AIL 722: Reinforcement Learning

Lecture 27: Overestimation and Double Q-Learning

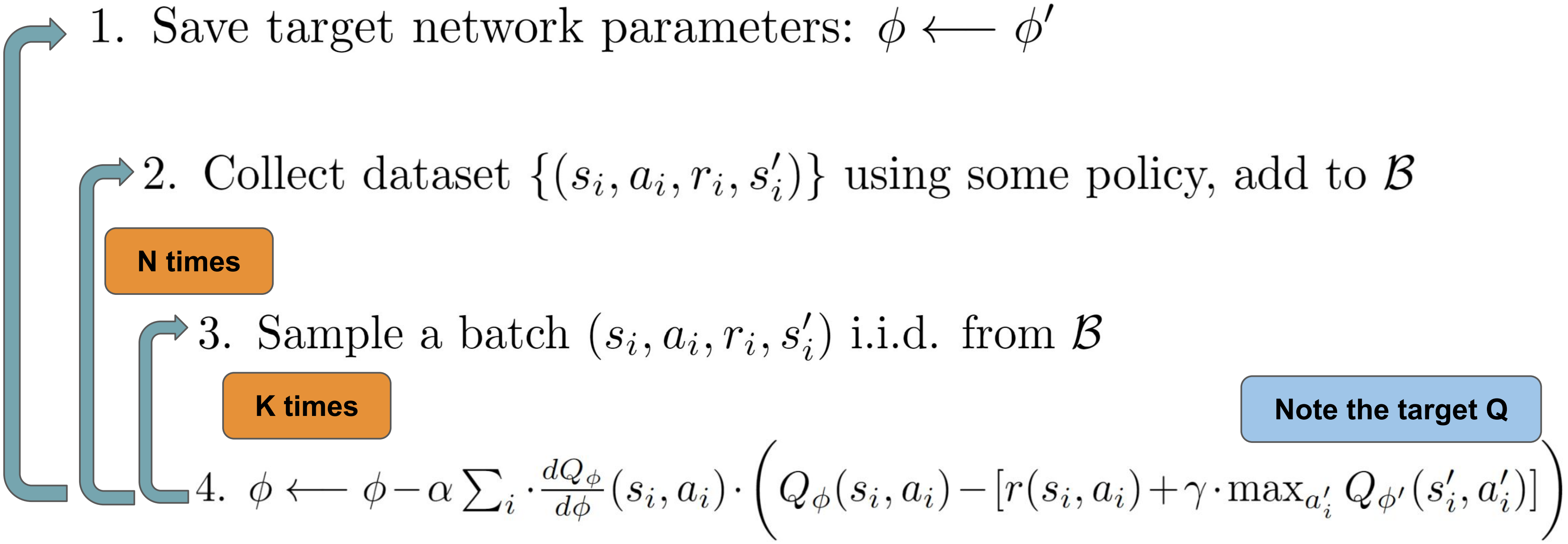
Raunak Bhattacharyya

ScAI YARDI SCHOOL OF ARTIFICIAL INTELLIGENCE
INDIAN INSTITUTE OF TECHNOLOGY DELHI

Recap & Today’s Outline

- Deep Q-Learning
- Interacting Processes
- Overestimation Bias

Replay Buffer and Target Network



1. Save target network parameters: $\phi \longleftarrow \phi'$

2. Collect dataset $\{(s_i, a_i, r_i, s'_i)\}$ using some policy, add to $\mathcal{B}$

N times

3. Sample a batch (s_i, a_i, r_i, s'_i) i.i.d. from $\mathcal{B}$

K times

Note the target Q

4. $\phi \longleftarrow \phi - \alpha \sum_i \cdot \frac{dQ_\phi}{d\phi}(s_i, a_i) \cdot \left(Q_\phi(s_i, a_i) - [r(s_i, a_i) + \gamma \cdot \max_{a'_i} Q_{\phi'}(s'_i, a'_i)] \right)$

