Chapter 7 Extensions

Deep Q-Learning

The Optimal Action-Value Function

Optimality of Observation-Based Policies

Chapter 7 assumes that, when the Markov property is satisfied, we may restrict attention to policies whose action distribution depends only on the current observation. We now prove this claim.

Consider a finite Markovian environment with

$$|r_t| \leq r_{\max}$$

for $t > 0$ and discount factor $0 \leq \gamma < 1$. Let π^* be the deterministic policy at which the policy-improvement construction in Chapter 7 stops. Thus,

$$\pi^*(o) = \arg\max_{a \in \mathcal{A}} Q^{\pi^*}(o, a)$$

for every observation o, as stated in Equation 7.17.

Let h be any history whose last observation is o, and let μ be any history-dependent policy, so that its action distribution $\mu(\cdot \mid h)$ may depend on the entire history. Write $V^{\mu}(h)$ for the expected discounted reward collected from h onward when future actions are selected according to μ. Then

$$V^{\mu}(h) \leq V^{\pi^*}(o). \tag{7E.1}$$

Proof.

We first define the value of a history. Fix a history h ending at time t. Starting with this history as the information currently available to the agent, let every future action be drawn from μ. Define

$$V^{\mu}(h) \stackrel{\text{def}}{=} \mathbb{E}[G_t \mid h]. \tag{7E.2}$$

The notation after the conditioning bar means that h is treated as the starting condition for the future interaction. Once h is fixed, the policy specifies its future action distributions and the environment specifies the distributions of the future observations and rewards, so the expectation is well defined.

When the environment is Markov and the policy depends only on the last observation, this continuation value depends on h only through that last observation and reduces to the ordinary value function of Equation 7.1. For a general history-dependent policy, two histories ending in the same observation may have different values.

Because $|r_t| \leq r_{\max}$ and $0 \leq \gamma < 1$, the absolute value of every discounted reward-to-go is at most

$$r_{\max} + \gamma r_{\max} + \gamma^2 r_{\max} + \cdots = \frac{r_{\max}}{1 - \gamma}.$$

Therefore,

$$|V^{\mu}(h)| \leq \frac{r_{\max}}{1 - \gamma} \tag{7E.3}$$

for every history h. The same bound holds for $V^{\pi^*}(o)$ at every observation o.

The history value also satisfies a one-step recursion. Let H_{t+1} denote the random history one step later. Its possible values have the form

$$h_{t+1} = (h, a_t, r_{t+1}, o_{t+1}).$$

Since

$$G_t = r_{t+1} + \gamma G_{t+1},$$

Equation 7E.2 first gives

$$V^{\mu}(h) = \mathbb{E}[r_{t+1} + \gamma G_{t+1}].$$

Here the starting history h is fixed. The randomness consists of the current action a_t drawn from $\mu(\cdot \mid h)$, the reward r_{t+1} and next observation o_{t+1} generated by the environment, and everything that happens after the resulting next history.

We now condition on the random next history H_{t+1}. By the law of total expectation from Equation 3.7,

$$V^{\mu}(h) = \mathbb{E}_{h_{t+1} \sim H_{t+1}}\big[\mathbb{E}[r_{t+1} + \gamma G_{t+1} \mid H_{t+1} = h_{t+1}]\big].$$

Once $H_{t+1} = h_{t+1}$ is fixed, the reward r_{t+1} is fixed because it is part of h_{t+1}. The expected remaining reward-to-go is $V^{\mu}(h_{t+1})$ by Equation 7E.2. Therefore,

$$\begin{aligned} &\mathbb{E}[r_{t+1} + \gamma G_{t+1} \mid H_{t+1} = h_{t+1}] \\ &= r_{t+1} + \gamma \mathbb{E}[G_{t+1} \mid H_{t+1} = h_{t+1}] \\ &= r_{t+1} + \gamma V^{\mu}(h_{t+1}). \end{aligned}$$

Substituting this equality into the outer expectation gives

$$V^{\mu}(h) = \mathbb{E}_{h_{t+1} \sim H_{t+1}}[r_{t+1} + \gamma V^{\mu}(h_{t+1})]. \tag{7E.4}$$

The distribution of H_{t+1} in this expectation is produced by drawing a_t from $\mu(\cdot \mid h)$ and then letting the environment generate r_{t+1} and o_{t+1}.

