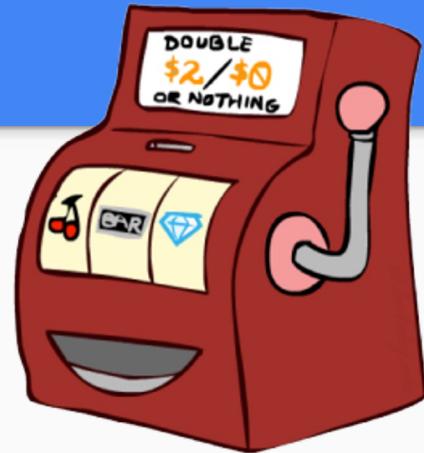
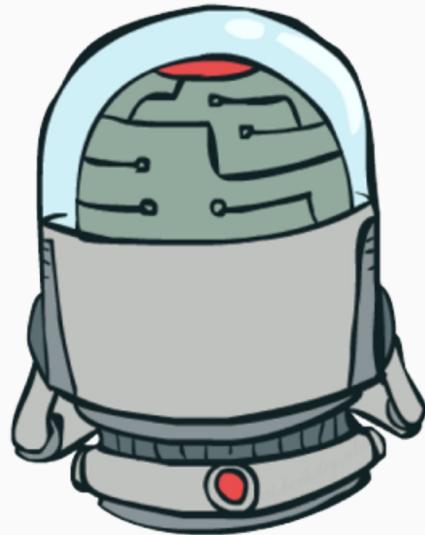


# Aprendizagem por Reforço

slides baseados na disciplina CS188 de UC Berkeley

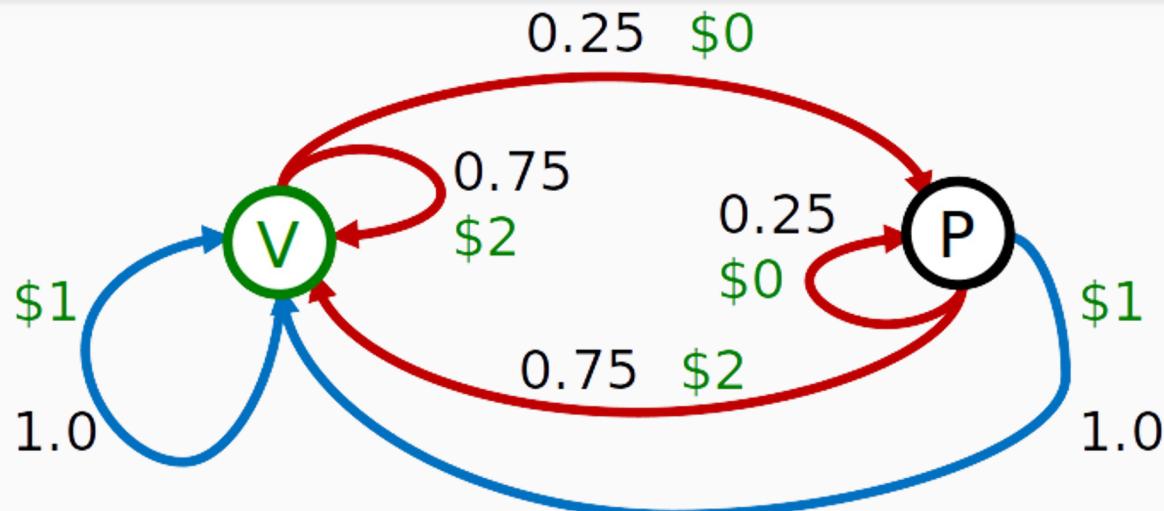
# Máquinas de Cassino



# MDP

Ações: azul, vermelho

Estados: venceu, perdeu



Pode jogar 100 vezes e não tem desconto

# Planejamento Offline

A solução do MDP é offline

Determina os valores de  $V^*$  através de cálculos.

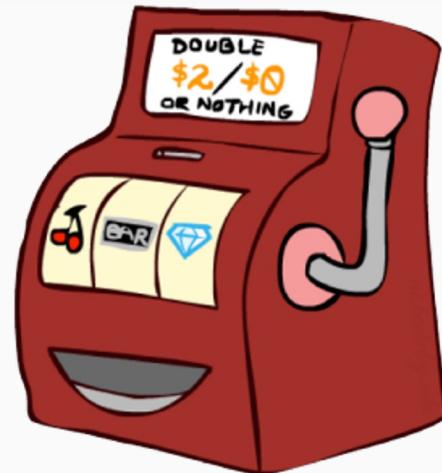
Precisa conhecer o MDP

Não precisa jogar o jogo.

