APRENDIZAGEM POR REFORÇO

Luís Morgado

2015

APRENDIZAGEM AUTOMÁTICA

Aprendizagem = Melhoria de desempenho,
para uma dada tarefa,
com a experiência

- Melhorar o desempenho para uma dada **tarefa *T***
- Com base numa medida de **desempenho *D***
- Com base na **experiência *E***

EXEMPLOS

Aprender a jogar xadrez

- ***T***: Jogar xadrez
- ***D***: Percentagem de jogos ganhos
- ***E***: Jogos realizados

Aprender a reconhecer escrita manual

- ***T***: Reconhecer e classificar caracteres escritos manualmente
- ***D***: Percentagem de caracteres reconhecidos correctamente
- ***E***: Conjunto de exemplos de caracteres e respectiva classificação

Aprender a conduzir um veículo

- ***T***: Conduzir com base na informação proveniente de câmaras de vídeo
- ***D***: Distância média percorrida sem erros
- ***E***: Sequências de imagens e de comandos de condução obtidos através da observação de um condutor humano

APRENDIZAGEM AUTOMÁTICA

Aprendizagem ≠ Memorização

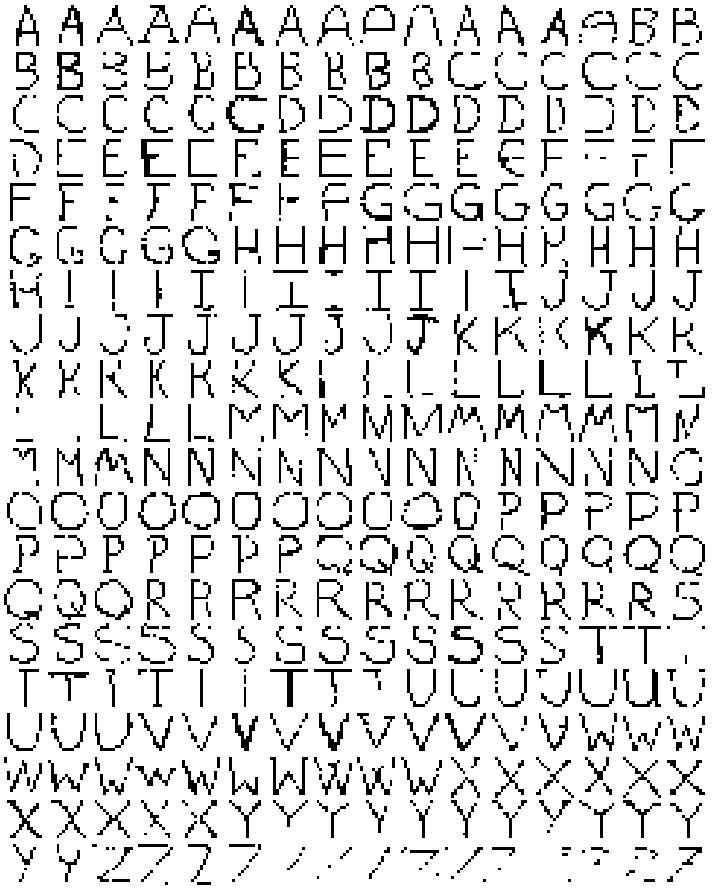
- Aprendizagem
  - **Generalização**
  - Formação de **abstracções** (modelos)
    - Protótipos
    - Conceitos
    - Padrões comportamentais

APRENDIZAGEM AUTOMÁTICA

- APRENDIZAGEM **CONCEPTUAL**
  - **O que é?**
    - Conceito
  - SUPERVISIONADA
  - NÃO SUPERVISIONADA
- APRENDIZAGEM **COMPORTAMENTAL**
  - **O que fazer?**
    - Comportamento (acção)
  - POR REFORÇO

APRENDIZAGEM CONCEPTUAL

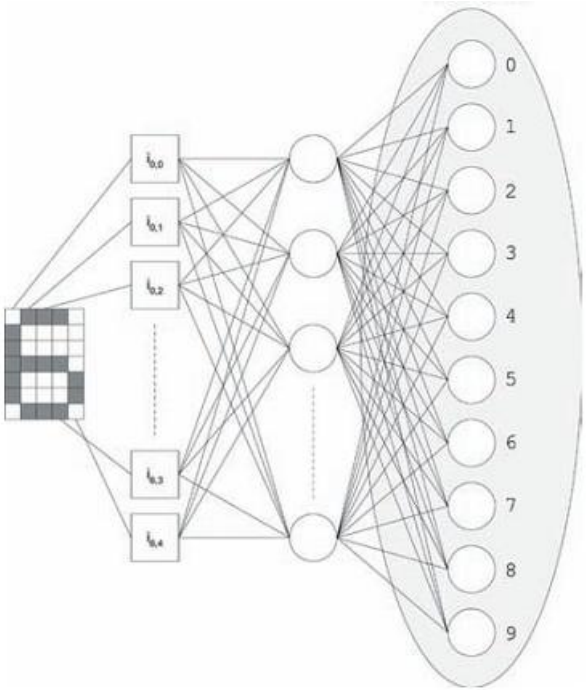
Conjunto de treino



[Fox *et al.*, 1994]

Modelo interno

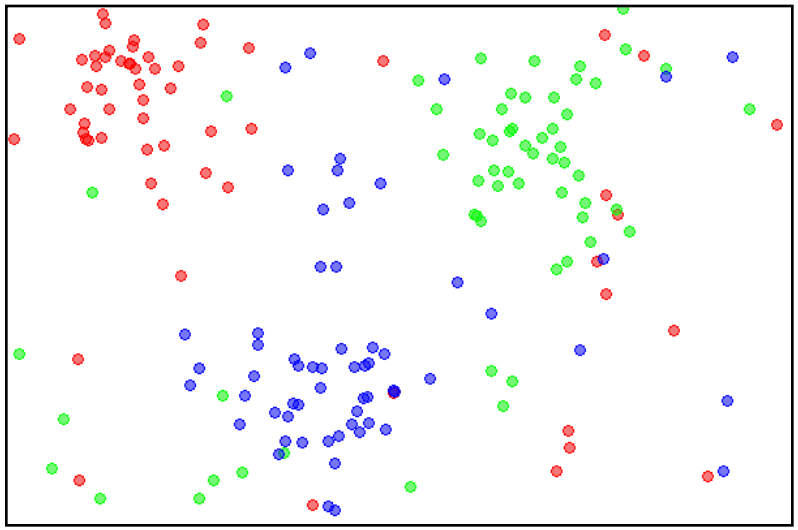
(e.g. redes reuronais)

