Lecture 24 - April 14, 2023

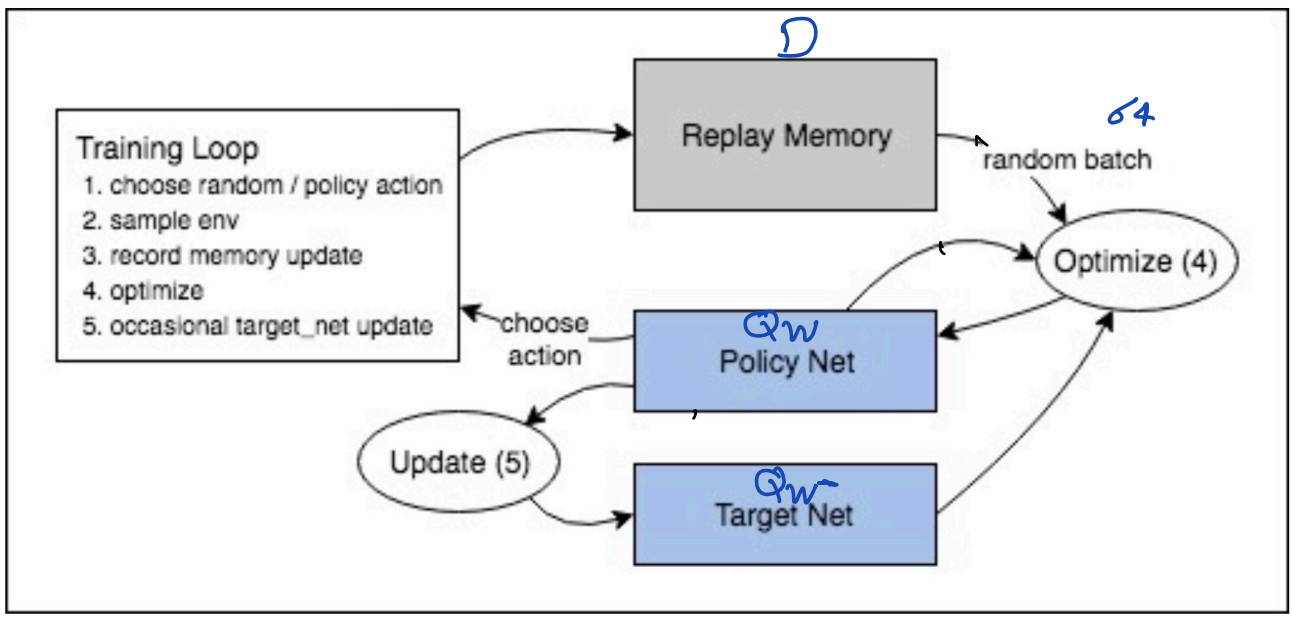
- Deep Q-Network (DQN) → Finite action
  - Deep Q-Network (DQN)
  - Double DQN
  - Prioritized DQN
  - Dueling DQN
  - Noisy-Net DQN
- Deep Policy Gradients (DPG) → large/continus Action
  - REINFORCE
  - REINFORCE with Baseline
  - Advantage Actor Critic (A2C)
  - Deep Deterministic Policy Gradiat (DDPG)

