# DS 102 Discussion 9

Monday, 13 April, 2020

In this discussion, we'll review the concepts of the value function $V(s)$ and Q-function $Q(s,a)$ introduced in Lectures 23 and 24, and practice going through the computations needed to solve them (*i.e.*, compute them, similar to what you'll be doing on HW 5).

First, a brief overview of Markov Decision Process (MDP) terminology:

- $s \in S$: states
- $a \in A$: actions we can take from states
- $\mathbb{P}(s' \mid s, a)$: transition function, capturing the distribution over states we will end up in if we take action $a$ from state $s$
- $R(s, a, s')$: reward function, which we receive at each iteration when we take action $a$ from state $s$ to end up in state $s'$.
- $\gamma \in [0, 1]$: discount factor for rewards received after the current iteration
- $\pi : S \to A$: policy, describing a strategy of what action to take from a state

The value function $V^\pi(s)$ of a policy $\pi$ gives the expected (discounted) reward received when starting from state $s$ and using strategy $\pi$:

$$V^\pi(s) = \sum_{a \in A} \pi(a \mid s) \sum_{s' \in S} \mathbb{P}(s' \mid s, a) \left[ R(s, a, s') + \gamma V^\pi(s') \right].$$

This equation is also known as the **Bellman equation**.

We are often interested in the value function of a particular policy: the one that is optimal from state $s$. This is the **optimal value function** $V^*(s)$:

$$V^*(s) = \max_{a \in A} \sum_{s' \in S} \mathbb{P}(s' \mid s, a) \left[ R(s, a, s') + \gamma V^*(s') \right].$$

Similarly, the **optimal Q-function** $Q^*(s, a)$ gives the expected (discounted) reward received when starting from state $s$, taking action $a$, then taking the optimal actions thereafter:

$$Q^*(s, a) = \sum_{s' \in S} \mathbb{P}(s' \mid s, a) \left[ R(s, a, s') + \gamma V^*(s') \right].$$

A typical goal in reinforcement learning is to find a policy $\pi^*$ that maximizes our expected discounted reward. Building up to that goal, we first need to understand how to evaluate the optimal value function and optimal Q-function.

1. We have the following grid representation of a problem:

| | | | |
|---|---|---|---|
| | | | 1 |
| | $\times$ | `start` | $-100$ |
| | | | |

where `start` represents our initial state, $\times$ is a state we can't access, and the 1 and $-100$ states are terminal states with corresponding rewards. The reward received when moving to any other state is zero.

(a) Assume state transitions are deterministic, meaning that an action in a particular direction always moves us in that direction (unless it's toward the $\times$ state, in which case we stay in the same state). Compute the optimal value function at each state, when $\gamma = 0.9$.

(b) Compute the optimal Q-function at our initial state for the actions of going up, down, left, and right.

(c) Based on the optimal Q-function you just computed, what would be the optimal move to make from `start`?

(d) Now suppose the state transitions are stochastic, such that there is a 0.8 probability of going in the direction you specified, and a 0.1 probability of going in either of the directions perpendicular to what specified. For example, if you decide to go up, you go up with 0.8 probability, go left with a 0.1 probability, and go right with a 0.1 probability. What is the best action to perform from `start`?