UVA | SCHOOL of DATA SCIENCE

Extensions to Q-Learning 2: Dueling Deep Q-Network

Reinforcement Learning

School of Data Science

University of Virginia

Last updated: June 23, 2025

Dueling Networks

Deep RL before this paper used conventional architectures (CNN, LSTM)

Focus here is new architecture better suited to model-free RL

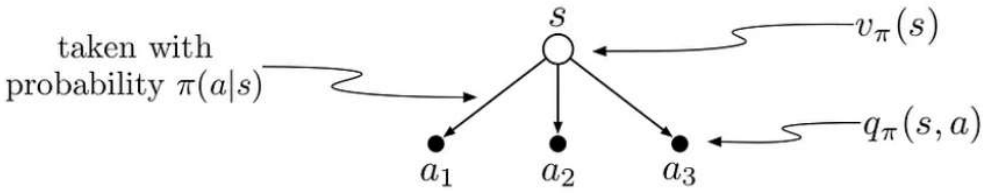
Dueling architecture separates state values and action advantages

Paper: *Dueling Network Architectures for Deep Reinforcement Learning.* Wang et. al.

Value and Advantage Function Definitions

Relationship of value functions:

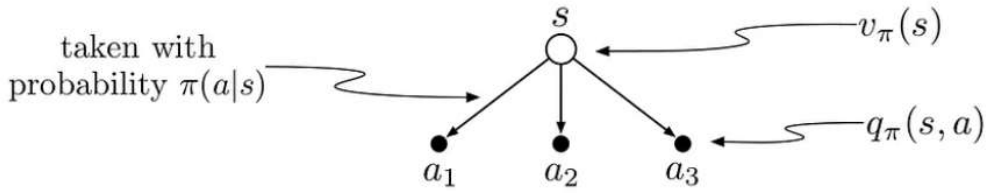
$$
\begin{aligned}
Q^{\pi}(s,a) &= \mathbb{E}\left[R_t \middle| s_t = s, a_t = a, \pi\right] \\
V^{\pi}(s) &= \mathbb{E}_{a\sim\pi(s)}\left[Q^{\pi}(s,a)\right].
\end{aligned}
$$



Value and Advantage Function Definitions

Relationship of value functions:

$$Q^{\pi}(s,a) = \mathbb{E}\left[R_t \middle| s_t = s, a_t = a, \pi\right]$$
$$V^{\pi}(s) = \mathbb{E}_{a\sim\pi(s)}\left[Q^{\pi}(s,a)\right].$$



The advantage function isolates effect of action taken

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s)$$

V(*s*) measures how good it is to be in state *s*

A(s,a) measures relative importance of each action

Sometimes, the action doesn't matter

Dueling Architecture

Separate streams for value and advantage functions

Common convolution feature

Top figure is Q-network

Bottom figure is dueling Q-network

Value and advantage streams are combined

For each network, outputs are Q(s,a)