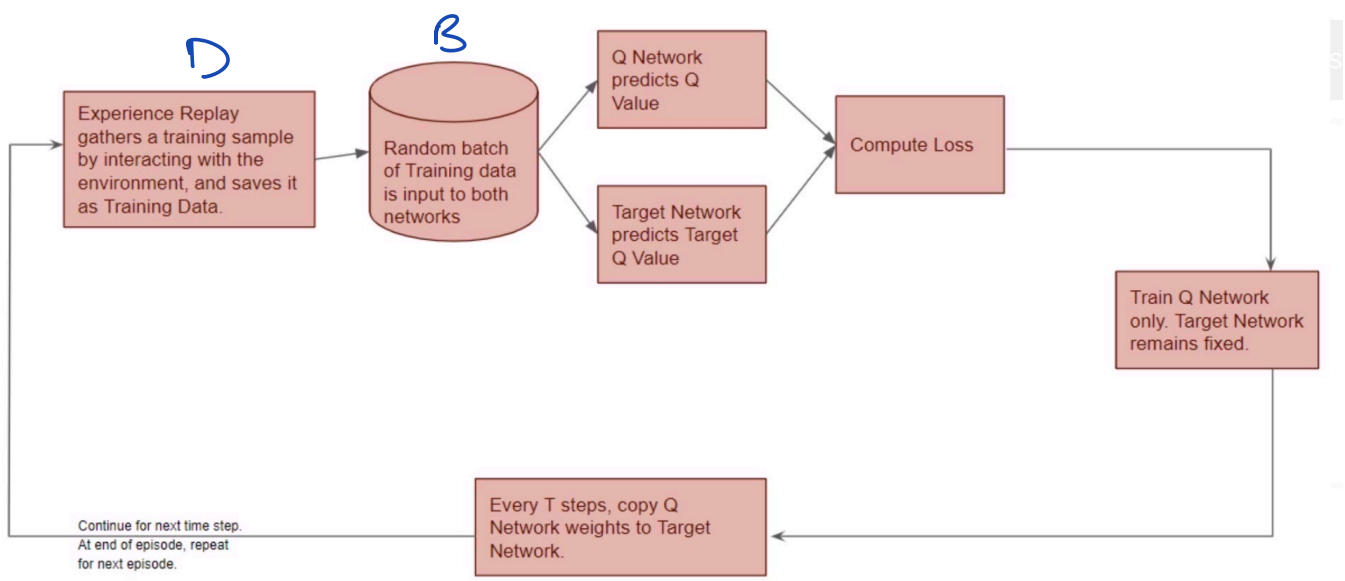
Project 3 → Due April 17

HW5 → Due April 18

TA's office hour: Wendsdays, 2pm-3pm (in-person)

Fridays, 2pm-3pm (virtual)


Training Loop
1. choose random / policy action
2. sample env
3. record memory update
4. optimize
5. occasional target_net update
Replay Memory
random batch
Optimize (4)
choose action
Policy Net
Update (5)
Target Net

