

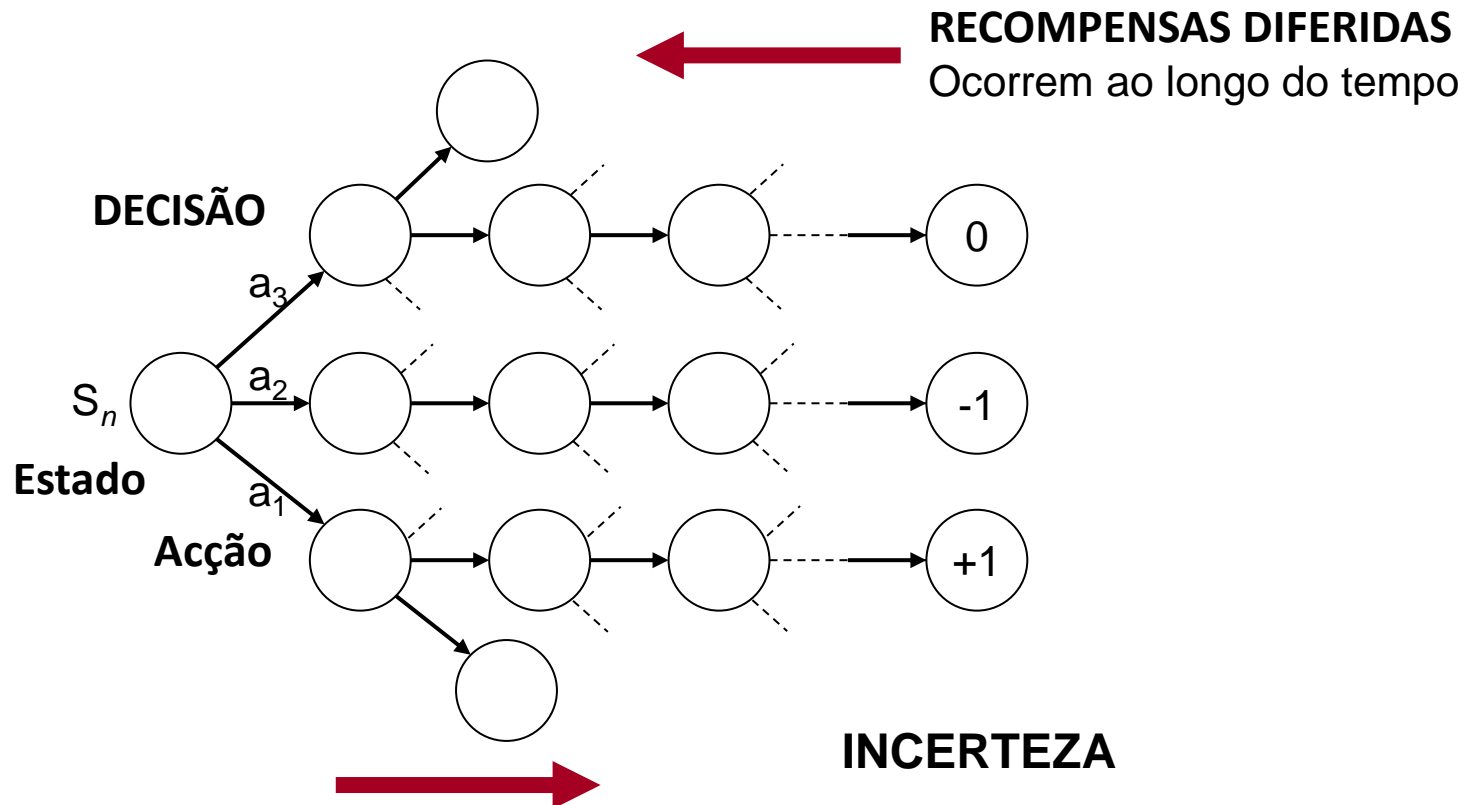
PROCESSOS DE DECISÃO SEQUENCIAL

Luís Morgado

2024

Problemas de Decisão Sequencial

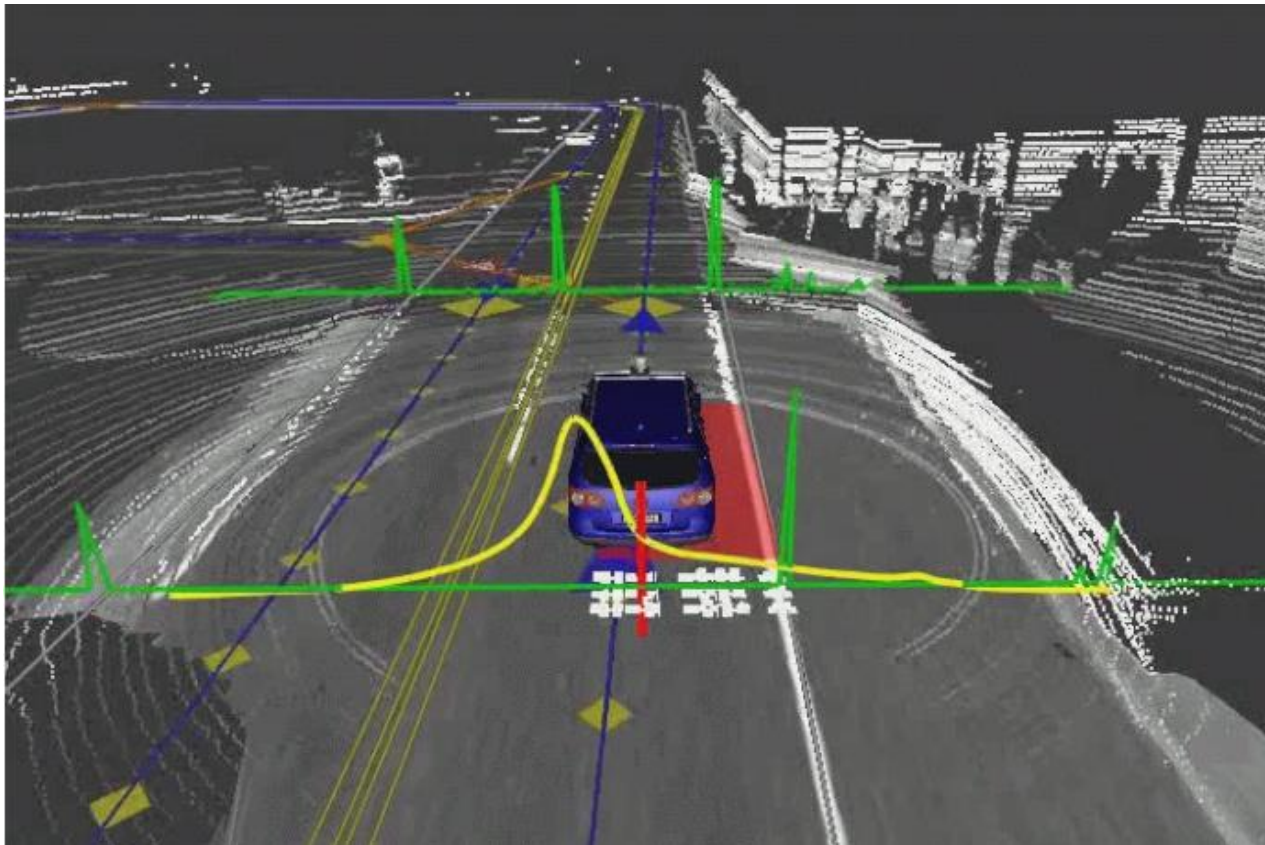
O valor das acções de um agente não depende de decisões simples, baseadas no estado actual, mas de uma sequência de acções encadeadas no tempo, podendo os resultados das acções ser incertos, ou seja, não totalmente controlados (*não determinísticos*)



Raciocínio com Incerteza

A incerteza resulta da impossibilidade de se obter informação completa relativa ao domínio do problema