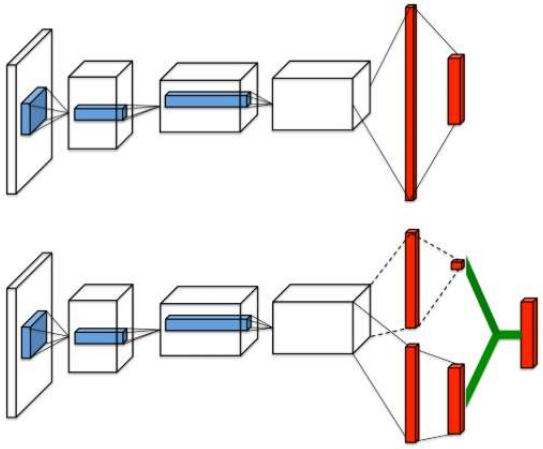
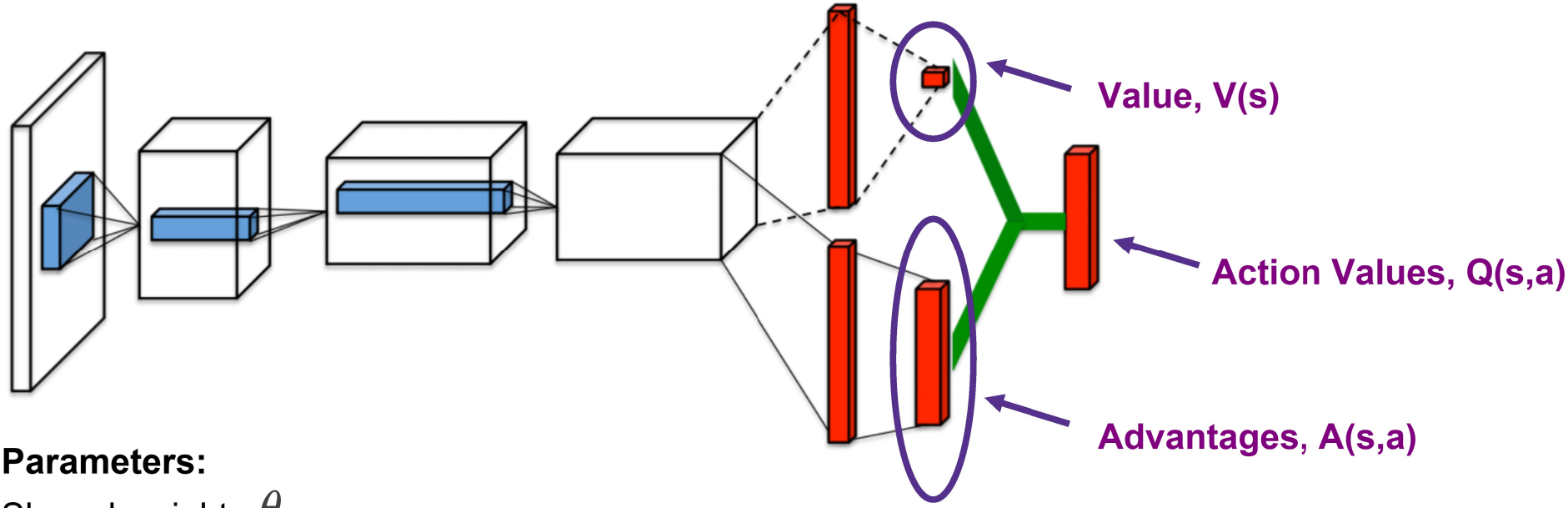


Figure 1. A popular single stream Q-network (**top**) and the dueling Q-network (**bottom**). The dueling network has two streams to separately estimate (scalar) state-value and the advantages for each action; the green output module implements equation (9) to combine them. Both networks output Q-values for each action.

Dueling Architecture, contd.



Parameters:

Shared weights θ

Value stream weights β

Advantage stream weights α

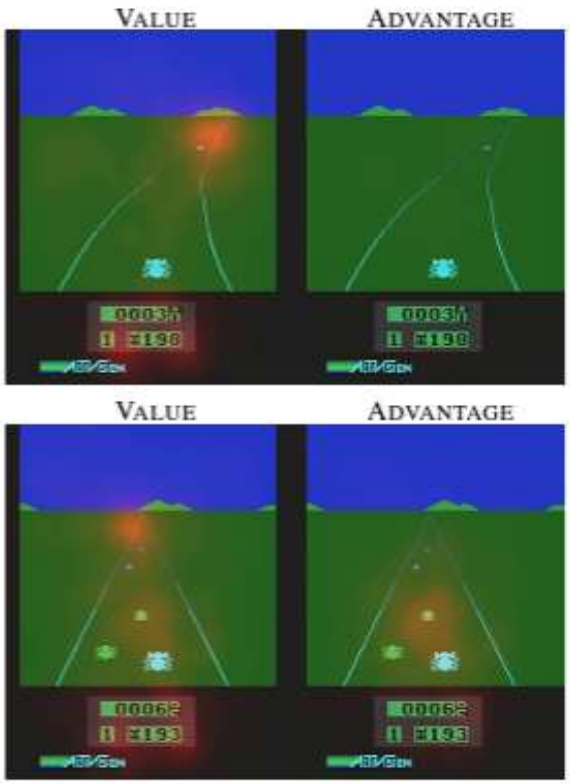
Value and Advantage Function Saliency

Figure shows saliency map for two time steps

Value stream pays attention to horizon and score

Advantage stream: when no cars are on road, action doesn't matter (no attention paid)

