Deep Q Networks

Easwar Subramanian

TCS Innovation Labs, Hyderabad

cs5500.2020@iith.ac.in

September 23, 2023

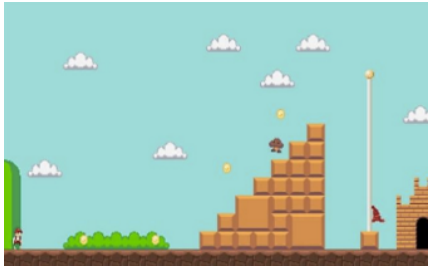
Overview

TCS Research & Innovation

Function Approximation Methods

On the need for Function Approximators

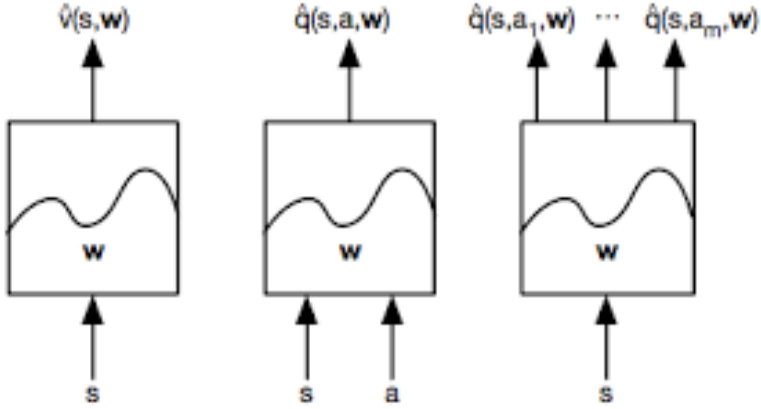
- To solve large scale RL problems
  - ★ Game of Backgammon : 10^{20} states
  - ★ Game of Go : 10^{170} states
  - ★ Even Atari games have large state space



$|\mathcal{S}|$ is very large : Curse of Dimensionality

Slide Credit: David Silver : UCL lectures

Neural Network Approximators



Policy Evaluation Using Neural Networks

The value of a policy π is given by

$$\begin{aligned} V^{\pi}(s) &= \mathbb{E}_{\pi}\left(\sum_{k=0}^{\infty}\gamma^{k} r_{t+k+1} | s_t = s\right) \\ &= \mathbb{E}_{\pi}\left[r_{t+1} + \gamma V^{\pi}(s_{t+1}) | s_t = s\right] \end{aligned}$$

Training data is collected as,

- MC Update : $\left(s_i, \underbrace{\left[\sum_{k=t}^{H}\left(\gamma^{k} r_{t+k+1} | s_t = s\right)\right]}_{=y_i}\right)$
- Fitted V Iteration : $\left(s_i, \underbrace{r + V_{\phi}^{\pi}(s_i')}_{=y_i}\right)$

On the Convergence of Fitted Iterations

Question : What can we say about the convergence of fitted iteration methods ?

- Does fitted V iteration converge to V^π ?
- Does neural fitted iteration converge to Q_* ?

Convergence in DP setup

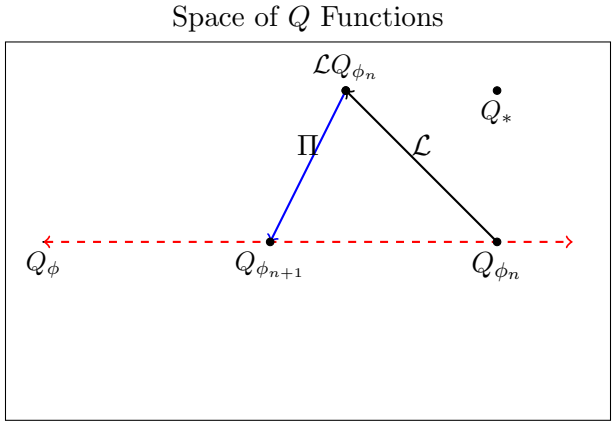
- Use the fixed point equation below to define a **contraction** operator $\mathcal{L}$ (contraction in L_∞ norm)

$$Q_*(s,a) \leftarrow \left[\sum_{s' \in \mathcal{S}} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} + \gamma \max_{a'} Q_*(s', a') \right) \right]$$

Convergence in TD setup

- State and action spaces are finite
- All state-action pairs are visited infinitely often
- Robbins-Monroe condition: $\sum_t \alpha_t = \infty$, $\sum_t \alpha_t^2 < \infty$

