# Double Q-learning

Vincent Antaki

McGill University

Problem :

- Q-learning sometime has difficulty with stochastic environment.
- Overestimation of action value introduced by positive bias leads to suboptimal policy.

Idea :

- Use two estimators instead of one.

# *Double Q-learning*, H. van Hasselt (2011)

**Lemma 1.** Let $X = \{X_1, ... X_m\}$ be a set of random variables and let $\mu^A = \{\mu_1^A, ..., \mu_M^A\}$ and $\mu^B = \{\mu_1^B, ..., \mu_M^B\}$ be two sets of unbiased estimators such that $E\{\mu_i^A\} = E\{\mu_i^B\} = E\{X_i\}$, for all $i$. Let $\mathbb{M} \stackrel{\text{def}}{=} \{j | Q\{X_j\} = max_i E\{X_i\}\}$ be the set of elements that maximize the expected values. Let $a^*$ be an element that maximizes $\mu^A : \mu_{a^*}^A = \max_i \mu_i^A$. Then, $E\{\mu_{a^*}^B\} = E\{X_{a^*}\} \leq \max_i E\{X_i\}$. Furthermore, the inequality is strickt if and only if $P(a^* \notin \mathbb{M}) > 0$

# *Double Q-learning*, H. van Hasselt (2011)

Case 1 : Assume $a^* \in \mathbb{M}$

- Then $E\{\mu^B_{a^*}\} = E\{X_{a^*}\} \stackrel{\text{def}}{=} \max_i E\{X_i\}$

Case 2 : Assume $a^* \notin \mathbb{M}$, choose $j \in \mathbb{M}$

- Then $E\{\mu^B_{a^*}\} = E\{X_{a^*}\} < E\{X_j\} \stackrel{\text{def}}{=} \max_i E\{X_i\}$

## *Double Q-learning*, H. van Hasselt (2011)

Combining the two cases yield :

$$E\{\mu^B_{a^*}\} = P(a^* \in \mathbb{M})E\{\mu^B_{a^*}|a^* \in \mathbb{M}\} + P(a^* \notin \mathbb{M})E\{\mu^B_{a^*}|a^* \notin \mathbb{M}\}$$

$$= P(a^* \in \mathbb{M})\max_i E\{X_i\} + P(a^* \notin \mathbb{M})E\{\mu^B_{a^*}|a^* \notin \mathbb{M}\}$$

$$\leq P(a^* \in \mathbb{M})\max_i E\{X_i\} + P(a^* \notin \mathbb{M})\max_i E\{X_i\} = \max_i E\{X_i\}$$

N.B. The inequality is strict if and only if $P(a^* \notin \mathbb{M}) > 0$.

## *Double Q-learning*, H. van Hasselt (2011)

- Unlike the single estimator, the double is unbiased when variables are i.i.d.
- In that case, all expected value are equal and $P(a^* \in \mathbb{M}) = 1$.

# First Experience - Multiarm Bandit

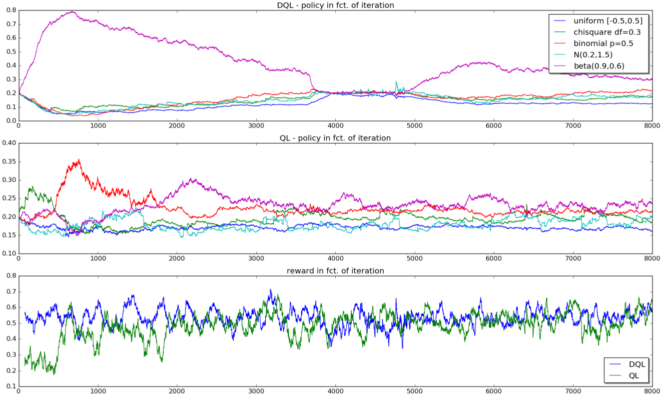
The agent choose between a set of options which provides by sampling a given distribution.*

