Convergence of Q-Learning via Stochastic Approximation

zhuobie

1 Stochastic Approximation (SA) Convergence Framework

1.1 Model and Notation

Let $L : \mathbb{R}^d \to \mathbb{R}^d$ be an operator, and we wish to find its fixed point $\theta^\star$ such that

$$L\theta^\star = \theta^\star.$$

Consider the asynchronous stochastic approximation (SA) iteration, where at time t only coordinate i is updated:

$$\theta_{t+1}(i) \;\leftarrow\; (1-\alpha_t(i))\,\theta_t(i) \;+\; \alpha_t(i)\,\big[(L\theta_t)(i) + \eta_t(i)\big], \tag{1}$$

with stepsize $\alpha_t(i) \geq 0$ (and $\alpha_t(i) = 0$ if the i-th component is not updated). The random term $\eta_t(i)$ represents zero-mean noise. Let

$$\mathcal{F}_t = \{\theta_0, \theta_1, \ldots, \theta_t\} \cup \{\eta_0, \eta_1, \ldots, \eta_{t-1}\} \cup \{\alpha_0, \alpha_1, \ldots, \alpha_t\}$$

denote the history of the algorithm up to time t.

1.2 Noise Assumption (A1)

For every i, t, assume:

(a) $\mathbb{E}[\eta_t(i) \mid \mathcal{F}_t] = 0$;

(b) there exist constants $c_1, c_2 \geq 0$ such that

$$\mathbb{E}[\eta_t(i)^2 \mid \mathcal{F}_t] \leq c_1 + c_2\|\theta_t\|^2.$$

1.3 Step-Size and Operator Conditions

(1) (**Step sizes**) For every i,

$$\sum_{t=0}^{\infty} \alpha_t(i) = \infty, \qquad \sum_{t=0}^{\infty} \alpha_t(i)^2 < \infty.$$

(2) (**Contraction**) L is a contraction under $\|\cdot\|_\infty$, i.e.,

$$\|L\theta_1 - L\theta_2\|_\infty \leq c\|\theta_1 - \theta_2\|_\infty, \qquad c < 1,$$

and thus admits a unique fixed point $\theta^\star$.

1.4 SA Convergence Theorem (Bertsekas–Tsitsiklis, 1996)

Under iteration (1), if the step-size, noise, and contraction assumptions hold, then

$$\theta_t \xrightarrow[t\to\infty]{\text{a.s.}} \theta^\star.$$

Intuitively, the algorithm performs a noisy, asynchronous evaluation of the contraction operator L; the step-size schedule ensures the noise averages out while the contraction pulls all iterates toward the unique fixed point.

2 Q-Learning Convergence: Reduction to SA and Assumption Verification

2.1 Writing Q-Learning as a Stochastic Approximation

For a finite MDP, let $Q_t \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{A}|}$. The Q-learning update is

$$Q_{t+1}(X_t, A_t) \;\leftarrow\; (1-\alpha_t)\, Q_t(X_t, A_t) \;+\; \alpha_t\Big[R_t + \gamma \max_{a'} Q_t(X_{t+1}, a')\Big],$$

and $Q_{t+1}(x,a) = Q_t(x,a)$ otherwise. Identify the correspondence:

$$\theta_t \equiv Q_t, \qquad L \equiv T^\star, \qquad (T^\star Q)(x,a) = r(x,a) + \gamma \sum_{x'} P(x'|x,a) \max_{a'} Q(x', a').$$

Then the update becomes exactly of the SA form:

$$Q_{t+1}(i) \;\leftarrow\; (1-\alpha_t(i))Q_t(i) \;+\; \alpha_t(i)\Big[(T^\star Q_t)(i) + \eta_t(i)\Big], \tag{2}$$

where the noise term is

$$\eta_t(X_t, A_t) \;=\; \Big(R_t + \gamma \max_{a'} Q_t(X_{t+1}, a')\Big) \;-\; (T^\star Q_t)(X_t, A_t).$$

Thus Q-learning is a stochastic approximation to the Bellman optimality operator $T^\star$.

2.2 Verifying Each SA Assumption

(2) Contraction property of $L = T^\star$. The Bellman optimality operator $T^\star$ is γ-contractive in $\|\cdot\|_\infty$:

$$\|T^\star Q_1 - T^\star Q_2\|_\infty \le \gamma \|Q_1 - Q_2\|_\infty,$$

ensuring a unique fixed point $Q^\star$.

(1) Step-size and visitation. Choose $\alpha_t(x, a)$ satisfying

$$\sum_{t=0}^{\infty} \alpha_t(x, a) = \infty, \qquad \sum_{t=0}^{\infty} \alpha_t(x, a)^2 < \infty,$$

and assume sufficient exploration so that each (x, a) is visited infinitely often, giving infinitely many non-zero updates.

A1(a) Zero-mean noise. Let $\mathcal{F}_t$ be the history up to choosing $\alpha_t(X_t, A_t)$ but before observing (X_{t+1}, R_t). Then

$$\mathbb{E}[\eta_t(X_t, A_t) \mid \mathcal{F}_t] = \mathbb{E}\left[R_t + \gamma \max_{a'} Q_t(X_{t+1}, a') \mid \mathcal{F}_t\right] - (T^\star Q_t)(X_t, A_t) = 0,$$

since the conditional expectation equals the Bellman target.

A1(b) Bounded conditional variance. Assume the conditional variance of rewards is bounded: $\mathrm{Var}[R_t \mid X_t, A_t] \le \sigma_R^2$. Then

$$\mathbb{E}\big[\eta_t(X_t, A_t)^2 \mid \mathcal{F}_t\big] \le 2\,\mathrm{Var}[R_t \mid X_t, A_t] + 2\gamma^2\,\mathrm{Var}\left[\max_{a'} Q_t(X_{t+1}, a') \mid X_t, A_t\right].$$

Moreover,

$$\mathrm{Var}\left[\max_{a'} Q_t(X_{t+1}, a') \mid X_t, A_t\right] \le \mathbb{E}\left[\max_{a'} Q_t(X_{t+1}, a')^2 \mid X_t, A_t\right] \le \|Q_t\|_2^2,$$

so that

$$\mathbb{E}\big[\eta_t(X_t, A_t)^2 \mid \mathcal{F}_t\big] \le 2\sigma_R^2 + 2\gamma^2\,\|Q_t\|_2^2.$$

Hence A1(b) holds with $c_1 = 2\sigma_R^2$ and $c_2 = 2\gamma^2$.

2.3 Conclusion: Almost-Sure Convergence of Q-Learning

All SA assumptions are satisfied:

- step-sizes and sufficient exploration,
- noise A1(a,b),
- γ-contraction of $T^\star$.

Therefore, by the stochastic approximation theorem,

$$Q_t \xrightarrow[t\to\infty]{\text{a.s.}} Q^\star.$$

Intuitively, each update is a noisy sample of the Bellman optimality operator. The learning-rate schedule averages out the noise, and the contraction property ensures convergence to the unique fixed point $Q^\star$.