Reinforcement Learning

Meng Oon Lee

September 12, 2021

1 Introduction

Discounted return at time step t:

$$G_t = \sum_{k=0} \gamma^k R_{t+k+1}, \quad \gamma \in [0,1] \tag{1}$$

One-step dynamics:

$$p(s', r|s, a) = \mathbb{P}(S_{t+1} = s', R_{t+1} = r|S_t = s, A_t = a), \tag{2}$$
$$s, s' \in S, \quad a \in A, \quad r \in R$$

Deterministic policy:

$$\pi : S \to A, \quad s \in S, \quad a \in A \tag{3}$$

Stochastic policy:

$$\pi : S \times A \to \pi(a|s) \in [0,1], \quad s \in S, \quad a \in A \tag{4}$$

State-value function:

$$v_\pi(s) = \mathbb{E}_\pi[G_t|S_t = s], \quad s \in S \tag{5}$$

Bellman expectation equation:

$$v_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s], \quad s \in \mathcal{S}, \quad \gamma \in [0,1] \tag{6}$$

Action-value function:

$$q_\pi(s, a) = \mathbb{E}_\pi[G_t|S_t = s, A_t = a], \quad s \in \mathcal{S}, \quad a \in \mathcal{A} \tag{7}$$

Optimal policy:

$$\pi_*(s) = \underset{a \in A(s)}{\arg\max}\, q_*(s, a), \quad s \in A(s) \tag{8}$$

Temporal-difference (TD):

1. TD Prediction:
$$V(S_t) \leftarrow (1-\alpha)V(S_t) + \alpha G_t \tag{9}$$

2. Sarsa (on-policy TD control):
$$Q(S_t, A_t) \leftarrow (1-\alpha)Q(S_t, A_t) + \alpha[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1})] \tag{10}$$

3. Q-learning (off-policy TD control):
$$Q(S_t, A_t) \leftarrow (1-\alpha)Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \max_a Q(S_{t+1}, a)], \tag{11}$$
$$a \in A(s)$$