

Programação Dinâmica

Samara Ribeiro Silva

Instituto Tecnológico de Aeronáutica, Laboratório de Inteligência Artificial para Robótica Móvel (CT-213). Professor Marcus Ricardo Omena de Albuquerque Máximo, São José dos Campos, São Paulo, 28 de junho de 2021.

A implementação foi realizada conforme as seguintes fórmulas:

- Para a Avaliação de Política

$$v_{k+1}(s) = \sum_{a \in A} \pi(a|s)r(s, a) + \sum_{a \in A} \pi(a|s)r(s, a)p(s'|s, a)v_k(s')$$

- Para a Iteração de Valor

$$v_{k+1}(s) = \max \left(r(s, a) + \sum_{s' \in S} p(s'|s, a)v_k(s') \right)$$

Já para a implementação da Iteração de Política foi realizada adaptando o código da avaliação de política com a inserção do `greedy_policy` a cada 3 iterações.

Nas figuras 1 a 6 é possível observar os resultados obtidos com os testes. Nos testes mostrados nas 3 primeiras figuras utilizou-se a probabilidade de executar corretamente a ação escolhida $p = 1$ e $\gamma = 1$. Nas demais foram utilizados $p = 0.8$ e $\gamma = 0.9$.

Analisando os resultados é possível observar que o valor do estado final em todos os testes apresenta o valor 0. Os resultados dos testes da iteração de valor e da iteração de política são iguais devido ambos poderem convergir para a política e valores ótimos.

Com a alteração dos valores de p e γ , observou que houve um aumento em módulo dos valores, isso ocorreu devido ao incremento devido à incerteza inserida pela diminuição da probabilidade de executar corretamente a ação escolhida. Já a diminuição do γ torna a escolha por uma recompensa imediata mais atraente.

Figura 1: Resultados obtidos com o teste da *policy evaluation*, com $p = 1$ e $\gamma = 1$

```

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 2: Resultados obtidos com o teste da *value iteration*, com $p = 1$ e $\gamma = 1$

```

Value iteration:
Value function:
[ -10.00, -9.00, -8.00, * , -6.00, -7.00]
[ -9.00, -8.00, -7.00, * , -5.00, -6.00]
[ -8.00, -7.00, -6.00, -5.00, -4.00, -5.00]
[ -7.00, -6.00, -5.00, * , -3.00, -4.00]
[ * , -5.00, -4.00, -3.00, -2.00, * ]
[ -7.00, -6.00, * , -2.00, -1.00, 0.00]
Policy:
[ RD , RD , D , * , D , DL ]
[ RD , RD , D , * , D , DL ]
[ RD , RD , RD , R , D , DL ]
[ R , RD , D , * , D , L ]
[ * , R , R , RD , D , * ]
[ R , U , * , R , R , SURD ]
-----

```

Figura 3: Resultados obtidos com o teste da *policy iteration*, com $p = 1$ e $\gamma = 1$

```

Policy iteration:
Value function:
[ -10.00,  -9.00,  -8.00,  *,  -6.00,  -7.00]
[  -9.00,  -8.00,  -7.00,  *,  -5.00,  -6.00]
[  -8.00,  -7.00,  -6.00,  -5.00,  -4.00,  -5.00]
[  -7.00,  -6.00,  -5.00,  *,  -3.00,  -4.00]
[   *,  -5.00,  -4.00,  -3.00,  -2.00,  * ]
[  -7.00,  -6.00,  *,  -2.00,  -1.00,  0.00]
Policy:
[  RD ,  RD ,  D ,  * ,  D ,  DL ]
[  RD ,  RD ,  D ,  * ,  D ,  DL ]
[  RD ,  RD ,  RD ,  R ,  D ,  DL ]
[  R ,  RD ,  D ,  * ,  D ,  L ]
[  * ,  R ,  R ,  RD ,  D ,  * ]
[  R ,  U ,  * ,  R ,  R ,  SURD ]
-----

```

Figura 4: Resultados obtidos com o teste da *policy evaluation*, com $p = 0.8$ e $\gamma = 0.98$

```

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 5: Resultados obtidos com o teste da *value iteration*, $p = 0.8$ e $\gamma = 0.98$

```

Value iteration:
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 6: Resultados obtidos com o teste da *policy iteration*, $p = 0.8$ e $\gamma = 0.98$

```

Policy iteration:
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 ]
[ R , D , D , R , D , D ]
[ R , D , D , * , D , L ]
[ * , R , R , D , D , * ]
[ R , U , * , R , R , S ]
-----

```