



## Exemplo: Ambiente 7 x 1

s	1	2	3	4	5	6	7
	-						+

**Perda** **Ganho**

  Estados terminais  
(não admitem transições)

Modelo do mundo para um *Processo de Decisão de Markov* (PDM)

$S$  – conjunto de estados do mundo

$A(s)$  – conjunto de acções possíveis no estado  $s \in S$

$T(s, a, s')$  – probabilidade de transição de  $s$  para  $s'$  através de  $a$

$R(s, a, s')$  – recompensa esperada na transição de  $s$  para  $s'$  através de  $a$

<b>s</b>	1	2	3	4	5	6	7
	-						+

Acções possíveis	A(s)
1. Não fazer nada	1. Não fazer nada
2. Fazer o teste	2. Fazer o teste
3. Fazer o teste e, se for positivo, fazer o tratamento	3. Fazer o teste e, se for positivo, fazer o tratamento
4. Fazer o teste e, se for positivo, não fazer o tratamento	4. Fazer o teste e, se for positivo, não fazer o tratamento
5. Fazer o teste e, se for negativo, fazer o tratamento	5. Fazer o teste e, se for negativo, fazer o tratamento
6. Fazer o teste e, se for negativo, não fazer o tratamento	6. Fazer o teste e, se for negativo, não fazer o tratamento

**A(s)**  
←  
→

$$T(s,a,s') = 1; s, s' \in S$$

	T(s,a,s')																		
<b>s</b>	1		2		3		4		5		6		7						
<b>a</b>	←	→	←	→	←	→	←	→	←	→	←	→	←	→	←	→			
<b>s'</b>	1	2	1	3	2	4	3	5	4	6	5	7	6	7					
	0	0	1	1	1	1	1	1	1	1	1	1	0	0					

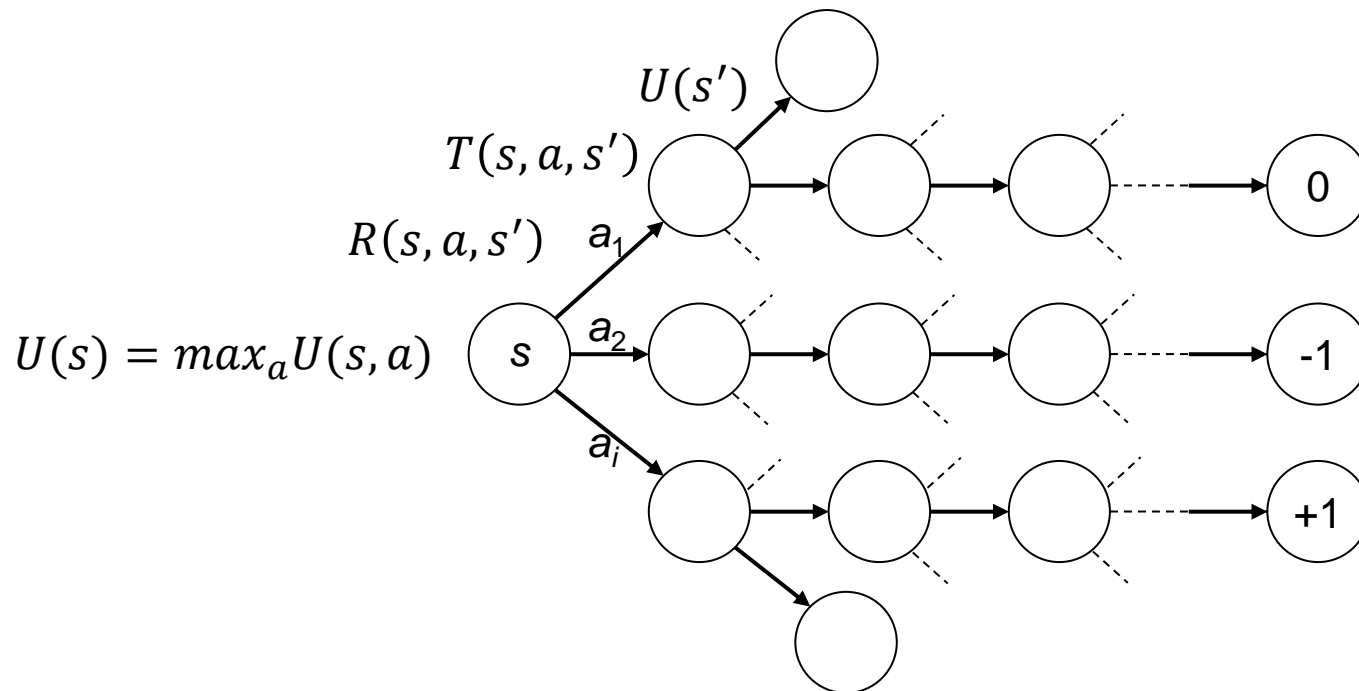
		R(s,a,s')													
		1		2		3		4		5		6		7	
s	a														
		←	→	←	→	←	→	←	→	←	→	←	→	←	→
s'		1	2	1	3	2	4	3	5	4	6	5	7	6	7
		0	0	-1	0	0	0	0	0	0	0	0	1	0	0

