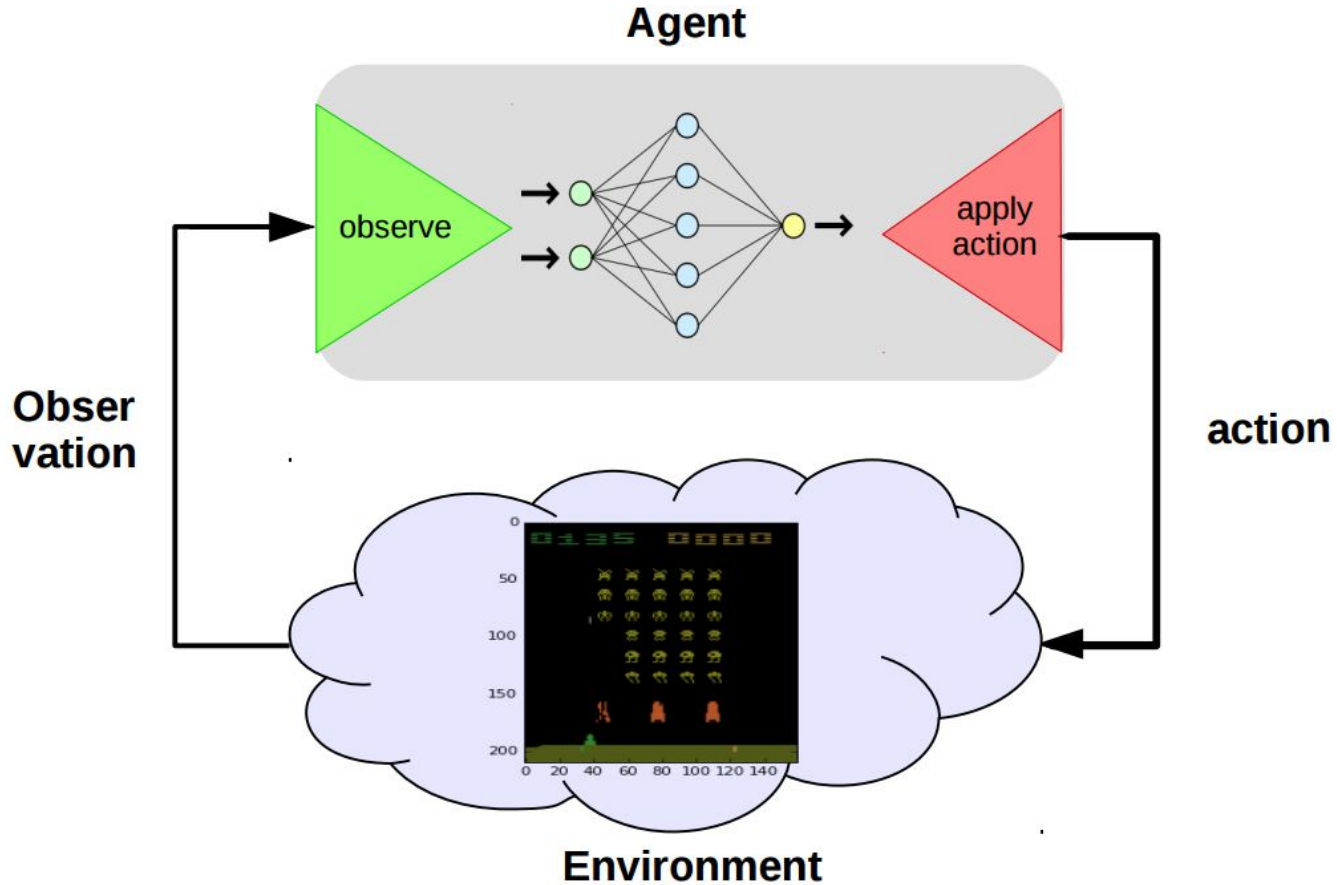


Reinforcement learning

Episode 4

Deep reinforcement learning

MDP



Basic definitions

$$G_t = \sum_{t'=t}^T \gamma^{(t'-t)} r_{t'}$$

$$Q^\pi(s, a) = E_\pi[G_t | s_t = s, a_t = a]$$

$$V^\pi(s) = E_\pi[G_t | s_t = s]$$

Recurrent relations

$$Q^\pi(s, a) = E_{s_{t+1}}[r_t + \gamma V^\pi(s_{t+1})]$$

$$Q^\pi(s, a) = E_{s_{t+1}, a_{t+1} \sim \pi}[r_t + \gamma Q^\pi(s_{t+1}, a_{t+1})]$$

Optimal policy

For all π, s, a : $Q^{\pi^*}(s, a) \geq Q^{\pi}(s, a)$

$$\pi^*(s) = \operatorname{argmax}_a Q^{\pi^*}(s, a)$$

Bellman optimality equation

$$Q^*(s_t, a) = E_{s_{t+1}}[r_t + \max_{a'} Q^*(s_{t+1}, a')]$$

Q-learning

Training step

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t))$$

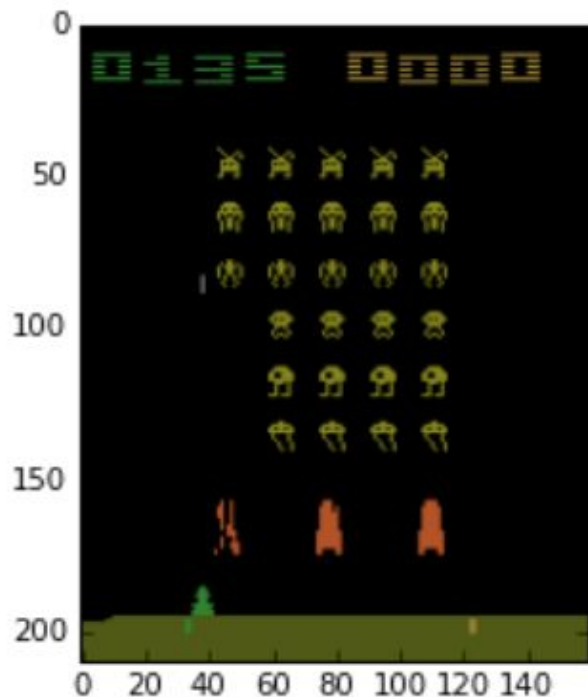
Q-learning as MSE minimization

$$L = (r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t))^2$$

$$\nabla L = 2 \cdot \underbrace{(r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t))}_{\text{Const}}$$

What's wrong here?

Real world



How many states are there?
approximately

$$|S| = 2^{210 \cdot 160 \cdot 8 \cdot 3}$$

Problem:

State space is usually large,
sometimes continuous.

Two solutions:

- Binarize state space (last week)
- Approximate agent with a function (crossentropy method)

Which one would you prefer for atari?

Problem:

State space is usually large,
sometimes continuous.

