

Markov Decision Process

Sargur N. Srihari
srihari@cedar.buffalo.edu

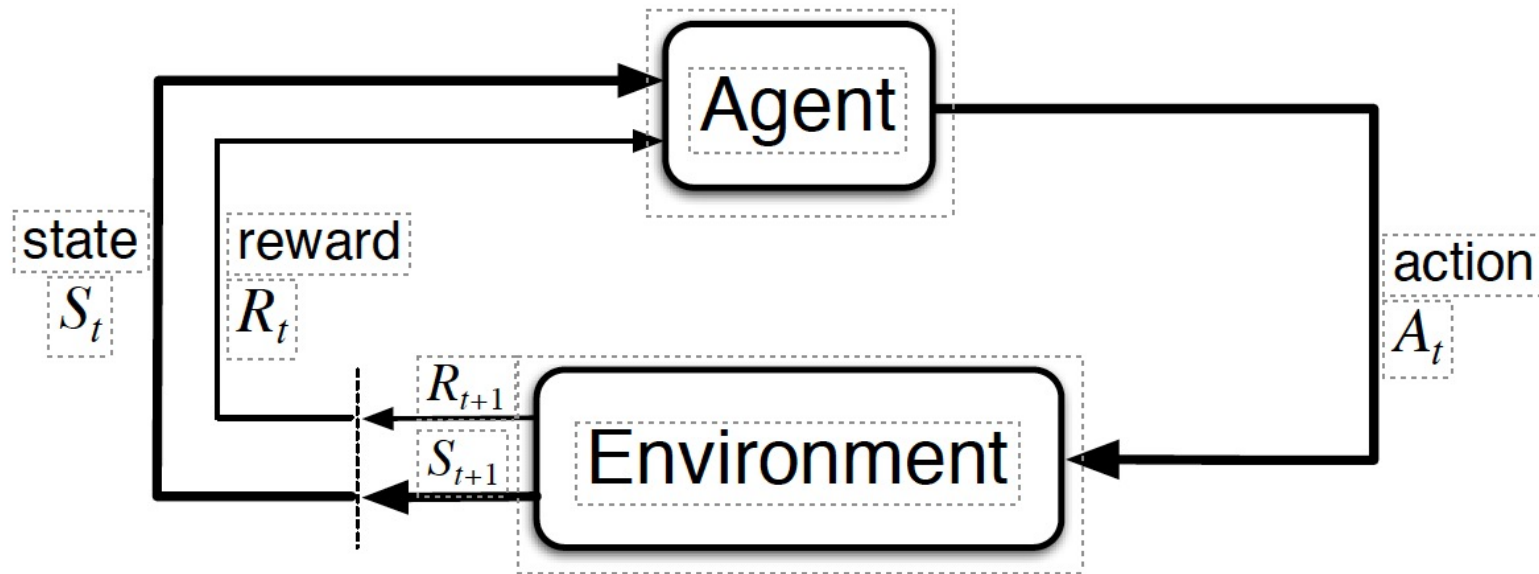
Topics in Markov Decision Process

1. How MDP differs from Bandit
2. Agent-Environment Interaction
3. Finite Markov Process Definition

How MDP differs from Bandit

- Actions influence not just immediate rewards, but also subsequent situations, or states, and through those future rewards
- Whereas in bandit problems we estimated the value $q^*(a)$ of each action a , in MDPs we estimate the value $q^*(s, a)$ of each action a in each state s , or we estimate the value $v^*(s)$ of each state given optimal action selections

Agent-environment Interaction in MDP



Agent: Learner and decision maker

Everything outside the agent, is the Environment.

MDP and agent give rise to a sequence or trajectory that begins:

$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \dots$

Finite Markov Decision Process

- Finite no. of States (\mathcal{S}), actions (\mathcal{A}), rewards (\mathcal{R})
 - Reward, State conditional distribution $p(s', r | s, a)$

$$p(s', r | s, a) \triangleq \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\}$$

- Note that

$$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r | s, a) = 1, \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}(s)$$

$$p(s' | s, a) \triangleq \Pr\{S_t = s' \mid S_{t-1} = s, A_{t-1} = a\} = \sum_{r \in \mathcal{R}} p(s', r | s, a)$$

