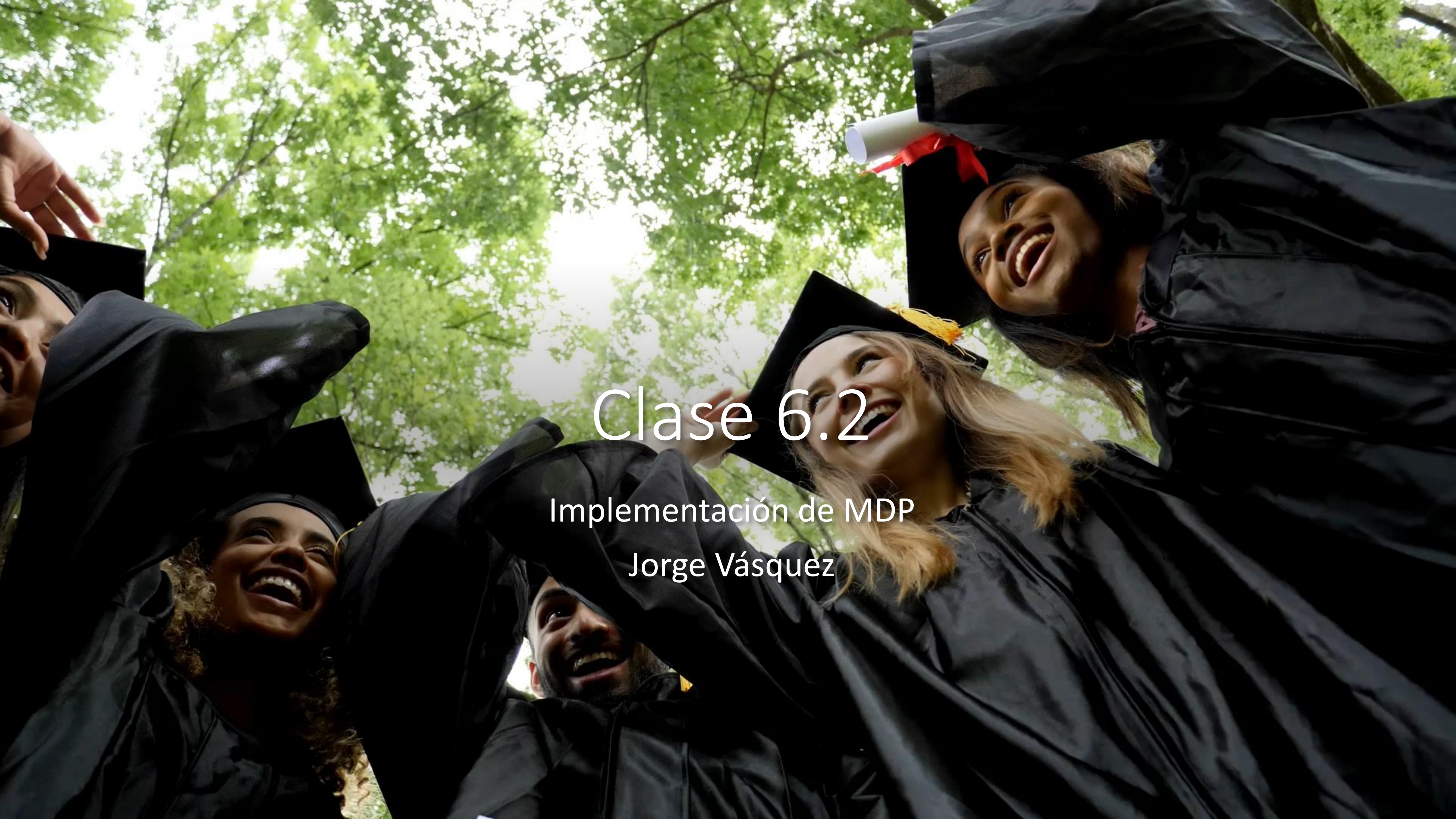


Clase 6.2

Implementación de MDP

Jorge Vásquez

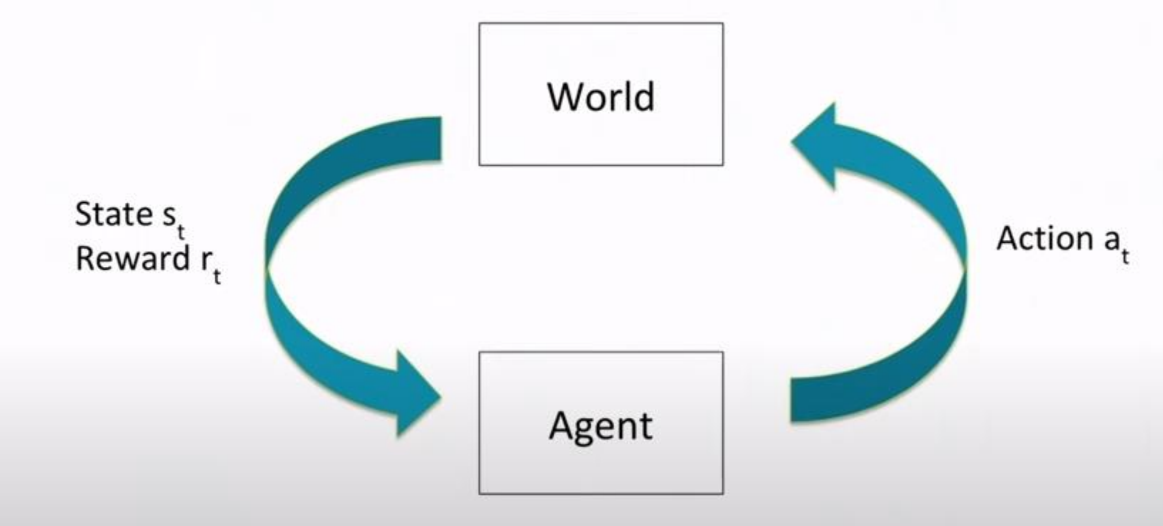
Modelos, Políticas, Valores

Modelo	Política	Función de Valor
• Modelo matemático de dinámica y recompensas.	• Función de mapeo del agente, desde estados a acciones	• Recompensas futuras por estar en estado y/o acción siguiendo una política

Modelando

- Proceso de Markov
- Proceso de Recompensas Markoviano
- Proceso de Decisiones de Markov
- Evaluación y Control de MDPs

Observación Completa: MDP



- MDP puede modelar un gran numero de problemas interesantes y configuraciones
  - Bandits: MDP con estado único

