EE488 Special Topics in EE
<Deep Learning and AlphaGo>

Sae-Young Chung

Project #2

December 4, 2017

KAIST

Task 1

Q(s,a)

```
[[ 0.   0. ]
 [ 0.2  0. ]
 [ 0.   0. ]
 [ 0.   0. ]
 [ 0.   0. ]
 [ 0.   0. ]
 [ 0.   0. ]
 [ 0.   0. ]
 [ 0.   0. ]
 [ 0.   0. ]
 [ 0.   0. ]
 [ 0.   0. ]
 [ 0.   0. ]
 [ 0.   0. ]
 [ 0.   0. ]
 [ 0.   0. ]
 [ 0.   0. ]
 [ 0.   0. ]
 [ 0.   0. ]
 [ 0.   0. ]
 [ 0.   0. ]]
```

n_epsides = 1
max_steps = 1000
alpha = 0.2
gamma = 0.9
epsilon.init = 1.
epsilon.final = 1.

test_n_episodes = 1
test_max_steps = 1000
test_epsilon = 0

Average number of runs: 198.0
Number of steps for testing: 130.0

Task 1

Q(s,a)

```
[[ 0.          0.        ]
 [ 0.5904      0.        ]
 [ 0.16272     0.00766092]
 [ 0.03427453  0.00070577]
 [ 0.00705775  0.        ]
 [ 0.          0.        ]
 [ 0.          0.        ]
 [ 0.          0.        ]
 [ 0.          0.        ]
 [ 0.          0.        ]
 [ 0.          0.        ]
 [ 0.          0.        ]
 [ 0.          0.        ]
 [ 0.          0.        ]
 [ 0.          0.        ]
 [ 0.          0.        ]
 [ 0.          0.        ]
 [ 0.          0.        ]
 [ 0.          0.        ]
 [ 0.          0.        ]
 [ 0.          0.2       ]
 [ 0.          0.        ]]
```

n_epsides = 5
max_steps = 1000
alpha = 0.2
gamma = 0.9
epsilon.init = 1.
epsilon.final = 1.

test_n_episodes = 1
test_max_steps = 1000
test_epsilon = 0

0: 42
1: 160
2: 38
3: 24
4: 138
Average number of runs: 80.4
Number of steps in testing 18

KAIST

Task 1

Q(s,a)

```
[[ 0.00000000e+00  0.00000000e+00]
 [ 2.00000000e-01  0.00000000e+00]
 [ 0.00000000e+00  0.00000000e+00]
 [ 0.00000000e+00  0.00000000e+00]
 [ 0.00000000e+00  0.00000000e+00]
 [ 0.00000000e+00  0.00000000e+00]
 [ 0.00000000e+00  0.00000000e+00]
 [ 0.00000000e+00  0.00000000e+00]
 [ 0.00000000e+00  0.00000000e+00]
 [ 0.00000000e+00  4.21286096e-05]
 [ 4.21286096e-06  5.48647792e-04]
 [ 1.97384075e-04  1.38686441e-03]
 [ 5.14062347e-04  4.85223854e-03]
 [ 2.48295317e-03  8.39126457e-03]
 [ 3.66914300e-03  1.87376210e-02]
 [ 8.45734113e-03  3.82843102e-02]
 [ 2.13014397e-02  6.05637447e-02]
 [ 2.77273672e-02  1.47868405e-01]
 [ 6.45323152e-02  3.14629171e-01]
 [ 1.17437507e-01  5.90400000e-01]
 [ 0.00000000e+00  0.00000000e+00]]
```

n_epsides = 5
max_steps = 1000
alpha = 0.2
gamma = 0.9
epsilon.init = 1.
epsilon.final = 1.

test_n_episodes = 1
test_max_steps = 1000
test_epsilon = 0

0: 30
1: 72
2: 86
3: 36
4: 262
Average number of runs: 97.2
Number of steps in testing 10

optimal

- How can an optimal policy be obtained with n_episodes = 5 only?
- Example)
  - First episode: Q(19,right) is updated and becomes positive
  - Second episode
    - 10→11→ …→ 17→ 18→ 19→ 18→ 17→ 18→ 17→ 16→ 17→ …

  - Therefore, n_episodes = 2 is enough, but the chance of getting an optimal policy will be lower.