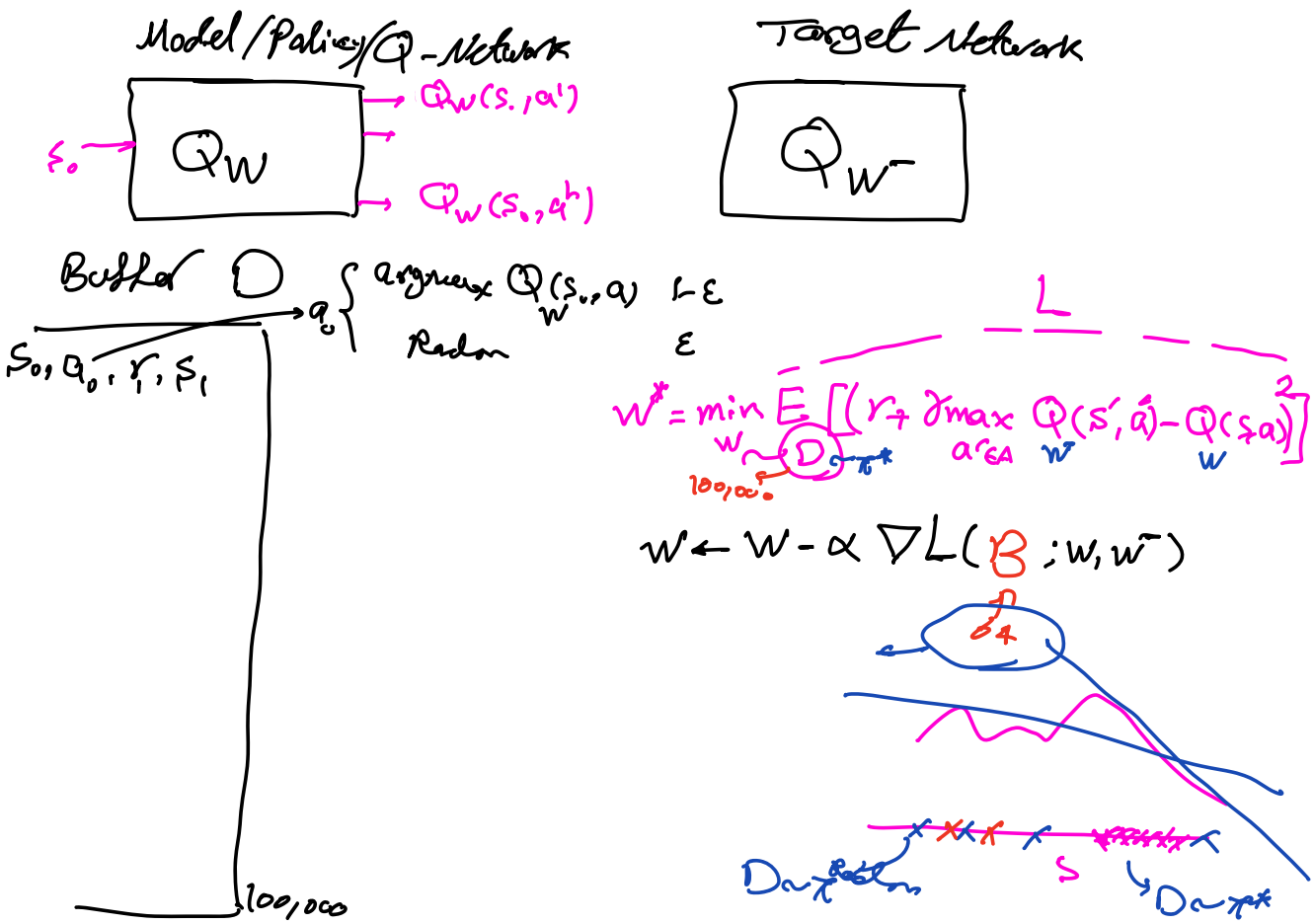

Experience Replay gathers a training sample by interacting with the environment, and saves it as Training Data.
Random batch of Training data is input to both networks
Q Network predicts Q Value
Target Network predicts Target Q Value
Compute Loss
Train Q Network only. Target Network remains fixed.
Every T steps, copy Q Network weights to Target Network.
Continue for next time step.
At end of episode, repeat for next episode.

DQN

Initialize Q_w, Q_{w^-} with random weights
$D \leftarrow \emptyset$
Repeat (for each episode):
  Initialize *s*
  Repeat (for each step of the episode):
    Choose *a* from *s* using policy derived from Q (e.g., e-greedy)
    Take action *a*, observe *r, s'*
    $D \leftarrow D \cup (s, a, s', r)$
    $s \leftarrow s'$
    If mod(*step,trainfreq*) == 0:
      sample batch *B* from *D*
      $w \leftarrow w - \alpha \nabla_w L(B; w, w^-)$
      if mod(*step,copyfreq*) == 0:
        $w^- \leftarrow w$

Where: $\nabla_w L(B; w, w^-) \approx -\frac{1}{|B|} \sum_{(s,a,s',r) \in B} \left(target(s'; w^-) - Q_w(s,a)\right) \nabla_w Q_w(s,a)$

$$target(s'; w^-) = r + \gamma \max_{a'} Q_{w^-}(s', a')$$

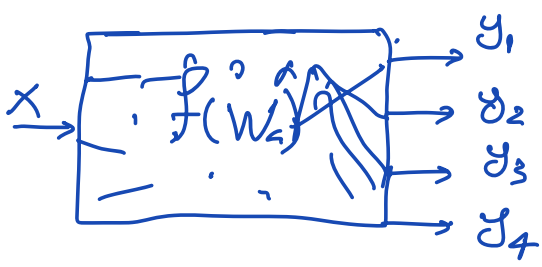


Double DQN

Initialize Q_w, Q_{w^-} with random weights
$D \leftarrow \emptyset$
Repeat (for each episode):
  Initialize *s*
  Repeat (for each step of the episode):
    Choose *a* from *s* using policy derived from Q (e.g., e-greedy)
    Take action *a*, observe *r, s'*
    $D \leftarrow D \cup (s, a, s', r)$
    $s \leftarrow s'$
    If mod(*step,trainfreq*) == 0:
      sample batch *B* from *D*
      $w \leftarrow w - \alpha \nabla_w L(B; w, w^-)$
      if mod(*step,copyfreq*) == 0:
        $w^- \leftarrow w$

