

# Procesos de Decisión de Markov

Stalin Muñoz y David A. Rosenblueth

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS)  
Universidad Nacional Autónoma de México (UNAM)

# Contenido

## Contenido

- 1 Conceptos preliminares
- 2 Políticas de un agente
- 3 Estados con utilidad en Cadenas de Markov
- 4 Procesos de decisión de Markov
- 5 Algoritmo de iteración de valores

## Conceptos preliminares

- 1 Estado = conjunto de asignaciones de valores para un conjunto de variables. Por ejemplo:  
 $\{\text{lloviendo} = \text{cierto}, \text{puerta} = \text{abierta}, \text{pasajeros} = 237, \dots\}$

## Conceptos preliminares

- 1 Estado = conjunto de asignaciones de valores para un conjunto de variables. Por ejemplo:  
 $\{\text{lloviendo} = \text{cierto}, \text{puerta} = \text{abierta}, \text{pasajeros} = 237, \dots\}$
- 2 Utilidad = función que asigna a un estado un valor real que representa de manera cuantitativa las preferencias del agente.

## Conceptos preliminares

- 1 Estado = conjunto de asignaciones de valores para un conjunto de variables. Por ejemplo:  
 $\{\text{lloviendo} = \text{cierto}, \text{puerta} = \text{abierta}, \text{pasajeros} = 237, \dots\}$
- 2 Utilidad = función que asigna a un estado un valor real que representa de manera cuantitativa las preferencias del agente.
- 3 Lotería = un experimento aleatorio en el que una variable toma un valor particular en un espacio muestral.

## Conceptos preliminares

- 1 Estado = conjunto de asignaciones de valores para un conjunto de variables. Por ejemplo:  
 $\{\text{lloviendo} = \text{cierto}, \text{puerta} = \text{abierta}, \text{pasajeros} = 237, \dots\}$
- 2 Utilidad = función que asigna a un estado un valor real que representa de manera cuantitativa las preferencias del agente.
- 3 Lotería = un experimento aleatorio en el que una variable toma un valor particular en un espacio muestral.
- 4 Utilidad de una lotería = Valor esperado de utilidad para un experimento aleatorio.

## Conceptos preliminares

- 1 Estado = conjunto de asignaciones de valores para un conjunto de variables. Por ejemplo:  
 $\{\text{lloviendo} = \text{cierto}, \text{puerta} = \text{abierta}, \text{pasajeros} = 237, \dots\}$
- 2 Utilidad = función que asigna a un estado un valor real que representa de manera cuantitativa las preferencias del agente.
- 3 Lotería = un experimento aleatorio en el que una variable toma un valor particular en un espacio muestral.
- 4 Utilidad de una lotería = Valor esperado de utilidad para un experimento aleatorio.
- 5 Cadena de markov = Modelo probabilístico de la evolución temporal de la distribución de probabilidad para los estados de un sistema que tiene la propiedad de Markov.

## Conceptos preliminares

- 1 Estado = conjunto de asignaciones de valores para un conjunto de variables. Por ejemplo:  
 $\{\text{lloviendo} = \text{cierto}, \text{puerta} = \text{abierta}, \text{pasajeros} = 237, \dots\}$
- 2 Utilidad = función que asigna a un estado un valor real que representa de manera cuantitativa las preferencias del agente.
- 3 Lotería = un experimento aleatorio en el que una variable toma un valor particular en un espacio muestral.
- 4 Utilidad de una lotería = Valor esperado de utilidad para un experimento aleatorio.
- 5 Cadena de markov = Modelo probabilístico de la evolución temporal de la distribución de probabilidad para los estados de un sistema que tiene la propiedad de Markov.
- 6 Decision = Acción que toma un agente racional para maximizar su utilidad.



## Política óptima

- La **política** de un agente es una función  $\pi : S \rightarrow A$  que mapea los estados del problema en acciones del agente.

