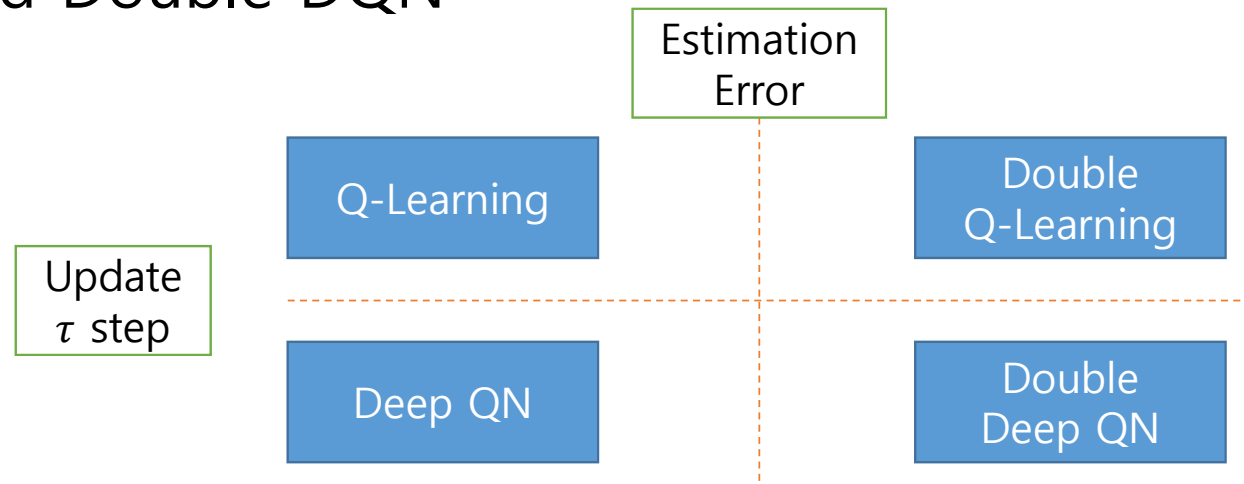
# Deep Reinforcement Learning with Double Q-learning

proceeding of AAAI 2016, Deep Mind

Youngseok Yoon

# Contents

- Background
- DQN and Over optimism
- Double Q-learning and Estimation error
- Double Q-learning and Double DQN
- Experiments



# 1. Background

- Q-Learning update Rule

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha(Y_t^{\mathrm{Q}} - Q(S_t, A_t; \boldsymbol{\theta}_t))\nabla_{\boldsymbol{\theta}_t} Q(S_t, A_t; \boldsymbol{\theta}_t)$$

$$Y_t^{\mathrm{Q}} \equiv R_{t+1} + \gamma \max_a Q(S_{t+1}, a; \boldsymbol{\theta}_t)$$

- DQN update Rule

$$Y_t^{\mathrm{DQN}} \equiv R_{t+1} + \gamma \max_a Q(S_{t+1}, a; \boldsymbol{\theta}_t^-)$$

# 2. DQN and Overoptimism

- The estimation error and Over optimism
  - Lower bound: $\max_a Q_t(s,a) \geq V_*(s) + \sqrt{\frac{C}{m-1}}$.
  - Expectation: $\mathbb{E}\left[\max_a Q_t(s,a) - V_*(s)\right] = \frac{m-1}{m+1}$
  - Proof in paper appendix.
- Q-learning's overestimations were first investigated by Thrun and Schwartz (1993).