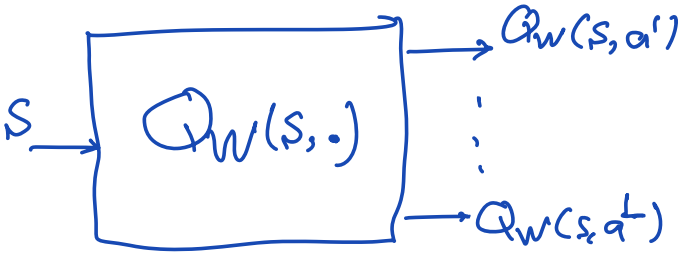
Where: $\nabla_w L(B; w, w^-) \approx -\frac{1}{|B|} \sum_{(s,a,s',r) \in B} \left(target(s', a'; w, w^-) - Q_w(s, a) \right) \nabla_w Q_w(s, a)$

$$target(s', a'; w, w^-) = r + \gamma Q_{w^-}(s', \arg\max_{a'} Q_w(s', a'))$$



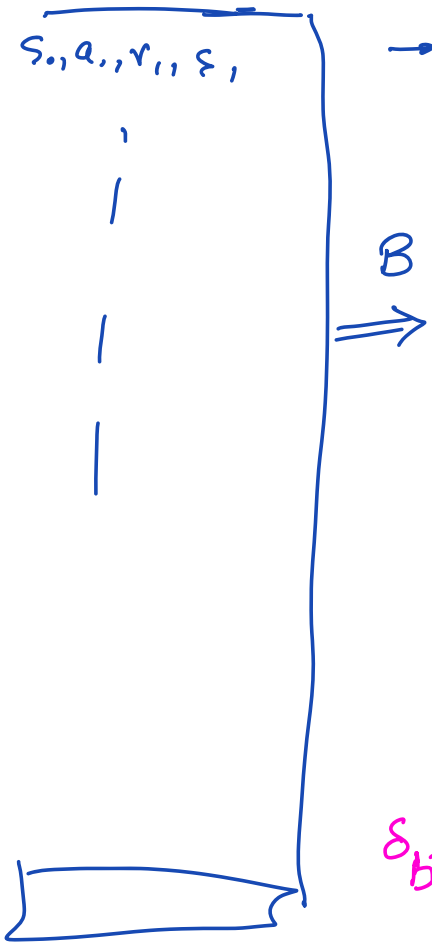
Training
x_1 y_1 = cat

Inference



$D \sim B$ s_0, a_0, r_1, s_1

$$D \quad w \leftarrow w + \alpha \frac{1}{|B|} \sum_{s,a,r,s' \in B} \left(r + \gamma \max_{a'} Q_{\bar{w}}(s',a') - Q_w(s,a) \right) \nabla_w Q_w(s,a)$$



$$\rightarrow \delta_1 = | r_1 + \gamma \max_{a'} Q_{\bar{w}}(s_1,a') - Q_w(s_0,a_0) |$$

$$\delta_2 = | r_2 + \gamma \max_{a'} Q_{\bar{w}}(s_2,a') - Q_w(s_1,a_1) |$$

$$\delta_b = | r_b + \gamma \max_{a'} Q_{\bar{w}}(s_1,a') - Q_w(s_0,a_0) |$$

Prioritized DQN:

$$P(i) = \frac{\delta_i + \epsilon}{\sum_{j=1}^{|D|} \delta_j + \epsilon}$$

Ranked →

$$P(i) \propto \frac{1}{\text{Rank}(i)}$$

Real Large

1

1

Small

100,000

$$P_i = \frac{P(i)^{\alpha}}{\sum_{j=1}^{|D|} P(j)^{\alpha}} \rightarrow \text{level of priority}$$