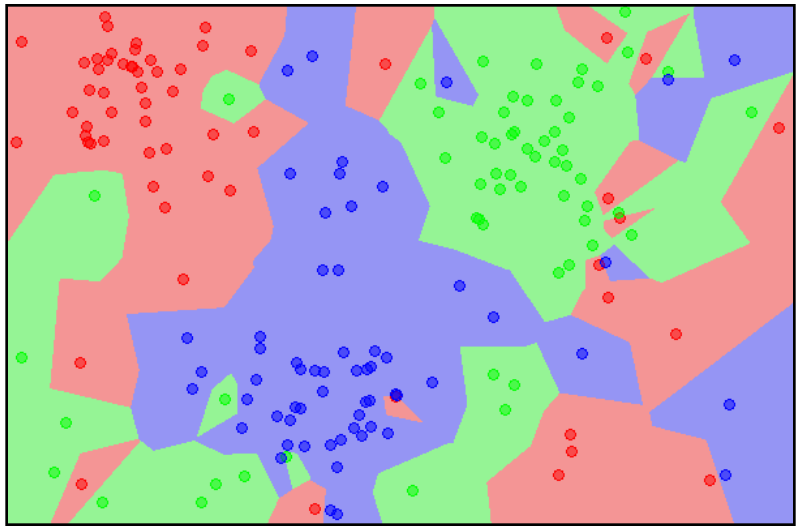


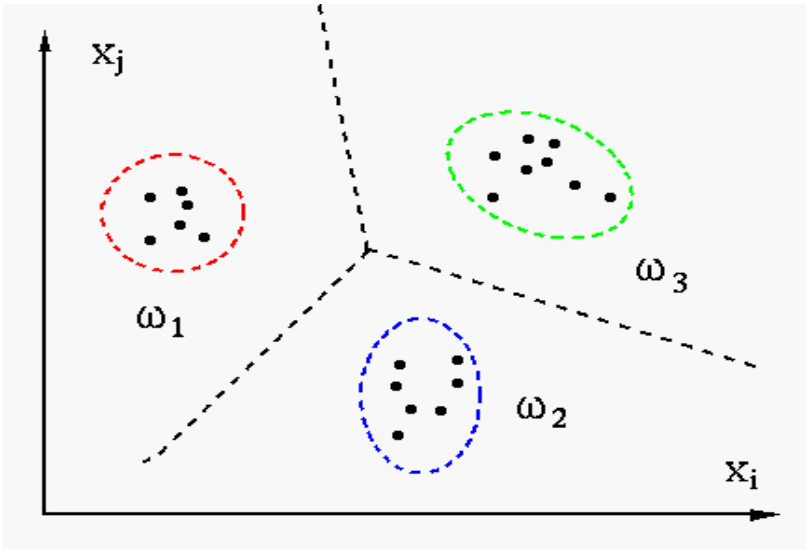
[Poole &Mackworth, 2010]

APRENDIZAGEM CONCEPTUAL

FORMAÇÃO DE CONCEITOS

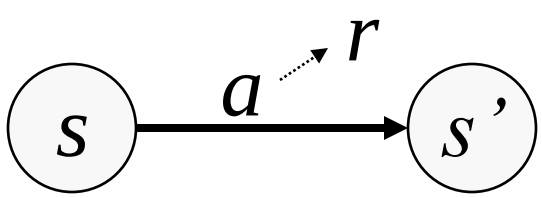




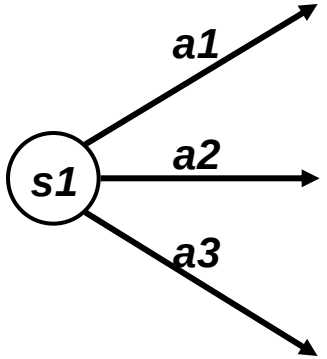


Aprendizagem por Reforço

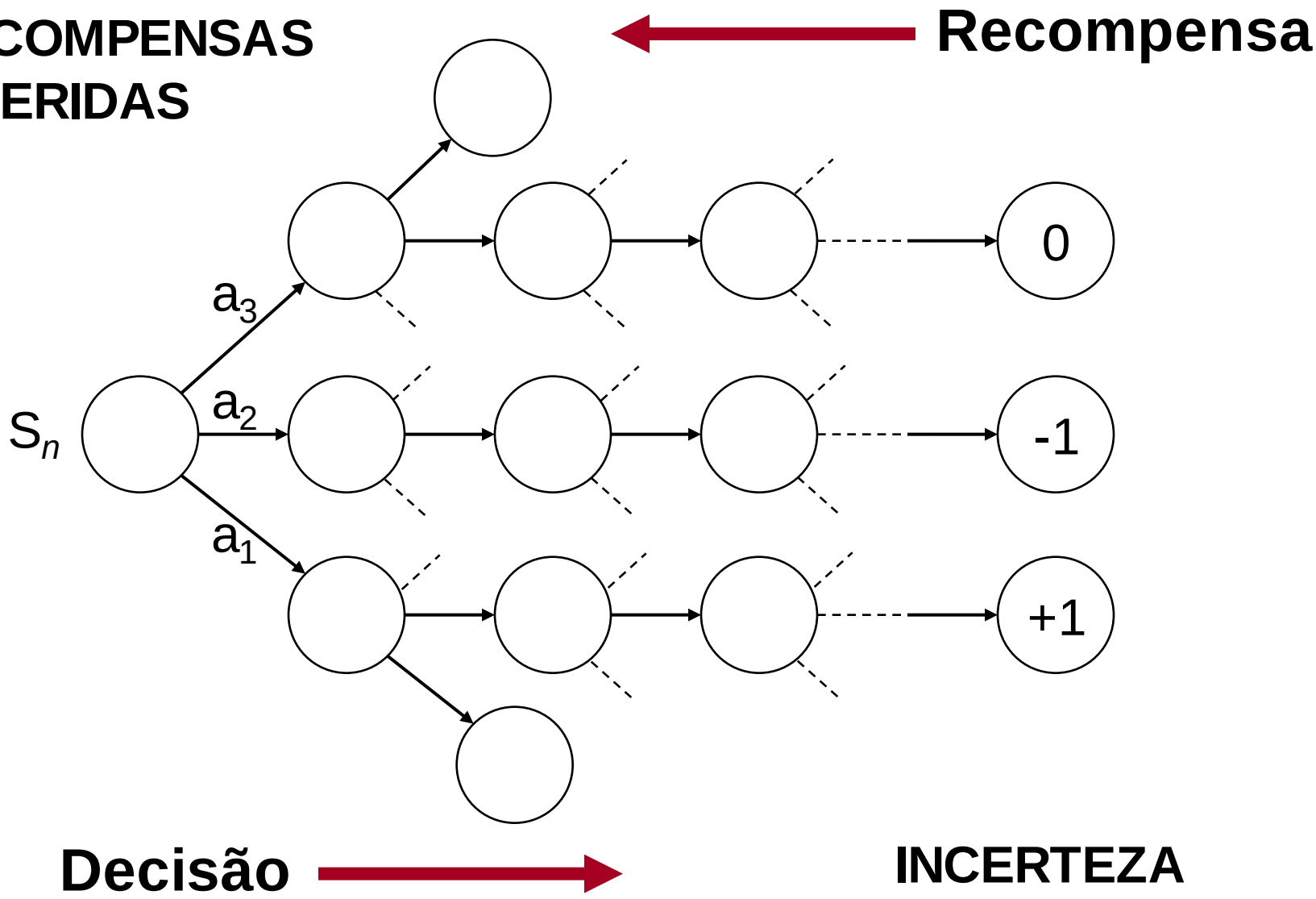
- Aprendizagem a partir da **interacção** com o ambiente
  - **Estado**
  - **Acção**
  - **Reforço**
    - Ganho / perda



