Convergence of Q-Learning

Q-learning (Watkins '89)

Q-learning (off-policy)

Let $\tilde{Q}$ be the estimate of Q. Given an experience (current state $x_k = i$, current action $a_t = u$, received reward $r(i,u)$, and next state $x_{k+1} = j$), Q-learning updates $\tilde{Q}$ as follows:

$$\begin{aligned}\tilde{Q}(i,u) \longleftarrow & (1-\beta_k)\tilde{Q}(i,u) + \beta_k(r(i,u) + \alpha \max_v \tilde{Q}(j,v)) \\ = & \tilde{Q}(i,u) + \beta_k \left(r(i,u) + \alpha \max_v \tilde{Q}(j,v) - \tilde{Q}(i,u)\right)\end{aligned}$$

A Modified Q-Learning Algorithm

To simplify the proof, Consider a modified Q-learning such that we use N experiences at each iteration.

- $\gamma_{i,u}$ is the faction of experiences such that the current state is i and current action is u. We assume $\gamma_{i,u} > 0$ for any $(i, u) \in (\mathcal{S}, \mathcal{A})$.
- $\hat{p}_{ij}(u)$ is the fraction of experiences that the next state is j given the current state and action are (i, u).
- Assume $r(i, u)$ is deterministic.

Remark

The modified algorithm with learning rate β approximates the original Q-learning with learning rate $\frac{\beta}{N}$. When β is small, the Q-values does not change much over N updates under the original Q-learning.

Convergence (Pseudo Contraction Mapping)

We will compare $\|Q_k - Q^*\|_\infty$ and $\|Q_{k+1} - Q^*\|_\infty$. Note that

$$Q_{k+1}(i,u) = (1-\beta)Q_k(i,u) + \beta\left(r(i,u) + \alpha\sum_j \hat{p}_{ij}(u)\max_v Q_k(j,v)\right)$$

$$Q^*(i,u) = (1-\beta)Q^*(i,u) + \beta\left(r(i,u) + \alpha\sum_j p_{ij}(u)\max_v Q^*(j,v)\right),$$

so

$$Q_{k+1}(i,u) - Q^*(i,u) = (1-\beta)\left(Q_k(i,u) - Q^*(i,u)\right) + \alpha\beta\left(\sum_j \hat{p}_{ij}(u)\max_v Q_k(j,v) - \sum_j p_{ij}(u)\max_v Q^*(j,v)\right).$$

Convergence (Pseudo Contraction Mapping)

$$
\begin{aligned}
& Q_{k+1}(i,u) - Q^*(i,u) \\
=& (1-\beta)\left(Q_k(i,u) - Q^*(i,u)\right) \\
& + \alpha\beta\left(\sum_j p_{ij}(u) \max_v Q_k(j,v) - \sum_j p_{ij}(u) \max_v Q^*(j,v)\right) \\
& + \alpha\beta\left(\sum_j \hat{p}_{ij}(u) \max_v Q_k(j,v) - \sum_j p_{ij}(u) \max_v Q_k(j,v)\right) \\
=& (1-\beta)\left(Q_k(i,u) - Q^*(i,u)\right) + \alpha\beta \sum_j \left(\hat{p}_{ij}(u) - p_{ij}(u)\right) \max_v Q_k(j,v) \\
& + \alpha\beta \sum_j p_{ij}(u) \left(\max_v Q_k(j,v) - \max_v Q^*(j,v)\right).
\end{aligned}
$$

$E[\hat{P}_{ij}(u)] = P_{ij}(u)$

$E[① \times W_k] = 0$

$E[② \times W_k] = 0$

Convergence (Pseudo Contraction Mapping)

$E[① \times W_k] = E[E[① \times W_k | Q_k]] = 0$

$$Q_{k+1}(i,u) - Q^*(i,u)$$
$$= (1-\beta)\left(Q_k(i,u) - Q^*(i,u)\right) + \alpha\beta \sum_j p_{ij}(u)\left(\max_v Q_k(j,v) - \max_v Q^*(j,v)\right) + W_k,$$

① ②

where $W_k = \alpha\beta \sum_j (\hat{p}_{ij}(u) - p_{ij}(u)) \max_v Q_k(j,v)$. Note that $E[\hat{p}_{ij}(u)] = p_{ij}(u)$ so $E[W_k | Q_k] = 0$.