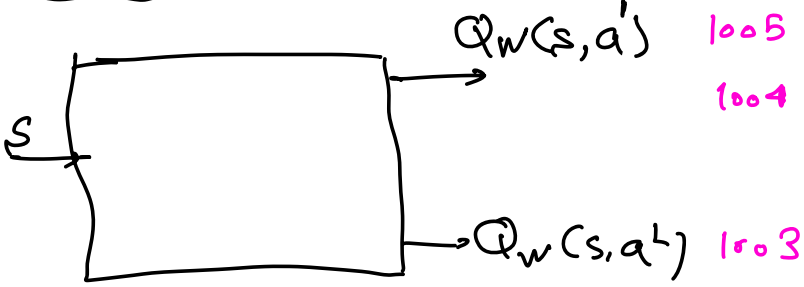
Random B $\Rightarrow$ $w \leftarrow w + \alpha \frac{1}{|B|} \sum_{s,a,r,s' \in B} \left(r + \gamma \max_{a'} Q_{\bar{w}}(s',a') - Q_w(s,a)\right) \nabla_w Q_w(s,a)$

Priority $\Rightarrow$ $w \leftarrow w + \alpha \frac{1}{|B|} \sum_{s,a,r,s' \in B} \frac{1}{P_{s,a,s',r}} \left(r + \gamma \max_{a'} Q_{\bar{w}}(s',a') - Q_w(s,a)\right) \nabla_w Q_w(s,a)$

- Priodically updating δ_i
- only for minbatch selected
- Stochastic update $\rightarrow$ 1000 accordingly δ_i

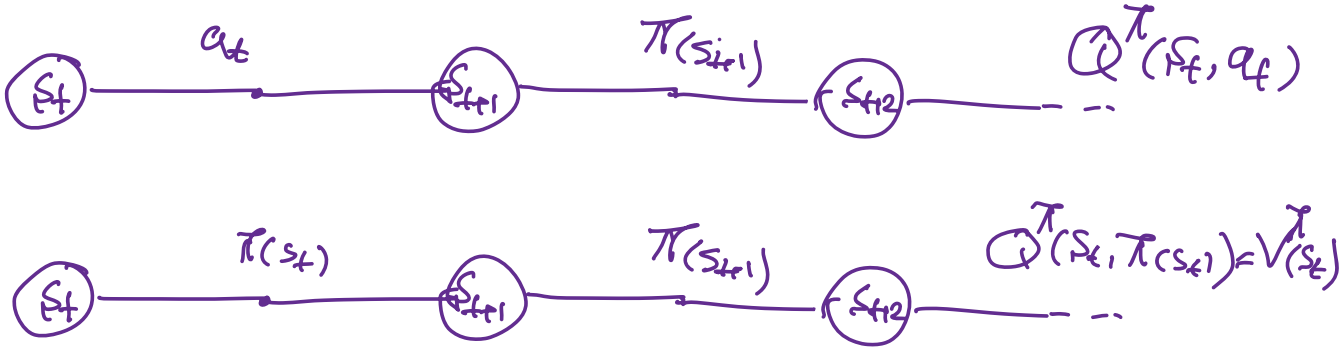
Dueling DQN

Advantage Function



advantage Function

$$Q(s,a) = V(s) + A(s,a)$$

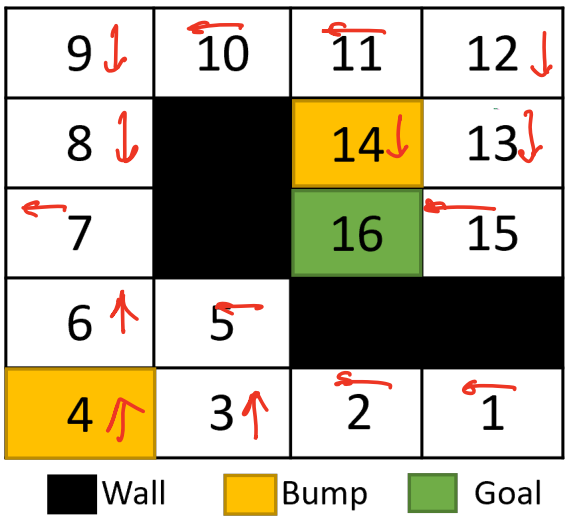


$V^{\pi}(15) = 99$

$Q^{\pi}(15, U) = -1 - 1 + 99 = 97$

$Q^{\pi}(15, R) = -1 + 99 = 98$

$Q^{\pi}(15, D) = -1 + 99 = 98$



$$\begin{cases} A^{\pi}(15, U) = \underbrace{Q^{\pi}(15, U)}_{97} - \underbrace{V^{\pi}(15)}_{99} = -2 \\ A^{\pi}(15, R) = -1 \\ A^{\pi}(15, D) = -1 \\ A^{\pi}(15, L) = Q^{\pi}(15, L) - V^{\pi}(15) = 0 \end{cases}$$

$$\underset{a \in A}{\text{Argmax}}\, Q^{\pi}(s, a) = \underset{a \in A}{\text{argmax}}\, A^{\pi}(s, a) + V^{\pi}(s)$$

