Tema 4: Procesos de Decisión de Markov

José Luis Ruiz Reina

Departamento de Ciencias de la Computación e Inteligencia Artificial
Universidad de Sevilla

Inteligencia Artificial

Contenido

Procesos de Decisión de Markov

- Tratamos ahora secuencias de acciones cuyos efectos son inciertos.
  - Similares a espacios de estados, pero el efecto de una acción está descrito mediante una distribución de probabilidad
  - El resultado de una acción sobre un estado ya *no es determinista*
  - Además, se introduce la noción de “recompensa” en un estado.

Procesos de Decisión de Markov

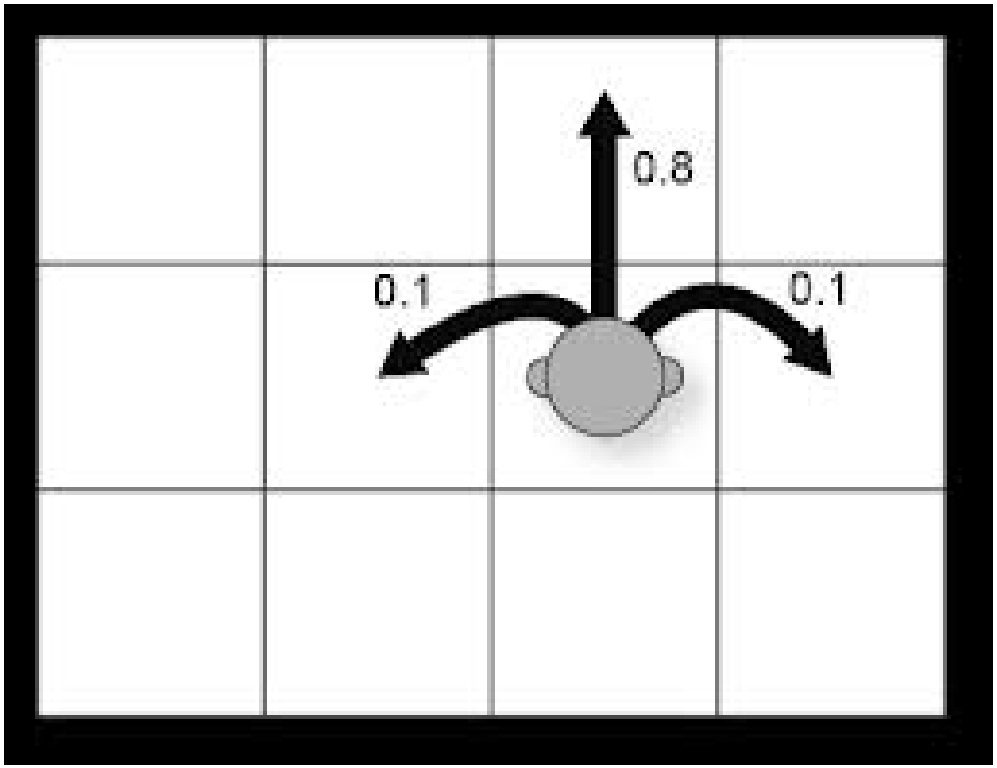
- Se busca cuál es la mejor acción a aplicar en cada momento:
  - Pero ya no tiene sentido buscar de antemano una secuencia de acciones
  - Mejor verlo como “problema de decisión secuencial”
- ¿Cómo decidir la mejor acción en cada momento?
  - Sabiendo de antemano sus posibles efectos y con qué probabilidad
  - Y la recompensa en cada situación

Conceptos probabilísticos

- Repasar los siguientes conceptos básicos de teoría de la probabilidad:
  - Función de probabilidad
  - Variable aleatoria
  - Distribución de probabilidad
  - Valor esperado de una variable aleatoria

