Q Learning

The goal of reinforcement learning is to compute an optimal policy π^*. As previously explained, the optimal policy can be easily computed from $Q(s,a)$. The following algorithm can be used to compute $Q(s,a)$.

0. Start with $\hat{Q}$ as an estimate for Q. For example, for all states s and actions a set $\hat{Q}(s,a) = 0$.

1. Repeat:

- **1.1** Select a state s and an action a.
- **1.2** Let $r = r(s,a)$ be the reward.
- **1.3** Let $s' = \delta(s,a)$ be the state one moves to from s if action a is taken.
- **1.4** Update:

$$\hat{Q}(s,a) = r + \gamma \max_{a'} \hat{Q}(s',a')$$

The method of selecting a state and an action at 1.1 is arbitrary. The algorithm is guaranteed to converge to the right Q if each (s,a) pair is visited infinitely many times.

Convergence

Let t be the iteration number. In terms of t the iteration in 1.4 can be written as:

$$\hat{Q}_{t+1}(s,a) = r + \gamma \max_{a'} \hat{Q}_t(s',a')$$

Consider the following error terms:

$$\begin{aligned}
e_t(s,a) &= |\hat{Q}_t(s,a) - Q(s,a)| \\
e_t(s) &= \max_a \ e_t(s,a) \\
e_t &= \max_s \ e_t(s) \qquad\qquad = \max_{s,a} |\hat{Q}_t(s,a) - Q(s,a)|
\end{aligned}$$

At iteration t the algorithm selects s,a, computes r,s', and updates $\hat{Q}(s,a)$. The key technical lemma here states that:

$$\boxed{e_{t+1}(s,a) \leq \gamma \ e_t}$$

Proof:

$$\begin{aligned}
e_{t+1}(s,a) &= |\hat{Q}_{t+1}(s,a) - Q(s,a)| \\
&= |(r + \gamma(\max_{a'} \hat{Q}_t(s',a')) - (r + \gamma(\max_{a'} Q(s',a'))| \\
&= \gamma |\max_{a'} \hat{Q}_t(s',a') - \max_{a'} Q(s',a')| \\
&\leq \gamma \ \max_{a'} |\hat{Q}_t(s',a') - Q(s',a')| \\
&= \gamma \ \max_{a'} \ e_t(s',a') \\
&\leq \gamma \ e_t
\end{aligned}$$

The inequality between the third and the fourth line is:

$$|\max_a f(a) - \max_a g(a)| \leq \max_a |f(a) - g(a)|$$

It follows from the triangle inequality in the l_∞ norm, and can also be easily proved directly.

The value of $e_t(s,a)$ is not guaranteed to decrease as a function of t, and may even increase. But as shown below the value of e_t cannot increase. It is guaranteed to decrease by a factor of γ in a cycle where all (s,a) pairs are visited. The first observation is that for *all* pairs s,a:

$$e_{t+j}(s,a) \le e_t \quad \text{for } j \ge 0$$

For the proof it is enough to show that it holds for $j = 1$, and then apply induction. The case $j = 1$ follows from the lemma. Now consider the relation between the error at iteration t and the error at iteration $t + \Delta$. We can make the following observations:

1. $e_{t+j} \le e_t$ for $j = 0, 1, \ldots, \Delta$.

2. $e_{t+\Delta}(s,a) \le \gamma\ e_t$ for all (s,a) that were visited in the Δ iterations.

3. $e_{t+\Delta} \le \gamma\ e_t$ if all (s,a) were visited in the Δ iterations.

Conclusion: $\hat{Q}(s,a)$ approaches $Q(s,a)$ if each (s,a) is visited infinitely many times.