

# 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

### Contenido

Hoy presentaremos un formalismo matemático para toma de decisiones en entornos con incertidumbre.

Los procesos de decisión de Markov.

- Conceptos preliminares
- Políticas de un agente
- Estados con utilidad en Cadenas de Markov
- Procesos de decisión de Markov
- Algoritmo de iteración de valores

## Conceptos preliminares

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

Un estado describe las variables relevantes del problema

## 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.

Asumiremos que existe una función de utilidad definida para todo estado del problema.

Esta función asigna un número real a los estados del problema. Entre mayor sea la preferencia del agente, mayor es el número asignado por la función de utilidad.

## 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.

Nuestra lotería representado un evento probabilístico.

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

La utilidad de la lotería entendida como el valor esperado de utilidad de la misma.

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

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

Las cadenas de Markov nos servirán como base para definir los procesos de decisión 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.

## Conceptos preliminares

En lo que trataremos asumimos que el agente actúa en forma racional.

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

Nos interesa encontrar una política para nuestro agente artificial.

La política de un agente puede entenderse formalmente como una función que le indica al agente que acción tomar en cada situación.

## 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.

El agente racional tomará las decisiones que más le convienen.

- 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

Pero, ¿qué es lo que más le conviene al agente?  
Aunque esto puede depender mucho del problema, podemos decir que hay maneras estándar de tratar la optimalidad para la utilidad esperada de un agente en un entorno de decisiones secuenciales.

- 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

Una política miope se limitará a tomar la acción que consigue el mayor beneficio inmediato.

Esta estrategia puede ser insuficiente en entornos complejos.

- 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

Una política de horizonte finito tratará de maximizar la suma de las utilidades para una ventana de tiempo en el futuro.

Por ejemplo, si el agente toma la acción que maximiza la utilidad para los próximos 7 días y no solo maximiza la utilidad del día de mañana.

- 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

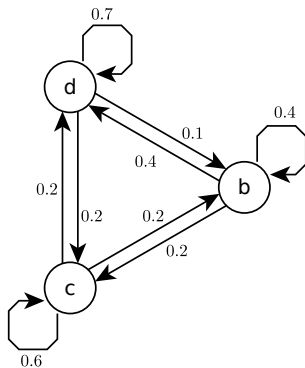
- La **política** de un agente es una función  $\pi : S \rightarrow A$  que mapea los estados del problema en acciones del agente.
- La **política óptima** es aquella política que maximiza el valor esperado de utilidad.
- En un problema de decisiones secuenciales, o iteradas, existen diferentes formas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

## Política óptima

La política horizonte infinito considera todos los tiempos futuros. Esta política es la más conveniente para el análisis matemático. Y es la que vamos a utilizar.

- 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



## Procesos de Decisión de Markov

### └ Estados con utilidad en Cadenas de Markov

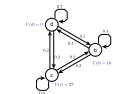
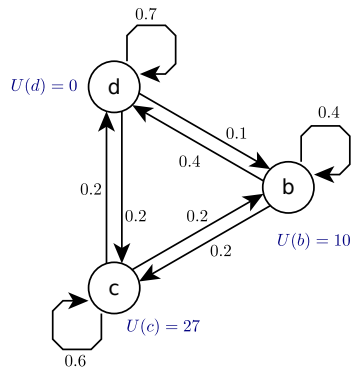
#### └ Utilidad esperada en cadenas de Markov



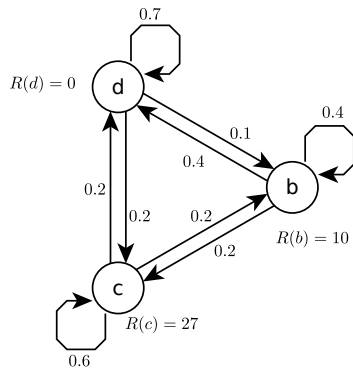
Utilizamos cadenas de Markov para modelar procesos secuenciales estocásticos.