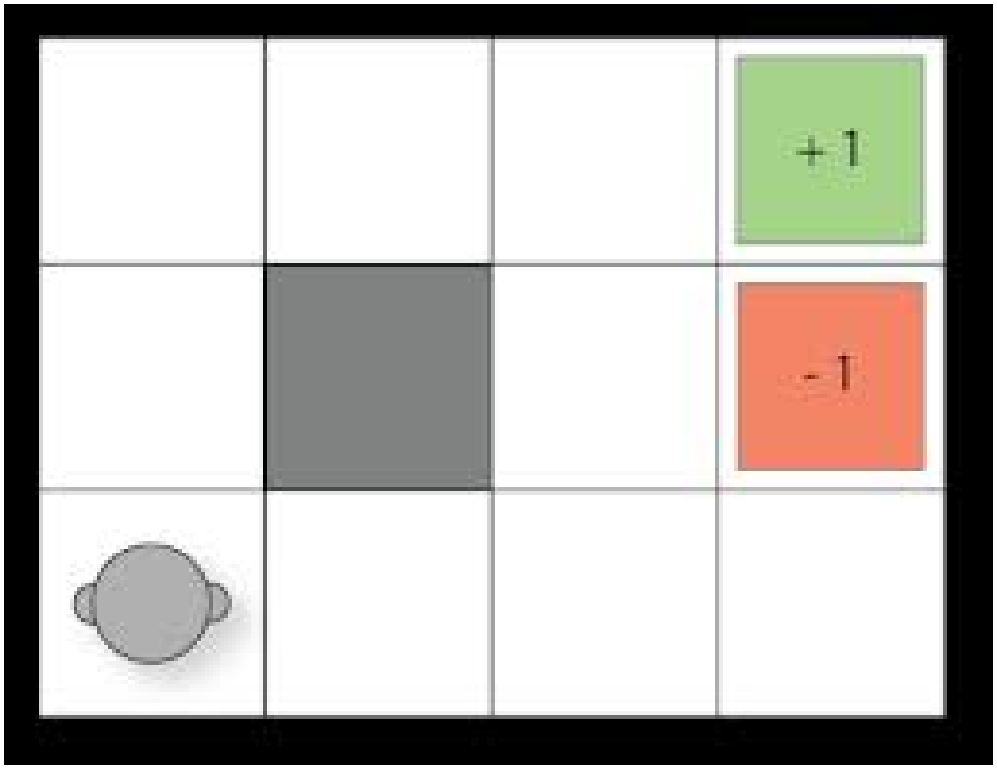
Ejemplo de MDP: Cuadrícula

Un robot en una cuadrícula puede intentar moverse en las cuatro direcciones: *arriba*, *abajo*, *izquierda*, *derecha* (si choca con la pared, queda en el mismo sitio). Cada acción consigue su objetivo con probabilidad 0.8, pero a veces se mueve en direcciones en angulo recto a la que se quería (con probabilidad 0.1 para cada lado).



Ejemplo de MDP: Cuadrícula

En la siguiente cuadrícula, vista como espacio de estados, los *estados* son las casillas, las acciones son las descritas anteriormente y el objetivo es llegar al estado final +1, evitando el -1



Ejemplo de MDP: Cuadrícula

- Si el mundo fuera determinista, una solución sería: *arriba*, *arriba*, *izquierda*, *izquierda*, *izquierda*.
  - Sin embargo, esa secuencia sólo sigue el camino deseado con probabilidad $\mathbf{(0{,}8)^5}$. Podría ocurrir también (con probabilidad más baja) que esa secuencia no le llevara al objetivo deseado.
- Otra novedad respecto a espacio de estados: cada estado tiene una **recompensa** asociada.
  - En el ejemplo: suponemos que todos los estados tienen recompensa -0.04, excepto los dos estados terminales, que tienen recompensa +1, y -1, respectivamente. En los estados terminales, ya no se puede aplicar ninguna acción.
- El objetivo es decidir en cada estado qué acción aplicar, de manera que se maximice el *total de recompensas* de los estados por donde se pasa.
  - Más adelante matizaremos qué queremos decir con el *total de recompensas*

