Reinforcement Learning Exercise

Exercises 05

1 Temporal-difference methods

(a)

1. Bruteforce: iterative value iteration → guarantees to converge → gives out desired values
2. True value of each state: probability of terminating on the right if starting from that state
   We know $x_C = 0.5$ and can formulate the others as folllowed:

$$\begin{aligned} x_D &= p(\text{go to } C \text{ from } D) \times x_C + p(\text{go to } E \text{ from } E) \times x_E \\ &= \frac{x_C + x_E}{2} \\ x_A &= \frac{x_{\text{LeftBox}} + x_B}{2} \\ x_B &= \frac{x_A + x_C}{2} \\ x_E &= \frac{x_{\text{RightBox}} + x_D}{2} \end{aligned}$$

   Solving the above linear equations gives the desired results.

$\Rightarrow$ Method 2 is much easier → I assume method 2 was used

(b)

The action selection is taken with respect to an ϵ-greedy algorithm, while the action value function update is determined using a different policy. So with Q-learning we improve a policy which is differnet from that used to generate the data. Hence it is a off-policy method.