Recap & Today's Outline

- Deep Q-Learning
- Interacting Processes
- Overestimation Bias
- Visualising Q values on ALE
- Examples of Overestimation
- Double Q-Learning

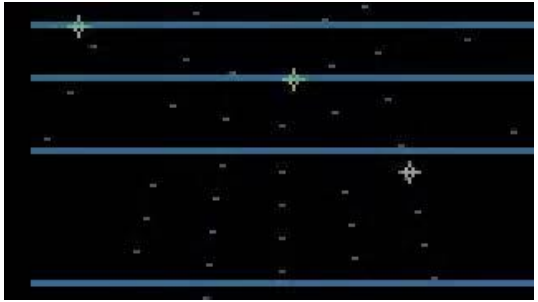
Arcade Learning Environment

Pong, Source: Youtube

Breakout, Source: Youtube

Q*bert, Source: Youtube

Seaquest, Source: Youtube



Beamrider, Source: Youtube

Enduro, Source: Youtube

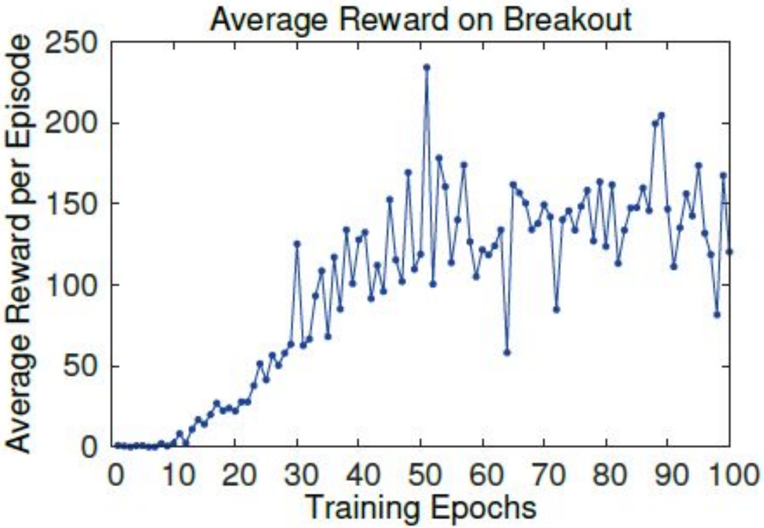
Arcade Learning Environment, Bellemare et. al., JAIR 2013

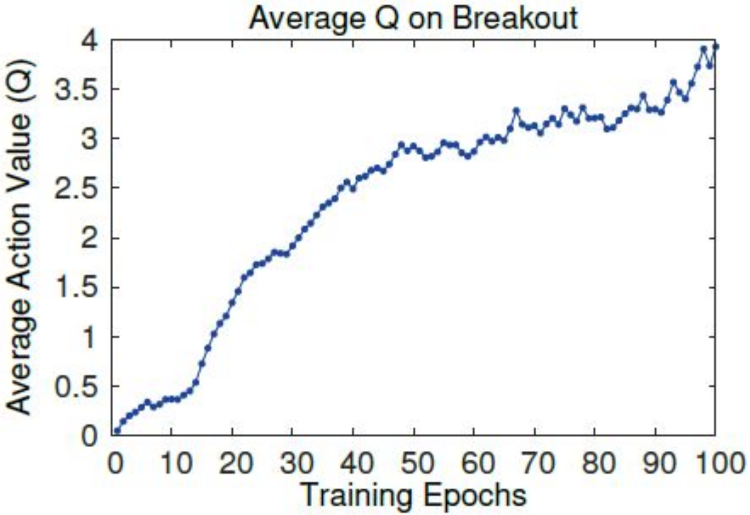
Arcade Learning Environment

- Goal: Single algo, with fixed set of hyperparams, to learn to play each game separately from interaction, given only screen pixels as input