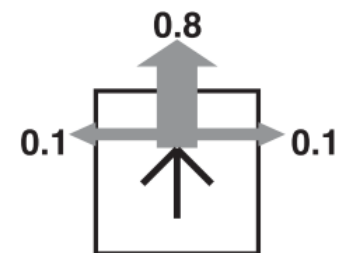
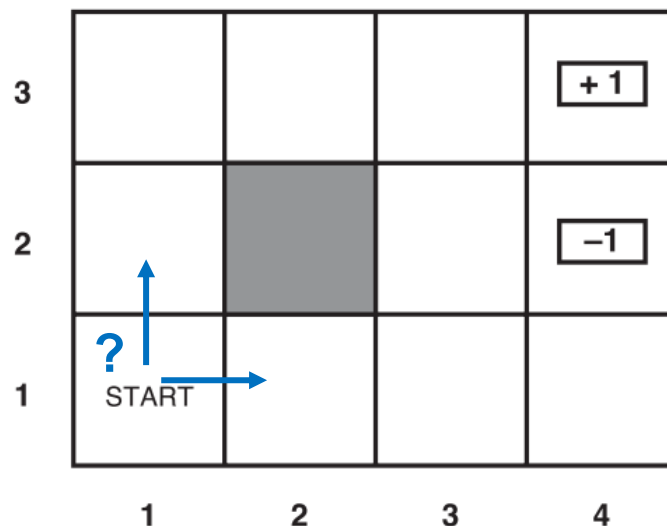
Exemplo: Navegação em veículos autónomos



Processos de Decisão Sequencial

- Problema da decisão ao longo do tempo
 - Valor de uma acção depende de uma sequência de decisões
 - Possibilidade de ganhos e perdas
 - Incerteza na decisão
 - Efeito cumulativo

Exemplo:

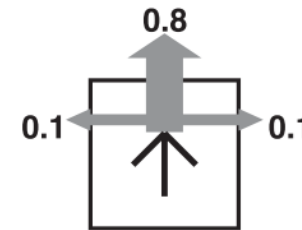
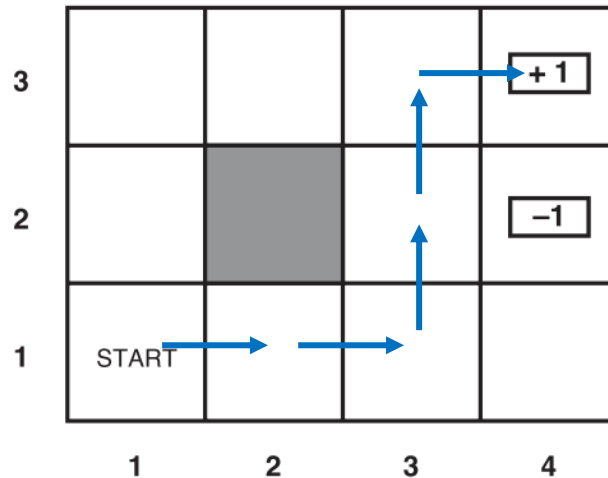


Desvios possíveis

**Incerteza no resultado
de uma acção**

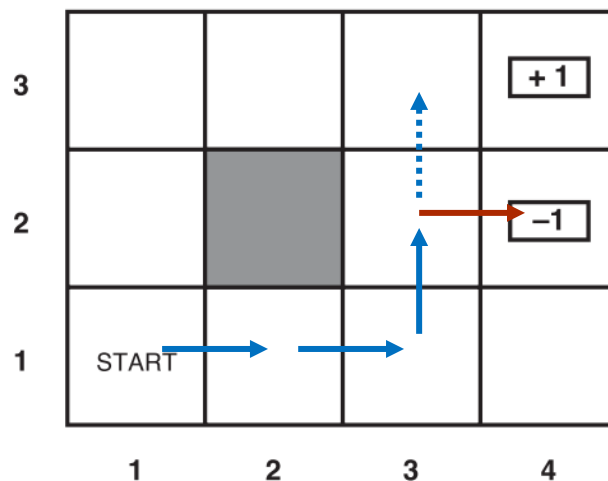
Processos de Decisão Sequencial

Exemplo:



Desvios possíveis

**Incerteza no resultado
de uma acção**



O efeito da incerteza no resultado da acção, resulta na possibilidade de **transições de estado não deterministas**, ou seja, sobre as quais o sistema não têm controlo total

Processos de Decisão Sequencial

Num processo de decisão sequencial, a evolução entre estados ocorre por efeito de acções que, no caso geral, podem ser *não-deterministas*, ou seja, **o seu resultado pode não ser completamente previsível**, pode existir incerteza no resultado da acção (*não determinismo* da acção).

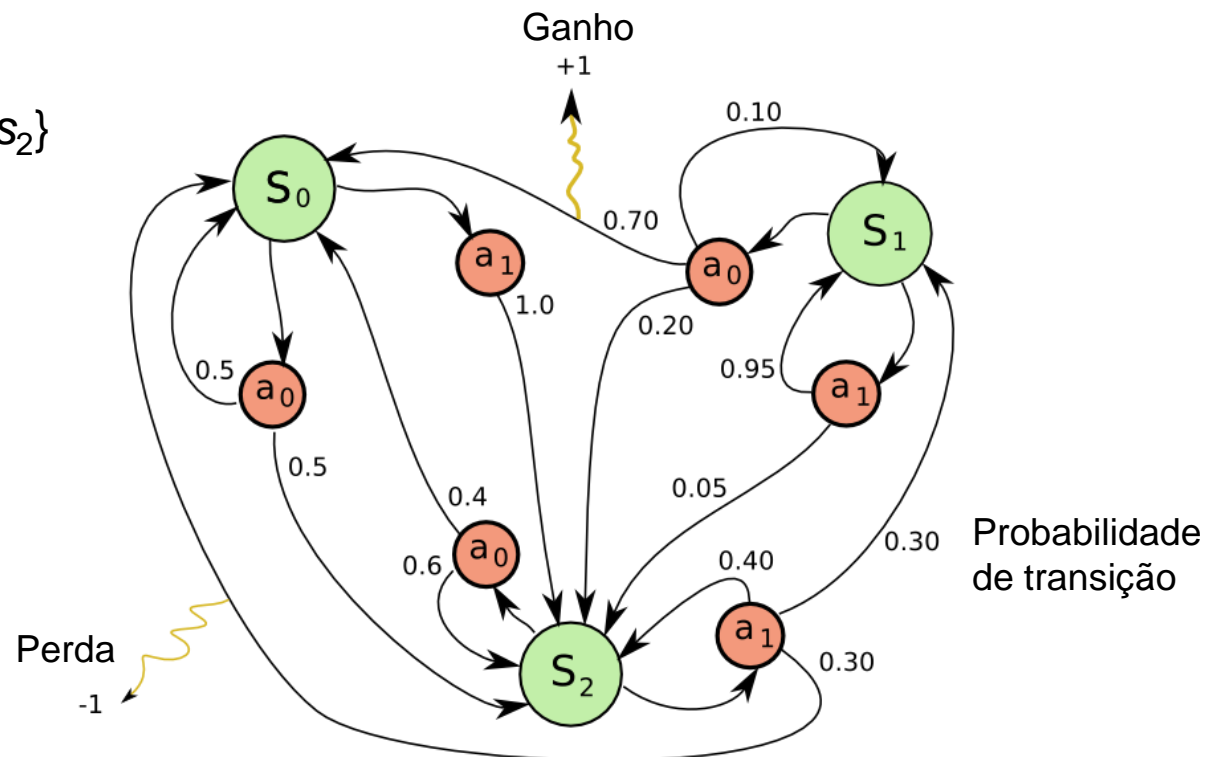
A este tipo de processos correspondem *espaços de estados não-deterministas*, nos quais as transições entre estados ocorrem por efeito das acções, com uma determinada **probabilidade de transição**.

A cada transição pode estar associada uma **recompensa**, que representa o **ganho ou perda** associado a essa transição de estado.

Exemplo:

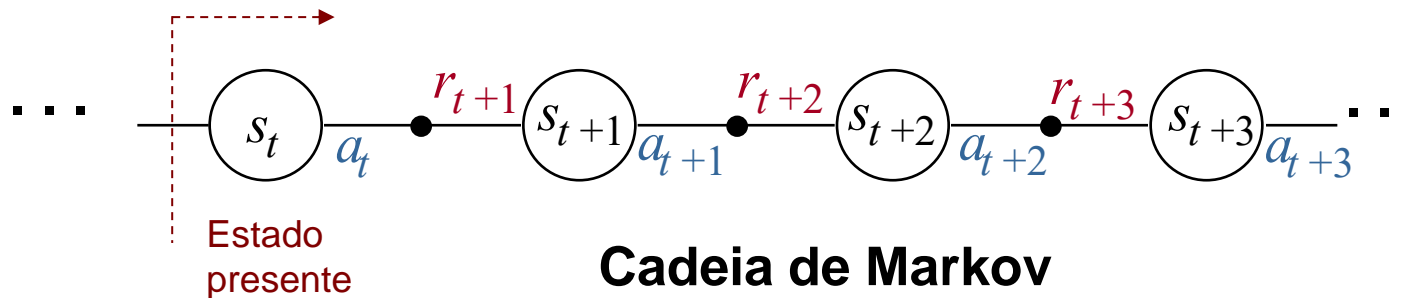
Estados = $\{s_0, s_1, s_2\}$

Acções = $\{a_0, a_1\}$



Propriedade de Markov

- Andrey Markov
 - Matemático Russo (1856 – 1922)
- Um processo estocástico tem a ***propriedade de Markov*** se a distribuição probabilística condicional dos **estados futuros** de um processo depender exclusivamente do **estado presente**
- **A previsão dos estados seguintes só depende do estado presente**