Projections and Convergence



Convergence Guarantee For Fitted Iteration Methods

- Define operator $\mathcal{L} : \mathcal{Q} \rightarrow \mathcal{Q}$ such that

$$\mathcal{L}Q = r + \gamma \max_{a'} Q(s', a')$$

- Backup operator $\mathcal{L}$ is a contraction in L_∞ norm s
- Projection operator (Π) are contractions in L_2 norm
- What about the composition $(\Pi \circ \mathcal{L})Q$?
  - ★ Need not be a contraction with respect to any norm

Sad Corollary

No guarantees on convergence to optimal value functions (on the manifold) exist for fitted iteration methods

Convergence of Monte Carlo Based Algorithm

Algorithm Monte Carlo Based Value Function Fitting

1: Initialize number of iterations N
2: **for** $i = 1$ to N **do**
3: Perform a roll-out from an initial state s_i (could be any state from $\mathcal{S}$)
4: Calculate targets y_i using Monte-Carlo roll outs

$$y_i = \left[\sum_{k=0}^{H} \left(\gamma^k r^i_{t+k+1} | s_t = s_i \right) \right]$$

5: Form input-output pairs (s_i, y_i) (N datapoints in total)
6: **end for**
7: Perform supervised regression with loss function

$$L(\phi) = \frac{1}{2} \sum_{i=1}^{N} \left[V^{\pi}_{\phi}(s_i) - y_i \right]^2$$

Convergence of Monte Carlo Based Algorithm

- Step 7 is gradient descent and it will converge at least local optimum
- Important : **Convergence guarantee is in the parameter space (ϕ) and not in value function space**

Algorithm Fitted Q Iteration

1: Initialize number of iterations N
2: **for** $j = 1$ to N **do**
3: Sample K transitions (s, a, r, s') using any behaviour policy μ
4: **for** $i = 1$ to K **do**
5: Calculate targets y_i using one step TD approximation

$$y_i = \left[r + \gamma \max_{a'} Q_{\phi_j}(s'_i, a') \right]$$

6: Form input-output pairs $(s_i, , y_i)$ (K Datapoints in total)
7: **end for**
8: Perform supervised regression (Optimizer : RProp) using loss function

$$L(\phi_j) = \frac{1}{2} \sum_{i=1}^{K} \left[Q_{\phi_j}(s_i, a_i) - y_i \right]^2$$

and get a new function approximator with new weights ϕ_{j+1}
9: **end for**

Online Q learning / Incremental Q learning

Question : Can we do the gradient update for every transition (s, a, r, s') ?

- We use the fitted Q iteration and set $K = 1$
- This is also the Watkins Q-learning update (used with function approximators)

Algorithm Online Q Learning

1: **for** $n = 1$ to N **do**
2: Take an action a and obtain the transition (s, a, r, s') using ϵ-greedy policy
3: Calculate target y using one step TD approximation

$$y = \left[r + \gamma \max_{a'} Q_{\phi_n}(s', a') \right]$$

4: Compute $g^{(n)} = \nabla_{\phi}(Q_{\phi_n}(s, a) - y)^2$
5: Set $\phi_{n+1} = \phi_n - \alpha g^{(n)}$
6: **end for**

Convergence Guarantee on Online Q learning

Algorithm Online Q Learning

1: **for** $n = 1$ to N **do**
2: Take an action a and obtain the transition (s, a, r, s') using ϵ-greedy policy
3: Calculate target y using one step TD approximation

$$y = \left[r + \gamma \max_{a'} Q_{\phi_n}(s', a') \right]$$

4: Compute $g^{(n)} = \nabla_\phi (Q_{\phi_n}(s, a) - y)$
5: Set $\phi_{n+1} \leftarrow \underbrace{\phi_n - \alpha g^{(n)}}_{\text{Is this GD ?}}$
6: **end for**

- Take a closer look at the one step gradient

$$g^{(n)} \leftarrow \phi_n - \alpha \nabla_\phi (Q_\phi(s, a) - \underbrace{\boxed{r + \gamma \max_{a'} Q_\phi(s', a')}}_{\text{moving target}})$$

