PROCESSOS DE DECISÃO SEQUENCIAL

Luís Morgado

ISEL-DEETC

O PROBLEMA DA TOMADA DE DECISÃO

TOMADA DE DECISÃO

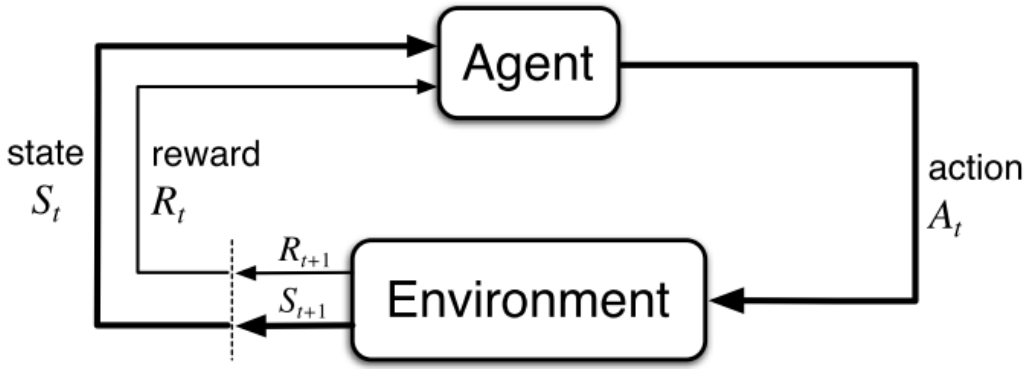
- Processo cognitivo que resulta na selecção de uma opção entre várias alternativas
- Baseada em alguma forma de **valor** ou preferência
  - Reflecte **propósito** (objectivo)
- **Teorias de racionalidade ilimitada**
  - Capacidade de representação e cálculo ilimitada
  - Representações completas e consistentes
  - Optimização
- **Teorias de racionalidade limitada**
  - Capacidade de representação e cálculo limitada
  - Utilização de heurísticas
  - Soluções aproximadas

O CONCEITO DE AGENTE

AGENTE

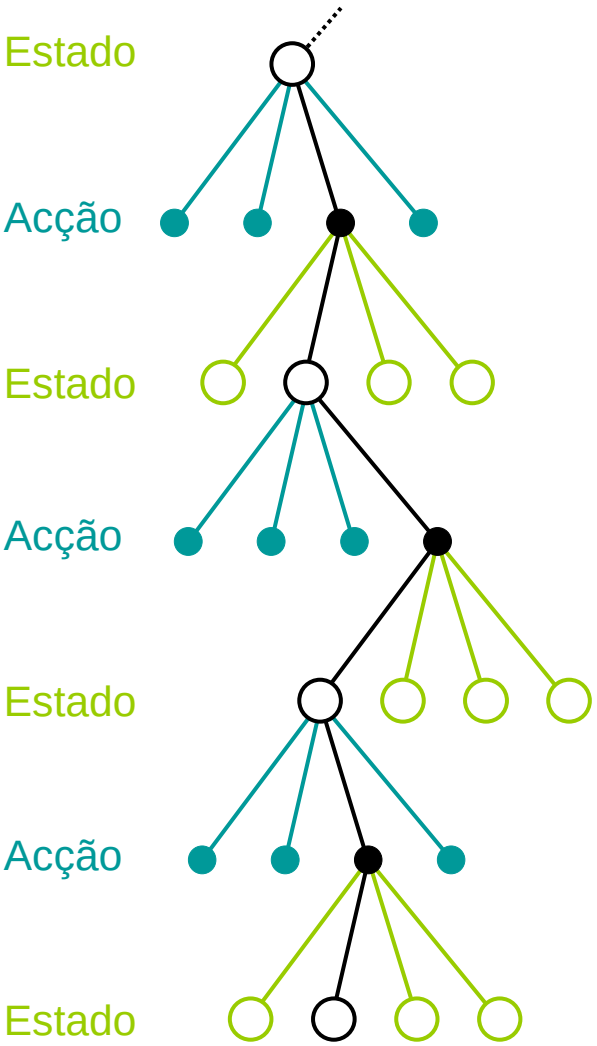
- Sistema computacional capaz de percepção, decisão e acção para concretização autónoma de objectivos, maximizando uma medida de desempenho
- Cognição, racionalidade

Contexto geral de tomada de decisão

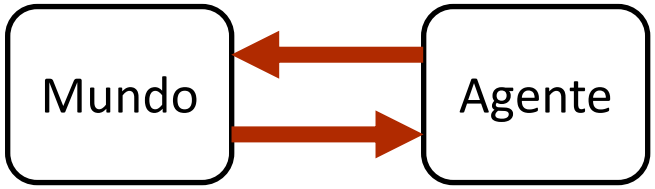


[Sutton & Barto, 2020]