Cadeia de Markov

Sequência de evolução ao longo dos estados possíveis
(estado, acção, recompensa)

Processos de Decisão de Markov

Representação do mundo sob a forma de PDM

S – conjunto de estados do mundo

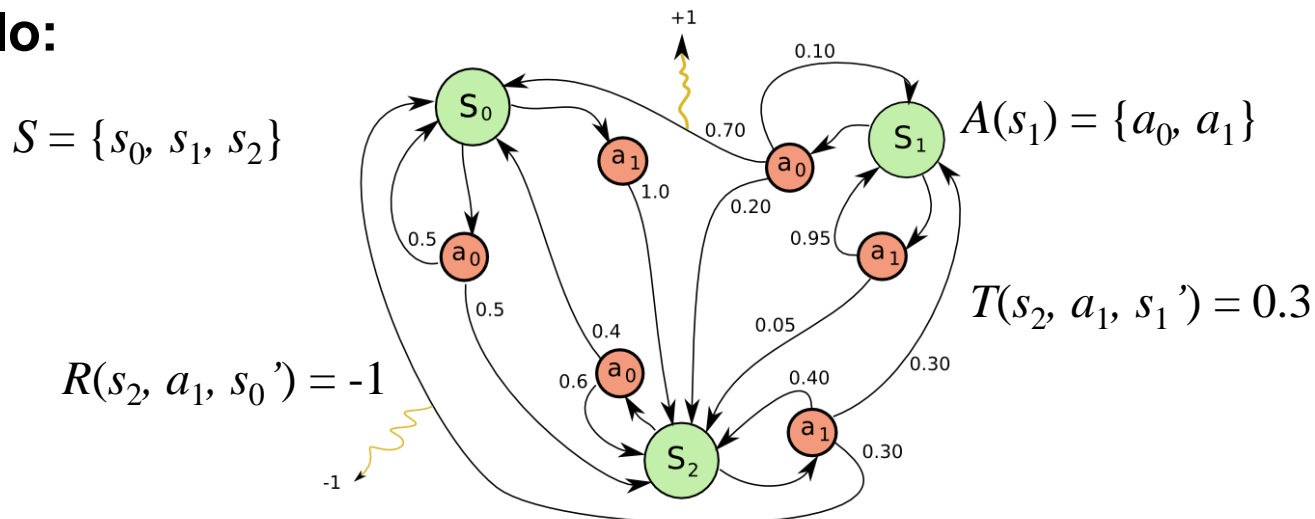
$A(s)$ – conjunto de acções possíveis num estado $s \in S$

$T(s, a, s')$ – 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

$t = 0, 1, 2, \dots$ – tempo discreto

Exemplo:



Tomada de Decisão Sequencial

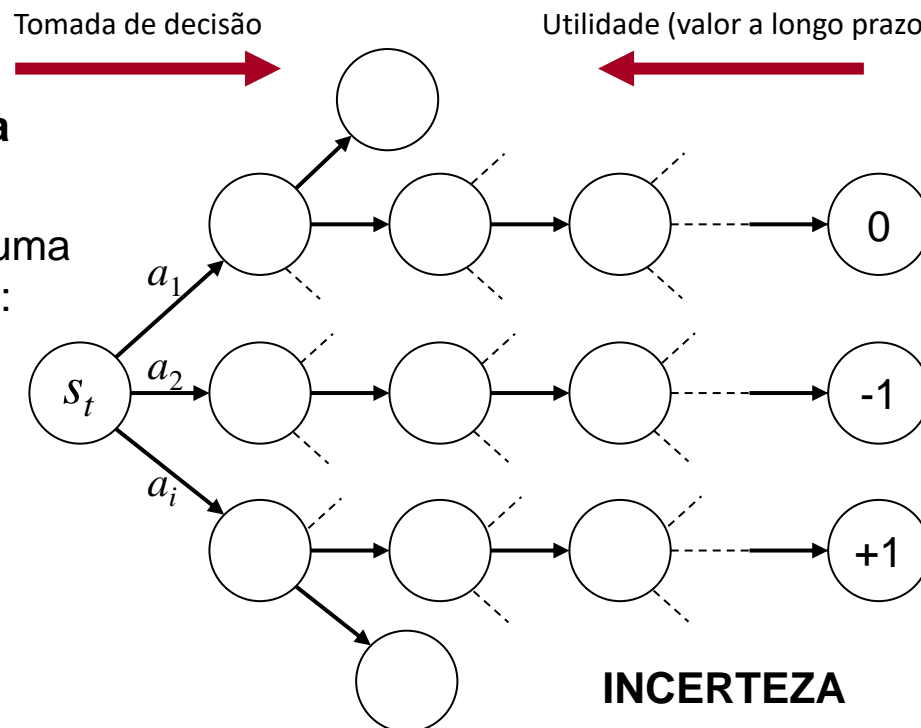
Num processo de decisão geral, para cada estado s_n , cada decisão de acção (a_1, a_2, \dots, a_i) pode levar a **múltiplas ramificações de estados e decisões futuras**

O **valor** (*utilidade*) desses estados e decisões pode não ser conhecido de imediato, acontecendo apenas **diferido no tempo**, em função dos encadeamentos de estados e decisões futuras

No caso geral, a própria evolução dos estados observados em função das acções realizadas pode não ser determinista, ou seja, estar sujeita a **incerteza**

Como decidir qual a acção a realizar?

É necessário definir uma medida de desempenho: *utilidade* (valor)

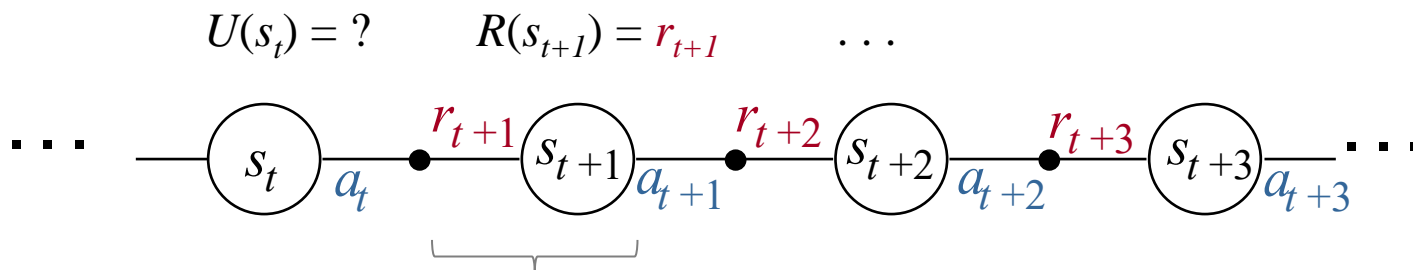


Utilidade (Valor) de Estado

Medida de desempenho para tomada de decisão que reflecte o efeito cumulativo da evolução de estado através das **recompensas** observadas, expressando **ganhos** e **perdas**

- **Recompensa**

- Valor finito positivo (**ganho**) ou negativo (**perda**)
- Expressa ganho ou perda num determinado estado
- Diferentes formas de recompensa
 - $R(s)$ – Recompensa depende apenas do estado
 - $R(s, a)$ – Recompensa depende do estado e da acção realizada
 - $R(s, a, s')$ – Recompensa depende do estado, da acção e do estado seguinte
- Tem um âmbito **local (curto prazo)**