Propiedades Markovianas

- Información del Estado: historia con suficiente estadística
- Estado St es Markov si y solo si:

$$p(s_{t+1}|s_t, a_t) = p(s_{t+1}|h_t, a_t)$$

- El futuro es independiente del presente dado en el pasado

Propiedades Markovianas

- Información del Estado: historia con suficiente estadística
- Estado St es Markoviano si y solo si:

$$p(s_{t+1} | s_t, a_t) = p(s_{t+1} | h_t, a_t)$$

- El futuro es independiente del presente dado

Proceso de Márkov

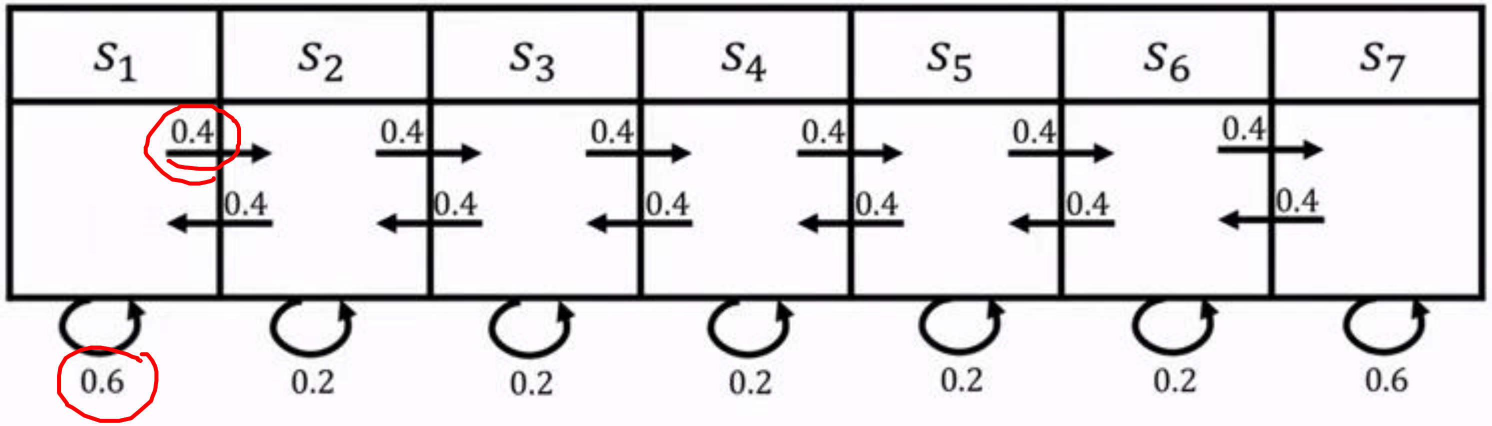
- Es un proceso *random* sin memoria
  - Secuencia de estados *random* markovianas
- Definición de Proceso de Márkov
  - S es un estado finito de estado s ∈ S ✓
  - P es un modelo de transición/dinámica que especifica $P(s_{t+1} = s' | s_t = s \ )$ ✓
- nota: sin recompensas, ~~sin acciones~~
- Si un numero finito de estado, pueden expresar P como una matriz seria:

Q P
R
V

$$P = \begin{pmatrix} P(s_1|s_1) & P(s_2|s_1) & \cdots & P(s_N|s_1) \\ P(s_1|s_2) & P(s_2|s_2) & \cdots & P(s_N|s_2) \\ \vdots & \vdots & \ddots & \vdots \\ P(s_1|s_N) & P(s_2|s_N) & \cdots & P(s_N|s_N) \end{pmatrix}$$

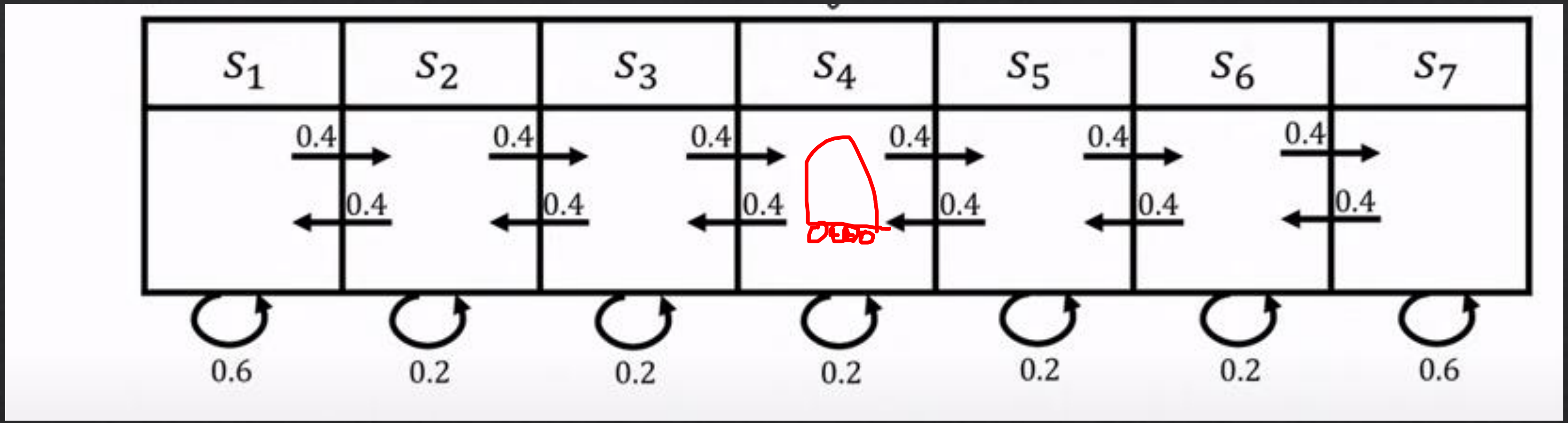
Matriz

Proceso de Transición de Márkov P, numérico



$$P = \begin{pmatrix} 0.6 & 0.4 & 0 & 0 & 0 & 0 & 0 \\ 0.4 & 0.2 & 0.4 & 0 & 0 & 0 & 0 \\ 0 & 0.4 & 0.2 & 0.4 & 0 & 0 & 0 \\ 0 & 0 & 0.4 & 0.2 & 0.4 & 0 & 0 \\ 0 & 0 & 0 & 0.4 & 0.2 & 0.4 & 0 \\ 0 & 0 & 0 & 0 & 0.4 & 0.2 & 0.4 \end{pmatrix}$$

Proceso de Transición de Márkov P, episodios



1 time step, 4 time steps, 4 pasos

- Episodios de muestras empezando de s4
  - s4, s5, s6, s7, s7, s7, ….
  - S4, s4, s5, s4, s5, s6,….
  - S4,s3,s2,s1,…