- Expectations

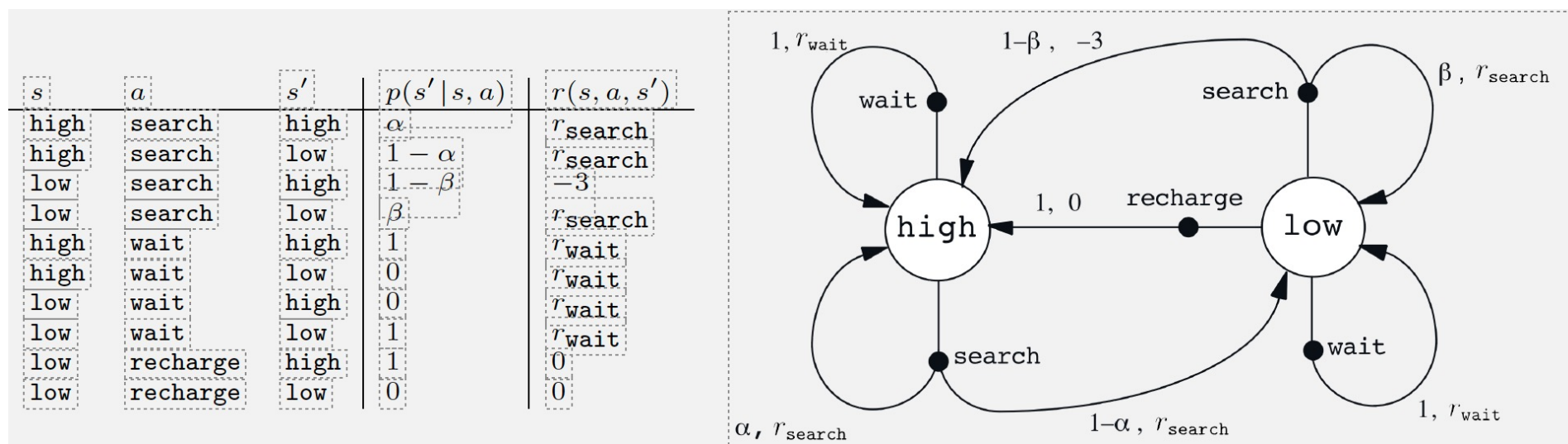
$$r(s, a) \triangleq \mathbb{E}[R_t \mid S_{t-1} = s, A_{t-1} = a] = \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r | s, a)$$

$$r(s, a, s') \triangleq \mathbb{E}[R_t \mid S_{t-1} = s, A_{t-1} = a, S_t = s'] = \sum_{r \in \mathcal{R}} r \frac{p(s', r | s, a)}{p(s' | s, a)}$$

Recycling Robot

- Collect empty cans, Sensors detect cans,
- Arm to pick up, Runs on battery
- Decisions to find cans based on battery charge
- States are charge levels $S=\{high, low\}$
 - In each state, the agent can decide whether to
 1. actively search for a can for a period of time
 2. remain stationary, wait for someone to bring can
 3. head back to home base to recharge its battery
- $A(high)=\{search, wait\}$, $A(low)=\{search, wait, recharge\}$
- Rewards zero most of the time, but positive when the robot secures a can, or large and negative if the battery runs all the way down

Finite MDP for Recycling Robot



Goals and Rewards

- Goal: maximization of the expected value of the cumulative sum of a received scalar reward
- Reward signal is way of communicating to the robot what you want it to achieve, not how it is achieved

Returns and Episodes

- Simplest case is to maximize expected return:

$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \dots + R_T ,$$

- Where T is the final time step

- Notion of final time step useful in
 - Plays of a game, Trips through a maze
- Each sequence is called an *episode*
- Expected Discounted return

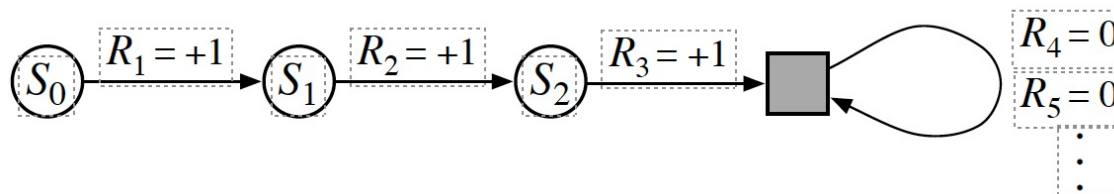
$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

- Reward after k steps is worth γ^{k-1} times immediate reward

$$\begin{aligned} G_t &\doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \dots \\ &= R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \dots) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned}$$

Unified Notation for Episodic and Continued

- Absorbing state



$$G_t \triangleq \sum_{k=t+1}^T \gamma^{k-t-1} R_k,$$

Policies

- A policy is a mapping from states to probabilities of selecting each possible action
- If agent is following policy π at time t , then $\pi(a|s)$ is the probability that $A_t = a$ if $S_t = s$
- Like p , π is an ordinary function; $\pi(a|s)$ defines a distribution over $a \in A(s)$ for each $s \in S$
- Reinforcement learning methods specify how the agent's policy is changed as a result of its experience

State Value Function for policy π

- Value of a state s under a policy π , denoted $v_\pi(s)$, is the expected return when starting in s and following π thereafter
- For MDPs we can define v_π formally as

$$v_\pi(s) \stackrel{\text{def}}{=} \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s \right], \text{ for all } s \in \mathcal{S}$$

Action-value function for policy π

- Value of taking action a in state s under a policy π , denoted $q_\pi(s, a)$, is the expected return starting from s , taking the action a , and thereafter following policy π :

$$q_\pi(s, a) \triangleq \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a \right]$$