- Demands good learning algo, not practically feasible to overfit the domain by relying on tuning

Visualizing Values

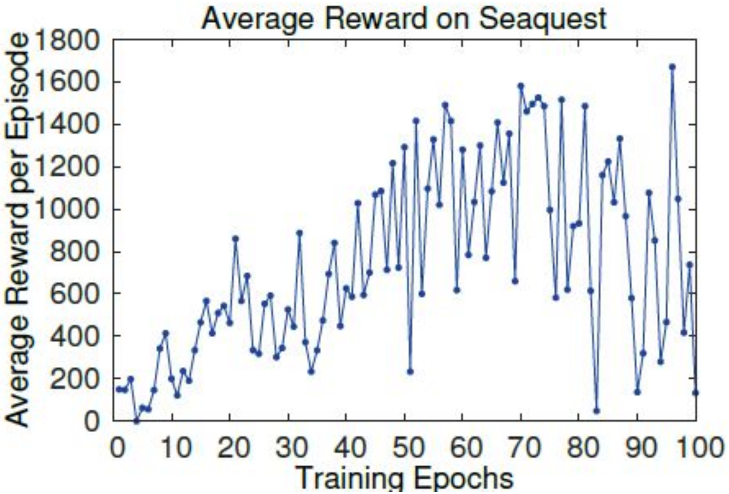


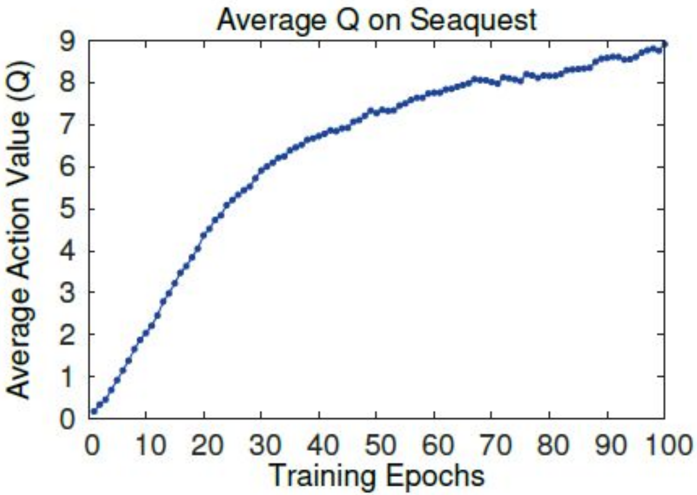