$S_1 \to S_2 \to S_3 - S_4$

Proceso de Recompensas de Markov, MRP

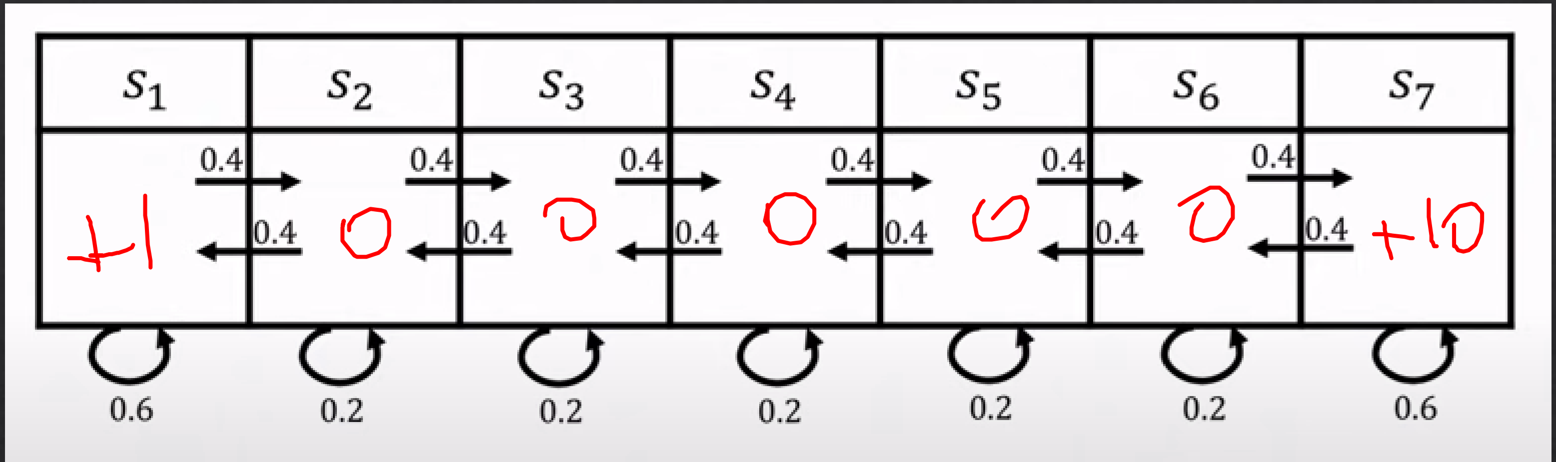
- Función de Recompensas es parte del Proceso de Recompensas de Markov (MRP)
- MRP es el Proceso de Markov + Recompensas
  - S es un conjunto finito de estado s ∈ S
  - P es modelo dinámico o de transición que especifica $P(s'|s)$
  - R es una función de recompensas $R(s)=E[r_t|s_t]$
  - Factor de Descuento $\gamma \in [0,1]$
  - No hay acciones todavía
- N numero de estados, R puede ser un vector

S.I.
DATA
P( )

$\underbrace{\text{MARKOV } P(s'|s) + \text{fn de recop.}}_{\text{MRP}}$

r(s)

Proceso de Recompensas de Markov, MRP



- Recompensas: En s1 es +1, s7=10, 0 en todos los otros estados.

Ganancia (Retorno) y Función de Valor

- Definición de Horizonte ✓ $H \; (\gamma \to H)$
  - Numero de timesteps para cada episodio
  - MRP finito
  - Puede ser infinito (acciones)
- Definición de Ganancia *Gt*
  - Suma de recompensas descontadas desde t a H
- Función de Valor V(s)
  - Recompensas esperadas comenzando en estado s

inm + futuras dependientes

$$G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \cdots$$

$E[G_t | s]$

$$V(s) = \mathbb{E}[G_t | s_t = s] = \mathbb{E}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \cdots | s_t = s]$$

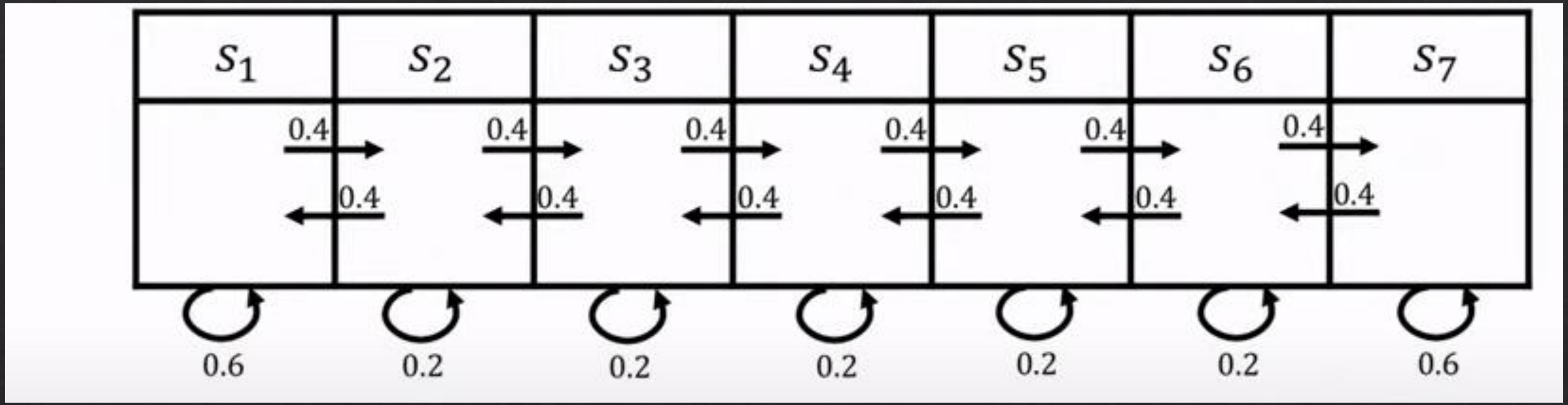
Factor de Descuento

- Conveniente Matemáticamente
- Evita ganancias y valores infinitas
- Humanos actuamos con un factor de descuento menor a 1
- $\gamma = 0$ solo importa la recompensa inmediata
- $\gamma = 1$ recompensas futuras son igual de beneficiosas que las inmediatas
- si los episodios son siempre finitos, se puede usar $\gamma = 1$ (reemplazo H por gamma) → generalización

