Convergence of Q-learning in Tabular Cases

Boyang Liu

December 8, 2019

Abstract

The basic Q-learning concept is assumed to be known for readers. This is a self-contained proof (except Q-learning) of Q-learning convergence in tabular case.

keywords: Cauchy sequence, Banach Fixed Point Theorem, Bellman Update

1 Cauchy Sequence

We first introduce the definition of Cauchy sequence.

Definition 1.1 (Cauchy Sequence)**.** Given a metric space $(\mathcal{X}, d)$, We say that a sequence of real numbers x_n is a *Cauchy sequence* provided that for every $\epsilon > 0$, there is a natural number N so that when $n, m \geq N$, we have that $|x_n - x_m| \leq \epsilon$

Theorem 1.1. For real number sequence, a sequence converges if and only if the sequence is Cauchy.

We omit the proof since it is easy and intuitive. A proof of above theorem can be found in here.

A metric space $(\mathcal{X}, d)$ is called complete if every cauchy sequence converges to an element of $\mathcal{X}$.

An counter example would be rational numbers. The rational numbers $\mathcal{Q}$ are not complete since the sequence of rational numbers could converge to irrational number.

In Q-learning, we are dealing with real number, and real number is a complete metric space (proof omitted). Thus, we can full utilize the properties of complete metric space.

2 Banach Fixed Point Theorem

Banach fixed point theorem is also known as *contraction mapping theorem* or *contractive mapping theorem*. It plays the central role in Q learning. We will first introduce the theorem and the definition of contraction mapping.

Definition 2.1 (Contraction Mapping)**.** Let $(\mathcal{X}, d)$ to be a metric space, a function $f : \mathcal{X} \to \mathcal{X}$ is called a contraction if there exists $\lambda \in (0, 1)$ such that

$$d(fx, fx^{'}) \leq \lambda d(x, x^{'}) \qquad \forall x, x^{'} \in \mathcal{X}$$

Clearly, the contraction function is continuous since the smallest λ is Lipschitz constant of f.

Theorem 2.1 (Banach Fixed Point Theorem)**.** Let $(\mathcal{X}, d)$ be a non-empty complete metric space with a contraction mapping $f : \mathcal{X} \to \mathcal{X}$. Then f admits a unique fixed point x^* in $\mathcal{X}$ (i.e. $f(x^*) = x^*$). Furthermore, x^* can be found as follows: start with an arbitrary element x_0 in $\mathcal{X}$, and define a sequence x_n by $x_n = f(x_{n-1})$, then $x_n \to x^*$

Proof. we first prove the uniqueness of the fixed point by contradiction.

If we have two fixed point $f(p) = p$ and $f(q) = q$, then

$$d(p, q) = d(fp, fq) \leq \lambda d(p, q),$$

which is a contradiction unless $d(p, q) = 0$. Hence if a fixed point exists, it is unique. Thus it remains to proof the existence.

Note we have $d(f^2x, f^2y) \leq \lambda d(fx, fy) \leq \lambda^2 d(x, y)$, by induction, we have $d(f^kx, f^ky) \leq \lambda^k d(x, y)$ for $k \geq 1$. Given $x \in \mathcal{X}$ set $a_n := f^n(x)$, We have for natural number $m > n$:

$$\begin{aligned}
&d(a_m, a_n) \leq d(a_m, a_{m-1}) + ... + d(a_{n+2}, a_{n+1} + d(a_{n+1}, a_n)) \\
&= d(f^m x, f^{m-1}x) + ... + d(f^{n+2}x, f^{n+1}x) + d(f^{n+1}x, f^n x) \\
&\leq \lambda^{m-1} d(fx, x) + ... + \lambda_{n+1} d(fx, x) + \lambda^n d(fx, x) \\
&= [\lambda^{m-1} + ... + \lambda^n] d(fx, x) \\
&= \lambda^n (\sum_{i=0}^{m-n-1} \lambda^i) d(fx, x) \\
&\leq \lambda^n (\sum_{i=0}^{\infty} \lambda^i) d(fx, x) \\
&= \lambda^n \frac{1}{1-\lambda} d(fx, x) \qquad \lambda \in (0, 1)
\end{aligned}$$

Since we have $\lambda < 1$ we could make the RHS to be arbitrary small by making n sufficiently large.

