Deep Reinforcement Learning with Double Q-Learning

By
Saptarashmi Bandyopadhyay
First Year PhD Student
Department of Computer Science and Engineering
Pennsylvania State University, University Park

Outline

- Motivation of the Paper
- New Contributions of the Paper

Introduction

- A conference paper 'Deep Reinforcement Learning with Double Q-Learning'
- Authors:
  - Dr. Hado Van Hasselt and Dr. Arthur Guez Research Scientist, Google DeepMind
  - Prof. David Silver, University College, London and currently leading the Reinforcement Learning Group at Google DeepMind

For reference

- Please read the following reference paper in addition to this paper for detailed information
- Double Q-learning by Hado van Hasselt, Advances in Neural Information Processing Systems, 23:2613-2621, 2010

Introduction

- A conference paper 'Deep Reinforcement Learning with Double Q-Learning'
- Authors:
  - Dr. Hado Van Hasselt and Dr. Arthur Guez Research Scientist, Google DeepMind
  - Prof. David Silver, University College, London and currently leading the Reinforcement Learning Group at Google DeepMind

Motivation

- Q Learning algorithm in combination with deep neural network experiences considerable over-estimations for some games in the Atari 2600 domain
- Prominent issues:
  - Are such over-estimations common?
  - Do they affect performance?
  - Can these over-estimations be prevented?

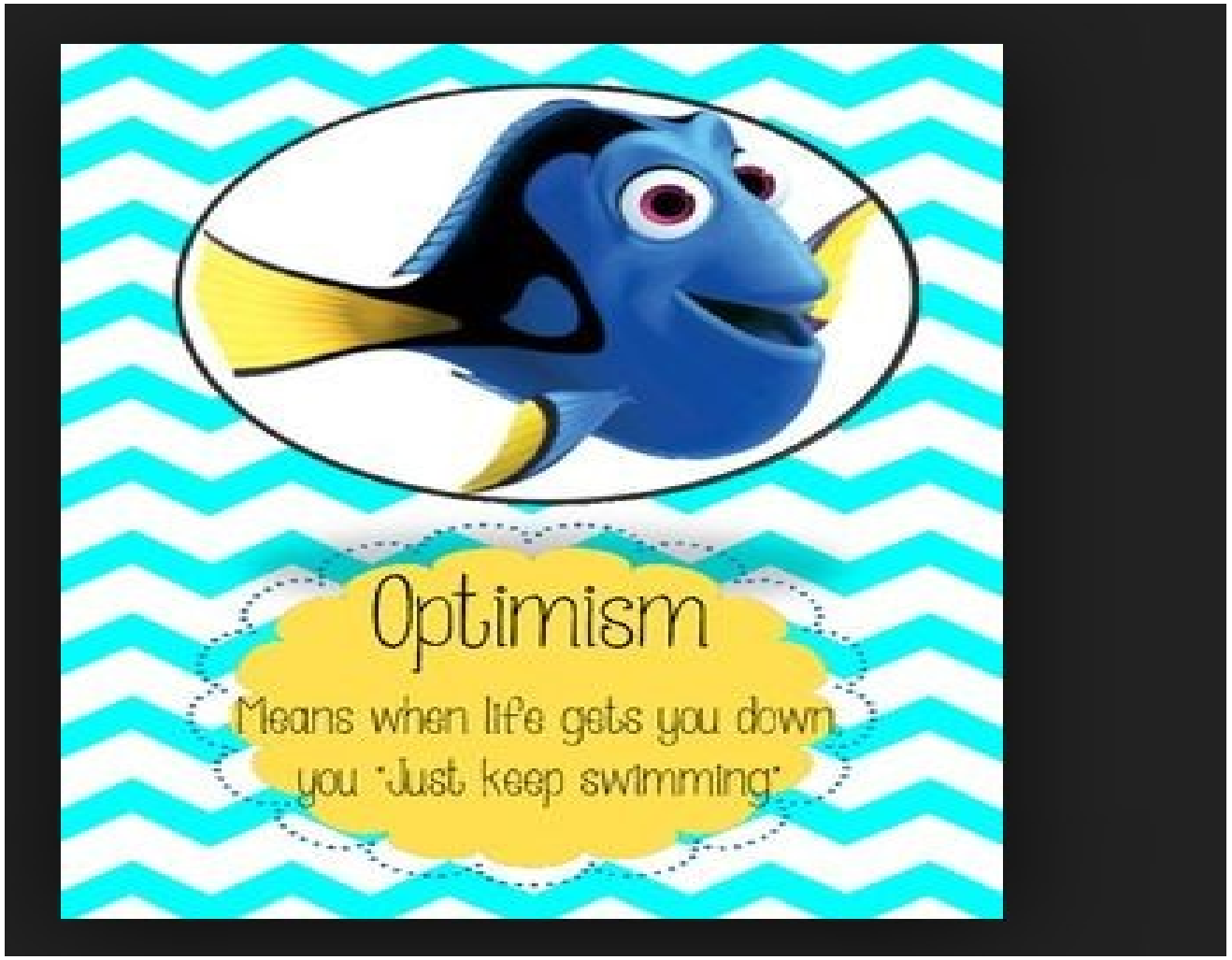
Over-estimation in Reinforcement Learning

