**1. RL for Optimal Energy Management of a Solar Microgrid**

1. RL algo (3 step ahead Q-learning) used to optimize the battery scheduling in dynamic environment of load and available solar power.
2. Increasing utility of battery and solar power generator leads to optimal performance of solar microgrid.
3. Microgrid - localized grouping of electricity sources and loads that normally operates synchronous with traditional centralized grid connected and disconnected and functions autonomously as physical.
4. Modelling of consumer agent
   1. Dynamic variation of load, solar power and battery considered as external environment
   2. Consumer modeled as individual agent (who uses RL for decision making)
   3. RL deals with leading in sequential decision making(SDM) - Markov Decision Process (MDP) has become standard for SDM
   4. In MDP - environment modelled as system states. Effect of actions taken in a state dependent on that state and not prior history.
   5. THE GOAL is to control system such that some performance criterion is maximized.
   6. Notions used:
      1. Solar power output: Psp
      2. Load: Dt
      3. Level of battery charged: Rt
   7. RL algorithm
      1. Models consumers adaptation to dynamically changing environment by performing actions of battery scheduling in an MDP environment.
      2. Agents observe environment and take action. It gets reward/punishment from the environment.
      3. Agent takes next action to optimize the reward in long run.
      4. After number of iteration, agent finds optimal policy to achieve long term objective.
      5. Environment characterized by states: S(t) at time t.
         1. Intrinsic value
         2. Dependent upon a certain immediate reward or cost - R(t) [Generated when state is entered]
   8. MDP - way to model sequential decision making under uncertainty
      1. Initial state s0 and each state will have reward r associated with it.
      2. Transition function T(s’|a,s) indicates probability of transitioning from state s to s’ when action a is taken. A discount factor Y in range 0 - 1 is applied to future rewards.
      3. Represents that current reward is more important than future, as if this is near to zero then future rewards are almost ignored. On the other hand, near one places great value in future rewards.
      4. Reward from policy: Sum of discounted expected utility of each state visited by that policy
      5. Optimal policy maximizes total expected discounted reward.
   9. Q learning
      1. Model free RL agent that explores env and finds next reward and best agent can do from next state.
      2. Only needs to know what states exists and what actions are possible in each state.
      3. Assign each state an estimated value - Q value
      4. When we visit a state and take an action, we receive a reward
      5. We visit states infinitely and update action values till it becomes convergent