Procesos de Decisión de Markov

Un *Proceso de Decisión de Markov* viene definido por:

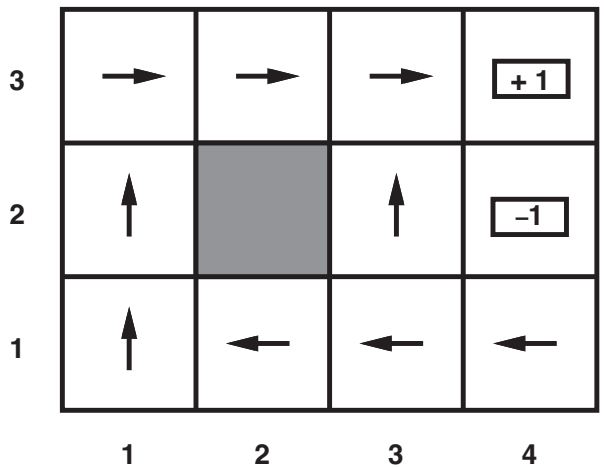
- Un conjunto $\boldsymbol{S}$ de **estados** (con un estado inicial $\boldsymbol{s_0}$)
- Para cada estado, un conjunto $\boldsymbol{A(s)}$ de **acciones** aplicables a ese estado.
- Un **modelo de transición**, dado por una distribución de probabilidad $\boldsymbol{P(s'|s, a)}$ para cada par de estados $\boldsymbol{s'}$, $\boldsymbol{s}$ y acción $\boldsymbol{a}$ aplicable a $\boldsymbol{s}$ (indicando la probabilidad de que aplicando $\boldsymbol{a}$ a $\boldsymbol{s}$ se obtenga $\boldsymbol{s'}$).
- Una función de **recompensa** $\boldsymbol{R(s)}$.

Propiedad de Markov: el efecto (incierto) de una acción sobre un estado sólo depende de la acción y del propio estado (y no de estados anteriores)

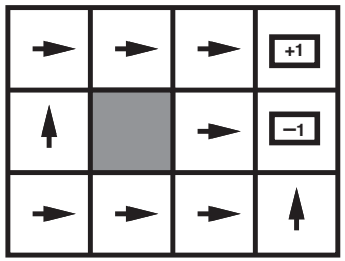
Políticas

- En este contexto, una solución no puede ser una secuencia de acciones, ya que el efecto de cada acción es incierto.
- Más bien buscamos una *política* que en cada posible estado recomiende una acción a aplicar: por cada estado que pasemos, aplicamos la acción que nos recomienda esa política
  - Formalmente: una **política** es una función π definida sobre el conjunto de estados $\boldsymbol{S}$, de manera que $\pi(\boldsymbol{s}) \in \boldsymbol{A}(\boldsymbol{s})$
- Una misma política puede generar secuencias de acciones distintas (aunque unas con más probabilidades que otras).
- Se busca la política **óptima**: aquella que maximice la *recompensa media esperada* para las posibles secuencias de acciones que se puedan generar

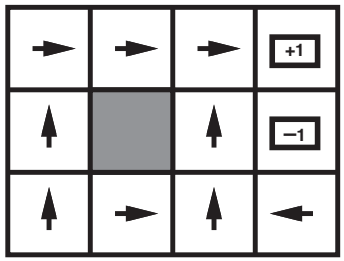
Ejemplo de políticas en la cuadrícula



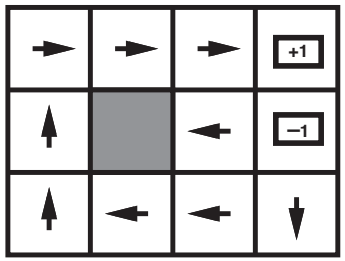
(a)