## Política óptima

- La **política** de un agente es una función  $\pi : S \rightarrow A$  que mapea los estados del problema en acciones del agente.
- La **política óptima** es aquella política que maximiza el valor esperado de utilidad.

## Política óptima

- La **política** de un agente es una función  $\pi : S \rightarrow A$  que mapea los estados del problema en acciones del agente.
- La **política óptima** es aquella política que maximiza el valor esperado de utilidad.
- En un problema de decisiones secuenciales, o iteradas, existen diferentes formas de definir la política óptima:

## Política óptima

- La **política** de un agente es una función  $\pi : S \rightarrow A$  que mapea los estados del problema en acciones del agente.
- La **política óptima** es aquella política que maximiza el valor esperado de utilidad.
- En un problema de decisiones secuenciales, o iteradas, existen diferentes formas de definir la política óptima:
  - **miope**: solo le interesa maximizar la utilidad esperada del tiempo siguiente,

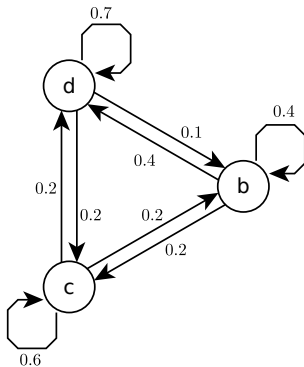
## Política óptima

- La **política** de un agente es una función  $\pi : S \rightarrow A$  que mapea los estados del problema en acciones del agente.
- La **política óptima** es aquella política que maximiza el valor esperado de utilidad.
- En un problema de decisiones secuenciales, o iteradas, existen diferentes formas de definir la política óptima:
  - **miope**: solo le interesa maximizar la utilidad esperada del tiempo siguiente,
  - **horizonte finito**: le interesa maximizar la utilidad en una ventana de tiempo definido, y

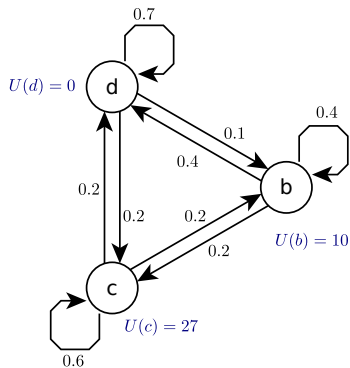
## Política óptima

- La **política** de un agente es una función  $\pi : S \rightarrow A$  que mapea los estados del problema en acciones del agente.
- La **política óptima** es aquella política que maximiza el valor esperado de utilidad.
- En un problema de decisiones secuenciales, o iteradas, existen diferentes formas de definir la política óptima:
  - **miope**: solo le interesa maximizar la utilidad esperada del tiempo siguiente,
  - **horizonte finito**: le interesa maximizar la utilidad en una ventana de tiempo definido, y
  - **horizonte infinito**: se maximiza la utilidad para todo tiempo futuro.