- Objective: To learn the best policies for sequential decision problems by optimizing a cumulative future reward signal
- Q learning sometimes learns extremely high action values due to a maximization step over estimated action values, thus being biased to over-estimated action values

Causes of Over-estimation

- In-accurate action values irrespective of the source of approximation error
  - Challenge: Learning is preferred to imprecise
- Inflexible function approximation
- Noise

Is Optimism Bad?



[2]

BUT!!!!!!!

- If the over-estimations are
  - Not uniform
  - Not concentrated at states that we are exploring
- Quality of the resulting policy is affected negatively!
- Can give sub-optimal policies (Thrun and Schwartz (1993))

Why Atari 2600 Games?

- Flexible function approximators with possible low asymptotic approximation error
- Determinism prevents the negative impact of noise
- Still over-estimation occurs!

Solution

- Double DQN algorithm
- Benefits of the new algorithm
  - More accurate action value estimates
  - Higher scores on several games
- So, it is important to reduce over-estimations in DQN

Revision of Q Learning

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

Value of an action a in state s for a given policy π and γ ∈ [0, 1] is the discount factor that is a trade-off between local and global reward

Optimal value is $Q_*(s, a) = \max_\pi Q_\pi(s, a)$

Q Learning

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

Y_t^Q :Target value;

R_{t+1} :Immediate reward;

S_{t+1} :Resulting state

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

Q-Learning update of network parameters after taking action A_t in state S_t where α is a scalar step size

Deep Q Networks

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

The target value is similar except θ^- i.e. parameters are copied every τ steps from the online network to ensure that $\theta_t^- = \theta_t$ is fixed for all steps

Double Q Learning

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

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

- 2 value functions are learned. So 2 sets of weight θ_t and θ_t'
- θ_t' is used to fairly evaluate the value of the policy

Estimation Errors

- Upward bias is induced irrespective of the source of error
- Any method will have errors as initial true values are unknown

Reward of the games

- Manual annotation of final rewards for distinct endings
- Magnitude of the scores provide sentiment polarity of good/bad endings
- Small negative reward in each non-terminating step to encourage the user to complete the game quickly

Lower Bound to Over-estimation

Theorem 1. *Consider a state s in which all the true optimal action values are equal at $Q_*(s,a) = V_*(s)$ for some $V_*(s)$. Let Q_t be arbitrary value estimates that are on the whole unbiased in the sense that $\sum_a (Q_t(s,a) - V_*(s)) = 0$, but that are not all correct, such that $\frac{1}{m}\sum_a (Q_t(s,a) - V_*(s))^2 = C$ for some $C > 0$, where $m \geq 2$ is the number of actions in s. Under these conditions, $\max_a Q_t(s,a) \geq V_*(s) + \sqrt{\frac{C}{m-1}}$.*

- Tight lower bound
- Lower bound on the absolute error of Double Q Learning estimate is 0

Double Q-Learning

- Over-estimation is reduced by breaking up the max operation in target into action selection and action evaluation
- Target network gives the second value function without additional networks
- Greedy policy is evaluated according to online network but target network is used to estimate its value

