INTERNATIONAL CENTER
FOR COMPUTATIONAL LOGIC

Technische Universität Dresden faculty of computer science, ICCL

USING REINFORCEMENT LEARNING TO PLAY ANGRY BIRDS

colloquium

Peter Hirsch

Dresden, 2017/9/26

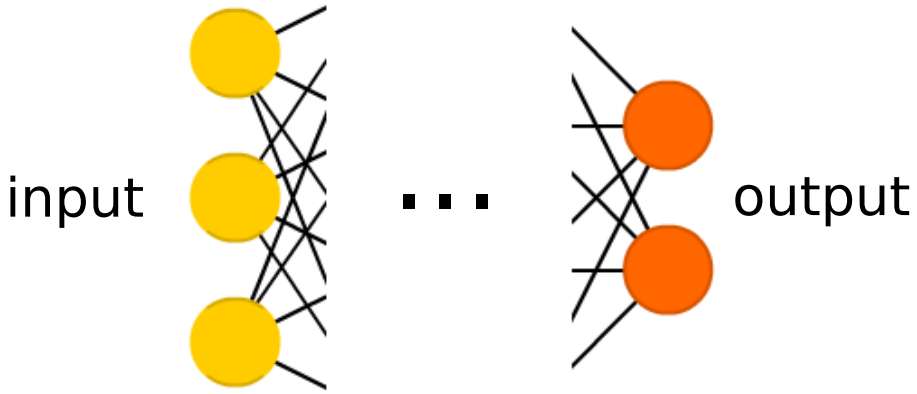
DRESDEN
concept
Exzellenz aus
Wissenschaft
und Kultur