We now compare this value with V^{π^*}. Write $o(h)$ for the last observation in a history h. Consider all numbers of the form

$$V^{\mu}(h) - V^{\pi^*}(o(h)).$$

They are bounded above by Equation 7E.3. We use the **supremum**, or smallest upper bound, of these numbers. An upper bound is a number at least as large as every number in the collection; the supremum is the smallest such upper bound. Define

$$d \overset{\text{def}}{=} \sup_{h}\big(V^{\mu}(h) - V^{\pi^*}(o(h))\big). \tag{7E.5}$$

The bounds above imply

$$-\frac{2r_{\max}}{1-\gamma} \le d \le \frac{2r_{\max}}{1-\gamma},$$

so d is a finite real number. By its definition, for every history h,

$$V^{\mu}(h) - V^{\pi^*}(o(h)) \le d. \tag{7E.6}$$

Fix one history h ending at observation o. Expand its value with Equation 7E.4, and add and subtract $\gamma V^{\pi^*}(o_{t+1})$ inside the expectation:

$$\begin{aligned} V^{\mu}(h) = \mathbb{E}\big[r_{t+1} + \gamma V^{\pi^*}(o_{t+1})\big] \\ +\gamma \mathbb{E}\big[V^{\mu}(h_{t+1}) - V^{\pi^*}(o_{t+1})\big]. \end{aligned} \tag{7E.7}$$

Both expectations use the same randomness as Equation 7E.4: the current action is drawn from $\mu(\cdot \mid h)$, and the environment generates the next reward and observation. We bound the two expectations separately.

For the first expectation, use the law of total expectation to average first over the current action selected from $\mu(\cdot \mid h)$:

$$\begin{aligned} &\mathbb{E}\left[r_{t+1} + \gamma V^{\pi^*}(o_{t+1})\right] \\ &= \sum_{a \in \mathcal{A}} \mu\,(a \mid h)\mathbb{E}\left[r_{t+1} + \gamma V^{\pi^*}(o_{t+1}) \mid h, a_t = a\right]. \end{aligned} \tag{7E.8}$$

This is where the Markov property is used. Once the current observation o and action a are fixed, the earlier history has no effect on the distribution of the next observation and reward. Therefore, Equation 7.4 gives

$$\begin{aligned} &\mathbb{E}\left[r_{t+1} + \gamma V^{\pi^*}(o_{t+1}) \mid h, a_t = a\right] \\ &= \mathbb{E}\left[r_{t+1} + \gamma V^{\pi^*}(o_{t+1}) \mid o_t = o, a_t = a\right] \\ &= Q^{\pi^*}(o, a), \end{aligned} \tag{7E.9}$$

where the last equality is the Bellman equation 7.6 for Q^{π^*}. Substituting this equality into Equation 7E.8 identifies the first expectation in Equation 7E.7 with a weighted average of the action values:

$$\begin{aligned} &\mathbb{E}\left[r_{t+1} + \gamma V^{\pi^*}(o_{t+1})\right] \\ &= \sum_{a \in \mathcal{A}} \mu\,(a \mid h) Q^{\pi^*}(o, a). \end{aligned} \tag{7E.10}$$

A weighted average cannot be larger than the largest number being averaged. Since π^* is greedy with respect to its own action values,

$$\begin{aligned} \sum_{a \in \mathcal{A}} \mu\,(a \mid h) Q^{\pi^*}(o, a) &\leq \max_{a \in \mathcal{A}} Q^{\pi^*}(o, a) \\ &= V^{\pi^*}(o). \end{aligned} \tag{7E.11}$$

The final equality follows from Equations 7.8 and Equation 7.17. Equations 7E.10 and Equation 7E.11 together give the direct bound

$$\mathbb{E}\left[r_{t+1} + \gamma V^{\pi^*}(o_{t+1})\right] \leq V^{\pi^*}(o). \tag{7E.12}$$

For the second expectation in Equation 7E.7, the next history h_{t+1} ends in observation o_{t+1}. Equation 7E.6 therefore gives

$$V^{\mu}(h_{t+1}) - V^{\pi^*}(o_{t+1}) \leq d$$

for every possible next history. Taking an expectation preserves this upper bound, so

$$\mathbb{E}\left[V^{\mu}(h_{t+1}) - V^{\pi^*}(o_{t+1})\right] \leq d. \tag{7E.13}$$