TOMADA DE DECISÃO SEQUENCIAL

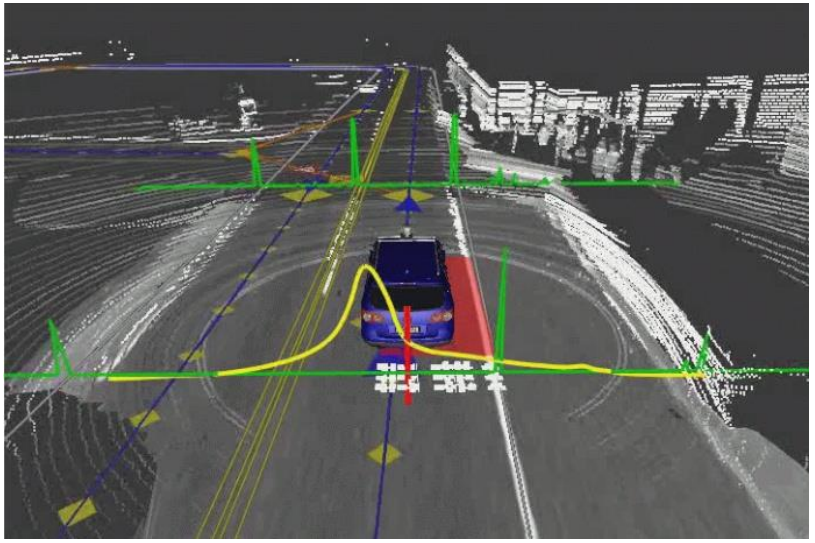


Como prever e controlar o desenrolar da interacção entre agente e ambiente ao longo do tempo para um **objectivo de longo prazo**?



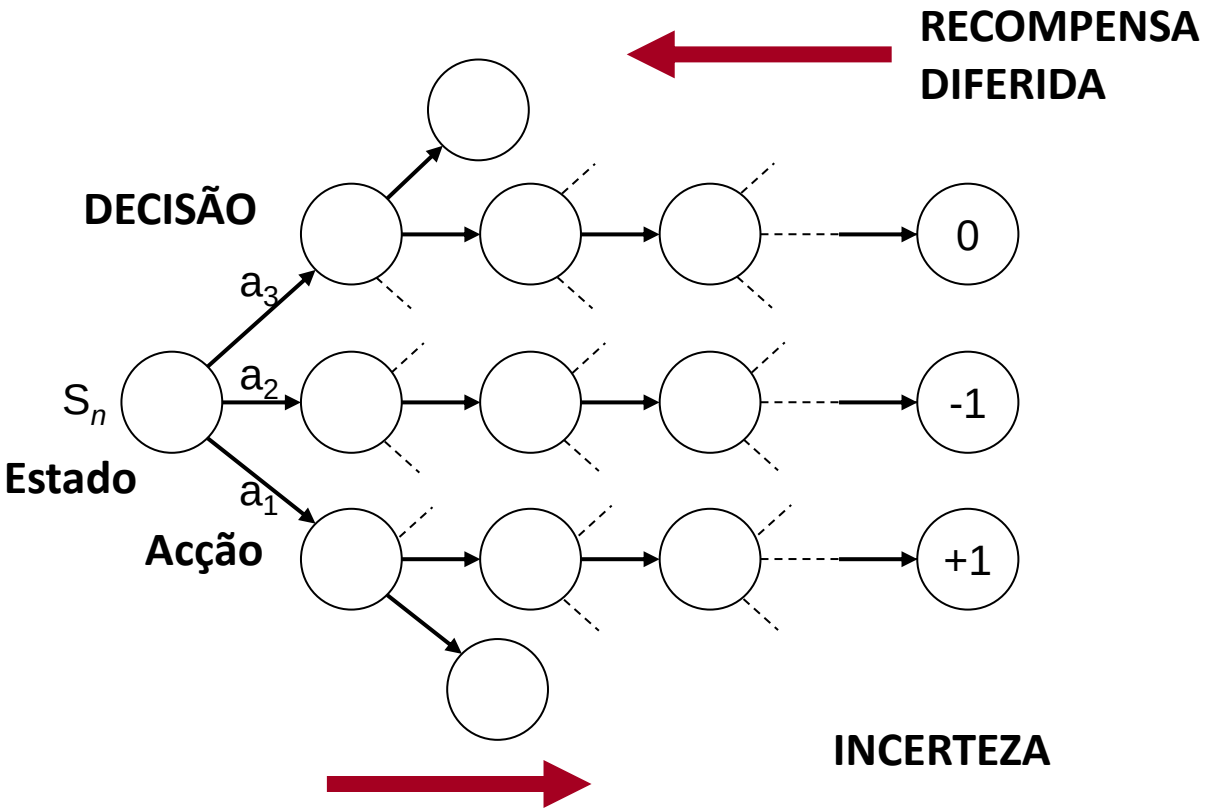
[Sutton & Barto, 1998]