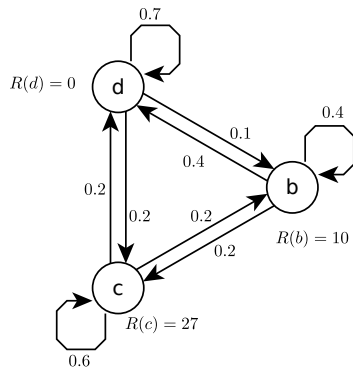
Observamos que la probabilidad de llegar a un estado al tiempo siguiente solo depende del estado en el que nos encontramos en el tiempo actual.

Representamos la cadena de Markov con un grafo de transición de estados.





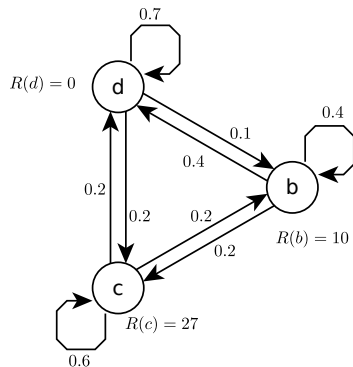




- Nos interesa conocer la utilidad de cada estado pero considerando también tiempos futuros.

Considerando que nos interesa encontrar una política horizonte infinito, plantearemos una utilidad que integra el valor esperado de utilidad para todos los estados futuros.

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



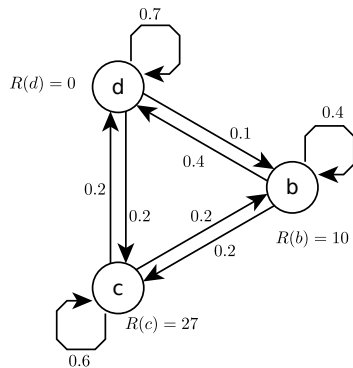
- Nos interesa conocer la utilidad de cada estado pero considerando también tiempos futuros.
- Vamos a utilizar un factor de descuento  $\gamma \in (0, 1]$  para ponderar la utilidad obtenida en tiempos futuros.

## Utilidad esperada en cadenas de Markov

Una recompensa vale más hoy que mañana, y vale más mañana que pasado mañana.

Aún no tomamos en cuenta las acciones del agente, pero esto lo haremos más adelante.

- 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.



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

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

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

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

Recordemos que para una loteria podemos calcular su utilidad sumando los productos de las utilidades de las salidas con su probabilidad respectiva.

Aquí podemos plantear la utilidad de un estado de la misma manera.

Para obtener la utilidad de un estado  $s$  sumamos los productos de la utilidad de cada estado  $s'$  al que podemos llegar desde  $s$  con su respectiva probabilidad.

Aquí el valor de utilidad  $V(s)$  no será la recompensa instantanea sino la utilidad horizonte infinito.

## 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 expresión para el valor horizonte infinito de utilidad nos lleva a una ecuación donde incorporamos tanto la recompensa instantanea como el valor esperado de utilidad para tiempos futuros.

Aplicamos el factor de descuento  $\gamma$  para las recompensas esperadas en tiempos futuros.

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

## 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:

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

- Matricialmente:

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

## Utilidad esperada en cadenas de Markov

Observamos que la expresión del término que corresponde con la utilidad de estados futuros puede expresarse como un producto matricial.

El producto de la transpuesta de la matriz de probabilidades de transición  $T$  con el vector de utilidades horizonte infinito  $V(s)$

Aquí la matriz  $T$  es la matriz de probabilidades de transición de una cadena de Markov.

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

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

- 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:

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

## 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:

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

Esta ecuación lineal puede resolverse invirtiendo una matriz.

## 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

Ahora vamos a incorporar las acciones del agente en la ecuación.

Tenemos un proceso estocástico donde el agente no puede anticipar con certidumbre el efecto de sus acciones.

A pesar de esto, las acciones sí tienen un efecto probabilístico.

Es decir, las probabilidades de transición de estado no sólo dependerán del estado actual, sino también de la acción que se toma.

De no ser así, cualquier política sería igual de buena y no habría problema que resolver.



