Instituto Tecnológico de Aeronáutica - ITA
Inteligência Artificial para Robótica Móvel - CT213
Aluno: Matheus Felipe Ramos Borges

Relatório do Laboratório 10 - Programação Dinâmica

1 Breve Explicação em Alto Nível da Implementação

O objetivo deste laboratório é implementar algoritmos clássicos de *Reinforcement Learning* baseados em programação dinâmica para resolver problemas de decisão sequencial em um ambiente do tipo *GridWorld*. Especificamente, o agente deve aprender a política ótima para alcançar o objetivo maximizando as recompensas acumuladas.

1.1 Avaliação de Política

A **avaliação de política** tem como objetivo estimar a *função de valor* $v_\pi(s)$ associada a uma política fixa π. Essa função representa a expectativa do retorno futuro (soma de recompensas descontadas) ao iniciar em um estado s e seguir a política π.

Para isso, utilizamos a **Equação de Bellman** para políticas fixas:

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s'} P(s'|s,a) \left[R(s,a) + \gamma \cdot v_\pi(s') \right]$$

Essa equação é aplicada de forma iterativa até que a função valor convirja, ou seja, até que as mudanças entre iterações sucessivas sejam menores que um limiar ϵ.

1.2 Iteração de Valor

A **iteração de valor** busca encontrar diretamente a função de valor ótima $v_*(s)$, sem necessidade de manter uma política explícita durante o processo. A atualização é feita utilizando a versão *ótima* da Equação de Bellman:

$$v_{k+1}(s) = \max_{a \in \mathcal{A}} \sum_{s'} P(s'|s,a) \left[R(s,a) + \gamma \cdot v_k(s') \right]$$

Após a convergência da função de valor, a política ótima π^* pode ser extraída selecionando, em cada estado, a ação que maximiza a expectativa de retorno.

1.3 Iteração de Política

A **iteração de política** alterna entre duas etapas principais: *avaliação de política* e *melhoria de política*. Inicialmente, assume-se uma política arbitrária. A função de valor correspondente é calculada e, em seguida, a política é melhorada tornando-a *gananciosa* (greedy) com relação à função de valor:

$$\pi_{k+1}(s) = \arg\max_{a \in \mathcal{A}} \sum_{s'} P(s'|s,a) \left[R(s,a) + \gamma \cdot v_{\pi_k}(s') \right]$$

Esse processo é repetido até que a política deixe de mudar, ou seja, até atingir a **política ótima** π^*.

2 Tabelas Comprovando Funcionamento do Código

2.1 Caso $p_c = 1,0$ e $\gamma = 1,0$

2.1.1 Avaliação de Política

```
Value function:
[  -384.09,  -382.73,  -381.19,    *    ,  -339.93,  -339.93]
[  -380.45,  -377.91,  -374.65,    *    ,  -334.92,  -334.93]
[  -374.34,  -368.82,  -359.85,  -344.88,  -324.92,  -324.93]
[  -368.76,  -358.18,  -346.03,    *    ,  -289.95,  -309.94]
[    *    ,  -344.12,  -315.05,  -250.02,  -229.99,    *    ]
[  -359.12,  -354.12,    *    ,  -200.01,  -145.00,     0.00]
Policy:
[  SURDL  ,  SURDL  ,  SURDL  ,    *    ,  SURDL  ,  SURDL  ]
[  SURDL  ,  SURDL  ,  SURDL  ,    *    ,  SURDL  ,  SURDL  ]
[  SURDL  ,  SURDL  ,  SURDL  ,  SURDL  ,  SURDL  ,  SURDL  ]
[  SURDL  ,  SURDL  ,  SURDL  ,    *    ,  SURDL  ,  SURDL  ]
[    *    ,  SURDL  ,  SURDL  ,  SURDL  ,  SURDL  ,    *    ]
[  SURDL  ,  SURDL  ,    *    ,  SURDL  ,  SURDL  ,    S    ]
----------------------------------------------------------------
```