And so is action space;

Two solutions:

- Binarize state space  Too many bins or handcrafted features
- Approximate agent with a function  Let's pick this one

From tables to approximations

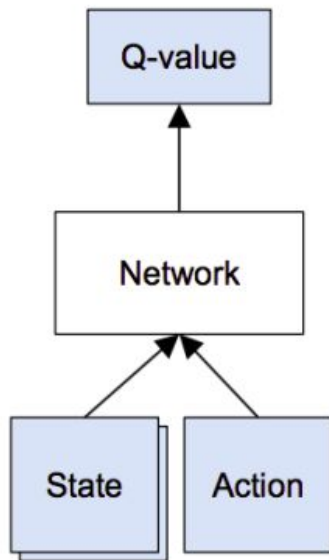
- Before:
 - For all states, for all actions, remember $Q(s,a)$
- Now:
 - Approximate $Q(s,a)$ with some function
 - e.g. linear model over state features

$$\operatorname{argmin}_{w,b} (Q(s_t, a_t) - [r_t + \gamma \cdot \max_{a'} Q(s_{t+1}, a')])^2$$

Question: should we use **classification** or **regression** model?
(e.g. logistic regression Vs linear regression)

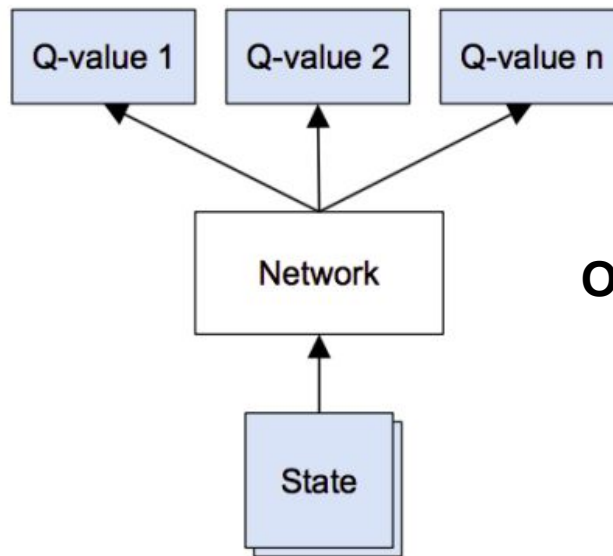
Possible architectures

**Continuous
control or large
number of
actions**



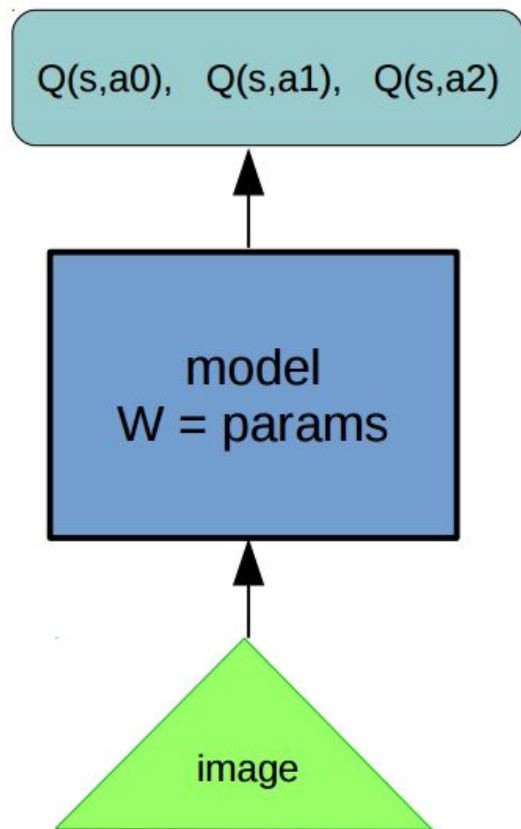
Given **(s,a)**
Predict $Q(s,a)$

**One pass for all
actions**



Given **s** predict all q-values
 $Q(s,a_0)$, $Q(s,a_1)$, $Q(s,a_2)$

Approximate Q-learning



Q-values:

$$\hat{Q}(s_t, a_t) = r + \gamma \cdot \max_{a'} Q(s_{t+1}, a')$$

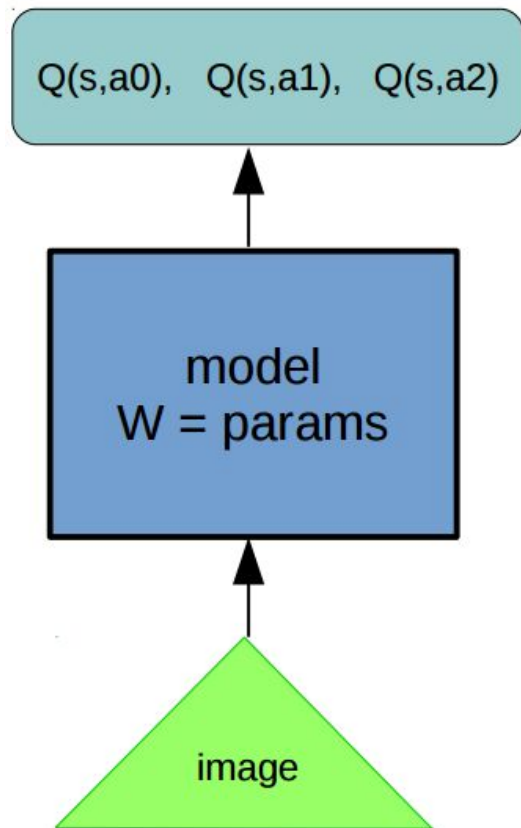
Objective:

$$L = (Q(s_t, a_t) - [r + \gamma \cdot \max_{a'} \underbrace{Q(s_{t+1}, a')}_{\text{Const}}])^2$$

Gradient step:

$$w_{t+1} = w_t - \alpha \cdot \frac{\delta L}{\delta w}$$

Approximate Q-learning



Objective:

$$L = \left(Q(s_t, a_t) - \hat{Q}(s_t, a_t) \right)^2$$

consider const

Q-learning:

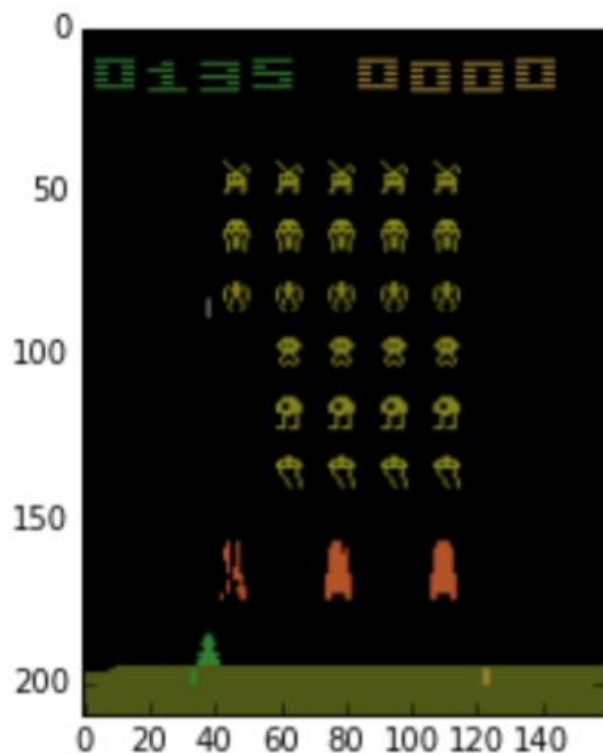
$$\hat{Q}(s_t, a_t) = r + \gamma \cdot \max_{a'} Q(s_{t+1}, a')$$