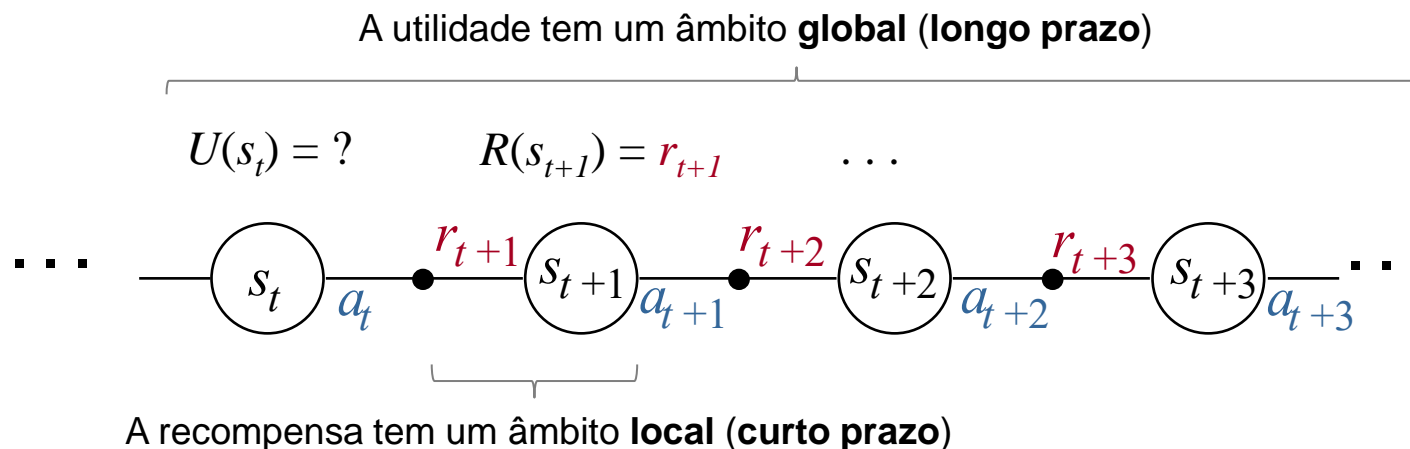
A recompensa tem um âmbito **local (curto prazo)**

Utilidade (Valor) de Estado

Medida de desempenho para tomada de decisão que reflecte o efeito cumulativo da evolução de estado através das **recompensas** observadas, expressando **ganhos e perdas**

- **Utilidade**

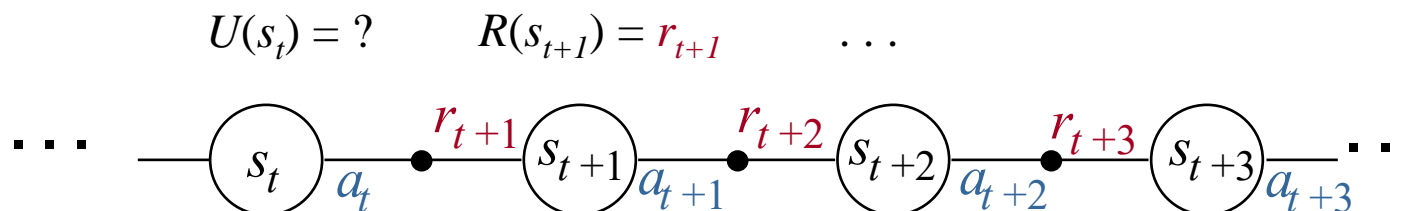
- $U(s)$
- Valor de um determinado estado s para realização do objectivo pretendido
- Reflecte o valor que pode ser obtido a partir dum estado ao longo de uma sequência de evolução de estado
- Tem um âmbito **global (longo prazo)**



Utilidade (Valor) de Estado

Cálculo da utilidade

(considerando que a recompensa depende apenas do estado)



A utilidade de um estado representa o valor que pode ser obtido a partir desse estado ao longo de uma sequência de evolução de estado

Uma forma possível de cálculo de utilidade é realizando a **soma das recompensas** ao longo da sequência de estados:

$$U(s_t) = R(s_t) + R(s_{t+1}) + R(s_{t+2}) + \dots$$

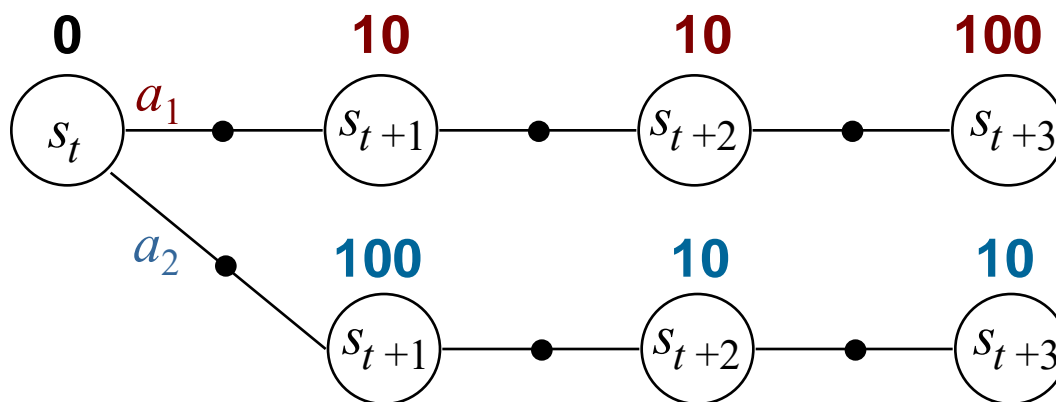
Utilidade (Valor) de Estado

Soma das recompensas

$$U(s_t) = R(s_t) + R(s_{t+1}) + R(s_{t+2}) + \dots$$

Exemplo 1:

Qual a melhor decisão no estado s_t , realizar a_1 ou realizar a_2 ?



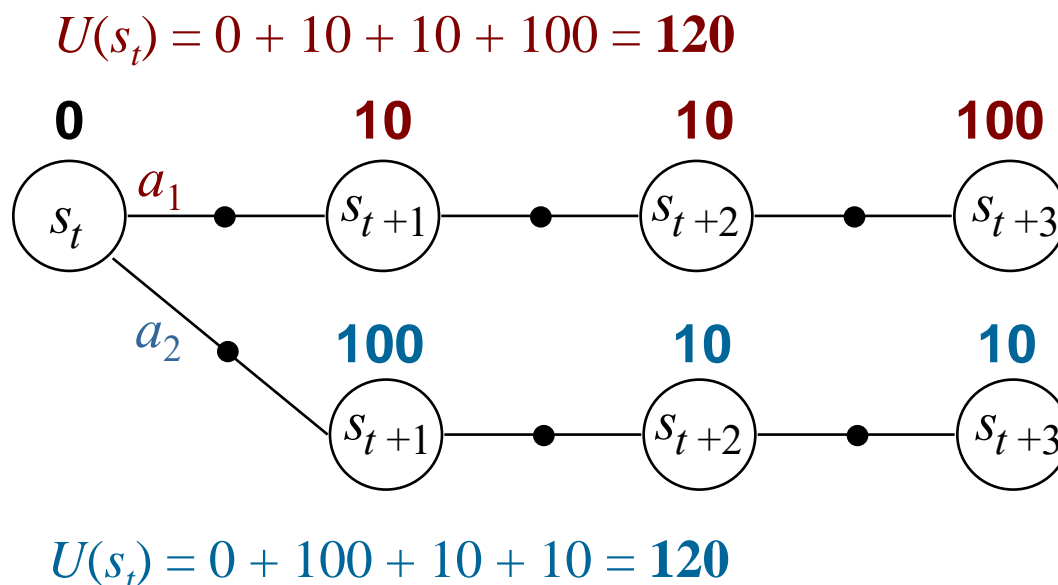
Utilidade (Valor) de Estado

Soma das recompensas

$$U(s_t) = R(s_t) + R(s_{t+1}) + R(s_{t+2}) + \dots$$

Exemplo 1:

Ambas as acções a_1 e a_2 tem a mesma utilidade, pelo que não existe preferência por qualquer das acções



Utilidade (Valor) de Estado

Soma das recompensas

$$U(s_t) = R(s_t) + R(s_{t+1}) + R(s_{t+2}) + \dots$$

- Problemas
 - Recompensas estão limitadas a uma gama finita de valores, caso contrário **a soma das recompensas pode ser infinita**
 - Impossibilita a decisão
 - Não reflecte o efeito da passagem do **tempo**
 - As recompensas mantêm o seu valor independente do momento em que ocorram

