A. Preliminaries.

1. Contradiction Mapping.

$(\mathbb{X}, d)$ is a metric space. $T: \mathbb{X} \to \mathbb{X}$ is a mapping. If $\exists$ $L>0$ s.t. $\forall x, y \in \mathbb{X}$, $d(Tx, Ty) \leq L \cdot d(x, y)$, T is L-Lipschitz in $\mathbb{X}$ w.r.t. d. Specifically, if $L<1$, then T is a contraction mapping.

2. Fixed point.

$(\mathbb{X}, d)$ is a metric space. $T: \mathbb{X} \to \mathbb{X}$ is a mapping. If $Tx = x$, then x is a fixed point of T.

3. Contradiction Mapping Theorem.

① $\mathbb{X}$ is complete.

② T contradiction mapping.

Then there exists and exists only one fixed point of T.

4. The optimal Bellman operator (OBO) is a contradiction mapping w.r.t. maximum norm.

Proof.

Recall the OBO H satisfies

$$H q(s,a) = \sum_{s' \sim \mathbb{P}(\cdot \mid s,a)} \mathbb{P}(s,a,s') \cdot \left[R(s,a,s') + \gamma \cdot \max_{a' \in \mathcal{A}} q(s',a') \right],$$

for some Q-function q.

For any two Q-functions q_1 and q_2,

$$\|Hq_1 - Hq_2\|_\infty$$

$$= \max_{(s,a)\in S\times A} \left| \sum_{s'\sim P(\cdot|s,a)} P(s'|s,a) \left[\gamma \cdot \max_{a'\in A} q_1(s',a') - \gamma \cdot \max_{a'\in A} q_2(s',a') \right] \right|$$

$$\le \max_{(s,a)\in S\times A} \gamma \sum_{s'\sim P(\cdot|s,a)} P(s'|s,a) \left| \max_{a'\in A} q_1(s',a') - \max_{a'\in A} q_2(s',a') \right|. \quad (1)$$

$$\le \max_{(s,a)\in S\times A} \gamma \cdot \sum_{s'\sim P(\cdot|s,a)} P(s'|s,a) \left| \max_{a'\in A} q_1(s',a') - q_2(s',a') \right|$$

$$\le \max_{(s,a)\in S\times A} \gamma \cdot \sum_{s'\sim P(\cdot|s,a)} P(s'|s,a) \cdot \left| \max_{(s'',a')\in S\times A} q_1(s'',a') - q_2(s'',a') \right|$$

$$= \gamma \cdot \|q_1 - q_2\|_\infty,$$

where Eq (1) follows from Jensen's inequality. ∎

5. Q-Learning

$$Q^\pi_{t+1}(s_t,a_t) = Q^\pi_t(s_t,a_t) - \alpha \cdot (Q^\pi_t(s_t,a_t) - \text{sample}_t),$$

$$= (1-\alpha) \cdot Q^\pi_t(s_t,a_t) + \alpha \cdot \text{sample}_t,$$

where $\text{sample}_t = R_t(s_t,a_t,s_{t+1}) + \gamma \cdot \max_{a'\in A} Q(s_{t+1},a')$, and $\{s_t\}$ is a sequence of states obtained by following policy π which satisfies

$$P_\pi(A_t = a \mid S_t = s) > 0$$

for any state action pairs $(s,a) \in S \times A$.

B. Content.

Theorem 1. Convergence of Q-Learning

Consider a finite MDP. (S, A, P, R, γ), where S and A are finite.

and R is bounded and deterministic. If we use Q-learning algorithm to solve it with learning rate α_t, which satisfies

① $0 \le \alpha_t \le 1$

② $\sum_t \alpha_t = +\infty$.

③ $\sum_t \alpha_t^2 < +\infty$.

Then. Q_{t+1} converges to Q^* w.p. 1.

Proof of Theorem 1:

Theorem 2.