PROBLEMAS DE DECISÃO SEQUENCIAL

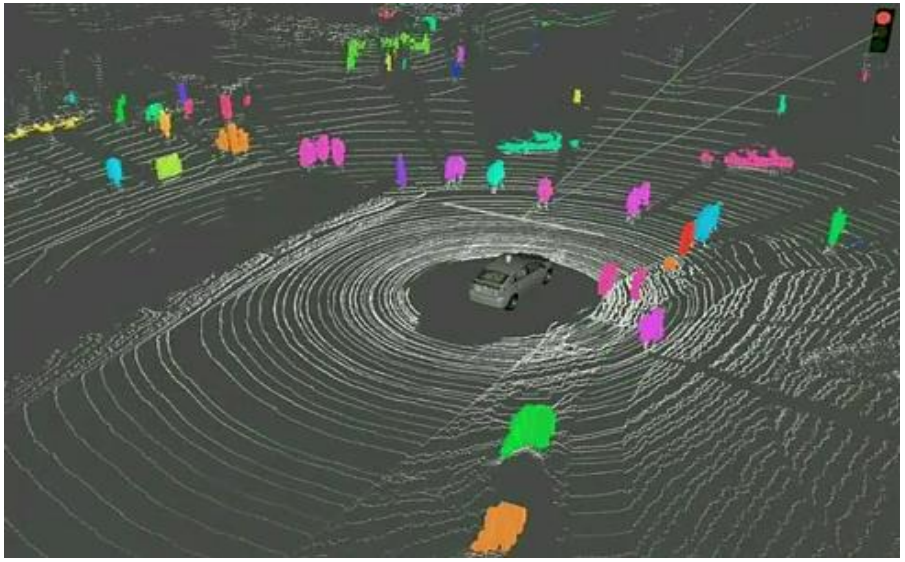


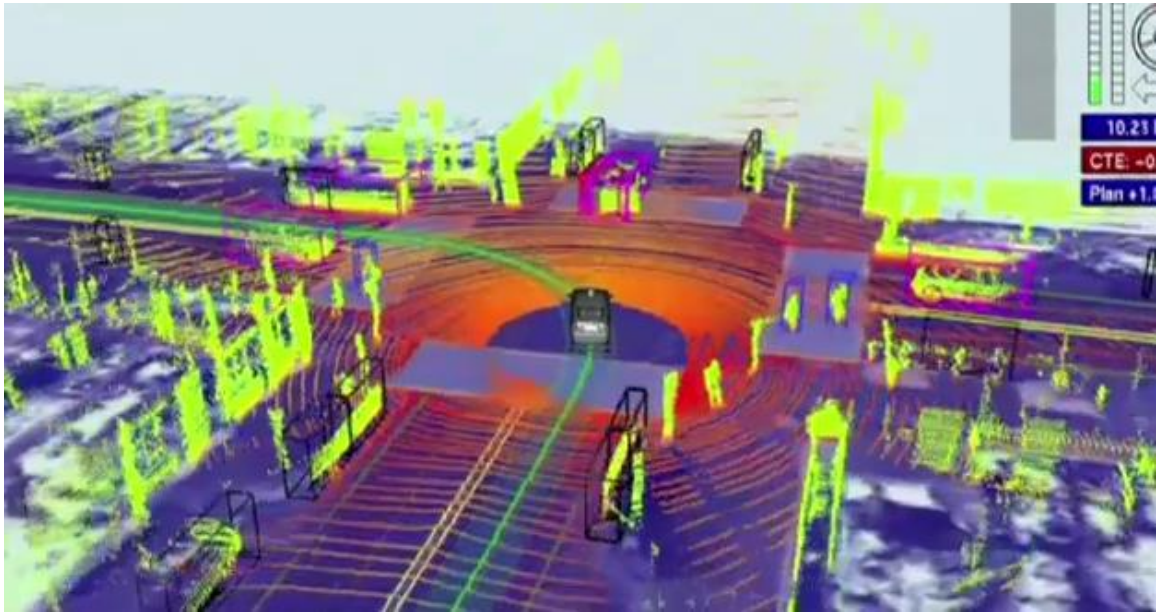
PROBLEMAS DE DECISÃO SEQUENCIAL