## Utilidad esperada en cadenas de Markov

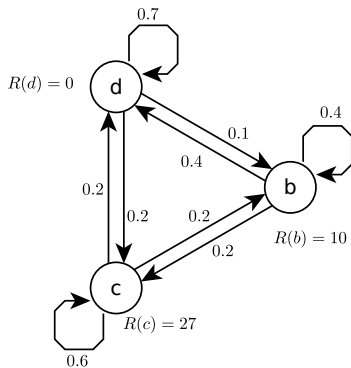


## Utilidad esperada en cadenas de Markov

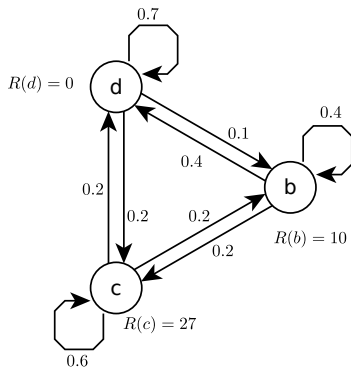




## Utilidad esperada en cadenas de Markov

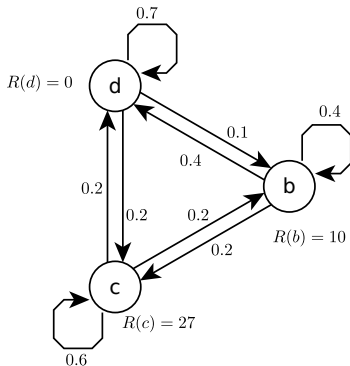


## Utilidad esperada en cadenas de Markov



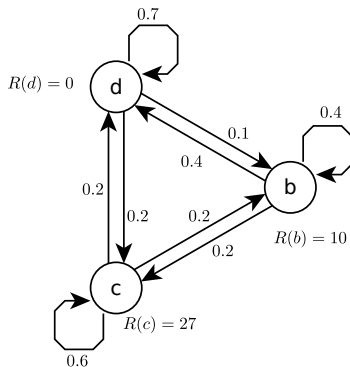
- Nos interesa conocer la utilidad de cada estado pero considerando también tiempos futuros.

## Utilidad esperada en cadenas de Markov



- Nos interesa conocer la utilidad de cada estado pero considerando también tiempos futuros.
- Vamos a utilizar un factor de descuento  $\gamma \in (0, 1]$  para ponderar la utilidad obtenida en tiempos futuros.

## Utilidad esperada en cadenas de Markov



Podemos tratar cada nodo del grafo como una lotería. Para el estado  $s$  tenemos que el valor esperado de utilidad es:

$$\sum_{s'} P(S^{t+1}=s' \mid S^t=s) V(s')$$

## Utilidad esperada en cadenas de Markov

- La utilidad *horizonte infinito* considera la recompensa instantanea y el valor esperado de recompensa para tiempos futuros con su factor de descuento:

$$V(s) = R(s) + \gamma \sum_{s'} P(S^{t+1}=s' \mid S^{t+1}=s) V(s')$$

## Utilidad esperada en cadenas de Markov

- La utilidad *horizonte infinito* considera la recompensa instantánea y el valor esperado de recompensa para tiempos futuros con su factor de descuento:

$$V(s) = R(s) + \gamma \sum_{s'} P(S^{t+1}=s' \mid S^{t+1}=s) V(s')$$

- Matricialmente:

$$V(s) = R(s) + \gamma T^T V(s)$$

## Utilidad esperada en cadenas de Markov

- La utilidad *horizonte infinito* considera la recompensa instantanea y el valor esperado de recompensa para tiempos futuros con su factor de descuento:

$$V(s) = R(s) + \gamma \sum_{s'} P(S^{t+1}=s' \mid S^t=s) V(s')$$

- Matricialmente:

$$V(s) = R(s) + \gamma T^T V(s)$$

- El cual podemos resolver:

$$V(s) = (I - \gamma T^T)^{-1} R(s)$$

## Incorporando las acciones del agente

- En los **procesos de decisión de Markov** las acciones del agente cambian las probabilidades de transición de estado.



## Incorporando las acciones del agente

- En los **procesos de decisión de Markov** las acciones del agente cambian las probabilidades de transición de estado.
- Asumiremos que el agente es **racional** y las acciones son aquellas que maximizan la utilidad horizonte infinito.

## Incorporando las acciones del agente

- En los **procesos de decisión de Markov** las acciones del agente cambian las probabilidades de transición de estado.
- Asumiremos que el agente es **racional** y las acciones son aquellas que maximizan la utilidad horizonte infinito.
- El proceso de decisión de Markov se modela con la **ecuación de Bellman**:

$$V(s) = R(s) + \gamma \max_a \sum_{s'} P(s' | s, a) V(s')$$

# Un grafo de decisiones

d

b

c

# Un grafo de decisiones

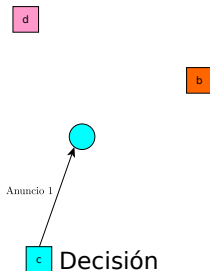
d

b

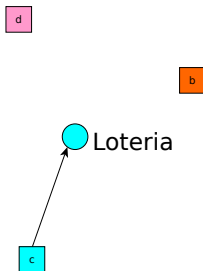
c

Decisión

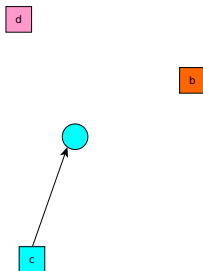
## Un grafo de decisiones



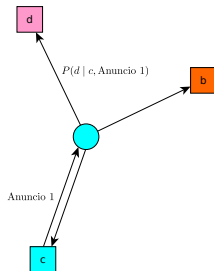
## Un grafo de decisiones



## Un grafo de decisiones

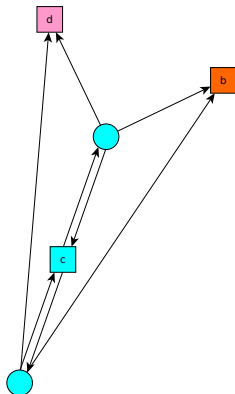


## Un grafo de decisiones

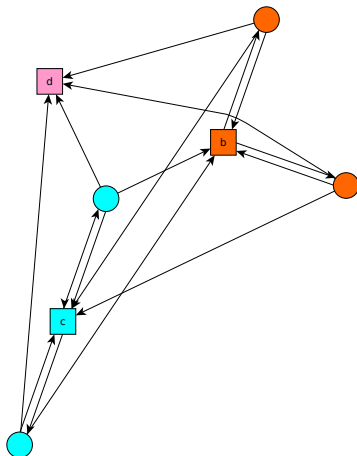




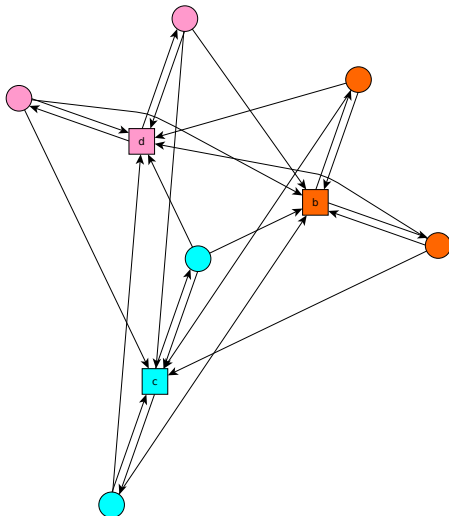
## Un grafo de decisiones



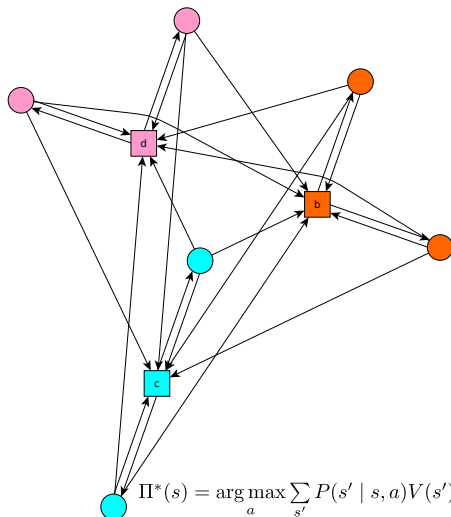
## Un grafo de decisiones



## Un grafo de decisiones



## Política óptima



## Política óptima: ¿cómo encontrar $\Pi^*(s)$ ?

- Deseamos resolver  $\Pi^* : S \rightarrow A$ :

$$\Pi^*(s) = \operatorname{argmax}_a \sum_{s'} P(s' \mid s, a) V(s')$$

## Política óptima: ¿cómo encontrar $\Pi^*(s)$ ?

- Deseamos resolver  $\Pi^* : S \rightarrow A$ :

$$\Pi^*(s) = \operatorname{argmax}_a \sum_{s'} P(s' \mid s, a) V(s')$$

- Algunos algoritmos para encontrar  $\Pi^*(s)$  son:

## Política óptima: ¿cómo encontrar $\Pi^*(s)$ ?

- Deseamos resolver  $\Pi^* : S \rightarrow A$ :

$$\Pi^*(s) = \operatorname{argmax}_a \sum_{s'} P(s' \mid s, a) V(s')$$

- Algunos algoritmos para encontrar  $\Pi^*(s)$  son:
  - **Iteración de valores**: Primero resolvemos  $V(s)$  en la ecuación de Bellman. Usamos  $V(s)$  para encontrar  $\Pi^*(s)$ .

## Política óptima: ¿cómo encontrar $\Pi^*(s)$ ?

- Deseamos resolver  $\Pi^* : S \rightarrow A$ :

$$\Pi^*(s) = \operatorname{argmax}_a \sum_{s'} P(s' \mid s, a) V(s')$$

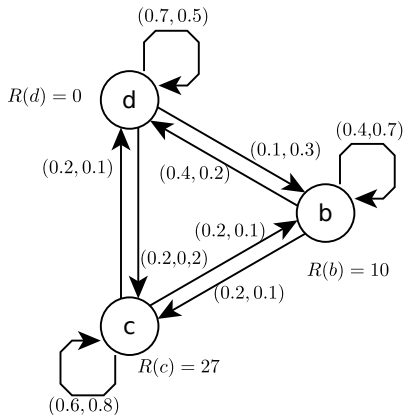
- Algunos algoritmos para encontrar  $\Pi^*(s)$  son:
  - **Iteración de valores**: Primero resolvemos  $V(s)$  en la ecuación de Bellman. Usamos  $V(s)$  para encontrar  $\Pi^*(s)$ .
  - **Iteración de políticas**: Se fija la política y se resuelve un sistema de ecuaciones para encontrar  $V(s)$ , repitiendo hasta que la política no cambie.



# Algoritmo de iteración de políticas

Probabilidades de transición parametrizadas por una política

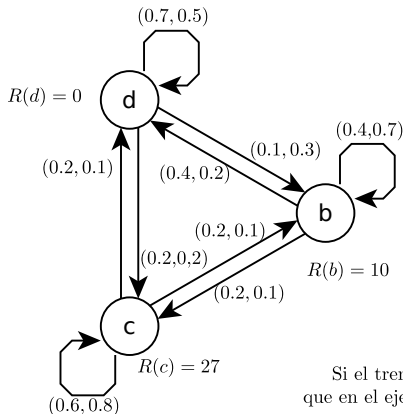
$$T(\pi) = \begin{bmatrix} f_{1,1}(\pi) & f_{1,2}(\pi) & f_{1,3}(\pi) \\ f_{2,1}(\pi) & f_{2,2}(\pi) & f_{2,3}(\pi) \\ f_{3,1}(\pi) & f_{3,2}(\pi) & f_{3,3}(\pi) \end{bmatrix}$$





# Algoritmo de iteración de políticas

Probabilidades de transición parametrizadas por una política



$$T(\pi) = \begin{bmatrix} f_{1,1}(\pi) & f_{1,2}(\pi) & f_{1,3}(\pi) \\ f_{2,1}(\pi) & f_{2,2}(\pi) & f_{2,3}(\pi) \\ f_{3,1}(\pi) & f_{3,2}(\pi) & f_{3,3}(\pi) \end{bmatrix}$$

$$\pi_0 = \begin{bmatrix} \neg \text{Anuncio} \\ \neg \text{Anuncio} \\ \neg \text{Anuncio} \end{bmatrix}$$

$$T = \begin{bmatrix} (0.7, 0.5) & (0.4, 0.2) & (0.2, 0.1) \\ (0.1, 0.3) & (0.4, 0.7) & (0.2, 0.1) \\ (0.2, 0.2) & (0.2, 0.1) & (0.6, 0.8) \end{bmatrix}$$

$$T(\pi_0) = \begin{bmatrix} 0.7 & 0.4 & 0.2 \\ 0.1 & 0.4 & 0.2 \\ 0.2 & 0.2 & 0.6 \end{bmatrix}$$