# Exemplo: Ambiente 7 x 1

s	1	2	3	4	5	6	7
	-						+

Perda

Ganho



$$U(s, a) = \sum_{s' \in \text{suc}(s)} T(s, a, s') [R(s, a, s') + \gamma U(s')]$$

Exemplo: Ambiente 7 x 1

s	1	2	3	4	5	6	7
	-						+

Ambiente determinista  
 $T(s,a,s') = 1; s, s' \in S$

Estados possíveis  
 $S = \{1, 2, 3, 4, 5, 6, 7\}$

Acções possíveis	A(s)
	←
	→

Cálculo da utilidade

$\gamma$	$\Delta_{max}$
0,5	0,0

$$U(s) = \max_a U(s, a)$$
$$U(s, a) = \sum_{s' \in suc(s)} T(s, a, s') [R(s, a, s') + \gamma U(s')]$$

Modelo de transição

T(s,a,s')														
s	1		2		3		4		5		6		7	
a	←	→	←	→	←	→	←	→	←	→	←	→	←	→
s'	1	2	1	3	2	4	3	5	4	6	5	7	6	7
	0	0	1	1	1	1	1	1	1	1	1	1	0	0

Modelo de recompensa

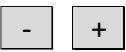
		R(s,a,s')																									
s		1		2		3		4		5		6		7													
a		←	→	←	→	←	→	←	→	←	→	←	→	←	→	←	→										
s'		1	2	1	3	2	4	3	5	4	6	5	7	6	7												
		0	0	-1	0	0	0	0	0	0	0	0	1	0	0												

Iteração	s	1	2	3	4	5	6	7
	U	0,000	0,000	0,000	0,000	0,000	0,000	0,000
1	U	0,000	0,000	0,000	0,000	0,000	1,000	0,000
	δ	0,000	0,000	0,000	0,000	0,000	1,000	0,000
2	U	0,000	0,000	0,000	0,000	0,500	1,000	0,000
	δ	0,000	0,000	0,000	0,000	0,500	0,000	0,000
3	U	0,000	0,000	0,000	0,250	0,500	1,000	0,000
	δ	0,000	0,000	0,000	0,250	0,000	0,000	0,000
4	U	0,000	0,000	0,125	0,250	0,500	1,000	0,000
	δ	0,000	0,000	0,125	0,000	0,000	0,000	0,000
5	U	0,000	0,063	0,125	0,250	0,500	1,000	0,000
	δ	0,000	0,063	0,000	0,000	0,000	0,000	0,000
6	U	0,000	0,063	0,125	0,250	0,500	1,000	0,000
	δ	0,000	0,000	0,000	0,000	0,000	0,000	0,000

Utilidade final

# Exemplo: Ambiente 7 x 1

s	1	2	3	4	5	6	7
	-						+



Estados terminais  
(não admitem transições)

$\gamma$
0,5

Utilidade final

U	0,000	0,063	0,125	0,250	0,500	1,000	0,000
$\delta$	0,000	0,000	0,000	0,000	0,000	0,000	0,000



$$\pi^*(s) = \operatorname{argmax}_a U(s, a)$$

$$U(s, a) = \sum_{s' \in \text{Suc}(s)} T(s, a, s') [R(s, a, s') + \gamma U(s')]$$

Ambiente determinista:  
Cada acção só tem um estado  
sucessor com  $T(s, a, s') = 1$

Ambiente determinista

$$T(s, a, s') = 1; s, s' \in S$$

Estados possíveis

$$S = \{1, 2, 3, 4, 5, 6, 7\}$$

Acções  
possíveis

A(s)
←
→

Modelo de transição

	T(s,a,s')													
s	1		2		3		4		5		6		7	
a	←	→	←	→	←	→	←	→	←	→	←	→	←	→
s'	1	2	1	3	2	4	3	5	4	6	5	7	6	7
	0	0	1	1	1	1	1	1	1	1	1	1	0	0

Modelo de recompensa

	R(s,a,s')													
s	1		2		3		4		5		6		7	
a	←	→	←	→	←	→	←	→	←	→	←	→	←	→
s'	1	2	1	3	2	4	3	5	4	6	5	7	6	7
	0	0	-1	0	0	0	0	0	0	0	0	1	0	0

Cálculo da política

	Política													
s	1		2		3		4		5		6		7	
a			←	→	←	→	←	→	←	→	←	→		
U(s,a)			-1,00	0,063	0,031	0,125	0,063	0,250	0,125	0,500	0,250	1,000		
a*			→	→	→	→	→	→	→	→	→	→		

Política óptima