= \max_a \mathbb{E}_{\pi_*} \{G_t \mid S_t = s, A_t = a\} \\
&= \max_a \mathbb{E}_{\pi_*} \left\{ R_{t+1} + \gamma G_{t+1} \,\middle|\, S_t = s, A_t = a \right\}
\end{aligned}
$$

Ecuaciones de optimalidad de Bellman

$$
\begin{aligned}
v_*(s) &= \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s,a) \\
&= \max_a \mathbb{E}_{\pi_*} \{G_t \mid S_t = s, A_t = a\} \\
&= \max_a \mathbb{E}_{\pi_*} \left\{ R_{t+1} + \gamma G_{t+1} \,\middle|\, S_t = s, A_t = a \right\} \\
&= \max_a \mathbb{E} \left[R_{t+1} + \gamma v_*(s_{t+1}) \mid S_t = s, A_t = a \right]
\end{aligned}
$$

Ecuaciones de optimalidad de Bellman

$$
\begin{aligned}
v_*(s) &= \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s, a) \\
&= \max_a \mathbb{E}_{\pi_*} \{G_t \mid S_t = s, A_t = a\} \\
&= \max_a \mathbb{E}_{\pi_*} \left\{ R_{t+1} + \gamma G_{t+1} \,\middle|\, S_t = s, A_t = a \right\} \\
&= \max_a \mathbb{E} \left[R_{t+1} + \gamma v_*(s_{t+1}) \mid S_t = s, A_t = a \right] \\
&= \max_a \sum_{s', r} p(s', r \mid s, a) \left[r + \gamma v_*(s') \right]
\end{aligned}
$$

Ecuaciones de optimalidad de Bellman

$$
\begin{aligned}
v_*(s) &= \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s,a) \\
&= \max_a \mathbb{E}_{\pi_*} \{G_t \mid S_t = s, A_t = a\} \\
&= \max_a \mathbb{E}_{\pi_*} \left\{ R_{t+1} + \gamma G_{t+1} \,\middle|\, S_t = s, A_t = a \right\} \\
&= \max_a \mathbb{E}\left[R_{t+1} + \gamma v_*(s_{t+1}) \mid S_t = s, A_t = a\right] \\
&= \max_a \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma v_*(s')\right]
\end{aligned}
$$

- Función de valor óptima sin referencia a una política específica.

- Similarmente, para q_*:

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

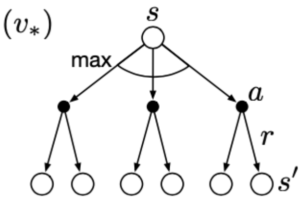
- Similarmente, para q_*:

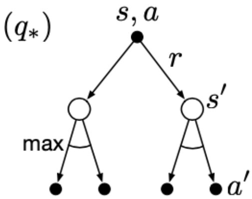
$$
\begin{aligned}
q_*(s,a) &= \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \mid S_t = s, A_t = a\right] \\
&= \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma \max_{a'} q_*(s', a')\right]
\end{aligned}
$$

- Similarmente, para q_*:

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

- Diagramas de backup:





- Si hay N estados, las ecuaciones de optimalidad de Bellman son un conjunto de N ecuaciones no lineales en N incógnitas.

- Si hay N estados, las ecuaciones de optimalidad de Bellman son un conjunto de N ecuaciones no lineales en N incógnitas.
- Para MPD finitos, estas ecuaciones tienen una solución única que es independiente de la política.

- Si hay N estados, las ecuaciones de optimalidad de Bellman son un conjunto de N ecuaciones no lineales en N incógnitas.
- Para MPD finitos, estas ecuaciones tienen una solución única que es independiente de la política.
- Si la dinámica del ambiente $p(s', r \mid s, a)$ es conocida, podemos en principio resolver las ecuaciones.

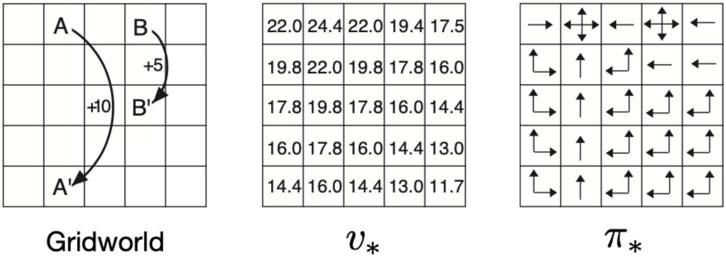
- Si hay N estados, las ecuaciones de optimalidad de Bellman son un conjunto de N ecuaciones no lineales en N incógnitas.
- Para MPD finitos, estas ecuaciones tienen una solución única que es independiente de la política.
- Si la dinámica del ambiente $p(s', r \mid s, a)$ es conocida, podemos en principio resolver las ecuaciones.
- A partir de v_* se puede determinar una política óptima

- Si hay N estados, las ecuaciones de optimalidad de Bellman son un conjunto de N ecuaciones no lineales en N incógnitas.
- Para MPD finitos, estas ecuaciones tienen una solución única que es independiente de la política.
- Si la dinámica del ambiente $p(s', r \mid s, a)$ es conocida, podemos en principio resolver las ecuaciones.
- A partir de v_* se puede determinar una política óptima : política greedy con respecto a v_*.

- Si hay N estados, las ecuaciones de optimalidad de Bellman son un conjunto de N ecuaciones no lineales en N incógnitas.
- Para MPD finitos, estas ecuaciones tienen una solución única que es independiente de la política.
- Si la dinámica del ambiente $p(s', r \mid s, a)$ es conocida, podemos en principio resolver las ecuaciones.
- A partir de v_* se puede determinar una política óptima : política greedy con respecto a v_*.
- En la práctica, no conocemos $p(s', r \mid s, a)$.

- Si hay N estados, las ecuaciones de optimalidad de Bellman son un conjunto de N ecuaciones no lineales en N incógnitas.
- Para MPD finitos, estas ecuaciones tienen una solución única que es independiente de la política.
- Si la dinámica del ambiente $p(s', r \mid s, a)$ es conocida, podemos en principio resolver las ecuaciones.
- A partir de v_* se puede determinar una política óptima : política greedy con respecto a v_*.
- En la práctica, no conocemos $p(s', r \mid s, a)$.
- Si conocemos q_* podemos conocer una política óptima sin conocer la dinámica del ambiente.

Ejemplo: Gridworld



22.0	24.4	22.0	19.4	17.5
19.8	22.0	19.8	17.8	16.0
17.8	19.8	17.8	16.0	14.4
16.0	17.8	16.0	14.4	13.0
14.4	16.0	14.4	13.0	11.7

Figure 3.5: Optimal solutions to the gridworld example.