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 não-determinísticos

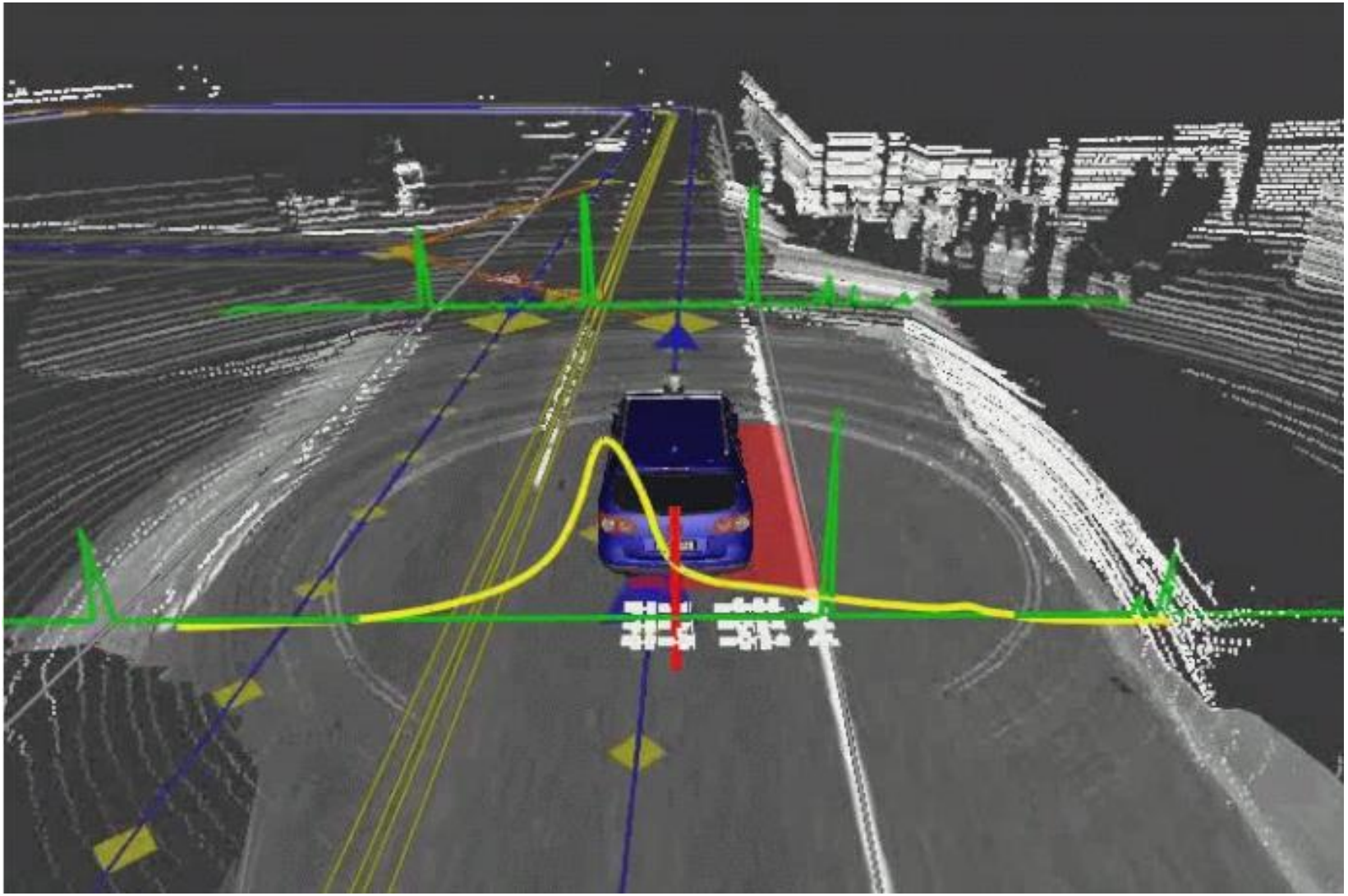


RACIOCÍNIO COM INCERTEZA





RACIOCÍNIO COM INCERTEZA



[Montemerlo, 2008]