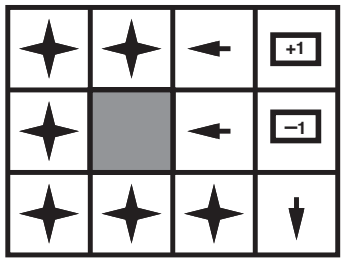
$R(s) < -1.6284$



$-0.4278 < R(s) < -0.0850$



$-0.0221 < R(s) < 0$



$R(s) > 0$

(b)

Valoración de secuencia de estados en el tiempo

- Supongamos que mediante aplicación de una secuencia de acciones, se ha generado una secuencia de estados $q_0 q_1 q_2 \cdots$
- ¿Cómo valoramos una secuencia de estados? Idea: a partir de las recompensas, pero *penalizando* el largo plazo
- Valoración mediante **recompensa con descuento**:

$$V([q_0 q_1 q_2 \cdots]) = R(q_0) + \gamma R(q_1) + \gamma^2 R(q_2) + \cdots$$

donde γ es el llamado **factor de descuento**

- En el ejemplo de la cuadrícula:
  - Con $\gamma = 0{,}8$, la secuencia de estados $(1,1),(2,1),(3,1),(3,2),(3,3),(3,4)$ tiene una valoración $-0{,}04 - 0{,}04 \cdot 0{,}8 - 0{,}04 \cdot 0{,}8^2 - 0{,}04 \cdot 0{,}8^3 - 0{,}04 \cdot 0{,}8^4 + 1 \cdot 0{,}8^5 = 0{,}193216$

Valoraciones de secuencias: observaciones

- Suponemos horizonte infinito
  - No hay un plazo fijo de terminación
  - Procesos estacionarios: la política óptima a partir de un momento sólo depende del estado en ese momento
- ¿Esto implica que las valoraciones pueden ser infinitas? En general, no:
  - Si hay estados terminales, los asimilamos a estados a partir del cual las recompensas son cero.
  - Aún con posibles secuencias infinitas, si las recompensas están acotadas por una cantidad $\boldsymbol{R_{max}}$ y $\gamma < \mathbf{1}$, entonces la valoración de una secuencia no puede ser mayor de $\boldsymbol{R_{max}}/(\mathbf{1} - \gamma)$ (*¿por qué?*)

Valoración de estados respecto de una política

- Dada una política π y un estado $\boldsymbol{s}$, podemos valorar $\boldsymbol{s}$ respecto de π teniendo en cuenta la valoración de las secuencias de estados que se generan si se sigue dicha política a partir de $\boldsymbol{s}$
- Ejemplo: si en la cuadrícula estamos en el estado $\mathbf{(1,4)}$ y aplicamos la política (a) del gráfico anterior, podríamos generar distintas secuencias, cada una con una probabilidad y una valoración. Entre otras:
  - $\mathbf{(1,4)}$, $\mathbf{(1,3)}$, $\mathbf{(1,2)}$, $\mathbf{(1,1)}$, $\mathbf{(2,1)}$, $\mathbf{(3,1)}$, $\mathbf{(3,2)}$, $\mathbf{(3,3)}$, $\mathbf{(3,4)}$, con probabilidad $\mathbf{0{,}8^8 = 0{,}168}$ y valoración 0.0013 (siendo $\gamma = \mathbf{0{,}8}$ y $\boldsymbol{R} = \mathbf{-0{,}04}$)
  - $\mathbf{(1,4)}$, $\mathbf{(1,3)}$, $\mathbf{(2,3)}$, $\mathbf{(3,3)}$, $\mathbf{(3,4)}$ con probabilidad $\mathbf{0{,}8^3 \cdot 0{,}1 = 0{,}0512}$ y valoración 0.291
  - $\mathbf{(1,4)}$, $\mathbf{(1,3)}$, $\mathbf{(2,3)}$, $\mathbf{(2,4)}$ con probabilidad $\mathbf{0{,}8 \cdot 0{,}1^2 = 0{,}008}$ y valoración -0.609
  - $\mathbf{(1,4)}$, $\mathbf{(2,4)}$ con probabilidad $\mathbf{0{,}1}$ y valoración -0.84
  - $\ldots$