Summary : Convergence Discussion

- Projection (Π) of the backup operator ($\mathcal{L}$) of optimal Q function need not be a contraction in any norm
- Fitted V iteration or fitted Q iteration need not converge because of the moving target problem
- In online Q learning algorithm,
  - ★ Samples obtained are sequentially correlated
  - ★ Moving target problem
- **Convergence guarantees exist only in tabular case**

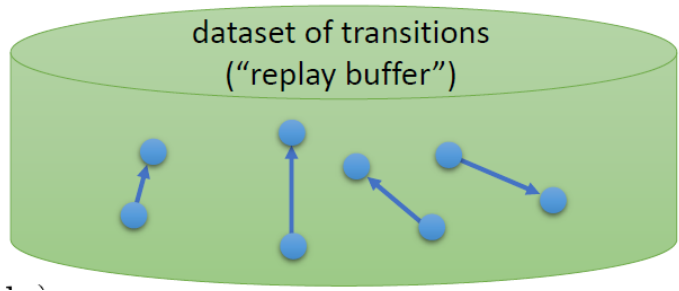
Towards a Stable Deep Q Network Algorithm

Desiredata

- Online algorithm like Q-learning in tabular case
- No sequential correlation in data samples
- Some stability with respect to gradient updates

Replay Buffers

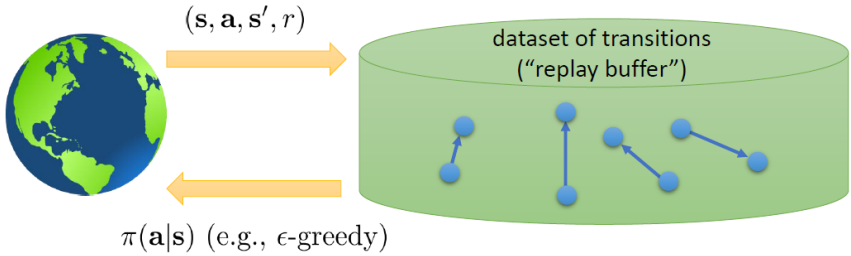
- Use the idea from fitted Q-iteration to collect and store transitions $(s, a.s', r)$



- Stored transition dataset is called **Replay Buffer** denoted by D
- Replay buffers are of fixed size (N)

Figure Source: Sergey Levine : UCB

Replay Buffers



- In an online setting, use ϵ-greedy policy to periodically feed the buffer with newer experiences
- Use FIFO like mechanism to maintain size
- Sample a random minibatch of transitions (B transitions) to perform gradient descent (random sampling ensure samples for SGD are no longer correlated)
- Variance of the gradient estimate is also low compared to gradient computed using one sample

Figure Source: Sergey Levine : UCB

Moving Target Problem - Target Networks

- Use an older set of weights to compute the targets
- Called **Target Network**
- Loss term is given by

$$L_i(\phi_i) = \left[\mathbb{E}_{(s,a,r,s') \in D} \left(Q_{\phi_i}(s,a) - \underbrace{r + \max_{a'} Q_{\phi_i'}(s',a')}_{\text{target}} \right)^2 \right]$$

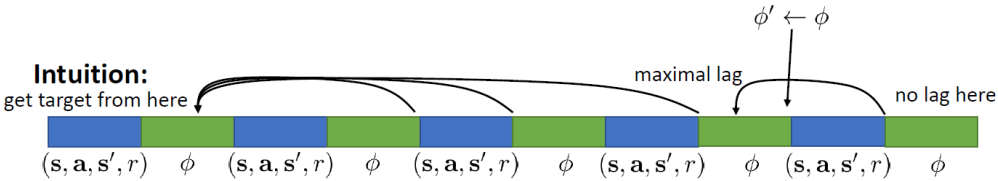
- Target network is kept constant for a while (every C steps) before being changed
  - ★ Every C steps the weights of the original network is copied to target network

DQN Algorithm

Algorithm DQN Algorithm

Initialize replay memory D to capacity N
Initialize action value function Q with parameters ϕ
Initialize target action value function $\widehat{Q}$ with parameters $\phi' = \phi$
for episodes $= 1$ to M **do**
    Initialize start state s_1
    for steps $t = 1$ to T **do**
        Select action a_t using ϵ-greedy policy
        Execute action a_t and store transition (s_t, a_t, r_t, s_{t+1}) in D
        Sample random minibatch (size B) of transitions from D
        for b $= 1$ to B **do**
            Calculate targets for each transitions (Bellman backup or reward)
        end for
        Perform a gradient descent step on $(y_i - Q_\phi(s_t, a_t))^2$ w.r.t ϕ
        Every C steps set $\widehat{Q} = Q$
    end for
end for