Avg on held out set of states

Playing Atari with Deep RL, Mnih et. al., ArXiv 2013

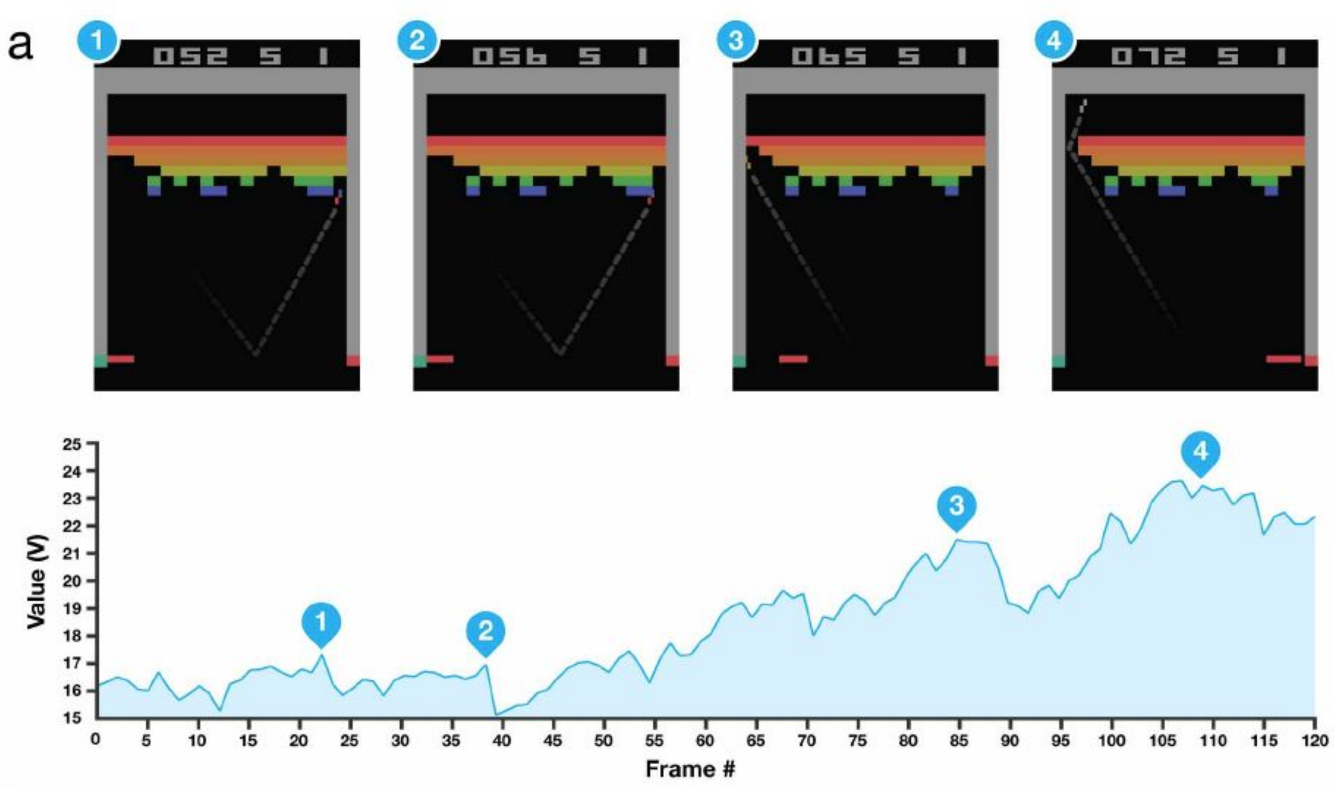
Visualizing Values





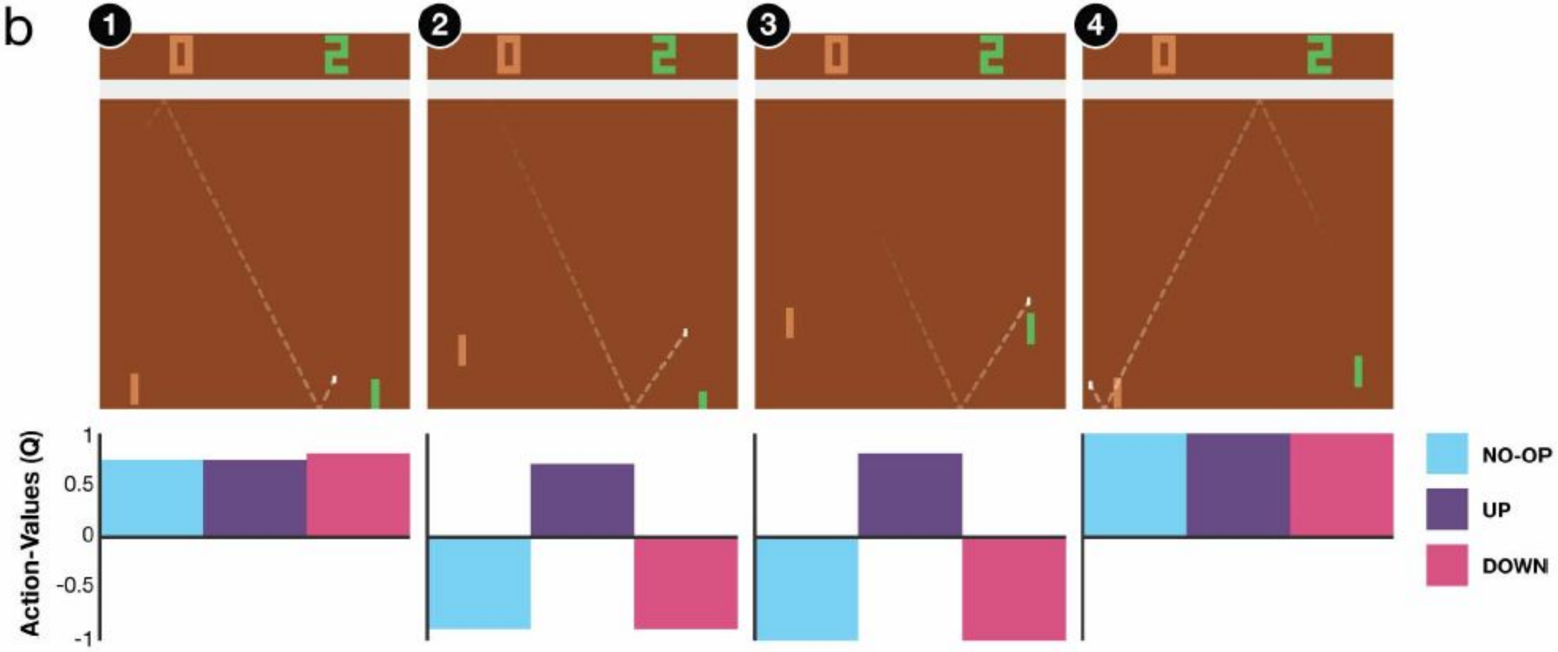
Playing Atari with Deep RL, Mnih et. al., ArXiv 2013

Visualizing Values



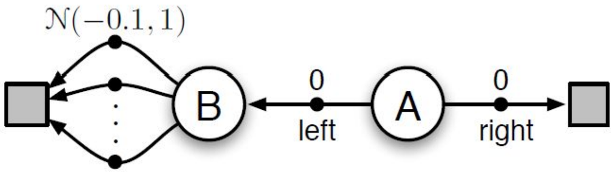
Human-level control through Deep RL, Mnih et. al., Nature 2015

