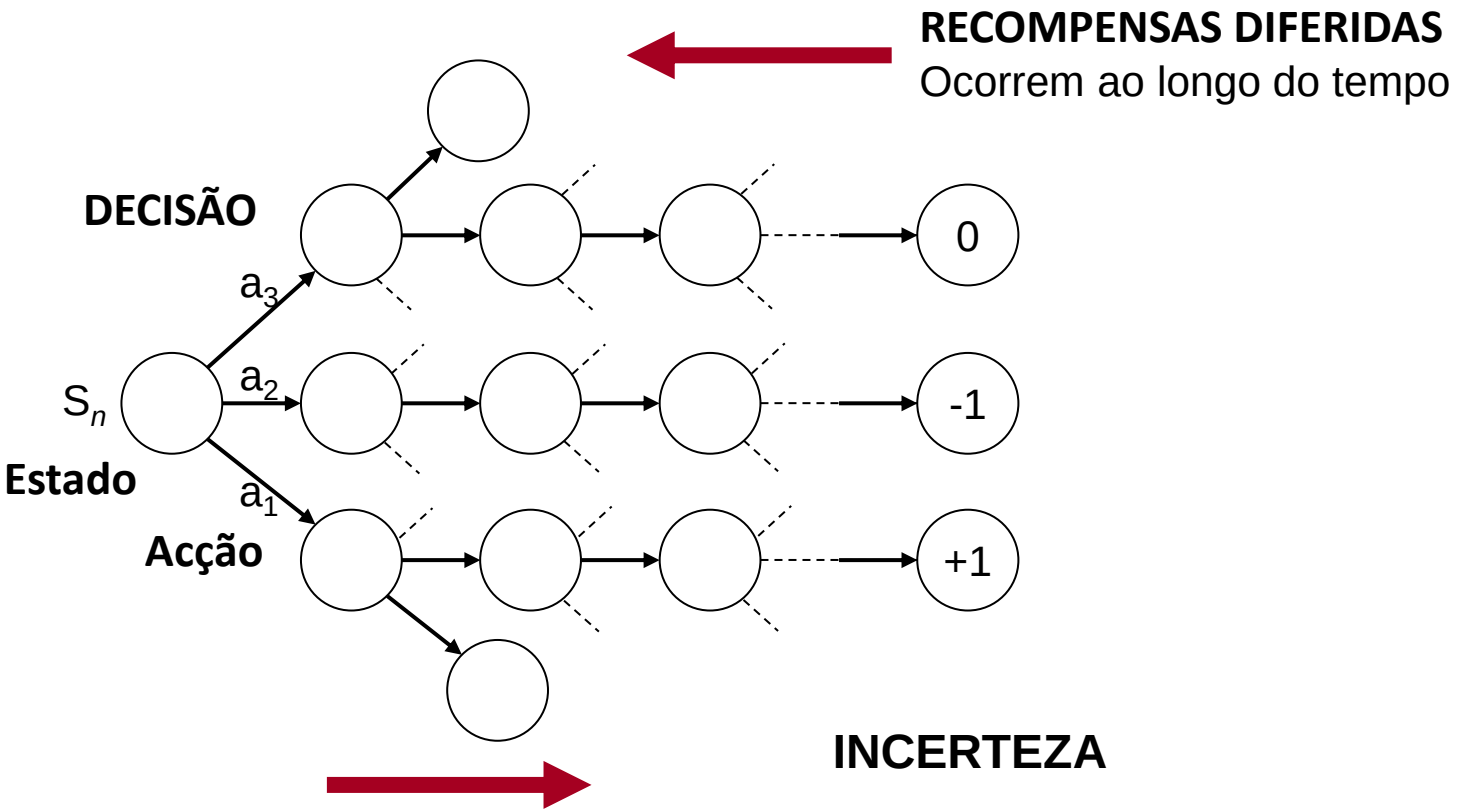
APRENDIZAGEM POR REFORÇO

Luís Morgado

2024

Problemas de Decisão Sequencial

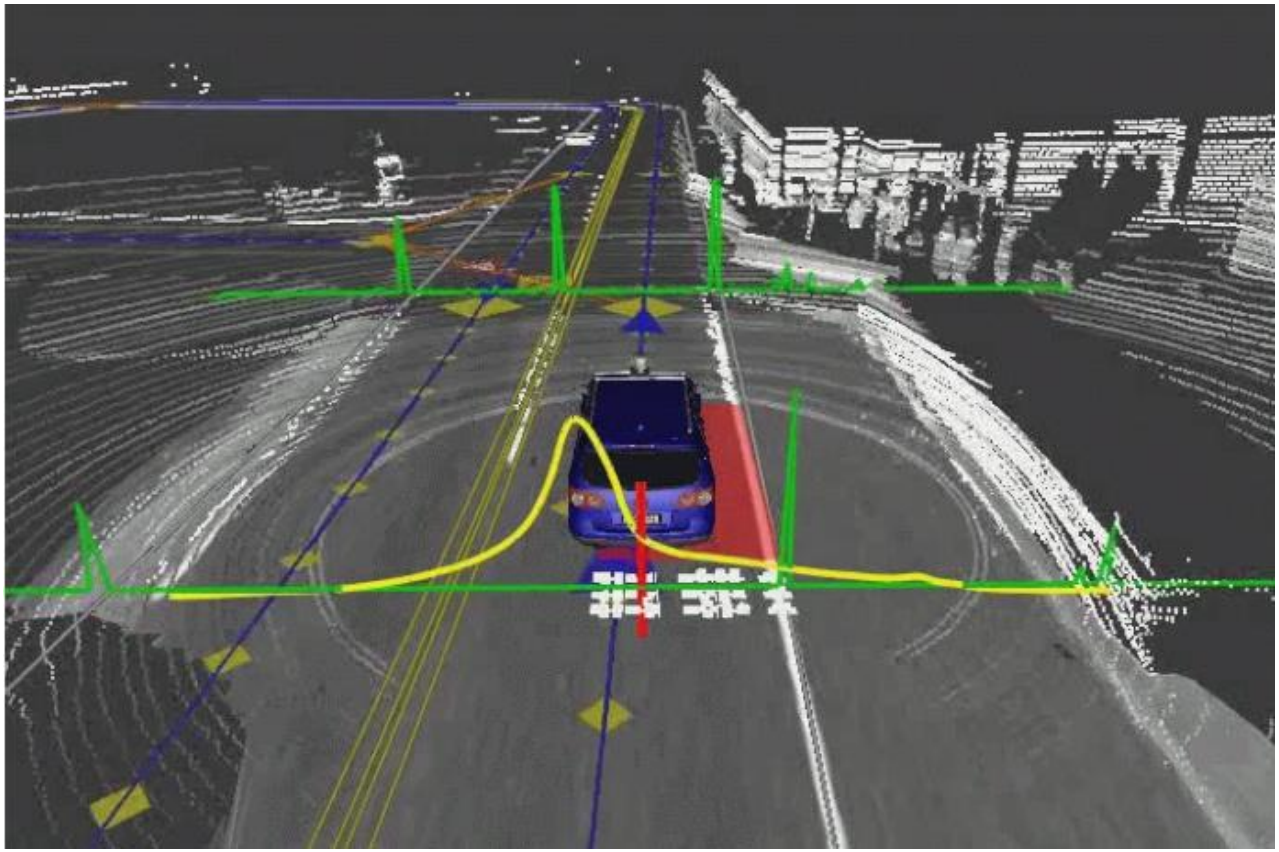
Problemas em que o valor das acções de um agente (*ganhos* e *perdas*) não depende de decisões simples, baseadas apenas no estado actual, mas de uma sequência de acções encadeadas no tempo, podendo os resultados das acções ser incertos, ou seja, não totalmente controlados (*não determinísticos*)



Tomada de Decisão com Incerteza

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

Exemplo: Navegação em veículos autónomos



[Montemerlo, 2008]