## 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 el proceso de decisión de Markov nuestro agente es racional.  
Es decir, desea maximizar su recompensa horizonte infinito.

- 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**:

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

## Incorporando las acciones del agente

El proceso de decisión de Markov queda descrito por la ecuación de Bellman.

Observamos que aparece un término no lineal  $\max$  el cual modela el efecto de la acción racional del agente, aquí representada con la letra  $a$ .

Adicionalmente se observa que la probabilidad de transición no solo está condicionada al estado actual, sino también a la acción 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



En teoría de decisiones encadenamos nodos de decisión con loterías formando árboles de decisión.

Los procesos de decisión de Markov pueden interpretarse de manera similar, pero dado que resultarían árboles infinitos con probabilidades de transición estacionarias es mejor tratarlos como grafos de decisión.

Aquí comenzamos con los estados los que representaremos con nodos de decisión.

d

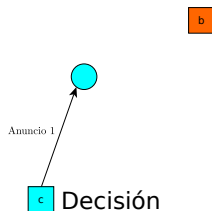
b

**c** Decisión

- Un grafo de decisiones

Por ejemplo, un pasajero se encuentra en el vagón comedor. Nuestro agente, es decir el tren inteligente, puede decidir entre un conjunto de acciones para tratar de cambiar la probabilidad de que el pasajero continúe o transite a otro vagón.

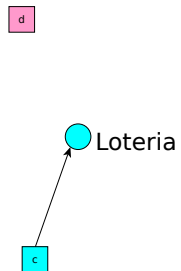
d



- └ Un grafo de decisiones

Una acción posible es hacer un anuncio publicitario en el vagón. Esto no definirá la acción del pasajero, pero si tendrá un efecto en la probabilidad de transición de estado.

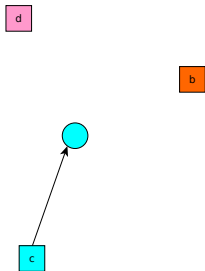




La transición la representamos con un nodo de lotería.

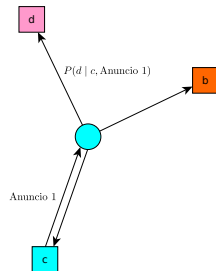


## Un grafo de decisiones



La lotería definirá el estado siguiente.

## Un grafo de decisiones



- Un grafo de decisiones

En este caso podemos transitar a cualquiera de los vagones: quedarnos en el comedor, ir al dormitorio o al bar.

Cada arista de la lotería tiene anotada la probabilidad de transición.

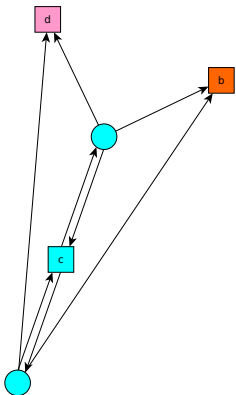
Para no saturar la figura solo ilustramos la probabilidad de transitar al dormitorio.

Observamos que esta probabilidad esta condicionada a encontrarnos en el vagón comedor y que el tren haga el anuncio 1.



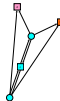


## Un grafo de decisiones

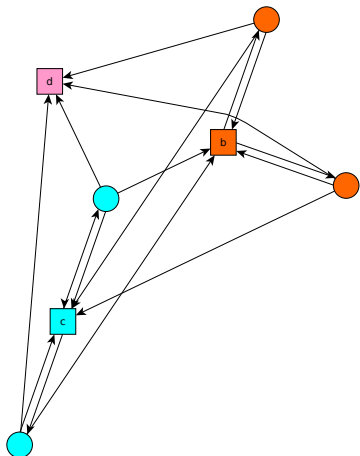


## Un grafo de decisiones

También es posible que el tren no haga anuncio alguno.  
En este caso tenemos otra lotería para transitar de estado.