Figura 1: Avaliação de Política para $p_c = 1,0$ e $\gamma = 1,0$

2.1.2 Iteração de Valor

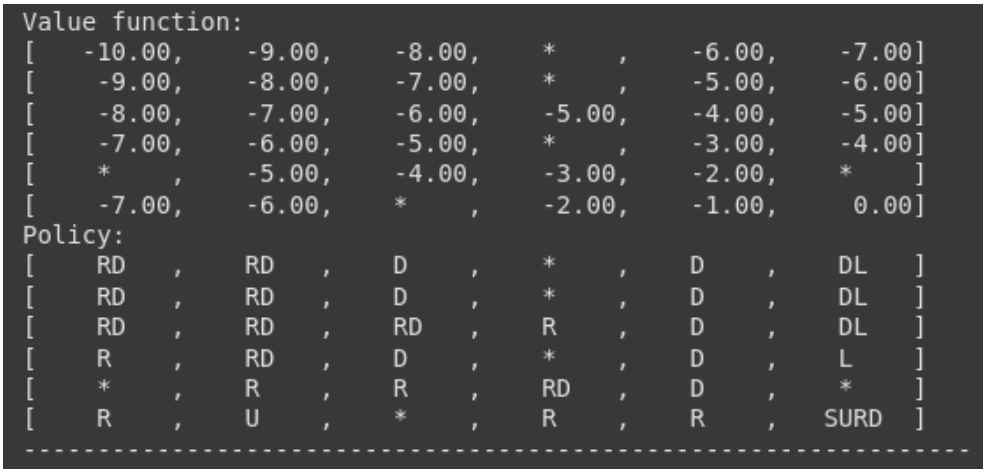

Figura 2: Interação de Valor para $p_c = 1,0$ e $\gamma = 1,0$

2.1.3 Iteração de Política

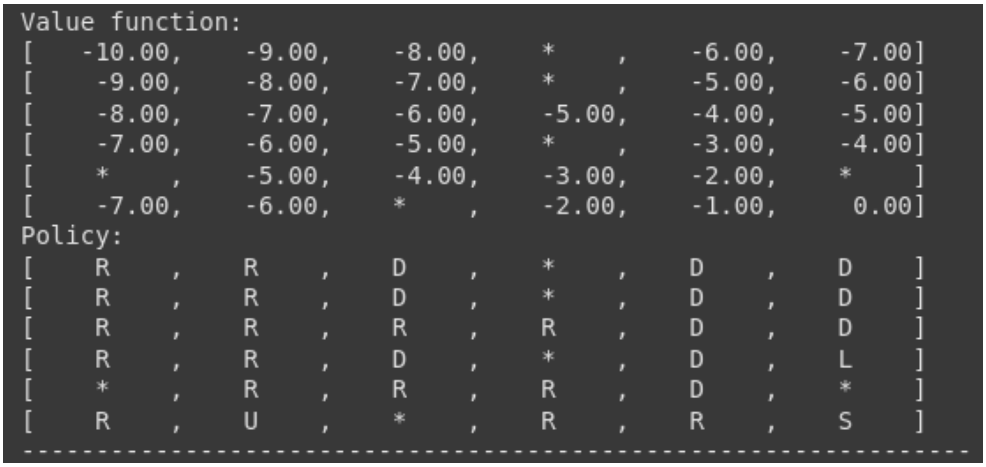

Figura 3: Iteração de política para $p_c = 1,0$ e $\gamma = 1,0$

2.2 Caso $p_c = 0,8$ e $\gamma = 0,98$

2.2.1 Avaliação de Política

```
Value function:
[   -47.19,   -47.11,   -47.01,    *    ,   -45.13,   -45.15]
[   -46.97,   -46.81,   -46.60,    *    ,   -44.58,   -44.65]
[   -46.58,   -46.21,   -45.62,   -44.79,   -43.40,   -43.63]
[   -46.20,   -45.41,   -44.42,    *    ,   -39.87,   -42.17]
[    *    ,   -44.31,   -41.64,   -35.28,   -32.96,    *    ]
[   -45.73,   -45.28,    *    ,   -29.68,   -21.88,     0.00]
Policy:
[  SURDL  ,  SURDL  ,  SURDL  ,    *    ,  SURDL  ,  SURDL  ]
[  SURDL  ,  SURDL  ,  SURDL  ,    *    ,  SURDL  ,  SURDL  ]
[  SURDL  ,  SURDL  ,  SURDL  ,  SURDL  ,  SURDL  ,  SURDL  ]
[  SURDL  ,  SURDL  ,  SURDL  ,    *    ,  SURDL  ,  SURDL  ]
[    *    ,  SURDL  ,  SURDL  ,  SURDL  ,  SURDL  ,    *    ]
[  SURDL  ,  SURDL  ,    *    ,  SURDL  ,  SURDL  ,    S    ]
----------------------------------------------------------------
```

