Lec-3, IT 567, 24-25

Policy:- agent's way of behaving at a given time, mapping from perceived states of the environment to actions to be taken when in these states.

→ sufficient to det. the behavior

→ can be stochastic → specifying prob. for each action.

$f(s)$

Reward:- defines the goal of a RL problem. The environment sends to the RL agent a single no. called the reward.

↳ defines good or bad events.

may also be a stochastic function of the state of the environment & the actions taken $f(s,a)$

If an action selected by the policy is followed by low reward, then the policy may be changed to select some other action in that situation in the future.

→ Value function :- reward signal indicates what is good in an immediate sense, a value func. specifies what is good in the long run.

* Value of a state :- is the total amount of reward an agent can expect to accumulate over the

future, starting from that state.

→ Rewards are given by the environment (directly) but values must be estimated & re-estimated from the seq. of observations an agent makes over its entire lifetime.

RL algos:- mostly methods for efficiently estimating values of states.

$V(s)$ $\forall$ possible s.

→ model of the environment:- we know how the environment will behave, for ex, given a state & action, the model might predict the resultant next state & next reward.

Google Deepmind

For deciding on a course of action, models can be used for it.

→ Model-Based → use model & plan

→ Model-free

↳ trial & error based purely

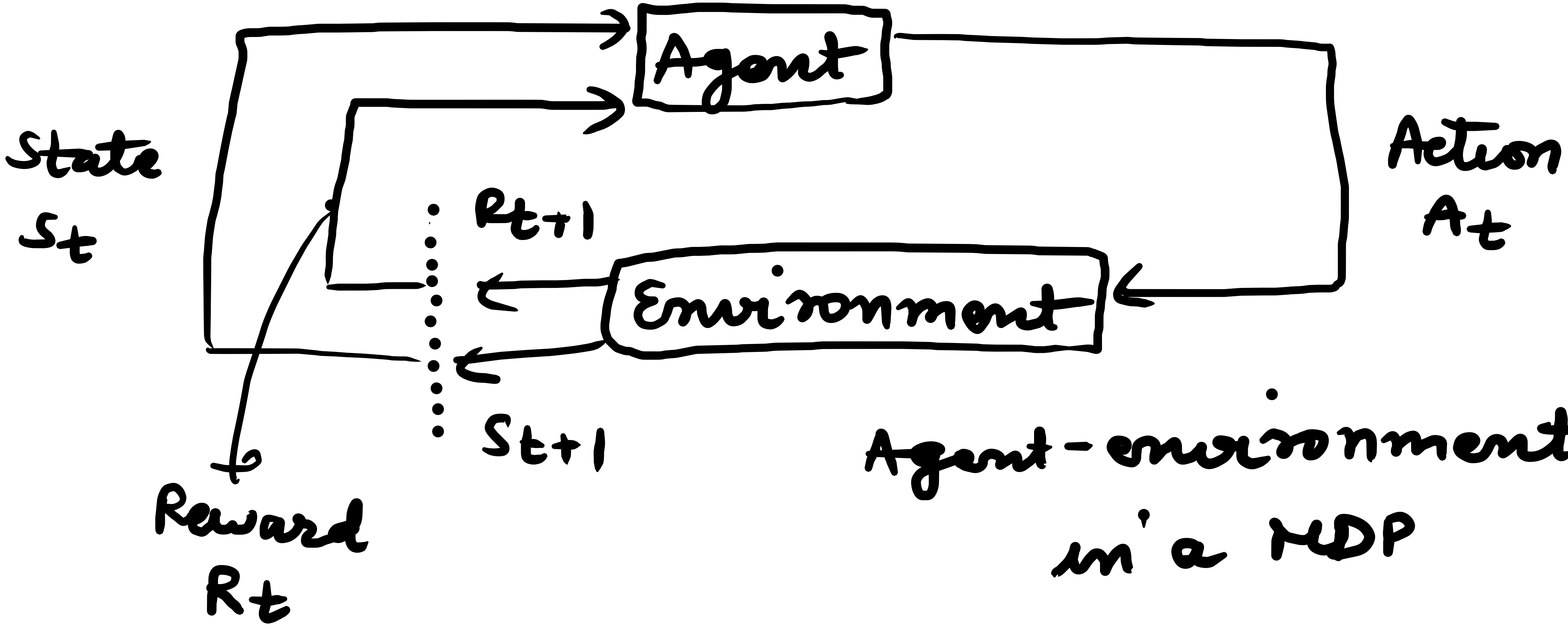
Chapter 3:

1. The learner & decision maker is called the agent

2. The thing it interacts with, comprising everything outside the agent, is called the environment.

MDPs :- Markov decision process.

MDPs are a mathematically idealized form of the RL problem, for which precise theoretical statements can be made.



Agent-environment interaction in a MDP

We use R_{t+1} instead of R_t to denote the reward due to A_t because it emphasizes that the next reward & next state R_{t+1} & S_{t+1} are jointly determined. – used widely in the literature.

→ Also, MDPs are a classical formalization of sequential decision making, whose actions influence not just immediate rewards, but also subsequent situations or states, & through those future rewards.