Theory Project

Lemma 1

Lemma 1. *Let $X = \{X_1, \ldots, X_M\}$ be a set of random variables and let $\mu^A = \{\mu_1^A, \ldots, \mu_M^A\}$ and $\mu^B = \{\mu_1^B, \ldots, \mu_M^B\}$ be two sets of unbiased estimators such that $E\{\mu_i^A\} = E\{\mu_i^B\} = E\{X_i\}$, for all i. Let $\mathcal{M} \stackrel{\text{def}}{=} \{j \mid E\{X_j\} = \max_i E\{X_i\}\}$ be the set of elements that maximize the expected values. Let a^* be an element that maximizes μ^A: $\mu_{a^*}^A = \max_i \mu_i^A$. Then $E\{\mu_{a^*}^B\} = E\{X_{a^*}\} \leq \max_i E\{X_i\}$. Furthermore, the inequality is strict if and only if $P(a^* \notin \mathcal{M}) > 0$.*

Proof.

$\boldsymbol{E\{\mu_{a^*}^B\} \leq \max_i E\{X_i\}}$

$\because E\{\mu_i^B\} = E\{X_i\} \ \therefore E\{\mu_{a^*}^B\} = E\{X_{a^*}\}$

Case 1. $a^* \in M$, then $E\{X_{a^*}\} = \max_i E\{X_i\}$

Case 2. $a^* \notin M$, then $E\{X_{a^*}\} < \max_i E\{X_i\}$

$\Rightarrow E\{\mu_{a^*}^B\} = E\{X_{a^*}\} \leq \max_i E\{X_i\}$

$\boldsymbol{E\{\mu_{a^*}^B\} < \max_i E\{X_i\} \Leftrightarrow P(a^* \notin M) > 0}$

$E\{\mu_{a^*}^B\} = P(a^* \in M)E\{X_{a^*}|a^* \in M\} + P(a^* \notin M)E\{X_{a^*}|a^* \notin M\}$

$E\{X_{a^*}|a^* \in M\} = \max_i E\{X_i\}$

$(\Rightarrow) E\{\mu_{a^*}^B\} < \max_i E\{X_i\} \Rightarrow P(a^* \notin M) > 0$

$E\{\mu_{a^*}^B\} < \max_i E\{X_i\}$

if $P(a^* \notin M) = 0$
then $E\{\mu_{a^*}^B\} = P(a^* \in M)E\{X_{a^*}|a^* \in M\} + P(a^* \notin M)E\{X_{a^*}|a^* \notin M\}$
$= 1 * E\{X_{a^*}|a^* \in M\} + 0 * E\{X_{a^*}|a^* \notin M\}$

$= \max_i E\{X_i\}$, contradiction with $E\{\mu_{a^*}^B\} < \max_i E\{X_i\}$, so $P(a^* \notin M) > 0$

$(\Leftarrow) P(a^* \notin M) > 0 \Rightarrow E\{\mu_{a^*}^B\} < \max_i E\{X_i\}$

$P(a^* \notin M) > 0$

$\max_i E\{X_i\} - E\{\mu_{a^*}^B\}$

$= E\{X_{a^*}|a^* \in M\} - (P(a^* \in M)E\{X_{a^*}|a^* \in M\} + P(a^* \notin M)E\{X_{a^*}|a^* \notin M\})$
$= P(a^* \notin M)(E\{X_{a^*}|a^* \in M\} - E\{X_{a^*}|a^* \notin M\})$
$> 0 \ (\because P(a^* \notin M) > 0$ *and* $E\{X_{a^*}|a^* \in M\} - E\{X_{a^*}|a^* \notin M\} > 0)$

so $E\{\mu_{a^*}^B\} < \max_i E\{X_i\}$

Lemma 2

Lemma 2. *Consider a stochastic process* (ζ_t, Δ_t, F_t), $t \geq 0$, *where* $\zeta_t, \Delta_t, F_t : X \to \mathbb{R}$ *satisfy the equations:*

