Markov Decision Process

Discrete-time Stochastic Control Process

- Control: Make decisions to achieve the goals of the task
- Stochastic: Agent's action only *partially* affects the evolution of the task
- Discrete-time: time progresses in finite intervals

Markov Decision Processes can be represented by 4-tuplet, (S, A, R, P):

Component	Description
States, S	Set of all possible states
Actions, A	Set of actions that can be taken in each of the states
Rewards, R	Set of rewards for each (s,a) pair
Probabilities, P	Probabilities of passing from one state to another when taking each action

MDP has no memory:

$$P[s_{t+1} | S_t = s_t] = P[S_{t+1} | S_t = s_t, S_{t-1} = s_{t-1}, \cdots, S_0 = s_0]$$

The next state only depends on current state and none of the previous states.

Types of MDP

Finite vs. Infinite

- Finite: States, Actions and Rewards are finite. (Chess)
- Infinite: >= 1 of the three is infinite. (Driving Car)

Episodic vs. Continuing

- Episodic: Terminates under certain condition
- Continuing: Simply keeps going

Trajectory vs. Episode

- Trajectory: Elements that are generated when the agent moves from one state to another
  > $\tau = S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3$
- Episode: Trajectory from initial state of the task to the terminal state
  > $\tau = S_0, A_0, R_1, S_1, A_1, R_2, S_2, \cdots, R_T, S_T$

Reward vs. Return

- Reward: Immediate result that our action produces (Note that short-term rewards can worsen long-term results.)
  > R_t

- Return: Sum of rewards that our agent obtains from certain point in time, t, until the task is completed, T. (The main task is to maximize Return)

  > $G_t = R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$

Discount Factor

Incentivizes the agent to reach goal through the shortest route.

> $G_t = R_1 + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-1} R_T$

where $\gamma \in [0, 1]$ is the discount factor. Thus, the longer it takes, the less the return.

Policy

Takes state as input and decides what action to take (action taken in s):

> $\pi : S \rightarrow A$

Probability of Action a in State s:

> $\pi(a|s)$

Deterministic Policy

Get only one action

> $\pi(s) = a_1$

Stochastic Policy

Get probability for each action.

> $\pi(s) = [p(a_1), p(a_2), p(a_3), \cdots, p(a_n)] = [0.3, 0.2, 0.5]$

Optimal Policy

π_* is the policy that chooses that action that maximizes the Return.

State Value

$v_\pi(s) = E[G_t | S_t = s]$

where $v_\pi(s) = E[R_{t+1} + \gamma R_{t+2} + \cdots + \gamma R^{T-t-1} | S_t = s]$ following policy π.

State-Action Value

$q_\pi(s, a) = E[G_t | S_t = s, A_t = a]$

where $q_\pi(s, a) = E[R_{t+1} + \gamma R_{t+2} + \cdots + \gamma R^{T-t-1} | S_t = s, A_t = a]$ following policy π.

Bellman Equations

State Value

$v_\pi(s) = E[G_t | S_t = s]$

$= E[R_{t+1} + \gamma R_{t+2} + \cdots + \gamma R^{T-t-1} | S_t = s]$

$= E[R_1 + \gamma G_{t+2} | S_t = s]$

$= \sum_a \pi(a|s) \sum_{s',r} p(s, r | s, a)[r + \gamma v_\pi(s')]$

following policy π.

State-Action Value

$q_\pi(s,a) = E[G_t \mid S_t = s, A_t = a]$

$= E[R_{t+1} + \gamma R_{t+2} + \cdots + \gamma R^{T-t-1} \mid S_t = s, A_t = a]$

$= E[R_1 + \gamma G_{t+2} \mid S_t = s, A_t = a]$

$= \sum_{s',r} p(s', r \mid s, a)[r + \gamma \sum_{a'} \pi(a' \mid s') q_\pi(s', a')]$

following policy π.

Bellman Optimality Equations

The optimal policy π_* is the one that chooses **actions** that *maximizes* $v(s)$ or $q(s,a)$:

$v_*(s)$
$= E_{\pi_*}[G_t \mid S_t = s]$
$= \max_a \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_*(s')]$

where $\pi_*(s) = \arg\max_a \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_*(s)]$, and

$q_*(s,a)$
$= E_{\pi_*}[G_t \mid S_t = s, A_t = a]$
$= \sum_{s',r} p(s', r \mid s, a)[r + \gamma \max_{a'} q_*(s', a')]$

where $\pi_*(s) = \arg\max_a q_*(s,a)$.