	Valor
Vermelho	150
Azul	100



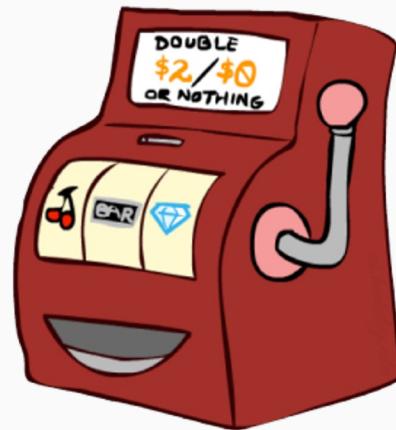
# Exemplo



\$2	\$2	\$0	\$2	\$2
\$2	\$2	\$0	\$0	\$0

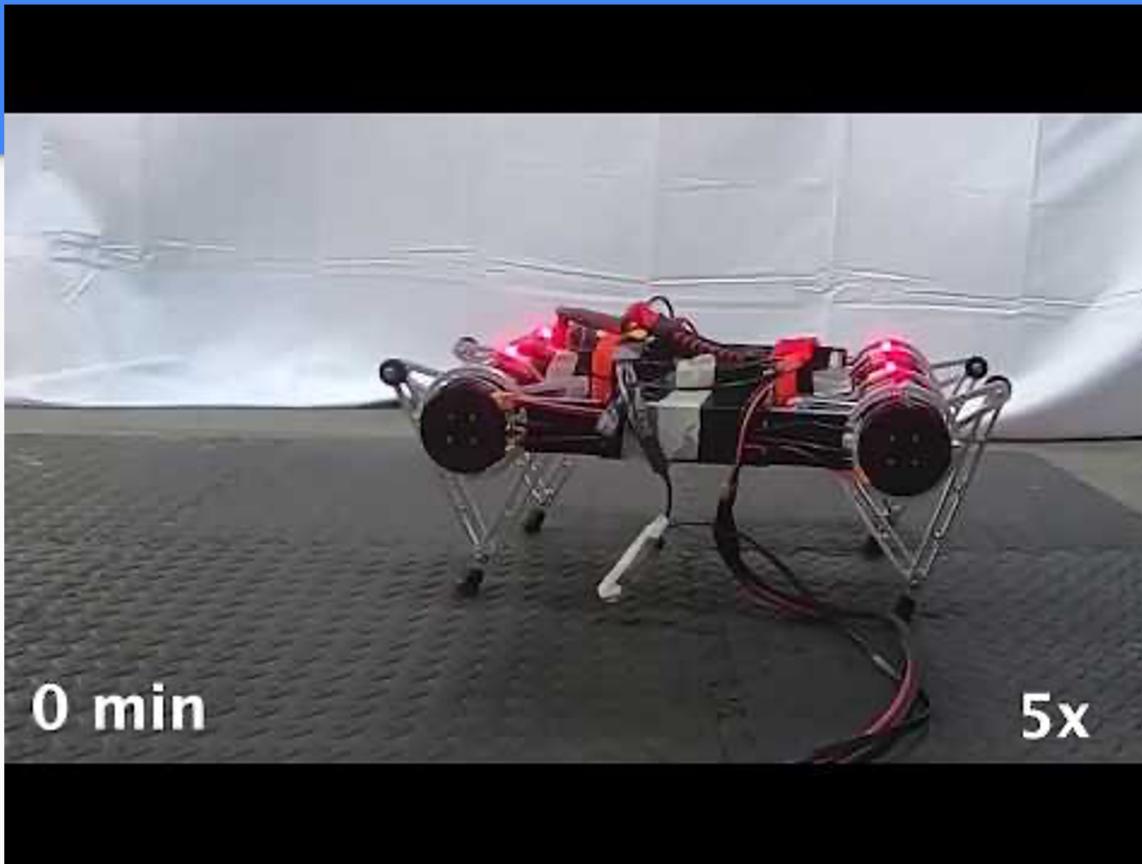
# Exemplo

E se não soubermos as probabilidades de retorno da máquina **vermelha**?