PROCESSOS DE DECISÃO SEQUENCIAL

ESPAÇO DE ESTADOS NÃO-DETERMINÍSTICO

Estados

Acções

Transições

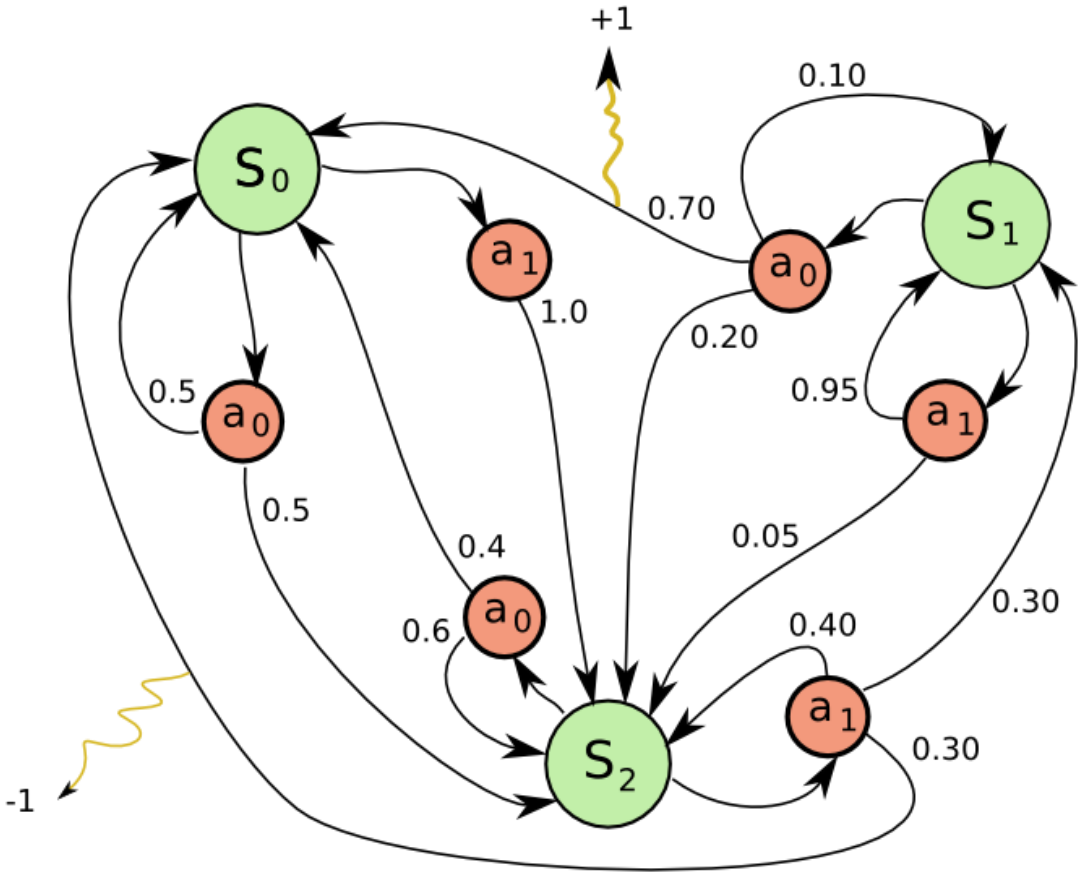
Recompensas

Modelo de Transição

– $T(s,a,s')$

Modelo de Recompensa

– $R(s,a,s')$



PROCESSOS DE DECISÃO SEQUENCIAL

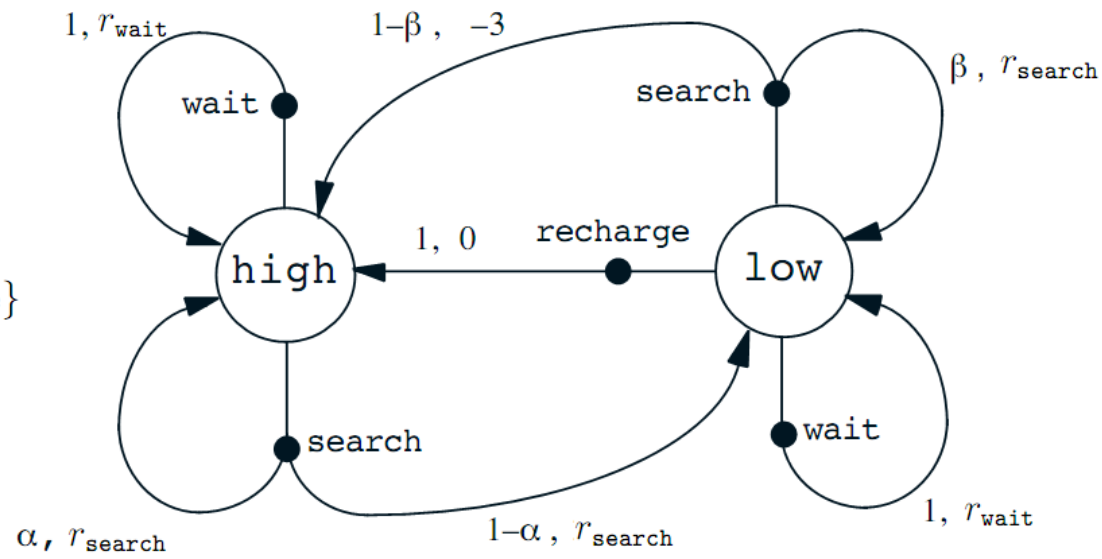
EXEMPLO: ROBOT DE RECICLAGEM

s	s'	a	$p(s'\|s,a)$	$r(s,a,s')$
high	high	search	α	$r_{\texttt{search}}$
high	low	search	$1-\alpha$	$r_{\texttt{search}}$
low	high	search	$1-\beta$	-3
low	low	search	β	$r_{\texttt{search}}$
high	high	wait	1	$r_{\texttt{wait}}$
high	low	wait	0	$r_{\texttt{wait}}$
low	high	wait	0	$r_{\texttt{wait}}$
low	low	wait	1	$r_{\texttt{wait}}$
low	high	recharge	1	0
low	low	recharge	0	0.