$G_t = r_0 + \gamma \sum r_i$ → 0

$[0,1]$

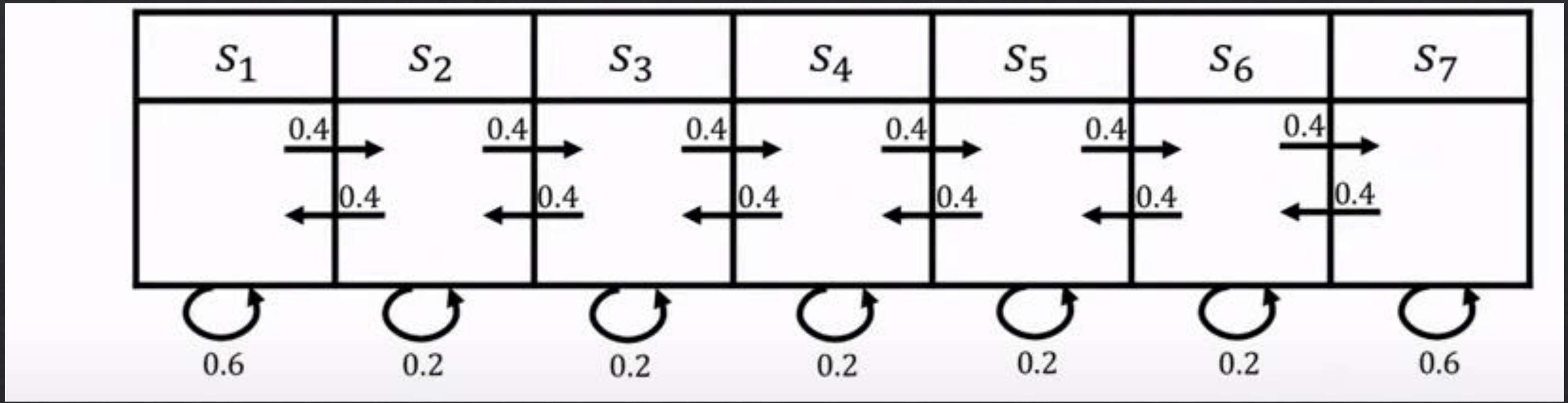
MRP en Rover



- Recompensas:
  - +1 en s1, +10 en s7, 0 en el resto
  - Recompensas de muestras para episodios de cuatro pasos , con $\gamma = 1/2$
  - s4, s5, s6, s7: 0 +1/2*0+1/4*0+1/8*10=1.25

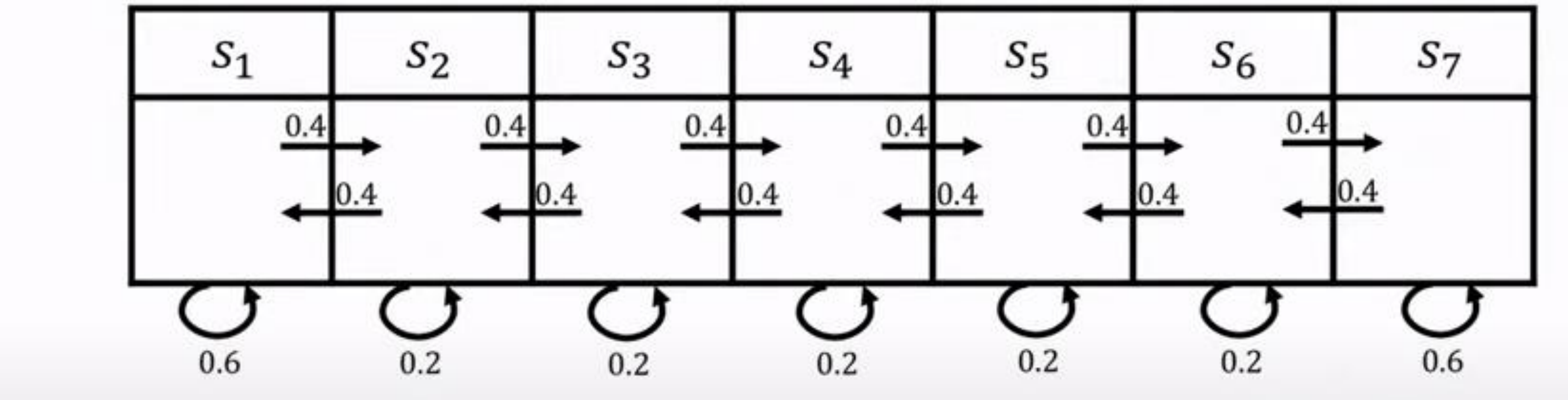
$\gamma^3 10 = 1.25$

MRP en Rover



- Recompensas:
  - +1 en s1, +10 en s7, 0 en el resto
  - Recompensas de muestras para episodios de cuatro pasos , con γ =1/2
  - s4, s5, s6, s7: 0 +1/2*0+1/4*0+1/8*10=1.25
  - s4, s4, s5, s4: 0 +1/2*0+1/4*0+1/8*0=0
  - s4, s3, s2, s1: 0 +1/2*0+1/4*0+1/8*10=1.25

MRP en Rover



- Recompensas: +1 en s1, +10 en s7, 0 en el resto
- Función de Valor: retorno esperado partiendo desde estado s

$$V(s) = \mathbb{E}[G_t | s_t = s] = \mathbb{E}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \cdots | s_t = s]$$

- Recompensas de muestras para episodios de cuatro pasos , con $\gamma = 1/2$
  - s4, s5, s6, s7: 0 +1/2*0+1/4*0+1/8*10=1.25
  - s4, s4, s5, s4: 0 +1/2*0+1/4*0+1/8*0=0
  - s4, s3, s2, s1: 0 +1/2*0+1/4*0+1/8*10=1.25
- V = [1.53 0.37 0.13 0.22 0.85 3.59 15.31]

Computar el Valor de un MRP

- Puede ser estimado mediante Simulación
  - Generar un numero largo de episodios
  - Recompensas promedio
  - Convergencia
  - No requiere supuestos de Markov

Computar el Valor de un MRP

1. Puede ser estimado mediante Simulación
    - Propiedad de Markov genera estructura adicional
    - La función de valor de MRP satisface:

$$V(s) = \underbrace{R(s)}_{\text{Immediate reward}} + \underbrace{\gamma \sum_{s' \in S} P(s'|s)V(s')}_{\text{Discounted sum of future rewards}}$$

Computar el Valor de un MRP, Ecuación de Bellman

2. En forma Analitica
   - Para estados finitos en MRP, podemos expresar V(s) usando la ecuación de matrices:

$$\begin{pmatrix} V(s_1) \\ \vdots \\ V(s_N) \end{pmatrix} = \begin{pmatrix} R(s_1) \\ \vdots \\ R(s_N) \end{pmatrix} + \gamma \begin{pmatrix} P(s_1|s_1) & \cdots & P(s_N|s_1) \\ P(s_1|s_2) & \cdots & P(s_N|s_2) \\ \vdots & \ddots & \vdots \\ P(s_1|s_N) & \cdots & P(s_N|s_N) \end{pmatrix} \begin{pmatrix} V(s_1) \\ \vdots \\ V(s_N) \end{pmatrix}$$

$$V = R + \gamma P V$$

Computar el Valor de un MRP, Solución Analítica

2. En forma Analitica
   - Para estados finitos en MRP, podemos expresar V(s) usando la ecuación de matrices:

$$\begin{pmatrix} V(s_1) \\ \vdots \\ V(s_N) \end{pmatrix} = \begin{pmatrix} R(s_1) \\ \vdots \\ R(s_N) \end{pmatrix} + \gamma \begin{pmatrix} P(s_1|s_1) & \cdots & P(s_N|s_1) \\ P(s_1|s_2) & \cdots & P(s_N|s_2) \\ \vdots & \ddots & \vdots \\ P(s_1|s_N) & \cdots & P(s_N|s_N) \end{pmatrix} \begin{pmatrix} V(s_1) \\ \vdots \\ V(s_N) \end{pmatrix}$$

$$V = R + \gamma P V$$

$$V - \gamma P V = R$$

$$(I - \gamma P) V = R$$

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

Algoritmos Iterativo para computar Valor en MRP

3. Programación Dinámica

Initialize $V_0(s) = 0$ for all s

For $k = 1$ until convergence

- For all s in S

$$V_k(s) = R(s) + \gamma \sum_{s' \in S} P(s'|s) V_{k-1}(s')$$

Con esto, analizamos MDP

- Otra forma de mirar MDP es desde MRP + acciones
  - S es un conjunto (finito) de estados s ∈ S
  - A es un conjunto (finito) de acciones a ∈ A
  - P es un modelo dinámico o de transición para **cada acción**

    $$P(s_{t+1} = s' | s_t = s, a_t = a)$$

  - R es una función de recompensas

    $$R(s_t = s, a_t = a) = \mathbb{E}[r_t | s_t = s, a_t = a]$$

  - γ es el factor de descuento γ ∈ [0,1]
- MDP es una tupla (S,A,P,R, γ)