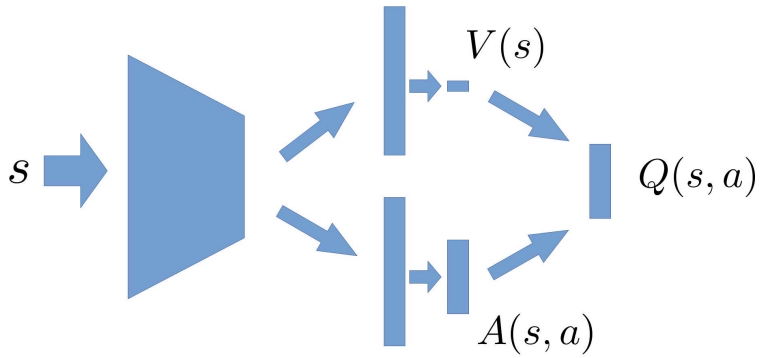
$V^{\pi}(11) = -100$

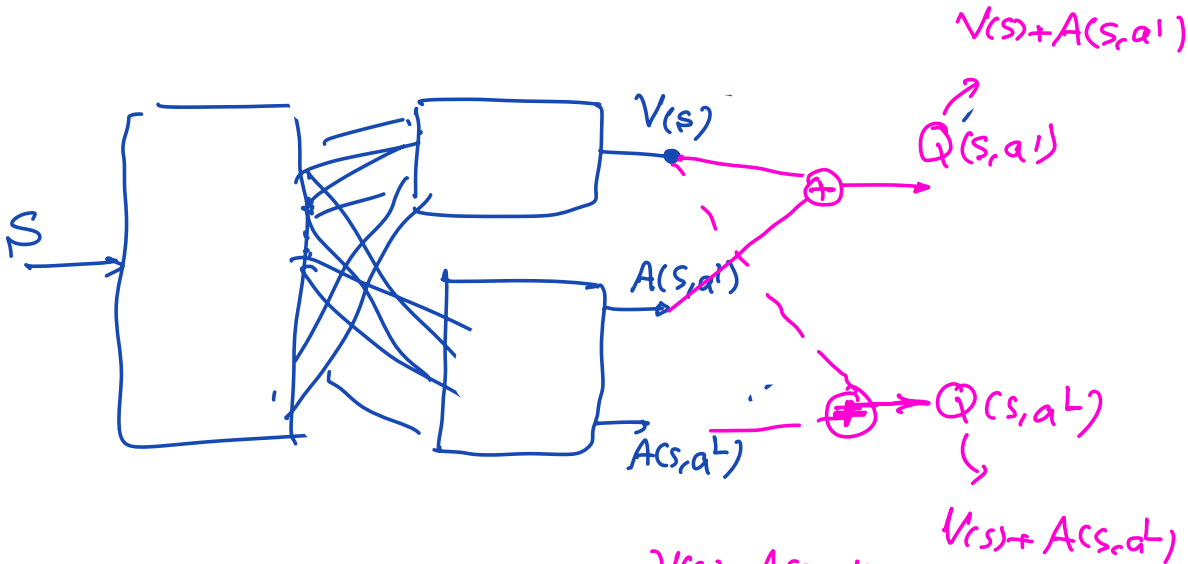
$Q^{\pi}(11, R) = 98$

$Q^{\pi}(11, L) = -100$

$Q^{\pi}(11, U) = 88$

$$\Rightarrow \begin{cases} A^{\pi}(11, R) = \underbrace{Q^{\pi}(11, R)}_{98} - \underbrace{V^{\pi}(11)}_{-100} \\ \qquad = 198 \\ A^{\pi}: \\ \vdots \end{cases}$$