SARSA:

$$\hat{Q}(s_t, a_t) = r + \gamma \cdot Q(s_{t+1}, a_{t+1})$$

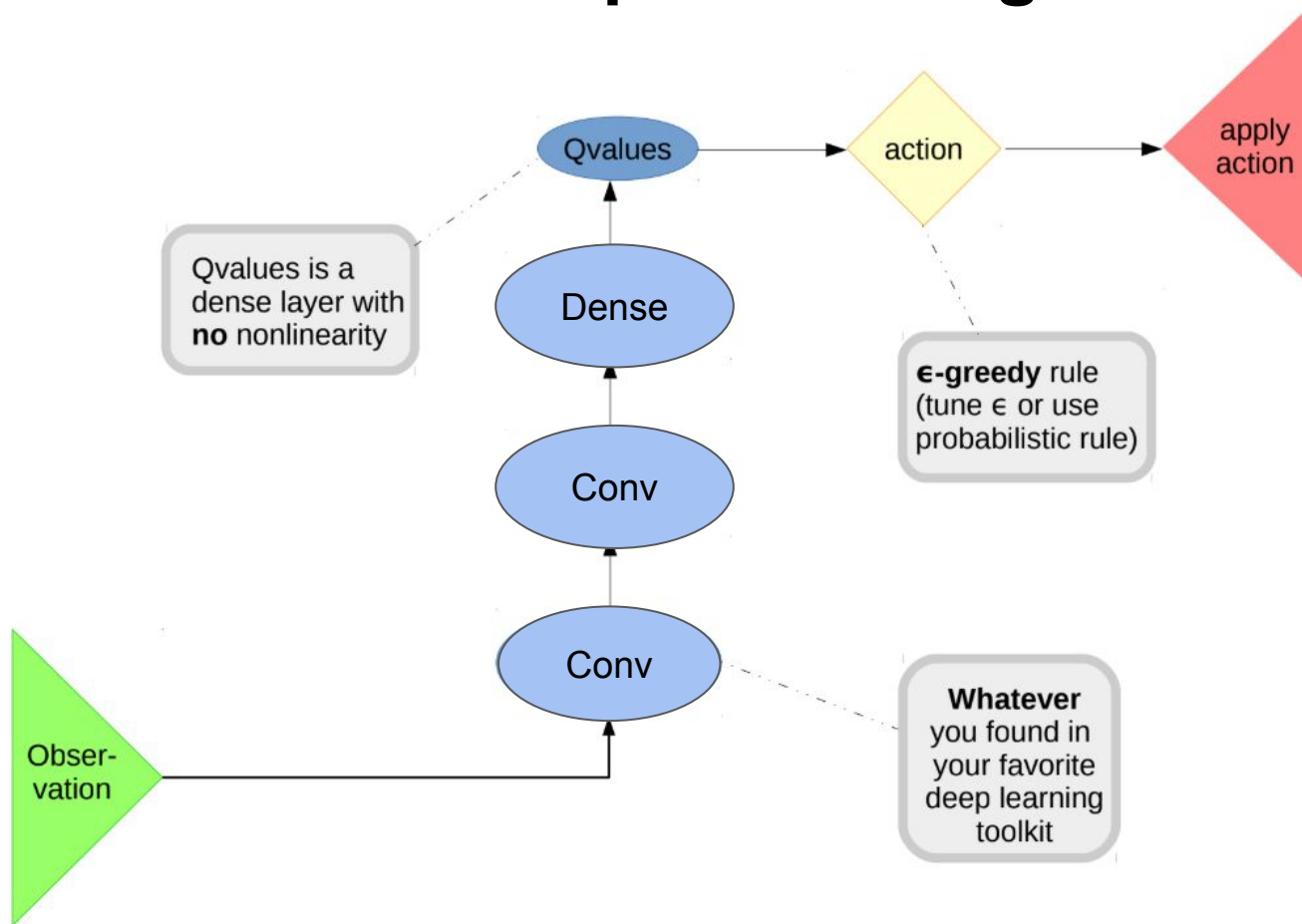
Expected Value SARSA:

$$\hat{Q}(s_t, a_t) = r + \gamma \cdot E_{a' \sim \pi(a|s)} Q(s_{t+1}, a')$$

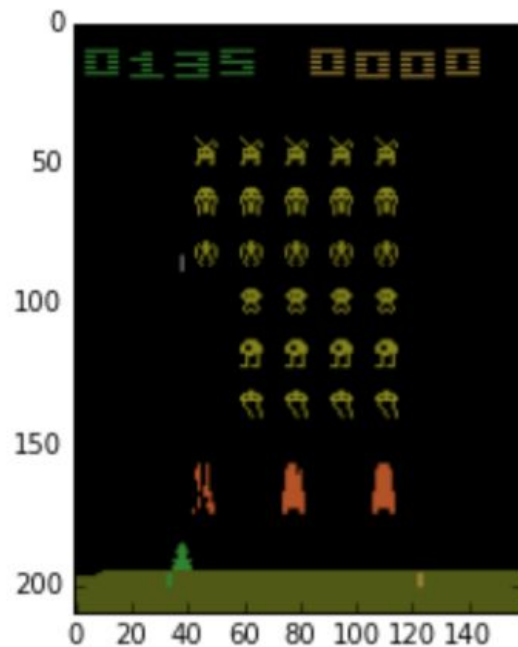


What kind of network digests images well?

Basic deep Q-learning



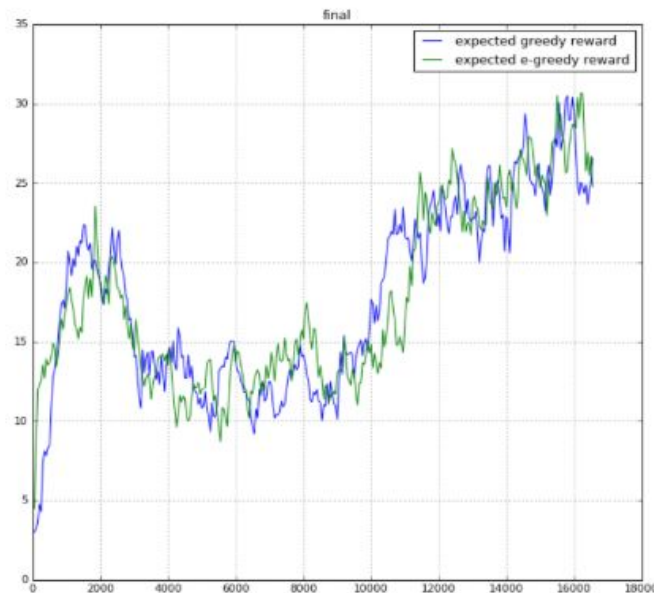




How bad it is if agent spends
next 1000 ticks under the left rock?
(while training)

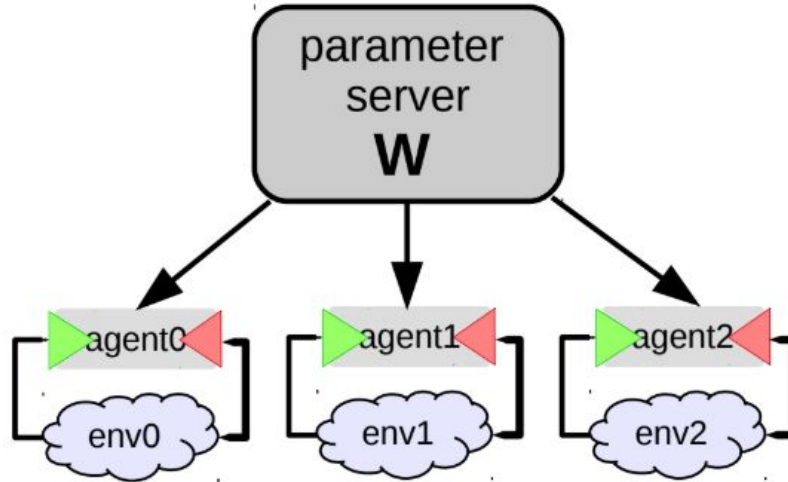
Problem

- Training samples are **not** “i.i.d”,
- Model forgets parts of environment it hasn't visited for some time
- Drops on learning curve
- **Any ideas?**



Multiple agent trick

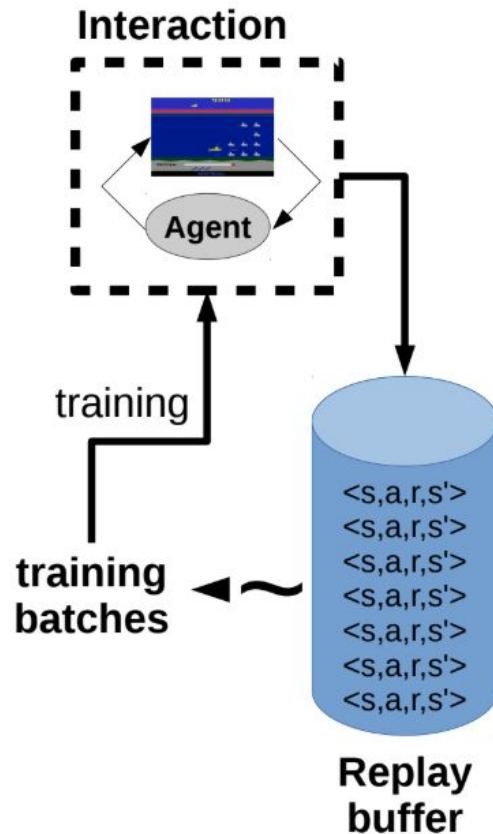
Idea: Throw in several agents with shared W .