$$\Delta_{t+1}(x_t) = (1 - \zeta_t(x_t))\Delta_t(x_t) + \zeta_t(x_t)F_t(x_t) \ , \tag{8}$$

where $x_t \in X$ *and* $t = 0, 1, 2, \ldots$. *Let* P_t *be a sequence of increasing* σ*-fields such that* ζ_0 *and* Δ_0 *are* P_0*-measurable and* ζ_t, Δ_t *and* F_{t-1} *are* P_t*-measurable,* $t = 1, 2, \ldots$. *Assume that the following hold: 1) The set X is finite. 2)* $\zeta_t(x_t) \in [0, 1]$, $\sum_t \zeta_t(x_t) = \infty$, $\sum_t (\zeta_t(x_t))^2 < \infty$ *w.p.1 and* $\forall x \neq x_t : \zeta_t(x) = 0$. *3)* $||E\{F_t|P_t\}|| \leq \kappa||\Delta_t|| + c_t$, *where* $\kappa \in [0, 1)$ *and* c_t *converges to zero w.p. 1. 4)* $Var\{F_t(x_t)|P_t\} \leq K(1 + \kappa||\Delta_t||)^2$, *where* K *is some constant. Here* $||\cdot||$ *denotes a maximum norm. Then* Δ_t *converges to zero with probability one.*

Conditions

1. The set X is finite.
2. $\zeta_t(x_t) \in [0,1], \Sigma_t\zeta_t(x_t) = \infty,\ \Sigma_t\zeta_t(x_t)^2 < \infty\ \ w.p.1,\ \forall x \neq x_t, \zeta_t(x) = 0$
3. $\big||E\{F_t|P_t\}|\big| \leq \kappa\big||\Delta_t|\big| + c_t$, where $\kappa \in [0,1)$ and $c_t \to 0\ w.p.1$
4. $Var\{F_t(x_t)|P_t\} \leq K\big(1 + \kappa\big||\Delta_t|\big|\big)^2$, *where K is some constant.*

Proof.

The Stochastic Approximation (Jaakkola, Jordan, Singh, 1993):

$$\Delta_{t+1}(x) = (1 - \alpha_t(x)) \cdot \Delta_t(x) + \beta_t(x) \cdot \varepsilon_t(x),\ \ \forall x \in X \quad \ldots \text{(SA1)}$$

▸ **Stochastic Approximation (Jaakkola, Jordan, Singh, 1993)**:
If the following conditions are satisfied, then $\Delta_t \to 0$, w.p.1:

1. $\sum_t \alpha_t(x) = \infty$, $\sum_t \alpha_t(x)^2 < \infty$, $\sum_t \beta_t(x) = \infty$, and $\sum_t \beta_t(x)^2 < \infty$, w.p.1
2. $\mathbb{E}\big[\beta_t(x) \,|\, \mathcal{H}_t\big] \leq \mathbb{E}\big[\alpha_t(x) \,|\, \mathcal{H}_t\big]$, w.p.1
3. $\Big|\mathbb{E}\big[\varepsilon_t(x) \,|\, \mathcal{H}_t\big]\Big| < \rho\|\Delta_t\|_\infty, \rho \in (0,1)$
4. $\mathbb{V}\big[\varepsilon_t(x) \,|\, \mathcal{H}_t\big] \leq C(1 + \|\Delta_t\|_\infty)^2$

Let $\alpha_t = \zeta_t,\ \beta_t = \zeta_t,\ \epsilon_t = F_t,\ C = K$ (some constant),

By the Stochastic Approximation (Jaakkola, Jordan, Singh, 1993), check the four conditions:

1. ∵ the condition 2 in lemma 2: $\Sigma_t\zeta_t(x) = \infty,\ \Sigma_t\zeta_t(x)^2 = \infty$, w.p.1, this condition is satisfied.
2. ∵ $\alpha_t = \beta_t = \zeta_t$, this condition is satisfied.
3. ∵ c_t converges to 0 w. p. 1 and $\kappa \in [0,1)$ and some $\rho \in (0,1)$ can satisfy $\kappa < \rho$, $c_t < (\rho - \kappa)\big||\Delta_t|\big|_\infty$, ∴ from the condition 3 in lemma 2:

   $|E[F_t|P_t]| \leq \big||E[F_t|P_t]|\big|_\infty \leq \kappa\big||\Delta_t|\big|_\infty + c_t < \rho\big||\Delta_t|\big|_\infty$, this condition is satisfied.

