MS&E338 Reinforcement Learning	Lecture 10 - May-6-2020

Q-Learning

Lecturer: Ben Van Roy *Scribe: Yu Gu and George Gui*

1 Q-learning and RTVI

Q-learning is an update rule that can be used to update Q value given some data. Given (h, a, r, h'), we can update Q value by:

$$\tilde{Q}(s, a) \leftarrow \tilde{Q}(s, a) + \alpha(r + \max_{a' \in \mathcal{A}} \tilde{Q}(s', a') - \tilde{Q}(s, a))$$

Note: if h' is the terminal history, just make $\max_{a' \in \mathcal{A}} \tilde{Q}(s', a') = 0$.
Recall another update way of RTVI we've discussed in the previous lecture. We can rewrite the RTVI with the similar form:

$$\tilde{Q}(s, a) \leftarrow \tilde{Q}(s, a) + \alpha(r_{ah} + \sum_{o \in \emptyset} \rho(o|h, a) \max_{a' \in \mathcal{A}} \tilde{Q}(f(s, a, o), a') - \tilde{Q}(s, a))$$

If $\alpha = 1$, then this is the basic form RVTI. Comparing these two equations, Q-learning is replacing the expectation of right hand side to a simple transition sample, which is more efficient in planning. Also, Q-learning is feasible to apply in RL, while RTVI is not, because we do not know the observation probabilities.

2 Stochastic Approximation

2.1 Basic Version

Given a sequence of i.i.d vectors: $X_1, X_2, ... \in \mathbb{R}^N$, where $E[X_k] = \bar{x}$, and $E[||X||_2^2] < \infty$. Let θ be an estimator for $\bar{x}$. One way is direct averaging:

$$\theta_{k+1} = \frac{1}{k+1} \sum_{i=1}^{k+1} x_i \to \bar{x}$$

The other way is by stochastic updating, which we will focus in this section.

$$\begin{aligned}\theta_{k+1} &= \theta_k + \frac{1}{k+1}(x_{k+1} - \theta_k) \\ &= \theta_k + \alpha_k(x_{k+1} - \theta_k)\end{aligned}$$

Lemma 1 (Sufficient condition for convergence)**.** *The above process converges to $\bar{x}$ if $\sum \alpha_k = \infty$ and $\sum \alpha_k^2 < \infty$.*

Notes: α_k can be stochastic.

Intuition of stochastic approximation x_{k+1} can be understood as a noisy version of $\bar{x}$:

$$x_{k+1} - \theta_k = \bar{x} - \theta_k + \omega_{k+1}$$

where $\omega_{k+1} = x_{k+1} - \bar{x}$ and $E[\omega_{k+1}|\theta_k] = 0$.

2.2 General Version

Let F be a contraction mapping w.r.t. Euclidian norm:

$$||F(\theta) - F(\bar{\theta})||_2 \leq \gamma ||\theta - \bar{\theta}||_2$$
$$\gamma \in [0, 1)$$

Let θ^* be the fixed point:

$$\theta^* = F(\theta^*)$$

Theorem 2 (Supermartingale convergence theorem)**.** *There are several versions.*

V1: If $X_1, X_2, ... \in \mathbb{R}_+$ and $E_k[X_{k+1}] \leq X_k$, then X_k converges w.p 1.

V2: If $X_1, X_2, ..., Y_1, Y_2, ... \in \mathbb{R}_+$ and $\sum Y_k < \infty$ and $E_k[X_{k+1}] \leq X_k + Y_k$, then X_k converges w.p. 1.

V3: If $X_1, X_2, ..., Y_1, Y_2, ..., Z_1, Z_2 \in \mathbb{R}_+$ and $\sum Y_k < \infty$ and $E_k[X_{k+1}] \leq X_k + Y_k - Z_k$, then X_k converges w.p. 1 and $\sum_k Z_k < \infty$.