When cars are on road, advantage stream pays attention to car in front (bottom right)



Identifiability

Given Q = V + A, we cannot recover V and A uniquely
(adding any *c* to V, and subtracting *c* from A, leaves Q unchanged)

Add a constraint: subtract the average advantage from A:

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

For actions with above average advantage, the second term in (*) will be positive
For action with greatest advantage, (*) term will be largest across all actions

Review of Dueling Architecture Code

We can see an implementation of Dueling Q-Network here:

Paper: *Reinforcement Learning for optimal sepsis treatment policies* (2017)

Authors: Raghu, Komorowski, Ahmed, Celi, Szolovits, Ghassemi

GitHub repo:
https://github.com/aniruddhraghu/sepsisrl/blob/master/continuous/q_network.ipynb

See section:

advantage and value streams

Implementation

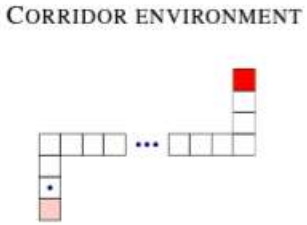
Our two streams (V, A) for Q are part of the model architecture

Training step runs the same as a standard Q-network: backpropagation

Evaluation

Sets up simple environment called *corridor*.

Redness of state signifies reward



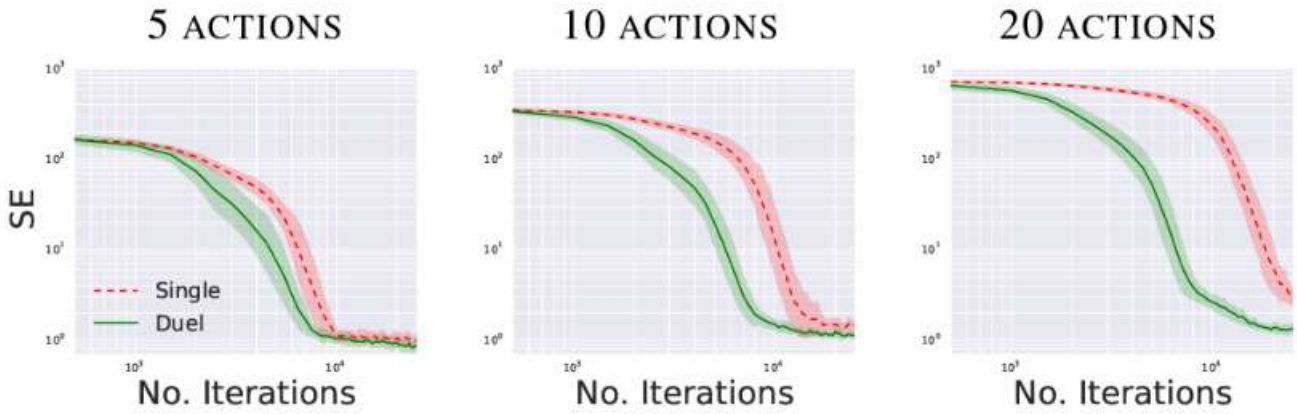
Can move up, down, left, right, or no move.

ε-greedy policy, measure performance by squared error (SE) of Q-value estimates vs. true action values

Single and Duel networks use MLP with three layers

For duel network, after first hidden layer, network branches off to two streams

Evaluation, contd.



Dueling network converges faster than Single

More pronounced as |A| increases

As |A| increases, SE is lower for Dueling network. It performs better.

Summary

We studied a novel architecture for Q function

Decomposes Q into value V and action advantage A streams

$$Q^{\pi}(s,a) = V^{\pi}(s) + A^{\pi}(s,a)$$

(we subtract average advantage for identifiability)

Summary

We studied a novel architecture for Q function

Decomposes Q into value V and action advantage A streams

$$Q^{\pi}(s,a) = V^{\pi}(s) + A^{\pi}(s,a)$$

(we subtract average advantage for identifiability)

Advantage function measures incremental importance over V(s):

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s)$$

Dueling network converges faster than usual DQN, Double Q-Network

Errors shown to be lower for Atari games