Reinforcement Learning

The value functions

Lecturer: *Nicolò Cesa-Bianchi* version January 19, 2023

This material is partially based on the book draft "Reinforcement Learning: Foundations" by Shie Mannor, Yishay Mansour, and Aviv Tamar.

We consider an MDP with finite state space $\mathcal{S}$, finite action space $\mathcal{A}$ such that $\mathcal{A}(s) = \mathcal{A}$ for all $s \in \mathcal{S}$, transition kernel $\{p(\cdot \mid s,a) : s \in \mathcal{S},\, a \in \mathcal{A}\}$, and time-dependent reward function $r_t : \mathcal{S} \times \mathcal{A} \to [-1,1]$.

We now define some quantities that will help us define the notion of optimal policy in an MDP. Consider the stochastic horizon case (for simplicity, without terminal reward) and an arbitrary stochastic policy $\boldsymbol{\pi} = (\pi_0, \pi_1, \dots)$. The **state-value function** $V_t^{\boldsymbol{\pi}} : \mathcal{S} \to \mathbb{R} \cup \{\infty\}$ gives the expected return obtained by running $\boldsymbol{\pi}$ from any state $s_t \in \mathcal{S}$ at time $t \geq 0$,

$$V_t^{\boldsymbol{\pi}}(s_t) = \mathbb{E}\left[\sum_{\tau=t}^{T} r_\tau(s_\tau, a_\tau)\right]$$

where $a_\tau \sim \pi_\tau(\cdot \mid s_\tau)$. The **action-value** function $Q_t^{\boldsymbol{\pi}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R} \cup \{\infty\}$ at time $t \geq 0$ is defined by

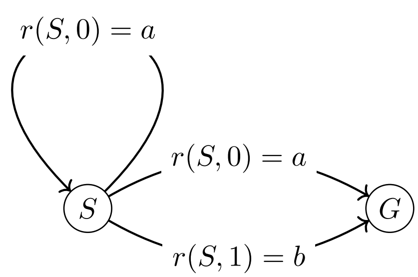
$$Q_t^{\boldsymbol{\pi}}(s,a) = r_t(s,a) + \sum_{s' \in \mathcal{S}} V_{t+1}^{\boldsymbol{\pi}}(s') p(s' \mid s,a)$$

This is the expected return of executing action a in state s at time t and then following policy $\boldsymbol{\pi}$. Hence,

$$V_t^{\boldsymbol{\pi}}(s) = \sum_{a \in \mathcal{A}} \left(r_t(s,a) + \sum_{s' \in \mathcal{S}} V_{t+1}^{\boldsymbol{\pi}}(s') p(s' \mid s,a) \right) \pi_t(a \mid s) = \sum_{a \in \mathcal{A}} Q_t^{\boldsymbol{\pi}}(s,a) \pi_t(a \mid s) \qquad s \in \mathcal{S}$$

For deterministic policies, the above equation becomes $V_t^{\boldsymbol{\pi}}(s) = Q_t^{\boldsymbol{\pi}}\big(s, \pi_t(s)\big)$ for any $s \in \mathcal{A}$ and $t \geq 0$.

Example. Consider the following game with two states, S (the initial state) and G (the goal state), and two actions, 0 and 1. Action 1 deterministically leads to the goal state with a reward of b. Action 0 always carries a reward of a with $0 < a < b$, leads to the goal state with probability p, and remains in state S with probability $1-p$.



We consider two deterministic and stationary Markov policies. Since everything is stationary, we can omit the subscripts t from V_t^π and Q_t^π. Policy π keeps on playing action 0 until the goal state is reached. Policy π' plays action 1 and immediately reaches the goal state. Clearly, $V^{\pi'}(S) = b$. On the other hand,

$$V^{\boldsymbol{\pi}}(S) = ap \sum_{k=1}^{\infty} k(1-p)^{k-1} = \frac{a}{p}$$

Hence, $V^{\boldsymbol{\pi}}(S) > V^{\boldsymbol{\pi}'}(S)$ if and only if $p < \frac{a}{b}$.