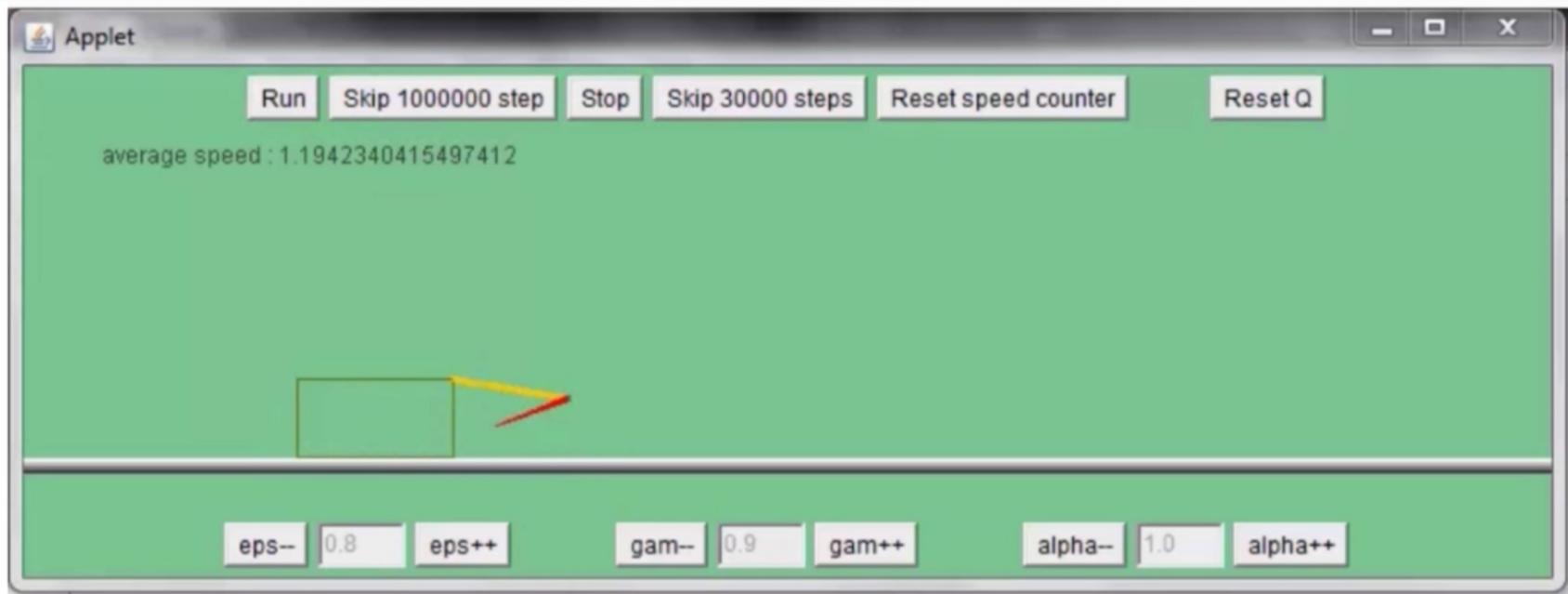


\$0	\$0	\$0	\$2	\$0
\$2	\$0	\$0	\$0	\$0

# Aprendendo a Caminhar



# Robô “Rastejador”



# Aprendizagem por Reforço

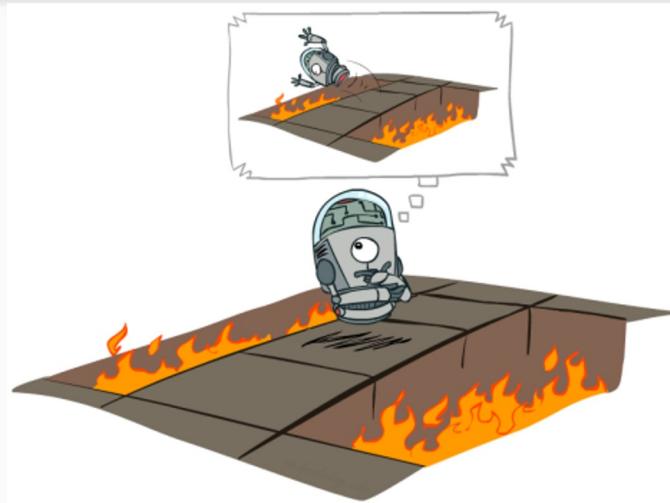
Processos de Decisão de Markov (MDP).

- Conjunto de estados  $s \in S$
- Conjunto de ações  $a \in A$
- Função de transição  $T(s, a, s')$  ou  $P(s'|s, a)$
- Função de recompensa  $R(s, a, s')$  (e fator  $\gamma$ )
- Um estado inicial

Objetivo ainda é encontrar política  $\pi^*$

Diferença:  $T(s, a, s')$  e  $R(s, a, s')$  são desconhecidas.

# MDP (offline) e RL (online)



solução offline



Aprendizagem  
Online

# Aprendizagem Baseada em Modelos

Aprender modelo baseado em experiência e solucionar o problema usando o modelo aproximado.

