

Problema Grid World com Q-Learning

Alunos:

- Gustavo Larsen
- João David
- Lucas Michels
- Luis Felipe Mondini
- Thiago Saraiva



Q-Learning: Conceito Completo

O **Q-Learning** é uma técnica de **Aprendizado por Reforço** que permite que um agente aprenda a melhor política, ou seja, a melhor ação a tomar em cada estado, interagindo com o ambiente sem conhecê-lo previamente.

A ideia central é maximizar a recompensa total acumulada ao longo do tempo, aprendendo por tentativa e erro: o agente toma uma ação, observa o resultado e atualiza seu conhecimento na tabela Q.

Essa tabela $Q(s, a)$ estima a qualidade de executar uma ação a em um estado s , e é atualizada iterativamente com base nas novas recompensas e estimativas futuras.

Fórmula do Q-Learning e Diferença Temporal

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \right]$$

Ela calcula a diferença temporal (TD error) entre o valor esperado e o valor real, atualizando a estimativa Q proporcionalmente.

Componentes da fórmula

$Q(s_t, a_t)$: valor atual estimado da ação a_t no estado s_t .

α : taxa de aprendizado, controla a rapidez da atualização.

r_{t+1} : recompensa recebida após a ação.

γ : fator de desconto, valoriza recompensas futuras.

$\max_{a'} Q(s_{t+1}, a')$: melhor valor esperado no próximo estado.



Após 100 episódios

↓ 1.03	↓ 1.63	→ -0.31	↓ 2.50
↓ 2.60	↓ 4.09	■	↓ 6.23
→ 4.31	↓ 6.09	↓ 7.78	↓ 9.62
→ 5.42	→ 7.97	→ 9.99	★

Após 1000 episódios

↓ 1.81	↓ 3.12	→ 0.36	↓ 3.45
↓ 3.12	↓ 4.58	■	↓ 6.79
→ 4.58	↓ 6.20	↓ 8.00	↓ 9.92
→ 6.20	→ 8.00	→ 10.00	★

Após 2000 episódios

↓ 1.81	↓ 3.12	→ 0.88	↓ 3.93
↓ 3.12	↓ 4.58	■	↓ 7.10
→ 4.58	↓ 6.20	↓ 8.00	↓ 9.94
→ 6.20	→ 8.00	→ 10.00	★

Entendendo a Tabela Q no Grid World

Política após 2000 episódios (melhor ação)

↓ 1.81	↓ 3.12	→ 0.88	↓ 3.93
↓ 3.12	↓ 4.58	■	↓ 7.10
→ 4.58	↓ 6.20	↓ 8.00	↓ 9.94
→ 6.20	→ 8.00	→ 10.00	★

A tabela Q representa os estados do ambiente Grid World, um grid 4x4 onde o agente se move.

Células marcadas com ■ são obstáculos, onde o agente não pode andar, e a célula ★ é o estado objetivo que o agente deseja alcançar.

As demais células representam os estados possíveis para o agente explorar e aprender a melhor ação a tomar.

Obstáculos

Locais bloqueados no grid, impossíveis de atravessar.

Objetivo

Estado final que o agente busca alcançar.

Estados Livres

Posições onde o agente pode se mover e aprender ações.

Setas na Tabela: Ações Aprendidas

Em cada célula da tabela, uma seta indica a melhor ação que o agente aprendeu para maximizar a recompensa futura.

As ações possíveis são movimentos para cima, baixo, esquerda e direita, representadas por setas direcionais.

Por exemplo, uma seta “→ 3.25” indica que a melhor ação naquele estado é mover para a direita, com valor Q associado de 3.25.

Ações

- ↑ subir
- ↓ descer
- ← esquerda
- → direita

Significado das Setas

Indicadores visuais da política aprendida pelo agente para cada estado.

Valores Q: Qualidade das Ações

Os números ao lado das setas representam o valor Q da ação naquele estado, indicando a qualidade e o retorno futuro esperado.

Valores mais altos indicam ações mais promissoras para maximizar a recompensa acumulada.

Assim, o agente prioriza ações com valores Q maiores para alcançar o objetivo com eficiência.

Valor Q

Estimativa da recompensa futura ao escolher uma ação.

Importância

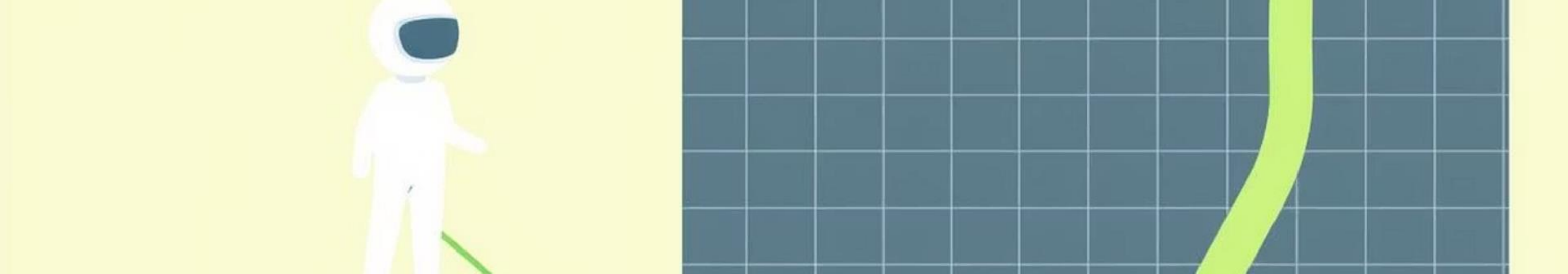
Guiam a escolha da melhor ação em cada estado.

