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)?

$m_1 = 1$

$m_2 = 2$

$N = 21$

arbitrary

say that end number.

$\{28, 25, 22, 19, 16, 13, 10, 7, 4, 1\}$

$\Rightarrow 3a+1 \quad (a \geq 0)$

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.

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

Winning Stategy : To make end number 89. In my turn.

- Variation of policy
  - There is a *deterministic* policy and a *random* policy, where the former gives an single action for each state and the latter may give a distribution of multiple action for each state.
  - There is a *stationary* policy and a *non-stationary* policy. The stationary policy is what we have discussed, i.e. $\pi : \mathcal{S} \to \mathcal{A}$. On the other hand, the non-stationary policy is $\pi : \mathcal{S} \times \mathcal{T} \to \mathcal{A}$.
  - Non-stationary policy means th output action may be different on the same state, if the current time step is diffferent.
  - For a infinite horizon problems, the optimal policy is guaranteed to be a stationary policy. For a finite horizon problems, the optimal policy may be a non-stationary policy. Dealing with non-stationary policy is painful task in general. In this case, it is often desirable to include time information to state description.

Exercise 3 ?

There is only finite number of deterministic stationary policy. How many is it?

$$|\Pi| =$$

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 $= \{1, 2, 3, 4 \ldots 31\}$

action state $= \{a_1, a_2\}$

reward $= R(30, a_1) = 1 = R(29, a_2)$ others, $R(s, a) = 0$

transition $P_{ss'}^{a} = \mathbb{P}[S_{t+1} = s' \mid S_t = s, A_t = a] = 1$

optimal policy $= \pi : S(30) \to a_1$

(π^*)

$\pi : S(29) \to a_2$

CamScanner로 스캔하기