Junior Antunes Koch
E-mail: junioru.desu@gmail.com

Q-Learning

This is classified as an Reinforcement Learning method which is based on learning due experimentation by the agent. There is no policy, it will be "discovered" by the agent as it "observes" the environment. In other words no model is needed in Q-Learning.

The subject to experimentation is called *agent*, it can ocupy a set of *states* and make a set os *actions*. When the agent make an action the states is changed and a reward is provided to it. With each transition a function (Q) measures the quality of the action based on the rewards. The agent then creates some kind of memory stored in the Q matrix. Every time a final state is reached an episode ends and there is nothing else for the agent to learn. A typical Q function is like this:

$$Q^{'}(s_t, a_t) = (1 - \alpha)Q(s_t, a_t) + \alpha(R_t + \gamma MAX(Q(s_{t+1}, a_{t+1}))), \tag{1}$$

where $Q^{'}(s_t, a_t)$ is the updated quality value, $Q(s_t, a_t)$ is the old quality value, α is the learning rate, γ is the exploration factor, R_t is the reward matrix and $MAX(Q(s_{t+1}, a_{t+1}))$ is the maximum value of the possible actions for the next state.

Let us check the meaning and how the above equation can be constructed. In each time step we can see the amount of information store in memory as

$$\frac{\partial Q}{\partial t} = R + \gamma \frac{\partial Q}{\partial x}, \tag{2}$$

where x is a point in space $X \equiv S \times A$ and $\frac{\partial Q}{\partial x}$ can be seen as the gradient of Q, a tangent vector of the space of configurations. The λ is a proportionality coeficient. It is possible to rewrite this equation taking $t = x^0$ and $x = x^1$,

$$\frac{\partial Q}{\partial x^0} - \gamma \frac{\partial Q}{\partial x^1} = R \tag{3}$$

or in a more general way considering Q as a real valued contravariant vector, where $M \equiv X \times T$ has a Minkowskian metric

$$\eta_{ab} = \begin{pmatrix} 1 & 0 \\ 0 & -\gamma^2 \end{pmatrix} \tag{4}$$

and

$$\partial_a = \begin{pmatrix} \partial_0 & 0 \\ 0 & -\gamma\partial_1 \end{pmatrix} \tag{5}$$

we have

$$\partial_a Q^a = R \text{ or } \vec{\nabla} \bullet \vec{Q} = R. \tag{6}$$

This way R represents the reward time and spatial evolution. The meaning of the above equation is that of a continuity equation and R works as an information source. Q^a is some kind of information field 1-tensor.

From this we can develop the whole information theory even with a general metric tensor to compute curvatures. By integration we obtain

$$\Delta Q = \int_T (R + \gamma \frac{\partial Q}{\partial x^1}) dx^0 \tag{7}$$

and doing it numerically

$$Q^{'} = Q + R\Delta t + \gamma \frac{\partial Q}{\partial x} \Delta t. \tag{8}$$

So the time step is linked to the learning rate and to get the same equation we add the term $(1 - \alpha)$ to control how much of the past stored infirmation we retain to the future.