Experience replay

Idea: store several past interactions
 $\langle s, a, r, s' \rangle$
Train on random subsamples

- Closer to i.i.d
pool contains several sessions
- Older interactions were obtained
under weaker policy



Experience replay

- You approximate $Q(s,a)$ with a neural network
- You use **experience replay** when training

Question: which of those algorithms will fail?

- Q-learning
- SARSA
- CEM
- Expected Value SARSA

Experience replay

- You approximate $Q(s,a)$ with a neural network
- You use **experience replay** when training

Agent trains off-policy on an older version of him

Question: which of those algorithms will fail?

Off-policy methods work, On-policy is super-slow (fail)

- | | |
|--------------|------------------------|
| – Q-learning | – CEM |
| – SARSA | – Expected Value SARSA |

When training with on-policy methods,

- use no (or small) experience replay
- compensate with parallel game sessions



Left or right?

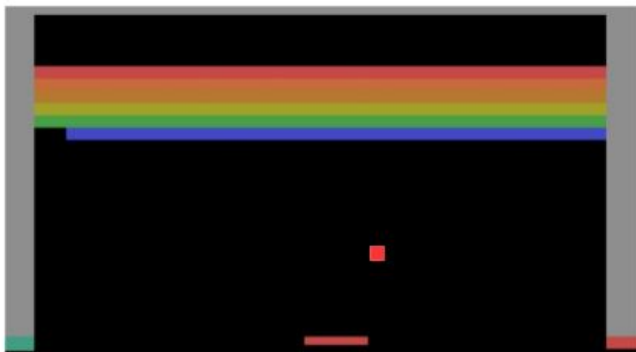
N-gram trick

Idea:

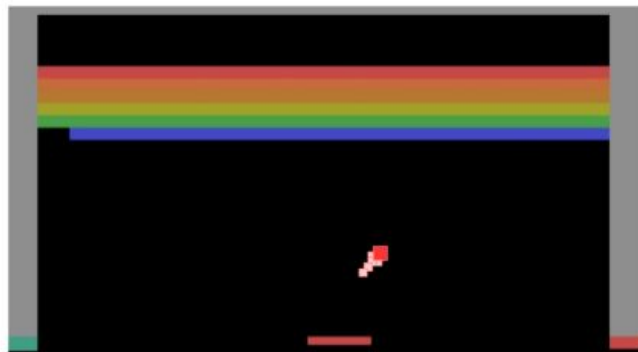
$$s_t \neq o(s_t)$$

$$s_t \approx (o(s_{t-n}), a_{t-n}, \dots, o(s_{t-1}), a_{t-1}, o(s_t))$$

e.g. ball movement in breakout



· One frame

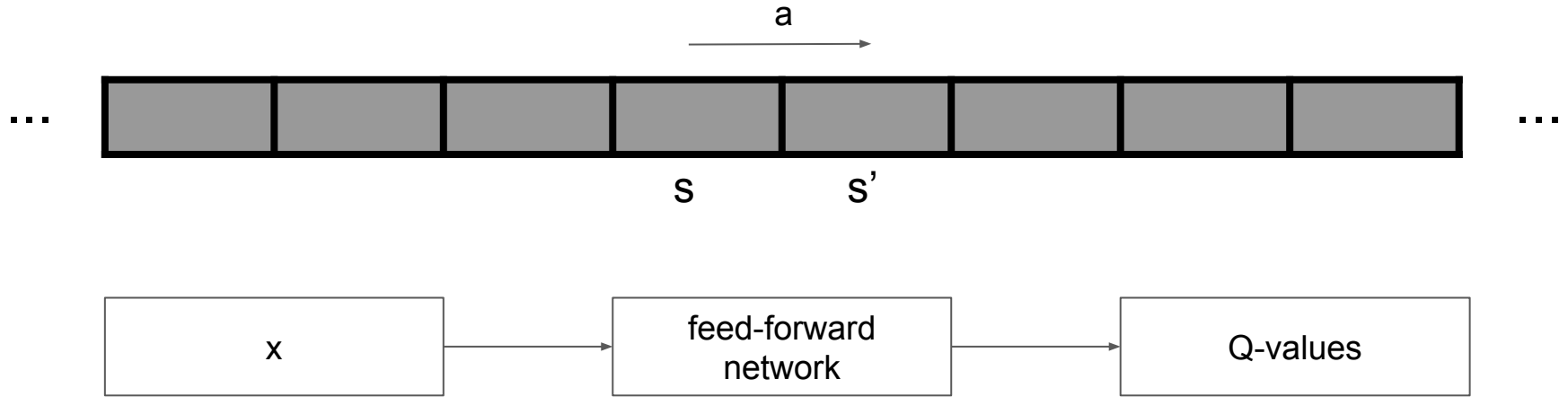


· Several frames

N-gram trick

- Nth-order markov assumption
- Works for velocity/timers
- Fails for anything longer than N frames
- Impractical for large N

Autocorrelation



Target is based on prediction

$Q(s, a)$ correlates with $Q(s', a)$

Target network

Idea: use network with frozen weights to compute the target

$$L(\Theta) = E_{s \sim S, a \sim A} [(Q(s, a, \Theta) - (r + \gamma \max_{a'} Q(s', a', \Theta^-)))^2]$$

where Θ^- is the frozen weights

↑
Const

Hard target network:

Update Θ^- every **n** steps and set its weights as Θ

Target network

Idea: use network with frozen weights to compute the target

$$L(\Theta) = E_{s \sim S, a \sim A} [(Q(s, a, \Theta) - (r + \gamma \max_{a'} Q(s', a', \Theta^-)))^2]$$

where Θ^- is the frozen weights

↑
Const

Hard target network:

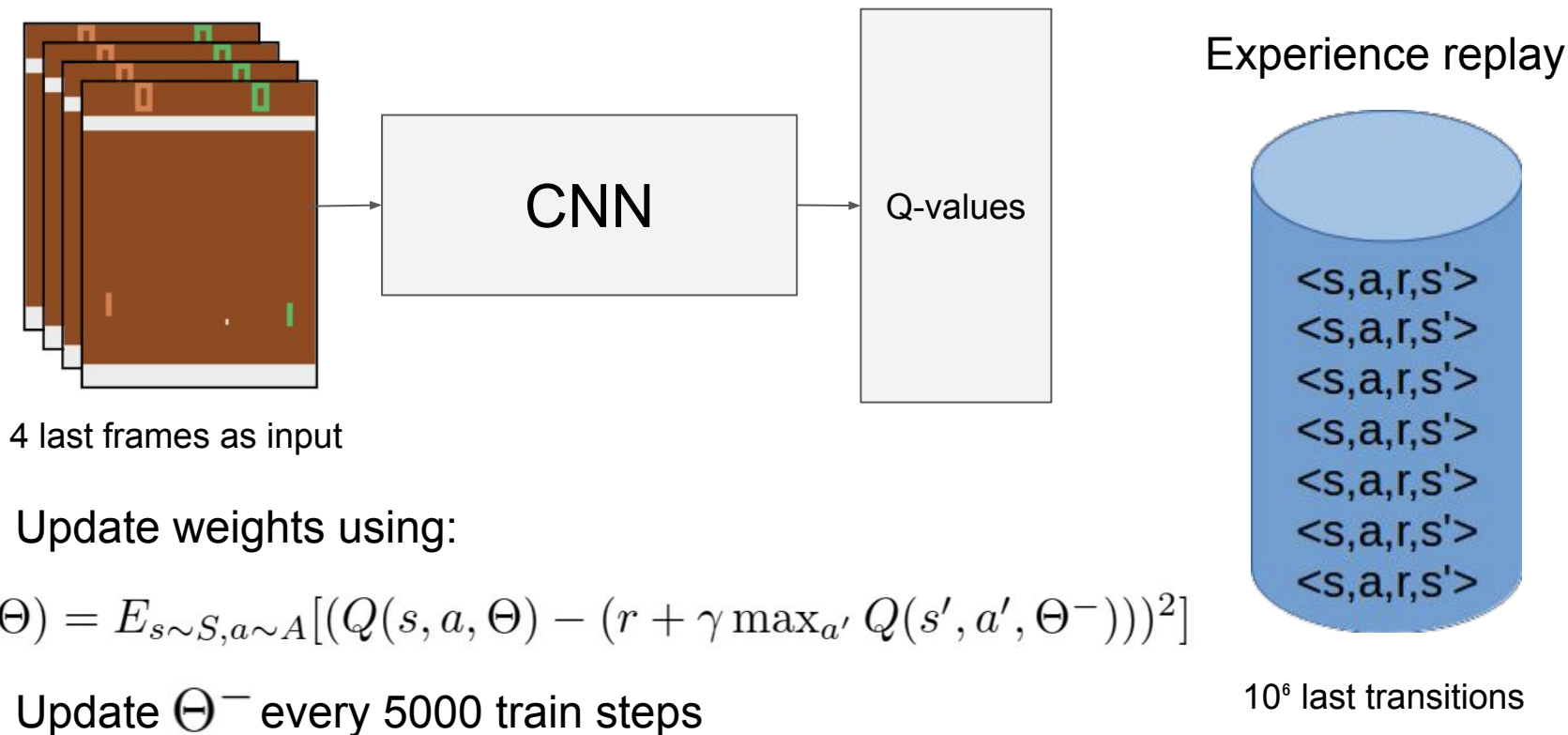
Update Θ^- every **n** steps and set its weights as Θ

Soft target network:

Update Θ^- every step:

$$\Theta^- = (1 - \alpha)\Theta^- + \alpha\Theta$$

Playing Atari with Deep Reinforcement Learning (2013, Deepmind)



Problem of overestimation

We use “max” operator to compute the target

$$L(s, a) = (Q(s, a) - (r + \gamma \max_{a'} Q(s', a')))^2$$

We have a problem

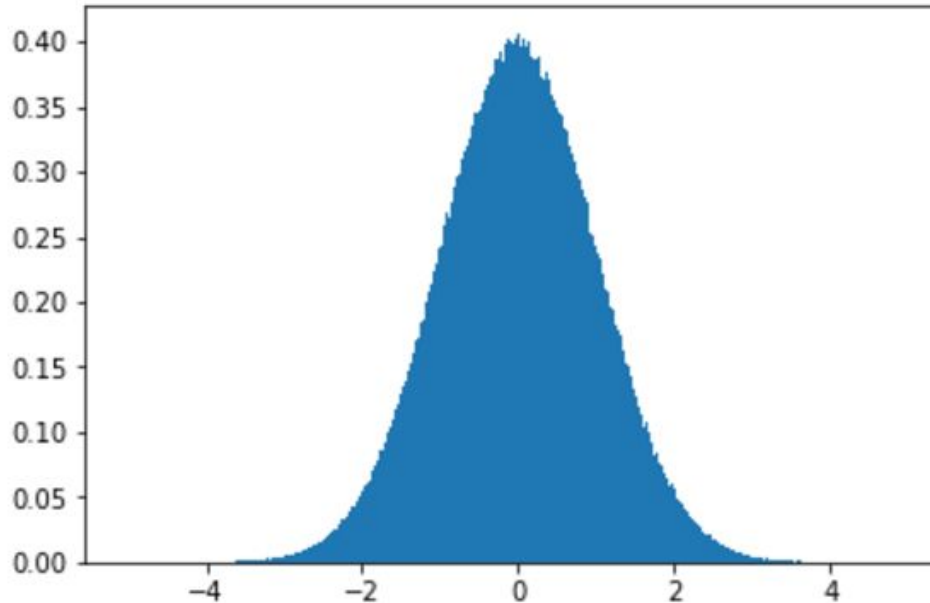
(although we want $E_{s \sim S, a \sim A}[L(s, a)]$ to be equal zero)

