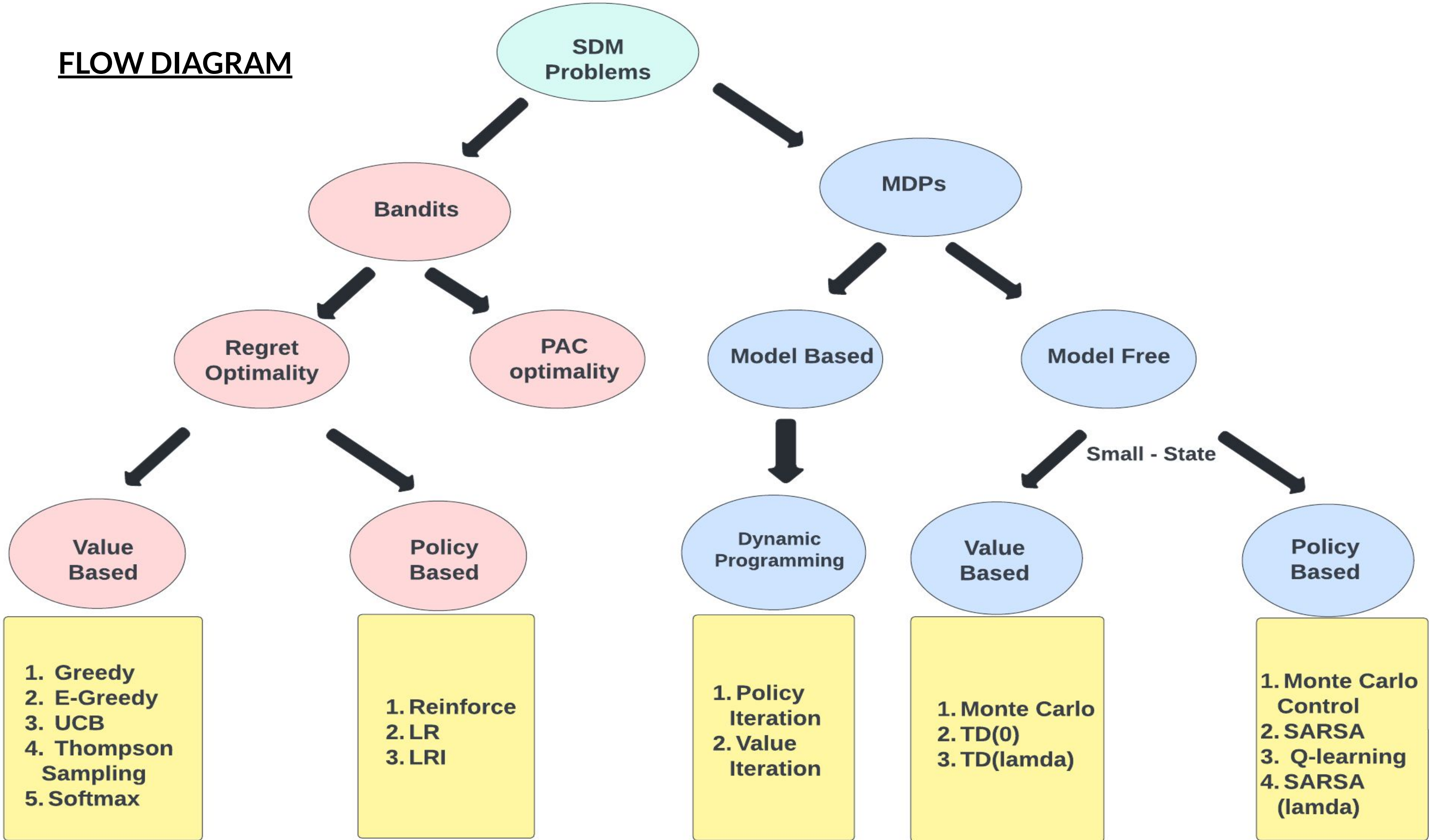
RL - Summary

Name - Arvind Pandit, Sumanth Raikar
Roll No. - 211022001,211022005

FLOW DIAGRAM

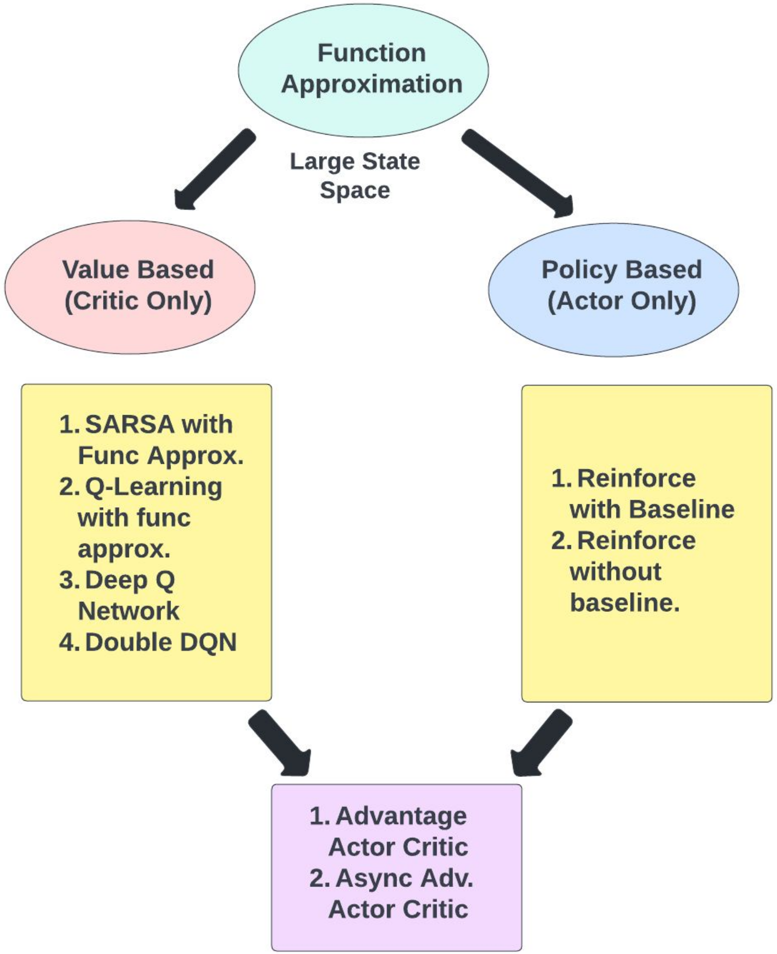


Bandits Algorithms

- ❖ Bandit Algorithms are based on the Regret optimality where the goal is to minimize the regret or maximize average cumulative reward by choosing an optimal arm.
- ❖ In bandit settings the available choices and rewards tomorrow are not affected by decisions taken today.
- ❖ For long term planning bandits are not useful because it only try to maximize the cumulative reward.

MDPs (Small State Space)

- ❖ In MDPs, if the model information is available such as <S,A,P,R> then we formulate it as a Dynamic Programming problem and solve it by using Value Iteration or Policy Iteration.
- ❖ If model information is not available we approach for the Tabular methods such as Monte- Carlo, TD(lamda), SARSA, Q-learning.
- ❖ There is a notion of
  - ➢ On-Policy - Evaluate or improve the policy that is used to make decisions.
  - ➢ Off-Policy - Evaluate or improve a policy different from that used to generate the data.
- ❖ This methods cannot work for large/continuous State-Action spaces due to memory insufficiency to maintain the Value Table.



MDPs (Large State Space)

- To overcome the limitations of Tabular methods function Approximation uses the linear or nonlinear function to approximate the Value or Policy.
- Critic Only - It approximates the Value function of a given state and takes the action which maximizes the immediate + future rewards.
- Actor Only - It directly approximates the Policy(action) for the given state.
- Actor-Critic methods use both Value as Policy function approximation.

Difficulty Faced

- Implementation of the Tile Coding and RBF based FA algorithms.
- Eligibility Traces, Off-Policy MC .
- Suddenly introduction of Neural Networks in Algorithms.
- Creating an custom environment can be helpful for the better understanding and formulation of the real-world problem.

THANK YOU