Alternative Target Network



Polyak Averaging

- Replace target network update step (Step 14) by

$$\phi' : \phi' \leftarrow \tau\phi' + (1 - \tau)\phi$$

- Typical value for $\tau = 0.99$

Figure Source: Sergey Levine : UCB

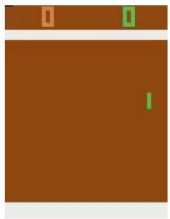
Efficacy of DQN Algorithm

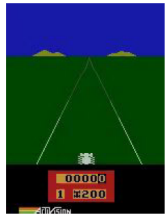
Historical Notes [2]

- Mnih et al. introduced Deep Q-Network (DQN) algorithm, applied it to ATARI games
- Used deep learning / ConvNets, published in early stages of deep learning craze (one year after AlexNet)
- Popularized ATARI (Bellemare et al., 2013) as RL benchmark
- Outperformed baseline methods, which used hand-crafted features

[2] Slide content from Schulman

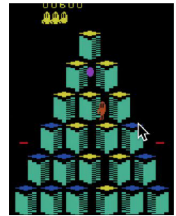
DQN on Atari [2]


Pong



Enduro

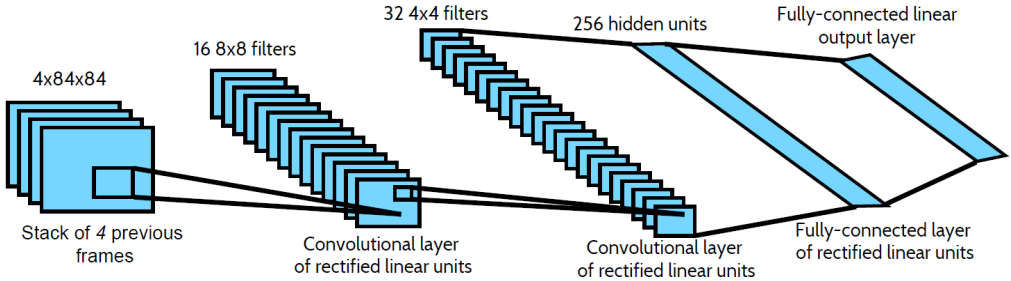
Beamrider


Q*bert

- 49 ATARI 2600 games
- From pixels to actions
- The change in score is the reward
- Same algorithm
- Same function approximator
- Same hyperparameters
- Roughly human-level performance on 29 out of 49 games

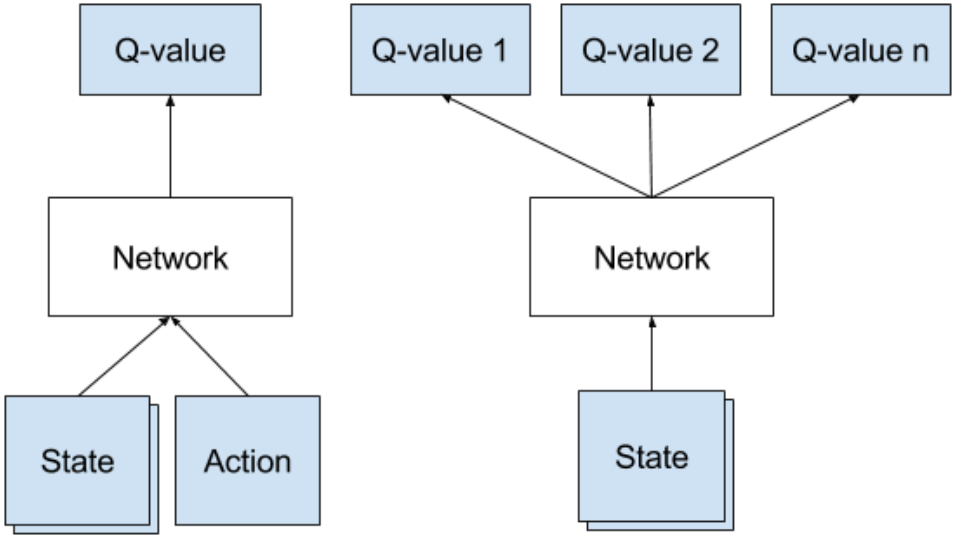
[2] Slide content from Minh

Atari DQN Architecture

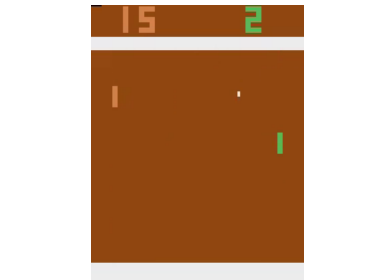


- Convolutional neural network architecture
- History of 4 frames as input
- One output per action ($Q(s, a)$) – expected reward for action a