Equation 7E.12 bounds the first expectation in Equation 7E.7 by $V^{\pi^*}(o)$. Equation 7E.13 bounds the second expectation there by d. Substituting these two bounds into Equation 7E.7, one at a time, gives

$$\begin{aligned} V^{\mu}(h) &= \mathbb{E}\left[r_{t+1} + \gamma V^{\pi^*}(o_{t+1})\right] + \gamma\mathbb{E}\left[V^{\mu}(h_{t+1}) - V^{\pi^*}(o_{t+1})\right] \\ &\leq V^{\pi^*}(o) + \gamma\mathbb{E}\left[V^{\mu}(h_{t+1}) - V^{\pi^*}(o_{t+1})\right] \\ &\leq V^{\pi^*}(o) + \gamma d. \end{aligned}$$

The first inequality uses Equation 7E.12. For the second inequality, Equation 7E.13 says

$$\mathbb{E}\left[V^{\mu}(h_{t+1}) - V^{\pi^*}(o_{t+1})\right] \leq d.$$

Since $\gamma \geq 0$, multiplying both sides of this inequality by γ gives

$$\gamma\mathbb{E}\left[V^{\mu}(h_{t+1}) - V^{\pi^*}(o_{t+1})\right] \leq \gamma d.$$

This is why the expectation no longer appears in the last line of the preceding calculation: the whole expected difference has been replaced by its upper bound d.

For the fixed history h, its last observation is o, so $o = o(h)$. Subtracting $V^{\pi^*}(o) = V^{\pi^*}(o(h))$ from both sides of

$$V^{\mu}(h) \leq V^{\pi^*}(o) + \gamma d$$

gives

$$V^{\mu}(h) - V^{\pi^*}(o(h)) \leq \gamma d \qquad (7E.14)$$

for every history h. Thus γd is an upper bound on all the differences whose smallest upper bound is d. By the definition of the supremum in Equation 7E.5 and applying the supremum to both sides of Equation 7E.14, it becomes

$$d \leq \gamma d,$$

where the right side didn't change because γd is a constant that does not depend on h, so taking the supremum over h leaves it unaffected.

Subtracting γd from both sides gives

$$(1 - \gamma)d \leq 0.$$

Since $0 \leq \gamma < 1$, the number $1 - \gamma$ is positive. Dividing by it gives

$$d \leq 0.$$

Finally, Equation 7E.6 gives, for every history h ending at observation o,

$$V^{\mu}(h) - V^{\pi^*}(o) \leq d \leq 0.$$

Hence

$$V^{\mu}(h) \leq V^{\pi^*}(o),$$

which proves Equation 7E.1.

Thus, no history-dependent policy can have a larger continuation value than π^* after any history, so restricting attention to observation-based policies causes no loss of optimality.

Q-Learning

Full Convergence Proof of Tabular Q-learning

Chapter 7 introduces the Q-learning target in Equation 7.22 and the tabular update in Equation 7.23. This note proves the convergence result used there. The result applies to tabular Q-learning in a finite Markovian environment; it does not assert convergence for the neural-network version developed later in Chapter 7.

Let Q_t be the Q-table after t environment steps. At time t, the agent observes o_t, selects a_t using any behavior policy, and receives the next reward and observation (r_{t+1}, o_{t+1}). It forms the target

$$y_t = r_{t+1} + \gamma \max_{a' \in \mathcal{A}} Q_t(o_{t+1}, a')$$

and updates only the visited entry:

$$Q_{t+1}(o_t, a_t) = Q_t(o_t, a_t) + \alpha_t(o_t, a_t)[y_t - Q_t(o_t, a_t)].$$

Every other entry remains unchanged. If the transition to o_{t+1} ends the episode, the maximum term is omitted. To cover such transitions in the proof below, we treat the end of an episode as a move to an absorbing observation with zero reward whose action values are fixed at zero and never updated. Then the maximum term is zero exactly when the episode ends, so the Bellman optimality operator defined below and every step of the proof apply to

episode-ending transitions as well. The fixed entries already equal their Q^* value of zero, so they are excluded from the requirement to be visited infinitely often.

Convergence of tabular Q-learning. Suppose that:

- the environment is Markovian;
- the observation and action spaces are finite;
- rewards are bounded: there is a finite number $r_{\max}$ such that $|r_t| \leq r_{\max}$;
- the discount factor satisfies $0 \leq \gamma < 1$;
- every observation-action pair is visited infinitely often;
- the step sizes satisfy $0 \leq \alpha_k(o,a) \leq 1$ and, for every observation-action pair (o,a),

$$\sum_k \alpha_k\,(o,a) = \infty, \qquad \sum_k \alpha_k\,(o,a)^2 < \infty,$$

where k counts visits to that particular pair.