4. Because $\kappa < 1$, from the condition 4 in lemma 2:

   $Var\{F_t(x_t)|P_t\} \leq K\left(1 + \kappa\big||\Delta_t|\big|_\infty\right)^2 \leq K\left(1 + 1 \cdot \big||\Delta_t|\big|_\infty\right)^2$, this condition is satisfied.

Lemma 2 satisfies those conditions, so by the Stochastic Approximation (Jaakkola, Jordan, Singh, 1993), $\Delta_{t+1}(x_t) = \big(1 - \zeta_t(x_t)\big)\,\Delta_t(x) + \zeta_t(x)F_t(x_t)$ converges to 0 w. p. 1.

Theorem 1

Theorem 1. *Assume the conditions below are fulfilled. Then, in a given ergodic MDP, both Q^A and Q^B as updated by Double Q-learning as described in Algorithm 1 will converge to the optimal value function Q^* as given in the Bellman optimality equation (2) with probability one if an infinite number of experiences in the form of rewards and state transitions for each state action pair are given by a proper learning policy. The additional conditions are: 1) The MDP is finite, i.e. $|S \times A| < \infty$. 2) $\gamma \in [0, 1)$. 3) The Q values are stored in a lookup table. 4) Both Q^A and Q^B receive an infinite number of updates. 5) $\alpha_t(s, a) \in [0, 1]$, $\sum_t \alpha_t(s, a) = \infty$, $\sum_t (\alpha_t(s, a))^2 < \infty$ w.p.1, and $\forall (s, a) \neq (s_t, a_t) : \alpha_t(s, a) = 0$. 6) $\forall s, a, s' : Var\{R_{sa}^{s'}\} < \infty$.*

Algorithm 1 Double Q-learning

1: Initialize Q^A,Q^B,s
2: **repeat**
3: Choose a, based on $Q^A(s, \cdot)$ and $Q^B(s, \cdot)$, observe r, s'
4: Choose (e.g. random) either UPDATE(A) or UPDATE(B)
5: **if** UPDATE(A) **then**
6: Define $a^* = \arg\max_a Q^A(s', a)$
7: $Q^A(s, a) \leftarrow Q^A(s, a) + \alpha(s, a)\left(r + \gamma Q^B(s', a^*) - Q^A(s, a)\right)$
8: **else if** UPDATE(B) **then**
9: Define $b^* = \arg\max_a Q^B(s', a)$
10: $Q^B(s, a) \leftarrow Q^B(s, a) + \alpha(s, a)(r + \gamma Q^A(s', b^*) - Q^B(s, a))$
11: **end if**
12: $s \leftarrow s'$
13: **until** end

Conditions

1. The MDP is finite, $i.e. |S \times A| < \infty$
2. $\gamma \in [0,1)$
3. The Q values are stored in a lookup table.
4. Both Q^A and Q^B receive an infinite number of updates.
5. $\alpha_t(s,a) \in [0,1], \Sigma_t \alpha_t(s,a) = \infty, \Sigma_t \alpha_t(s,a)^2 < \infty \; w.p.1 \; and \; \forall (s,a) \neq (s_t, a_t): \alpha_t(s,a) = 0$
6. $\forall s, a, s': Var\{R_{sa}^{s'}\} < \infty$

Proof.