Problem of overestimation

Normal distribution

$3 \cdot 10^6$ samples

mean: ~ 0.0004



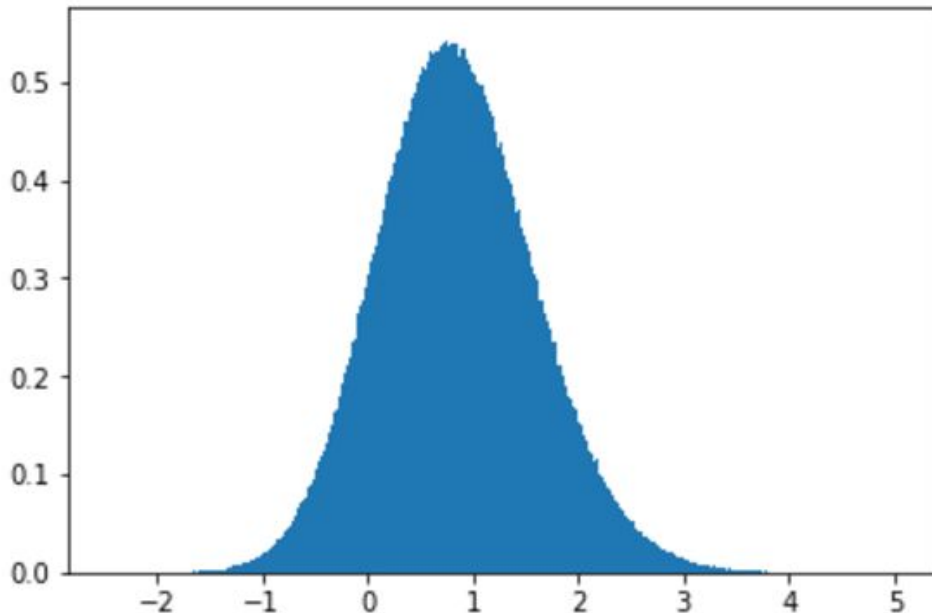
Problem of overestimation

Normal distribution

$3 \cdot 10^6 \times 3$ samples

Then take maximum of every tuple

mean: ~ 0.8467



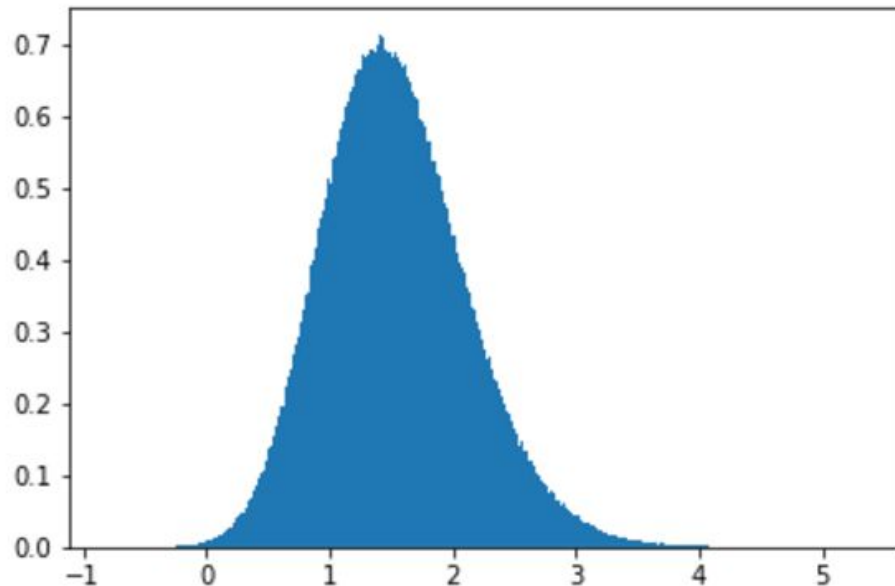
Problem of overestimation

Normal distribution

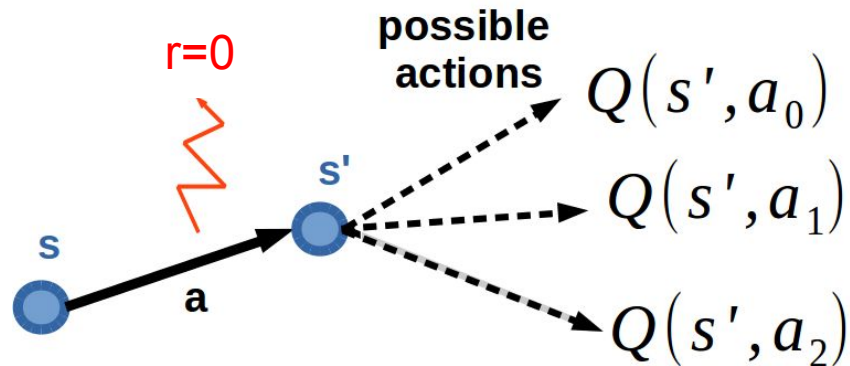
$3 \times 10^6 \times 10$ samples

Then take maximum of every tuple

mean: ~ 1.538



Problem of overestimation

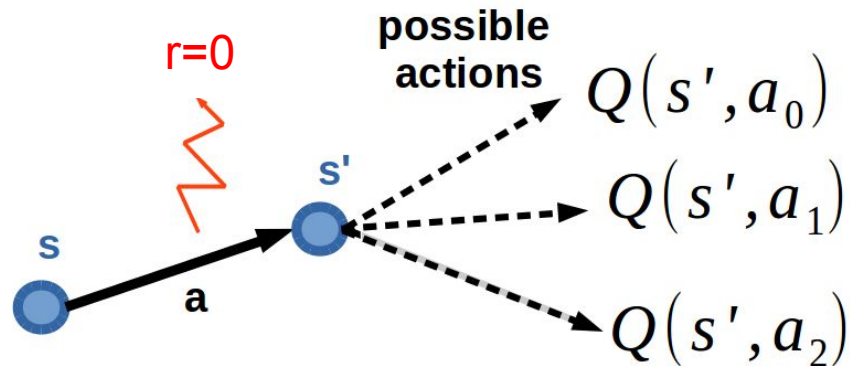


Suppose true $Q(s', a')$ are equal to $\mathbf{0}$ for all a'

But we have an approximation (or other) error $\sim N(0, \sigma^2)$

So $Q(s, a)$ should be equal to $\mathbf{0}$

Problem of overestimation



But if we update $Q(s, a)$ towards $r + \gamma \max_{a'} Q(s', a')$
we will have overestimated $Q(s, a) > \mathbf{0}$ because

$$E[\max_{a'} Q(s', a')] \geq \max_{a'} E[Q(s', a')]$$

Double Q-learning (NIPS 2010)

$$y = r + \gamma \max_{a'} Q(s', a') \quad - \text{Q-learning target}$$

$$y = r + \gamma Q(s', \operatorname{argmax}_{a'} Q(s', a')) \quad - \text{Rewritten Q-learning target}$$

Idea: use two estimators of q-values: Q^A, Q^B

They should compensate mistakes of each other because they will be independent
Let's get argmax from another estimator!

$$y = r + \gamma Q^A(s', \operatorname{argmax}_a Q^B(s', a')) \quad - \text{Double Q-learning target}$$

Double Q-learning (NIPS 2010)

Algorithm 1 Double Q-learning

```
1: Initialize  $Q^A, Q^B, s$ 
2: repeat
3:   Choose  $a$ , based on  $Q^A(s, \cdot)$  and  $Q^B(s, \cdot)$ , observe  $r, s'$ 
4:   Choose (e.g. random) either UPDATE(A) or UPDATE(B)
5:   if UPDATE(A) then
6:     Define  $a^* = \arg \max_a Q^A(s', a)$ 
7:      $Q^A(s, a) \leftarrow Q^A(s, a) + \alpha(s, a) (r + \gamma Q^B(s', a^*) - Q^A(s, a))$ 
8:   else if UPDATE(B) then
9:     Define  $b^* = \arg \max_a Q^B(s', a)$ 
10:     $Q^B(s, a) \leftarrow Q^B(s, a) + \alpha(s, a) (r + \gamma Q^A(s', b^*) - Q^B(s, a))$ 
11:   end if
12:    $s \leftarrow s'$ 
13: until end
```

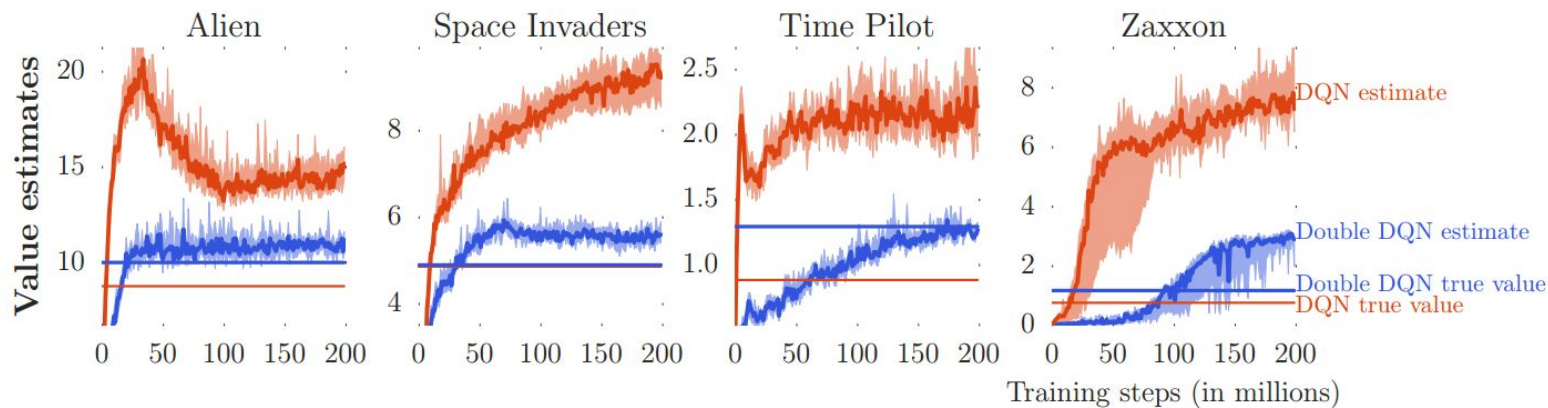
Can we combine this algorithm with DQN?