# 2. DQN and Overoptimism

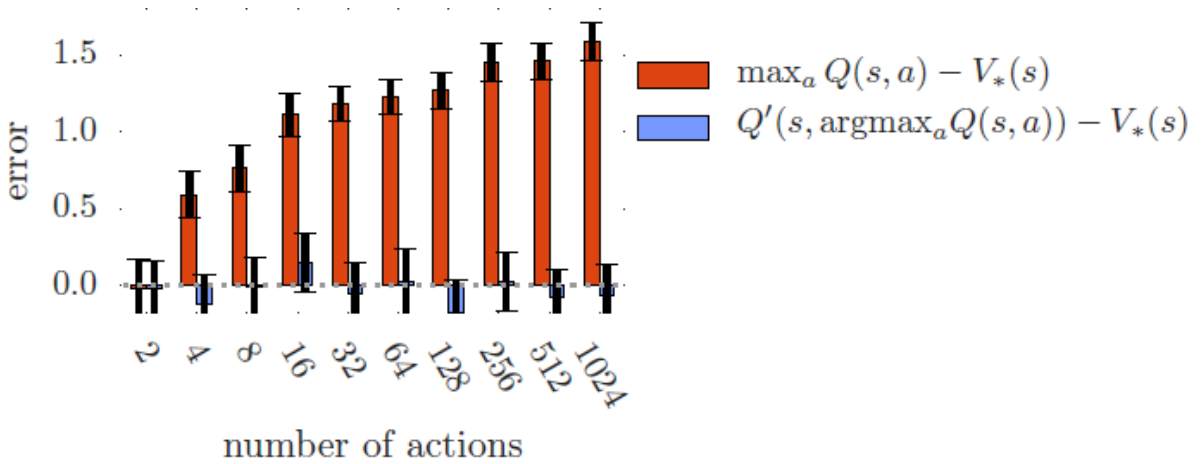



Figure 1: The orange bars show the bias in a single Q-learning update when the action values are $Q(s,a) = V_*(s) + \epsilon_a$ and the errors $\{\epsilon_a\}_{a=1}^m$ are independent standard normal random variables. The second set of action values $Q'$, used for the blue bars, was generated identically and independently. All bars are the average of 100 repetitions.

# 3. Double Q-learning and Estimation error

- Q-Learning

$$Y_t^{\mathrm{Q}} \equiv R_{t+1} + \gamma \max_a Q(S_{t+1}, a; \boldsymbol{\theta}_t)$$

$$Y_t^{\mathrm{Q}} = R_{t+1} + \gamma Q(S_{t+1}, \operatorname*{argmax}_a Q(S_{t+1}, a; \boldsymbol{\theta}_t); \boldsymbol{\theta}_t)$$

- Double Q-learning

$$Y_t^{\mathrm{DoubleQ}} \equiv R_{t+1} + \gamma Q(S_{t+1}, \operatorname*{argmax}_a Q(S_{t+1}, a; \boldsymbol{\theta}_t); \boldsymbol{\theta}'_t)$$

# 3. Double Q-learning and Estimation error

- Estimation error
  - Assume $\sum_a (Q_t(s,a) - V_*(s)) = 0$, $\frac{1}{m}\sum_a (Q_t(s,a) - V_*(s))^2 = C$
  - We can have

$$Q_t(s,a_1) = V_*(s) + \sqrt{C\frac{m-1}{m}}, \quad Q_t(s,a_i) = V_*(s) - \sqrt{C\frac{1}{m(m-1)}}, \text{ for } i > 1.$$

  - Then, we can get $|Q'_t(s, \text{argmax}_a\, Q_t(s,a)) - V_*(s)| = 0$, if $Q'_t(s,a_1) = V_*(s)$
  - So, the lower bound is zero.

