Dora Q-Learning - making better use of explorations

Esther Nicart[1,2], Bruno Zanuttini[1], Bruno Grilhères[3], Fredéric Praca[2]

[1] Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France
first.last-name@unicaen.fr

[2] Cordon Electronics DS2i, 27000 Val de Reuil, France first.last-name@cordonweb.com

[3] Airbus Defence and Space, Élancourt, France first.last-name@airbus.com

Eligibility traces for Q-Learning(λ) [Watkins (1989), Peng (1996), Baird (1995), Cichosz (1995), Sutton & Barto (1998), Wang et al (2013), ...] (hereafter referred to as *Q(λ)*) record the stack of (state, action) pairs enacted during a learning episode, enabling any rewards observed to be back-propagated down the stack, thus speeding up learning.

In standard Q(λ), after an *explore* action, the eligibility trace is cut (reset to an empty stack), meaning that any good results found further on can take a long time to percolate back to the initial state. We present here *Dora*, an adaptation of Q(λ) which makes better use of results found when exploring, and therefore learns consistently faster.

In Dora, our aim is to avoid cutting the trace on an *explore* if possible. This idea is quite simple and natural, but to the best of our knowledge, it has not been developed like this before. We note that the principle of Dora could be argued to resemble that of *experience replay* [Long-Ji Lin, 1991], but Dora is not model-based, has fewer parameters, and consumes less memory, whilst still giving excellent results.

In both Q(λ) and Dora, whenever the algorithm chooses to *explore* in a given state s (that is, it tries an action a which does not currently have the greatest estimated Q-value $\hat{Q}_t(s, \cdot)$) and ends up in a state s', earning an immediate reward r, we update, as usual, the Q-value of a in s using the temporal difference $\delta_t = r + \gamma \max_b \hat{Q}_t(s', b) - \hat{Q}_t(s, a)$, resulting in $\hat{Q}_{t+1}(s, a) = \hat{Q}_t(s, a) + \alpha_t(s, a)\delta_t$.

Standard Q(λ) would clear the trace here, but with Dora, if the new experience now makes a a greedy action in s, that is if $\hat{Q}_{t+1}(s, a) \geq \max_b \hat{Q}_t(s, b)$, we continue just as though a was an *exploit* in s (which in retrospect is the case). Precisely, we do *not* cut the trace and we back-propagate the temporal difference $\hat{Q}_t(s, a) + \alpha_t(s, a)\delta_t - \max_b \hat{Q}_t(s, b)$. Only if $\hat{Q}_{t+1}(s, a)$ is still less than $\max_b \hat{Q}_t(s, b)$, do we admit that a really was an *explore*, and clear the trace.

Observe that if a becomes a greedy action in s, not only do we back-propagate the temporal difference at time t, but we also keep the trace intact, enabling the propagation of results from further in the run back across this "*explore*" join. This is in sharp contrast with Q(λ), which does neither.

Experiments

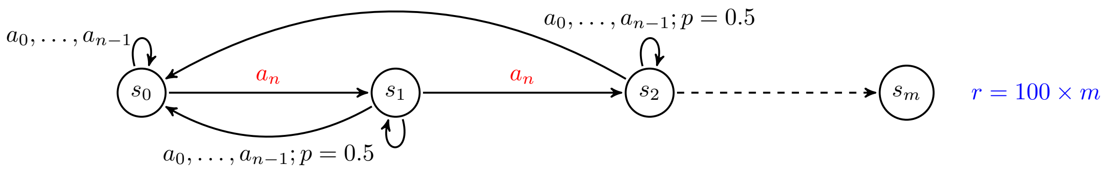
We tested Dora against Q(λ) taking the average results over at least 50 runs. The discount value γ and the decay rate λ were both set to 0.9. We measured the quality of their learning by recording the evolution of the distance of their current *value function* V^t from the optimal V^* at each time-step. We measured this distance in three ways (where s_0 is the starting state) :

episodic-distance$(V^*, V^t) = |V^*(s_0) - V^t(s_0)|$; infinite-distance$(V^*, V^t) = \max_s\{|V^*(s) - V^t(s)|\}$; and L$_2$-distance$(V^*, V^t) = (\sum_s (V^*(s) - V^t(s))^2)^{1/2}$.

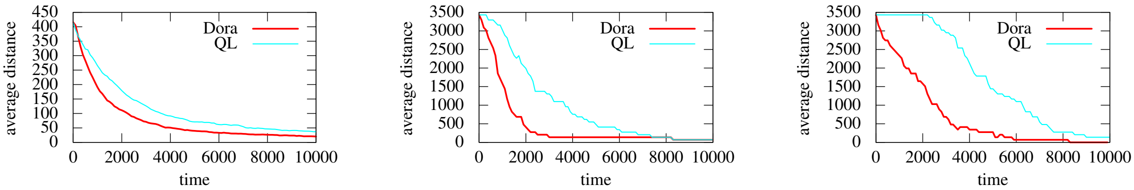
We first tested their comparative performance on randomly generated MDPs from 10 to 100 states, and 5 to 10 actions. The rewards for each couple (s, a) were integers generated randomly between 0 and 100, and the transitions unbounded. We observed that Dora consistently learnt faster than Q(λ) (e.g. Figure 1b), the larger the MDP, the more significant her advantage.

Intuitively, we thought that Dora should perform even better on "long thin" or "cliff-walk" MDPs with long trajectories, and lots of exploration potentially necessary to reach the goal (see Figure 1a), and indeed, we found that in this case, Dora significantly outperforms Q(λ) (Figure 1c), especially with a decreasing ϵ (Figure 1d).

Another natural baseline for assessing the efficiency of Dora is a naive version with no trace-clearing at all (even on explorations) as mentioned in Sutton & Barto (1998). We ran some preliminary experiments which



(a) "Long thin" MDP (n actions, m states) with $p(s_i, a_1 \ldots a_{n-1}, s_0) = 0.5$ and $p(s_i, a_1 \ldots a_{n-1}, s_i) = 0.5$ (zero reward) ; $p(s_i, a_n, s_{i+1}) = 1.0$. State s_m is the only one to offer a reward of $100 \times m$



(b) Random MDP ; 80 states 6 actions ; 50 runs ; decreasing ϵ ; infinite distance

(c) "Long thin" MDP ; 15 states 2 actions ; 50 runs ; constant $\epsilon = 0.2$; episodic distance

(d) "Long thin" MDP ; 15 states 2 actions ; 50 runs ; decreasing ϵ ; episodic distance

FIGURE 1 – The "long thin" MDP, and a very small, but typical selection of the results comparing standard Q(λ) with Dora QL, showing their average distance from the optimal policy.

suggest that on generic MDPs this gives much worse results than both Q(λ) and Dora. A more thorough investigation we leave for future work.

Ongoing Work

We plan to run experiments in a wider variety of settings, for example, with rewards which reduce the optimism rather than increase it, and with several optimal paths in a "long thin" MDP. We also conjecture that there are families of "long thin" MDPs for which we can *prove* that Dora learns exponentially faster than Q(λ) or other algorithms. Apart from the improved algorithm, we believe that such formal results would help gain insight into the interplay between exploration and back-propagation.

Acknowledgements

We wholeheartedly thank the reviewers for their thought-provoking questions and comments.