Mismo Ejemplo para MDP

MRP

s_1	s_2	s_3	s_4	s_5	s_6	s_7
						

$$P(s'|s, a_1) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \quad P(s'|s, a_2) = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Políticas del MDP

$\pi(s)$

s_1	s_2	......	s_n

- La Política especifica que acción tomar en cada estado
  - Puede ser determinística o estocástica
- Para generalización, se consideran distribuciones de probabilidades
  - Dado un estado, distribuciones por acciones
- Política:

$$\pi(a|s) = P(a_t = a | s_t = s)$$

$P(\pi(s')|s)$

$$\pi(a|s) = P(a|s)$$

MDP con Políticas

- MDP + $\pi(a|s)$ = MRP
- Específicamente, es un MRP (S,R^{π}, P^{π},γ) donde:

M1 → fn. dl recomp

M2 → Modelo de Tr

$$R^{\pi}(s) = \sum_{a \in A} \pi(a|s) R(s,a)$$

$$P^{\pi}(s'|s) = \sum_{a \in A} \pi(a|s) P(s'|s,a)$$

- Podemos entonces evaluar el valor de una política para un MDP

Evaluación de Políticas en MDP, usando Algoritmo Iterativo

Initialize $V_0(s) = 0$ for all s

For $k = 1$ until convergence

- For all s in S

$$V_k^{\pi}(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s'|s, \pi(s)) V_{k-1}^{\pi}(s')$$

- Esto se llama **Bellman Backup** para una política particular.

Evaluación de Políticas en MDP, Ejemplo

s_1	s_2	s_3	s_4	s_5	s_6	s_7
+1		0				

- Acá hay dos acciones discretas
- Recompensas: +1 en s1, +10 en s7, y 0 el resto
- Digamos $\pi(s) = a1$ $\forall s$. $\gamma = 0$
- ¿Cual es el valor de esta política?

0

$$V_k^{\pi}(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s'|s, \pi(s)) V_{k-1}^{\pi}(s')$$

$V_k^{\pi}(s) = r(s, a_1) \rightarrow 0$

Evaluación de Políticas en MDP, Ejemplo

- Dinámica: p(s6|s6,a1)=0.5, p(s7|s6,a1)=0.5
- Recompensas: para todas las acciones, +1 en s1, +10 en estado s7, 0 el resto
- Con $\pi(s) = a1 \;\; \forall s$, asumamos V_k = [1 0 0 0 0 0 10] y k=1, γ=0.5

For all s in S

$$V_k^\pi(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s'|s, \pi(s)) V_{k-1}^\pi(s')$$

$$= 0 + 0.5\,[\cancel{0.5 \cdot 0} + 0.5 \cdot 10]$$

$$= 0.5\,(5)$$

$$= 2.5$$

Control de MDPs

- Computar la Política Optimal

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

- También existe un valor optimal $\longrightarrow V^*$
- Política optimal es determinística

Preguntas

- Dado
  - 7 estados discretos $s_1, s_2, \ldots, s_7$
  - 2 acciones $\leftarrow, \rightarrow$
  - ¿Cuantas políticas determinísticas hay?

$$\begin{array}{ccccccccc} & s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 \\ \pi_1 = & \leftarrow & \rightarrow & \rightarrow & \rightarrow & \rightarrow & \rightarrow & \rightarrow \end{array} \qquad 2^7$$

$\pi_2 =$

  - ¿Es la política optimal única?

It. Valor $V^* \longrightarrow \pi^* \longrightarrow \boxed{\text{NO}}$

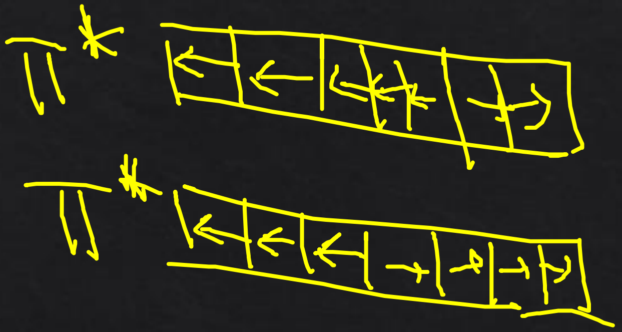
$\hookrightarrow$ It. Política

Control de MDPs

- Computar la Política Optimal

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

- Existe un valor optimal único
- Una Política Optimal para un MDP en un horizonte infinito es:
  - Determinística
  - Estacionaria
  - Pero no única, puede haber acciones-estados con valores optímales iguales



Cómo encontramos esta política?

Búsqueda de Políticas

"Métodos de Sol"

1
- Buscar la mejor política desde el valor
- El numero de políticas es $|A|^{|S|}$
- Iteración de Política es mucho más eficiente

Iteración de Política (PI)

Set $i = 0$

Initialize $\pi_0(s)$ randomly for all states s

While $i == 0$ or $\|\pi_i - \pi_{i-1}\|_1 > 0$ (L1-norm, measures if the policy changed for any state):

- $V^{\pi_i} \leftarrow$ MDP V function policy **evaluation** of π_i
- $\pi_{i+1} \leftarrow$ Policy **improvement**
- $i = i + 1$

Valores Estados-Acción Q

(s, a) Q-values

- El valor de un estado-acción de una política

$$Q^{\pi}(s,a) = R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^{\pi}(s')$$

- Toma acción a, luego sigue la política π

Mejora de Políticas

$(s, a) \rightarrow$ Qvalues

π^*

- Computar valor de estado-acción de una política π_i
  - Para s en S y a en A

$$Q^{\pi_i}(s, a) = R(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) V^{\pi_i}(s')$$

$|a|^{|s|}$

s_1, a_1

s_1, a_2

- Computar una nueva política π_{i+1}, para cada s ∈ S

$$\pi_{i+1}(s) = \arg\max_a Q^{\pi_i}(s, a) \ \forall s \in S$$

Mejora de Políticas

$$Q^{\pi_i}(s,a) = R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^{\pi_i}(s')$$

Evaluación de Política con Prog. Dinámica

Initialize $V_0^\pi(s) = 0$ for all s

For $k = 1$ until convergence

- For all s in S

$$V_k^\pi(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s'|s, \pi(s)) V_{k-1}^\pi(s')$$