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 $s \in S$

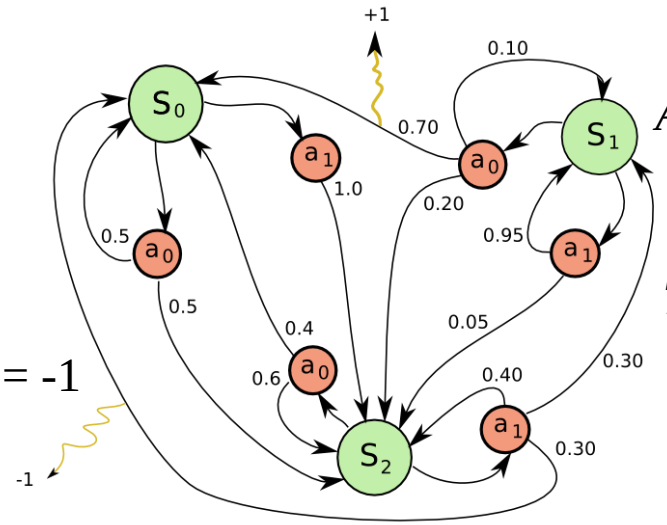
$T(s, a, s')$ – probabilidade de transição de s para s' através de a

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

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

Exemplo:

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



$A(s_1) = \{a_0, a_1\}$

$T(s_2, a_1, s_1') = 0.3$

$R(s_2, a_1, s_0') = -1$

Processos de Decisão de Markov

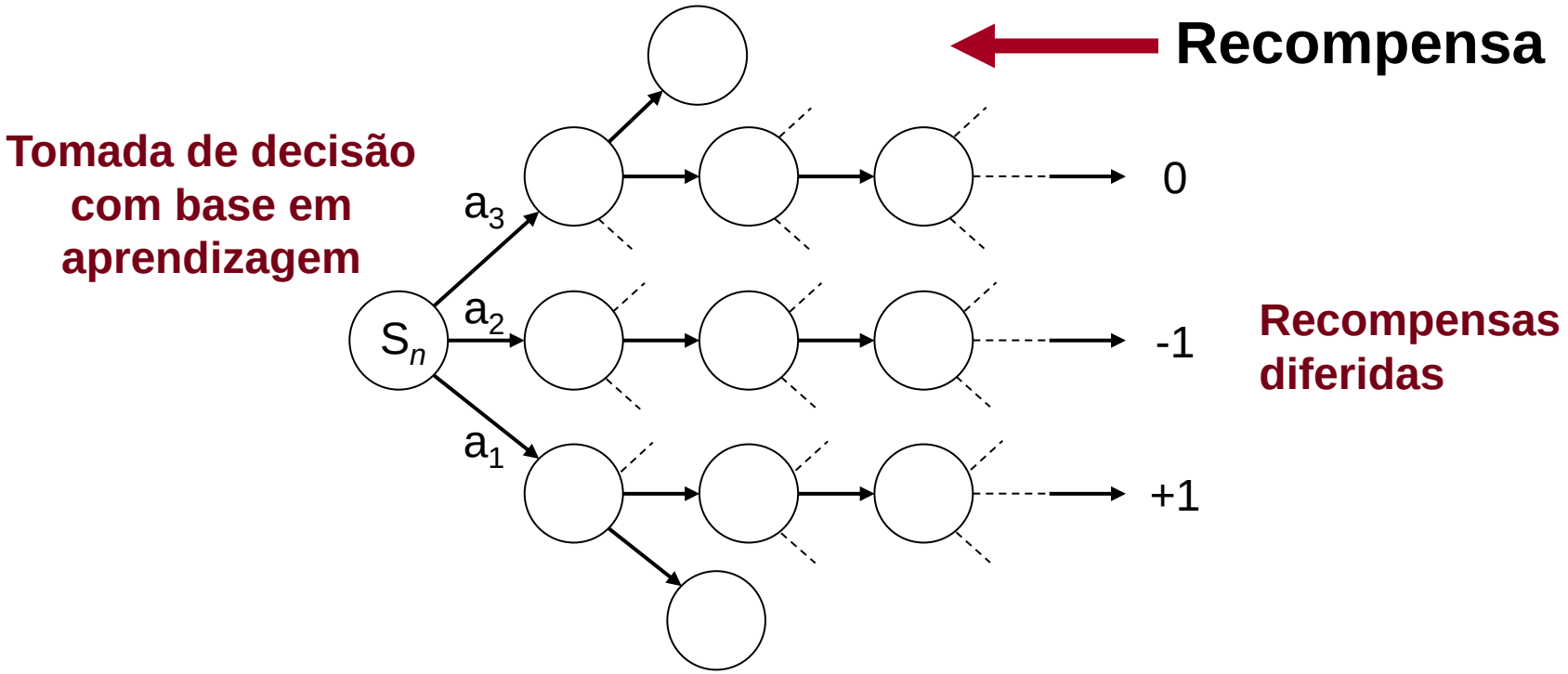
Problemas

- Dimensão dos espaços de estados (*problema da dimensionalidade*)
- Dificuldade de definição de modelos do mundo (por exemplo a partir de dados experimentais)
- Modelos de transição ***T(s, a, s')*** e de recompensa ***R(s, a, s')*** desconhecidos

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

? ?