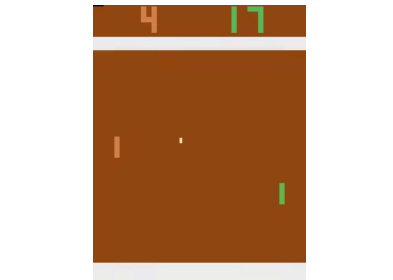
Profile of Q Function Approximator



Demonstration - Ping Pong

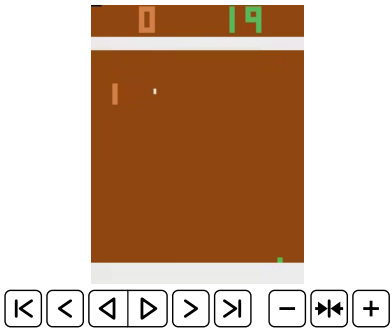


Random Policy

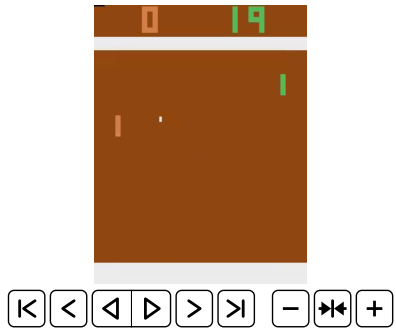


After 5.2 Millon Epochs

Demonstration - Ping Pong

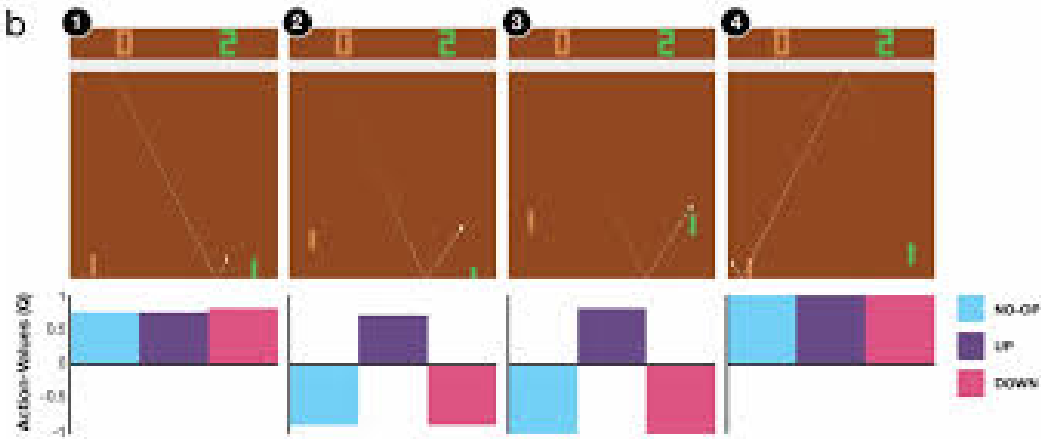


After 8 Million Epochs

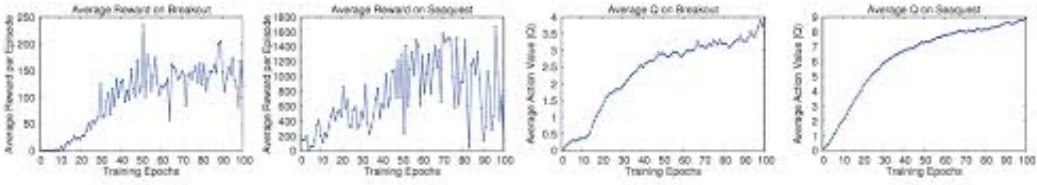


After 9.5 Millon Epochs

Are the Q-Values Meaningful ?