$\mathcal{S} = \{\texttt{high}, \texttt{low}\}$ (carga da bateria)

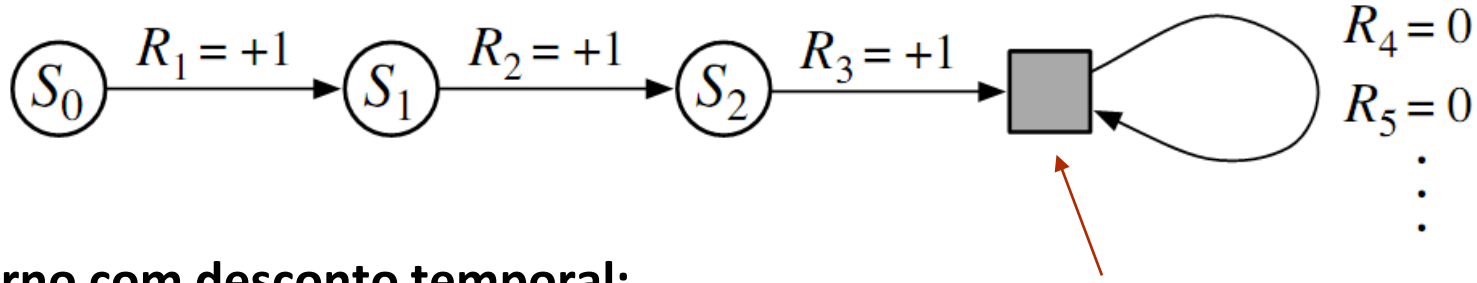
$\mathcal{A}(\texttt{high}) = \{\texttt{search}, \texttt{wait}\}$

$\mathcal{A}(\texttt{low}) = \{\texttt{search}, \texttt{wait}, \texttt{recharge}\}$



[Sutton & Barto, 2012]

CADEIAS DE MARKOV



Retorno com desconto temporal:

$$G_t = \sum_{k=0}^{T-t-1} \gamma^k R_{t+k+1}, \quad \gamma \in [0,1] \text{ - Factor de desconto}$$

PROPRIEDADE DE MARKOV

Um processo estocástico tem a **propriedade de Markov** se a distribuição probabilística condicional dos estados futuros de um processo **depender exclusivamente do estado presente**

[Sutton & Barto, 2012]

RETORNO DE HORIZONTE INFINITO

- Não está limitado a uma gama finita de valores
- Necessário ponderar a distância no tempo das recompensas
  - Recompensas descontadas (*discounted reward*)
  - Factor de desconto $\boldsymbol{\gamma} \in [0,1]$

$$R_t = r_{t+1} + \gamma \cdot r_{t+2} + \gamma^2 \cdot r_{t+3} + \gamma^3 \cdot r_{t+4} + \ldots = \sum_{i=1}^{\infty} \gamma^{i-1} \cdot r_{t+i}$$

Recompensas descontadas no tempo

PROCESSOS DE DECISÃO DE MARKOV

- **Representação do mundo sob a forma de PDM**

  S – conjunto de estados do mundo

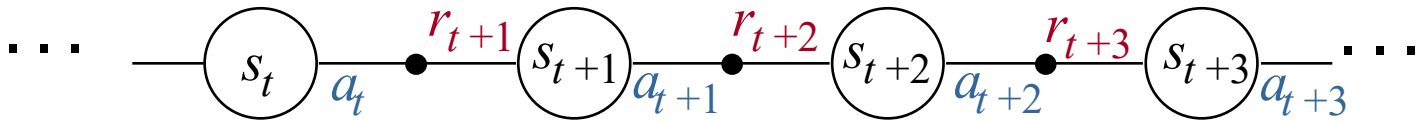
  $A(s)$ – conjunto de acções possíveis no estado $s \in S$

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

  $R(s,a,s')$ – retorno esperado na transição de s para s' através de a

  γ – taxa de desconto para recompensas diferidas no tempo

  $t = 0, 1, 2, \ldots$ – tempo discreto



Cadeia de Markov

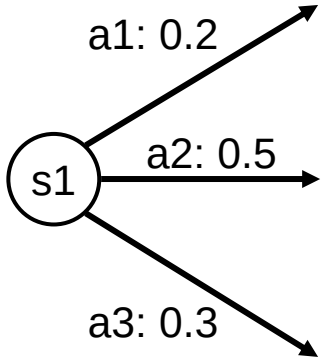
POLÍTICA COMPORTAMENTAL

- Forma de representação do comportamento do agente
- Define qual a acção que deve ser realizada em cada estado
- Política **determinista**

