Q-Learning

Thuong Dang

October 2025

1 Vocabularies

Let us consider an example: the gridworld game. Assume that we have a grid with empty cells, walls, the starting point and the target point. Our goal is to reach the target point from the starting point with the following rules:

1. Valid moves include: left , right, up, down
2. The penalty for an empty cell is -1
3. The penalty for a wall is -5
4. The reward for the target point is $+10$.

And our task is to maximize the reward on the grid. To formulate this into a mathematical problem, we introduce the following terminology. We denote S the *set of states*, A the *set of actions*, and $\pi : S \to A$ a *policy*. And $r : S \times A \to \mathbb{R}$ the *rewards* if we choose an action $a \in A$ at state $s \in S$. Note that we want r to be *bounded*.

For example, in gridworld game, S is the set of all grid points. We are currently at $s = (1, 1)$, and it is our current state. Our valid moves at this state include $\{\text{left}, \text{right}, \text{up}, \text{down}\}$. And they are possible actions at this state. And if we choose to go right, then $\pi(s) = \text{right}$. If this action leads us to the wall, the reward is -5, if empty cell, the reward is -1, if it leads us to the goal, then the reward is $+10$.

2 Q-learning

We want to maximize the future reward at any time t. This is a sum $r_t + r_{t+1} + ...$, but it can lead to divergent series, so we to introduce a positive constant $0 < \gamma < 1$ and a series

$$r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + ...$$

Because r is bounded, this (absolute) series is bounded by a geometric series, and hence, it is convergent. Another reason to introduce γ is the longer the time is, the less immediate reward we can get.

For a policy π, the sum of rewards above is indeed a stochastic process, we would put this sum inside an expectation of a probability distribution, and define:

$$Q^{\pi}(s,a) = E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + ...|s_t = s, a_t = a]$$

The *Bellman operator* is defined to be

$$(\mathcal{T}Q)(s,a) = E[r_t + \gamma \max_{a'} Q(s',a')|s_t = s, a_t = a]$$

The Bellman operator is a contraction mapping (See the last section for a proof), and the Banach fixed point theorem tells us that there exists a *unique* fixed point Q^* of $\mathcal{T}$. And it means $\mathcal{T}Q^* = Q^*$. It means, Q^* satisfies

$$Q^*(s,a) = E[r_t + \gamma r_{t+1} + ...] = E[r_t + \gamma \max_{a'} Q^*(s',a')]$$

And this tells us that, the policy π^* corresponds to Q^* is the *optimal* policy. Why is that? For any policy π, we have

$$\max_{a'} Q^*(s',a') \geq Q^*(s',\pi(s'))$$

And this implies

$$E[r_t + \gamma \max_{a'} Q^*(s',a')] \geq E[r_t + \gamma Q^*(s',\pi(s'))]$$

Applying this again, we have

$$Q^*(s',\pi(s')) = E[r_{t+1} + \gamma \max_{a''} Q^*(s'',a'')|s_{t+1} = s', a_{t+1} = \pi(s')] \geq$$

$$\geq E[r_{t+1} + \gamma Q^*(s'',\pi(s''))|s_{t+1} = s', a_{t+1} = \pi(s')]$$

Substituting back, we have

$$Q^*(s,a) \geq E[r_t + \gamma r_{t+1} + \gamma^2 Q^*(s'',\pi(s''))]$$

Continuing this process, by induction, we have

$$Q^*(s,a) \geq Q^{\pi}(s,a)$$

And it means, Q^* will give the optimal reward, and the policy π^* corresponds to Q^* is the optimal policy.