- Aprendizagem de **comportamentos**
  - O que fazer
  - Relação entre situações e acções



O Problema da Decisão Sequencial



Processos de Decisão Sequencial

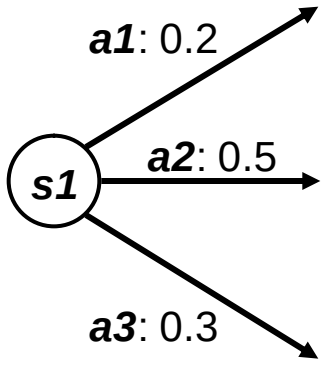
Política Comportamental

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

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

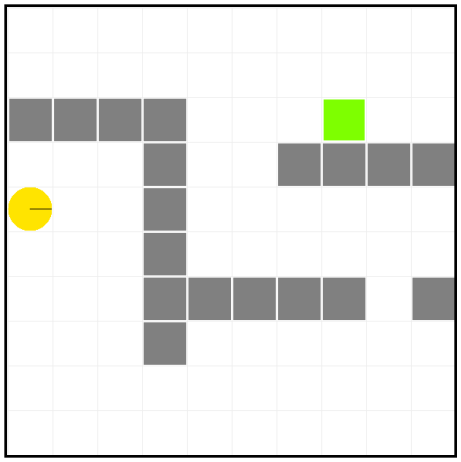
- Política **não determinista**

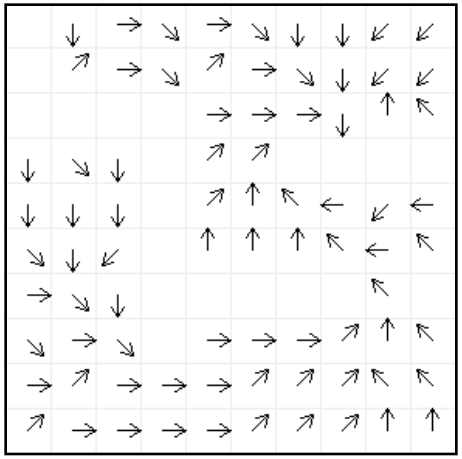
$$\pi : S \times A(s) \rightarrow [0,1] \; ; \; s \in S$$