Obviously, the sequence $(a_n)_{n=0}^{\infty}$ is cauchy. Since $\mathcal{X}$ is complete this cauchy converges to $x^* \in \mathcal{X}$, that is

$$x^* = \lim_{n \to \infty} f^n(x)$$

By the property of limit that we can have

$$x^* = \lim_{n \to \infty} f^n(x) = \lim_{n \to \infty} f^{n+1}(x) = f(x^*)$$

So x^* is a fixed point.

□

3 Q-learning

We denote MDP as tuple $(\mathcal{X}, \mathcal{A}, P, r)$, where

- $\mathcal{X}$ is finite state space
- $\mathcal{A}$ is finite action space
- P is transition probabilities
- $r : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$ represents the reward function

We also have a constant $\gamma \in (0,1]$. By the bellman update, we have

$$Q^*(x,a) = \sum_y P_a(x,y)[r(x,a,y) + \gamma V^*(y)]$$

Based on that, we can define the Q-learning update.

Theorem 3.1 (Bellman Update)**.** The optimal Q-function is a fixed point of a contraction operator $\mathbf{H}$, defined for a generic function $q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ as

$$(\mathbf{H}q)(x,a) = \sum_y P_a(x,y)[r(x,a,y) + \gamma \max_{b \in \mathcal{A}} q(y,b)]$$

We first prove that the operator is a contraction in the sup-norm.

$$||\mathbf{H}q_1 - \mathbf{H}q_2||_\infty \leq \gamma ||q_1 - q_2||_\infty$$

Proof.

$$\begin{aligned}
&\|\mathbf{H}q_1 - \mathbf{H}q_2\|_\infty = \\
&= \max_{x,a} \left| \textstyle\sum_{y \in \mathcal{X}} \mathrm{P}_a(x,y) \left[r(x,a,y) + \gamma \max_{b \in \mathcal{A}} q_1(y,b) - r(x,a,y) + \gamma \max_{b \in \mathcal{A}} q_2(y,b) \right] \right| = \\
&= \max_{x,a} \gamma \left| \textstyle\sum_{y \in \mathcal{X}} \mathrm{P}_a(x,y) \left[\max_{b \in \mathcal{A}} q_1(y,b) - \max_{b \in \mathcal{A}} q_2(y,b) \right] \right| \leq \\
&= \max_{x,a} \gamma \textstyle\sum_{y \in \mathcal{X}} \mathrm{P}_a(x,y) \left| \max_{b \in \mathcal{A}} q_1(y,b) - \max_{b \in \mathcal{A}} q_2(y,b) \right| \leq \\
&= \max_{x,a} \gamma \textstyle\sum_{y \in \mathcal{X}} \mathrm{P}_a(x,y) \max_{z,b} \left| q_1(z,b) - q_2(z,b) \right| \\
&= \max_{x,a} \gamma \textstyle\sum_{y \in \mathcal{X}} \mathrm{P}_a(x,y) \left\| q_1 - q_2 \right\|_\infty \\
&= \gamma \left\| q_1 - q_2 \right\|_\infty
\end{aligned} \tag{1}$$

□

The above theorem states that if we knew the transition probability, we can directly use the update rule to find he optimal Q-function.

If we do not know the transition probability, q-learning has the following update rule

$$Q_{t+1}(x_t, a_t) = Q_t(x_t, a_t) + \alpha_t(x_t, a_t) \left[r_t + \gamma \max_{b \in \mathcal{A}} Q_t(x_{t+1}, b) - Q_t(x_t, a_t) \right] \tag{2}$$

Now, we formally give the Q-learning convergence theorem.

Theorem 3.2. Given a finite MDP $(\mathcal{X}, \mathcal{A}, P, r)$, the Q-learning algorithm. given by the update rule

$$Q_{t+1}(x_t, a_t) = Q_t(x_t, a_t) + \alpha_t(x_t, a_t) \left[r_t + \gamma \max_{b \in \mathcal{A}} Q_t(x_{t+1}, b) - Q_t(x_t, a_t) \right] \tag{3}$$

$$(0 \leq \alpha_t(x,a) < 1) \tag{4}$$

converges w.p.1 to the optimal Q-function as long as

$$\sum_t \alpha_t(x,a) = \infty \quad \sum_t \alpha_t^2(x,a) < \infty \tag{5}$$

for all $(x,a) \in \mathcal{X} \times \mathcal{A}$.

We will apply the following theorem to help us to prove the Q-learning convergence.

