D3 Dynamic Programming

Jaemin Park

2021-01-20

차 례

Exercises

Exercise 1

How would you generalize this game with arbitrary value of m_1 (minimum increment),m_2 (maximum increment), and N (the winning number)?

m_1 is 1(in baskin robbins game)
m_2 is 2(in BR game)
N is 31
When S = N-m_1, action - m_1,
When S = N-m_2, action - m_2

optimal policy is to make state as N-k(m_1+m_2),(k= natural number)

Exercise 2

(????)

Two players are to play a game. The two players take turns to call out integers. The rules are as follows. Describe A's winning strategy.

1. A must call out an integer between 4 and 8, inclusive.
2. B must call out a number by adding A's last number and an integer between 5 and 9, inclusive.
3. A must call out a number by adding B's last number and an integer between 2 and 6, inclusive.
4. Keep playing until the number larger than or equal to 100 is called by the winner of this game.

Since player who called larger or equal to 100 is winner, A's winning number is 89,90 and B's winning number is 89-93.

Winning number of A - {89,90},{78,79},{67,68},{56,57},{45,46},{34,35},{23,24},{12,13}

Winning number of B - {89-93},{78-82},{67-71},{56-60},{45-49},{34-38},{23-27},{12-16}

Since A starts with number between 4 and 8, then whatever A start with, B always can call B's winning number.

It is hard to find A's winning strategy in this situation.

Exercise 3

There is only finite number of deterministic stationary policy. How many is it? $|\Pi| = ?$

If we say number of states as S, and number of possible actions as A,
then number of deterministic stationary policy is A^S

Exercise 4

(???)

Formulate the first example in this lecture note using the terminology including state, action, reward, policy, transition. Describe the optimal policy using the terminology as well.

State = {1,2,3,4⋯30,31}
Action = {a_1(increase 1), a_2(increase 2)}
Reward = R(30| a_1) = R(29| a_2) = 1
otherwise R(s,a) = 0
Transition =
$P^a_{ss'} = P(S_{t+1} = s+1 | S_t = s, A_t = a1) = P(S_{t+1} = s+2 | S_t = s, A_t = a2) = 1$
otherwise, 0
Policy - Each state we can choose 2 actions a_1, a_2. number of policy is 2^{31}

Optimal policy is making reward as 1. So when state matches s%3 = 0, then optimal policy is using action a_2, if state s%3 = 0 then optimal policy is using action a_1.

Excercise 5

From the first example,

- Assume that your opponent increments by 1 with prob 0.5 and by 2 with prob 0.5
- Assume that the winning number is 0 instead of 31
- your opponent played first and she called out 1
- your current a policy π_0 is that
  - if the current state s $\leq$5 then increment by 2
  - if the current state s > 5 then increment by 1

Evaluate $V^{\pi_0}(1)$.

```
import numpy as np
current_state = 1
me_call = True
while True:
  if(current_state<=5):
    current_state+=2
  else:
    current_state+=1
  if(current_state>=10):
    break
  me_call = False
  prob = np.random.uniform(0,1)
  if prob<0.5:
    current_state+=1
  else:
    current_state+=2
  if(current_state>=10):
    break
  me_call = True
if(me_call):
  ("winner is me")
else:
  ("winner is opponent")
```

```
## 'winner is me'
```