Lecture D3. Dynamic Programming

Sim, Min Kyu, Ph.D., mksim@seoultech.ac.kr

SEOULTECH 서울과학기술대학교 데이터사이언스학과

I. Motivation

Motivation - Reaching to a number (a.k.a. Baskin Robbins)

- A and B are to play a game. They take turn to call out integers.
  1. The serving player must call out an integer between 1 or 2.
  2. The opponent player 1) takes the other player's number and 2) increments it by 1 or 2, then 3) call out the number.
  3. Keep playing back and forth until someone calling out the number 31. The person calling out 31 is winner. (Incrementing 1 from 30 or incrementing 2 from 29 is winning).

- Do you want to go first or second? What is your winning strategy?

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)?

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

II. Some terminology

State

- The *state space* is the integer between 1 and 31.
- $\mathcal{S} = \{1, 2, 3, \cdots, 31\}$.

Action

- In each state, a player may choose among two possible *actions*.
- Namely, we may write a_1 and a_2, where
  - a_1 means the action of incrementing the previous number by 1 and
  - a_2 means the action of incrementing the previous number by 2.
- The *action space* $\mathcal{A} = \{a_1, a_2\}$.
- For each state, the player is to choose one among the possible action.
- Among the possible action, there exists an *optimal action*. The existence of optimal action is provable.

Random component

- In a fully *deterministic system*, the transition is governed by the previous state. In other words,

$$S_{t+1} = f(S_t)$$

- In *DTMC* and *MRP*, the transition was governed both by the previous state and some randomness. In other words,

$$S_{t+1} = f(S_t, \text{some randomness})$$

- In this problem (*Dynamic Programming*), the transition is governed by the previous state and the player's action. In other words,

$$S_{t+1} = f(S_t, A_t)$$

  There is no random component in transition.
- In *MDP*, the transition is affeced by randomness again. In other words,

$$S_{t+1} = f(S_t, A_t, \text{some randomness})$$

Reward function

- MRP fashion
  - In this problem, the reward is given only on the terminal state.
  - Using MRP's notation, you may describe it using *reward function*, $R(s) = \mathbb{E}[r_t | S_t = s]$.
  - Namely, $R(31) = 1$, and $R(s) = 0$ for all other s.
- DP fashion
  - Alternatively, since this problem has the action component, it is more natural to include action to the *reward function*.
  - Formally, $R(s, a) = \mathbb{E}[r_t | S_t = s, A_t = s]$.
  - Namely, $R(30, a_1) = R(29, a_2) = 1$ and all other $R(s, a) = 0$.

Policy

- For a particular state, there is an optimal action. But you feel that identifying an optimal action for a single state does not suffice. It is not sufficient in 'solving a problem.'
- A *policy* specifies which action to take on each state.
- A *policy* must include all contingent action plan for all possible scenario.
- *Strategy* and *policy* are interchangeable term in sequential optimization problem. But *strategy* is preferred term in economics, and *policy* is more preferred term in engineering.

Optimal Policy

- Solving a problem in real sense is to find *an optimal action for all possible states*.
- In other words, the *optimal policy* must include all contingent action plan for all possible scenario. Each action must be optimal for in any case.
- Among the all possible *policies*, there exists an *optimal policy* that maximizes the expected return(discounted sum of rewards).

Formulation

- Policy is a new thing. How to formularize?
- A policy function $\pi(\cdot)$ maps a state into actions. Namely, $\pi : \mathcal{S} \to \mathcal{A}$
  - For example, if your policy includes an action plan of playing a_1 on state s_3, then $a_1 = \pi(3)$.
- A policy may include randomized actions with a distribution. In this case we call *random* policy, as opposed to the previous *deterministic* policy.
  - For example, if your policy function $\pi(\cdot)$ says you should play a_1 with prob. 0.3 and a_2 with prob. 0.7 on the state s_3, then $\mathbb{P}(\pi(s_3) = a_1) = 0.3$ and $\mathbb{P}(\pi(s_3) = a_2) = 0.7$.

The goal of DP in terms of policy

- The goal of sequential optimization is to find a policy that maximizes the state-value function $V_t(s)$.
- For a policy π, there is a corresponding value function, written as $V_t^{\pi}(s)$.
- A policy π is an optimal policy if it maximizes $V_t^{\pi}(s)$ among all possible π.
- We notate *optimal* policy as π^*.
- That is,

$$\pi^* = argmax_{\pi \in \Pi} V_t^{\pi}(s), \forall s$$

Class of policy

- *Deterministic* policy vs *Random* policy
  - *Deterministic* policy gives an single action for each state.
  - *Random* policy gives a distribution of multiple actions for each state.
- *Stationary* policy vs *non-stationary* policy
  - The stationary policy is what we have discussed, i.e. $\pi : \mathcal{S} \to \mathcal{A}$.
  - The non-stationary policy is $\pi : \mathcal{S} \times \mathcal{T} \to \mathcal{A}$.
  - Non-stationary policy means the action in the same state can be different if time is different.

Optimality of stationary policy

- 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 desirable to include time information to state description.

III. Exercises

Exercise 3

There exist only finite number of deterministic stationary policies. If the number of state is $|\mathcal{S}|$ *and the number of action is* $|\mathcal{A}|$*, then 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.

"If you torture your data enough, it will confess the truth. - Alex Shapiro"