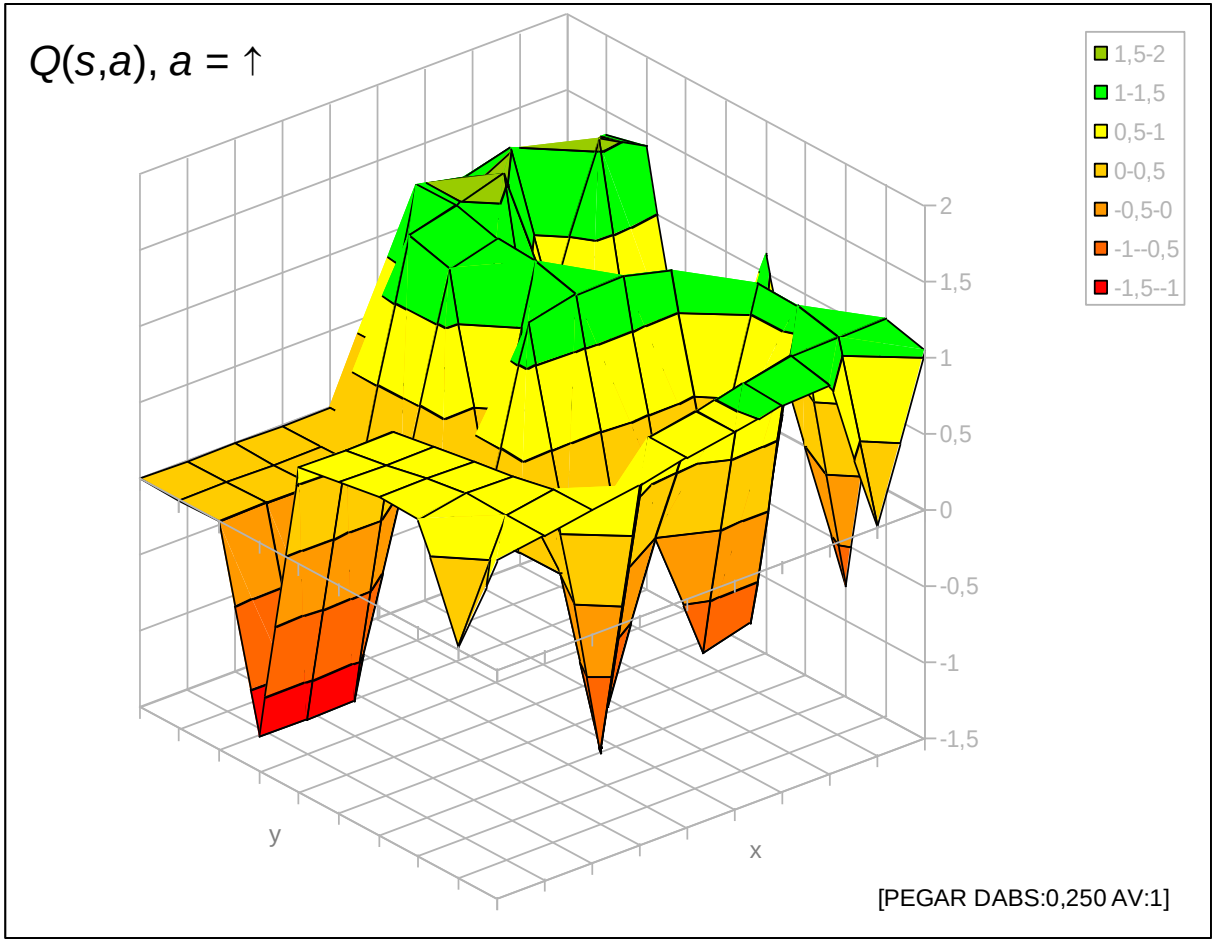


APRENDIZAGEM POR REFORÇO

FORMAÇÃO DE POLÍTICAS COMPORTAMENTAIS

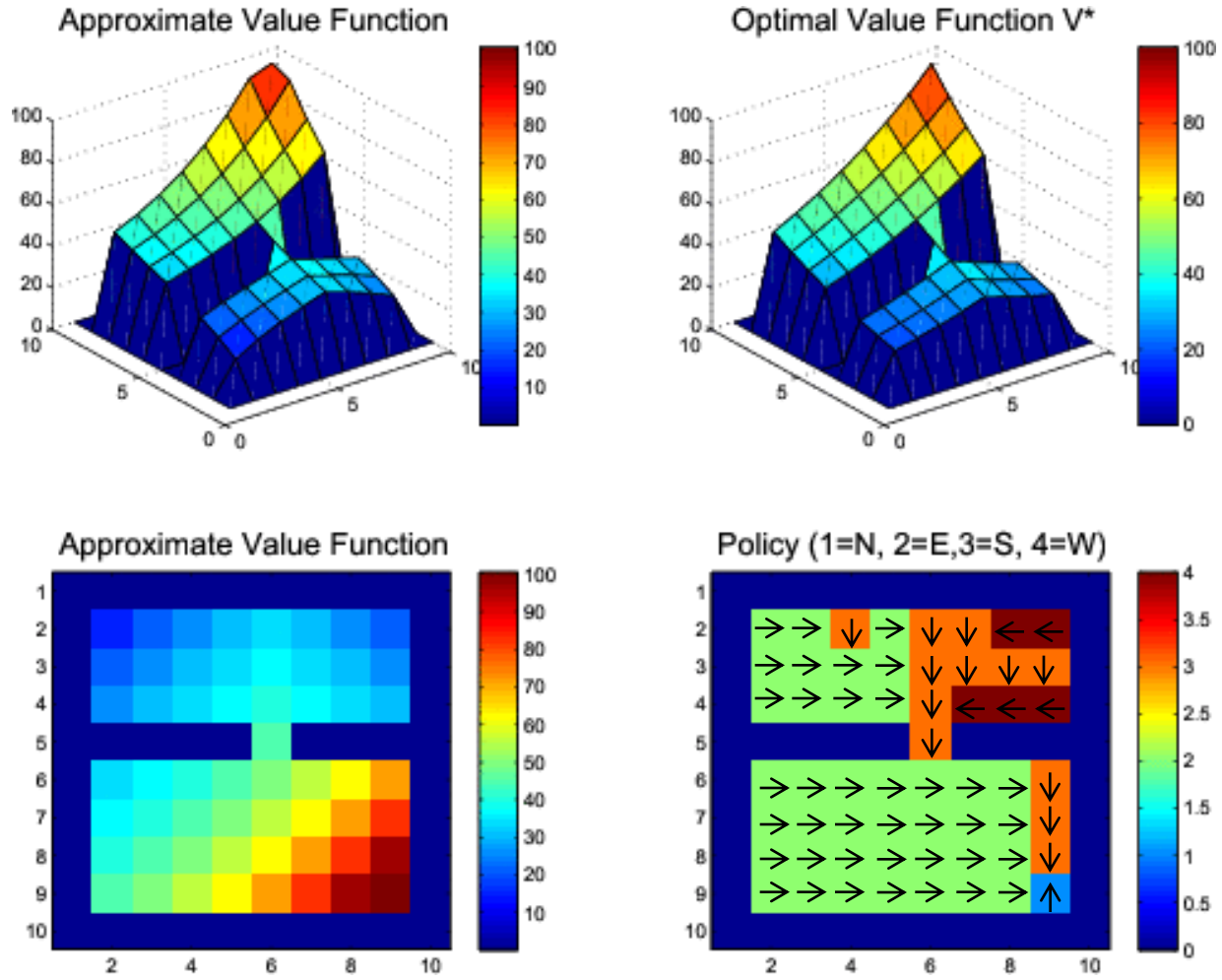






Processos de Decisão Sequencial

Política comportamental e Função valor



[Mahadevan, 2009]

Valor Cumulativo (Utilidade)

- Recompensas aditivas
  - $V^{\pi}(s_0) = R(s_0) + R(s_1) + R(s_2) + \ldots$
- Recompensas descontadas (no tempo)
  - $V^{\pi}(s_0) = R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \ldots$
  - Factor de desconto temporal
    - $\gamma \in [0,1]$
  - Recompensas não estão limitadas a uma gama finita de valores

Aprendizagem por Reforço

Seja s_0, a_0, r_1, s_1, a_1, r_2, s_2, a_2, r_3,... uma sequência observada de acções seleccionadas com base na política π

Para cada instante temporal t,

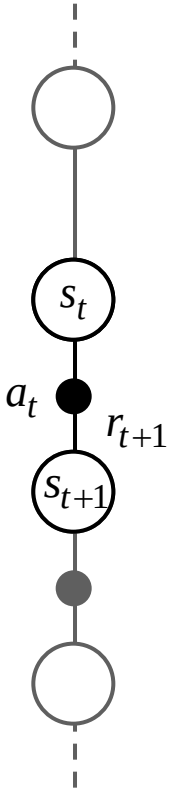
$$V^{\pi}(s_0) = \mathrm{E}\langle r_1 + \gamma r_2 + \gamma^2 r_3 + \ldots \rangle$$

$$V^{\pi}(s_t) = \mathrm{E}\langle r_{t+1} + \gamma V^{\pi}(s_{t+1}) \rangle$$ Equação de Bellman

Este valor esperado não é conhecido, mas é conhecido o **valor actual** (estimado)

$$r_{t+1} + \gamma V(s_{t+1})$$

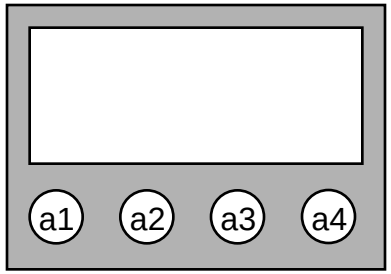
Na aprendizagem por reforço a actualização é feita na direcção desse valor de forma progressiva



[Sutton & Barto, 1998]

Aprendizagem de Valor de Acção

- Exemplo: escolha repetida de diferentes acções
- Por cada acção é obtida uma recompensa
- Resultado depende só da acção escolhida
- **Motivação**
  - **Maximizar a recompensa de longo prazo**



Aprendizagem de Valor de Acção

- Como determinar o valor (Q) de cada acção?
- Valor médio para uma acção k após n tentativas

$$Q_n^k = \frac{r_1^k + r_2^k + r_3^k + ... + r_n^k}{n}$$

- De forma incremental

$$Q_n^k = Q_{n-1}^k + \frac{1}{n}\left[r_n^k - Q_{n-1}^k\right]$$

Aprendizagem de Valor de Acção

- Problemas não estacionários?
- Acumulação não linear
  - Por exemplo, exponencialmente amortecida

$$Q_n^k = Q_{n-1}^k + \alpha[r_n^k - Q_{n-1}^k]$$

$\alpha \in$ [0,1] - Factor de aprendizagem

Dilema Explorar / Aproveitar (*Explore / Exploit*)