3 Learning

For the learning process, we want to approximate Q^* with the Bellman equation by the Banach's fixed point theorem (See the last section):

$$Q^*(s,a) = E[r + \gamma \max_{a'} Q(s',a')]$$

At any time t, our prediction is an action a, and the target should be $(r + \gamma \max_{a'} Q(s', a'))$, and it is treated as a constant where r is the reward at s with a chosen action a, the loss is

$$L = [r + \gamma \max_{a'} Q(s', a') - Q(s, a)]^2$$

Taking the derivative in terms of $Q(s, a)$, we get

$$\frac{\partial L}{\partial Q(s, a)} = -2(r + \gamma \max_{a'} Q(s', a') - Q(s, a))$$

And we can apply stochastic gradient descent to introduce a learning rate α and update:

$$Q(s, a) = (1 - \alpha)Q(s, a) + \alpha(r + \gamma \max_{a'} Q(s', a'))$$

After enough iterations, we can obtain the optimal policy π^* by defining

$$\pi^*(s) = \text{argmax}_{a \in A} Q(s, a)(\forall s \in S)$$

4 The Bellman optimality operator is a contraction

In this section, we will prove that the Bellman operator $\mathcal{T}$ is a contraction. For any bounded function $Q : S \times A \to \mathbb{R}$, recall the Bellman optimality operator

$$(\mathcal{T}Q)(s, a) \;:=\; \mathbb{E}\Big[\, r(s, a) + \gamma \max_{a' \in A} Q(s', a') \;\Big|\; s_t = s,\, a_t = a \Big]$$

We equip the space of bounded functions in $S \times A$ with the sup norm $\|Q\|_\infty := \sup_{(s,a) \in S \times A} |Q(s, a)|$ and denote $\mathcal{B}(S \times A)$ the space of these functions. Note that $(\mathcal{B}(S \times A), \|\cdot\|_\infty)$ is a *complete metric space*. And if we can prove $\mathcal{T}$ is a contraction mapping, it follows from the Banach fixed point theorem that there exists a unique fixed point of $\mathcal{T}$. Moreover, starting from any initial bounded function $Q_0 \in \mathcal{B}(S \times A)$, the iterations $Q_{k+1} = \mathcal{T}Q_k$ converge to Q^*. And the foundation for our Sections 2 and 3 is reduced to the following

Theorem 1 (Contraction of $\mathcal{T}$)**.** *For any bounded $Q_1, Q_2 : S \times A \to \mathbb{R}$,*

$$\|\mathcal{T}Q_1 - \mathcal{T}Q_2\|_\infty \;\le\; \gamma \, \|Q_1 - Q_2\|_\infty.$$

In particular, $\mathcal{T}$ is a γ-contraction in $(\mathcal{B}(S \times A), \|\cdot\|_\infty)$.

Proof. Fix $(s, a) \in S \times A$. Using the definition of $\mathcal{T}$ and linearity of expectation,

$$(\mathcal{T}Q_1)(s, a) - (\mathcal{T}Q_2)(s, a) = \gamma \, \mathbb{E}\Big[\max_{a'} Q_1(s', a') - \max_{a'} Q_2(s', a') \;\Big|\; s, a \Big].$$

Take absolute values and apply the inequality

$$\left|\max_{a'} x_{a'} - \max_{a'} y_{a'}\right| \;\le\; \max_{a'} |x_{a'} - y_{a'}| \quad \text{(for any real families } (x_{a'})_{a'}, (y_{a'})_{a'}),$$

to obtain

$$\left|(\mathcal{T}Q_1)(s,a) - (\mathcal{T}Q_2)(s,a)\right| \;\le\; \gamma\,\mathbb{E}\Big[\max_{a'} |Q_1(s',a') - Q_2(s',a')| \,\Big|\, s,a\Big].$$

Since $\max_{a'} |Q_1(s',a') - Q_2(s',a')| \le \|Q_1 - Q_2\|_\infty$ for every s', the expectation preserves the bound:

$$\left|(\mathcal{T}Q_1)(s,a) - (\mathcal{T}Q_2)(s,a)\right| \;\le\; \gamma\,\|Q_1 - Q_2\|_\infty.$$

Taking the supremum over (s,a) yields the desired contraction:

$$\|\mathcal{T}Q_1 - \mathcal{T}Q_2\|_\infty \le \gamma \|Q_1 - Q_2\|_\infty. \qquad \square$$