Ec. Bellman

Evaluación de Política con Prog. Dinámica

Initialize $V_0^\pi(s) = 0$ for all s

For $k = 1$ until convergence

- For all s in S

$$\boxed{V_k^\pi(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s'|s, \pi(s)) V_{k-1}^\pi(s')}$$

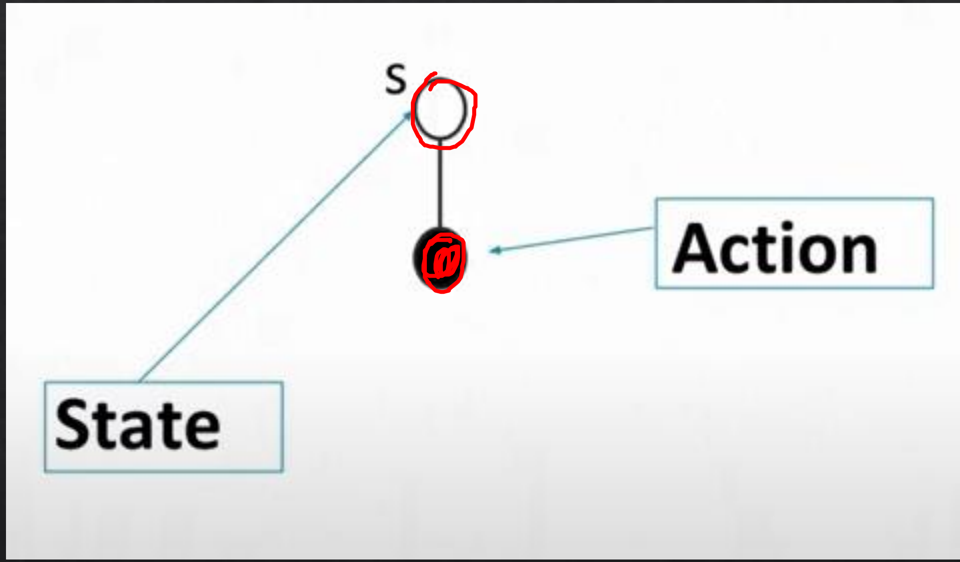
$V_k^\pi(s)$ is exact value of k-horizon value of state s under policy π

$V_k^\pi(s)$ is an estimate of infinite horizon value of state s under policy π

$$\boxed{V^\pi(s) = \mathbb{E}_\pi[G_t|s_t = s]} \approx \mathbb{E}_\pi[r_t + \gamma V_{k-1}|s_t = s]$$

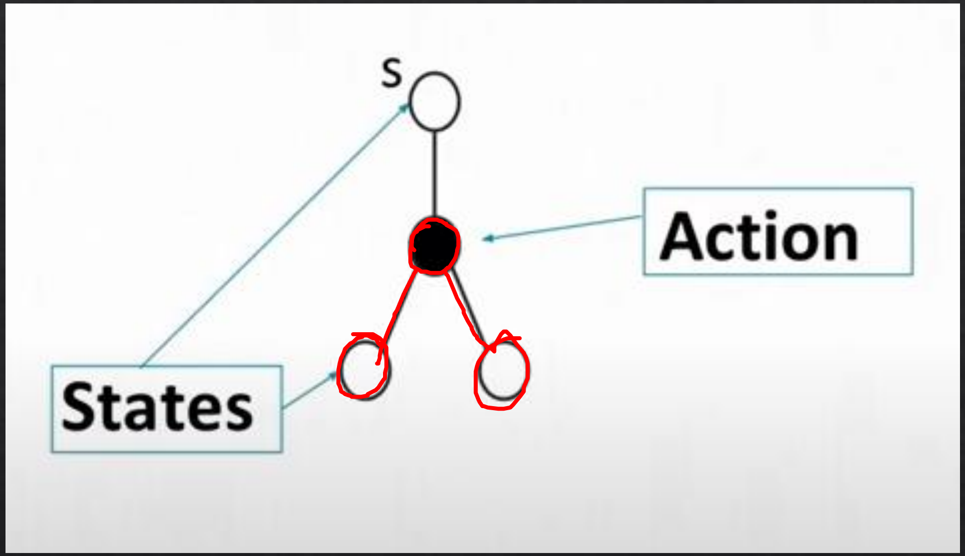
Evaluación de Política con Prog. Dinámica

$$V^{\pi}(s) \leftarrow \mathrm{E}_{\pi}[r_t + \gamma V_{k-1} | s_t = s]$$



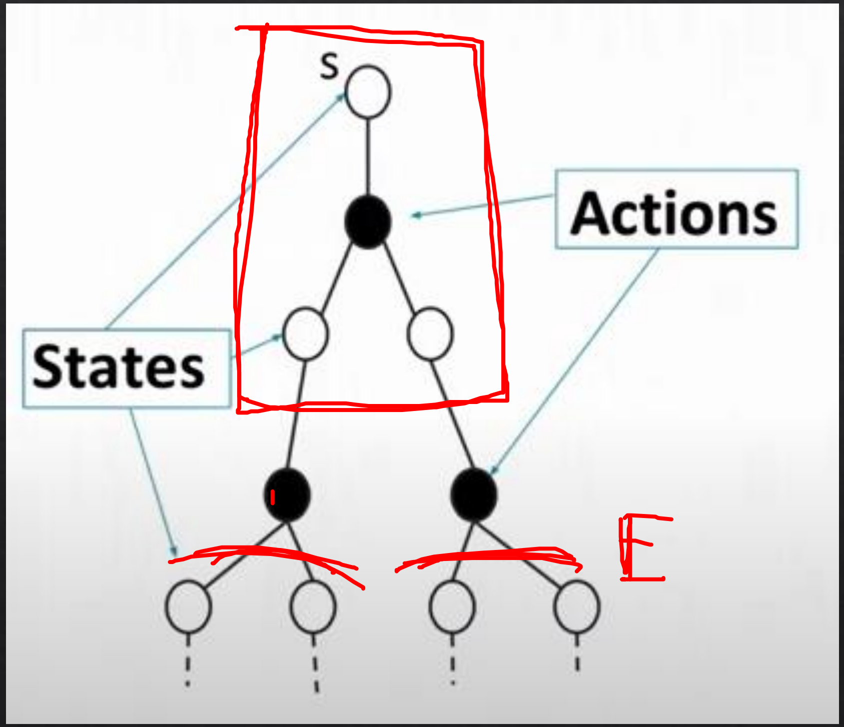
Evaluación de Política con Prog. Dinámica

