Reinforcement learning: Markov Decision Processes
Written by Olga Bane (May 2020)

Markov decision processes (MDPs) describe **sequential decision making** whereaby action influence immediate reward and subsequent states and rewards. There is therefore a need to **tradeoff immediate and delayed reward** in MDP, as our goal (in reinforcement learning in general) is to maximize the total reward.

In **bandits**, we estimated the value of each action a as $q_*(a)$.
In **MDPs**, we estimate the value of each action a in each state s as $q_*(s, a)$ or the value of each state $v_*(s)$ given optimal action selections.

Agents and environments

An **agent** is the learner and decision maker that interacts with the **environment**. The agent and the environment interact continuously at each time step: the agent performs an action based on the environment's **state**, and the environment presents a **new state** and a **reward**, which the agent tries to maximize through its actions.

The trajectory therefore begins: $S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2...$

In a *finite* MDP, the sets of states, actions, and rewards are finite. States and Rewards at time t (R_t, S_t) are random variables that have well defined discrete probability distributions, which depend only on the preceding *immediate* State and Action (S_{t-1}, A_{t-1}).

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

$p(s', r|s, a)$ defines the dynamics of the MDP. $p = S * R * S * A$, and specifies a probability distribution with AUC = 1. The must include all past agent-environment interactions that make a difference for the future (called the **Markov property**).

Goals and rewards

The reward describes the **goal** of the agent. The agent should not be rewarded for steps that we *think* would help the agent achieve the goal, but rather only for reaching the goal.

Returns and episodes

The agent's goal is to **maximize total expected total reward**, G_t. In the simplest case, the agent-environment interaction breaks into subsequences (*episodes*), where the final timestep T is defined. The next episode begins independently of how the previous ended. Time of termination T is a random variable that varies from episode to episode.

$$G_t = R_{t+1} + R_{t+2} + ... + R_T$$

An agent-environment interaction could alternatively go on continuously, without the ability to define episodes or the final timestep T. This poses a difficulty - if T could be *infinite*, then *the return that we are trying to maximize could also be infinite.*

Next:
1. Discounted returns