## Un grafo de decisiones

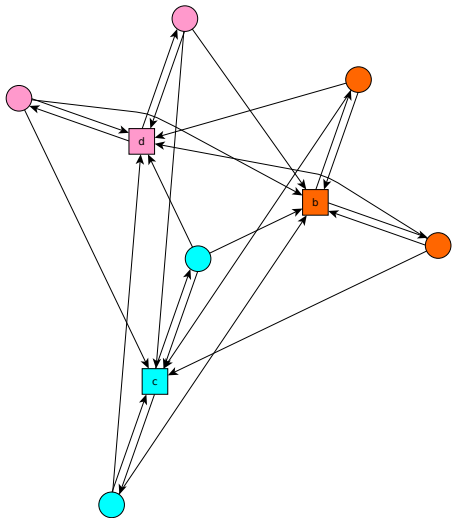


## Un grafo de decisiones

Así para el estado  $b$  tenemos acciones similares y las loterías respectivas.



## Un grafo de decisiones



## Procesos de Decisión de Markov

- └ Procesos de decisión de Markov

- └ Un grafo de decisiones

Un grafo de decisiones

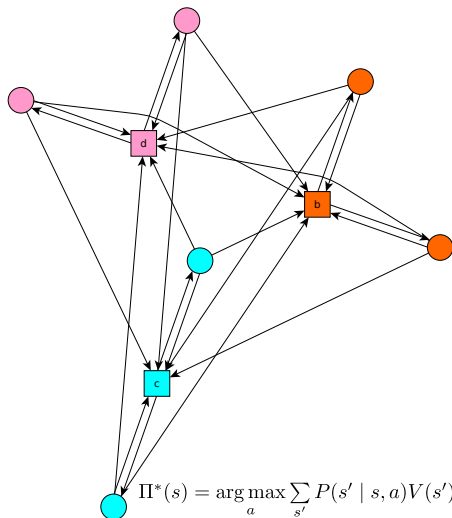


Y para el estado  $d$ .

La pregunta relevante que hacer es ¿cuál es la política óptima?

Es decir, ¿que acción debe tomar el tren en cada vagón, tal que maximice la utilidad horizonte infinito?

## Política óptima



$$\Pi^*(s) = \arg \max_a \sum_{s'} P(s' | s, a) V(s')$$

## Política óptima



De la ecuación de Bellman podemos ver que la política óptima  $\Pi^*(s)$  se obtiene con la acción que maximiza el valor esperado de utilidad para cada estado.

## 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 $\Pi^*(s)$ ?

El problema de diseño del agente racional consiste en encontrar la política óptima  $\Pi^*$ . Dada un estado  $s$  que acción debe tomar el agente.

## 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:

## Procesos de Decisión de Markov └ Procesos de decisión de Markov

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

Existen varias formas de resolver la política óptima.

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

- Deseamos resolver  $\Pi^* : S \rightarrow A$ :  
$$\Pi^*(s) = \operatorname{argmax}_a \sum_{s'} P(s' \mid s, a) V(s')$$
- 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  $\Pi^*(s)$ .

## Procesos de Decisión de Markov └ Procesos de decisión de Markov

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

El algoritmo de iteración de valores encuentra primero  $V(s)$  iterando la ecuación de Bellman hasta que los valores convergen.

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  $\Pi^* : S \rightarrow A$ :

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

- 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.

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

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

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

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

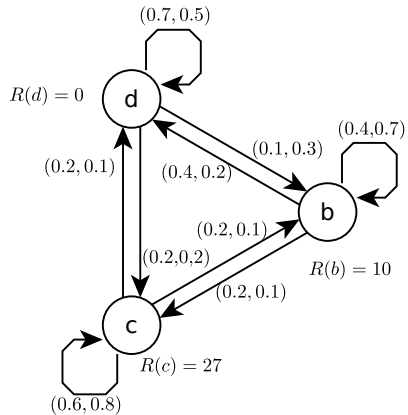
- **Iteración de valores**: Primero resolvemos  $V(s)$  en la ecuación de Bellman. Usamos  $V(s)$  para encontrar  $\Pi^*(s)$ .
- **Iteración de políticas**: Se fija la política y se resuelve un sistema de ecuaciones para encontrar  $V(s)$ , repitiendo hasta que la política no cambie.