$$V^{\pi}(s) \leftarrow \mathrm{E}_{\pi}[r_t + \gamma V_{k-1} | s_t = s]$$



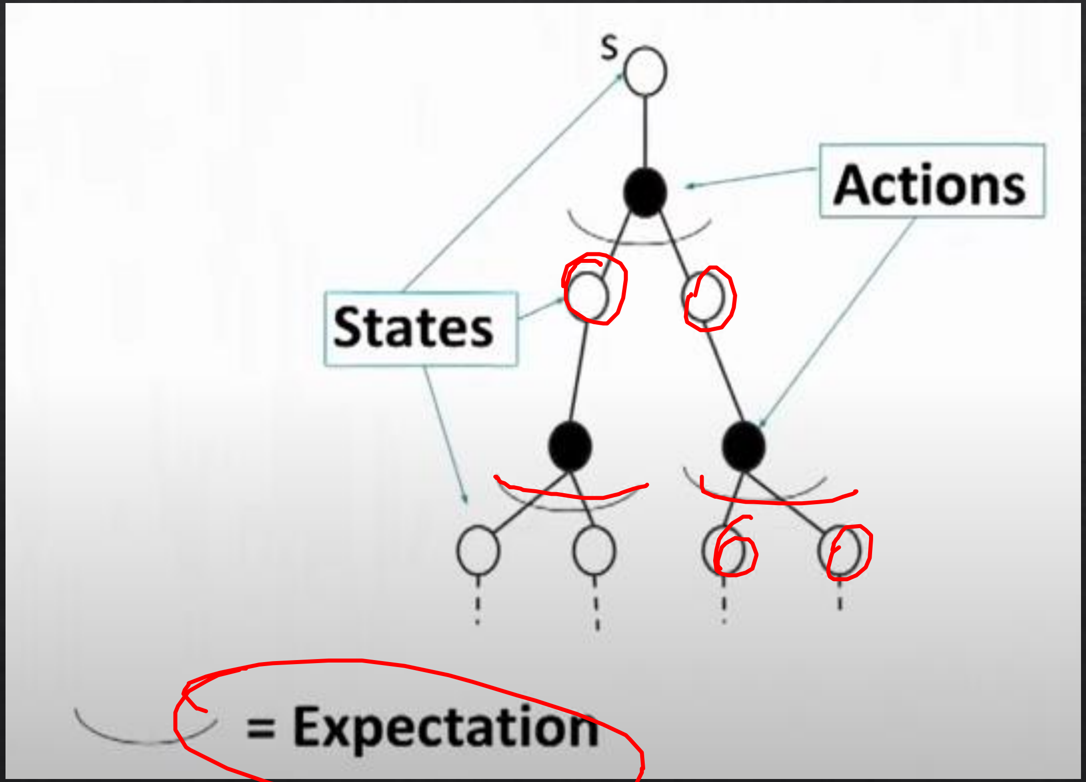
Evaluación de Política con Prog. Dinámica

$$V^{\pi}(s) \leftarrow \mathrm{E}_{\pi}[r_t + \gamma V_{k-1} | s_t = s]$$



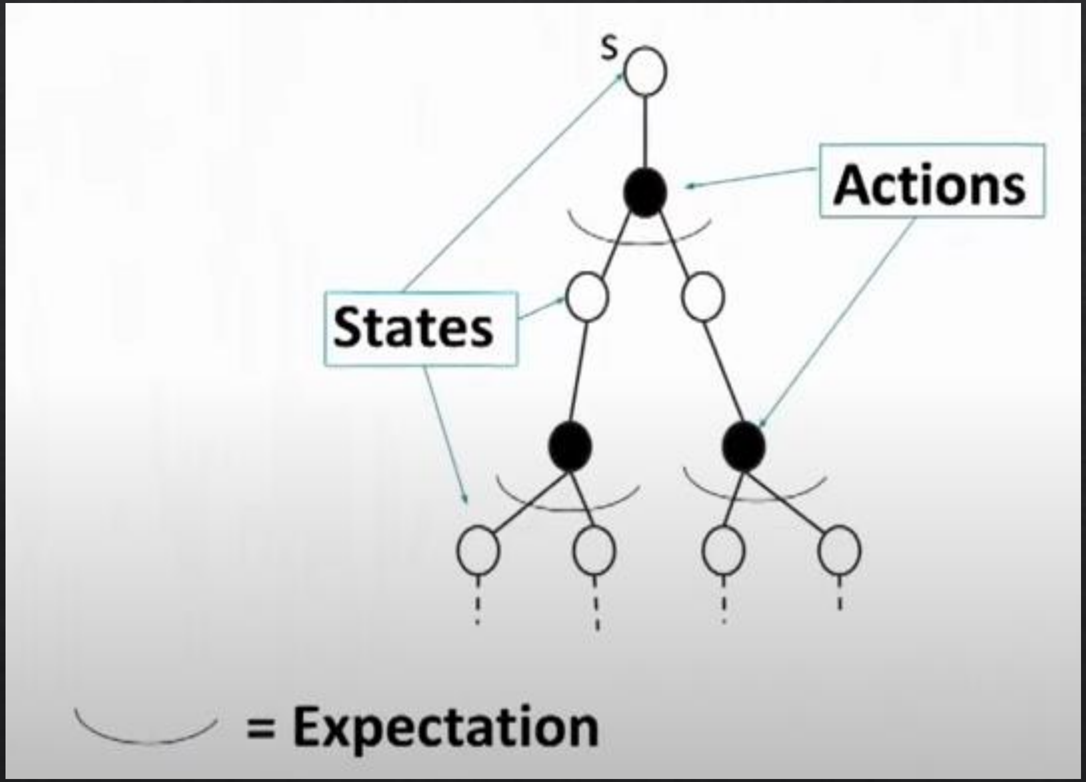
Evaluación de Política con Prog. Dinámica

