# Reference on Reinforcement Learning

## Contents

**GLIE** GLIE stands for "Greedy in the Limit of Infinite Exploration". The learning policies in RL can be divided into two broad categories: a *decay exploration* strategy which become more and more greedy and *persistent exploration* which always maintain a fix exploration rate. The advantage of the first one is that we can eventually converge to the optimal policy. The second one may have the advantage always be adaptive but may not converge to the optimal. (In here, we talk about convergence in the sense that the behavior will become optimal. It is possible that some of the algorithm converge to the correct Q-value but still behave randomly with some probability by using persistent exploration strategy, Q-learning with fix $\epsilon$-greedy for example). We may want to consider this in the context of on-policy&off-policy.

If a *decay exploration* strategy has the following two charaters:

1. each action is executed infinitely often in every state that is visited infinitely often, and
2. in the limit, the learning policy is greedy with respect to the Q-value function with probability 1.

Than we can consider this decay exploration strategy GLIE. Some example of GLIE include Boltzmann Selection, $\epsilon$-greedy.

Detail see [1].

**On-policy&Off-policy** An RL algorithm can be essentially divided into two parts, the *learning policy* and *update rule*. The first one is a non-stationary policy that maps experience (state visited, action chosen, rewawrd received) to into a currently choice of action. The second part is how the algorithm uses experience to change its estimation of the optimal value function. In off-policy algorithm, the *update rules* doesn't have relationship with *learning policy*, that is the *update rules* doesn't care the what action agent take. Q-learning can be consider as the off-policy algorithm.

$$Q_{t+1}(s,a) = (1-\alpha)Q_t(s,a) + \alpha(r_t + \gamma \max_{a'} Q(s',a'))$$

We can see that the Q-value is update based on the $\max_{a'} Q(s',a')$, which doesn't depend on the action the agent was taking.

However, if we take a look of SARSA(0), which is very similar to Q-learning.

$$Q_{t+1}(s,a) = (1-\alpha)Q_t(s,a) + \alpha(r_t + \gamma Q(s',a'))$$

We can see the update is based on the Q-value of the next action of the agent. Thus it is an on-policy algorithm. The convergence condition are heavey depend on the *learning policy*, The Q-value of SARSA(0) can only converge to optimality in the limit only if the learning policy behavior optimal in the limit. The SARSA(0) and Q-learning will be same if we use greedy action selection strategy.

Detail see [1].

**Boltzmann Selection** sdfsf

**$\epsilon$-greedy** $\epsilon$-greedy is a common used exploration strategy for Model-free reinforcement learning. This strategy can be seen as a combination of greedy strategy and random strategy. Every time when agent choose an action to perform, it choose greedily with probability 1-$\epsilon$ (exploitation), and randomly with probability $\epsilon$ (exploration). However, based on the changes of $\epsilon$ value, this method have two versions. First, the $\epsilon$ value can decrease as the learning process goes on, this is the *decay exploration* method. Another one which are more commonly used is that we fix the value of $\epsilon$.

The difference between those methods is that, the first one can be GLIE if the $\epsilon$-value goes 0 eventually but the second one cannot. Suppose we have a counter of how many times a state have been visited, $n_t(s)$ and a constant $c$. As long as the $\epsilon$ value for the a state $\epsilon_t(s) = \frac{c}{n_t(s)}$ where $0 < c < 1$, this method can be consider GLIE. However, in practice, we usually use fixed value for $\epsilon$ and after the convergence of the estimation value of action, we switch to greedy selection.

# References

[1] Satinder Singh, Tommi Jaakkola, Michael L Littman, and Csaba Szepesvári. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 38(3):287–308, 2000.