Use lemma 2 to prove in Update A, Q^A will converge to Q^*:

Let $P_t = \{Q_0^A, Q_0^B, s_0, a_0, r_1, s_1, \dots, s_t, a_t\}$, $X = S \times A$, $\Delta_t = Q_t^A - Q^*$

$Q_{t+1}^A(s_t, a_t) = Q_t^A(s_t, a_t) + \alpha(s_t, a_t)(r_t + \gamma Q_t^B(s_{t+1}, a^*) - Q^A(s_t, a_t))$

$\Rightarrow Q_{t+1}^A(s_t, a_t) = (1 - \alpha(s_t, a_t))Q_t^A(s_t, a_t) + \alpha(s_t, a_t)(r_t + \gamma Q_t^B(s_{t+1}, a^*))$

subtract $Q^*(s_t, a_t)$ from both sides of the equation:

$\Rightarrow \Delta_{t+1}(s_t, a_t) = (1 - \alpha(s_t, a_t))\Delta_t(s_t, a_t) + \alpha(s_t, a_t)(r_t + \gamma Q_t^B(s_{t+1}, a^*) - Q^*(s_t, a_t))$

Let $\zeta_t = \alpha_t$, $F_t(s_t, a_t) = r_t + \gamma Q_t^B(s_{t+1}, a^*) - Q_t^*(s_t, a_t)$,

$a^* = \underset{a}{\operatorname{argmax}}\, Q^A(s_{t+1}, a^*)$, $\gamma = \kappa$

Check the four conditions in lemma 2 is satisfied:

1. Satisfied by the condition 1 of theorem 1.
2. Satisfied by the condition 5 of theorem 1.
3. Let $F_t^Q(s_t, a_t) = r_t + \gamma Q_t^A(s_{t+1}, a^*) - Q_t^*(s_t, a_t)$, the value of F_t in Q-learning
   $F_t(s_t, a_t) = F_t^Q(s_t, a_t) + \gamma(Q_t^B(s_{t+1}, a^*) - Q_t^A(s_{t+1}, a^*))$,
   Let $\Delta^{BA} = Q_t^B(s_{t+1}, a^*) - Q_t^A(s_{t+1}, a^*)$

   Show that $\left|\left|E\{F_t^Q(s_t, a_t)|P_t\}\right|\right|_\infty \leq \gamma \left|\left|\Delta_t\right|\right|_\infty$ first:

   $\left|\left|E\{F_t^Q(s_t, a_t)|P_t\}\right|\right|_\infty = \left|\left|E\{r_t + \gamma Q_t^A(s_{t+1}, a^*) - Q^*(s_t, a_t)\}\right|\right|_\infty$

$$= \max_{a_t} \left| E\left\{ r_t + \gamma \max_a Q_t^A(s_{t+1}, a) - Q^*(s_t, a_t) \right\} \right|$$

The update for Q^* *is*: $Q_{t+1}^*(s,a) = (1-\alpha)Q_t^*(s,a) + \alpha(r + \gamma \cdot \max_a Q_t^*(s',a))$,

$\because Q_{t+1}^* = Q_t^* \ \therefore Q^*(s,a) = r + \gamma \cdot \max_a Q^*(s',a))$

$$= \max_{a_t} \left| E\left\{ r_t + \gamma \max_{a_t} Q_t^A(s_{t+1}, a_t) - (r_t + \gamma \cdot \max_{a_t} Q^*(s_{t+1}, a_t)) \right\} \right|$$

$$= \gamma \max_{a_t} \left| E\left\{ \max_{a_t} Q_t^A(s_{t+1}, a_t) - \max_{a_t} Q^*(s_{t+1}, a_t) \right\} \right|$$

$$\leq \gamma \max_{a_t} |E\{Q_t^A(s_{t+1}, a_t) - Q^*(s_{t+1}, a_t)\}|$$

$$\leq \gamma \max_a E|Q_t^A(s_{t+1}, a_t) - Q^*(s_{t+1}, a_t)| \ \text{(Jensen's inequality)}$$

$$\leq \gamma \left|\left| Q_t^A - Q^* \right|\right|_\infty = \gamma \left|\left| \Delta_t \right|\right|_\infty$$