Task 1

Q(s,a)

```
[[ 0.         0.        ]
 [ 1.         0.81      ]
 [ 0.9        0.729     ]
 [ 0.81       0.6561    ]
 [ 0.729      0.59049   ]
 [ 0.6561     0.531441  ]
 [ 0.59049    0.4782969 ]
 [ 0.531441   0.43046721]
 [ 0.4782969  0.38742049]
 [ 0.43046721 0.34867844]
 [ 0.38742049 0.38742049]
 [ 0.34867844 0.43046721]
 [ 0.38742049 0.4782969 ]
 [ 0.43046721 0.531441  ]
 [ 0.4782969  0.59049   ]
 [ 0.531441   0.6561    ]
 [ 0.59049    0.729     ]
 [ 0.6561     0.81      ]
 [ 0.729      0.9       ]
 [ 0.81       1.        ]
 [ 0.         0.        ]]
```

n_epsides = 1000
max_steps = 1000
alpha = 0.2
gamma = 0.9
epsilon.init = 1.
epsilon.final = 1.

test_n_episodes = 1
test_max_steps = 1000
test_epsilon = 0

Average number of runs: 100.4
Number of steps in testing 10

Analysis

- Optimal policy π_*: go left if $s < 10$, go right if $s > 10$ (any action is OK if $s = 10$, e.g., go left, go right, or go left or right with probability 0.5 each)
- Optimal action value function

$$Q_*(s,a) = \begin{cases} \gamma^{s-1} & \text{if } s < 10 \text{ and } a = \text{left} \\ \gamma^{s+1} & \text{if } s < 10 \text{ and } a = \text{right} \\ \gamma^{19-s} & \text{if } s > 10 \text{ and } a = \text{right} \\ \gamma^{21-s} & \text{if } s > 10 \text{ and } a = \text{left} \\ \gamma^{9} & \text{if } s = 10 \end{cases}$$

- You can see γ^2 is multiplied if you take a sub-optimal action (e.g., going from $s = 7$ to $s = 8$) since coming back requires another step and therefore such a sub-optimal action costs two additional steps.

Recap – Random Walk

- Random walk example: $S_n = \sum_{i=1}^{n} X_i$, $X_i \sim$ i.i.d. ± 1 w.p. $\frac{1}{2}$ each

$$\mathbb{E}[S_n] = \sum_{i=1}^{n} \mathbb{E}[X_i] = 0$$

$$\mathbb{E}[S_n^2] = \mathbb{E}\left[\sum_{i=1}^{n} X_i^2\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}[X_i X_j]$$

$$= \sum_{i=1}^{n} \mathbb{E}[X_i^2] = n$$

$$\lim_{n\to\infty} \frac{\mathbb{E}[|S_n|]}{\sqrt{n}} = \sqrt{\frac{2}{\pi}}$$

Analysis

- Since $\lim_{n\to\infty} \mathbb{E}[|S_n|]/\sqrt{n} = \sqrt{\frac{2}{\pi}}$, we can expect you need an order of $10^2 = 100$ steps to reach $s = 0$ or $s = 20$.
- Indeed, simulation shows 100.2 steps were needed on average.
- To be precise, we need to analyze the average number of time steps an episode takes.

Task 1

Q(s,a)

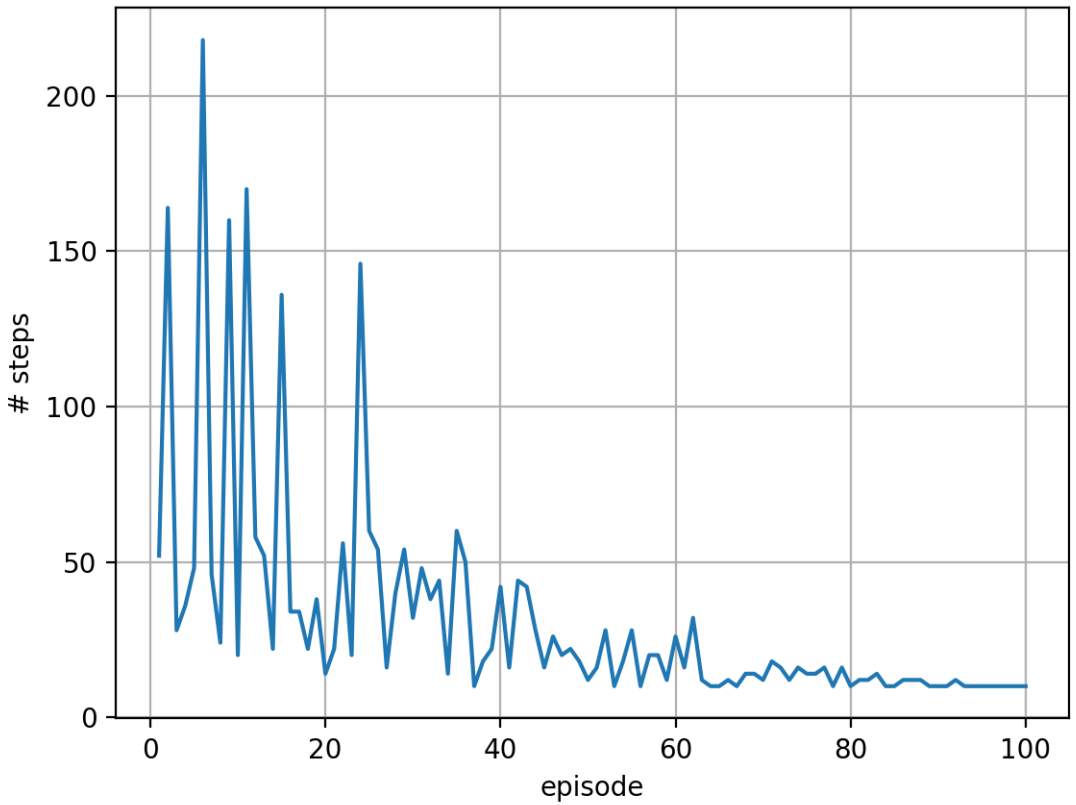
```
[[ 0.         0.        ]
 [ 1.         0.80851074]
 [ 0.89999999  0.72814571]
 [ 0.80999996  0.65308072]
 [ 0.7289998   0.58934071]
 [ 0.65609922  0.53089091]
 [ 0.59048789  0.47760669]
 [ 0.53143541  0.4281108 ]
 [ 0.47828431  0.38209722]
 [ 0.43043871  0.33787478]
 [ 0.38735979  0.27405712]
 [ 0.32946615  0.17605408]
 [ 0.23698705  0.10268889]
 [ 0.15853989  0.01539565]
 [ 0.03343938  0.00409224]
 [ 0.00246016  0.0199669 ]
 [ 0.00613811  0.04590899]
 [ 0.00577569  0.15608808]
 [ 0.00527213  0.44324029]
 [ 0.07355681  0.7902848 ]
 [ 0.         0.        ]]
```