Maximização

Agente busca maximizar esses valores para otimizar sua política.

Política após 2000 episódios (melhor ação)

↓ 1.81	↓ 3.12	→ 0.88	↓ 3.93
↓ 3.12	↓ 4.58	■	↓ 7.10
→ 4.58	↓ 6.20	↓ 8.00	↓ 9.94
→ 6.20	→ 8.00	→ 10.00	★






Como a Tabela Guia a Política do Agente

O agente consulta a tabela Q para decidir sua próxima ação, sempre escolhendo a ação com maior valor Q no estado atual.

Essa tabela representa o que o agente aprendeu após o treinamento, indicando o caminho ideal no grid considerando obstáculos e penalidades.

Um episódio é uma sequência completa de interações, começando no estado inicial e terminando ao alcançar o objetivo ou condição de parada.

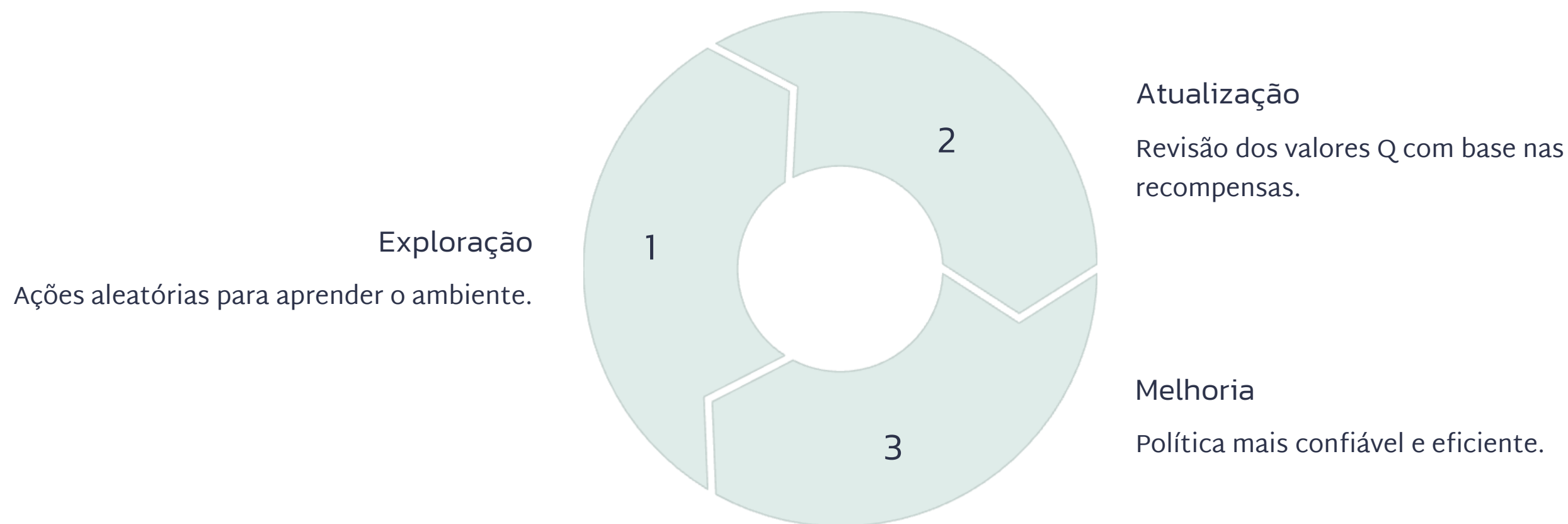
- | | | | | | |
|---|---|---|---|---|---|
|  | Episódios
Sequências de tentativas do agente para aprender o caminho ideal. |  | Estado Inicial
Posição (0,0) no Grid World. |  | Estado Final
Objetivo na posição (3,3) ou condição de parada. |
|---|---|---|---|---|---|

Evolução do Aprendizado ao Longo dos Episódios

No início, o agente explora ações aleatórias para descobrir quais são boas.

Com a experiência, ele atualiza a tabela Q usando a regra do Q-Learning, ajustando a confiança em cada ação com base nas recompensas e valores futuros.

Ao longo dos episódios, a política melhora: o agente escolhe ações melhores com mais frequência, e os valores Q refletem a qualidade real das ações.





Aprendizado por Tentativa e Erro

Cada episódio é uma experiência que ajuda o agente a aprender o caminho ideal para maximizar a recompensa.

O agente aprende por tentativa e erro, atualizando seu conhecimento sobre o valor das ações com base nas experiências acumuladas.

Quanto mais episódios, melhor a política e maior a confiança nos valores Q , indicados na tabela pelas setas e números.

Experiência

Aprendizado contínuo com cada tentativa.

Atualização

Melhora progressiva da política.

Confiança

Valores Q mais precisos e confiáveis.



Progresso na Tabela Q ao Longo do Treinamento

Após poucos episódios, as ações do agente são quase aleatórias, com valores Q baixos e política confusa.

Com muitos episódios, as setas indicam caminhos coerentes e os números maiores mostram maior confiança nas ações aprendidas.

Esse progresso visualiza o aprendizado e a consolidação da política ótima pelo agente.

Início

Ações aleatórias e baixa confiança.

Meio

Política em desenvolvimento, valores Q ajustados.

Fim

Política otimizada e valores Q confiáveis.

Problema Grid World com Q-Learning

Alunos:

- Gustavo Larsen
- João David
- Lucas Michels
- Luis Felipe Mondini
- Thiago Saraiva