Aprendizagem por Reforço



- Aprendizagem incremental a partir da experiência

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

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 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

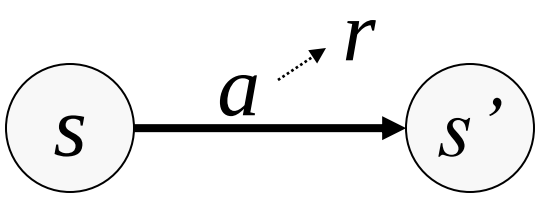
- Aprendizagem
  - **Generalização**
  - Formação de **abstracções** (modelos)
    - Protótipos
    - Conceitos
    - Padrões comportamentais

Aprendizagem Automática

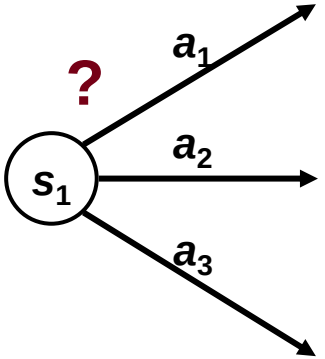
- Aprendizagem **conceptual**
  - **O que é?**
    - **Conceito**
  - Supervisionada
  - Não supervisionada
- Aprendizagem **comportamental**
  - **O que fazer?**
    - **Comportamento (acção)**
  - **Aprendizagem por reforço**

Aprendizagem por Reforço

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

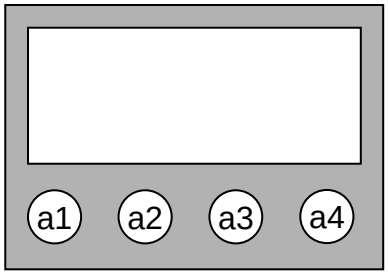


- Aprendizagem de **comportamentos**
  - O que fazer
  - Relação entre situações e acções
    - **Políticas** (de acção)

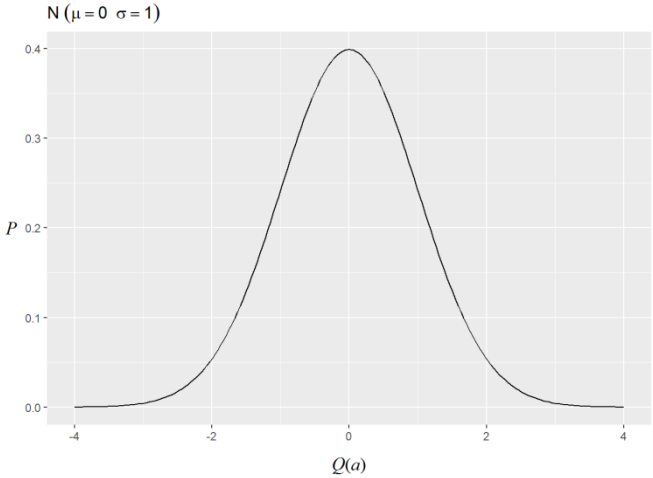


Aprendizagem de Valor de Acção

- **Exemplo**
  - Escolha repetida de diferentes acções
  - Por cada acção é obtida uma recompensa
    - De acordo com uma determinada distribuição de probabilidades
  - Resultado depende só da acção escolhida
- **Motivação**
  - **Maximizar a recompensa de longo prazo**



Exemplo: Decisão entre várias acções com valores desconhecidos



Exemplo: Distribuição aleatória dos valores $Q(a)$ das acções

Aprendizagem de Valor de Acção

- Como determinar o valor $Q(a)$ de cada acção?
- **Valor médio** para uma acção a após n tentativas
  - Cada tentativa produz uma recompensa r_n

$$Q_n(a) = \frac{r_1^a + r_2^a + \cdots + r_n^a}{n}$$

O cálculo da estimativa de valor de cada acção é realizado após *n* tentativas de realização da acção *a*