1. Setting 1 : $[\mathcal{U}(-0.5, 0.5), \mathcal{X}(\mathbf{0.3}), \mathcal{B}(p = \mathbf{0.5}), \mathcal{N}(\mathbf{0.2}, 1.5), \beta(0.9, \mathbf{0.6})]$
2. Setting 2 : $[\mathcal{N}(\mathbf{0}, 0.25), \mathcal{N}(\mathbf{0.2}, 0.25), \mathcal{N}(\mathbf{0.4}, 0.25), \mathcal{N}(\mathbf{0.6}, 0.25), \mathcal{N}(\mathbf{0.8}, 0.25), \mathcal{N}(\mathbf{1.0}, 0.25), \mathcal{N}(\mathbf{1.2}, 0.25)]$

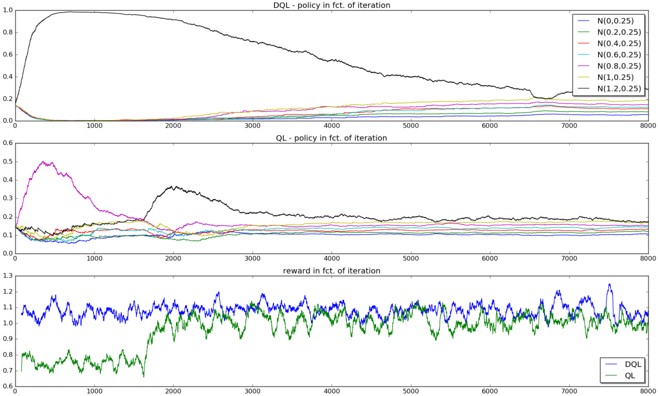
* Numbers in bold indicate the mean of the distribution

We set an epsilon value of 0.2, a learning rate of 0.05 and a discount factor of 0.9. We run each algorithm for 8000 iterations.

# Results - Setting 1

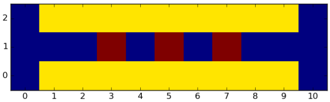


# Results - Setting 2



# Second Experience - The Windy Bridge

- The agent starts in the bottom left corner and the goal state is the the bottom right corner.
- At every iteration, the agent receive a -1 point reward.
- It most cross a bridge surrounded by a cliff to reach the goal and receive a 1000 points reward.
- Falling into a cliff brings a -10 points rewards and puts the agent back at its starting position
- There are 3 windy tiles on the bridge.
- With $p = 0.3$, a windy tile pushes the agent in a random direction.

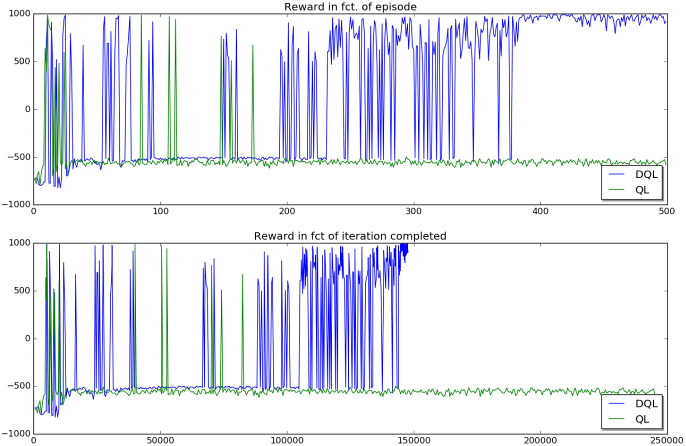


# Experience settings

- $\alpha = 0.05$
- $\gamma = 0.9$
- Maximum iteration by episode $= 500$
- Softmax temperature $= 4$
- Scheduled epsilon :

| epochs | $\epsilon$ |
|---|---|
| 0 | 0.3 |
| 100 | 0.2 |
| 200 | 0.1 |
| 300 | 0.05 |

# Results



Reward in fct. of episode
1000
500
0
−500
−1000
0
100
200
300
400
500
DQL
QL
Reward in fct of iteration completed
1000
500
0
−500
−1000
0
50000
100000
150000
200000
250000
DQL
QL