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

- Política **não determinista**

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



PROCESSOS DE DECISÃO DE MARKOV

Objectivo

- Maximizar o valor (retorno) de uma sequência de acções
  - Política comportamental
  - Valor de um estado com base numa política

$$V^{\pi}(s) = \mathrm{E}\langle r_1 + \gamma r_2 + \gamma^2 r_3 + \ldots \,|\, s_0 = s, \pi \rangle$$

Factor de desconto $\gamma \in [0,1]$ para recompensas diferidas no tempo

- Obter política óptima
- Definir uma relação de ordem parcial entre políticas, existindo, pelo menos, uma **política óptima**

$$\pi^* = \arg\max_{\pi} V^{\pi}$$

[Sutton & Barto, 2012]

EXEMPLO

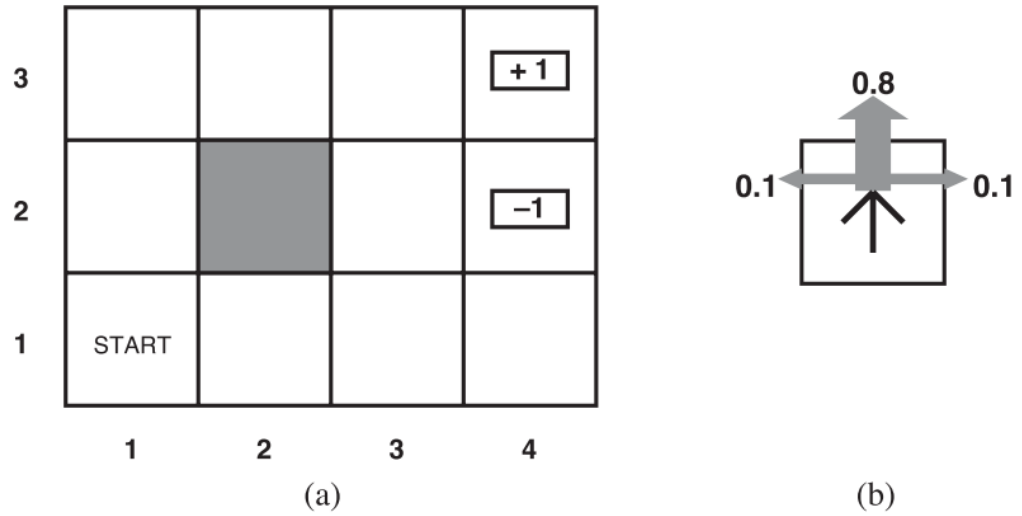


Figure 17.1 (a) A simple 4×3 environment that presents the agent with a sequential decision problem. (b) Illustration of the transition model of the environment: the "intended" outcome occurs with probability 0.8, but with probability 0.2 the agent moves at right angles to the intended direction. A collision with a wall results in no movement. The two terminal states have reward +1 and –1, respectively, and all other states have a reward of –0.04.

[Russel & Norvig, 2010]

