Motivation

1. Q-learning is known to overestimate action values under certain conditions.

2. The paper demonstrates that such overestimation happens in standard DRL benchmark, significantly lower performance and can be partially prevented.

Double Q-learning

1. Under a given policy π, the true value of an action a in a state s can be given by:

$$Q_\pi(s,a) \equiv \mathbb{E}\left[R_1 + \gamma R_2 + \ldots | S_0 = s, A_0 = a, \pi\right]$$

2. When we parameterize the value function $Q\left(s, a; \boldsymbol{\theta}_t\right)$ with a function approximator, then the update for the parameter is:

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

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

3. In DQN, the target for training Q is:

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

4. In either targets, the same Q-function is used for action selection and action evaluation to construct the target.

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

5. This makes it likely to select action with over-estimated values. To prevent this, double Q-learning decouple the selection from the evaluation.

$$Y_t^{\text{ Double } Q} \equiv R_{t+1} + \gamma Q\left(S_{t+1}, \underset{a}{\operatorname{argmax}} Q\left(S_{t+1}, a; \boldsymbol{\theta}_t\right); \boldsymbol{\theta}_t'\right)$$

6. Note that the selection of the action, in the argmax, is still due to the online weights θ_t. This means that as in Q-learning, we are still estimating the value of the greedy policy according to the current values, as defined by θ_t. However, the evaluation of the action is done by another set of weight θ_t'.

7. The paper would like to make minimal changes to the existing DQN algorithm. They thus propose to use the target network to evaluate the action chosen by argmax over the Q parameterized by current θ_t. That is, they set θ_t' to be $\boldsymbol{\theta}_t^{-}$.

Experimental Result

1. In a continuous bandit case, they show that the overestimation is not specific to one value function.

2. And the more high capacity the function class used to approximate Q is, the more severe the over-estimation. This means that since we use large neural networks in modern DRL, we're in trouble wrt over-estimation.

3. They show that over-estimation in DQN happens in the Atari domain. And applying double Q-learning reduces the degree of over-estimation and increases performance overall.

4. It is worth noting that even with double Q-learning, over-estimation still occurs in the Atari domains.