On Tracking the Training Process



Double DQN

Maximization Bias

- Consider single-state MDP with 2 actions.
- Both actions have zero mean rewards (the agent does not know this information)
- Let $\hat{Q}(\cdot, a_1)$ and $\hat{Q}(\cdot, a_2)$ be (unbiased) finite sample estimates of Q for action $a_!$ and a_2 respectively
- The agent will prefer the action which has maximum $\hat{Q}$ based on sample estimates, although both actions have same expected mean reward

Maximization Bias

- Consider single-state MDP with 2 actions.
- One action has $-\epsilon$ (ϵ positive and small) mean reward and the other action has zero mean reward.
- Let $\hat{Q}(\cdot, a_1)$ and $\hat{Q}(\cdot, a_2)$ be (unbiased) finite sample estimates of Q for action a_1 and a_2 respectively
- In this case, the agent might choose action a_1 , depending on how the maximum $\hat{Q}$ values that is based on sample estimates show up, although action a_2 is clearly better in expectation

Extending the analogy, the (tabular or deep) Q-learning algorithm may actiually pick the suboptimal action for target computation and this causes over estimation of Q-values.

Issues in Tabular Q-Learning Approach

- The problem with vanilla tabular Q-Learning is that the same samples are being used to decide which action is the best (highest expected reward), and the same samples are also being used to estimate that action-value
- Break up the Q-learning update rule as follows

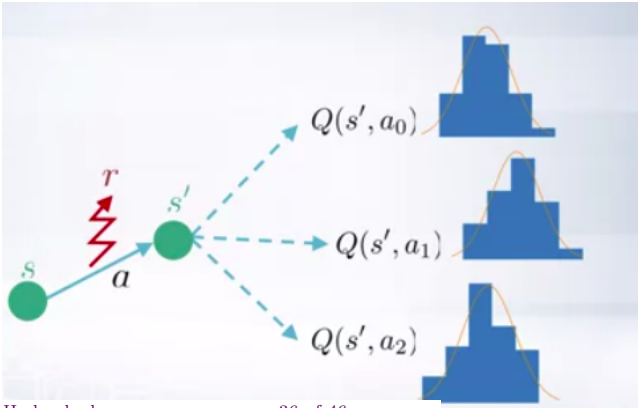
$$Q(s_t, a) \leftarrow Q(s_t, a) + \alpha \left(r_{t+1} + \gamma Q(s_{t+1}, \arg\max Q(s_{t+1}, a)) - Q(s_t, a) \right)$$

- If action value is overestimated, then it is chosen as the best action, and its overestimated value is used as the target

Slide Credit: Medium Blog : Ameet Deshpande

Persistence of the Problem in DQN

- For transition $(s_t, a_t, r_{t+1}, s_{t+1})$ the TD target of the Q-value update is $r_{t+1} + \gamma \arg\max_{a'} Q_\phi(s_{t+1}, a')$
- $Q_\phi(\cdot, \cdot)$ is a noisy estimate during training phase
- Therefore, $\max Q_\phi(\cdot, \cdot)$ would typically be overestimated during training



Evidence from Atari Games

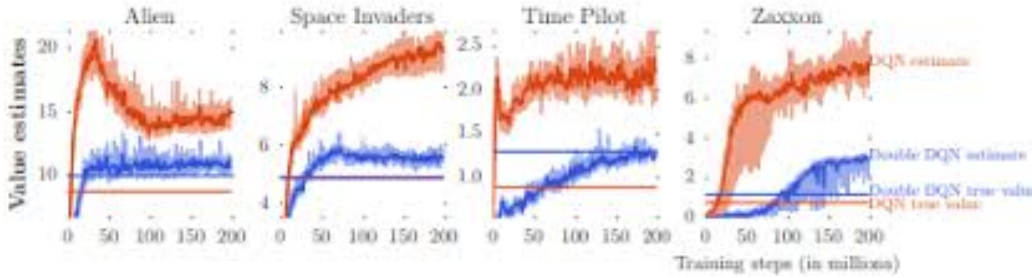


Figure Source: van Hasselt : DDQN paper

Possible Resolution ...

