1 Adaptations of Q-learning

Below is the classic Q-learning algorithm [3], taken from [2, Sec 6.5]:

Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(\textit{terminal}, \cdot) = 0$

Loop for each episode:
  Initialize S
  Loop for each step of episode:
    Choose A from S using policy derived from Q (e.g., ε-greedy)
    Take action A, observe R, S'
    $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma \max_a Q(S', a) - Q(S, A)\big]$
    $S \leftarrow S'$
  until S is terminal

Q1. Consider a 5x5 grid world (i.e. $|\mathcal{S}^+| = 25$ states) where an agent has actions { Left, Right, Up, Down, Stay } which move the agent to adjacent cells as expected or remain still. The agent starts on the bottom-left square, gets a constant reward $R = -0.0001$ every step regardless of the action taken or state reached, except $R = +100$ when reaching the top-right (terminal) cell, ending the episode. How does a Q-learning agent learn to act in this problem?

Q2. Now consider the classic game of Tic-Tac-Toe. What are the states, actions, and rewards? How can Q-learning be adapted to play and/or solve Tic-Tac-Toe? Hint: there are two distinct interpretations. One of them makes explicit use of these identities: $R_1 = -R_2$, and $Q_1(S, A) = -Q_2(S, A)$.

2 Counterfactual Regret Minimization

Counterfactual regret (CFR) minimization has been an important algorithm in Poker AI research for finding approximate Nash equilibria in two-player zero-sum games [4].

Players start with uniform random initial policies $\pi = (\pi_1, \pi_2)$, and empty tables $R(s, a)$ and $S(s, a)$ and every iteration proceeds with three steps (notation glossary below):

Evaluate π : compute counterfactual values $q^c_{\pi,i}(s, a)$ and $v_{\pi,i}(s)$ for all states s, and actions $a \in A(s)$, and accumulate immediate regret $r(s, a) = q^c_{\pi,i}(s, a) - v_{\pi,i}(s)$ for all states and actions

Update tables : For all s, $a \in A(s)$: updates the accumulated regret table $R(s, a) = R(s, a) + r(s, a)$, and average strategy tables $S(s, a) = S(s, a) + \eta^\pi_{\tau(s)}(s)\pi(s, a)$

Update policy : For all s, $a \in A(s)$: update the policy (using **regret matching** [1]), define $x^+ = \max(x, 0)$:

$$\pi(s, a) = \begin{cases} \frac{R^+(s,a)}{\sum_{a \in A(s)} R^+(s,a)} & \text{if denominator is positive;} \\ \frac{1}{|A(s)|} & \text{otherwise.} \end{cases}$$

The average policy, $\bar{\pi}(s,a) = \frac{S(s,a)}{\sum_{a \in A(s)} S(s,a)}$, converges to an approximate Nash equilibrium in two-player zero-sum games.

Notation glossary:

- s is an information state
- $A(s)$ is the set of legal actions at s
- $\tau(s)$ is the player to play at s
- $\pi(s)$ is the policy at state s (probability distribution over $A(s)$)
- $\pi(s,a)$ is the probability of taking action a at info. state s
- $h \in s$ is a legal history in state s
- z is a terminal history (final state)
- η is a reach probability. Specifically:
  - $\eta^\pi(h)$ is the probability of reaching history h given players are playing with π
  - $\eta_i^\pi(h)$ is only player i's contribution to the reach probability
  - $\eta_{-i}^\pi(h)$ is all other players' (*except* i) contribution to the reach probability
  - $\eta_i^\pi(s)$, for $i = \tau(s)$, is a shorthand for $\eta_i^\pi(h)$ for any $h \in s$, since they are all the same due to perfect recall)
  - $\eta^\pi(h,z)$ is the reach probability of playing from history h to z
- $u_i(z)$ is the utility to player i of terminal history z
- $Z(s,a)$ is the set of histories $h \in s$ paired with the terminal histories reachable by any history in s and after having taken action a
- $q_{\pi,i}^c(s,a)$, where $i = \tau(s)$, is defined to be:

$$q_{\pi,i}^c(s,a) = \sum_{h,z \in Z(s,a)} \eta_{-i}^\pi(h)\eta^\pi(h,z)u_i(z)$$

- $v_{\pi,i}^c(s) = \sum_{a \in A(s)} \pi(s,a) q_{\pi,i}^c(s,a)$

References

[1] S. Hart and A. Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

[2] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction.* MIT Press, 2nd edition, 2018.

[3] Christopher J.C.H. Watkins and Peter Dayan. Q-learning. *Maching Learning*, 8:279–292, 1992.

[4] M. Zinkevich, M. Johanson, M. Bowling, and C. Piccione. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, 2008.