Theorem 3.3. The random process $\{\Delta_t\}$ taking values in $\mathbb{R}^n$ and defined as

$$\Delta_{t+1}(x) = (1 - \alpha_t(x))\,\Delta_t(x) + \alpha_t(x)F_t(x)$$

converges to zero w.p. 1 under the following assumptions:

- $0 \leq \alpha_t \leq 1, \sum_t \alpha_t(x) = \infty$ and $\sum_t \alpha_t^2(x) < \infty$
- $\|\mathbb{E}\left[F_t(x)|\mathcal{F}_t\right]\|_W \leq \gamma \|\Delta_t\|_W$, with $\gamma < 1$
- $\mathrm{var}\left[F_t(x)|\mathcal{F}_t\right] \leq C\left(1 + \|\Delta_t\|_W^2\right)$, for $C > 0$

Now we are in good shape to proof the Q-learning convergence.

Proof.

$$Q_{t+1}\left(x_t, a_t\right) = \left(1 - \alpha_t\left(x_t, a_t\right)\right) Q_t\left(x_t, a_t\right) + \alpha_t\left(x_t, a_t\right)\left[r_t + \gamma \max_{b \in \mathcal{A}} Q_t\left(x_{t+1}, b\right)\right]$$

Subtracting from both sides the quantity $Q^*\left(x_t, a_t\right)$ and letting

$$\Delta_t(x, a) = Q_t(x, a) - Q^*(x, a)$$

yields

$$\Delta_{t+1}(x_t, a_t) = \left(1 - \alpha_t\left(x_t, a_t\right)\right) \Delta_t(x_t, a_t) + a_t(x_t, a_t)\left[r_t + \gamma \max_{b \in \mathcal{A}} Q_t\left(x_{t+1}, b\right) - Q^*\left(x_t, a_t\right)\right]$$

If we write

$$F_t(x, a) = r(x, a, X(x, a)) + \gamma \max_{b \in \mathcal{A}} Q_t(y, b) - Q^*(x, a)$$

where $X(x, a)$ is a random sample state from the Markov chain $(\mathcal{X}, P_a)$, we have

$$\begin{aligned}\mathbb{E}\left[F_t(x, a)|\mathcal{F}_t\right] &= \sum_{y \in \mathcal{X}} \mathrm{P}_a(x, y)\left[r(x, a, y) + \gamma \max_{b \in \mathcal{A}} Q_t(y, b) - Q^*(x, a)\right] = \\ &= \left(\mathbf{H}Q_t\right)(x, a) - Q^*(x, a)\end{aligned}$$

Using the fact that $Q^* = \mathbf{H}Q^*$,

$$\mathbb{E}\left[F_t(x, a)|\mathcal{F}_t\right] = \left(\mathbf{H}Q_t\right)(x, a) - \left(\mathbf{H}Q^*\right)(x, a) \tag{6}$$

By applying the contraction mapping, we have

$$\|\mathbb{E}\left[F_t(x, a)|\mathcal{F}_t\right]\|_\infty \leq \gamma \|Q_t - Q^*\|_\infty = \gamma \|\Delta_t\|_\infty \tag{7}$$

Now we proof the bounded variance:

$$\begin{aligned}&\mathrm{var}\left[F_t(x)|\mathcal{F}_t\right] = \\ &= \mathbb{E}\left[\left(r(x, a, X(x, a)) + \gamma \max_{b \in \mathcal{A}} Q_t(y, b) - Q^*(x, a) - \left(\mathbf{H}Q_t\right)(x, a) + Q^*(x, a)\right)^2\right] = \\ &= \mathbb{E}\left[\left(r(x, a, X(x, a)) + \gamma \max_{b \in \mathcal{A}} Q_t(y, b) - \left(\mathbf{H}Q_t\right)(x, a)\right)^2\right] = \\ &= \mathrm{var}\left[r(x, a, X(x, a)) + \gamma \max_{b \in \mathcal{A}} Q_t(y, b)|\mathcal{F}_t\right]\end{aligned} \tag{8}$$

Since r is bounded, clearly verifies that

$$\mathrm{var}\left[F_t(x)|\mathcal{F}_t\right] \leq C\left(1 + \|\Delta_t\|_W^2\right) \tag{9}$$

for some constance C. □

References

[1] Jaakkola, Tommi, Michael I. Jordan, and Satinder P. Singh. "Convergence of stochastic iterative dynamic programming algorithms." Advances in neural information processing systems. 1994.

[2] Daniel Murfet, course note MAST30026: Metric and Hilbert spaces.

[3] Melo, F. S. (2001). Convergence of Q-learning: A simple proof. Institute Of Systems and Robotics, Tech. Rep, 1-4..