Then, for every observation-action pair (o,a),

$$Q_t(o,a) \to Q^*(o,a)$$

with probability 1.

The two step-size requirements are the **Robbins-Monro conditions**. The first requires the total amount of learning assigned to an entry to be infinite, so the step sizes never shrink so fast that the entry stops moving before it reaches its target. The second makes the later steps small enough for random transition noise to average away. A standard example is $\alpha_k(o,a) = 1/(k+1)$, counting the first visit as $k = 0$ as in Equation 7.33.

Proof

We use the following timing convention. At the initial time $t = 0$, the environment has issued only the initial observation o_0. For every $t \geq 1$, it has already issued the pair (r_t, o_t). The agent observes o_t and chooses action a_t. After that action is executed, the environment issues the next pair (r_{t+1}, o_{t+1}).

Sometimes, to distinguish the current quantities from the immediately following ones, we denote the current observation-action pair by (o,a) and write the next pair issued by the environment using prime notation: (r', o').

The prime means "next." The pair (r', o') is issued together by the environment after action a is executed in observation o.

Let

$$\mathcal{X} \stackrel{\text{def}}{=} \mathcal{O} \times \mathcal{A}$$

be the finite set of all observation-action pairs. A Q-table is just a vector with one component for each element of $\mathcal{X}$.

We will measure the distance between two Q-tables using the **supremum norm**, also called the **max norm**:

$$\|Q_1 - Q_2\|_\infty \stackrel{\text{def}}{=} \max_{(o,a)\in\mathcal{X}} |Q_1(o,a) - Q_2(o,a)|.$$

So $\|Q_1 - Q_2\|_\infty \leq \varepsilon$ means that the two tables differ by at most ε in every cell.

Now define the **Bellman optimality operator** T. An **operator** is just a function whose input is itself a function or table. Applied to a Q-table Q, the operator T produces a new Q-table, written TQ as shorthand for T(Q). Because TQ is itself a table, $(\mathrm{T}Q)(o,a)$ denotes its entry for the observation-action pair (o,a). We define the output table one entry at a time:

$$(\mathrm{T}Q)(o,a) \stackrel{\text{def}}{=} \mathbb{E}\left[r' + \gamma \max_{a'\in\mathcal{A}} Q(o',a') \mid o,a\right]. \qquad (7E.15)$$

The expectation is over the environment's possible next emissions after the agent takes action a in observation o. Each next emission is the pair (r', o'): the next reward and the next observation. In a deterministic environment, the same pair (o, a) always leads to the same emitted pair (r', o'). In a stochastic environment, the same pair (o, a) may lead to different emitted pairs on different visits. In either case, once (o, a) is fixed, the behavior policy does not appear in this definition.

Equation 7.21 says exactly that

$$Q^* = \mathrm{T}Q^*. \tag{7E.16}$$

A table satisfying $Q = \mathrm{T}Q$ is called a **fixed point** of T: applying T to the table gives back the same table.

The proof that the tabular Q-learning update in Equation 7.23 converges to Q^* has four steps. First, we show that T is a contraction. Second, we show that its fixed point is unique. Third, we show that the sampled Q-learning update has the Bellman update as its expectation. Fourth, we show that the sampling noise averages away under the learning-rate assumptions above.

Step 1: The Bellman optimality operator is a contraction

A mapping T is called a **contraction** with factor γ if the distance between its outputs is at most γ times the distance between its inputs:

$$\|\mathrm{T}Q_1 - \mathrm{T}Q_2\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty. \tag{7E.17}$$

Because $\gamma < 1$, a contraction shrinks worst-case disagreement.