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

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

If the following assumptions hold:

① $0 \le \alpha_t \le 1$, $\sum_t \alpha_t = +\infty$, $\sum_t \alpha_t^2 < +\infty$.

② $\| \mathbb{E}[F_t(x) \mid \mathcal{F}_t] \|_\infty \le \gamma \cdot \|\Delta_t\|_\infty$, with $\gamma < 1$.

③ $\mathbb{V}[F_t(x) \mid \mathcal{F}_t] \le C \cdot (1 + \|\Delta_t\|_\infty)^2$, for $C > 0$,

where $\mathcal{F}_t = \{\Delta_t, \Delta_{t-1}, \dots, \Delta_1, F_{t-1}, F_{t-2}, \dots, F_1\}$ stands for the past at time t. Then $\{\Delta_t\}$ converges to 0 w.p. 1.

Recall the Q-learning update:

$$Q_{t+1}(S_t, a_t) = (1-\alpha_t)\cdot Q_t(S_t, a_t) + \alpha_t \cdot \left(R(S_t, a_t, S') + \gamma \cdot \max_{a' \in \mathcal{A}} Q_t(S', a')\right),$$

where $S' \sim \mathbb{P}(\cdot \mid S_t, a_t)$.

Substracting from both sides the quantity $Q^*(S_t, a_t)$ shows that

$$Q_{t+1}(S_t, a_t) - Q^*(S_t, a_t) = (1-\alpha_t)\cdot\left[Q_t(S_t, a_t) - Q^*(S_t, a_t)\right]$$
$$+ \alpha_t \left(R(S_t, a_t, S') + \gamma \cdot \max_{a' \in \mathcal{A}} Q_t(S', a') - Q^*(S_t, a_t)\right) \quad (2)$$

Let

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

and

$$F_t(s,a) = R(s,a,s') + \gamma \cdot \max_{a' \in \mathcal{A}} Q_t(s',a') - Q^*(s,a),$$

with $s' \sim \mathbb{P}(\cdot|s,a)$. Then Eq.(2) could be reformulated as

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

a. Assumption (1) in Theorem 2 is satisfied.

b. $\mathbb{E}[F_t(s_t,a_t)|\mathcal{F}_t]$

$$= \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_t,a_t)}\left[R(s_t,a_t,s') + \gamma \max_{a' \in \mathcal{A}} Q_t(s',a') - Q^*(s_t,a_t)\right].$$

$$= (HQ_t)(s_t,a_t) - Q^*(s_t,a_t).$$

$$= (HQ_t)(s_t,a_t) - (HQ^*)(s_t,a_t), \quad (3)$$

where Eq(3) follows from that Q^* is the fixed point of H.

Then

$$\|\mathbb{E}[F_t|\mathcal{F}_t]\|_\infty = \|HQ_t - HQ^*\|_\infty \le \gamma \|Q_t - Q^*\|_\infty = \gamma \|\Delta_t\|_\infty.$$

Assumption (2) in Theorem 2 is satisfied.

c. $\mathbb{V}_{s' \sim \mathbb{P}(\cdot|s_t,a_t)}[F_t(s_t,a_t)|\mathcal{F}_t]$

$$= \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_t,a_t)}\left[\left(F_t(s_t,a_t) - \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_t,a_t)}[F_t(s_t,a_t)|\mathcal{F}_t]\right)^2 \middle| \mathcal{F}_t\right]$$

$$= \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_t,a_t)}\Big[\Big[R(s_t,a_t,s') + \gamma \cdot \max_{a' \in \mathcal{A}} Q_t(s',a') - Q^*(s_t,a_t)$$

$$- \sum_{s' \sim \mathbb{P}(\cdot|s_t,a_t)} \mathbb{P}(s_t,a_t,s')\Big[R(s_t,a_t,s') + \gamma \max_{a' \in \mathcal{A}} Q_t(s,a') - Q^*(s_t,a_t)\Big]\Big]^2 \Big| \mathcal{F}_t\Big].$$