Utilidade (Valor) de Estado

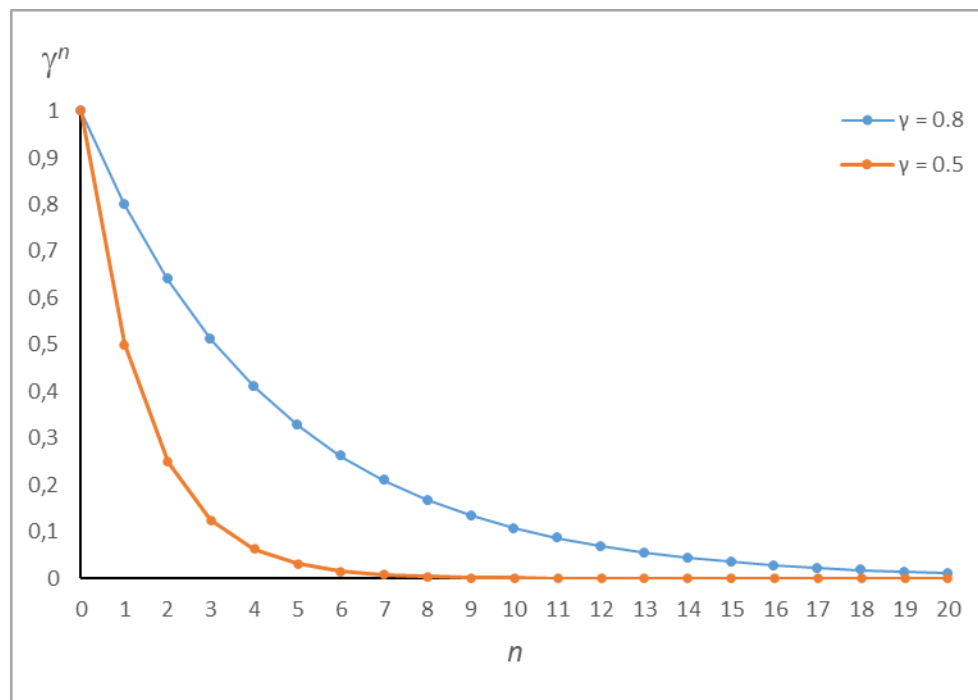
Recompensas descontadas (no tempo)

$$U(s_t) = R(s_t) + \gamma R(s_{t+1}) + \gamma^2 R(s_{t+2}) + \dots$$

Factor de desconto

$$\gamma \in [0,1]$$

- Recompensas não estão limitadas a uma gama finita de valores
- O factor de desconto reflecte o efeito da passagem do tempo
 - O momento em que a recompensa ocorre é relevante para a sua contribuição para a utilidade (valor) a longo prazo



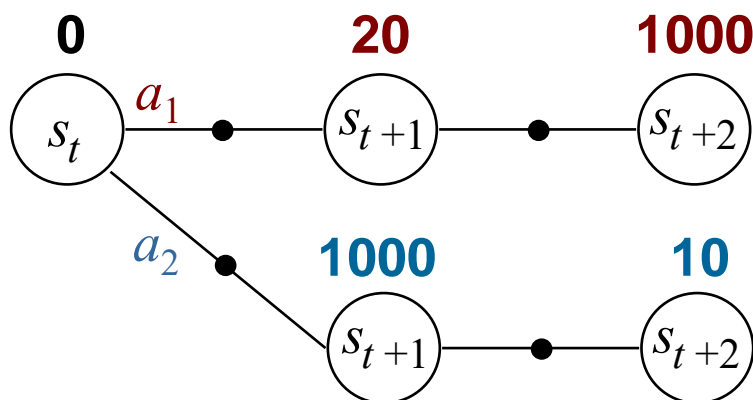
Utilidade (Valor) de Estado

Soma das recompensas

$$U(s_t) = R(s_t) + R(s_{t+1}) + R(s_{t+2}) + \dots$$

Exemplo 2:

Qual a melhor decisão no estado s_t , realizar a_1 ou realizar a_2 ?



Utilidade (Valor) de Estado

Recompensas descontadas (no tempo)

$$U(s_t) = R(s_t) + \gamma R(s_{t+1}) + \gamma^2 R(s_{t+2}) + \dots$$

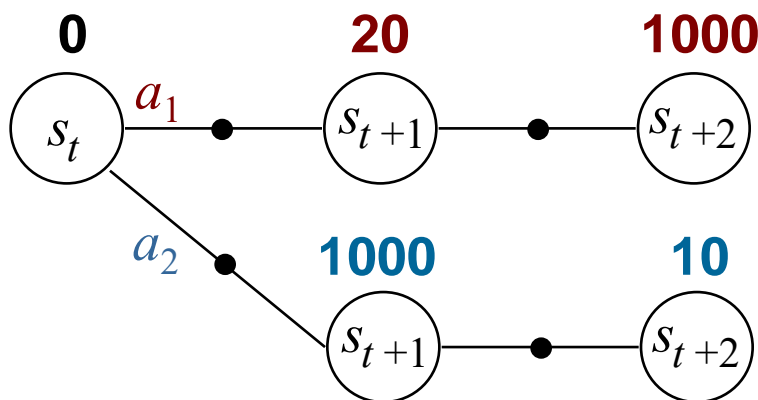
Exemplo 2:

A opção a_2 tem utilidade superior à opção a_1 pelo que a decisão racional (maximiza a utilidade) é realizar a_2

Factor de
desconto

$$\gamma = 0.5$$

$$U(s_t) = 0 + 0.5 * 20 + 0.5^2 * 1000 = 0 + 10 + 250 = \mathbf{260}$$



$$U(s_t) = 0 + 0.5 * 1000 + 0.5^2 * 10 = 0 + 500 + 2,5 = 502,5$$

Política (de tomada de decisão)

Função que define qual a acção que deve ser realizada em cada estado (*estratégia de acção*)

Política **determinista**

$$\pi : S \rightarrow A(s) ; s \in S$$

Para cada estado indica uma acção específica a realizar

Exemplo:

$$\pi(s_1) = a_1$$

Política **não determinista**

$$\pi: S \times A(s) \rightarrow [0,1] ; s \in S$$

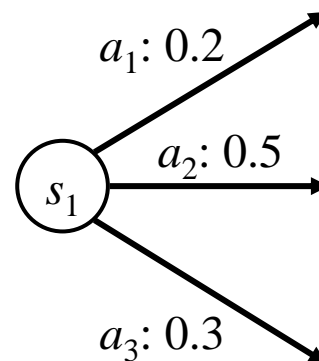
Para cada par (estado, acção) indica a probabilidade de no estado realizar a acção

Exemplo:

$$\pi(s_1, a_1) = 0.2$$

$$\pi(s_1, a_2) = 0.5$$

$$\pi(s_1, a_3) = 0.3$$



Política (de tomada de decisão)

Exemplo:

3	0.812	0.868	0.918	+ 1
2	0.762		0.660	-1
1	0.705	0.655	0.611	0.388
	1	2	3	4

Utilidade de estado para o ambiente
4 x 3, com $\gamma = 1$ e $R(s) = -0.04$ para
estados não terminais

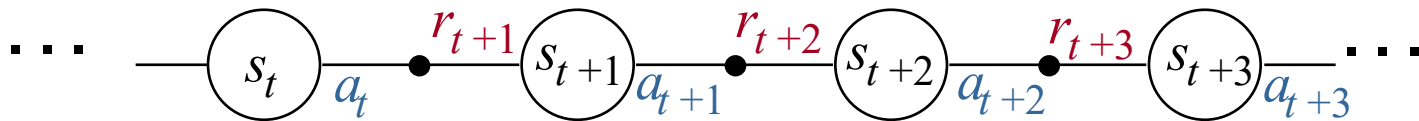
3	0.812	0.868	0.918	+ 1
2	0.762		0.660	-1
1	0.705	0.655	0.611	0.388
	1	2	3	4

Política ótima para a utilidade de
estado calculada

Utilidade (Valor) de Estado

Cálculo da utilidade de estado

$$U(s_t) = ? \quad R(s_t, a_t, s_{t+1}) = r_{t+1} \quad \dots$$



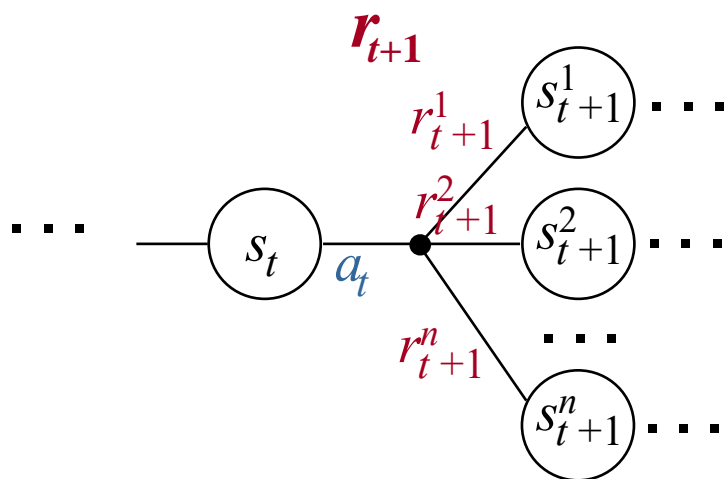
Soma das recompensas locais descontadas (no tempo)

$$U(s_t) = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots$$

Utilidade (Valor) de Estado

Cálculo da utilidade de estado considerando o não determinismo da acção
É necessário considerar a probabilidade de ocorrência de cada resultado possível