Visualizing Values



Human-level control through Deep RL, Mnih et. al., Nature 2015

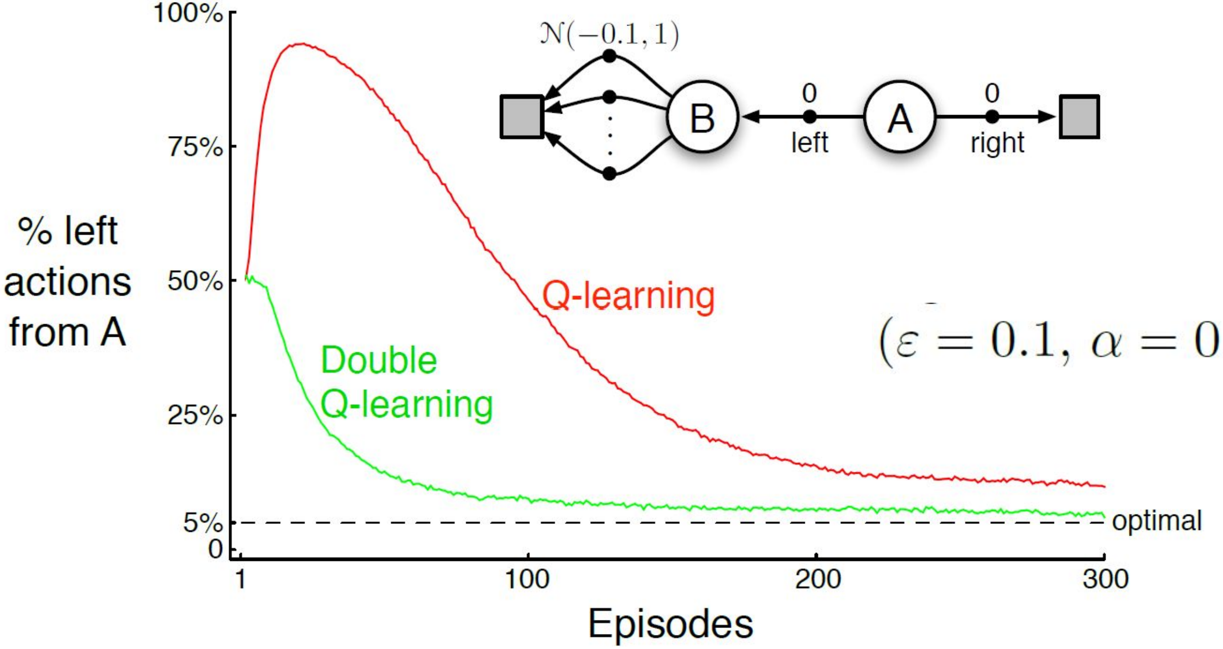
Problem: Overestimation



- Episodes always start in A
- Right transitions to terminal state and terminates
- Left transitions to B with reward 0
- Many possible actions from B
- All lead to termination
- Reward is drawn from N(-0.1,1)

Section 6.7, Reinforcement Learning: An Introduction, Sutton & Barto

Overestimation



$(\varepsilon = 0.1,\ \alpha = 0.1,\ \text{and}\ \gamma = 1)$

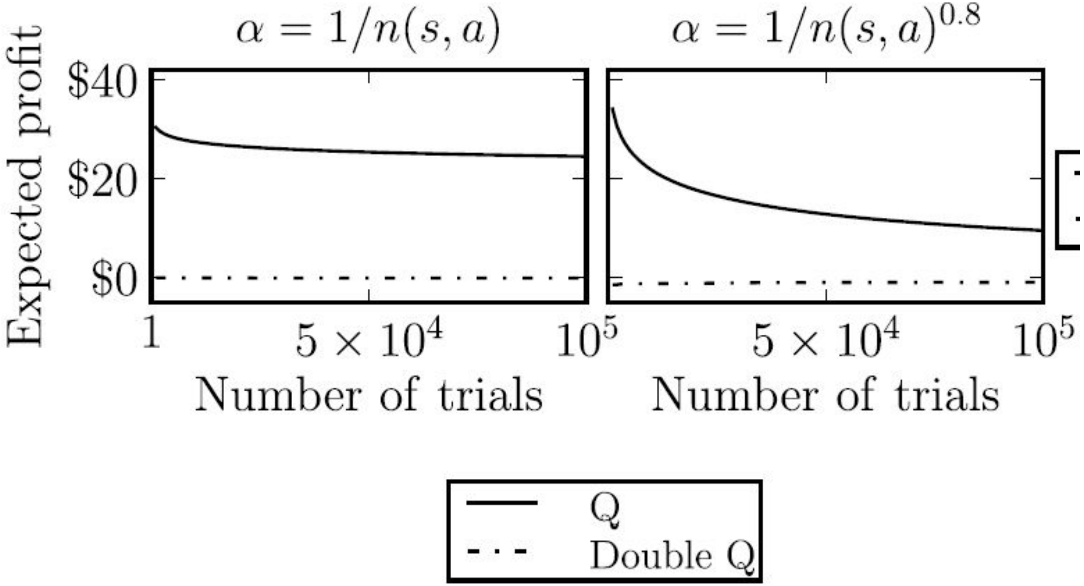
Section 6.7, Reinforcement Learning: An Introduction, Sutton & Barto

Overestimation: Roulette Example

- Single state, 171 actions
- $1 bet each try
- Expected payout 0.947$ on each bet
- One Stop action ends the game with $0
- Ignore available funds: Bet $1 every time