Then show that $\Delta^{BA} = Q_t^B(s_{t+1}, a^*) - Q_t^A(s_{t+1}, a^*)$ converges to 0:

Let $F_t^B(s_t, a_t) = r_t + \gamma Q_t^A(s_{t+1}, b^*) - Q_t^B(s_t, a_t)$,

$F_t^A(s_t, a_t) = r_t + \gamma Q_t^B(s_{t+1}, a^*) - Q_t^A(s_t, a_t)$

For Update A or Update B,

$\Delta_{t+1}^{BA}(s_t, a_t) = \Delta_t^{BA}(s_t, a_t) + \alpha(s_t, a_t) F_t^B(s_t, a_t)$ or

$\Delta_{t+1}^{BA}(s_t, a_t) = \Delta_t^{BA}(s_t, a_t) - \alpha(s_t, a_t) F_t^A(s_t, a_t)$

$E\{\Delta_{t+1}^{\mathrm{BA}}(\mathrm{s_t}, \mathrm{a_t})|\mathrm{P_t}\} = \Delta_t^{\mathrm{BA}}(s_t, a_t) + E[\alpha(s_t, a_t)(F_t^B(s_t, a_t) - F_t^A(s_t, a_t))|P_t]$

Let $\zeta^{BA} = \frac{1}{2}\alpha_t$, $F^{BA}(s_t, a_t) = \gamma\left(Q_t^A(s_t, b^*) - Q_t^B(s_t, a^*)\right) \geq 0 (\because \textit{the symmetry of } A \textit{ and } B)$

$= \Delta_t^{\mathrm{BA}}(s_t, a_t) + \zeta^{BA}(s_t, a_t) E[F^{BA}(s_t, a_t) - \Delta_t^{\mathrm{BA}}(s_t, a_t)|P_t]$

$= \left(1 - \zeta^{BA}(s_t, a_t)\right)\Delta_t^{\mathrm{BA}}(s_t, a_t) + \zeta^{BA}(s_t, a_t) E[F^{BA}(s_t, a_t)|P_t]$

$\because E[F^{BA}(s_t, a_t)|P_t] = \gamma E\{Q_t^A(s_{t+1}, b^*) - Q_t^B(s_{t+1}, a^*)|P_t\} \leq \gamma E\{Q_t^A(s_{t+1}, a^*) - Q_t^B(s_{t+1}, a^*)|P_t\}$

$\leq \gamma ||\Delta_t^{BA}||$

By lemma 2, $\Delta_t^{BA} \to 0$

Because $F_t(s_t, a_t) = F_t^Q(s_t, a_t) + \gamma\left(Q_t^B(s_{t+1}, a^*) - Q_t^A(s_{t+1}, a^*)\right)$,

$$\left|\left| E\{F_t(s_t, a_t)|P_t\} \right|\right|_\infty = \left|\left| E\{F_t^Q(s_t, a_t) + \gamma \Delta_t^{BA}\} \right|\right|_\infty \leq \gamma \left|\left| \Delta_t \right|\right|_\infty$$

This condition is satisfied.

4. From the condition 6 of theorem 1, $Var(R_{sa}^{s'})$ is bounded, so $Var(r_t)$, $Var(Q_t^B)$, $Var(Q_t^*)$ are bounded, too.

   $F_t(s_t, a_t) = r_t + \gamma Q_t^B(s_{t+1}, a^*) - Q_t^*(s_t, a_t)$

   $\Rightarrow Var\left(F_t(s_t, a_t)\right) = Var(r_t) + \gamma Var(Q_t^B) - Var(Q_t^*) < \infty$, this condition is satisfied.

By lemma 2, Q^A will converge to Q^*

Because of the symmetry of A and B, Q^B will converge to Q^*, so both Q^A and Q^B will converge to the optimal value.