Take any two Q-tables Q_1 and Q_2, and fix one pair (o, a). By the definition of T,

$$(\mathrm{T}Q_1)(o, a) = \mathbb{E}\left[r' + \gamma \max_b Q_1(o', b) \mid o, a\right],$$

and

$$(\mathrm{T}Q_2)(o, a) = \mathbb{E}\left[r' + \gamma \max_b Q_2(o', b) \mid o, a\right].$$

Both expectations average over the same possible next emitted pairs (r', o') from the same previous observation-action pair (o, a). Therefore the reward part cancels when we subtract:

$$\begin{aligned}(\mathrm{T}Q_1)(o, a) - (\mathrm{T}Q_2)(o, a) &= \mathbb{E}\left[r' + \gamma \max_b Q_1(o', b) - r' - \gamma \max_b Q_2(o', b) \mid o, a\right] \\ &= \mathbb{E}\left[\gamma \max_b Q_1(o', b) - \gamma \max_b Q_2(o', b) \mid o, a\right] \\ &= \gamma \mathbb{E}\left[\max_b Q_1(o', b) - \max_b Q_2(o', b) \mid o, a\right].\end{aligned}$$

The immediate reward r' cancels because it is the same emitted reward in both expressions being compared. The only remaining difference is in the future-value term, and that term is multiplied by the discount factor γ inside the Bellman operator.

Now take absolute values. The absolute value of an average is no larger than the average absolute value, so

$$|(\mathrm{T}Q_1)(o, a) - (\mathrm{T}Q_2)(o, a)| \leq \gamma \mathbb{E}\left[\left|\max_b Q_1(o', b) - \max_b Q_2(o', b)\right| \mid o, a\right]. \tag{7E.18}$$

The max operator is **non-expansive** under the supremum norm, as shown in the main chapter's derivation of Equation 7.27. It means that for any two functions f and g the following holds:

$$\left|\max_b f(b) - \max_b g(b)\right| \leq \max_b |f(b) - g(b)|$$

Therefore, for any next observation o' that can appear in an emitted pair (r', o'),

$$\left|\max_b Q_1(o', b) - \max_b Q_2(o', b)\right| \leq \max_b |Q_1(o', b) - Q_2(o', b)|$$

This inequality says that taking a maximum over actions cannot create more disagreement than the largest action-value disagreement already present at that observation.

And since the largest disagreement at this particular next observation cannot exceed the largest disagreement anywhere in the whole table,

$$\max_b |Q_1(o', b) - Q_2(o', b)| \leq \|Q_1 - Q_2\|_\infty.$$

Chaining the two inequalities bounds the disagreement of the maxima at any reachable next observation by $\|Q_1 - Q_2\|_\infty$, a quantity that is constant with respect to the expectation over the next emitted pair. Taking the expectation of both sides of that bound gives

$$\mathbb{E}\left[\left|\max_b Q_1(o', b) - \max_b Q_2(o', b)\right| \mid o, a\right] \leq \|Q_1 - Q_2\|_\infty.$$

Substituting this into Equation 7E.18 gives, for the fixed pair (o, a),

$$|(\mathrm{T}Q_1)(o, a) - (\mathrm{T}Q_2)(o, a)| \leq \gamma \|Q_1 - Q_2\|_\infty. \qquad (7E.19)$$

The pair (o, a) was arbitrary. Therefore Equation 7E.19 holds for every observation-action pair:

$$\forall (o, a) \in \mathcal{X}, \qquad |(\mathrm{T}Q_1)(o, a) - (\mathrm{T}Q_2)(o, a)| \leq \gamma \|Q_1 - Q_2\|_\infty.$$

The right-hand side does not depend on (o, a). A number that bounds the absolute value of every cell of the table $\mathrm{T}Q_1 - \mathrm{T}Q_2$ also bounds the largest of those absolute values, and by the definition of the max norm that largest value is $\|\mathrm{T}Q_1 - \mathrm{T}Q_2\|_\infty$. Therefore

$$\|\mathrm{T}Q_1 - \mathrm{T}Q_2\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty.$$

So, by definition of contraction given in Equation 7E.17, T is a contraction.

Step 2: The contraction has only one fixed point, and it is Q^*

Suppose two tables Q_A and Q_B are both fixed points of T:

$$Q_A = \mathrm{T}Q_A, \qquad Q_B = \mathrm{T}Q_B.$$

Then

$$\|Q_A - Q_B\|_\infty = \|\mathrm{T}Q_A - \mathrm{T}Q_B\|_\infty.$$

Since T is a contraction,

$$\|\mathrm{T}Q_A - \mathrm{T}Q_B\|_\infty \leq \gamma \|Q_A - Q_B\|_\infty.$$

Combining the two equations,

$$\|Q_A - Q_B\|_\infty \leq \gamma \|Q_A - Q_B\|_\infty.$$

Because $\gamma < 1$, this is possible only if

$$\|Q_A - Q_B\|_\infty = 0.$$

So $Q_A = Q_B$ which means the Bellman optimality operator cannot have two different fixed points.