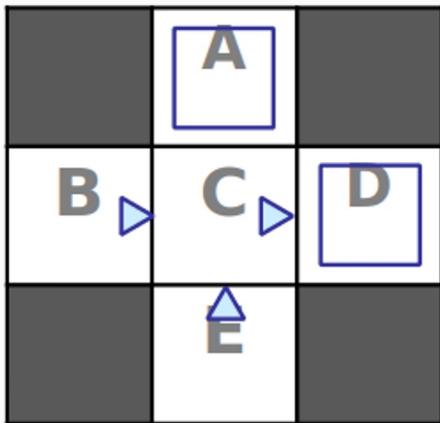
## Aprender um modelo MDP empírico

- Conte o número de  $s'$  resultantes de  $s, a$  e obtenha assim uma aproximação  $\hat{T}(s, a, s')$
- Descubra cada  $\hat{R}(s, a, s')$  obtido ao observar  $(s, a, s')$
- Use iteração de valor para solucionar o problema aproximado.



# Aprendizagem Baseada em Modelos

política de  
entrada  $\pi$



$$\gamma = 1$$

Episódios observados  
(Treinamento)

Episódio 1

B, dir, C, -1  
C, dir, D, -1  
D, sair, x,  
+10

Episódio 2

B, dir, C, -1  
C, dir, D, -1  
D, sair, x,  
+10

Episódio 3

E, cima, C, -1  
C, dir, D, -1  
D, sair, x, +10

Episódio 4

E, cima, C, -1  
C, dir, A, -1  
A, sair, x, -10

Modelo  
Aprendido

$\hat{T}(s, a, s')$

$\hat{T}(B, \text{dir}, C) = 1.00$   
 $\hat{T}(C, \text{dir}, D) = 0.75$   
 $\hat{T}(C, \text{dir}, A) = 0.25$   
...

$\hat{R}(s, a, s')$

$\hat{R}(B, \text{dir}, C) = -1$   
 $\hat{R}(C, \text{dir}, D) = -1$   
 $\hat{R}(D, \text{sair}, x) = +10$   
...

# Exemplo: Cálculo de Idade

P(A) é conhecido

$$E[A] = \sum_a P(a) \cdot a = 0.35 \times 20 + \dots$$

Se P(A) não for conhecido, colete amostras  $[a_1, a_2, \dots, a_N]$

Baseado em Modelo

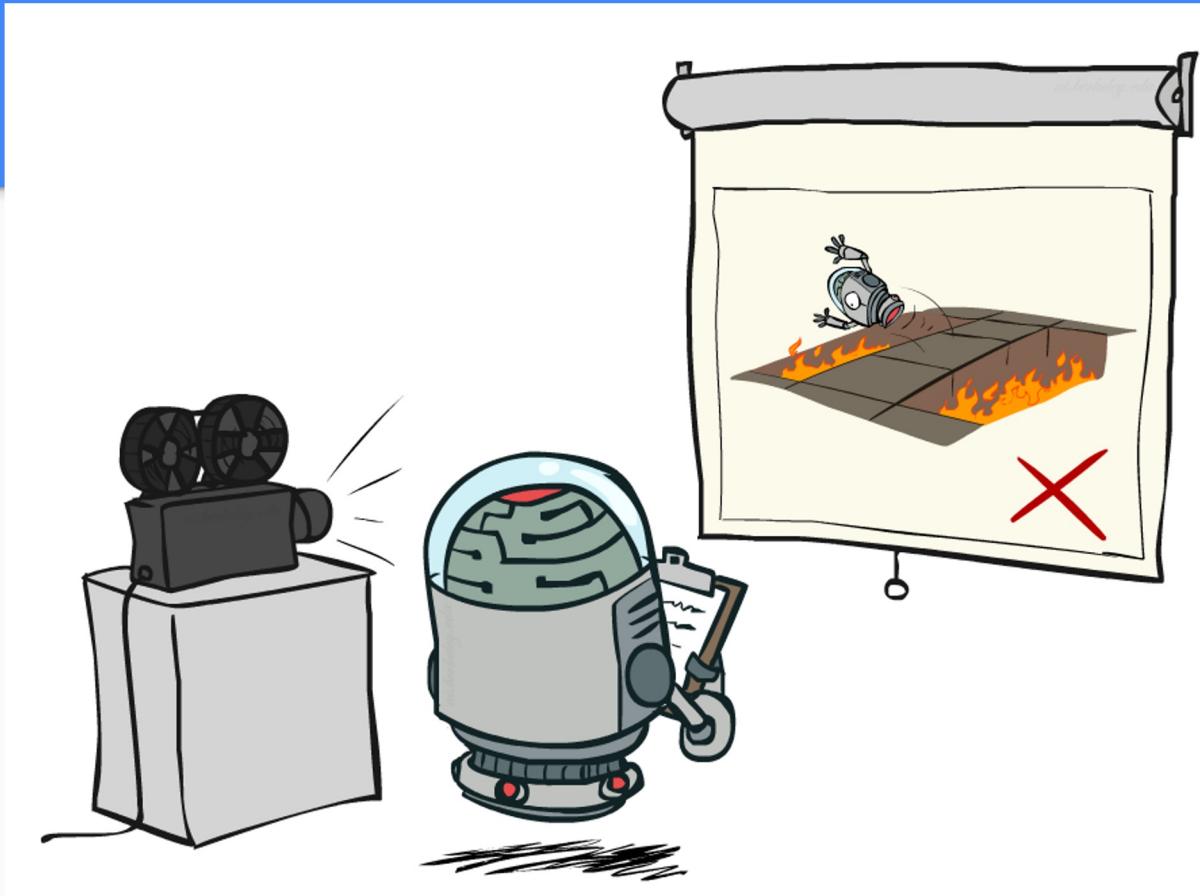
$$\hat{P}(a) = \frac{\text{num}(a)}{N}$$