Por exemplo a recompensa r_{t+1} pode ter diferentes valores possíveis, dependendo do efeito não determinista da acção a_t



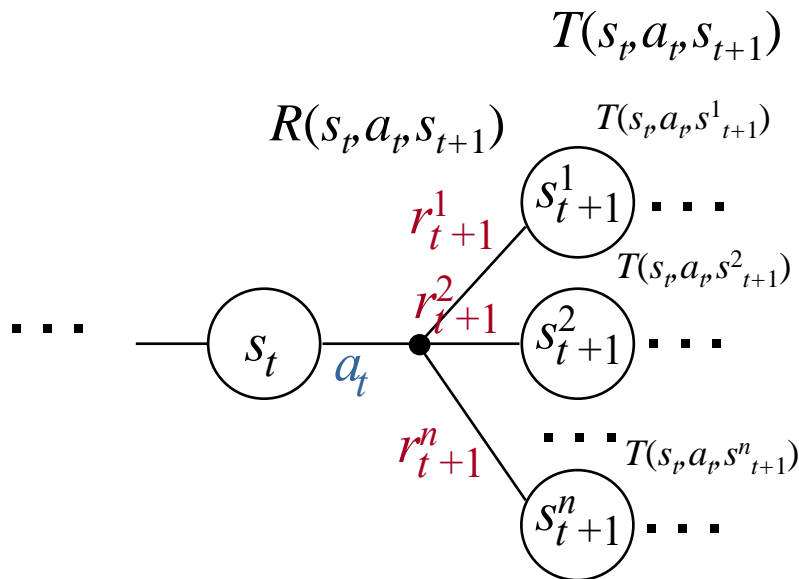
Para considerar a probabilidade de ocorrência de cada resultado possível deve ser calculado o **valor esperado**:

$$E\langle x \rangle = \sum_{i=0}^n p(x_i)x_i$$

$$U(s_t) = E\langle r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots \rangle$$

Utilidade (Valor) de Estado

Cálculo da utilidade de estado, tendo por base um **modelo de transição** de estado $T(s, a, s')$ e um **modelo de recompensa** $R(s, a, s')$



$R(s_t, a_t, s_{t+1})$ pode gerar diferentes resultados com diferentes probabilidades:

$$R(s_t, a_t, s_{t+1}^1) = r_{t+1}^1$$

$$R(s_t, a_t, s_{t+1}^2) = r_{t+1}^2$$

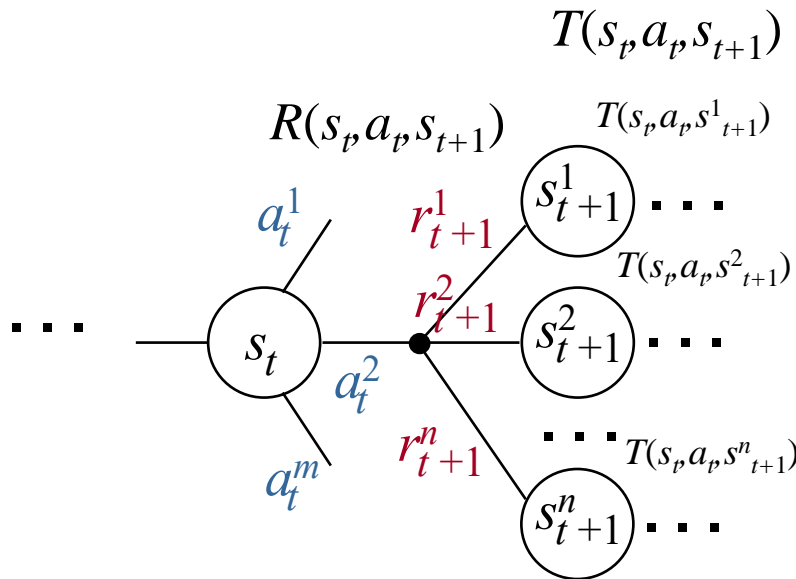
...

$$R(s_t, a_t, s_{t+1}^n) = r_{t+1}^n$$

$$\begin{aligned} U(s_t) &= \mathbb{E} \left\langle R(s_t, a_t, s_{t+1}) + \gamma R(s_{t+1}, a_{t+1}, s_{t+2}) + \gamma^2 R(s_{t+2}, a_{t+2}, s_{t+3}) + \dots \right\rangle \\ &= \mathbb{E} \left\langle R(s_t, a_t, s_{t+1}) + \gamma U(s_{t+1}) \right\rangle \end{aligned}$$

Utilidade (Valor) de Estado

Cálculo da utilidade de estado, tendo por base um **modelo de transição** de estado $T(s, a, s')$ e um **modelo de recompensa** $R(s, a, s')$



Valor esperado:

$$E\langle x \rangle = \sum_{i=0}^n p(x_i) x_i$$

$$U(s_t) = E\langle R(s_t, a_t, s_{t+1}) + \gamma U(s_{t+1}) \rangle$$

$$= \sum_{i=1}^m \pi(s_t, a_t^i) \sum_{j=1}^n T(s_t, a_t^i, s_{t+1}^j) [R(s_t, a_t^i, s_{t+1}^j) + \gamma U(s_{t+1}^j)]$$

O Princípio da Solução Óptima

A solução óptima é a que maximiza a utilidade de cada estado observado

- O cálculo da utilidade é possível através de *Programação Dinâmica*
 - Solução obtida por decomposição de um problema em sub-problemas
- Num PDM isso é viável devido à propriedade de Markov
- As utilidades dos estados podem ser determinadas em função das utilidades dos estados sucessores

$$\begin{aligned} U^\pi(s) &= \mathbb{E} \langle r_1 + \gamma r_2 + \gamma^2 r_3 + \dots \rangle \\ &= \mathbb{E} \langle r_1 + \gamma U^\pi(s') \rangle \\ &= \sum_a \pi(s, a) \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma U^\pi(s')] \end{aligned}$$

Equação de Bellman

O Princípio da Solução Óptima

Utilidade de estado para uma política π

$$U^\pi(s) = \sum_a \pi(s, a) \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma U^\pi(s')]$$

Política óptima π^*

Para cada estado s , escolher a acção a com maior utilidade

$$\pi^*(s) = \arg \max_a \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma U(s')]$$

Utilidade de estado para a política óptima π^*

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

Processos de Decisão de Markov

Utilidade da política óptima no caso geral

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

Se a recompensa só depende do estado

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

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

Processos de Decisão de Markov

Cálculo iterativo da utilidade de estado

Iniciar $U(s)$:

$$U(s) \leftarrow 0, \quad \forall s \in S$$

Iterar $U(s)$:

$$U_{i+1}(s) \leftarrow \max_a \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma U_i(s')], \quad \forall s \in S$$

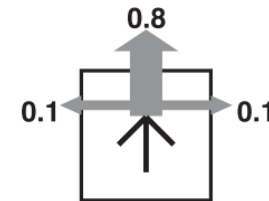
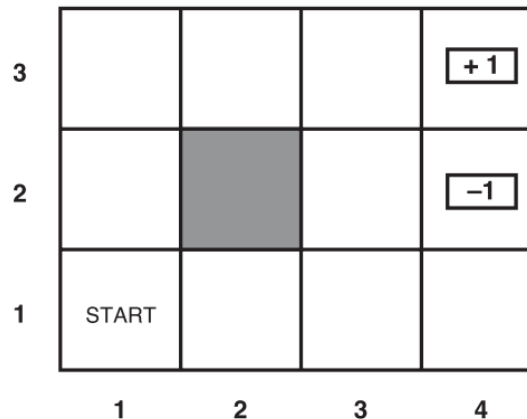
No limite:

$$U \rightarrow U^{\pi^*}$$

Cálculo da Utilidade de Estado

Exemplo:

Utilidade de estado para o ambiente 4 x 3, com $R(s) = -0.04$ para estados não terminais

