Lec-5, IT567, 24-25

$$r(s,a,s') = \sum_{r \in R} r \cdot p(\underline{r \mid s'}, a, s)$$

$$p(r \mid s', a, s) = \frac{p(r, s' \mid s, a)}{p(s' \mid s, a)}$$

(see eq. 1,2,3 in the left)

$$p(s', r \mid s, a)$$

$$p(x,y) = p(x)\, p(y \mid x)$$
$$= p(y)\, p(x \mid y) \quad (1)$$

$$p(x,y \mid z) = p(x \mid z)\, p(y \mid x, z)$$
$$= p(y \mid z)\, p(x \mid y, z) \quad (2)$$

$$p(r, s' \mid s, a) = p(r \mid s, a) \times p(s' \mid r, s, a)$$
$$= p(s' \mid s, a) \times p(r \mid s, a, s') \quad (3)$$

MDP framework:- is abstract & flexible & can be applied to many different problems in many " ways.

for extime steps:- may refer to arbitrary successive stages of decision making & acting.

actions:-
- → low-level controls such as voltage
  or
- → high-level decisions — whether or not to have lunch or go to graduate school.

<u>Goals & Rewards</u>:- Goal is formalized in terms of a special signal, called the reward passing from environment to agent. $R_t \in R$

" ↳ maximization of the expected value of the cumulative sum of a received scalar signal (called reward)"

→ most distinctive feature of RL. – Read last paragraph on pg 53.

Returns & Episodes:-

Goal is to maximize the cumulative reward it receives in the long run.

Return $\triangleq G_t = R_{t+1} + R_{t+2} + \dots + R_T$

where T → final time step (FTS)

FTS :- when the agent-environment interaction breaks naturally into subsequences which we call episodes (or trials)

Chess → terminal state

- win
- loss
- draw

ex- plays of a game, trips through a maze, or any sort of repeated interaction

each episode ends in a special state called terminal state, followed by a reset to a standard starting state or to a sample from a standard distribution of starting states.

episodes are independent

episodes can be considered to end in the same terminal state, with different rewards for the diff. outcomes.

These are called 'episodic tasks'.