$$E[A] \approx \sum_a \hat{P}(a) \cdot a$$

Livre de Modelo

$$E[A] \approx \frac{1}{N} \sum_i a_i$$

# Aprendizagem por Reforço Passivo X Ativo



# Aprendizagem por Reforço Passivo

Problema simplificado: avaliação de política

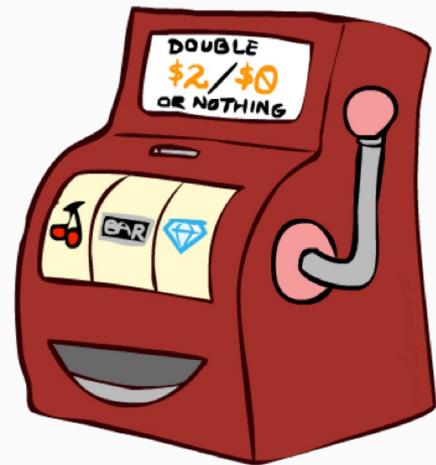
- Política  $\pi$  é fornecida como entrada
- Transições  $T(s, a, s')$  e recompensas  $R(s, a, s')$  são desconhecidas
- Objetivo: aprender os valores de  $V\pi(s)$

Aprendiz observa o que acontece ao seguir uma política no mundo real.  
Não faz escolha de ações.  
As ações são tomadas no mundo real (não é offline).

# Avaliação Direta

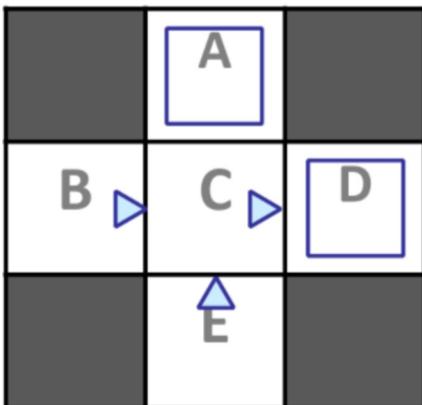
Para uma dada política  $\pi$ , calcular o valor de  $V\pi(s)$ .

1. Siga  $\pi$ ;
2. Observe o valor descontado da recompensa a partir de um determinado estado.
3. Calcule a média dos valores observados.



# Avaliação Direta

política de  
entrada  $\pi$



Episódios observados  
(Treinamento)

Episódio 1

- B, dir, C, -1
- C, dir, D, -1
- D, sair, x, +10

Episódio 2

- B, dir, C, -1
- C, dir, D, -1
- D, sair, x, +10

Episódio 3

- E, cima, C, -1
- C, dir, D, -1
- D, sair, x, +10

Episódio 4

- E, cima, C, -1
- C, dir, A, -1
- A, sair, x, -10

Valores de Saída

	A -10	
B +8	C +4	D+10
	E -2	

# Avaliação Direta

- Vantagens

1. Fácil de entender
2. Não precisa conhecer T e R
3. Eventualmente converge.

- Desvantagens

1. Não faz uso da estrutura do problema (conexões).
2. Valor de cada estado é calculado separadamente.
3. Demora muito para convergir.

Valores de Saída

	A -10	
B +8	C +4	D+10
	E -2	

tanto B quanto E  
passam por C, como  
podem ter valores  
diferentes?

