Published in 1994, *Proceedings of the National Conference on Artificial Intelligence (AAAI-94)*, p. 1464, AAAI/MIT Press, Cambridge, MA

When the Best Move Isn't Optimal: Q-learning with Exploration

George H. John*
Computer Science Department
Stanford University
Stanford, CA 94305
gjohn@cs.Stanford.EDU

The most popular delayed reinforcement learning technique, Q-learning (Watkins 1989), estimates the future reward expected from executing each action in every state. If these estimates are correct, then an agent can use them to select the action with maximal expected future reward in each state, and thus perform optimally. Watkins has proved that Q-learning produces an optimal policy (the function mapping states to actions) and that these estimates converge to the correct values given the optimal policy.

However, often the agent does not follow the optimal policy faithfully – the agent must also *explore* the world, taking suboptimal actions in order to learn more about its environment. The "optimal" policy produced by Q-learning is no longer optimal if its prescriptions are only followed occasionally. In many situations (*e.g.*, dynamic environments), the agent never stops exploring. In such domains Q-learning converges to policies which are suboptimal in the sense that there exists a different policy which would achieve higher reward when combined with exploration.

A bit of notation: $Q(x,a)$ is the expected future reward received after taking action a in state x. $V(x)$ is the expected future reward received after starting in state x. $\widehat{Q}$ and $\widehat{V}$ are used to denote the approximations kept by the algorithm. Each time the agent takes an action a moving it from state x to state y and generating a reward r, Q-learning updates the approximations according to the following rules:

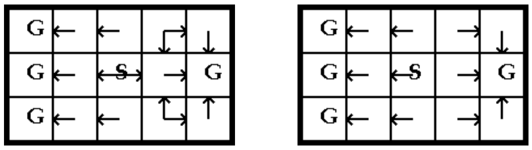
$$\begin{aligned}\widehat{Q}(x,a) &\leftarrow \beta(r+\gamma\widehat{V}(y)) + (1-\beta)\widehat{Q}(x,a) \\ \widehat{V}(x) &\leftarrow \max_a \widehat{Q}(x,a) \end{aligned} \quad (1)$$

where β is the learning rate parameter and γ is the discount rate parameter.

We propose replacing the $\widehat{V}$ update equation by

$$\widehat{V}(x) \leftarrow \sum_a P(a)\widehat{Q}(x,a) \ . \quad (2)$$

That is, update $\widehat{V}(x)$ with the *expected*, not the *maximal*, future reward, taking into account the exploration policy when calculating the expected reward. We call this new algorithm $\overline{Q}$-learning. When the agent always takes the best action, Equations 1 and 2 are equivalent. For some exploration strategies $P(a)$, the probability of taking action a, might be estimated using sample statistics if it is not possible to calculate in closed form. In our experiments this did not degrade performance.



Q: reward = 3.35 $\qquad$ $\overline{Q}$: reward = 3.44

Figure 1: Policies and average reward produced by Q and $\overline{Q}$-learning. "S" is the starting state and "G" indicates goal states which give a reward of 9 units. The algorithms were run for 10^6 steps. Here *random walk* exploration was used, where 30% of the time the agent took a random action instead of the policy-recommended action. $\gamma = .9$, $\beta = .5$.

Figure 1 shows an example motivating our approach. The agent begins in the middle of the world and must choose whether to approach the wall of goals on the left, or the single goal on the right. Q-learning is indifferent as to which action should be performed, because with no exploration either action is optimal – a goal is only two steps away in either direction. $\overline{Q}$ prefers to move left towards the wall of goals because it knows that the agent will explore, and because of this it is better to move left since a goal is still close if exploration causes it to move up or down. By thus taking exploration into account, $\overline{Q}$ achieves higher reward.

The results in Figure 1 are typical of our experiments. $\overline{Q}$-learning always generates policies giving greater average reward, but the improvement over Q-learning depends on the domain and the amount of exploration.

References

Watkins, C. J. C. H. 1989. *Learning from Delayed Rewards.* Ph.D. Dissertation, Cambridge University. Psychology Department.

*This material is based upon work supported under a National Science Foundation Graduate Research Fellowship.