$\Rightarrow E[W_k] = 0$

$E\left[\alpha\beta \sum_i (\hat{P}_{ij}(u) - P_{ij}(u)) \max_v Q_k(j,v) \,\middle|\, Q_k\right]$

constant

$= \alpha\beta \sum_i E[\hat{P}_{ij}(u) - P_{ij}(u) | Q_k] \max_v Q_k(j,v)$

$$\hat{P}_{ij}(u) = \frac{\sum_{m=1}^{N} 1(x_m, u_m, x_{m+1}) = (i, u, j)}{\sum_{m=1}^{N} 1(x_m, u_m, x_{m+1}) = (i, u, *)}$$

$$E[1(x_m, u_m, x_{m+1}) = (i, u, j)]$$

$$= P_{ij}(u) \cdot P_\mu(i, u)$$

$$P_{ij}(u) = P(x' = j \mid x = i, u = u)$$

$$E[1(x_m, u_m, x_{m+1}) = (i, u, *)] = P_\mu(i, u)$$

$|Q_k(i,u) - Q^*(i,u)| \le \|Q_k - Q^*\|_\infty$

Convergence (Pseudo Contraction Mapping)

$$E\left[(Q_{k+1}(i,u) - Q^*(i,u))^2\right]$$

$$= (1-\beta)^2 E\left[(Q_k(i,u) - Q^*(i,u))^2\right] +$$

$$\alpha^2\beta^2 E\left[\left(\sum_j p_{ij}(u)\left(\max_v Q_k(j,v) - \max_v Q^*(j,v)\right)\right)^2\right] + E[W_k^2] +$$

$$2(1-\beta)\alpha\beta E\left[(Q_k(i,u) - Q^*(i,u))\sum_j p_{ij}(u)\left(\max_v Q_k(j,v) - \max_v Q^*(j,v)\right)\right],$$

$\max |Q_k(j,v) - Q^*(j,v)| = \|Q_k - Q^*\|_\infty$

Convergence (Pseudo Contraction Mapping)

We note that

$$\left|\max_v Q_k(j,v) - \max_v Q^*(j,v)\right| \le \max_v |Q_k(j,v) - Q^*(j,v)| \quad \text{(Homework)}$$

so for any (i,u), we have

$$\begin{aligned} & E\left[(Q_{k+1}(i,u) - Q^*(i,u))^2\right] \\ \le & (1-\beta)^2 E\left[\|Q_k - Q^*\|_\infty^2\right] + \\ & \alpha^2\beta^2 E\left[\|Q_k - Q^*\|_\infty^2\right] + E[W_k^2] + 2(1-\beta)\alpha\beta E\left[\|Q_k - Q^*\|_\infty^2\right] \end{aligned}$$

Convergence (Pseudo Contraction Mapping)

The inequality above holds for any (i, u), so we have

$$
\begin{aligned}
& E\left[\|Q_{k+1} - Q^*\|_\infty^2\right] \\
& \leq \left((1-\beta)^2 + \alpha^2\beta^2 + 2(1-\beta)\alpha\beta\right) E\left[\|Q_k - Q^*\|_\infty^2\right] + E[W_k^2] \\
& \leq \left(1 - \beta + \beta^2\right) E\left[\|Q_k - Q^*\|_\infty^2\right] + E[W_k^2]
\end{aligned}
$$

Note that

$$
E[W_k^2] \leq \frac{c\beta^2 \log^2 N}{N}
$$

for some $c > 0$, where c is a constant independent of N, β and k. (details in the textbook)

Convergence (Pseudo Contraction Mapping)

Define $\rho = 1 - \beta + \beta^2$. Therefore,

$$
\begin{aligned}
E\left[\|Q_{k+1} - Q^*\|_\infty^2\right] \leq & \rho E\left[\|Q_k - Q^*\|_\infty^2\right] + \frac{c\beta^2 \log^2 N}{N} \\
\leq & \rho^2 E\left[\|Q_{k-1} - Q^*\|_\infty\right] + (1+\rho)\frac{c\beta^2 \log^2 N}{N} \\
\leq & \rho^{k+1} E\left[\|Q_0 - Q^*\|_\infty\right] + \left(\sum_{m=0}^{k} \rho^k\right) \frac{c\beta^2 \log^2 N}{N} \\
\leq & \rho^{k+1} E\left[\|Q_0 - Q^*\|_\infty\right] + \frac{c\beta^2 \log^2 N}{(1-\rho)N} \\
= & (1 - \beta + \beta^2))^{k+1} E\left[\|Q_0 - Q^*\|_\infty\right] + \frac{c\beta \log^2 N}{(1-\rho)N}
\end{aligned}
$$