# Avaliação de Política?

Porque não utilizar avaliação de política visto na última aula? O método converge mais rapidamente pois faz uso das conexões.

1. Inicializa-se  $V_0^\pi(s) = 0$  para todos os estados  $s$
2. Faça  $k = 0$
3. Repita até convergir:
  - a. para todo  $s \in S$

$$V_{k+1}^\pi(s) \leftarrow \sum_{s'} T(s, \pi(s), s') [R(s, \pi(s), s') + \gamma V_k^\pi(s')]$$

Avaliação de Política requer conhecimento de  $T$  e  $R$ . Como atualizar os valores de  $V$  sem conhecer  $T$  e  $R$ ?

# Avaliação de Política Baseada em Amostragem

- Melhorar as estimativas de V através de amostras.
- Fazer várias amostras da ação  $a$  a partir de  $s$ .

$$\text{amostra}_1 = R(s, \pi(s), s'_1) + \gamma V_k^\pi(s'_1)$$

$$\text{amostra}_2 = R(s, \pi(s), s'_2) + \gamma V_k^\pi(s'_2)$$

...

$$V_{k+1}^\pi(s) \leftarrow \frac{1}{n} \sum_i \text{amostra}_i$$

$$\text{amostra}_n = R(s, \pi(s), s'_n) + \gamma V_k^\pi(s'_n)$$

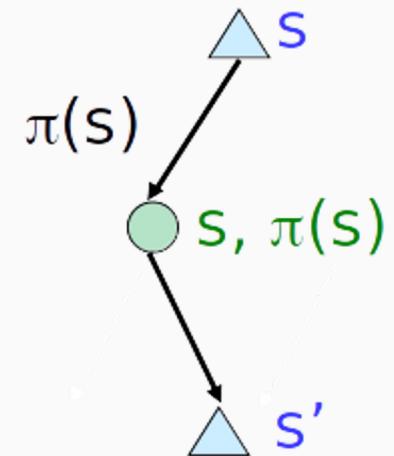
# Aprendizagem por Diferença Temporal (TD)

- Ideia: aprender através de cada tupla  $(s, a, s', r)$
- Estados  $s'$  com maior probabilidade  $T(s, a, s')$  irão aparecer com mais frequência.
- Diferença temporal dos valores aprendidos.
- Política ainda é fixa, estamos apenas avaliando.
- Modifica valor sempre que uma tupla é observada.

Amostra de  $V(s)$ :  $amostra = R(s, \pi(s), s') + \gamma V^\pi(s')$

Atualiza  $V(s)$ :  $V^\pi(s) \leftarrow (1 - \alpha)V^\pi(s) + \alpha \cdot amostra$

Ou o equivalente:  $V^\pi(s) \leftarrow V^\pi(s) + \alpha(amostra - V^\pi(s))$



# Médias Móveis Exponenciais

TD utiliza a seguinte atualização:

$$\bar{x}_n = (1 - \alpha) \cdot \bar{x}_{n-1} + \alpha \cdot x_n$$

Amostras mais recente são mais importantes

$$\bar{x}_n = \frac{x_n + (1 - \alpha) \cdot x_{n-1} + (1 - \alpha)^2 \cdot x_{n-2} + \dots}{1 + (1 - \alpha) + (1 - \alpha)^2 + \dots}$$

Médias móveis exponenciais faz você esquecer o passado (que deveria ser esquecido pois as amostras eram incorretas de toda forma...)

# Exemplo: Aprendizagem TD

Estados

	A	
B	C	D
	E	

$$\gamma = 1, \alpha = 1/2$$

Transições Observadas

B, dir, C, -2

C, dir, D, -2

	0	
0	0	8
.	0	

	0	
-1	0	8
.	0	

	0	
-1	3	8
.	0	

$$V^\pi(s) \leftarrow (1 - \alpha)V^\pi(s) + \alpha [R(s, \pi(s), s') + \gamma V^\pi(s')]$$

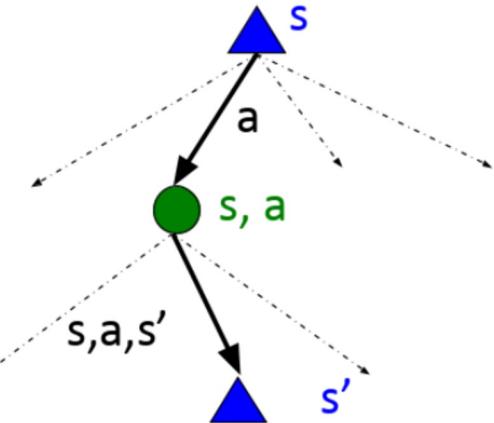
# Aprendizagem TD

- TD é um método livre de modelo que aprende os valores  $V$  para uma política  $\pi$  fixa.