El algoritmo de iteración de políticas itera la política hasta que converge sin preocuparse de encontrar los valores de  $V(s)$ .



## 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

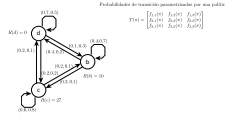
El algoritmo de iteración de políticas esta basado en el hecho de que dada una política conocida  $\pi$ , el proceso de decisión de Markov puede tratarse como una cadena de Markov.

Regresamos al ejemplo del tren.

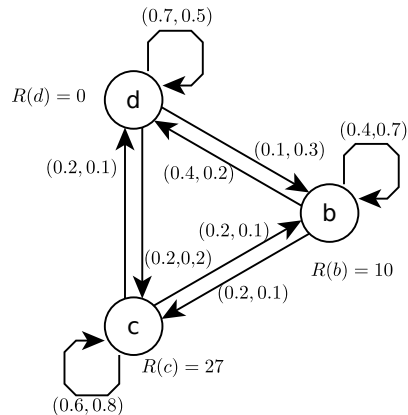
Ahora hemos anota las aristas con tuplas de probabilidades. Cada elemento de la tupla corresponde con una acción distinta.

En este ejemplo el primer elemento de la tupla corresponde con no anunciar,

el segundo elemento corresponde con la acción *Anuncio*.



## Algoritmo de iteración de políticas



Probabilidades de transición parametrizadas por una política

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

$$\pi = \begin{bmatrix} \pi(s=d) \\ \pi(s=b) \\ \pi(s=c) \end{bmatrix}$$

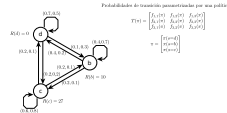
- └ Algoritmo de iteración de políticas

La política es una función.

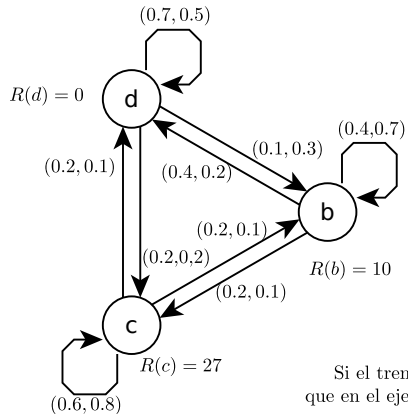
Para cada estado nos dice que acción tomar.

Esta función puede representarse como una lista de acciones.

La posición de la acción es la misma que la del estado correspondiente.



## Algoritmo de iteración de políticas



Probabilidades de transición parametrizadas por una política

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

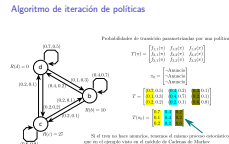
$$\pi_0 = \begin{bmatrix} \neg \text{Anuncio} \\ \neg \text{Anuncio} \\ \neg \text{Anuncio} \end{bmatrix}$$

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

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

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

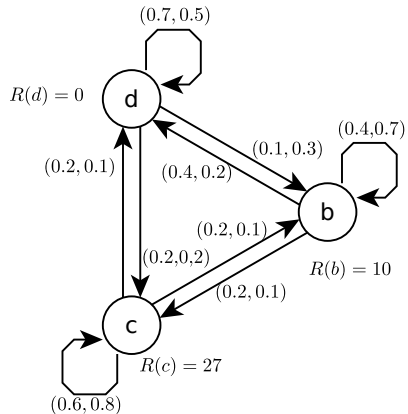
- └ Algoritmo de iteración de políticas



Podemos pensar en un arreglo  $T$  que tiene como entradas las tuplas. La política selecciona de cada tupla un elemento particular.

En este caso si la política para todo estado es no anunciar, entonces nuestra matriz de probabilidades de transición es la misma que en el ejemplo usado en el tema de cadenas de Markov.