Valoración de estados respecto de una política

- Idea: valorar un estado $\boldsymbol{s}$ respecto de una política π como la media esperada (es decir, ponderada por su probabilidad) de las valoraciones de todas las secuencias que se podrían obtener.
  - En el ejemplo anterior:
    $\mathbf{0{,}168 \cdot 0{,}0013 + 0{,}0512 \cdot 0{,}291 - 0{,}008 \cdot 0{,}609 - 0{,}1 \cdot 0{,}84 + \cdots}$
- La valoración de un estado respecto de una política π la notamos por $\boldsymbol{V^{\pi}(s)}$

Cálculo de valoración respecto de una política

- Afortunadamente, para obtener $\boldsymbol{V^{\pi}(s)}$ *no necesitaremos calcular todas las posibles secuencias* que podrían iniciarse en $\boldsymbol{s}$
- La clave está en la siguiente propiedad de $\boldsymbol{V^{\pi}}$, que relaciona la valoración de cada estado con la de sus "vecinos":

$$\boldsymbol{V^{\pi}(s)} = \boldsymbol{R(s)} + \gamma \cdot \sum_{\boldsymbol{s'}} (\boldsymbol{P(s'|s, \pi(s))} \cdot \boldsymbol{V^{\pi}(s')})$$

- Calcular $\boldsymbol{V^{\pi}}$ es resolver ese sistema de ecuaciones líneales:
  - Las incognitas son los $\boldsymbol{V^{\pi}(s)}$, una por cada estado $\boldsymbol{s}$
  - Hay tantas ecuaciones como estados

Cálculo de valoración respecto de una política (ejemplo)

En la política (a) de la cuadrícula en la figura anterior, éstas serían alguna de las ecuaciones que salen:

- $\boldsymbol{V^{\pi}(1,1) = -0{,}04 + \gamma \cdot (0{,}8 \cdot V^{\pi}(2,1) + 0{,}1 \cdot V^{\pi}(1,1) + 0{,}1 \cdot V^{\pi}(1,2))}$
- $\boldsymbol{V^{\pi}(1,2) = -0{,}04 + \gamma \cdot (0{,}8 \cdot V^{\pi}(1,1) + 0{,}2 \cdot V^{\pi}(1,2))}$
- $\boldsymbol{V^{\pi}(1,3) = -0{,}04 + \gamma \cdot (0{,}8 \cdot V^{\pi}(1,2) + 0{,}1 \cdot V^{\pi}(1,3) + 0{,}1 \cdot V^{\pi}(2,3))}$
- $\boldsymbol{\ldots}$
- $\boldsymbol{V^{\pi}(3,4) = +1}$

Resolviendo este sistema, obtenemos $\boldsymbol{V^{\pi}}$

Aproximaciones al cálculo de v^π

- Muestreo: Generar secuencias de estados aplicando las acciones que indica la política teniendo en cuenta la distribución de probabilidad que indica el modelo de transición y calcular la media de sus valoraciones.
- Iteración de valores:
  - Comenzamos por un valor arbitrario $V_0^\pi(s)$ para cada estado s
  - En cada iteración se aplica la fórmula que relaciona la valoración de un estado con la de sus "vecinos"

$$V_{i+1}^\pi(s) = R(s) + \gamma \cdot \sum_{s'} (P(s'|s, \pi(s)) \cdot V_i^\pi(s'))$$

Políticas óptimas y valoraciones de estados

- La **valoración de un estado s**, notada $V(s)$, se define como:

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

  - la mejor valoración que una política pueda conseguir a partir de un estado.
- Podemos además definir la **política óptima** π^*:

$$\pi^*(s) = \operatorname*{argmax}_{a \in A(s)} \sum_{s'} (P(s'|s,a) \cdot V(s'))$$

  - aplicar la acción que lleve a la mejor valoración esperada en el estado siguiente
- Se tiene que $V^{\pi^*} = V$.