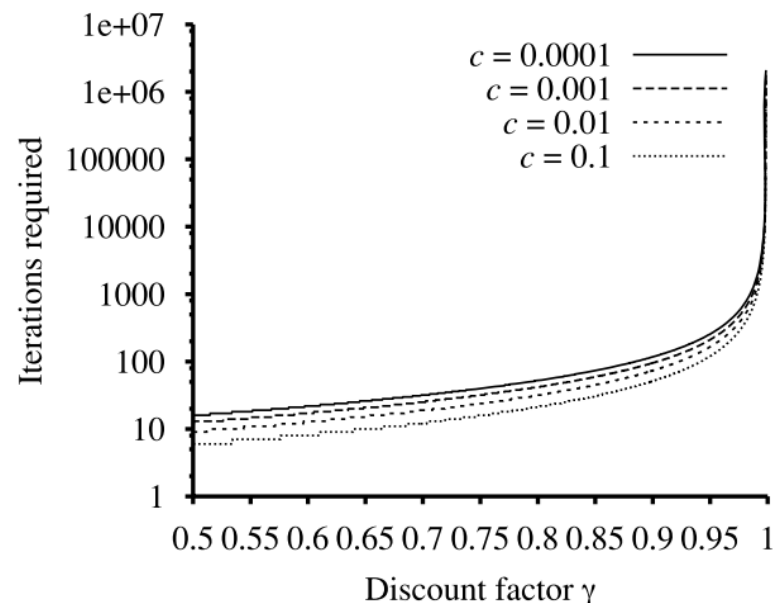
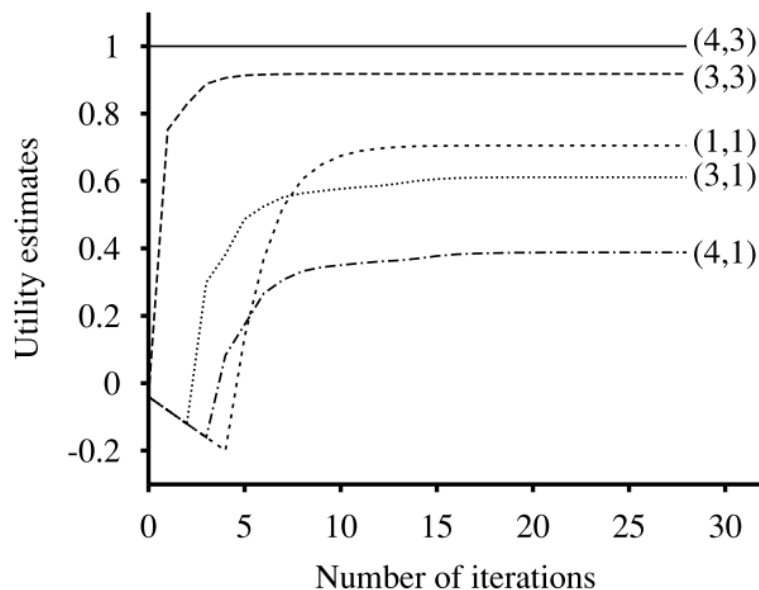
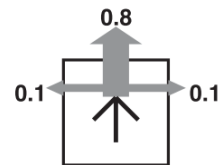
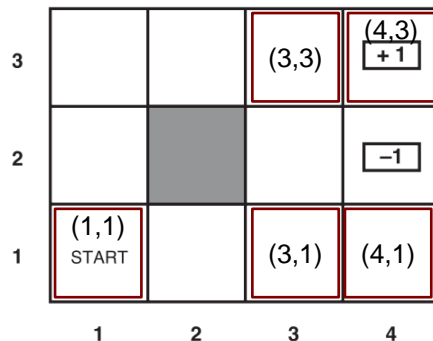


$$U_{i+1}(s) \leftarrow R(s) + \gamma \max_a \sum_{s'} T(s, a, s') U_i(s')$$

$$U(1,1) = -0.04 + \gamma \max \left\{ \begin{array}{ll} 0.8U(1,2) + 0.1U(2,1) + 0.1U(1,1), & (Up) \\ 0.9U(1,1) + 0.1U(1,2), & (Left) \\ 0.9U(1,1) + 0.1U(2,1), & (Down) \\ 0.8U(2,1) + 0.1U(1,2) + 0.1U(1,1) \} & (Right) \end{array} \right.$$

Cálculo da Utilidade de Estado

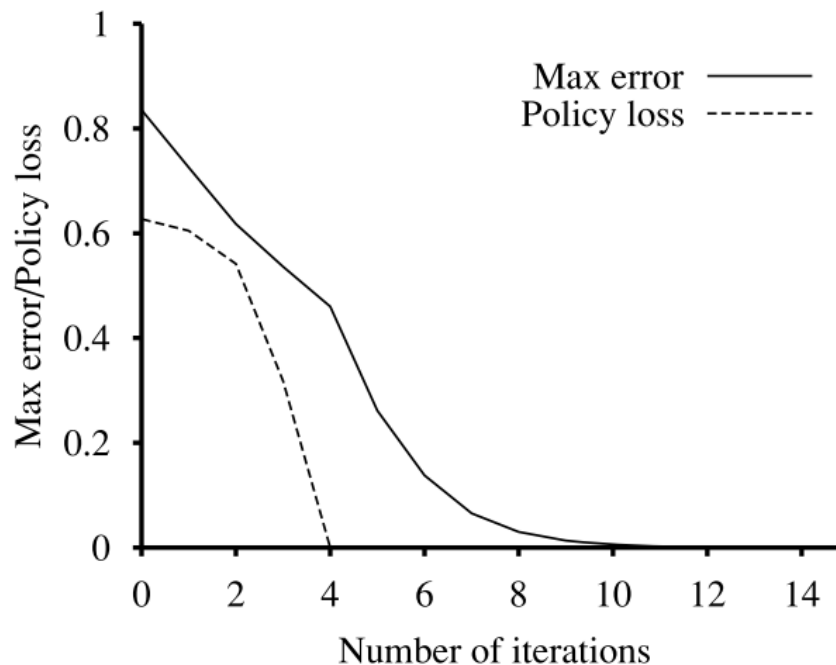
Exemplo de cálculo iterativo da utilidade para alguns estados do problema



Cálculo da Utilidade de Estado

O erro da política (acções de estado incorrectas) decresce de forma mais rápida do que o erro da utilidade

A política reflecte o gradiente da função de utilidade, pelo que tem um carácter relativo, possibilitando a convergência para a política óptima antes da convergência da função de utilidade



Processos de Decisão de Markov

Iteração da utilidade de estado

Iniciar $U(s)$:

$$U(s) \leftarrow 0, \quad \forall s \in S$$

Iterar $U(s)$:

$$U(s) \leftarrow \max_a \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma U(s')], \quad \forall s \in S$$

No limite:

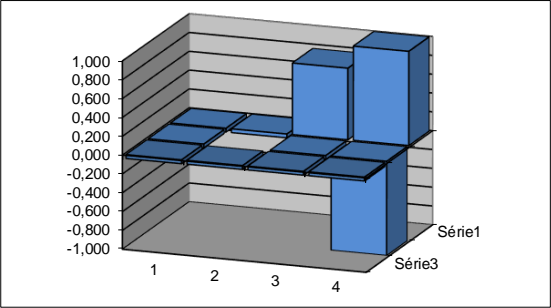
$$U \rightarrow U^{\pi^*}$$

Critério de paragem de iteração?

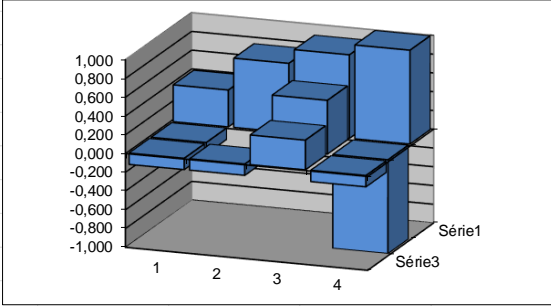
- Diferença máxima de actualização $\delta \leq \Delta_{\max}$ (limiar de convergência)