But we established in Equation 7E.16 that Q^* satisfies the Bellman optimality equation,

$$Q^* = \mathrm{T}Q^*.$$

Therefore Q^* is the unique fixed point of T.

This already answers one important question: Q-learning cannot converge to some other Bellman-consistent table, because no other Bellman-consistent table exists.

Step 3: The sampled Q-learning target has $(\mathrm{T}Q_t)(o_t, a_t)$ as its conditional expectation

Now connect the Bellman optimality operator to the target used by the actual Q-learning update.

Let Q_t be the table after t environment steps. Let h_t denote the history available before the action at time t is chosen. This history includes the current observation o_t and all earlier observations, actions, and rewards. Since the table is produced by the previous Q-learning updates, Q_t is determined by h_t.

The behavior policy chooses action a_t. It may be greedy, epsilon-greedy, random, hand-coded, changing over time, or dependent on the current table. The convergence proof does not require this policy to be optimal, because its randomness is used only for exploration, not for defining the value being learned: once the agent has sampled (o_t, a_t), the environment's response to that pair is the same regardless of why the agent chose a_t. The behavior policy only decides which transitions the agent gets to see; the Q-learning target decides what those transitions are used to learn. The proof therefore asks only one thing of the behavior policy—coverage: every observation-action pair must be visited infinitely often, because an entry that is never visited is never corrected.

After the agent executes a_t in observation o_t, the environment emits the next reward and the next observation,

$$r_{t+1}, o_{t+1}.$$

Q-learning forms the sampled target

$$y_t \overset{\text{def}}{=} r_{t+1} + \gamma \max_{a' \in \mathcal{A}} Q_t(o_{t+1}, a').$$

Then it updates only the visited table entry:

$$Q_{t+1}(o_t, a_t) = Q_t(o_t, a_t) + \alpha_t(o_t, a_t)[y_t - Q_t(o_t, a_t)], \tag{7E.20}$$

and leaves every other entry unchanged:

$$Q_{t+1}(o, a) = Q_t(o, a) \qquad \text{for } (o, a) \neq (o_t, a_t).$$

We now compute the conditional expectation of y_t after a_t has been selected and before the environment emits r_{t+1} and o_{t+1}. Condition on h_t and a_t. Then Q_t, o_t, and a_t are fixed. The only remaining randomness in y_t comes from the environment's next emitted pair (r_{t+1}, o_{t+1}).

By the definition of y_t,

$$\mathbb{E}[y_t \mid h_t, a_t] = \mathbb{E}\left[r_{t+1} + \gamma \max_{a' \in \mathcal{A}} Q_t(o_{t+1}, a') \mid h_t, a_t\right].$$

By the Markov property, once o_t and a_t are fixed, the conditional distribution of the next emitted reward and observation does not depend on the earlier history:

$$p(o_{t+1}, r_{t+1} \mid h_t, a_t) = p(o_{t+1}, r_{t+1} \mid o_t, a_t).$$

Since Q_t is fixed under the conditioning on h_t, the same replacement of the conditioning applies to the expression inside the expectation:

$$\mathbb{E}\left[r_{t+1} + \gamma \max_{a' \in \mathcal{A}} Q_t(o_{t+1}, a') \mid h_t, a_t\right] = \mathbb{E}\left[r_{t+1} + \gamma \max_{a' \in \mathcal{A}} Q_t(o_{t+1}, a') \mid o_t, a_t\right].$$

By the definition of the Bellman optimality operator,

$$(\mathrm{T}Q_t)(o_t, a_t) = \mathbb{E}\left[r_{t+1} + \gamma \max_{a' \in \mathcal{A}} Q_t(o_{t+1}, a') \mid o_t, a_t\right].$$

Therefore,

$$\mathbb{E}[y_t \mid h_t, a_t] = (\mathrm{T}Q_t)(o_t, a_t). \tag{7E.21}$$

By the definition of unbiasedness in Equation 2.1, this means that, conditioned on h_t and a_t, y_t is an unbiased estimate of the exact Bellman update $(\mathrm{T}Q_t)(o_t, a_t)$ for the visited entry.

This proves the claim of this step: conditioned on h_t and a_t, the sampled Q-learning target y_t has conditional expectation equal to the Bellman optimality operator applied to the current table at the visited entry. In words: every visit to the pair (o_t, a_t) produces one sample of the same underlying question—how good is this action in this observation if, after the first step, the agent acts greedily?—no matter which policy generated the visit. Some

samples are lucky and some are unlucky, but Q-learning never needs the average answer in advance; Step 4 shows that repeated samples are enough.