Table of Contents

INTERNATIONAL CENTER
FOR COMPUTATIONAL LOGIC

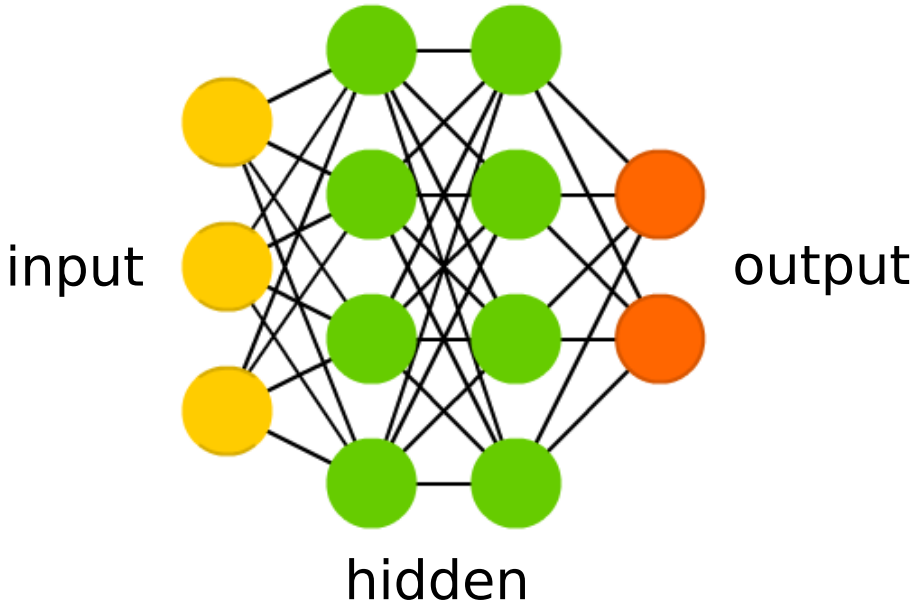
Angry Birds in a Nutshell

Neural Networks

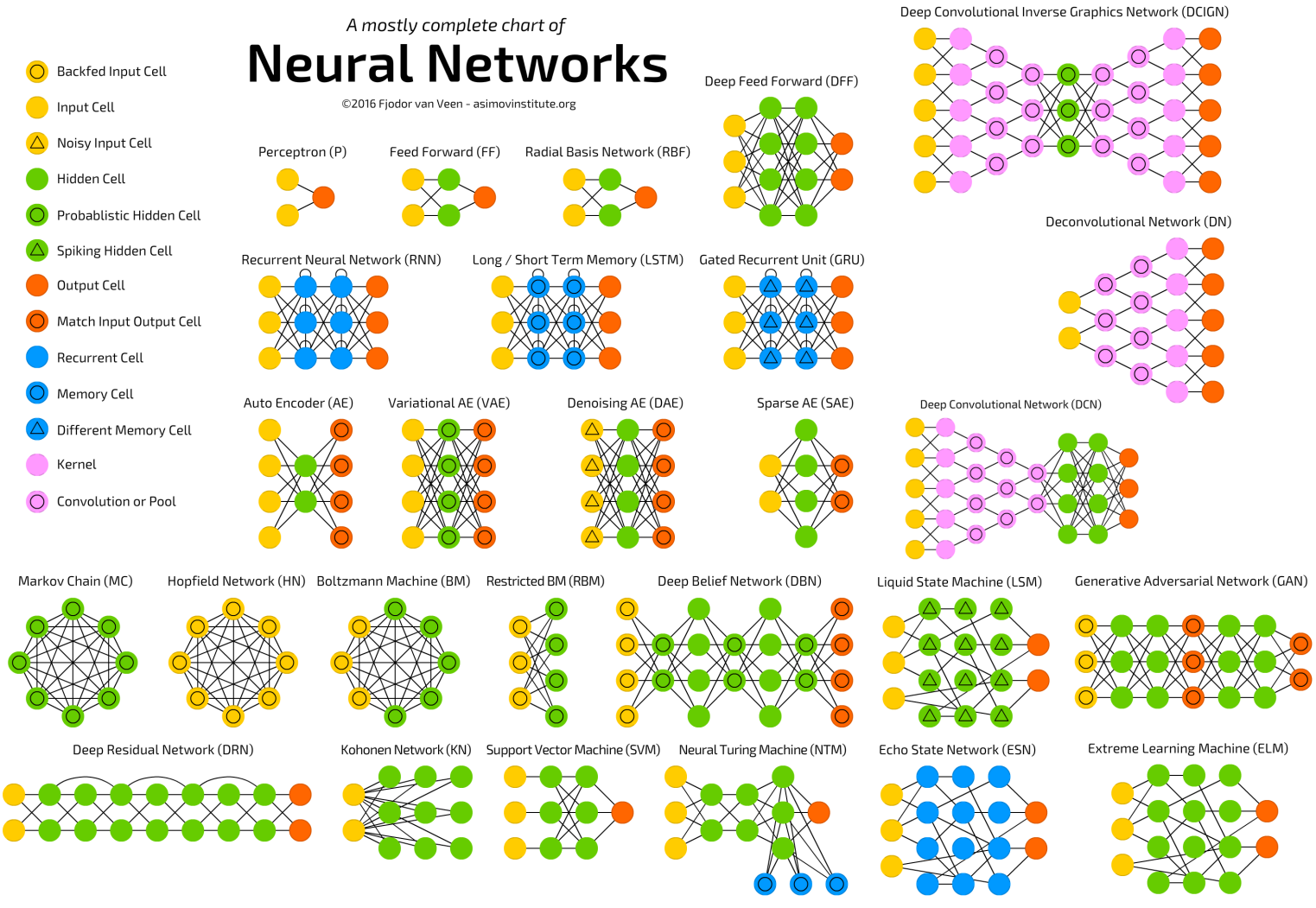