- There are two identical fair coins (we don't know they are fair)
- If a coin lands on head, we get one dollar; otherwise we lose a dollar
- Interested in answering the following questions
  - ★ Which coin will yield will more money in future flips ?
  - ★ How much can we expect to win or lose per flip using the coin from previous question ?
- Two ways to answer
  - ★ Flip each coin n few times and answer both questions
  - ★ Flip each coin n_1 times, answer the first question; collect fresh n_2 samples to answer second question based on the answer to the first question

The idea behind the second method is that we use separate samples to choose the best action and separate samples to use Q-values

Solution to Maximization Bias

- Have two different set of samples to decide the action and to evaluate the target
- The idea is to use two Q functions
- In the tabular Q-learning setting, for each transition quadruple $(s_t, a_t, r_{t+1}, s_{t+1})$ we flip a fair coin to decide any of the two update steps given below,

$$Q_1(s_t, a_t) \leftarrow Q_1(s_t, a_t) + \alpha \left(r_{t+1} + \gamma Q_2(s_{t+1}, \arg\max Q_1(s_{t+1}, a)) - Q_1(s_t, a_t)\right)$$

$$Q_2(s_t, a_t) \leftarrow Q_2(s_t, a_t) + \alpha \left(r_{t+1} + \gamma Q_1(s_{t+1}, \arg\max Q_2(s_{t+1}, a)) - Q_2(s_t, a_t)\right)$$

Double DQN

- In the DQN setting, we already have two Q networks (to address the moving target problem). So, we can take advantage of that by using the following update rule.
- In Double DQN, targets for the transition $(s_t, a_t, r_{t+1}, s_{t+1})$ is computed as follows,

$$Q^{original}(s_t, a_t) \leftarrow r + \gamma Q^{target}(s, \arg\max Q^{original}(s, a))$$

- In the above equation, the $\leftarrow$ actually means assigning targets which can then be picked up while sampling the replay buffer
- The fundamental idea is to use two Q_ϕ's (both are noisy estimates of true Q) in different ways so that the overestimation problem mellows down

Priortized Experience Replay

Priortized Experience Replay

- Replaying all transitions with equal probability is suboptimal
- Replay transitions in proportion to absolute Bellman error

$$\left| r + \gamma \max_{a'} Q_{\phi'}(s', a') - Q_{\phi}(s, a) \right|$$

- Leads to much faster learning

Priortized Experience Replay

- TD error for vanilla DQN is

$$\delta_i = r_t + \gamma \max_{a \in \mathcal{A}} Q_{\phi'}(s_{t+1}, a) - Q_{\phi}(s_t, a_t)$$

- TD error for DDQN is

$$\delta_i = r_t + \gamma Q_{\phi'}(s_{t+1}, \text{argmax}_{a \in \mathcal{A}} Q_{\phi}(s_{t+1}, a)) - Q_{\phi}(s_t, a_t)$$

- Priority for each entry in replay buffer D is given by $p_i = |\delta_i| + \epsilon$

Priortized Experience Replay

- Sample from replay buffer according to probability distribution

$$P(i) = \frac{p_i^\alpha}{\sum_k p_k^\alpha}$$

with α determining the level of prioritization

- In order to compute the expectation

$$\min_\phi \mathbb{E}_{(s_t,a_t,r_t,s_{t+1})\sim D}\left[\left(r_t + \gamma \max_{a\in\mathcal{A}} Q_{\phi'}(s_{t+1},a) - Q_\phi(s_t,a_t)\right)^2\right],$$

it is essential to use the importance sampling weights in each mini-batch of the gradient update

$$w_i = \left(\frac{1}{N}\cdot\frac{1}{P(i)}\right)^\beta$$

- The parameter β is an annealing term that is low (0.4 to 0.8) in the beginning of the training and tends to 1 towards the end of the training.
- The question on optimal choices of α and β during various phases of training depends on task at hand

Practical Tips for DQN [2]

- DQN is more reliable on some tasks than others. Test your impementation on reliable tasks like Pong and Breakout: if it doesn't achieve good scores, something is wrong
- Large replay buffers improve robustness of DQN, and memory efficiency is important
- DQN converges slowly - for ATARI it is often necessary to wait for 10-40 million frames (couple of hours to a day of training on GPU) to see results significantly better than random policy. **Be Patient**
- Always run at least two different seeds when experimenting
- Learning rate scheduling is beneficial. Try high learning rates in initial exploration period

[2] Slide content from Schulman

Practical Tips for DQN [2]

- Try non-standard exploration schedules
- Do use Double DQN with priortized experience replay – significant improvement
- Use Huber loss on Bellman error

$$L_\delta(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } |a| \leq \delta, \\ \delta(|a| - \frac{1}{2}\delta), & \text{otherwise.} \end{cases}$$