Over-estimation in Learning

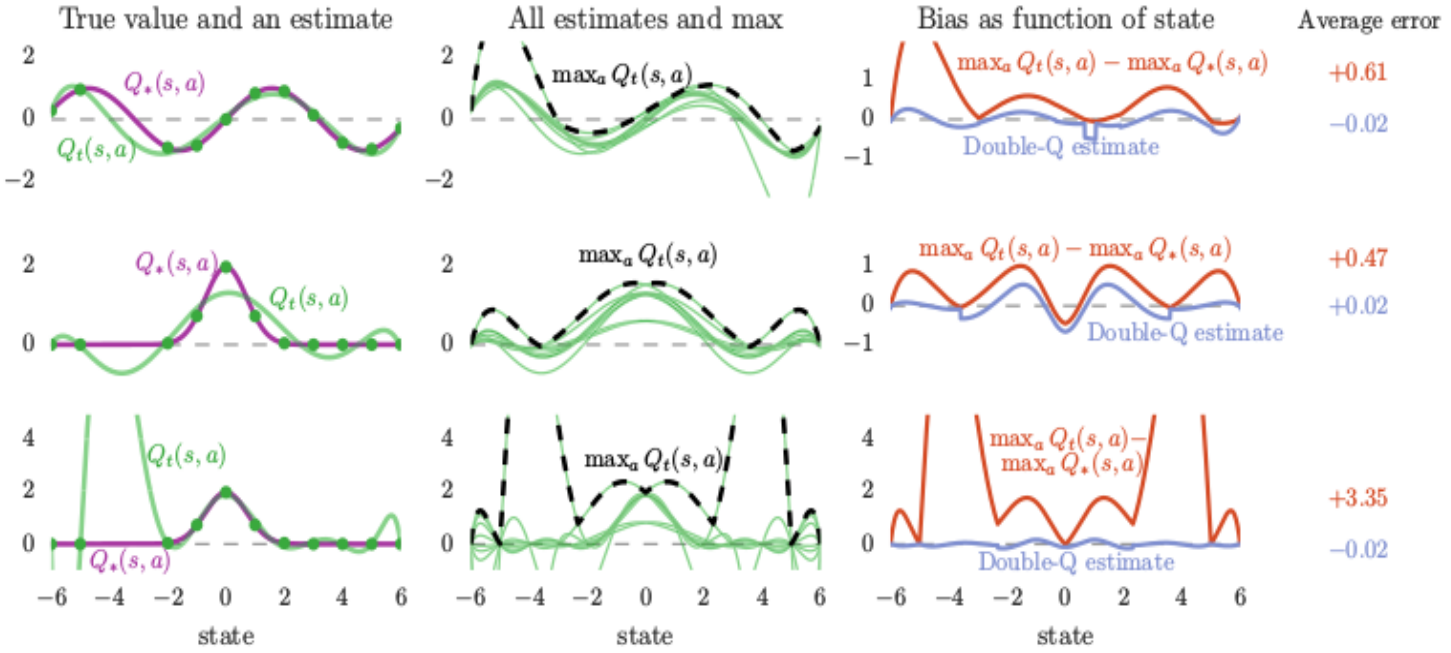


Figure 2: Illustration of overestimations during learning. In each state (x-axis), there are 10 actions. The **left column** shows the true values $V_*(s)$ (purple line). All true action values are defined by $Q_*(s,a) = V_*(s)$. The green line shows estimated values $Q(s,a)$ for one action as a function 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)

