5.6 Q-Learning

Saturday, February 13, 2021 1:52 PM

Q-Learning directly updates estimates of Q-factors associated with an optimal policy.

Let us define the optimal Q-factor,

$$Q^*(i,u) = \sum_{j=0}^{n} p_{ij}(u)\left(g(i,u,j) + J^*(j)\right), \quad i = 1,2,\dots,n$$

$$J^*(i) = \min_{u \in U(i)} Q^*(i,u)$$

Combining the above two, we get,

$$Q^*(i,u) = \sum_{j=0}^{n} p_{ij}(u)\left(g(i,u,j) + \min_{v \in U(j)} Q^*(j,v)\right).$$

Using the uniqueness of Bellman's eq,

$$\min_{u \in U(i)} Q(i,u) = \min_{u \in U(i)} Q^*(i,u).$$

In terms of Q factors,

$$Q(i,u) = \sum_{j=0}^{n} p_{ij}(u)\left(g(i,u,j) + \min_{v \in U(j)} Q(j,v)\right).$$

More generally,

$$Q(i,u) = (1-\gamma)Q(i,u) + \gamma \sum_{j=0}^{n} p_{ij}(u)\left(g(i,u,j) + \min_{v \in U(j)} Q(j,v)\right).$$

Q-Learning is an approximate version of the above update where the expectation is replaced by a single sample,

$$Q(i,u) = (1-\gamma)Q(i,u) + \gamma\left(g(i,u,j) + \min_{v \in U(j)} Q(j,v)\right).$$

<u>The Convergence of Q-Learning</u> –

We first write a general statement,

$$Q_{t+1}(i,u) = (1-\gamma_t(i,u))\,Q_t(i,u) + \gamma_t(i,u)\left(g(i,u,\bar{i}) + \min_{v \in U(\bar{i})} Q_t(\bar{i},v)\right)$$

$\gamma_t(i,u) = 0$ for $t \notin T^{i,u}$. and $Q_t(0,u) = 0 \ \forall t$.

Proposition-5: Suppose that –

$$\sum_{t=0}^{\infty} \gamma_t(i,u) = \infty, \quad \sum_{t=0}^{\infty} \gamma_t^2(i,u) < \infty, \quad \forall i, u \in U(i).$$

Then $Q(i,u)$ converges to $Q^*(i,u)$ with probability 1, in each of the following cases –

(a) If all policies are proper.

(b) If Assumptions 1 and 2 hold and if $Q_t(i,u)$ is bounded with probability 1.

Proof: Consider a mapping H,

$$(HQ)(i,u) = \sum_{j=0}^{n} p_{ij}(u)\left(g(i,u,j) + \min_{v \in U(j)} Q(j,v)\right),$$

Then Q-Learning is of the form,

$$Q_{t+1}(i,u) = (1-\gamma_t(i,u))\,Q_t(i,u) + \gamma_t(i,u)\left((HQ_t)(i,u) + w_t(i,u)\right).$$

where $w_t(i,u) = g(i,u,\bar{i}) + \min_{v \in U(\bar{i})} Q_t(\bar{i},v) - \sum_{j=0}^{n} p_{ij}(u)\left(g(i,u,j) + \min_{v \in U(j)} Q_t(j,v)\right).$

Note that $E[w_t(i,u) \mid \mathcal{F}_t] = 0$ and $E[w_t^2(i,u) \mid \mathcal{F}_t] \le K\left(1 + \max_{j,v} Q_t^2(j,v)\right).$

(a) We first consider the case when all policies are proper,

According to Chapter-2 there exist $\xi(i) > 0, i \neq 0$, and a scalar $\beta \in [0,1)$,

$$\sum_{j=1}^{n} p_{ij}(u)\,\xi(j) \le \beta\,\xi(i), \quad \forall i,\ u \in U(i).$$

For any two vectors Q and $\bar{Q}$,

$$\left|(HQ)(i,u) - (H\bar{Q})(i,u)\right| \le \sum_{j=1}^{n} p_{ij}(u)\left|\min_{v \in U(j)} Q(j,v) - \min_{v \in U(j)} \bar{Q}(j,v)\right|,$$

$$\le \sum_{j=1}^{n} p_{ij}(u) \max_{v \in U(j)} \left|Q(j,v) - \bar{Q}(j,v)\right|,$$

$$\le \sum_{j=1}^{n} p_{ij}(u)\, \|Q - \bar{Q}\|_\xi\, \xi(j),$$

$$\le \beta\, \|Q - \bar{Q}\|_\xi\, \xi(i).$$

Dividing both sides by $\xi(i)$ and taking min over $\forall i$ and $u \in U(i)$,

$$\|HQ - H\bar{Q}\|_\xi \le \beta\, \|Q - \bar{Q}\|_\xi$$

$\therefore$, H is a weighted max norm contraction. Convergence of Q-Learning follows from Proposition-4 of Chapter-4.

(b) In the second case, we impose assumptions 1 and 2 and remove the case of all proper policies,

Clearly H is a monotone mapping,

$$Q \le \bar{Q} \Rightarrow HQ \le H\bar{Q},$$

Also, for 'r' a positive scalar and 'e' being ~~a~~ vector of all components equal to 1,

$$HQ - re \le H(Q - re) \le H(Q + re) \le HQ + re.$$

Finally, $HQ = Q$ has a unique solution and Q_t converges to Q^* with probability 1 based on Proposition-6 from Chapter 4. ■

<u>Q-Learning for Discounted Problems.</u>

We may not convert the discounted problem to a stochastic shortest path problem. This can be handled directly,

$$Q(i,u) = (1-\gamma)Q(i,u) + \gamma\left(g(i,u,j) + \alpha \min_{v \in U(j)} Q(j,v)\right).$$

Here, 'α' is the discount factor.

Similar convergence results are observed as for the stochastic shortest-path case.

x ———————— x ———————— x