- **Quando é que se aprendeu o suficiente para começar a aplicar o que se aprendeu?**
- **Exploração** (*Exploration*)
  - Escolher uma acção que permita explorar o mundo para melhorar a aprendizagem
- **Aproveitamento** (*Exploitation*)
  - Escolher a acção que leva à melhor recompensa de acordo com a aprendizagem
  - Acção Sôfrega (*Greedy*)

Estratégias de Selecção de Acção

- Estratégia *greedy*

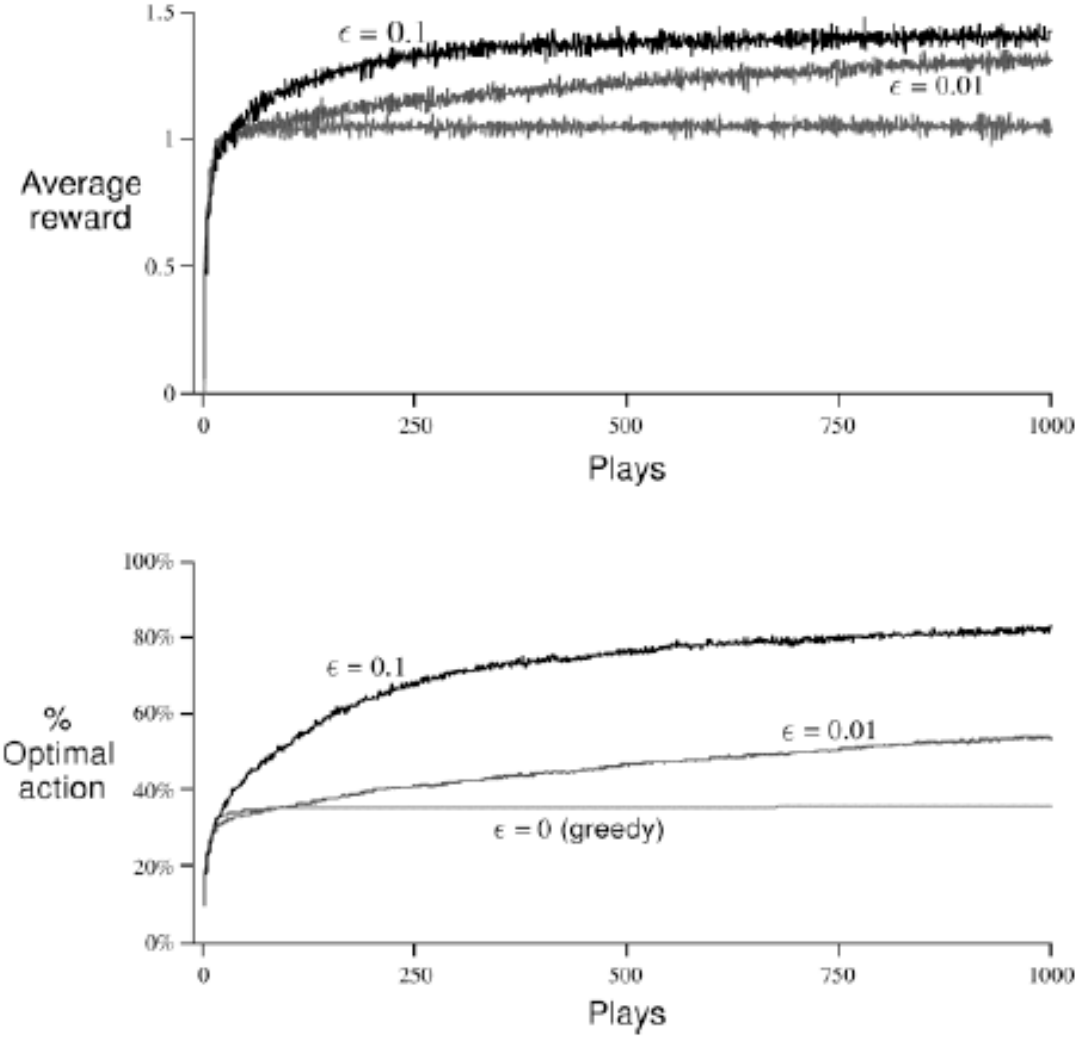
$$a_t = a^*_t = \text{argmax}_a \; Q_t(a)$$

- Estratégia ε-*greedy*

$$a_t = \begin{cases} a^*_t \text{ com probabilidade } 1 - \varepsilon \\ \text{acção aleatória com probabilidade } \varepsilon \end{cases}$$

- Balanceamento de Exploração / Aproveitamento

Exemplo

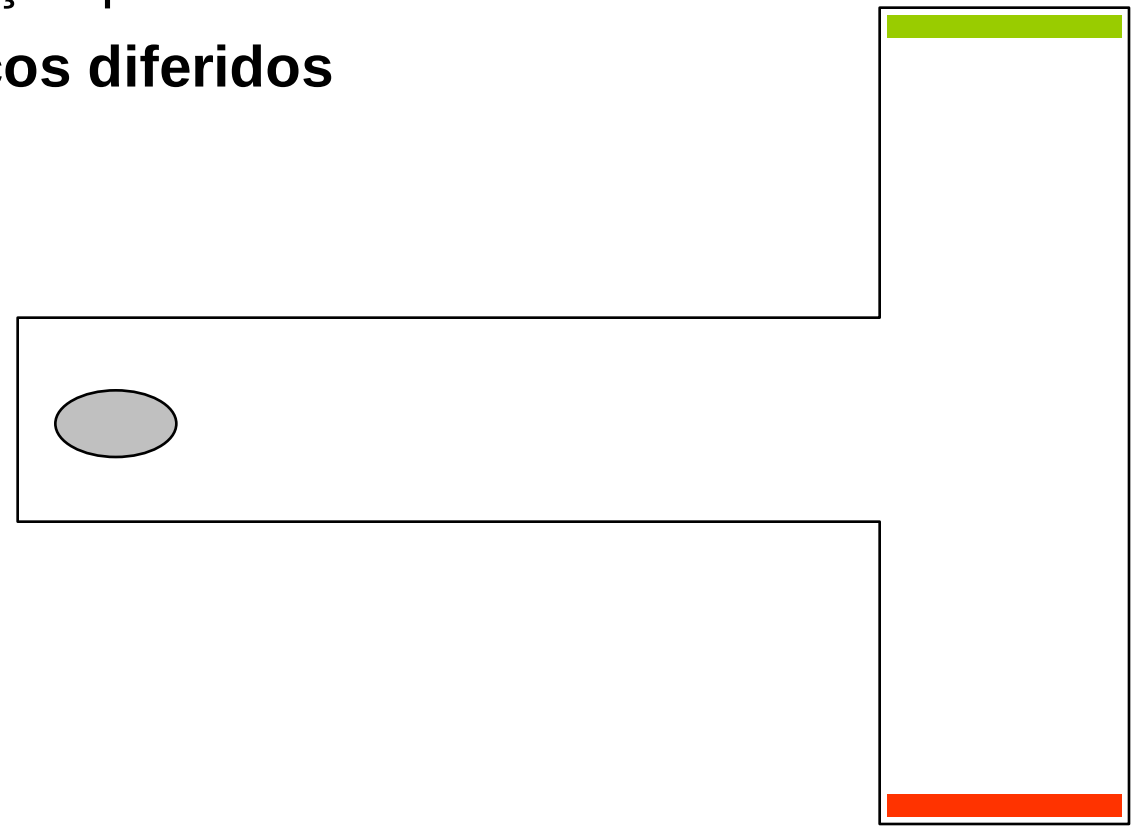


[Sutton & Barto, 1998]