Double Q-Learning, van Hasselt, NeurIPS 2010

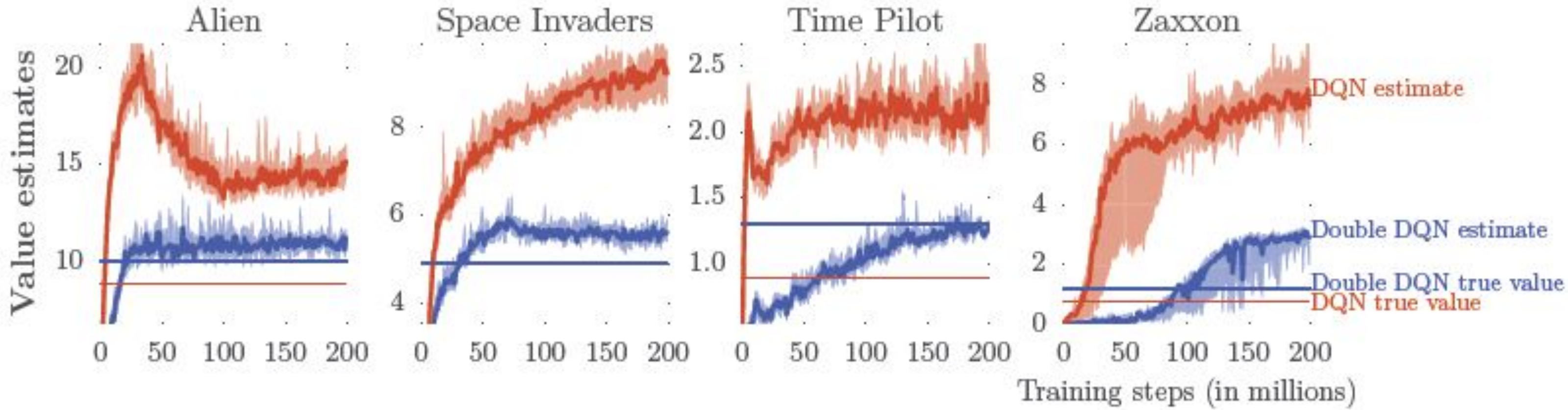
Overestimation: Roulette Example



- Linear decay
- Polynomial decay

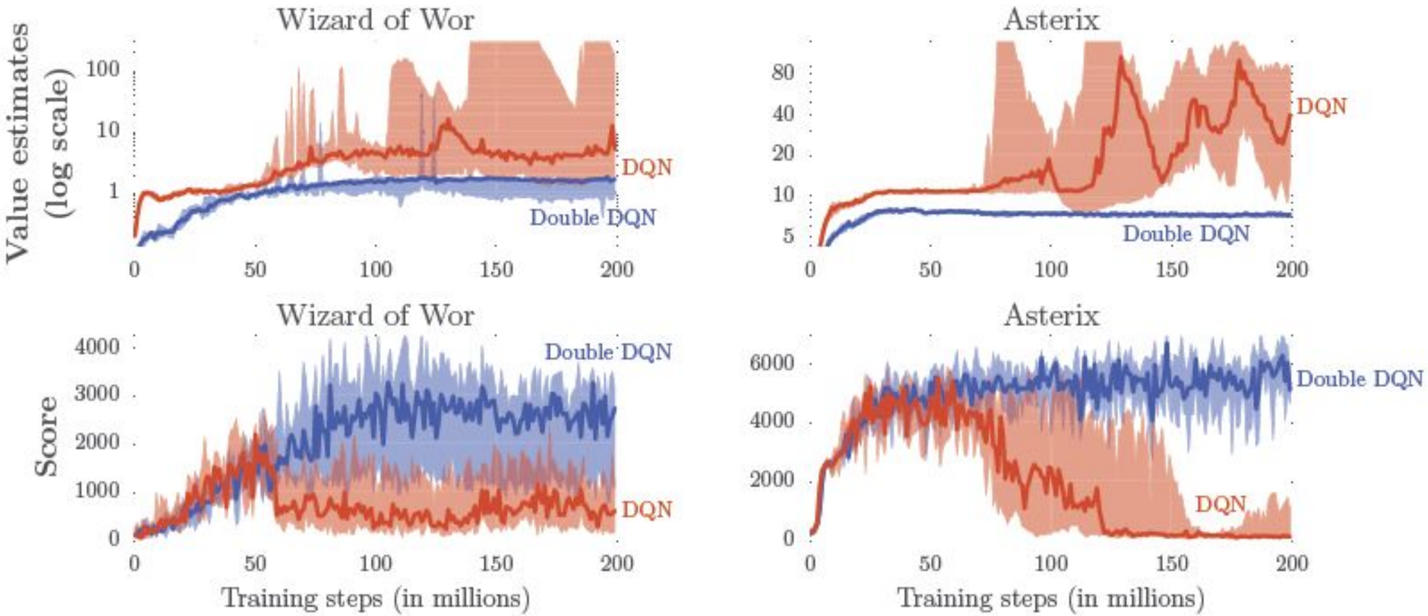
Double Q-Learning, van Hasselt, NeurIPS 2010