Objetivo: calcular V y π^*

Ecuaciones de Bellman

- De manera análoga a V^{π}, podemos describir $V(s)$ en función de las valoraciones de los estados "vecinos". Son las llamadas **ecuaciones de Bellman**:

$$V(s) = R(s) + \gamma \cdot \max_{a \in A(s)} \sum_{s'} (P(s'|s,a) \cdot V(s'))$$

- Nuevamente, es un sistemas de ecuaciones (una ecuación por estado).
- La solución a este sistema de ecuaciones nos da la valoración de cada estado y a partir de ésta, la política óptima

Ecuaciones de Bellman (ejemplo)

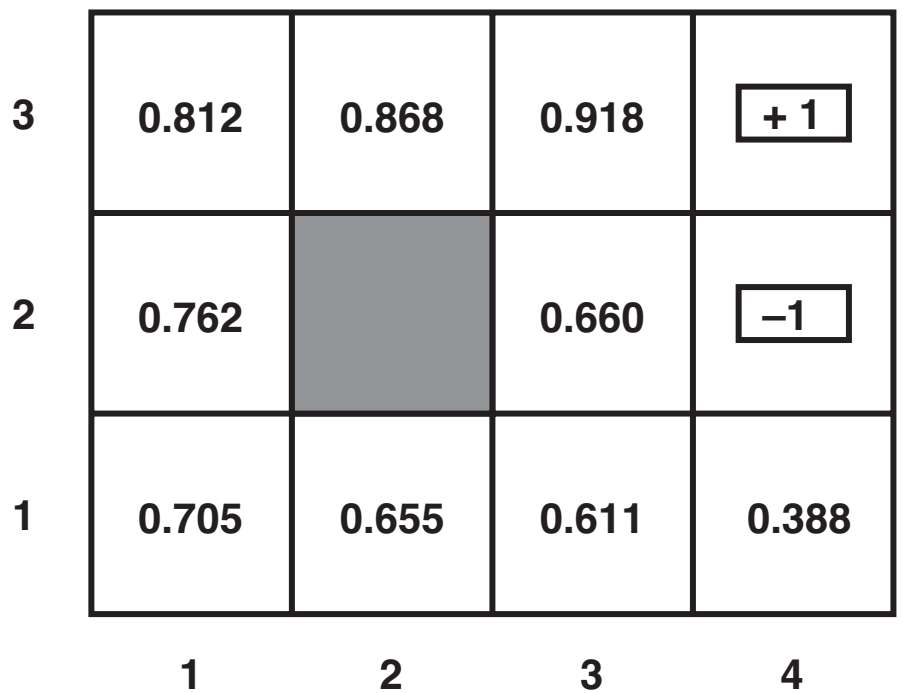
- Por ejemplo, en la cuadrícula, la ecuación correspondiente al estado de la casilla $(1,1)$ es:

$$\begin{aligned} V(1,1) = -0{,}04 + \gamma \cdot \max[\ & 0{,}8V(1,2) + 0{,}1V(2,1) + 0{,}1V(1,1), \\ & 0{,}9V(1,1) + 0{,}1V(1,2), \\ & 0{,}9V(1,1) + 0{,}1V(2,1), \\ & 0{,}8V(2,1) + 0{,}1V(1,2) + 0{,}1V(1,1)] \end{aligned}$$

- Las demás ecuaciones, son similares, una por cada estado

Ecuaciones de Bellman (ejemplo)

La siguiente figura muestra la valoración de cada estado en el problema de la cuadrícula (para $\gamma = \mathbf{1}$, y $\boldsymbol{R(s)} = -\mathbf{0{,}04}$), obtenida solucionando las correspondientes ecuaciones de Bellman.