Aprendizagem por Reforço

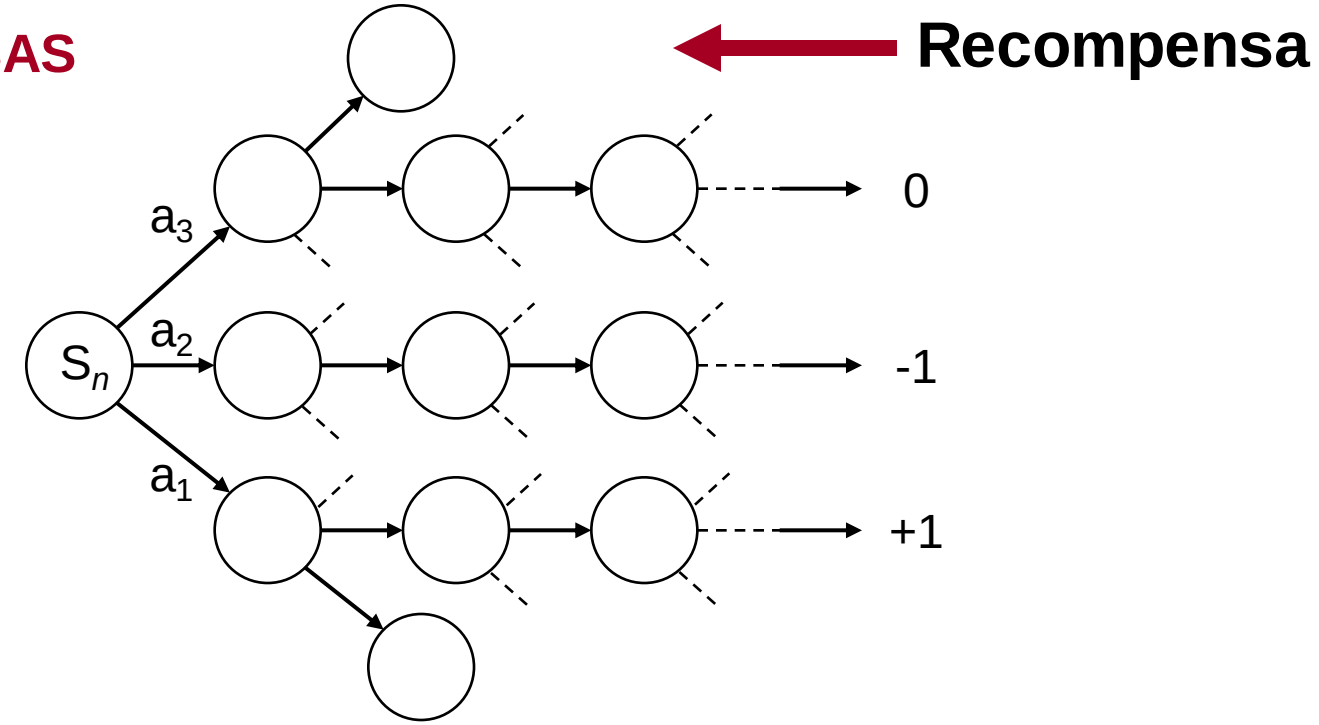
Os reforços podem não ser imediatos

- **Reforços diferidos**



Aprendizagem por Reforço

RECOMPENSAS DIFERIDAS



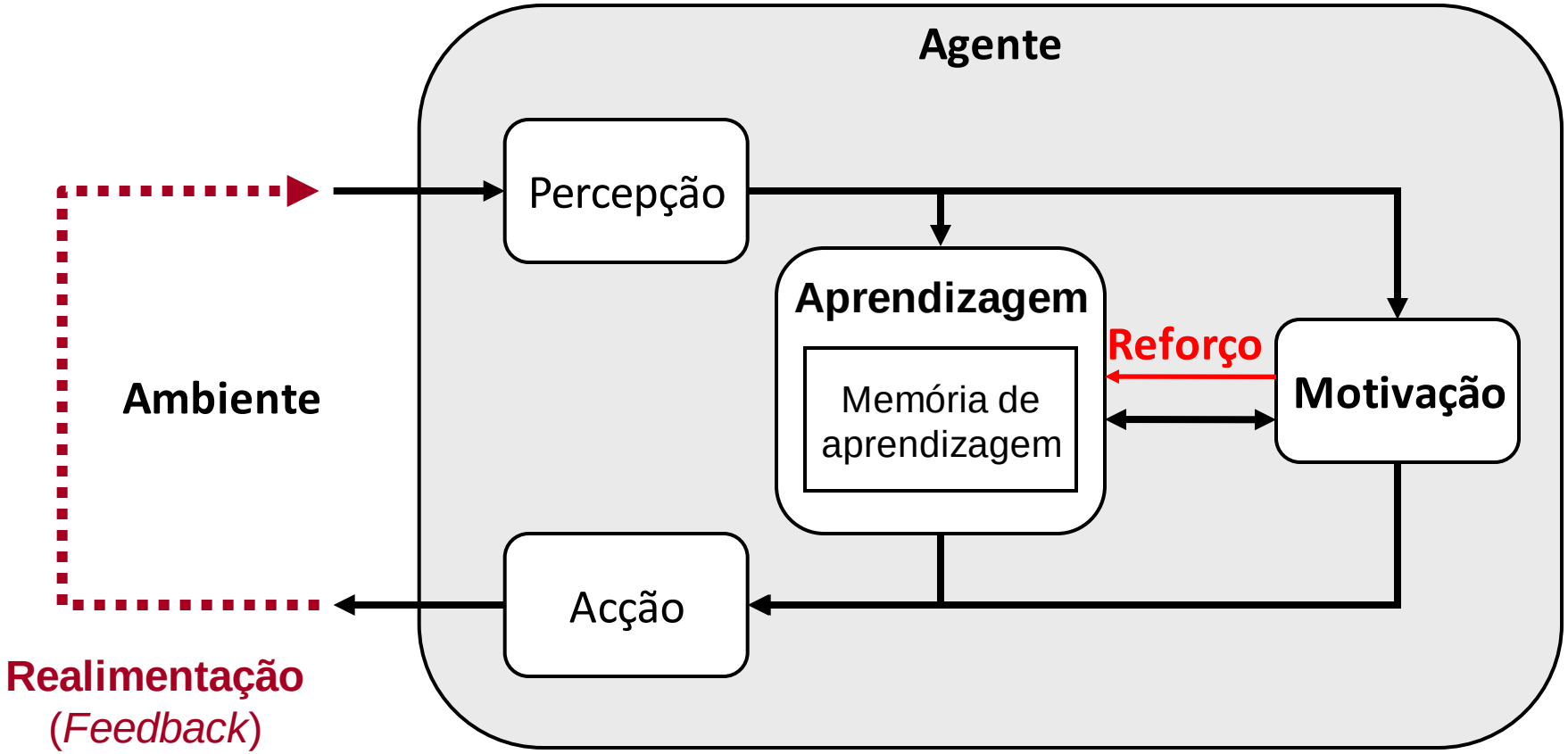
- Aprendizagem incremental a partir da experiência