# 3. Double Q-learning and Estimation error

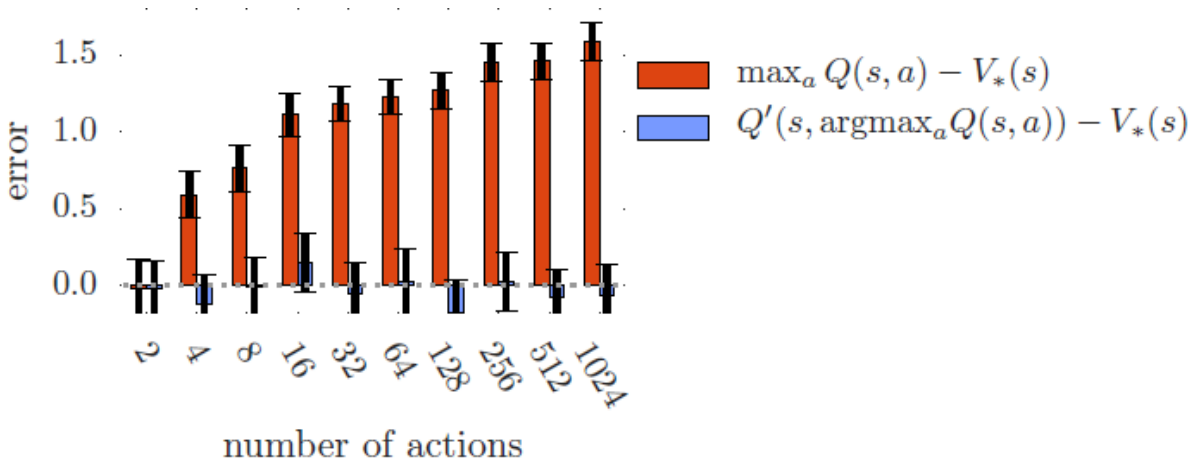



Figure 1: The orange bars show the bias in a single Q-learning update when the action values are $Q(s,a) = V_*(s) + \epsilon_a$ and the errors $\{\epsilon_a\}_{a=1}^m$ are independent standard normal random variables. The second set of action values $Q'$, used for the blue bars, was generated identically and independently. All bars are the average of 100 repetitions.

# 3. Double Q-learning and Estimation error

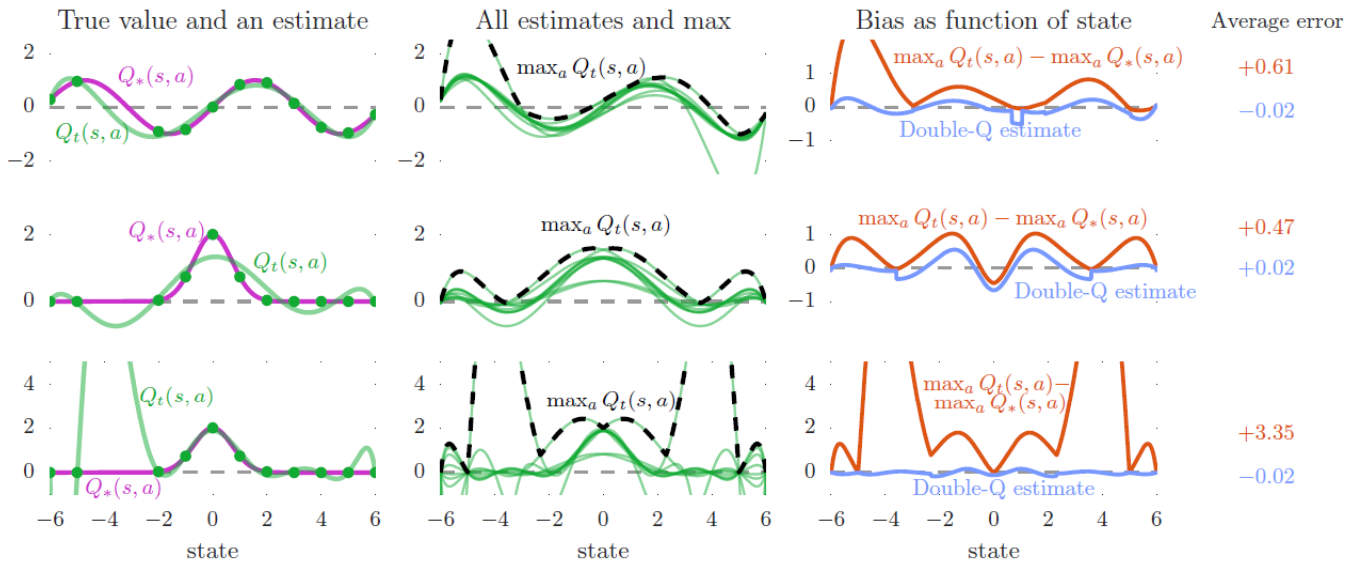



Figure 2: Illustration of overestimations during learning. In each state (x-axis), there are 10 actions. The **left column** shows the true values $V_*(s)$ (purple line). All true action values are defined by $Q_*(s,a) = V_*(s)$. The green line shows estimated values $Q(s,a)$ for one action as a function of state, fitted to the true value at several sampled states (green dots). The **middle column** plots show all the estimated values (green), and the maximum of these values (dashed black). The maximum is higher than the true value (purple, left plot) almost everywhere. The **right column** plots shows the difference in orange. The blue line in the right plots is the estimate used by Double Q-learning with a second set of samples for each state. The blue line is much closer to zero, indicating less bias. The three **rows** correspond to different true functions (left, purple) or capacities of the fitted function (left, green). (Details in the text)