n_epsides = 100
max_steps = 1000
alpha = 0.2
gamma = 0.9
epsilon.init = 1.
epsilon.final = 0.

test_n_episodes = 1
test_max_steps = 1000
test_epsilon = 0

Average number of runs: 26.6
Number of steps in testing 10

Task 1



Task 3 Training Example

episode: 99, train score: 2.000000, test score: 0.000000, test time steps: 3
episode: 199, train score: 1.000000, test score: 4.000000, test time steps: 33
episode: 299, train score: 1.000000, test score: 2.000000, test time steps: 15
episode: 399, train score: 2.000000, test score: 11.000000, test time steps: 105
episode: 499, train score: 2.000000, test score: 8.000000, test time steps: 200
episode: 599, train score: 15.000000, test score: 15.000000, test time steps: 145
episode: 699, train score: 14.000000, test score: 14.000000, test time steps: 200
episode: 799, train score: 2.000000, test score: 14.000000, test time steps: 200
episode: 899, train score: 15.000000, test score: 15.000000, test time steps: 189
episode: 999, train score: 5.000000, test score: 14.000000, test time steps: 200
episode: 1099, train score: 14.000000, test score: 15.000000, test time steps: 135
episode: 1199, train score: 1.000000, test score: 15.000000, test time steps: 157
episode: 1299, train score: 1.000000, test score: 15.000000, test time steps: 123
episode: 1399, train score: 2.000000, test score: 15.000000, test time steps: 115
episode: 1499, train score: 8.000000, test score: 14.000000, test time steps: 200
episode: 1599, train score: 6.000000, test score: 15.000000, test time steps: 117
episode: 1699, train score: 4.000000, test score: 11.000000, test time steps: 200
episode: 1799, train score: 14.000000, test score: 15.000000, test time steps: 140
episode: 1899, train score: 2.000000, test score: 15.000000, test time steps: 113
episode: 1999, train score: 15.000000, test score: 15.000000, test time steps: 125

Another Example

episode: 99, train score: 0.000000, test score: 1.000000, test time steps: 9
episode: 199, train score: 2.000000, test score: 3.000000, test time steps: 21
episode: 299, train score: 3.000000, test score: 2.000000, test time steps: 15
episode: 399, train score: 2.000000, test score: 8.000000, test time steps: 65
episode: 499, train score: 3.000000, test score: 12.000000, test time steps: 91
episode: 599, train score: 0.000000, test score: 11.000000, test time steps: 200
episode: 699, train score: 2.000000, test score: 13.000000, test time steps: 200
episode: 799, train score: 15.000000, test score: 15.000000, test time steps: 86
episode: 899, train score: 2.000000, test score: 13.000000, test time steps: 200
episode: 999, train score: 0.000000, test score: 14.000000, test time steps: 200
episode: 1099, train score: 14.000000, test score: 14.000000, test time steps: 200
episode: 1199, train score: 10.000000, test score: 15.000000, test time steps: 157
episode: 1299, train score: 6.000000, test score: 15.000000, test time steps: 145
episode: 1399, train score: 11.000000, test score: 15.000000, test time steps: 121
episode: 1499, train score: 0.000000, test score: 13.000000, test time steps: 200
episode: 1599, train score: 15.000000, test score: 13.000000, test time steps: 200
episode: 1699, train score: 8.000000, test score: 15.000000, test time steps: 141
episode: 1799, train score: 6.000000, test score: 15.000000, test time steps: 141
episode: 1899, train score: 12.000000, test score: 15.000000, test time steps: 115
episode: 1999, train score: 14.000000, test score: 14.000000, test time steps: 200