## Algoritmo de iteración de políticas



Probabilidades de transición parametrizadas por una política

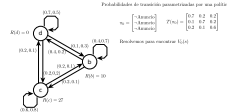
$$\pi_0 = \begin{bmatrix} \text{Anuncio} \\ \text{Anuncio} \\ \text{Anuncio} \end{bmatrix} \quad T(\pi_0) = \begin{bmatrix} 0.7 & 0.2 & 0.2 \\ 0.1 & 0.7 & 0.2 \\ 0.2 & 0.1 & 0.6 \end{bmatrix}$$

Resolvemos para encontrar  $V_0(s)$

# Procesos de Decisión de Markov

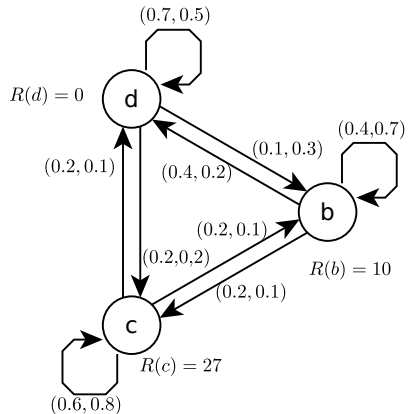
- └ Algoritmo de iteración de políticas

### Algoritmo de iteración de políticas



Encontraremos la utilidad horizonte infinito para cada estado.

## Algoritmo de iteración de políticas



Probabilidades de transición parametrizadas por una política

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

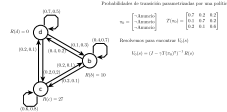
Resolvemos para encontrar  $V_0(s)$

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

# Procesos de Decisión de Markov

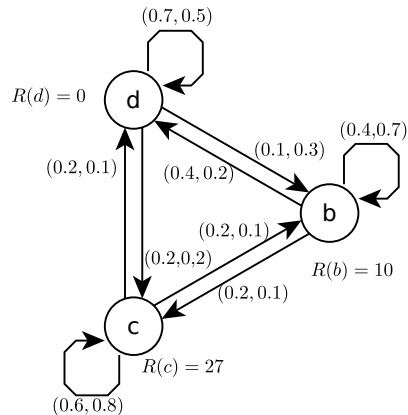
- └ Algoritmo de iteración de políticas

### Algoritmo de iteración de políticas



Observe que en la ecuación de Bellman matricial, la matriz de transición de la cadena de Markov aparece en forma transpuesta.

## Algoritmo de iteración de políticas



Probabilidades de transición parametrizadas por una política

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

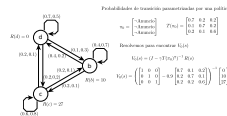
Resolvemos para encontrar  $V_0(s)$

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

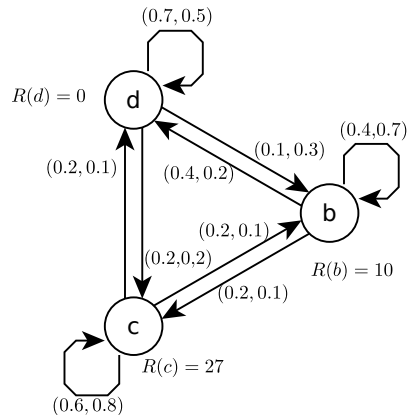
$$V_0(s) = \left( \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - 0.9 \begin{bmatrix} 0.7 & 0.1 & 0.2 \\ 0.2 & 0.7 & 0.1 \\ 0.2 & 0.2 & 0.6 \end{bmatrix} \right)^{-1} \begin{bmatrix} 0 \\ 10 \\ 27 \end{bmatrix}$$

## Algoritmo de iteración de políticas

Sustituimos...



## 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  $V_0(s)$

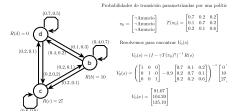
$$V_0(s) = (I - \gamma T(\pi_0)^T)^{-1} R(s)$$

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

$$V_0(x) = \begin{bmatrix} 91.07 \\ 104.20 \\ 135.10 \end{bmatrix}$$

## Algoritmo de iteración de políticas

Obtenemos los valores horizonte infinito.