$$s \rightarrow a \rightarrow r \rightarrow s' \rightarrow a' \rightarrow \ldots$$

Aprendizagem Associativa

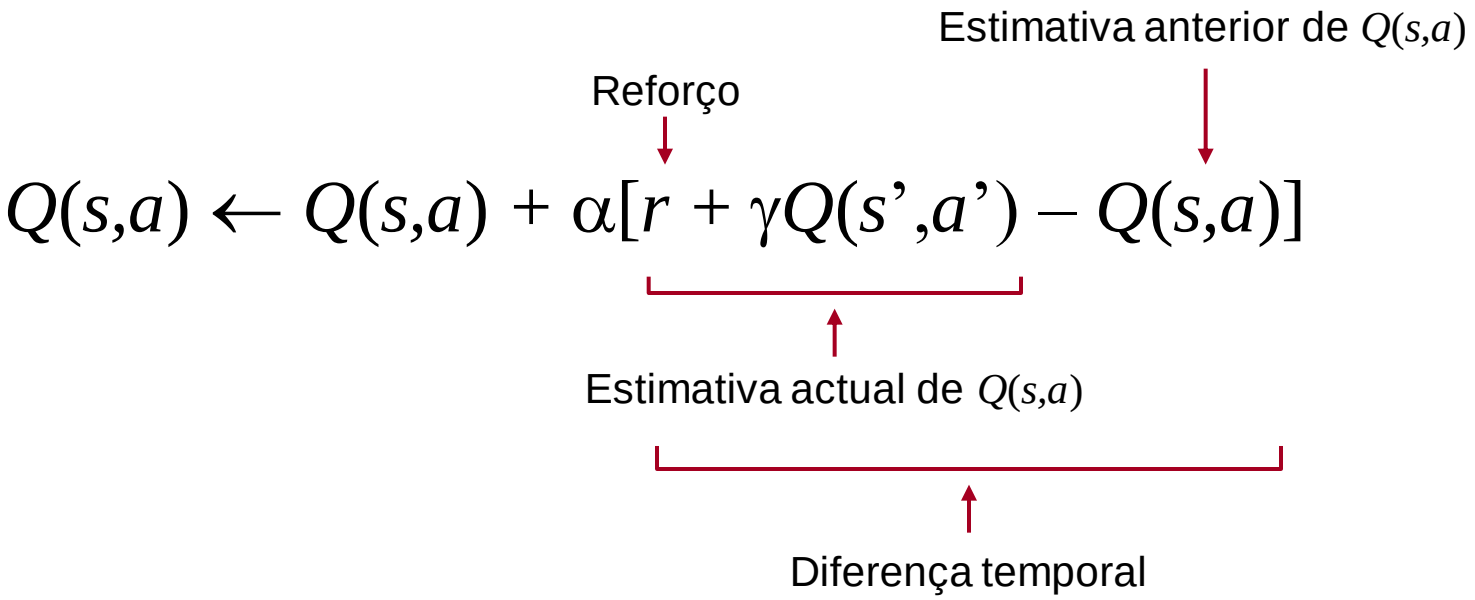
- Estado pode evoluir ao longo do tempo
  - Estados observados
    - $s \in S$
  - Acções realizadas
    - $a \in A$
  - Reforços obtidos
    - $r \in \mathbb{R}$
  - **Valor de num estado realizar uma acção**
    - $Q(s,a)$

Aprendizagem por Reforço



Aprendizagem por Diferença Temporal

Actualização de uma **estimativa** de valor de estado com base na sua mudança (**diferença temporal**) entre instantes sucessivos



Algoritmo *SARSA*

- Iniciar **Q**(***s***,***a***)
- Repetir (por cada episódio)
  - Iniciar ***s***
  - Escolher ***a*** de acordo com ***s*** com base numa política derivada de **Q** (por exemplo ε-*greedy*)
  - Repetir (por cada passo)
    - Executar acção ***a***, observar ***r*** e ***s***'
    - Escolher ***a'*** de acordo com ***s'*** com base numa política derivada de **Q** (por exemplo ε-*greedy*)
    - Actualizar Q:
      $\mathbf{Q}(\boldsymbol{s,a}) \leftarrow \mathbf{Q}(\boldsymbol{s,a}) + \alpha[\boldsymbol{r} + \gamma\,\mathbf{Q}(\boldsymbol{s',a'}) - Q(\boldsymbol{s,a})]$
    - $\boldsymbol{s} \leftarrow \boldsymbol{s}'$, $\boldsymbol{a} \leftarrow \boldsymbol{a}'$
  - Até ***s*** ser um estado terminal