$$= \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_t,a_t)} \Big[\Big[R(s_t,a_t,s') + \gamma \cdot \max_{a' \in \mathcal{A}} Q_t(s',a') - Q^*(s_t,a_t) - (\mathcal{T}Q_t)(s_t,a_t) + Q^*(s_t,a_t) \Big]^2 \mid \mathcal{F}_t \Big].$$

$$= \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_t,a_t)} \Big[\Big[R(s_t,a_t,s') + \gamma \cdot \max_{a' \in \mathcal{A}} Q_t(s',a') - (\mathcal{T}Q_t)(s_t,a_t) \mid \mathcal{F}_t \Big] \Big]$$

$$= \mathbb{V}_{s' \sim \mathbb{P}(\cdot|s_t,a_t)} \Big[R(s_t,a_t,s') + \gamma \cdot \max_{a' \in \mathcal{A}} Q_t(s',a') \mid \mathcal{F}_t \Big].$$

$$= \mathbb{V}_{s' \sim \mathbb{P}(\cdot|s_t,a_t)} \cdot \Big[R(s_t,a_t,s') + \gamma \cdot Q^*\big(s', \operatorname*{argmax}_{a' \in \mathcal{A}} Q_t(s',a')\big) + \Big(\gamma \cdot \max_{a' \in \mathcal{A}} Q_t(s',a') - \gamma Q^*\big(s', \operatorname*{argmax}_{a' \in \mathcal{A}} Q_t(s',a')\big) \Big) \Big].$$

($X_{t,1}$: the first term; $X_{t,2}$: the second term.)

Let random variable $X_{t,1}(s') = R(s_t,a_t,s') + \gamma \cdot Q^*\big(s', \operatorname*{argmax}_{a' \in \mathcal{A}} Q_t(s',a')\big)$

$$X_{t,2}(s') = \Big(\gamma \cdot \max_{a' \in \mathcal{A}} Q_t(s',a') \Big) - \gamma Q^*\big(s', \operatorname*{argmax}_{a' \in \mathcal{A}} Q_t(s',a')\big)$$

Then $\sup_{s' \in S} |X_{t,2}(s')| = \sup_{s' \in S} \Big| \Big(\gamma \cdot \max_{a' \in \mathcal{A}} Q_t(s',a') \Big) - \gamma Q^*\big(s', \operatorname*{argmax}_{a' \in \mathcal{A}} Q_t(s',a')\big) \Big|$

$$\leq \sup_{s' \in S} \gamma \Big| \max_{a' \in \mathcal{A}} Q_t(s',a') - Q^*(s',a') \Big|$$

$$\leq \sup_{s' \in S} \gamma \Big| \max_{(s'',a') \in S \times \mathcal{A}} Q_t(s'',a') - Q^*(s'',a') \Big|$$

$$= \gamma \, \| \Delta_t \|_\infty.$$

It follows that

$$\mathbb{V}_{s' \sim \mathbb{P}(\cdot \mid S_t, C_t)} \left[F_t(S_t, C_t) \mid \mathcal{F}_t \right]$$

$$= \mathbb{V}_{s' \sim \mathbb{P}(\cdot \mid S_t, C_t)} \left[X_{t,1} + X_{t,2} \mid \mathcal{F}_t \right]$$

$$\leq \sup_{s' \in S} \left| X_{t,1}(s') + X_{t,2}(s') \right|^2$$

$$\leq \sup_{s' \in S} \left(|X_{t,1}(s')| + |X_{t,2}(s')| \right)^2$$

$$\leq \left(C' + \gamma^2 \| \Delta_t \|_\infty \right)^2$$

$$= C \left(1 + \| \Delta_t \|_\infty \right)^2,$$

for some constants C' and C. Assumption ③ in Theorem 2 is satisfied and applying Theorem 2 immediately shows that $\lim_{t \to +\infty} \Delta_t = 0$, which further implies that $\lim_{t \to +\infty} Q_t = Q^*$ and concludes the proof.

□