Figura 4: Avaliação de Política para $p_c = 0,8$ e $\gamma = 0,98$

2.2.2 Iteração de Valor

```
Value function:
[   -11.65,   -10.78,    -9.86,    *    ,    -7.79,    -8.53]
[   -10.72,    -9.78,    -8.78,    *    ,    -6.67,    -7.52]
[    -9.72,    -8.70,    -7.59,    -6.61,    -5.44,    -6.42]
[    -8.70,    -7.58,    -6.43,    *    ,    -4.09,    -5.30]
[    *    ,    -6.43,    -5.17,    -3.87,    -2.76,    *    ]
[    -8.63,    -7.58,    *    ,    -2.69,    -1.40,     0.00]
Policy:
[    D    ,    D    ,    D    ,    *    ,    D    ,    D    ]
[    D    ,    D    ,    D    ,    *    ,    D    ,    D    ]
[    RD   ,    D    ,    D    ,    R    ,    D    ,    D    ]
[    R    ,    RD   ,    D    ,    *    ,    D    ,    L    ]
[    *    ,    R    ,    R    ,    D    ,    D    ,    *    ]
[    R    ,    U    ,    *    ,    R    ,    R    ,    S    ]
----------------------------------------------------------------
```

Figura 5: Interação de Valor para $p_c = 0,8$ e $\gamma = 0,98$

2.2.3 Iteração de Política

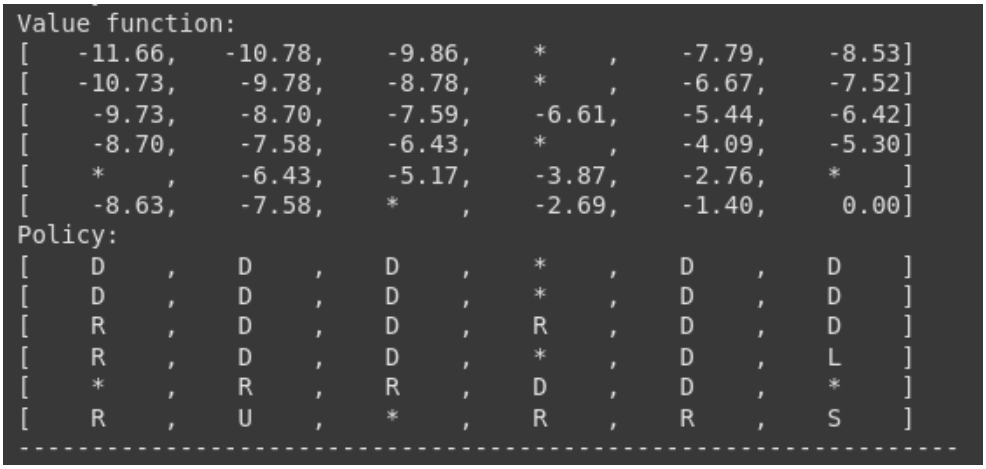

Figura 6: Avaliação de Política para $p_c = 0,8$ e $\gamma = 0,98$

3 Discussão dos Resultados

Os resultados obtidos demonstram que os algoritmos de programação dinâmica implementados apresentaram comportamento consistente com o esperado. No caso determinístico ($p_c = 1.0$, $\gamma = 1.0$), a convergência foi rápida e a política ótima resultante seguiu trajetórias diretas até o objetivo. Já no caso estocástico ($p_c = 0.8$, $\gamma = 0.98$), os algoritmos exigiram mais iterações para convergir, refletindo a incerteza nas ações e resultando em políticas mais conservadoras. As diferenças entre os dois cenários destacam como a estocasticidade e o fator de desconto afetam a propagação dos valores e a natureza das decisões tomadas pelo agente.