# Procesos de Decisión de Markov

- └ Algoritmo de iteración de políticas

Probabilidades de transición generadas por un polio

$\pi_0 = \begin{bmatrix} -\lambda \text{ (aero)} & 0 & 0 \\ -\lambda \text{ (aero)} & 0 & 0 \\ -\lambda \text{ (aero)} & 0 & 0 \end{bmatrix}$   $\hat{P}(\pi_0) = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.1 & 0.7 & 0.2 \\ 0.2 & 0.1 & 0.7 \end{bmatrix}$   $\hat{I}_0(\pi_0) = \begin{bmatrix} 0.187 \\ 0.032 \\ 0.033 \end{bmatrix}$

Examinamos la política  $\pi_0$  para la función objetivo

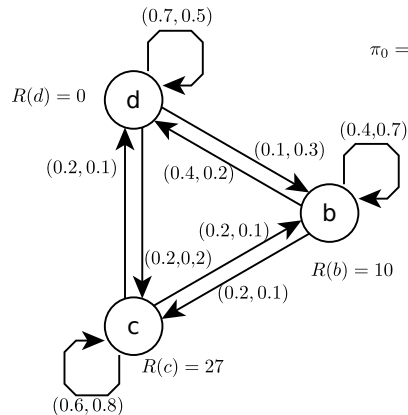
$\pi_0(\pi) = \arg \max_{\pi} P(\pi, \pi_0) / \pi(\pi)$

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

En este paso vamos a recalcular la política óptima  $\pi_1(s)$



## 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

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

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

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

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

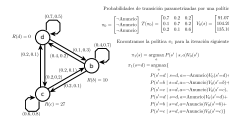
$$P(s'=c | s=d, a=\neg \text{Anuncio}) V_0(s'=c),$$

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

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

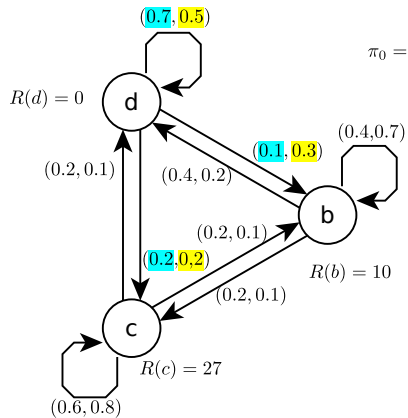
$$P(s'=c | s=d, a=\text{Anuncio}) V_0(s'=c) \}$$

## Algoritmo de iteración de políticas



Para el estado  $d$  encontraremos la acción que maximiza la utilidad.  
 Así quedan las sumatorias.

## 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

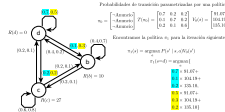
$$\pi_1(s) = \operatorname{argmax}_a P(s' | s, a) V_0(s')$$

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

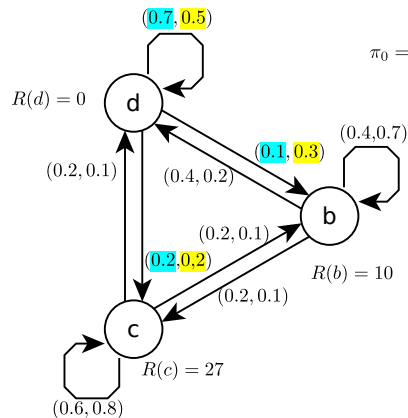
0.7	$\times 91.07 +$
0.1	$\times 104.19 +$
0.2	$\times 135.10,$
0.5	$\times 91.07 +$
0.3	$\times 104.19 +$
0.2	$\times 135.10\}$

- └ Algoritmo de iteración de políticas

Sustituimos valores...



