Convergence of Q-learning: a simple proof

Francisco S. Melo
Institute for Systems and Robotics,
Instituto Superior Técnico,
Lisboa, PORTUGAL
fmelo@isr.ist.utl.pt

1 Preliminaries

We denote a Markov decision process as a tuple $(\mathcal{X}, \mathcal{A}, \mathsf{P}, r)$, where

- $\mathcal{X}$ is the (finite) state-space;
- $\mathcal{A}$ is the (finite) action-space;
- P represents the transition probabilities;
- r represents the reward function.

We denote elements of $\mathcal{X}$ as x and y and elements of $\mathcal{A}$ as a and b. We admit the general situation where the reward is defined over triplets (x, a, y), *i.e.*, r is a function

$$r : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \longrightarrow \mathbb{R}$$

assigning a reward $r(x, a, y)$ everytime a transition from x to y occurs due to action a. We admit r to be a bounded, deterministic function.

The value of a state x is defined, for a sequence of controls $\{A_t\}$, as

$$J(x, \{A_t\}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(X_t, A_t) \mid X_0 = x\right].$$

The optimal value function is defined, for each $x \in \mathcal{X}$ as

$$V^*(x) = \max_{\mathcal{A}_t} J(x, \{A_t\})$$

and verifies

$$V^*(x) = \max_{a \in \mathcal{A}} \sum_{y \in \mathcal{X}} \mathsf{P}_a(x, y)\big[r(x, a, y) + \gamma V^*(y)\big].$$

From here we define the optimal Q-function, Q^* as

$$Q^*(x, a) = \sum_{y \in \mathcal{X}} \mathsf{P}_a(x, y)\big[r(x, a, y) + \gamma V^*(y)\big].$$

The optimal Q-function is a fixed point of a contraction operator $\mathbf{H}$, defined for a generic function $q : \mathcal{X} \times \mathcal{A} \longrightarrow \mathbb{R}$ as

$$(\mathbf{H}q)(x,a) = \sum_{y \in \mathcal{X}} \mathsf{P}_a(x,y)\big[r(x,a,y) + \gamma \max_{b \in \mathcal{A}} q(y,b)\big].$$

This operator is a contraction in the sup-norm, *i.e.*,

$$\left\|\mathbf{H}q_1 - \mathbf{H}q_2\right\|_\infty \leq \gamma \left\|q_1 - q_2\right\|_\infty . \tag{1}$$

To see this, we write

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

The Q-learning algorithm determines the optimal Q-function using point samples. Let π be some random policy such that

$$\mathbb{P}_\pi \left[A_t = a \mid X_t = x\right] > 0$$

for all state-action pairs (x,a). Let $\{x_t\}$ be a sequence of states obtained following policy π, $\{a_t\}$ the sequence of corresponding actions and $\{r_t\}$ the sequence of obtained rewards. Then, given any initial estimate Q_0, Q-learning uses the following update rule:

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

where the step-sizes $\alpha_t(x,a)$ verify $0 \leq \alpha_t(x,a) \leq 1$. This means that, at the $(t+1)^{\text{th}}$ update, only the component (x_t,a_t) is updated.[1]

This leads to the following result.

[1] There are variations of Q-learning that use a single transition tuple (x,a,y,r) to perform updates in multiple states to speed up convergence, as seen for example in [2].

Theorem 1. *Given a finite MDP* $(\mathcal{X}, \mathcal{A}, \mathsf{P}, r)$*, the Q-learning algorithm, given by the update rule*

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

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

$$\sum_t \alpha_t(x, a) = \infty \qquad \sum_t \alpha_t^2(x, a) < \infty \quad (3)$$

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

Notice that, since $0 \leq \alpha_t(x, a) < 1$, (3) requires that all state-action pairs be visited infinitely often.

To establish Theorem 1 we need an auxiliary result from stochastic approximation, that we promptly present.

Theorem 2. *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* $\sum_t \alpha_t^2(x) < \infty$*;*
- $\|\mathbb{E}[F_t(x) \mid \mathcal{F}_t]\|_W \leq \gamma \|\Delta_t\|_W$*, with* $\gamma < 1$*;*
- $\mathbf{var}[F_t(x) \mid \mathcal{F}_t] \leq C(1 + \|\Delta_t\|_W^2)$*, for* $C > 0$*.*

Proof See [1]. □

We are now in position to prove Theorem 1.

Proof of Theorem 1 We start by rewriting (2) as

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

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

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

yields

$$\begin{aligned}\Delta_t(x_t, a_t) = (1 - \alpha_t(x_t, a_t))\Delta_t(x_t, a_t))+ \\ + \alpha_t(x, a)\big[r_t + \gamma \max_{b \in \mathcal{A}} Q_t(x_{t+1}, b) - Q^*(x_t, a_t)\big].\end{aligned}$$

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 obtained from the Markov chain $(\mathcal{X}, \mathsf{P}_a)$, we have

$$\mathbb{E}\left[F_t(x,a) \mid \mathcal{F}_t\right] = \sum_{y \in \mathcal{X}} \mathsf{P}_a(x,y)\big[r(x,a,y) + \gamma \max_{b \in \mathcal{A}} Q_t(y,b) - Q^*(x,a)\big] = \\ = (\mathbf{H}Q_t)(x,a) - Q^*(x,a).$$

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

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

It is now immediate from (1) that

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

Finally,

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

which, due to the fact that r is bounded, clearly verifies

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

for some constant C.

Then, by Theorem 2, Δ_t converges to zero w.p.1, *i.e.*, Q_t converges to Q^* w.p.1. □

References

[1] Tommi Jaakkola, Michael I. Jordan, and Satinder P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6(6):1185–1201, 1994.

[2] Carlos Ribeiro and Csaba Szepesvári. Q-learning combined with spreading: Convergence and results. In *Proceedings of the ISRF-IEE International Conference: Intelligent and Cognitive Systems (Neural Networks Symposium)*, pages 32–36, 1996.