$$\text{output} = \sum \text{weights} \cdot \text{inputs}$$

Deep Neural Networks

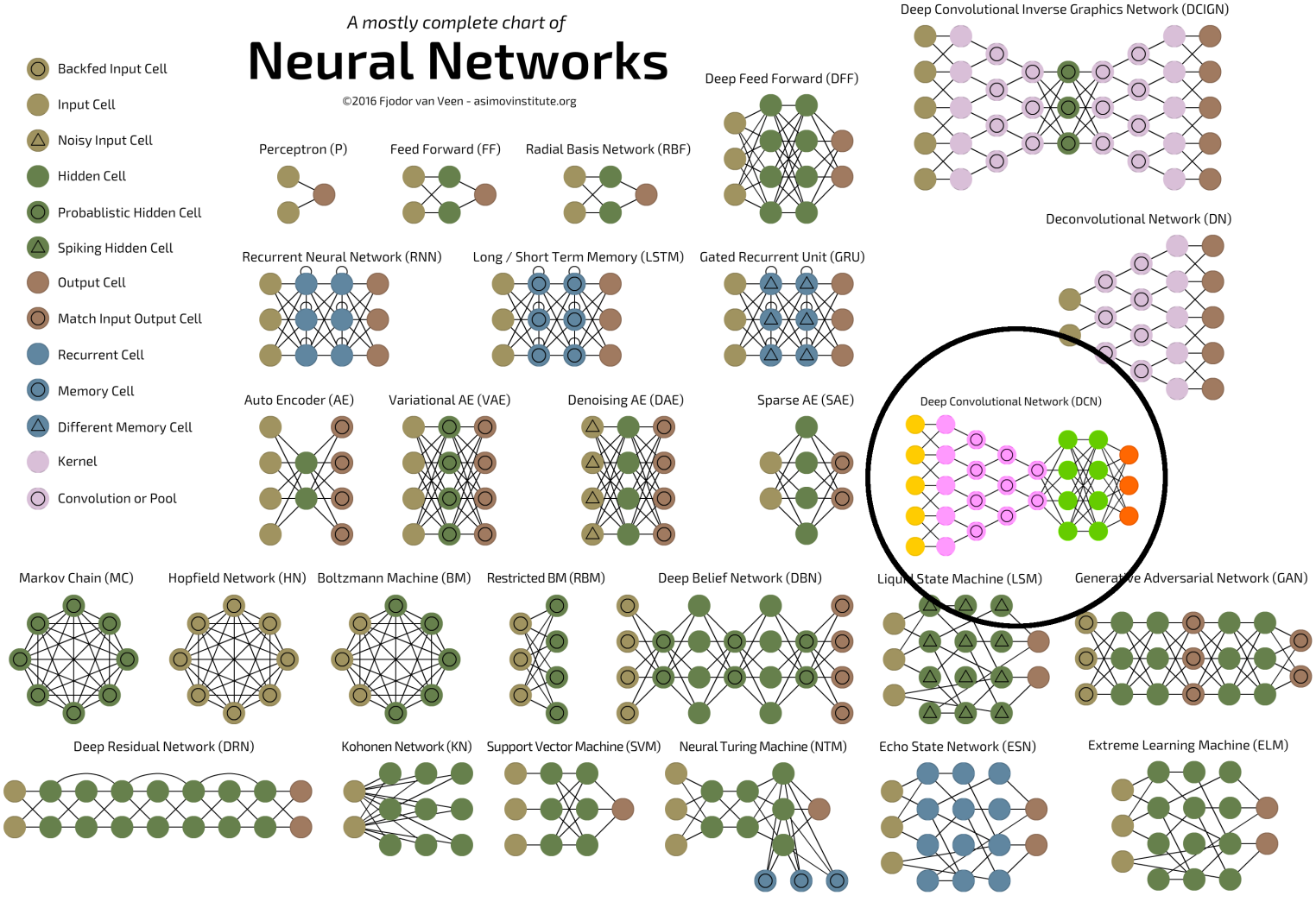


Deep Neural Networks