Deep Reinforcement Learning with Double Q-learning (Deepmind, 2015)

Idea: use main network to choose action!

$$y_{dqn} = r + \gamma \max_{a'} Q(s', a', \Theta^-)$$

$$y_{ddqn} = r + \gamma Q(s', \operatorname{argmax}_{a'} Q(s', a', \Theta), \Theta^-)$$



	DQN	Double DQN	Double DQN (tuned)
Median	47.5%	88.4%	116.7%
Mean	122.0%	273.1%	475.2%

Experience Replay

State	Action	Reward	Next state
s ₀	a ₀	0	s ₁
s ₁	a ₁	0	s ₂
...
s _(n-1)	a _(n-1)	0	s _n
s_n	a_n	100	s_(n+1)
s _(n+1)	a _(n+1)	0	s _(n+2)
...

Prioritized Experience Replay (2016, Deepmind)

Idea: sample transitions from xp-replay cleverly

We want to set probability for every transition. Let's use the absolute value of TD-error of transition as a probability!

$$\text{TD-error } \delta = Q(s, a) - (r + \gamma Q(s', \arg\max_{a'} Q(s', a', \Theta), \Theta^-))$$

$$p = |\delta|$$

$$P(i) = \frac{p_i^\alpha}{\sum_k p_k^\alpha} \text{ where } \alpha \text{ is the priority parameter (when } \alpha \text{ is 0 it's the uniform case)}$$

Do you see the problem?

Transitions become non i.i.d. and therefore we introduce the bias.

Prioritized Experience Replay (2016, Deepmind)

Solution: we can correct the bias by using importance-sampling weights

$$w_i = \left(\frac{1}{N} \cdot \frac{1}{P(i)} \right)^\beta \quad \text{where } \beta \text{ is the parameter}$$

So we sample using $P(i) = \frac{p_i^\alpha}{\sum_k p_k^\alpha}$ and multiply error by w_i

Prioritized Experience Replay (2016, Deepmind)

Additional details

We also normalize weights by $1 / \max_i w_i$ (here is no mathematical reason)

When we put transition into experience replay, we set maximal priority $p_t = \max_{i < t} p_i$

It is the bonus homework!

Let's watch a video...

<https://www.youtube.com/watch?v=UXurvvdY93o>

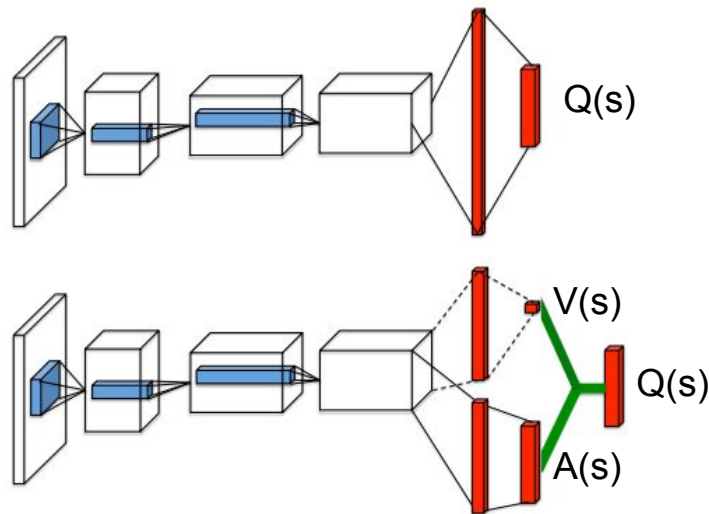
Dueling Network Architectures for Deep Reinforcement Learning (2016, Deepmind)

Idea: change the network's architecture.

Recall:

Advantage Function $A(s,a) = Q(s,a) - V(s)$

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

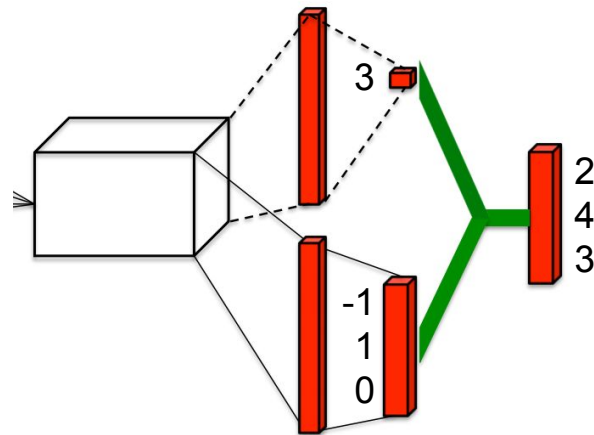
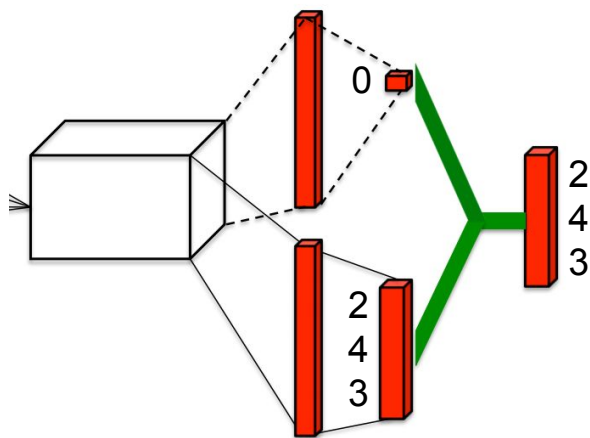


Do you see the problem?

Dueling Network Architectures for Deep Reinforcement Learning (2016, Deepmind)

Here is one extra freedom degree!

Example:



Which one is good?

Dueling Network Architectures for Deep Reinforcement Learning (2016, Deepmind)

Solution: require $\max_{a' \in |\mathcal{A}|} A(s, a'; \theta, \alpha)$ to be equal to zero!

So the **Q-function** computes as:

$$Q(s, a; \theta, \alpha, \beta) = V(s; \theta, \beta) + \left(A(s, a; \theta, \alpha) - \max_{a' \in |\mathcal{A}|} A(s, a'; \theta, \alpha) \right)$$

Dueling Network Architectures for Deep Reinforcement Learning (2016, Deepmind)

Solution: require $\max_{a' \in |\mathcal{A}|} A(s, a'; \theta, \alpha)$ to be equal to zero!

So the **Q-function** computes as:

$$Q(s, a; \theta, \alpha, \beta) = V(s; \theta, \beta) +$$

$$\left(A(s, a; \theta, \alpha) - \max_{a' \in |\mathcal{A}|} A(s, a'; \theta, \alpha) \right)$$

It's the homework!