Theorem 3. *Suppose $\theta_{k+1} = \theta_k + \alpha_k(F(\theta_k) - \theta_k + w_{k+1})$, $E[w_{k+1}|\theta_k] = 0$, $E[||w_{k+1}||_2^2|\theta_k] \leq c$ and $\alpha_k > 0$, $\sum \alpha_k = \infty$, $\sum \alpha_k^2 < \infty$ then $\theta_k \to \theta^*$ with probability 1.*

Proof. Let $d_k = ||\theta^* - \theta_k||_2^2$, then:

$$\begin{aligned}
E[d_{k+1}|\theta_k] &= E[||\theta^* - (\theta_k + \alpha_k(F(\theta_k) - \theta_k + w_{k+1}))||_2^2|\theta_k] && (1)\\
&= ||\theta^* - (\theta_k + \alpha_k(F(\theta_k) - \theta_k))||_2^2 + \alpha_k^2 E[||w_{k+1}||_2^2] && (2)\\
&\leq ||\theta^* - (\theta_k + \alpha_k(F(\theta_k) - \theta_k))||_2^2 + \alpha_k^2 c && (3)\\
&= d_k - 2\alpha_k(\theta^* - \theta_k)^T(F(\theta_k) - \theta_k) + \alpha_k^2||F(\theta_k) - \theta_k||_2^2 + \alpha_k^2 c && (4)\\
&\leq d_k - 2\alpha_k(\theta^* - \theta_k)^T(F(\theta_k) - \theta_k) + \alpha_k^2(||F(\theta_k) - \theta^*||_2 + ||F(\theta^*) - \theta_k||_2)^2 + \alpha_k^2 c && (5)\\
&\leq d_k - 2\alpha_k(1-\gamma)d_k + \alpha_k^2(1+\gamma)^2 d_k + \alpha_k^2 c && (6)
\end{aligned}$$

Additional notes from step (5) to step (6):

$$\begin{aligned}
(\theta^* - \theta_k)^T(F(\theta_k) - \theta_k) &= -(\theta^* - \theta_k)^T(\theta^* - F(\theta_k)) + d_k\\
&\geq -||\theta^* - \theta_k||_2||\theta^* - F(\theta_k)||_2 + d_k\\
&\geq -||\theta^* - \theta_k||_2 \cdot \gamma||\theta^* - \theta_k||_2 + d_k\\
&\geq (1-\gamma)d_k
\end{aligned}$$

Let $X_k = d_k$, $Y_k = \alpha_k^2 c$, and $Z_k = 2\alpha_k(1-\gamma)d_k - \alpha_k^2(1+\gamma)^2 d_k$. We know that $X_k \geq 0, Y_k \geq 0, \sum Y_k = \sum \alpha_k^2 c < \infty$ and since the second term of Z_k contains α_k^2 where $\alpha_k > 0$ and it is decreasing over time, there exists a K such that $Z_k \geq 0$ for any $k \geq K$. Then we can apply the supermartingale convergence theorem, which implies that d_k converges and $\sum Z_k < \infty$. Now if d_k does not converge to 0, since $\sum a_k = \infty$, $\sum a_k d_k$ will go to infinity, then $\sum Z_k = \infty$. Thus, d_k must satisfy to $d_k \to 0$. $\square$

2.3 Connection to Q-Learning

Recall that the update rule of Q-learning is:

$$\tilde{Q}(s,a) \leftarrow \tilde{Q}(s,a) + \alpha(r + \max_{a' \in \mathcal{A}} \tilde{Q}(s',a') - \tilde{Q}(s,a))$$

One way to think about it is that sampling history using relevance weights ν:

$$(\tilde{F}\tilde{Q})(s,a) = \frac{1}{\nu(H_s)} \sum_{h \in H_s} \nu(h)(\bar{r}_{ah} + \sum_{o \in O} \rho(o|h,a) \max_{a'} \tilde{Q}(f(s,a),a')$$

$$\tilde{Q} \leftarrow \tilde{Q} + \alpha(\tilde{F}\tilde{Q} - \tilde{Q} + noise) \tag{7}$$

We can think of Q-learning update in terms of stochastic approximation, then we have 0 mean noise.
Note: $\tilde{F}$ is not a contraction mapping w.r.t Euclidian norm, but w.r.t. a weighted maximum norm.