Cálculo de política óptima

A partir de la valoración anterior, podemos calcular la política óptima, usando la fórmula

$$\pi^*(s) = \underset{a \in A(s)}{\mathbf{argmax}} \sum_{s'} (P(s'|s,a) \cdot V(s')).$$

Por ejemplo, en el estado (1,1):

- Acción ***arriba***: $\mathbf{0{,}8 \cdot 0{,}762 + 0{,}1 \cdot 0{,}705 + 0{,}1 \cdot 0{,}655 = 0{,}7456}$
- Acción ***abajo***: $\mathbf{0{,}8 \cdot 0{,}705 + 0{,}1 \cdot 0{,}705 + 0{,}1 \cdot 0{,}655 = 0{,}7}$
- Acción ***izquierda***: $\mathbf{0{,}8 \cdot 0{,}705 + 0{,}1 \cdot 0{,}705 + 0{,}1 \cdot 0{,}762 = 0{,}7107}$
- Acción ***derecha***: $\mathbf{0{,}8 \cdot 0{,}655 + 0{,}1 \cdot 0{,}705 + 0{,}1 \cdot 0{,}762 = 0{,}6707}$

Luego la acción óptima en el estado $\mathbf{(1,1)}$ es moverse hacia *arriba*. El resto se muestra en la figura de la diapositiva 11 etiquetada con (a)

Iteración de valores

- ¿Cómo resolvemos las ecuaciones de Bellman?
  - Cada $\boldsymbol{V(s)}$ es una incognita en el sistema de ecuaciones.
  - Hay tantas ecuaciones como estados
  - Problema: *no es un sistema lineal* (debido al ***max***)
- Se aplica un método *iterativo*
  - Comenzamos con un valor arbitrario $\boldsymbol{V_0(s)}$, para cada estado $\boldsymbol{s}$
  - En cada iteración se aplican las ecuaciones de Bellman para *actualizar* los valores que se tienen hasta el momento:

$$\boldsymbol{V_{i+1}(s)} \longleftarrow \boldsymbol{R(s)} + \gamma \cdot \max_{\boldsymbol{a} \in \boldsymbol{A(s)}} \sum_{\boldsymbol{s'}} (\boldsymbol{P(s'|s,a)} \cdot \boldsymbol{V_i(s')})$$

Iteración de valores: propiedades

- Se puede demostrar que los valores que se van calculando en las diferentes iteraciones convergen asintóticamente hacia la solución (única si $\gamma < 1$) de las ecuaciones de Bellman
- La convergencia es rápida para valores de γ pequeños
- Además, podemos acotar con bastante precisión el error que cometemos . Más precisamente:
  - Sea $||V_{i+1} - V_i|| = \max_s |V_{i+1}(s) - V_i(s)|$
  - Se tiene que si $||V_{i+1} - V_i|| < \epsilon \cdot (1 - \gamma)/\gamma$, entonces $||V_{i+1} - V|| < \epsilon$, donde V es la solución exacta a las ecuaciones de Bellman
  - Éste será el criterio de parada en las iteraciones

Algoritmo de iteración de valores

- **Entrada:**
  - Un proceso de decisión de Markov: conjunto de estados $\boldsymbol{S}$, $\boldsymbol{A(s)}$, modelo de transición $\boldsymbol{P(s'|s,a)}$, recompensa $\boldsymbol{R(s)}$ y descuento γ
  - $\epsilon > \mathbf{0}$, cota de error máximo permitido
- **Salida:**
  - Una valoración $\boldsymbol{V(s)}$ para cada estado
- **Procedimiento:**
  - Inicio: Sea $\boldsymbol{V_0}$ una función sobre los estados, con valor 0 para cada estado, $\delta = \infty$ e $\boldsymbol{i}$ igual a $\mathbf{0}$
  - Repetir
    - Hacer $\boldsymbol{i} = \boldsymbol{i} + \mathbf{1}$ y $\delta = \mathbf{0}$
    - Para cada $\boldsymbol{s} \in \boldsymbol{S}$ hacer:
      - $\boldsymbol{V_i(s)} \longleftarrow \boldsymbol{R(s)} + \gamma \cdot \max\limits_{\boldsymbol{a} \in \boldsymbol{A(s)}} \sum\limits_{\boldsymbol{s'}} (\boldsymbol{P(s'|s,a)} \cdot \boldsymbol{V_{i-1}(s')})$
      - Si $|\boldsymbol{V_i(s)} - \boldsymbol{V_{i-1}(s)}| > \delta$ entonces $\delta = |\boldsymbol{V_i(s)} - \boldsymbol{V_{i-1}(s)}|$

    hasta que $\delta < \epsilon \cdot (\mathbf{1} - \gamma)/\gamma$
  - Devolver $\boldsymbol{V_i}$