Step 4: The noise averages away, so the error converges to zero

Step 3 showed that the sampled target y_t has conditional expectation $(\mathrm{T}Q_t)(o_t, a_t)$. The difference between the sampled target and this conditional expectation is the sampling noise:

$$w_{t+1} \stackrel{\text{def}}{=} y_t - (\mathrm{T}Q_t)(o_t, a_t). \tag{7E.22}$$

By Equation 7E.21,

$$\mathbb{E}[w_{t+1} \mid h_t, a_t] = 0. \tag{7E.23}$$

As t ranges over training steps, the terms w_{t+1} form the sequence of sampling-noise terms. Equation 7E.23 says that each term has zero conditional expectation given the history and selected action before the corresponding environment emission. Such a noise sequence is called a **martingale difference sequence**.

We also need the noise to be bounded. Let

$$B \stackrel{\text{def}}{=} \max\left\{\|Q_0\|_\infty, \frac{r_{\max}}{1-\gamma}\right\}.$$

Suppose $\|Q_t\|_\infty \le B$. Since rewards are bounded by $r_{\max}$,

$$|y_t| \le r_{\max} + \gamma B.$$

Because

$$B \ge \frac{r_{\max}}{1-\gamma},$$

we have

$$r_{\max} + \gamma B \le B.$$

Therefore,

$$|y_t| \le B.$$

The Q-learning update can be written as

$$Q_{t+1}(o_t, a_t) = (1-\alpha_t)Q_t(o_t, a_t) + \alpha_t y_t,$$

where α_t abbreviates $\alpha_t(o_t, a_t)$. Since $0 \le \alpha_t \le 1$, this is a weighted average of $Q_t(o_t, a_t)$ and y_t. If both values lie between $-B$ and B, then the updated value also lies between $-B$ and B. All other entries are unchanged. The base case holds by the definition of B: $\|Q_0\|_\infty \le B$. Therefore, by induction,

$$\|Q_t\|_\infty \le B$$

for all t.

The noise term is then bounded by an explicit constant. The same computation that bounded y_t also bounds the Bellman target: since rewards are bounded by $r_{\max}$ and $\|Q_t\|_\infty \le B$,

$$|(\mathrm{T}Q_t)(o, a)| \le r_{\max} + \gamma B \le B$$

for every pair (o, a). Therefore, by the triangle inequality,

$$|w_{t+1}| = |y_t - (\mathrm{T}Q_t)(o_t, a_t)| \le |y_t| + |(\mathrm{T}Q_t)(o_t, a_t)| \le 2B.$$

This is the bounded-noise constant that the convergence theorem below requires.

Now define the error table

$$e_t \stackrel{\text{def}}{=} Q_t - Q^*. \tag{7E.24}$$

For the visited pair (o_t, a_t), subtract $Q^*(o_t, a_t)$ from both sides of the Q-learning update. Since $Q^* = \mathrm{T}Q^*$,

$$Q^*(o_t, a_t) = (\mathrm{T}Q^*)(o_t, a_t).$$

Thus

$$\begin{aligned} e_{t+1}(o_t, a_t) &= Q_{t+1}(o_t, a_t) - Q^*(o_t, a_t) \\ &= Q_t(o_t, a_t) + \alpha_t[y_t - Q_t(o_t, a_t)] - Q^*(o_t, a_t) \\ &= (1 - \alpha_t)[Q_t(o_t, a_t) - Q^*(o_t, a_t)] + \alpha_t[y_t - Q^*(o_t, a_t)] \\ &= (1 - \alpha_t)e_t(o_t, a_t) + \alpha_t[y_t - (\mathrm{T}Q^*)(o_t, a_t)]. \end{aligned}$$

Using the definition of w_{t+1},

$$y_t = (\mathrm{T}Q_t)(o_t, a_t) + w_{t+1}.$$

Substituting this into the previous display gives the error recursion

$$e_{t+1}(o_t, a_t) = (1 - \alpha_t)e_t(o_t, a_t) + \alpha_t[(\mathrm{T}Q_t)(o_t, a_t) - (\mathrm{T}Q^*)(o_t, a_t) + w_{t+1}]. \qquad (7E.25)$$