$V(s)+A(s,a^1)$

$V(s) = 6$

$A(s,a^1) = 0$

$A(s,a^2) = 1$

$\longrightarrow$

$Q(s,a^1) = 6$

$Q(s,a^2) = 7$

possible solution

$V(s) = 1$

$A(s,a^1) = 5$

$A(s,a^2) = 6$

$V(s) = 5$

$A(s,a^1) = 1$

$A(s,a^2) = 2$

$$\max_{a \in A} Q(s,a) = V(s)$$

$\rightarrow a^{max}$

$$A(s, a^{max}) = V(s) - Q(s, a^{max}) = 0$$

$$Q_W(s,a) = V_W(s) + A_W(s,a) - \max_{a' \in A} A_W(s,a')$$

$V(s) = 6$

$A(s,a^1) = 0$

$A(s,a^2) = 1$

$\rightarrow$ Prediction NN

$Q(s,a^1) = 6 + 0 = 6$

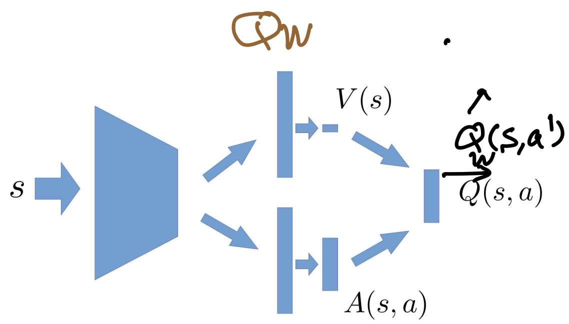
$Q(s,a^2) = 6 + 1 = 7$

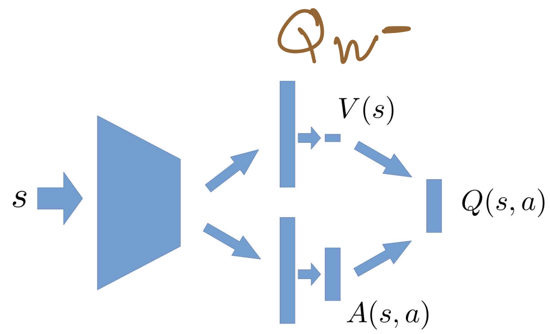
$$\max_{a \in A} Q(s,a) = V(s)$$

$\neq$

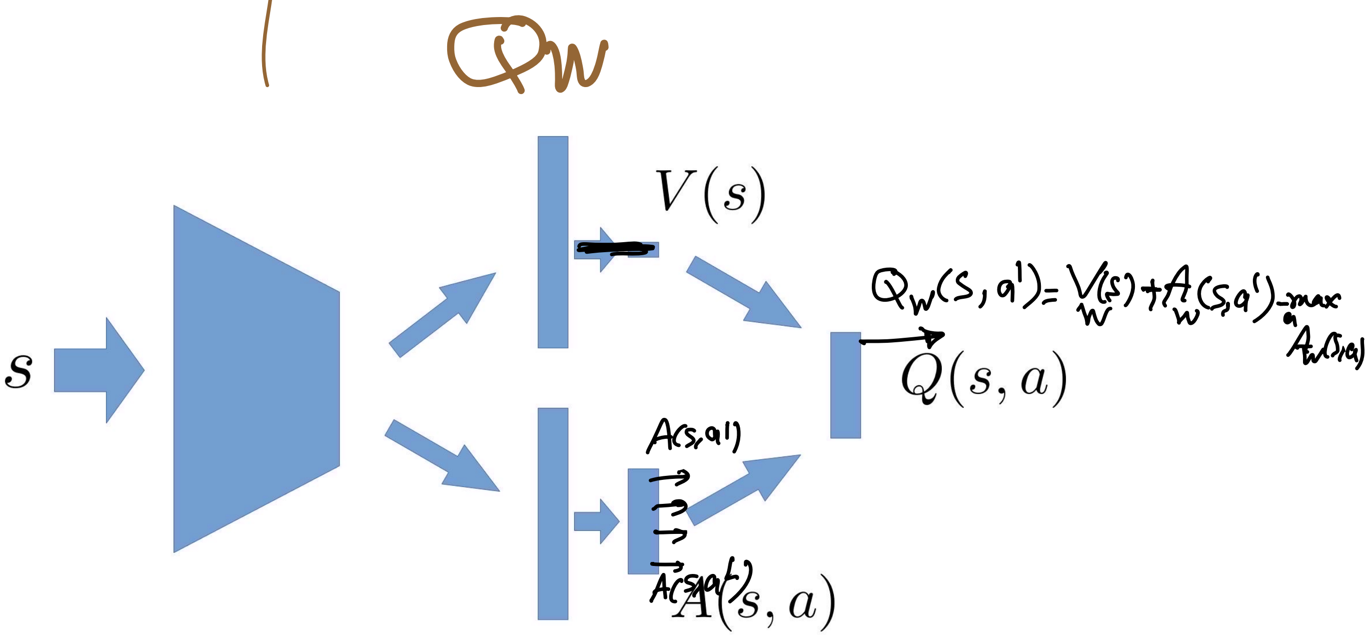
$$Q_W(s,a^1) = \underbrace{V_W(s)}_{6} + \underbrace{A_W(s,a^1)}_{0} - \underbrace{\max_{a' \in A} A_W(s,a')}_{1} = 5$$