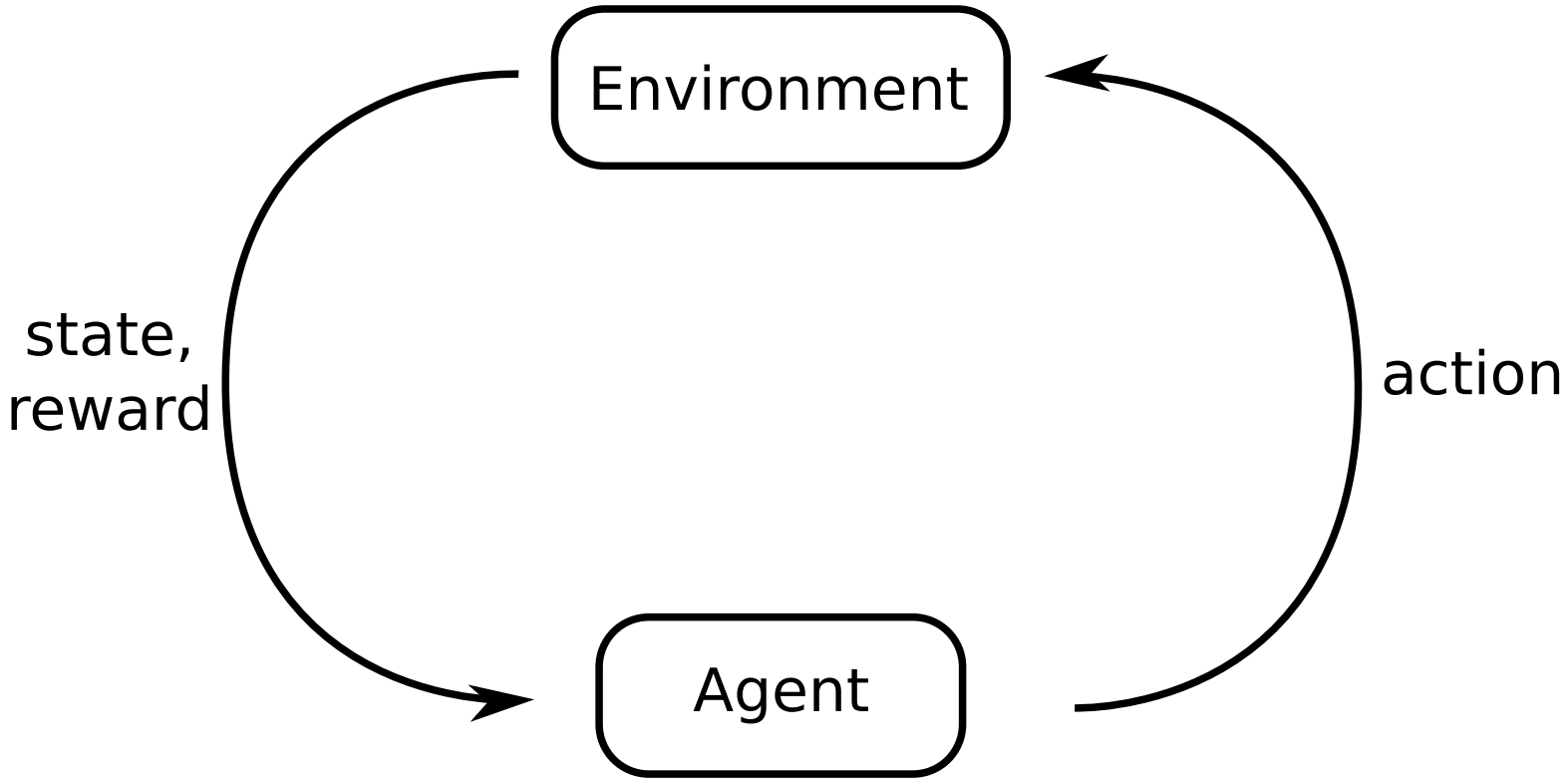


Deep Neural Networks

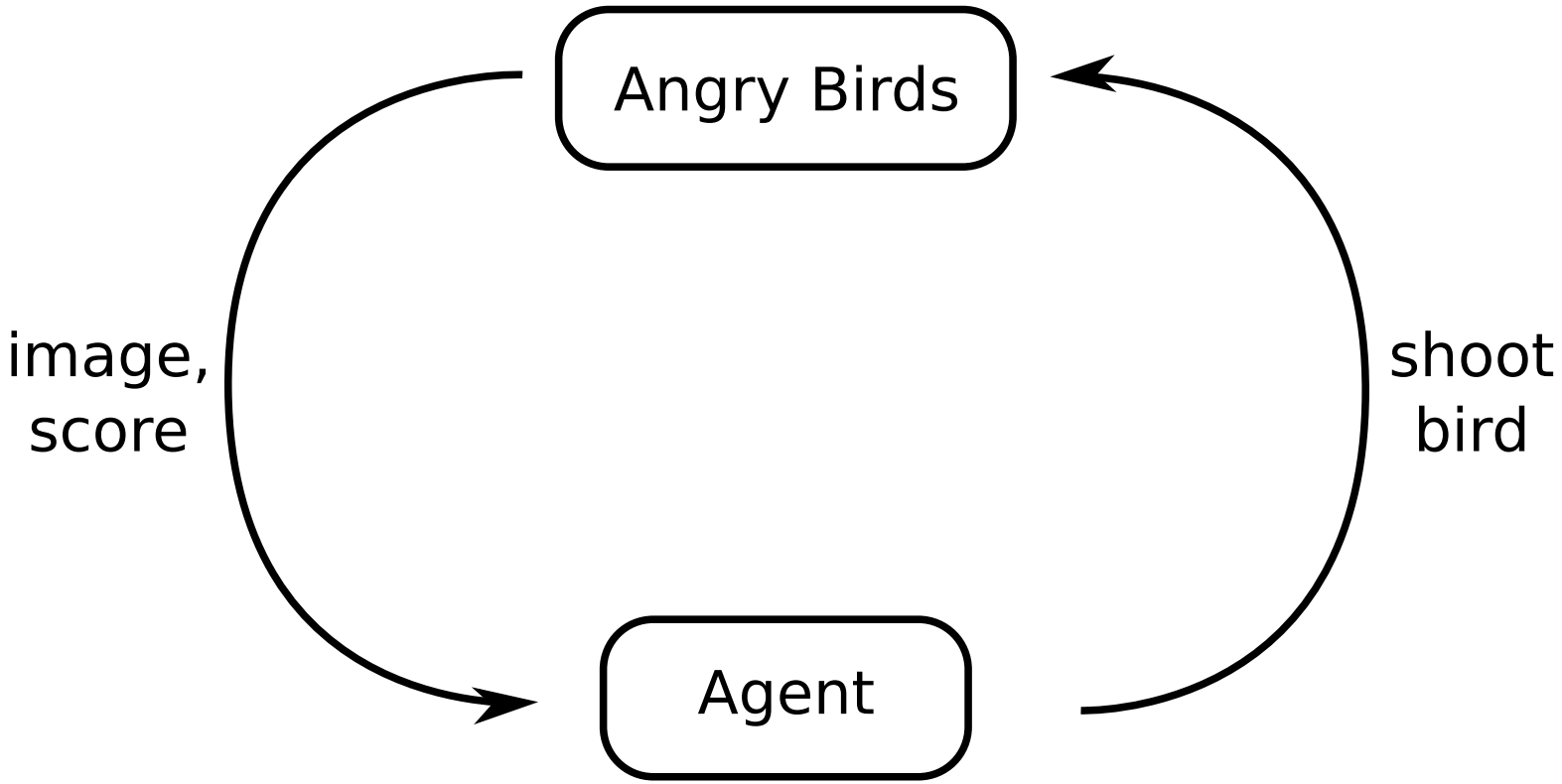
INTERNATIONAL CENTER FOR COMPUTATIONAL LOGIC



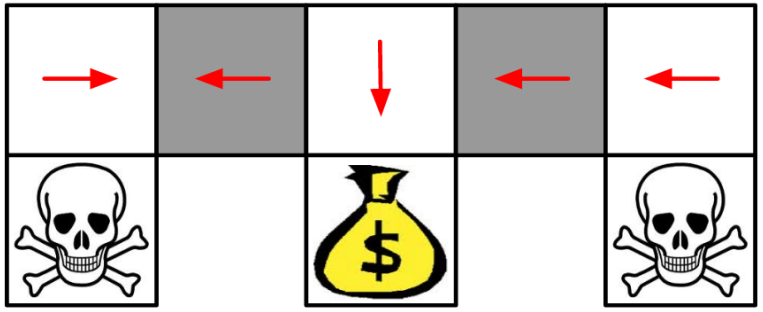