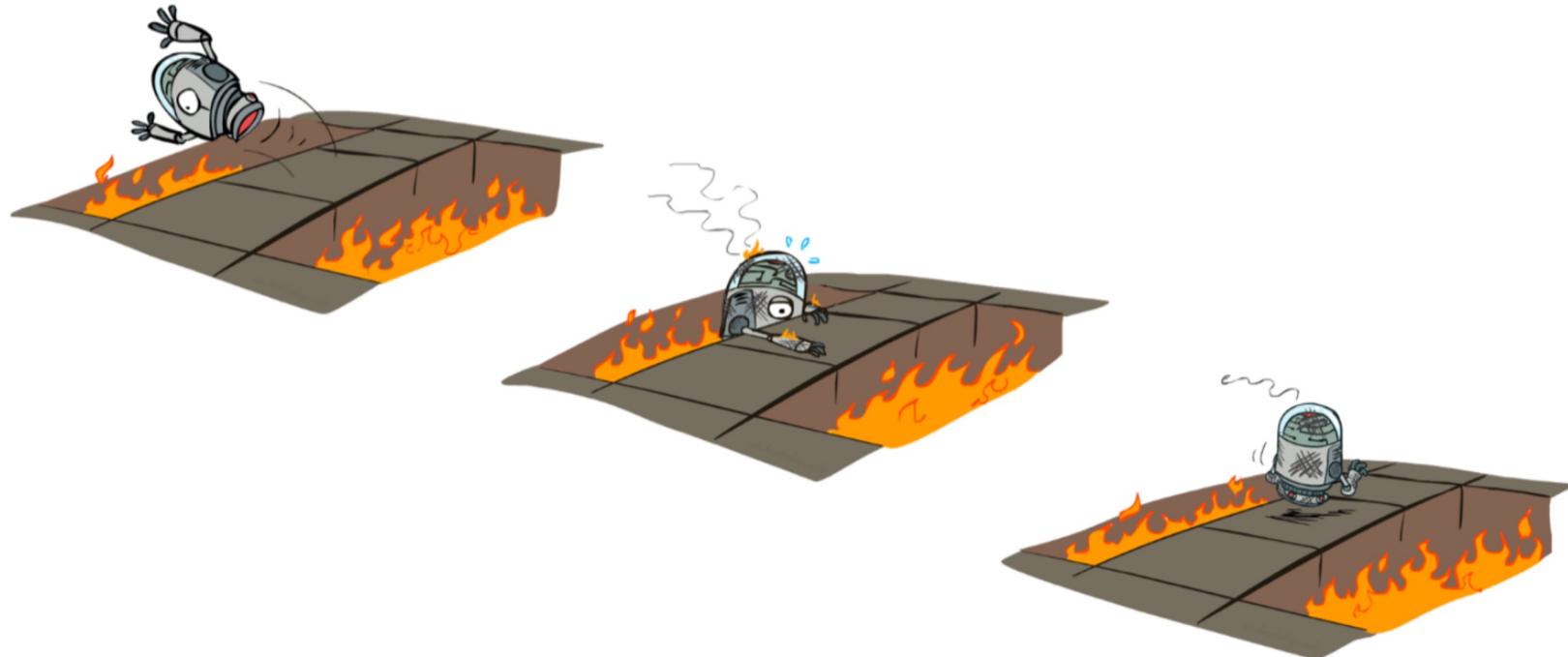
- Problema:** Como aprender uma nova política  $\pi$ ?

$$\pi_{i+1}(s) \leftarrow \operatorname{argmax}_a \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma V^{\pi_i}(s')]$$

- Solução:** Aprender os valores  $Q$  diretamente!



# Aprendizagem por Reforço Ativo



# Aprendizagem por Reforço Ativo

- $T(s, a, s')$  e  $R(s, a, s')$  ainda são desconhecidas.
- Nós escolhemos as ações.
- Objetivo: aprender os valores ótimos e as políticas ótimas.