Double Q-Learning for the 6 Atari Games

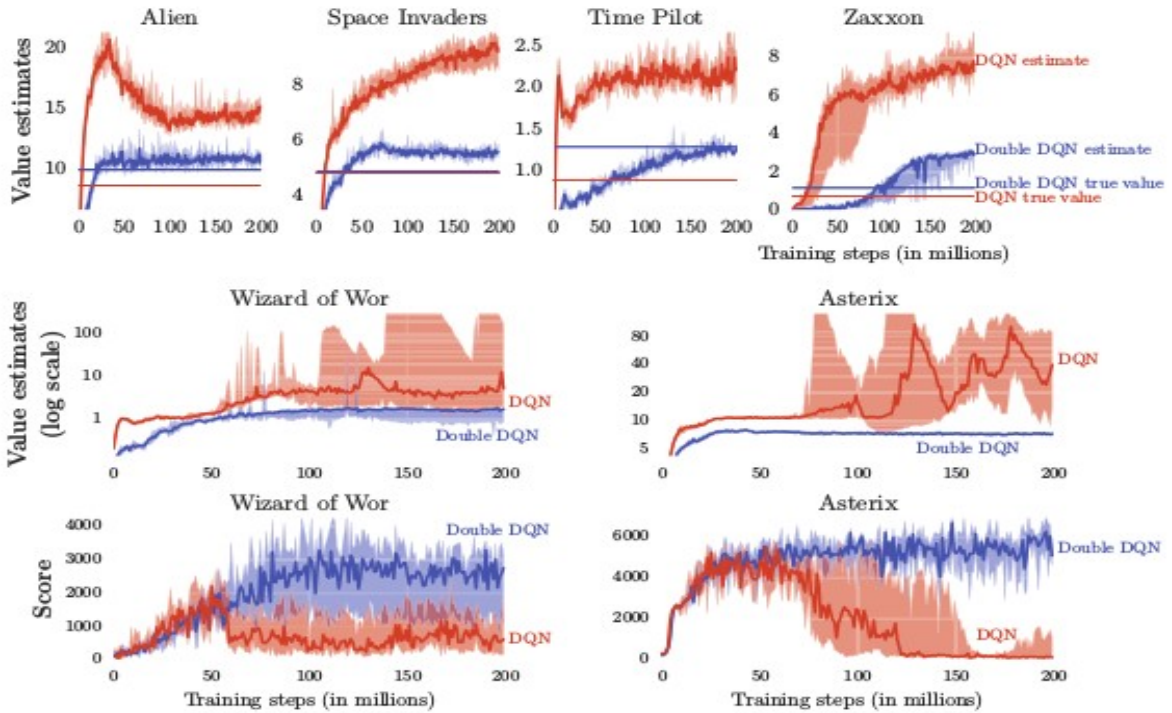


Figure 3: The **top** and **middle** rows show value estimates by DQN (orange) and Double DQN (blue) on six Atari games. The results are obtained by running DQN and Double DQN with 6 different random seeds with the hyper-parameters employed by Mnih et al. (2015). The darker line shows the median over 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.

Performance on 49 Atari Games

	no ops		human starts		
	DQN	DDQN	DQN	DDQN	DDQN (tuned)
Median	93%	**115%**	47%	88%	**117%**
Mean	241%	**330%**	122%	273%	**475%**

Table 1: Summarized normalized performance on 49 games for up to 5 minutes with up to 30 no ops at the start of each episode, and for up to 30 minutes with randomly selected human start points. Results for DQN are from Mnih et al. (2015) (no ops) and Nair et al. (2015) (human starts).

Mean and median are in % to show their relative improvement in score with respect to the human performed games.

Quality of Learned Policies

- Over-estimation not necessarily impacts quality negatively
- However, reducing over-estimations can improve the stability of the learning

Robustness to Human Starts

- Deterministic games can lead to memorization of action sequences without much need to generalize
- Double DQN has higher median and mean scores than DQN
- Double DQN is more robust to the evaluation

Summary of the paper

- Reasons of over-optimism in Q-Learning in large-scale problems
- Over-estimations are more common and severe in practice as observed with Atari games
- Double Q-Learning successfully reduces over-optimism, leading to more stable and reliable learning

Summary of the paper

- Double DQN algorithm uses existing architecture and deep neural network of DQN without extra networks
- Double DQN finds better policies in the Atari 2600 domain

Summary of the paper

- Double DQN algorithm uses existing architecture and deep neural network of DQN without extra networks
- Double DQN finds better policies in the Atari 2600 domain

References

1. Deep Reinforcement Learning with Double Q-Learning by Hado van Hasselt , Arthur Guez, and David Silver, Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16),2016.
2. https://www.pinterest.com/pin/306667055849545558/

Thank You