Overestimation: ALE



Deep RL with Double Q-Learning, van Hasselt et. al., AAAI 2016

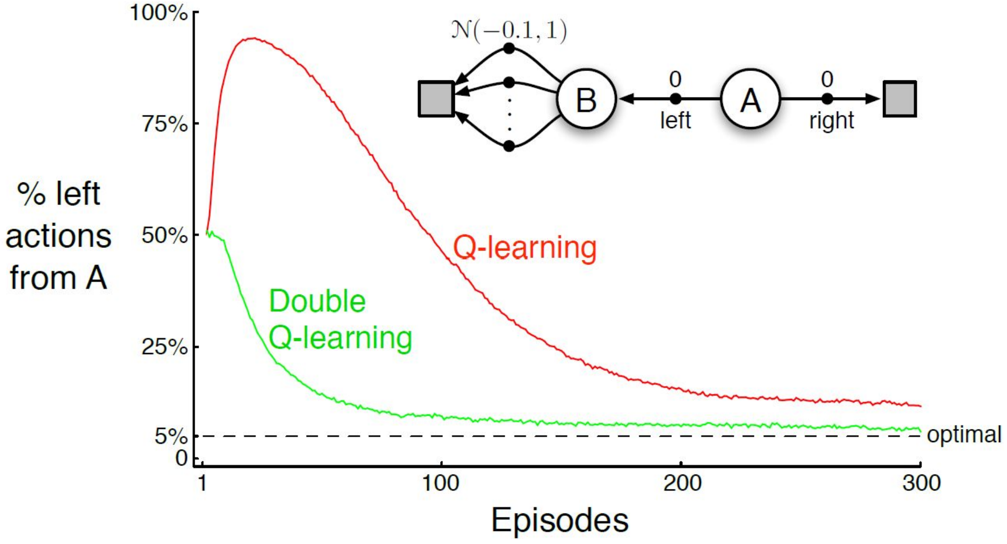
Impact on Performance



Deep RL with Double Q-Learning, van Hasselt et. al., AAAI 2016

Overestimation Bias

$$2.\ \ y_i = r(s_i, a_i) + \gamma \cdot \max_{a'_i} Q(s'_i, a'_i)$$



Section 6.7, Reinforcement Learning: An Introduction, Sutton & Barto

Summary & Announcements

- Summary
  - Visualising Values
  - Problem of overestimation
  - Double Q-Learning

- Announcements
  - Project Proposal due
    - Send as an email to Raunak
      - raunakbh@iitd.ac.in
    - Deadline: Friday, 18/10, 11.55 pm