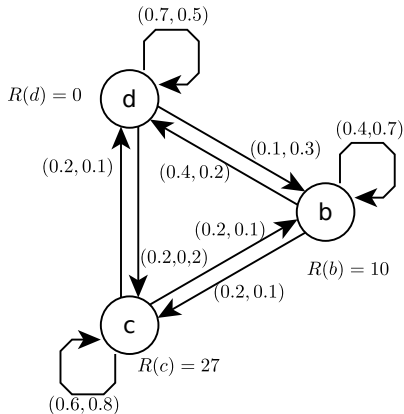
Si el tren no hace anuncios, tenemos el mismo proceso estocástico que en el ejemplo visto en el módulo de Cadenas de Markov

## Algoritmo de iteración de políticas

Probabilidades de transición parametrizadas por una política

$$\pi_0 = \begin{bmatrix} \neg \text{Anuncio} \\ \neg \text{Anuncio} \\ \neg \text{Anuncio} \end{bmatrix} \quad T(\pi_0) = \begin{bmatrix} 0.7 & 0.2 & 0.2 \\ 0.1 & 0.7 & 0.2 \\ 0.2 & 0.1 & 0.6 \end{bmatrix}$$

Resolvemos para encontrar  $V_0(s)$



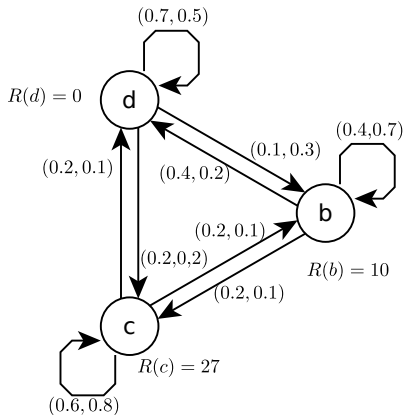
## Algoritmo de iteración de políticas

Probabilidades de transición parametrizadas por una política

$$\pi_0 = \begin{bmatrix} \neg \text{Anuncio} \\ \neg \text{Anuncio} \\ \neg \text{Anuncio} \end{bmatrix} \quad T(\pi_0) = \begin{bmatrix} 0.7 & 0.2 & 0.2 \\ 0.1 & 0.7 & 0.2 \\ 0.2 & 0.1 & 0.6 \end{bmatrix}$$

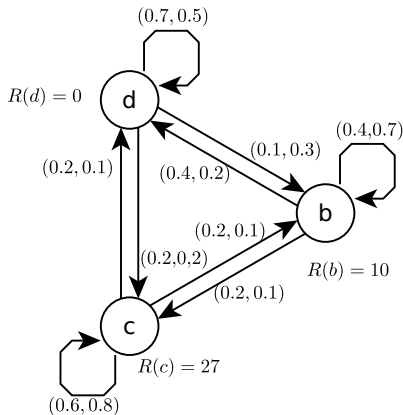
Resolvemos para encontrar  $V_0(s)$

$$V_0(s) = (I - \gamma T(\pi_0)^T)^{-1} R(s)$$



## Algoritmo de iteración de políticas

Probabilidades de transición parametrizadas por una política



$$\pi_0 = \begin{bmatrix} \neg \text{Anuncio} \\ \neg \text{Anuncio} \\ \neg \text{Anuncio} \end{bmatrix} \quad T(\pi_0) = \begin{bmatrix} 0.7 & 0.2 & 0.2 \\ 0.1 & 0.7 & 0.2 \\ 0.2 & 0.1 & 0.6 \end{bmatrix}$$

