Reward Function → $r_{t+1} = R(s_t, a_t, s_{t+1})$

↳ This formulation is called a Markov Decision Process

MDP: A method to select an action a given state s. Then observe a' & s' based on transition probs P.

★ Most DQNs are flat CNNs + batch normalization

★ DQN is the agent's learning in the env.

Markov Decision Processes

Special stochastic time control process for decision making which assumes random probability & a decision maker having complete control.

S: Set of states. At each time step, the state of the environment is an element $s \in S$.

A: Set of actions. At each time step, agent chooses an action $a \in A$ to perform.

$p(s_{t+1} | s_t, a_t)$: State transition model that describes how the env state changes when user performs an action a depending on current state s

$p(r_{t+1} | s_t, a_t)$: Reward model that describes the real valued reward value that the agent receives from the env after performing an action. (Depends on state & action)

γ: Discount factor that controls importance of future rewards.

— In MDP, we search for a policy function that the agent or decision maker will choose in next state s.