RL in a Nutshell



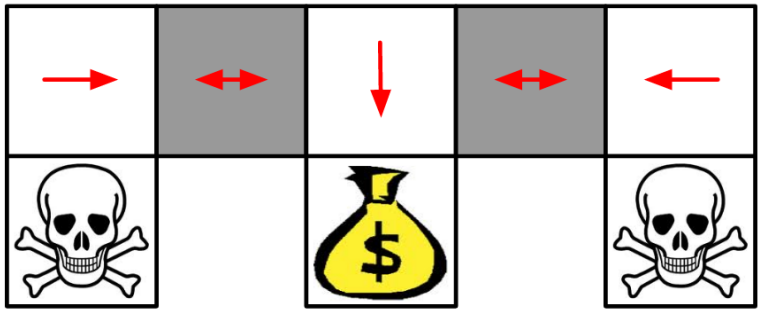
RL in a Nutshell



RL: Policy-based Methods



deterministic policy

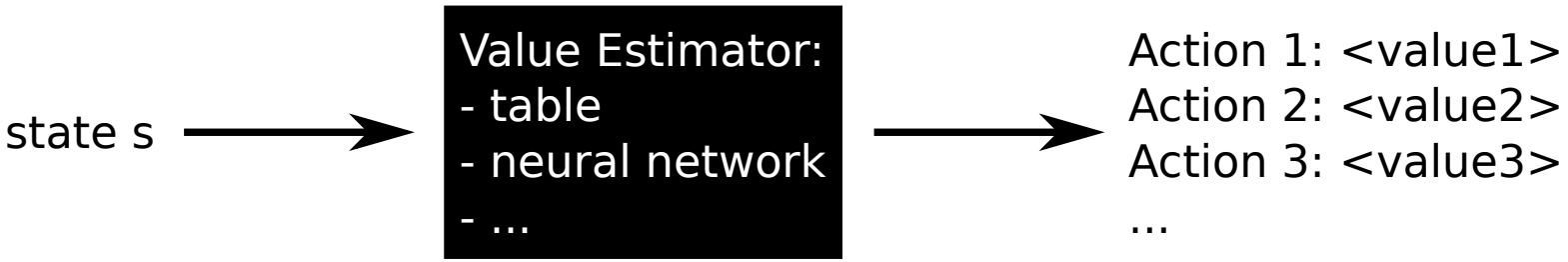


stochastic policy