Cálculo da Utilidade de Estado

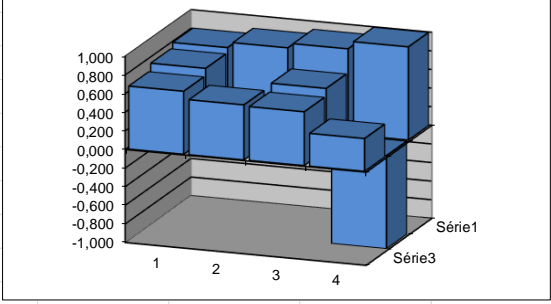
t					$\delta(s)$				δ	
0	0,000	0,000	0,000	1,000						
	0,000		0,000	-1,000						
	0,000	0,000	0,000	0,000						
1	-0,040	-0,040	0,760	1,000	-0,040	-0,040	0,760	0,000		
	-0,040		-0,040	-1,000	-0,040	0,000	-0,040	0,000		
	-0,040	-0,040	-0,040	-0,040	-0,040	-0,040	-0,040	-0,040	0,760	
2	-0,080	0,560	0,832	1,000	-0,040	0,600	0,072	0,000		
	-0,080		0,464	-1,000	-0,040	0,000	0,504	0,000		
	-0,080	-0,080	-0,080	-0,080	-0,040	-0,040	-0,040	-0,040	0,600	
3	0,392	0,738	0,890	1,000	0,472	0,178	0,058	0,000		
	-0,120		0,572	-1,000	-0,040	0,000	0,108	0,000		
	-0,120	-0,120	0,315	-0,120	-0,040	-0,040	0,395	-0,040	0,472	
4	0,577	0,819	0,906	1,000	0,185	0,082	0,017	0,000		
	0,250		0,629	-1,000	0,370	0,000	0,057	0,000		
	-0,160	0,188	0,394	0,100	-0,040	0,308	0,078	0,220	0,370	
5	0,698	0,849	0,914	1,000	0,121	0,030	0,007	0,000		
	0,472		0,648	-1,000	0,222	0,000	0,019	0,000		
	0,162	0,313	0,492	0,185	0,322	0,124	0,098	0,085	0,322	
6	0,756	0,861	0,916	1,000	0,058	0,012	0,003	0,000		
	0,613		0,656	-1,000	0,141	0,000	0,008	0,000		
	0,385	0,416	0,528	0,272	0,222	0,104	0,036	0,087	0,222	
7	0,785	0,865	0,917	1,000	0,029	0,004	0,001	0,000		
	0,687		0,658	-1,000	0,075	0,000	0,003	0,000		
	0,530	0,466	0,553	0,310	0,145	0,050	0,025	0,038	0,145	



Iteração 1

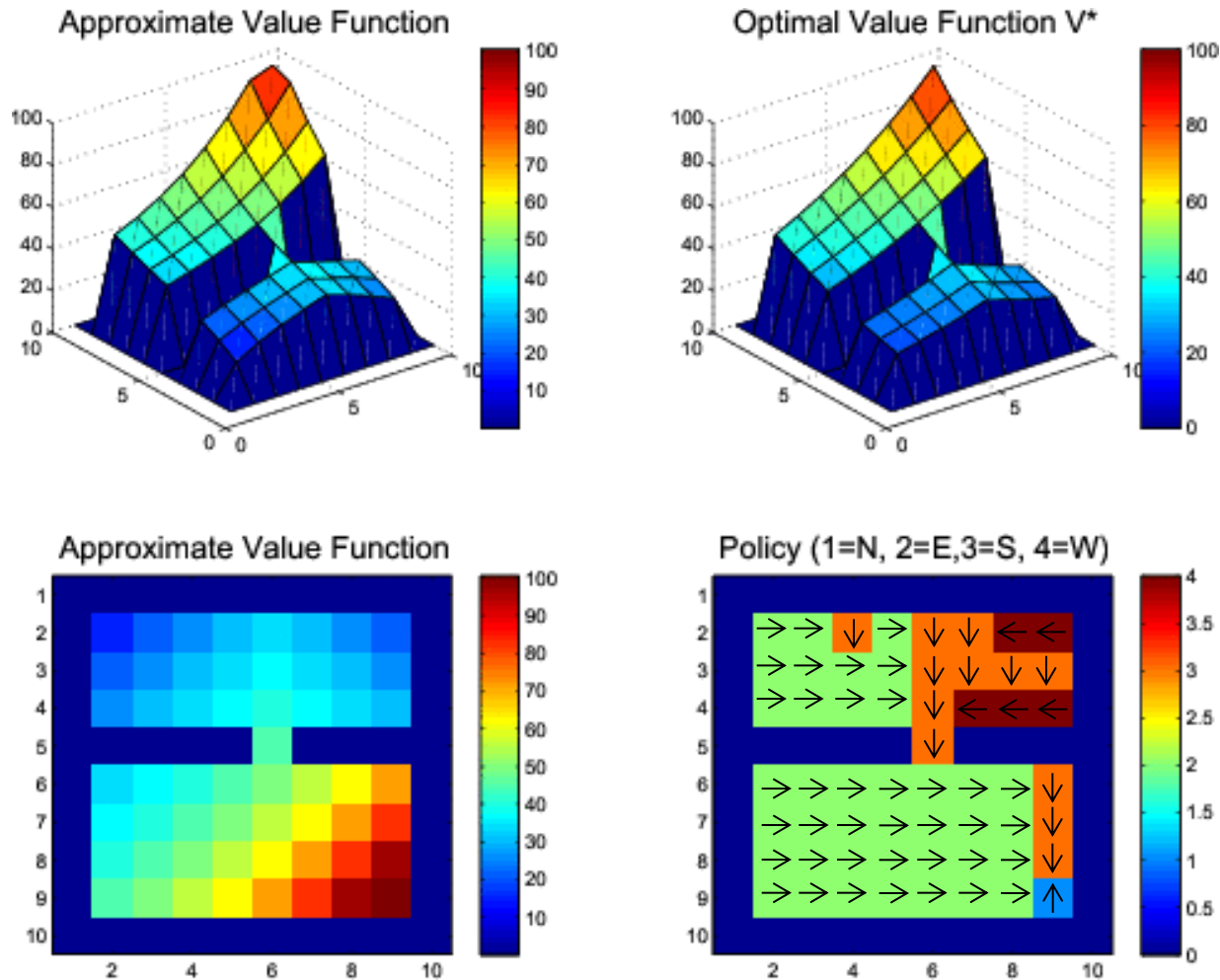


Iteração 3



Iteração 10

Processos de Decisão de Markov



[Mahadevan, 2009]

CÁLCULO DA UTILIDADE

Iteração da utilidade de estado

Iniciar $U(s)$:

$$U(s) \leftarrow 0, \forall s \in S$$

Iterar $U(s)$:

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

No limite:

$$U \rightarrow U^{\pi^*}$$

Critério de paragem de iteração?

- Diferença máxima de actualização $< \Delta_{\max}$ (limiar de convergência)

```
function utilidade:
```

```
     $U[s] \leftarrow 0, \forall s \in S$ 
```

```
    do:
```

```
         $U_{ant} \leftarrow U$ 
```

```
         $\delta \leftarrow 0$ 
```

```
        for  $s$  in  $S$ :
```

```
             $U[s] \leftarrow \max_{a \in A(s)} U_{acção}(s, a, U_{ant})$ 
```

```
             $\delta \leftarrow \max\{\delta, |U[s] - U_{ant}[s]|\}$ 
```

```
    while  $\delta > \Delta_{\max}$ :
```

```
    return  $U$ 
```

```
function  $U_{acção}(s, a, U)$ :
```

```
    return  $\sum_{s'} T(s, a, s') [R(s, a, s') + \gamma U[s']]$ 
```

Processos de Decisão de Markov

- **Propriedade de Markov**
 - Estados futuros dependem apenas do estado actual
 - São independentes de estados passados
- **Modelo do mundo - representação do problema**
 - Conjunto de estados
 - S
 - Conjunto de acções possíveis num estado
 - $A(s)$
 - Modelo de transição
 - $T(s,a,s')$ – também designado $P(s,a,s')$
 - Modelo de recompensa
 - $R(s,a,s')$ – no caso geral
 - $R(s, a)$ – se a recompensa só depende do estado e da acção
 - $R(s)$ – se a recompensa só depende do estado

Processos de Decisão de Markov

- Problemas
 - Dimensão dos espaços de estados
(*problema da dimensionalidade*)
 - Dificuldade de definição das dinâmicas
(por exemplo a partir de dados experimentais)
 - Modelos desconhecidos

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

? ?

- Resolução através de aprendizagem por reforço

Referências

[Russel & Norvig, 2010]

S. Russell, P. Norvig, “Artificial Intelligence: A Modern Approach”, 3rd Ed., Prentice Hall, 2010

[Sutton & Barto, 1998]

R. Sutton, A. Barto, “Reinforcement Learning: An Introduction”, MIT Press, 1998

[Mahadevan, 2009]

S. Mahadevan, “Learning Representation and Control in Markov Decision Processes: New Frontiers”, Foundations and Trends in Machine Learning, 1:4, 2009

[LaValle, 2006]

S. LaValle, “Planning Algorithms”, Cambridge University Press, 2006

[Kragic & Vincze, 2009]

D. Kragic, M. Vincze, “Vision for Robotics”, Foundations and Trends in Robotics, 1:1, 2009