The action-value function for $\boldsymbol{\pi}$ is $Q^{\boldsymbol{\pi}}(S,0) = \frac{a}{p}$ and $Q^{\boldsymbol{\pi}}(S,1) = b$. Similarly, $Q^{\boldsymbol{\pi}'}(S,0) = a+(1-p)b$ and $Q^{\boldsymbol{\pi}'}(S,1) = b$.

We now characterize the optimal policy in the finite horizon case. As we have shown, in this case we do not lose generality by restricting to deterministic policies. To avoid confusion, we use H to denote the horizon value and we call stage any time step $h = 0, \dots, H$. Then, the expected return (or state-value function) of a deterministic policy $\boldsymbol{\pi} = (\pi_0, \dots, \pi_H)$ at stage h is

$$V_h^{\boldsymbol{\pi}}(s_h) = \mathbb{E}\left[\sum_{t=h}^{H} r_t\big(s_t, \pi_t(s_t)\big)\right] = r_h\big(s_h, \pi_h(s_h)\big) + \sum_{s' \in \mathcal{S}} V_{h+1}^{\boldsymbol{\pi}}(s') p\big(s' \mid s_h, \pi_h(s_h)\big)$$

and the action-value function at stage h is

$$Q_h^{\boldsymbol{\pi}}(s,a) = r_h(s,a) + \sum_{s' \in \mathcal{S}} V_{h+1}^{\boldsymbol{\pi}}(s') p(s' \mid s, a)$$

Let $\boldsymbol{\pi}^*$ be the optimal deterministic policy, satisfying $V_h^{\boldsymbol{\pi}^*}(s) \ge V_h^{\boldsymbol{\pi}}(s)$ for all $s \in \mathcal{S}$, $h \in \{0, \dots, H\}$, and all deterministic policies $\boldsymbol{\pi}$. For brevity, we write V_h^* and Q_h^*. Note that the optimal policy π_h^* at stage h is computed as

$$\pi_h^*(s) = \operatorname*{argmax}_{a \in \mathcal{A}} Q_h^*(s,a) \qquad s \in \mathcal{S}$$

Using backward induction, it is easy to compute the optimal state-value and action-value functions for all $h = 0, \dots, H$. Let V_{H+1}^* and Q_{H+1}^* be constant zero functions. First, observe that

$$\begin{aligned} Q_H^*(s,a) &= r_H(s,a) \\ V_H^*(s) &= \max_{a \in \mathcal{A}} r_H(s,a) = \max_{a \in \mathcal{A}} Q_H^*(s,a) \end{aligned}$$

Now, given Q_h^* and V_h^* for $h \in \{1, \dots, H\}$, we can compute Q_{h-1}^* and V_{h-1}^* as follows. By definition of action-value function,

$$Q_{h-1}^*(s,a) = r_{h-1}(s,a) + \sum_{s' \in \mathcal{S}} V_h^*(s') p(s' \mid s, a)$$

For the state-value function, we compute the optimal expected return from stage $h-1$ by maximizing the sum of the optimal reward at stage $h-1$ and the optimal expected return V_h^* from stage h onwards,

$$\begin{aligned} V_{h-1}^*(s) &= \max_{a \in \mathcal{A}} \left(r_{h-1}(s,a) + \sum_{s' \in \mathcal{S}} V_h^*(s') p(s' \mid s, a) \right) \\ &= \max_{a \in \mathcal{A}} Q_{h-1}^*(s,a) \end{aligned}$$

This is a manifestation of Bellman's **principle of optimality:** the tail of an optimal policy is optimal for the "tail" problem. The system of equations

$$V_h^*(s) = \max_{a \in \mathcal{A}} \left(r_h(s,a) + \sum_{s' \in \mathcal{S}} V_{h+1}^*(s') p(s' \mid s, a) \right) \qquad s \in \mathcal{A},\ h = 0, \ldots, H$$

is called the Bellman optimality equations.