EXEMPLO

POLÍTICA COMPORTAMENTAL

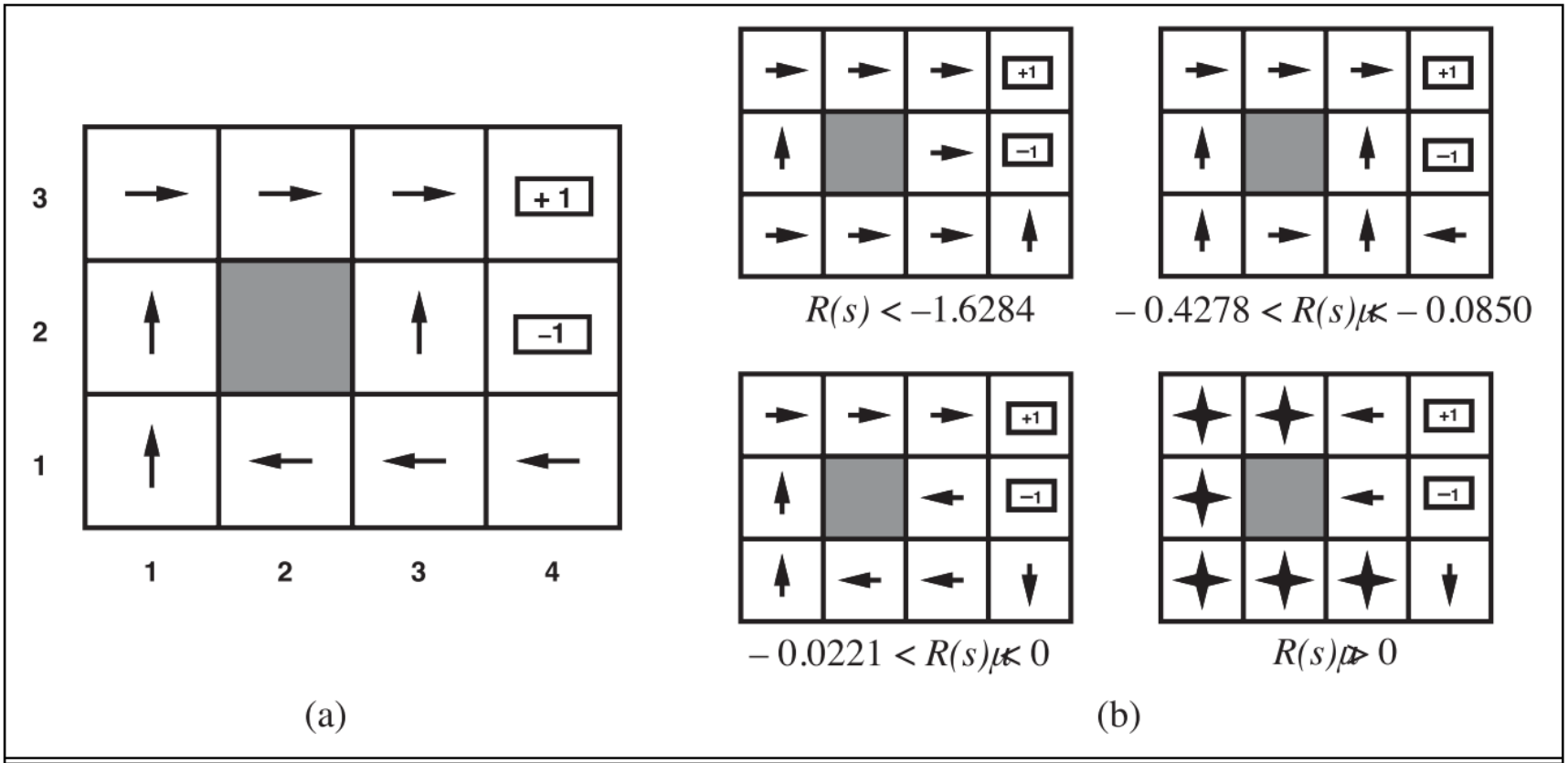


Figure 17.2 (a) An optimal policy for the stochastic environment with $R(s) = -0.04$ in the nonterminal states. (b) Optimal policies for four different ranges of $R(s)$.

[Russel & Norvig, 2010]

EXEMPLO

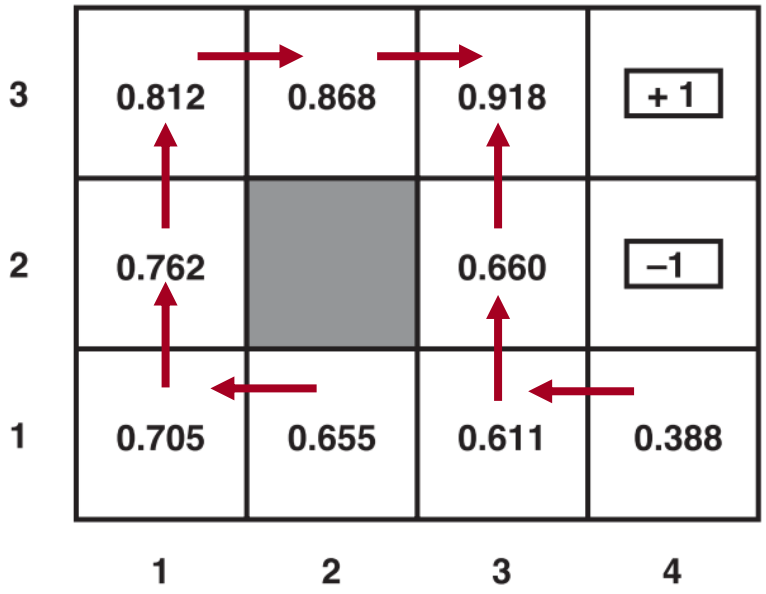


Figure 17.3 The utilities of the states in the 4×3 world, calculated with $\gamma = 1$ and $R(s) = -0.04$ for nonterminal states.

[Russel & Norvig, 2010]

PROCESSOS DE DECISÃO DE MARKOV

- **Utilidade** (**valor**) de estado
  - Medida de valor de estado
  - Reflecte congruência com uma finalidade definida (objectivo)
  - Depende da **política** comportamental utilizada

- $\boldsymbol{R(s)}$
  - Recompensa a **curto-prazo**
- $\boldsymbol{U(s)}$
  $\boldsymbol{V(s)}$
  - Recompensa a **longo-prazo**

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

[Sutton & Barto, 2012]
R. Sutton, A. Barto, “Reinforcement Learning: An Introduction”, 2nd Edition - Preview, MIT Press, 2012

[Sutton & Barto, 2020]
R. Sutton, A. Barto, “Reinforcement Learning: An Introduction”, 2nd Edition, MIT Press, 2020

[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