BELLMAN EQUATION

UZAY CETIN

1. Reinforcement Learning

1.1. **Bellman Equation.** How to measure $Q(s,a)$, the quality/goodness of an action a given the state s we are in? If we know it for every possible action a, we can maximize our outcome. Bellman equation, gives us a recursive formula for that.

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

Which means that, quality of action a in current state s equals to an immediate reward $R(s,a)$ plus a γ discounted $max_{a'} Q(s',a')$ maximum quality we can get with a new action a' from the new state s'. That is,

$$\text{Current quality} = \text{immediate reward} + \text{discounted Future Quality}$$

1.2. **Temporal difference.** In an ideal case, after many iterations, Bellman Equation will be true. Initially there will be non-zero temporal difference $TD(s,a)$.

$$TD(s,a) = [R(s,a) + \gamma max_{a'} Q(s',a')] - Q(s,a) \neq 0$$

1.3. **Q-Learning.** In the environment, there are non-zero rewards and initially $Q(s,a) = 0$ for all states and actions. During an episode,

- Based on our *predictions* $Q(s,a)$, we choose max of possible actions.
- After each action, we learn a new *target* value $R(s,a) + \gamma max_{a'} Q(s',a')$

That is, for $Q(s,a)$

- Previously, we had collected *prediction* values $Q(s,a) = Q(s,a)$
- Now a new *target* value comes, $Q(s,a) = R(s,a) + \gamma max_{a'} Q(s',a')$

The question is how to combine, previously collected prediction values with a new target value?

1.4. **Online Learning.** Previous question can be converted a simple moving average problem,

$$
\begin{aligned}
\overline{\mathbf{x_t}} &= \frac{1}{t}\sum_{i}^{t} x_i \\
&= \frac{1}{t}(x_1 + x_2 + \ldots + x_{t-1} + x_t) \\
&= \frac{1}{t}((t-1)\overline{\mathbf{x_{t-1}}} + x_t) \\
&= \overline{\mathbf{x_{t-1}}} + \frac{1}{t}(x_t - \overline{\mathbf{x_{t-1}}})
\end{aligned} \tag{1}
$$

Here, lets write $\alpha = \frac{1}{t}$, $\overline{\mathbf{x_t}} = Q_t(s,a)$ and $x_t = [R(s,a) + \gamma max_{a'} Q(s',a')]$ We get

$$
Q_t(s,a) = Q_{t-1}(s,a) + \alpha([R(s,a) + \gamma max_{a'} Q(s',a')] - Q_{t-1}(s,a)) \tag{2}
$$