For all unvisited entries,

$$e_{t+1}(o, a) = e_t(o, a) \qquad \text{for } (o, a) \neq (o_t, a_t).$$

In Equation 7E.25, the bracketed expression is the sum of two scalar terms. The **Bellman-error term** is the Bellman-operator difference

$$b_t \stackrel{\text{def}}{=} (\mathrm{T}Q_t)(o_t, a_t) - (\mathrm{T}Q^*)(o_t, a_t). \qquad (7E.26)$$

The other scalar term is the sampling-noise term w_{t+1} from Equation 7E.22. The contraction property of T bounds the Bellman-error term:

$$\begin{aligned} |b_t| &= |(\mathrm{T}Q_t)(o_t, a_t) - (\mathrm{T}Q^*)(o_t, a_t)| \\ &\leq \|\mathrm{T}Q_t - \mathrm{T}Q^*\|_\infty \\ &\leq \gamma\|Q_t - Q^*\|_\infty \\ &= \gamma\|e_t\|_\infty. \end{aligned} \qquad (7E.27)$$

So the error recursion has two parts. The first part is the Bellman-error term from Equation 7E.26, whose magnitude is at most $\gamma\|e_t\|_\infty$ by Equation 7E.27. Since $\gamma < 1$, this part shrinks the worst-case error. The second part is the bounded martingale-difference noise w_{t+1}.

We now use the standard asynchronous stochastic-approximation theorem, stated here in abbreviated form. Its full statement and step-by-step proof—together with the Robbins-Monro averaging lemma on which the proof rests—are given in the companion note *Asynchronous stochastic-approximation theorem*.

Asynchronous stochastic-approximation theorem. Consider a finite table of errors e_t. Suppose that entries are updated one at a time, every entry is updated infinitely often, and the learning rates satisfy $0 \leq \alpha_t \leq 1$ and, for each entry,

$$\sum_k \alpha_k\,(o, a) = \infty, \qquad \sum_k \alpha_k\,(o, a)^2 < \infty.$$

Suppose also that the noise terms are bounded martingale differences with respect to the histories available just before each update's noise is observed, that the non-noise part of each update is determined by that history, and that it has magnitude at most $\gamma\|e_t\|_\infty$ for some $0 \leq \gamma < 1$. Then

$$\|e_t\|_\infty \to 0$$

with probability 1.

The Q-learning error recursion satisfies these conditions:

- The table is finite.

- Every observation-action pair is visited infinitely often.
- The learning rates satisfy $0 \leq \alpha_t(o_t, a_t) \leq 1$ and the Robbins-Monro conditions.
- The noise term w_{t+1} is bounded—by the constant $2B$ computed above—and has conditional mean zero given h_t and a_t.
- The Bellman-error term is determined by the pre-transition information: once h_t and a_t are fixed, so are Q_t, o_t, and a_t, and therefore so is b_t.
- The Bellman-error term is bounded in magnitude by $\gamma\|e_t\|_\infty$.

Therefore,

$$\|e_t\|_\infty \to 0$$

with probability 1. Since $e_t = Q_t - Q^*$,

$$\|Q_t - Q^*\|_\infty \to 0.$$

Equivalently,

$$Q_t(o, a) \to Q^*(o, a)$$

for every observation-action pair (o, a).

This completes the convergence argument. Step 1 showed that the Bellman optimality operator is a contraction. Step 2 showed that its unique fixed point is Q^*. Step 3 showed that the sampled Q-learning target has this Bellman target as its conditional expectation. Step 4 showed that the remaining sampling noise averages away under the learning-rate and exploration assumptions. Therefore tabular Q-learning converges to Q^* with probability 1.

What the proof establishes

The behavior policy determines which observation-action pair is visited, but it does not determine the value toward which that pair is updated. Once (o_t, a_t) is fixed, the conditional average of the sampled target is the Bellman optimality update $(\mathrm{T}Q_t)(o_t, a_t)$. The operator T has Q^* as its unique fixed point and contracts the largest table error by the factor $\gamma < 1$.

The actual update sees a random sample rather than the conditional average. Repeated visits ensure that every entry continues to be corrected, while the Robbins-Monro step sizes make the accumulated sampling noise vanish. The asynchronous stochastic-approximation theorem combines these facts and yields convergence of the whole table to Q^*.

This explains the off-policy result. The behavior policy controls which transitions are collected; the Q-learning target controls what those transitions are used to learn. No importance weights are required because the target does not contain the behavior policy's action probabilities.