Instituto Tecnológico de Aeronáutica - ITA
Inteligência Artificial para Robótica Móvel - CT213
Aluno:

Relatório do Laboratório 10 - Programação Dinâmica

1 Breve Explicação em Alto Nível da Implementação

1.1 Avaliação de Política

Avaliação de política para uma política dada. A impelemtnação seguiu o descrito nas aulas e semelhante ao código do próximo método (com a diferença do acréscimo de uma soma nas ações dependente da política).

1.2 Iteração de Valor

Iteração de valor sem função política definida. Busca-se o valor máximo por ação e não o somatório em todas as ações (como na avaliação de política).

```
def value_iteration(grid_world, initial_value,
                num_iterations=10000, epsilon=1.0e-5):
        dimensions = grid_world.dimensions
        gamma = grid_world.gamma
        value = np.copy(initial_value)
        for k in range(0,num_iterations):
                next_value = np.zeros(dimensions)
                for i in range(dimensions[0]):
                        for j in range(dimensions[1]):
                                current_state = (i, j)
                                rsp = np.zeros(NUM_ACTIONS)
                                for action in range(NUM_ACTIONS):
                                        r = grid_world.reward(current_state, action)
                                        rsp[action] = r
                                        for next_state in grid_world.get_valid_sucessors(
                                                (i, j), action):
                                                transition_prob =
                                                        grid_world.transition_probability(
                                                        current_state, action, next_state)
                                                rsp[action] = rsp[action] +
                                                        gamma*transition_prob*
                                                        value[next_state[0], next_state[1]]
                                next_value[i,j] = np.amax(rsp)
                delta_value = next_value - value
                delta_value_max = np.amax(np.absolute(delta_value))
                value = next_value
                if delta_value_max < epsilon:
                        break
        return value
```

1.3 Iteração de Política

Uso das funções já implementadas:

```
def policy_iteration(grid_world, initial_value, initial_policy,
        evaluations_per_policy=3, num_iterations=10000,
        epsilon=1.0e-5):

        value = np.copy(initial_value)
        policy = np.copy(initial_policy)
        for k in range(0, num_iterations):
                new_value = policy_evaluation(grid_world=grid_world,
                        initial_value=value, policy=policy,
                        num_iterations=evaluations_per_policy)
                new_policy = greedy_policy(grid_world, new_value)

                delta_value = new_value - value
                delta_value_max = np.amax(np.absolute(delta_value))

                value = new_value
                policy = new_policy
                if delta_value_max < epsilon:
                        break
        return value, policy
```

2 Tabelas Comprovando Funcionamento do Código

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

2.1.1 Avaliação de Política

```
"""
[  -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    ]
"""
```

2.1.2 Iteração de Valor

```
"""
[   -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  ]
"""
```

2.1.3 Iteração de Política

```
"""
[   -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  ]
"""
```

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

2.2.1 Avaliação de Política

```
"""
[   -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    ]
"""
```

2.2.2 Iteração de Valor

```
"""
[   -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    ]
"""
```

2.2.3 Iteração de Política

```
"""
[   -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    ]
"""
```

3 Discussão dos Resultados

Dentre os resultados, é interessante observar que tanto a iteração de valores como a iteração de política alcançam o valor mínimo. Além disso, observa-se como é possível evoluir a política, quando comparado ao resultado da avaliação de política para uma aleatória, tem-se um resultado final muito superior.