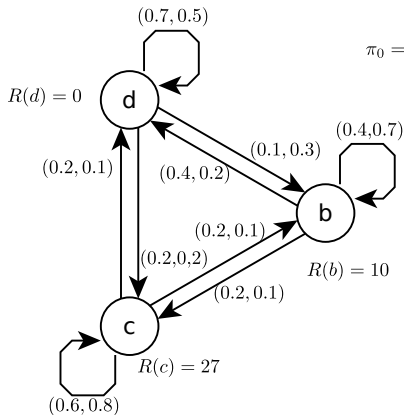
Resolvemos para encontrar  $V_0(s)$

$$V_0(s) = (I - \gamma T(\pi_0)^T)^{-1} R(s)$$

$$V_0(s) = \left( \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - 0.9 \begin{bmatrix} 0.7 & 0.1 & 0.2 \\ 0.2 & 0.7 & 0.1 \\ 0.2 & 0.2 & 0.6 \end{bmatrix} \right)^{-1} \begin{bmatrix} 0 \\ 10 \\ 27 \end{bmatrix}$$



# Algoritmo de iteración de políticas



Probabilidades de transición parametrizadas por una política

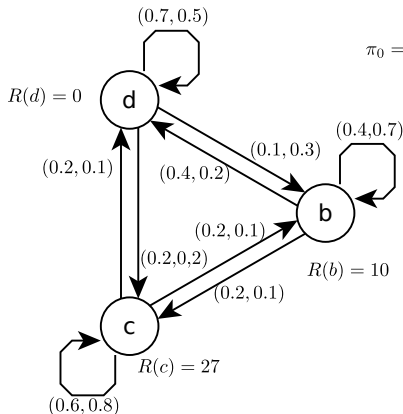
$$\pi_0 = \begin{bmatrix} \neg \text{Anuncio} \\ \neg \text{Anuncio} \\ \neg \text{Anuncio} \end{bmatrix} \quad T(\pi_0) = \begin{bmatrix} 0.7 & 0.2 & 0.2 \\ 0.1 & 0.7 & 0.2 \\ 0.2 & 0.1 & 0.6 \end{bmatrix} \quad V_0(s) = \begin{bmatrix} 91.07 \\ 104.20 \\ 135.10 \end{bmatrix}$$

Encontramos la política  $\pi_1$  para la iteración siguiente

$$\pi_1(s) = \operatorname{argmax}_a P(s' | s, a) V_0(s')$$



# Algoritmo de iteración de políticas



Probabilidades de transición parametrizadas por una política

$$\pi_0 = \begin{bmatrix} \neg \text{Anuncio} \\ \neg \text{Anuncio} \\ \neg \text{Anuncio} \end{bmatrix} \quad T(\pi_0) = \begin{bmatrix} 0.7 & 0.2 & 0.2 \\ 0.1 & 0.7 & 0.2 \\ 0.2 & 0.1 & 0.6 \end{bmatrix} \quad V_0(s) = \begin{bmatrix} 91.07 \\ 104.20 \\ 135.10 \end{bmatrix}$$

Encontramos la política  $\pi_1$  para la iteración siguiente

$$\pi_1(s) = \operatorname{argmax}_a P(s' | s, a) V_0(s')$$

$$\pi_1(s=d) = \operatorname{argmax}_a \{$$

$$P(s'=d | s=d, a=\neg \text{Anuncio}) V_0(s'=d) +$$

$$P(s'=b | s=d, a=\neg \text{Anuncio}) V_0(s'=b) +$$

$$P(s'=c | s=d, a=\neg \text{Anuncio}) V_0(s'=c),$$

$$P(s'=d | s=d, a=\text{Anuncio}) V_0(s'=d) +$$

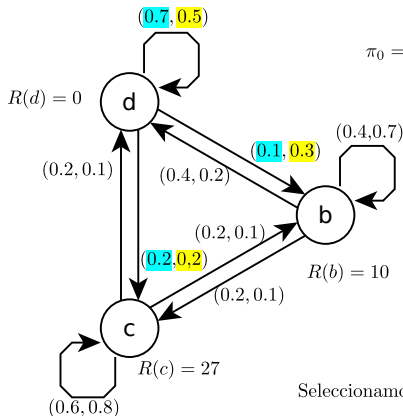
$$P(s'=b | s=d, a=\text{Anuncio}) V_0(s'=b) +$$