$$Q_W(s,a^2) = \underbrace{V_W(s)}_{6} + \underbrace{A_W(s,a^2)}_{1} - \underbrace{\max_{a'} A_W(s,a')}_{1} = 6$$





$$L(B; w, w^-) = \frac{1}{B} \sum_{s,a,r,s' \in B} \left(y_i - Q_w(s,a) \right)^2$$



100 5

1004

1008

$$Q_w(s,a) = V_w(s) + A_w(s,a) - \frac{1}{|A|}\sum_{a' \in A} A_w(s,a')$$

$V(s) = 6$

$A(s,a^1) = 0$

$A(s,a^2) = 1$

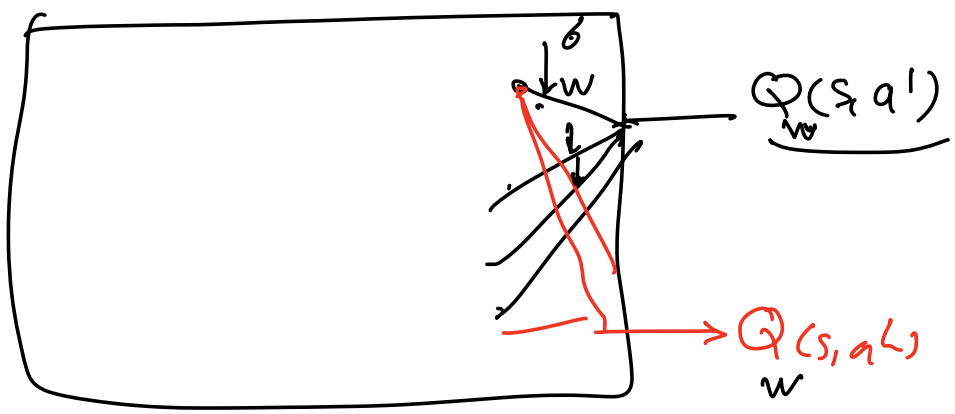
$$Q_w(s,a^1) = \underbrace{V_w(s)}_{6} + \underbrace{A(s,a^1)}_{0} - \frac{1}{2}\left(\underbrace{A(s,a^1)}_{0} + \underbrace{A(s,a^2)}_{1}\right)$$

$$= 5.5$$

$$Q_w(s,a^2) = \underbrace{V_w(s)}_{6} + \underbrace{A(s,a^2)}_{1} - \frac{1}{2}\left(\underbrace{A(s,a^1)}_{0} + \underbrace{A(s,a^2)}_{1}\right)$$

$$= 6.5$$

Noisy-Net DQN



$$y = \mu_w + \sigma_w \odot \varepsilon$$

$\varepsilon \leftarrow \mathcal{N}(0, 1)$

Instead of ε-greedy $\Rightarrow$ Random layer at the end of NN

$s_0, a \rightarrow \text{argmax}\, Q_w(s_0, a)$