- **Requer a realização de *n* tentativas para cada acção** e da memorização dos respectivos resultados
- **Só após *n* tentativas para cada acção** é possível obter uma estimativa do valor de cada acção

Este tipo de abordagem não permite aproveitar o conhecimento obtido de forma incremental

Aprendizagem de Valor de Acção

- Valor médio para uma acção a após n tentativas
  - Cada tentativa produz uma recompensa r_n

$$Q_n(a) = \frac{r_1^a + r_2^a + \cdots + r_n^a}{n}$$

Cálculo da estimativa de valor de cada acção **após *n* tentativas de realização da acção *a***

- De forma incremental, mantendo memória da estimativa anterior do valor de cada acção $Q_{n\text{-}1}(a)$

Estimativa anterior do valor de uma acção ***a***

$$Q_n(a) = Q_{n-1}(a) + \frac{1}{n}[r_n^a - Q_{n-1}(a)]$$

Diferença entre a recompensa observada e a estimativa de valor da acção ***a***

Cálculo incremental da estimativa de valor de cada acção

Aprendizagem de Valor de Acção

- **Problemas não estacionários?**
  - A distribuição de probabilidades muda com o tempo
- Estimação por acumulação não linear
  - Por exemplo, exponencialmente amortecida

$$Q(a)_n = Q(a)_{n-1} + \alpha[r_n^a - Q(a)_{n-1}]$$

O factor α substitui **1/*n*** possibilitando uma ponderação não linear das recompensas ao longo do tempo

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

Determina a relevância das recompensas em função do tempo

$\alpha \to 0$: Maior relevância das recompensas mais antigas

$\alpha \to 1$: Maior relevância das recompensas mais recentes

Dilema Explorar / Aproveitar

- **Aprendizagem por reforço**
  - **Explorar:** para obter conhecimento
  - **Aproveitar:** para maximizar valor
- **Exploração**
  - Escolher uma acção que permita explorar o mundo para melhorar a aprendizagem
- **Aproveitamento**
  - Escolher a acção que leva à melhor recompensa de acordo com a aprendizagem
    - **Acção sôfrega** (*greedy*)
- **Quando é que se *explorou* o suficiente para começar a *aproveitar* o que se aprendeu?**

Estratégias de Selecção de Acção

- Estratégia sôfrega (*greedy*)
  - Escolher a melhor acção a^* de acordo com a estimativa de actual valor de cada acção $Q_t(a)$

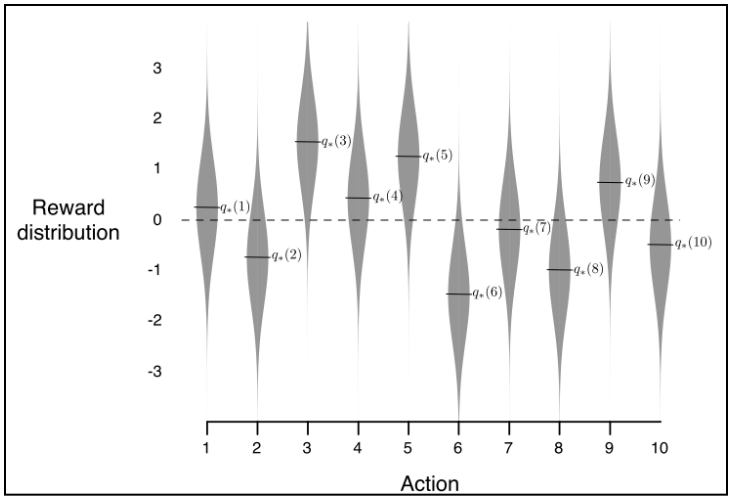
$$a_t = a^*_t = \text{argmax}_a \, Q_t(a)$$