# Iteração de Valor Q

## Iteração de Valor

Inicializa  $V_0(s) = 0$  e calcula os valores  $V_1(s)$ :

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma V_k(s')]$$

## Iteração de Q

Inicializa  $Q_0(s, a) = 0$  e calcula os valores  $Q_1(s, a)$ :

$$Q_{k+1}(s, a) \leftarrow \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma \max_{a'} Q(s', a')]$$

# Aprendizagem-Q

**Aprendizagem-Q:** Iteração de Valor Q baseado em amostragens.

$$Q_{k+1}(s, a) \leftarrow \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma \max_{a'} Q(s', a')]$$

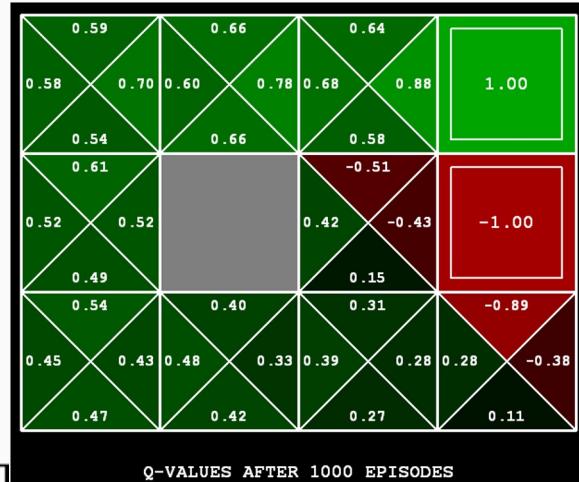
Aprende à medida que vai agindo:

1. Recebe amostra ( $s, a, s', r$ )
2. Considera o valor antigo de  $Q(s, a)$
3. Considera a estimativa da nova amostra:

$$\text{amostra} = R(s, a, s') + \gamma \max_{a'} Q(s', a')$$

4. Atualiza o valor de  $Q(s, a)$  de acordo com a média:

$$Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha[\text{amostra}]$$



# Aprendizagem-Q

**Aprendizagem-Q:** Iteração de Valor Q baseado em amostragens.

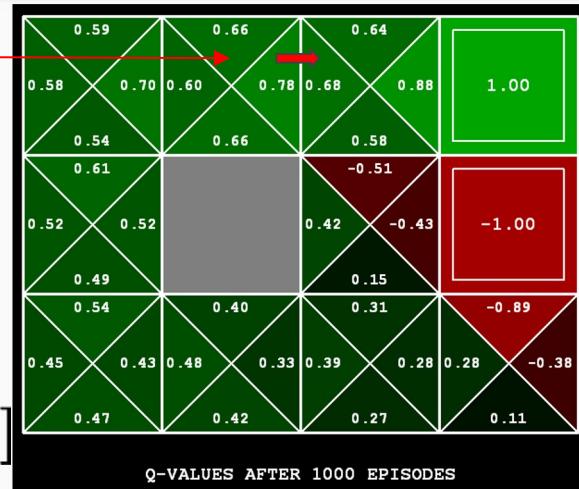
Aprende à medida que vai agindo:

1. Recebe amostra ( $s, a, s', r$ )
2. Considera o valor antigo de  $Q(s, a)$
3. Considera a estimativa da nova amostra:

$$\text{amostra} = R(s, a, s') + \gamma \max_{a'} Q(s', a')$$

4. Atualiza o valor de  $Q(s, a)$  de acordo com a média:

$$Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha[\text{amostra}]$$



$$\alpha=0.5 \quad R(s,a,s')=0 \quad \gamma=1$$

$$\text{amostra} = 0 + 1 * 0.88 \quad Q(s,a)=(1-0.5)(0.78)+0.5*0.88=0.83$$

# Exemplo: Aprendizagem-Q

- Converge para a política ótima independente do que o agente faz (Off-policy learning).
- Precisa explorar bastante.
- Como minimizar o arrependimento?



# Bibliografia

Reinforcement Learning – An Introduction – 23 nov 2018 por Richard S. Sutton , Andrew G. Barto , Francis Bach

<http://www.andrew.cmu.edu/course/10-703/textbook/BartoSutton.pdf>

FIM

# Exemplo: Aprendizagem-Q

- Converge para a política ótima independente do que o agente faz (Off-policy learning).
- Precisa explorar bastante.
- Como minimizar o arrependimento?



# Bibliografia

Reinforcement Learning – An Introduction – 23 nov 2018 por Richard S. Sutton , Andrew G. Barto , Francis Bach

<http://www.andrew.cmu.edu/course/10-703/textbook/BartoSutton.pdf>

# Links

[https://colab.research.google.com/github/simoninithomas/Deep\\_reinforcement\\_learning\\_Course/blob/master/Q\\_Learning\\_with\\_FrozenLakev2.ipynb](https://colab.research.google.com/github/simoninithomas/Deep_reinforcement_learning_Course/blob/master/Q_Learning_with_FrozenLakev2.ipynb)

FIM