Política Comportamental Óptima

- Função valor de estado-acção

$$Q^{\pi}(s,a)$$

- Valor óptimo

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a)$$

- Política óptima

$$\pi^*(s) = \arg\max_{a} Q^*(s,a)$$

Política "*greedy*" em relação a Q^*

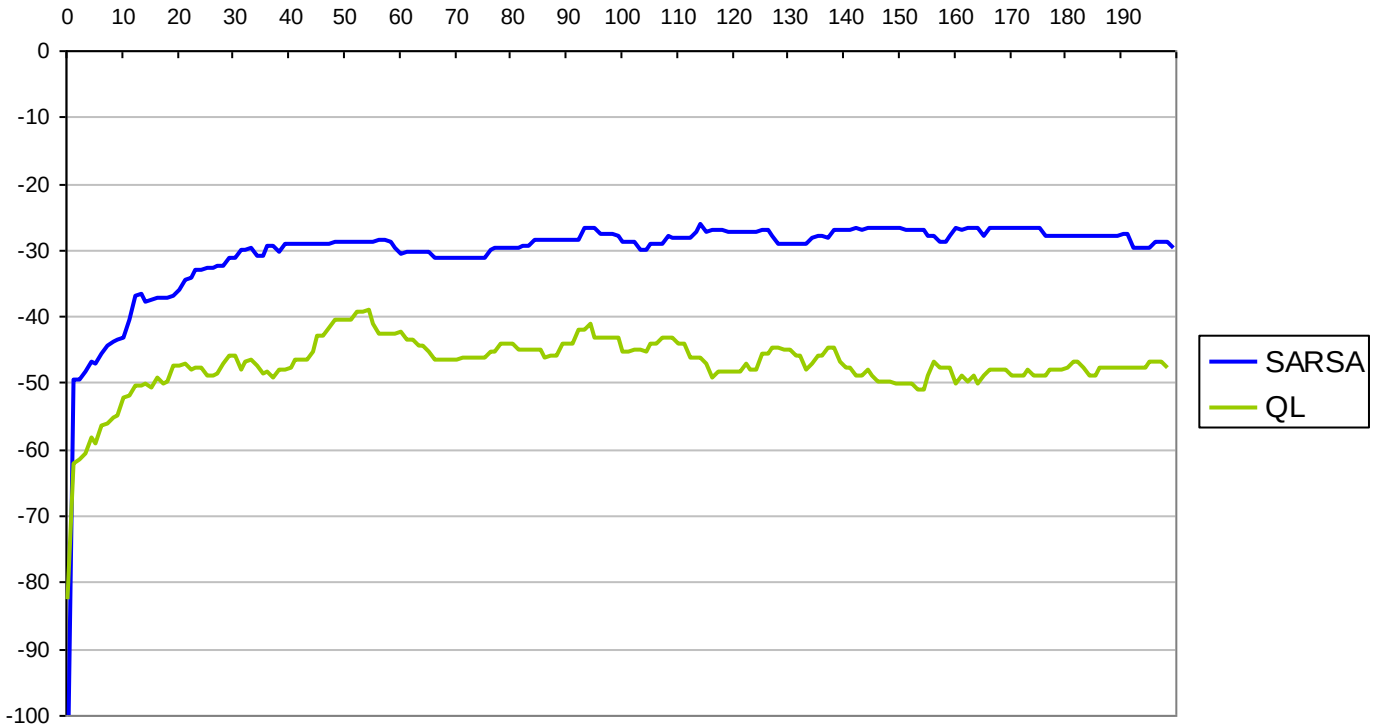
Algoritmo *Q-Learning*

- Iniciar **Q**(***s***,***a***)
- Repetir (por cada episódio)
  - Iniciar ***s***
  - Repetir (por cada passo)
    - Escolher ***a*** de acordo com ***s*** com base numa política derivada de **Q** (por exemplo ε-*greedy*)
    - Executar acção ***a***, observar ***r*** e ***s***'
    - Actualizar **Q**:
      $\mathbf{Q}(\boldsymbol{s},\boldsymbol{a}) \leftarrow \mathbf{Q}(\boldsymbol{s},\boldsymbol{a}) + \alpha[\boldsymbol{r} + \gamma \max_{\boldsymbol{a}'} \mathbf{Q}(\boldsymbol{s}',\boldsymbol{a}') - Q(\boldsymbol{s},\boldsymbol{a})]$
    - $\boldsymbol{s} \leftarrow \boldsymbol{s}'$
  - Até ***s*** ser um estado terminal

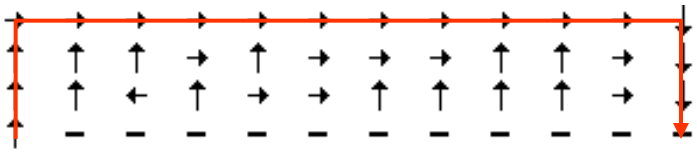
Processo de Aprendizagem

- Dois tipos de aprendizagem
  - **Política de selecção de acção única (*On-policy*)**
    - Utilização da mesma política de selecção de acção para comportamento e para propagação de valor
    - Exploração de todas as acções (e.g. política ε-*greedy*)
  - **Políticas de selecção de acção diferenciadas (*Off-policy*)**
    - Utilização da mesma política de selecção de acção para comportamento e para propagação de valor
    - **Optimização da função valor $Q(s,a)$**

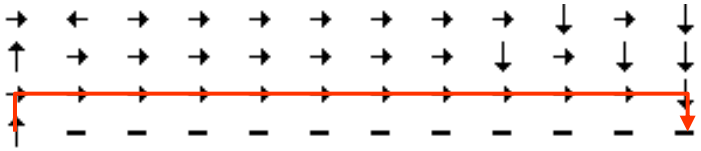
SARSA vs. *Q-Learning*



SARSA



QL



Dilema Explorar / Aproveitar

- Para convergir para o valor óptimo
  - Não se pode apenas explorar
  - Não se pode apenas aproveitar
- Estratégia Sôfrega (*Greedy*)
  - Mínimos/máximos locais
- Nunca se pode parar de explorar
  - Convergência assimptótica
- Deve-se progressivamente reduzir a exploração
  - GLIE (*Greedy in the Limit of Infinite Exploration)*

Algoritmo *Q-Learning*

- **Propriedades**
  - Os valores da matriz *Q* **convergem** no limite se:
    - Se cada par estado-acção (*s*,*a*) for visitado um número ilimitado de vezes
    - O parâmetro α tender para 0 no limite
  - No limite a estratégia *greedy* de aproveitamento de *Q*(*s*,*a*) converge para a **política óptima**
- **Estes requisitos podem ser satisfeitos através de:**
  - $\alpha(s,a) \approx 1/k$
    - Sendo k o número de vezes que a acção ***a*** foi seleccionada em ***s***
  - Estratégia de selecção de acção ε-*greedy* com $\varepsilon \approx 1/t$
    - Sendo ***t*** função do tempo ou do número de tentativas de aprendizagem

Aprendizagem por Reforço

- Problemas
  - Complexidade dos espaços de estados
  - Tempo de convergência
- Soluções
  - Utilização de memória episódica
  - Utilização de modelos do mundo
  - Arquitecturas híbridas
  - Generalização / Abstracção
  - Racionalidade confinada ("*bounded rationality*")

Referências

[Russel & Norvig, 2003]
S. Russell, P. Norvig, “Artificial Intelligence: A Modern Approach”, 2nd Edition, Prentice Hall, 2003

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

[Fox *et al.*, 1994]
G. Fox, R. Williams, P. Messina, “Parallel Computing Works”, Morgan Kaufmann, 1994

[Poole & Mackworth, 2010]
D. Poole, A. Mackworth, Artificial Intelligence: Foundations of Computational Agents, Cambridge University Press, 2010

[Scamell-Katz, 2009]
S. Scamell-Katz, “Breaking the Habit”, Retail & Shopper, 2009

[Chris Barnard, 2003]
C. Barnard, “Animal Behaviour: Mechanism, Development, Ecology and Evolution”, Prentice Hall, 2003