$$V^{\pi}(s) \leftarrow \mathrm{E}_{\pi}[r_t + \gamma V_{k-1} | s_t = s]$$



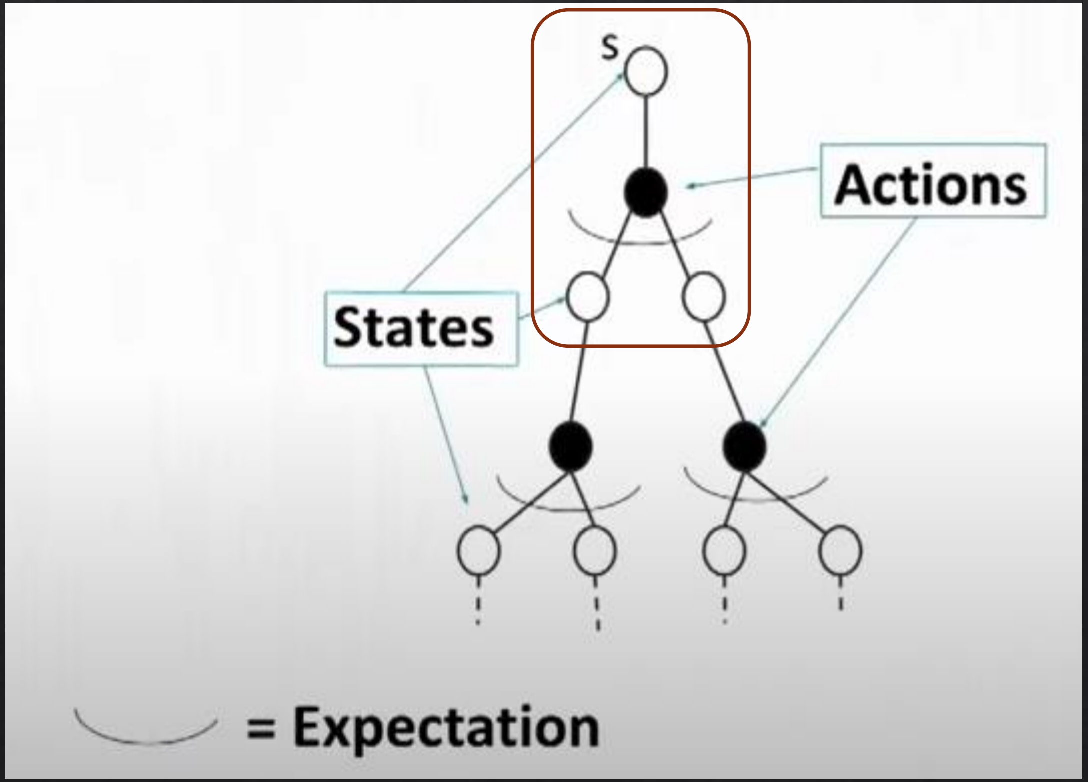
Evaluación de Política con Prog. Dinámica

$$V^{\pi}(s) \leftarrow \mathrm{E}_{\pi}[r_t + \gamma V_{k-1} | s_t = s]$$



Evaluación de Política con Prog. Dinámica

$$V^{\pi}(s) \leftarrow \mathrm{E}_{\pi}[r_t + \gamma V_{k-1} | s_t = s]$$



$$V_k^{\pi}(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s'|s, \pi(s)) V_{k-1}^{\pi}(s')$$

sabemos un modelo tipo $P(s'|s, a)$

En Resumen,

$$G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \cdots$$

Programación Dinámica:

$$V^\pi(s) \approx \mathbb{E}_\pi[r_t + \gamma V_{k-1} | s_t = s]$$

- Requiere de un Modelo MDP M
- Usa estimación de valores para recompensas futuras
- Requiere supuestos markovianos

Libre de Modelo (Model-free)
1. Evaluación de Política Montecarlo (MC)

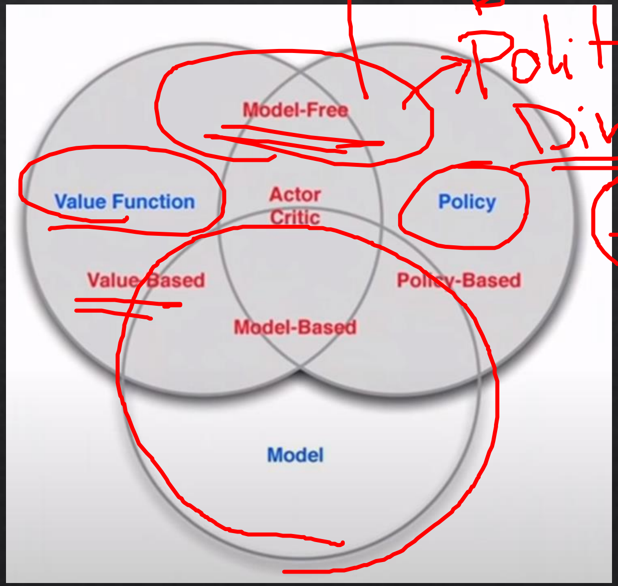
$$G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \cdots$$
$$V^\pi(s) = \mathbb{E}_{T\sim\pi}[G_t | s_t = s]$$

- Valor esperado sobre trayectorias T generadas siguiendo π
- Valor = ganancia promedio ✓
- No necesita dinámica Modelo Dinámico $P(s'|s,a)$ ht
- No asume estados markovianos
- Se usa para MDP episódicos
  - Promediar ganancia de un episodio completo $[T] \to t=1,\ldots,T$
  - Requiere que cada episodio termine H

$V \to \pi$

Busqueda Política Directa

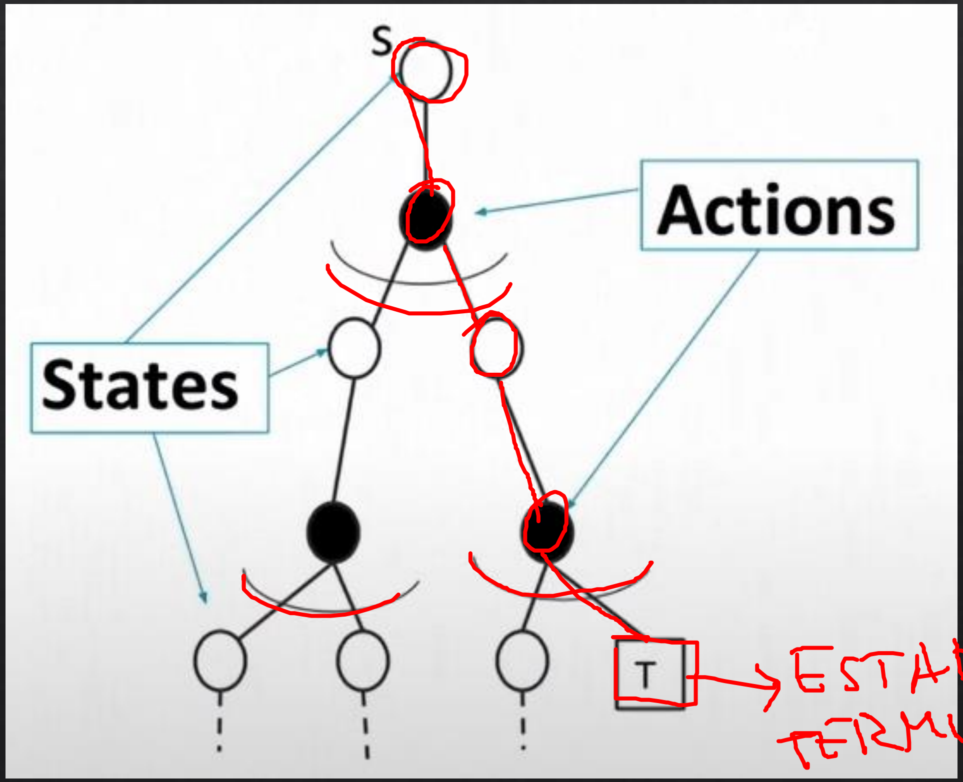
5.2



Libre de Modelo (Model-free)
1. Evaluación de Política Montecarlo (MC)

$$V^{\pi}(s) = V^{\pi}(s) + \alpha(G_{i,t} - V^{\pi}(s))$$

$= \frac{1}{n}$



ESTADO TERMINAL

Libre de Modelo (Model-free)
1. Aprendizaje con Diferencias Temporales

- Es libre de modelo
- Ocupa programación dinámica y Montecarlo