## 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

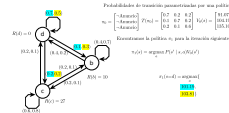
$$\pi_1(s) = \operatorname{argmax}_a P(s' | s, a) V_0(s')$$

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

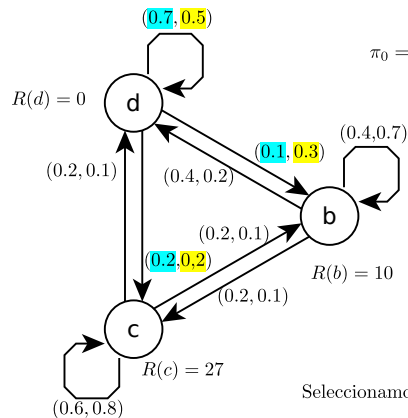
## Procesos de Decisión de Markov

### Algoritmo de iteración de políticas

Tomaremos la acción que da el valor máximo.



## Algoritmo de iteración de políticas



Probabilidades de transición parametrizadas por una política

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

Encontramos la política  $\pi_1$  para la iteración siguiente

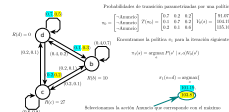
$$\pi_1(s) = \operatorname{argmax}_a P(s' | s, a) V_0(s')$$

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

$$\begin{matrix} 101.19, \\ 103.81 \end{matrix}$$

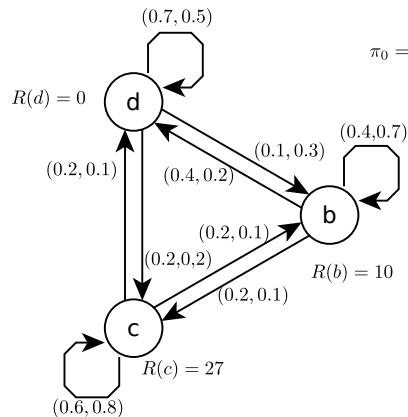
Seleccionamos la acción Anuncio que corresponde con el máximo

## Algoritmo de iteración de políticas



En este caso hacer el anuncio resulta en una mayor utilidad.

## Algoritmo de iteración de políticas



Probabilidades de transición parametrizadas por una política

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

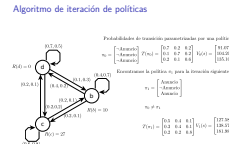
Encontramos la política  $\pi_1$  para la iteración siguiente

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

$$\pi_0 \neq \pi_1$$

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

- └ Algoritmo de iteración de políticas

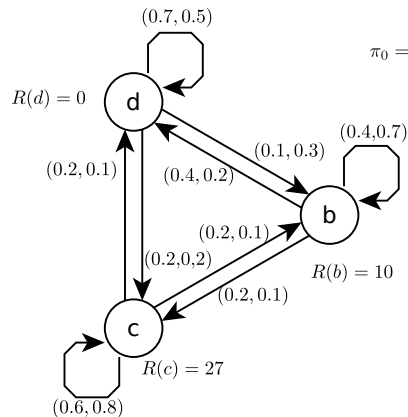


$\pi_1$  queda definida como hacer el anuncio si se esta en el dormitorio o el comedor pero no hacerlo si se está en el bar.

Observamos que  $\pi_0$  es diferente de  $\pi_1$  por lo que el algoritmo continuará iterando.

Calculamos las probabilidades de transición para  $\pi_1$  y resolvemos el sistema de ecuaciones obteniendo los nuevos valores de utilidad horizonte infinito  $V_1(s)$ .

## Algoritmo de iteración de políticas

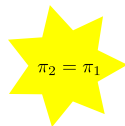


Probabilidades de transición parametrizadas por una política

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

Encontramos la política  $\pi_2$  para la iteración siguiente

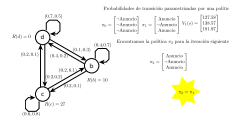
$$\pi_2 = \begin{bmatrix} \text{Anuncio} \\ \neg \text{Anuncio} \\ \text{Anuncio} \end{bmatrix}$$



## Procesos de Decisión de Markov

### Algoritmo de iteración de políticas

#### Algoritmo de iteración de políticas



Recalculamos la política óptima  $\pi_2$  observando que el algoritmo convergió.

El algoritmo termina.