- Estratégia ε-*greedy*

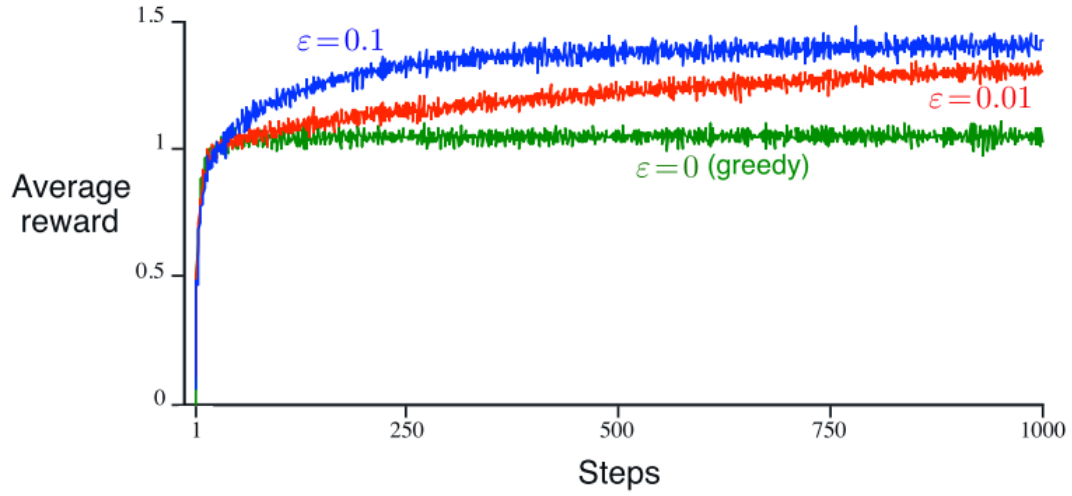
$$a_t = \begin{cases} a^*_t \text{ com probabilidade } 1 - \varepsilon \\ \text{acção aleatória com probabilidade } \varepsilon \end{cases}$$

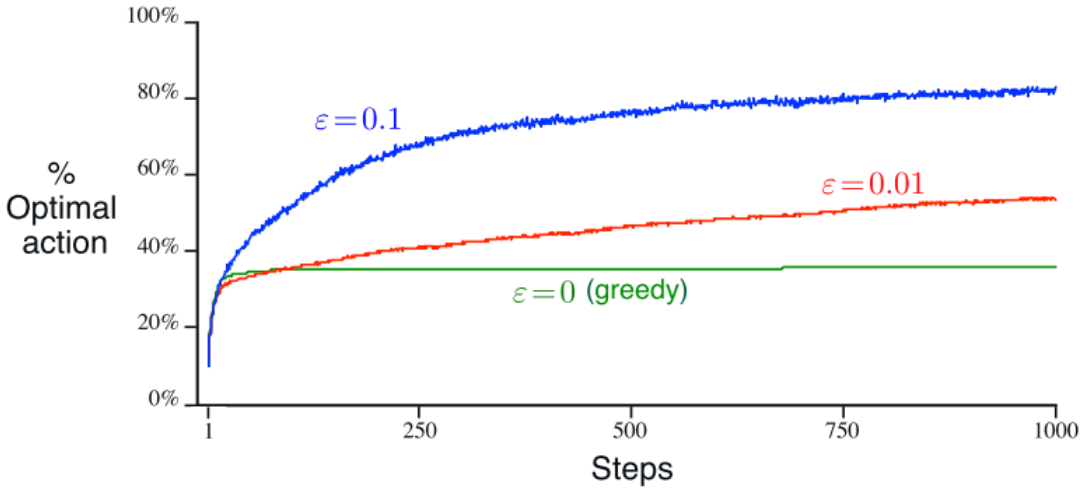
  - Balanceamento de **Exploração** / **Aproveitamento**

Exemplo: Estratégia *greedy* vs. ε-*greedy*



Distribuições das recompensas associadas a cada acção





[Sutton & Barto, 2020]

Aprendizagem por Reforço

- **Aprendizagem associativa**
  - Estados observados
    - $s \in S$
  - Acções realizadas
    - $a \in A$
  - Reforços obtidos
    - $r \in \mathbb{R}$
  - **Valor de num estado realizar uma acção**
    - $Q(s,a) \in \mathbb{R}$

$q = Q(s,a)$ valor de no estado s realizar a acção a

$$q = Q(s,a)$$

Associação de uma estimativa de valor $Q(s,a)$ a cada par estado-acção (s,a) que representa o valor de no estado s realizar a acção a