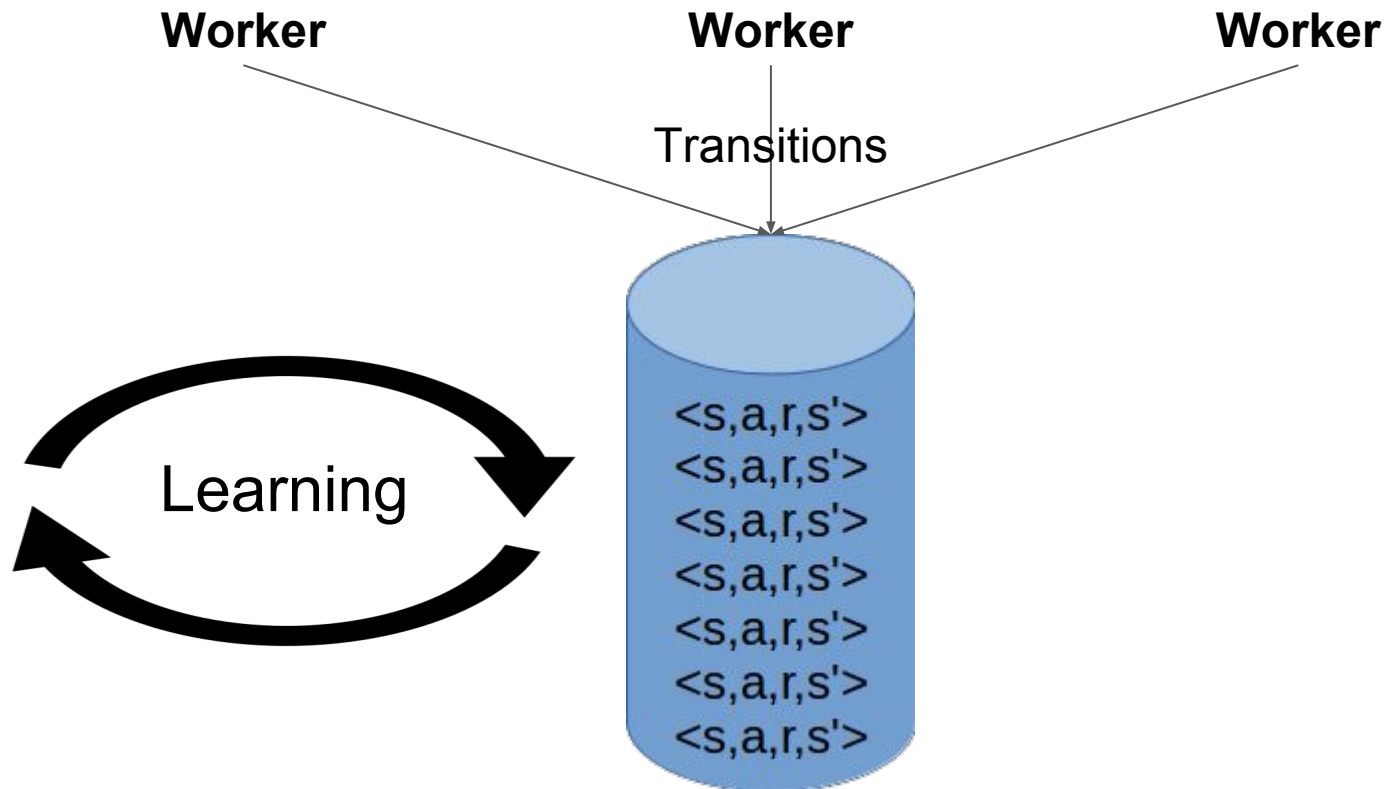
Authors of this papers also introduced this way to compute Q-values:

$$Q(s, a; \theta, \alpha, \beta) = V(s; \theta, \beta) +$$

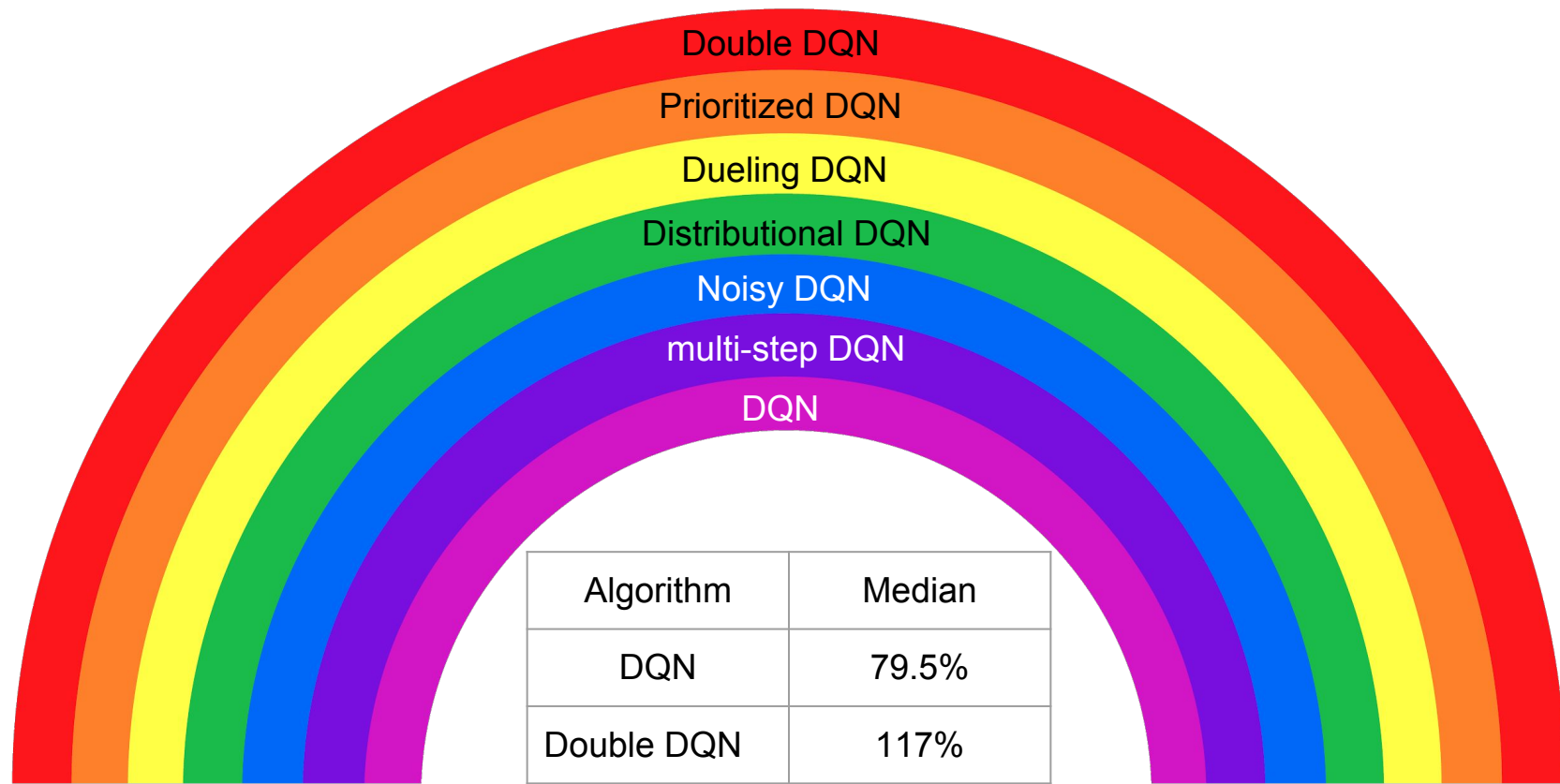
$$\left(A(s, a; \theta, \alpha) - \frac{1}{|\mathcal{A}|} \sum_{a'} A(s, a'; \theta, \alpha) \right)$$

They wrote that this variant increases stability of the optimization
(The fact that this loses the original semantics of Q doesn't matter)

Asynchronous Methods for Deep Reinforcement Learning (2016, Deepmind)



Rainbow (2017, Deepmind)



R2D2 (2018, Deepmind)

LSTM

Reward re-scaling

Distributed Prioritized
Experience Replay

Double DQN

n-step DQN

Dueling DQN

Median performance: **1920%** of human performance!

Thanks for your attention!