# 4. Double Q-learning and Double DQN

- DQN

$$Y_t^{\mathrm{Q}} = R_{t+1} + \gamma Q(S_{t+1}, \operatorname*{argmax}_a Q(S_{t+1}, a; \boldsymbol{\theta}_t); \boldsymbol{\theta}_t)$$

- Double Q-learning

$$Y_t^{\mathrm{DoubleQ}} \equiv R_{t+1} + \gamma Q(S_{t+1}, \operatorname*{argmax}_a Q(S_{t+1}, a; \boldsymbol{\theta}_t); \boldsymbol{\theta}'_t)$$

- Double DQN

$$Y_t^{\mathrm{DoubleDQN}} \equiv R_{t+1} + \gamma Q(S_{t+1}, \operatorname*{argmax}_a Q(S_{t+1}, a; \boldsymbol{\theta}_t), \boldsymbol{\theta}_t^-)$$

# 5. Experiments

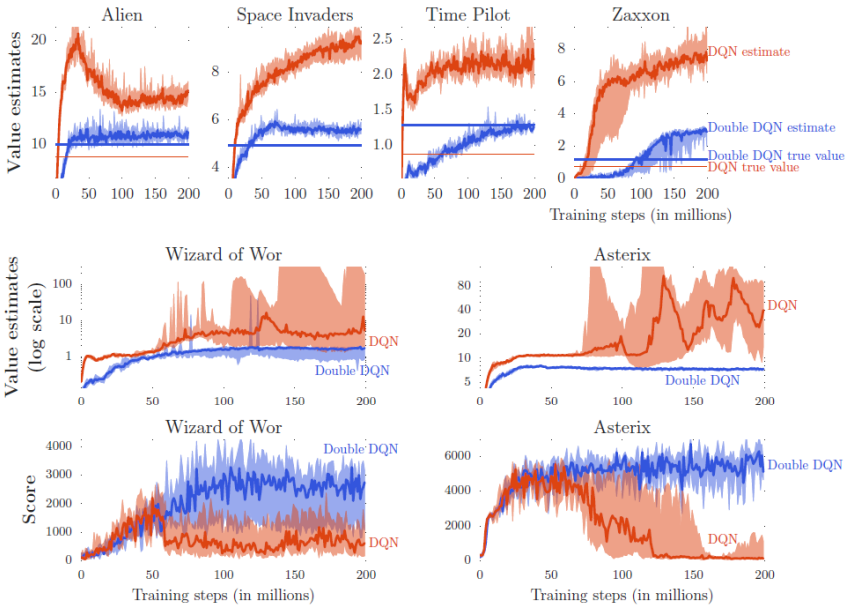



Figure 3: The **top** and **middle** rows show value estimates by DQN (orange) and Double DQN (blue) on six Atari games. The results are obtained by running DQN and Double DQN with 6 different random seeds with the hyper-parameters employed by Mnih et al. (2015). The darker line shows the median over seeds and we average the two extreme values to obtain the shaded area (i.e., 10% and 90% quantiles with linear interpolation). The straight horizontal orange (for DQN) and blue (for Double DQN) lines in the top row are computed by running the corresponding agents after learning concluded, and averaging the actual discounted return obtained from each visited state. These straight lines would match the learning curves at the right side of the plots if there is no bias. The **middle** row shows the value estimates (in log scale) for two games in which DQN's overoptimism is quite extreme. The **bottom** row shows the detrimental effect of this on the score achieved by the agent as it is evaluated during training: the scores drop when the overestimations begin. Learning with Double DQN is much more stable.

# 6. Discussion

- Over optimistic value are not necessarily a problem in and of themselves.
- It is known that sometimes it is good to be optimistic: optimism in the face of uncertainty is a well-known exploration technique
- If, however, the overestimations are not uniform and not concentrated at states about which we wish to learn more, then they might negatively affect the quality of the resulting policy. (Thrun and Schwartz (1993) give specific examples in which this leads to suboptimal policies, even asymptotically.)

Thank you!