Aprendizagem por Reforço

Valor de realizar uma acção num determinado estado

Valor Estado-Acção: $Q(s,a)$

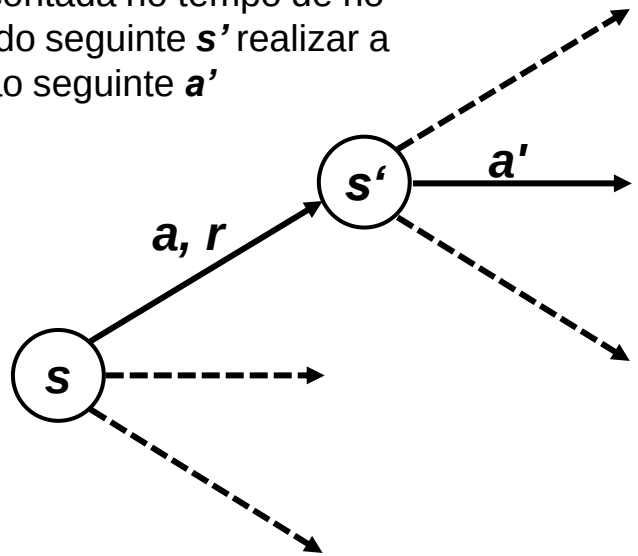
Reforço
(recompensa observada)

$$Q'(s,a) = r + \gamma Q(s',a')$$

Valor estimado com base na recompensa observada

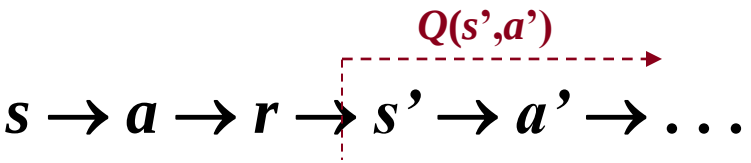
Estimativa actual de $Q(s,a)$

O valor estimado ***Q'(s,a)*** de realizar a acção ***a*** no estado ***s*** corresponde à soma do reforço obtido (recompensa), com a estimativa de valor ***Q(s',a')*** descontada no tempo de no estado seguinte ***s'*** realizar a acção seguinte ***a'***



Aprendizagem incremental a partir da experiência

Q(s',a') representa o valor futuro (a longo prazo)



Aprendizagem por Diferença Temporal

Estimativa incremental de valor de acção por diferença temporal entre o valor estimado com base na recompensa observada $\boldsymbol{Q'(s,a)}$ e o valor estimado anterior $\boldsymbol{Q(s,a)}$

Diferença temporal

$$Q(s,a) \leftarrow Q(s,a) + \alpha[Q'(s,a) - Q(s,a)]$$

$$\downarrow$$

$$Q'(s,a) = r + \gamma Q(s',a')$$

$$\downarrow$$

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s',a') - Q(s,a)]$$

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

Determina a relevância das recompensas em função do tempo

$\alpha \rightarrow 0$: Maior relevância das recompensas mais antigas

$\alpha \rightarrow 1$: Maior relevância das recompensas mais recentes

Aprendizagem por Diferença Temporal

Estimação de valor por acumulação de recompensas de forma não linear para lidar com **ambientes não estacionários** (regulada pelo factor de aprendizagem $\alpha \in [0,1]$)

Actualização de uma estimativa de **valor de estado-acção** $\boldsymbol{Q(s,a)}$ com base na sua mudança (*diferença temporal*) entre instantes sucessivos

Estimativa anterior de $Q(s,a)$

Reforço

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s',a') - Q(s,a)]$$

Estimativa actual de $Q(s,a)$

Diferença temporal

Algoritmo *SARSA*

Aprendizagem incremental a partir da experiência

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

1. Iniciar $Q(s,a)$
2. Repetir (por cada episódio)
3. Iniciar s
4. Escolher a de acordo com s com base numa política derivada de Q (por exemplo ε-*greedy*)
5. Repetir (por cada passo)
6. Executar acção a, observar r e s'
7. Escolher a' de acordo com s' com base numa política derivada de Q (por exemplo ε-*greedy*)
8. Actualizar Q:
$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma\, Q(s',a') - Q(s,a)]$$
9. Actualizar $s \leftarrow s'$, $a \leftarrow a'$
10. Até s ser um estado terminal

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)*

Referências

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

[Russel & Norvig, 2020]
S. Russell, P. Norvig, “Artificial Intelligence: A Modern Approach”, 4th Edition, Pearson, 2020

[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