Iteración de politicas

- Existe una manera alternativa de obtener $\boldsymbol{V}$, comenzando con una política inicial π_0 cualquiera:

Paso 1 Calcular $\boldsymbol{V}^{\pi_i}$

Paso 2 A partir de $\boldsymbol{V}^{\pi_i}$, calcular una nueva política π_{i+1} que en cada estado recomiende la acción que mejor valoración espera, respecto de $\boldsymbol{V}^{\pi_i}$

- Los dos pasos anteriores se iteran hasta que π_i y π_{i+1} coinciden, en cuyo caso hemos conseguido la política óptima, y su valoración asociada es precisamente $\boldsymbol{V}$.

Iteración de politicas: percepciones

- El proceso anterior termina (pues hay un número finito de políticas distintas) y devuelve la valoración buscada (ya que se llega a una solución de las ecuaciones de Bellman).
- ¿Cómo calcular V^{π_i} en el paso 1?:
  - Como ya hemos visto, resolviendo un sistema de ecuaciones lineal (diapositivas 16 y 17).
  - Aplicando el método iterativo (similar al que se usa con las ecuaciones de Bellman, diapositiva 18)
- ¿Cómo calcular π_{i+1} en el paso 2?:

$$\pi_{i+1}(s) = \underset{a \in A(s)}{\mathbf{argmax}} \sum_{s'} (P(s'|s,a) \cdot V^{\pi_i}(s'))$$

Algoritmo de iteración de políticas

- **Entrada:**
  - Un proceso de decisión de Markov: conjunto de estados $\boldsymbol{S}$, $\boldsymbol{A(s)}$, modelo de transición $\boldsymbol{P(s'|s,a)}$, recompensa $\boldsymbol{R(s)}$ y descuento γ
  - Un número $\boldsymbol{k}$ de iteraciones
- **Salida:** Una política óptima π y su valoración asociada
- **Procedimiento:**
  - Inicio: π una política aleatoria, asignando una acción a cada estado
  - Repetir
    - Calcular $\boldsymbol{V^{\pi}}$
    - Hacer ***actualizada*** igual a ***Falso***
    - Para cada estado $\boldsymbol{s}$:
      - Si
      $$\max_{a \in A(s)} \sum_{s'} (P(s'|s,a) \cdot V^{\pi}(s')) > \sum_{s'} (P(s'|s,\pi(s)) \cdot V^{\pi}(s')),$$
      entonces hacer $\pi(\boldsymbol{s})$ igual a $\operatorname{argmax}_{a \in A(s)} \sum_{s'} (P(s'|s,a) \cdot V^{\pi}(s'))$ y ***actualizada*** igual ***Verdad***

    hasta que ***actualizada*** sea ***Falso***
  - Devolver π y $\boldsymbol{V^{\pi}}$

Bibliografía

- Jurafsky, D. y Martin, J.H. *Speech and Language Processing* (Second Edition) (Prentice-Hall, 2009)
  - Cap. 6: “Hidden Markov and Maximum Entropy Models ”
- Russell, S. y Norvig, P. *Artificial Intelligence (A modern approach)* (Third edition) (Prentice Hall, 2009)
  - Cap. 17 (hasta 17.3): “Making complex decisions”
- Russell, S. y Norvig, P. *Inteligencia Artificial (Un enfoque moderno)* (Segunda edición) (Pearson Educación, 2004)
  - Cap. 17 (hasta 17.3): “Toma de decisiones complejas”