$$P(s'=c | s=d, a=\text{Anuncio}) V_0(s'=c) \}$$





## Algoritmo de iteración de políticas



Probabilidades de transición parametrizadas por una política

$$\pi_0 = \begin{bmatrix} \neg \text{Anuncio} \\ \neg \text{Anuncio} \\ \neg \text{Anuncio} \end{bmatrix} \quad T(\pi_0) = \begin{bmatrix} 0.7 & 0.2 & 0.2 \\ 0.1 & 0.7 & 0.2 \\ 0.2 & 0.1 & 0.6 \end{bmatrix} \quad V_0(s) = \begin{bmatrix} 91.07 \\ 104.19 \\ 135.10 \end{bmatrix}$$

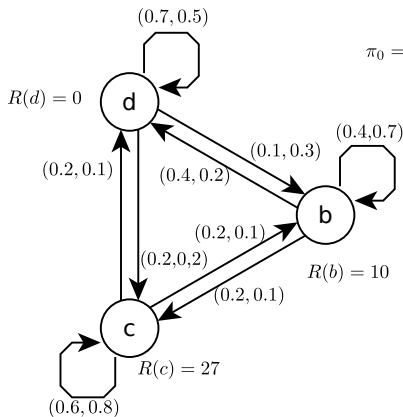
Encontramos la política  $\pi_1$  para la iteración siguiente

$$\pi_1(s) = \operatorname{argmax}_a P(s' | s, a) V_0(s')$$

$$\pi_1(s=d) = \operatorname{argmax}_a \{ 101.19, 103.81 \}$$

Seleccionamos la acción Anuncio que corresponde con el máximo

# Algoritmo de iteración de políticas



Probabilidades de transición parametrizadas por una política

$$\pi_0 = \begin{bmatrix} \neg \text{Anuncio} \\ \neg \text{Anuncio} \\ \neg \text{Anuncio} \end{bmatrix} \quad T(\pi_0) = \begin{bmatrix} 0.7 & 0.2 & 0.2 \\ 0.1 & 0.7 & 0.2 \\ 0.2 & 0.1 & 0.6 \end{bmatrix} \quad V_0(s) = \begin{bmatrix} 91.07 \\ 104.20 \\ 135.10 \end{bmatrix}$$

Encontramos la política  $\pi_1$  para la iteración siguiente

$$\pi_1 = \begin{bmatrix} \text{Anuncio} \\ \neg \text{Anuncio} \\ \text{Anuncio} \end{bmatrix}$$

$$\pi_0 \neq \pi_1$$

$$T(\pi_1) = \begin{bmatrix} 0.5 & 0.4 & 0.1 \\ 0.3 & 0.4 & 0.1 \\ 0.2 & 0.2 & 0.8 \end{bmatrix} \quad V_1(s) = \begin{bmatrix} 127.58 \\ 138.57 \\ 181.98 \end{bmatrix}$$

# Algoritmo de iteración de políticas

Probabilidades de transición parametrizadas por una política

$$\pi_0 = \begin{bmatrix} \neg \text{Anuncio} \\ \neg \text{Anuncio} \\ \neg \text{Anuncio} \end{bmatrix} \quad \pi_1 = \begin{bmatrix} \text{Anuncio} \\ \neg \text{Anuncio} \\ \text{Anuncio} \end{bmatrix} \quad V_1(s) = \begin{bmatrix} 127.58 \\ 138.57 \\ 181.97 \end{bmatrix}$$

Encontramos la política  $\pi_2$  para la iteración siguiente

$$\pi_2 = \begin{bmatrix} \text{Anuncio} \\ \neg \text{Anuncio} \\ \text{Anuncio} \end{bmatrix}$$