(with state approximation, gray states not distinguishable)

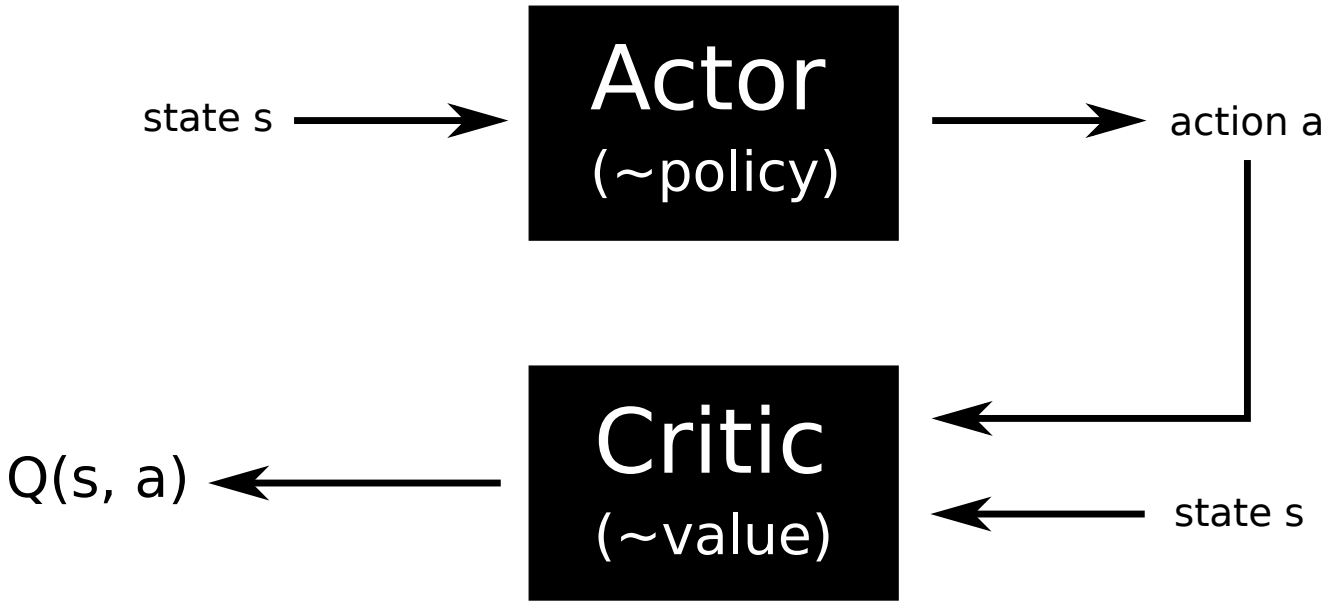
RL: Value-based Methods

INTERNATIONAL CENTER
FOR COMPUTATIONAL LOGIC



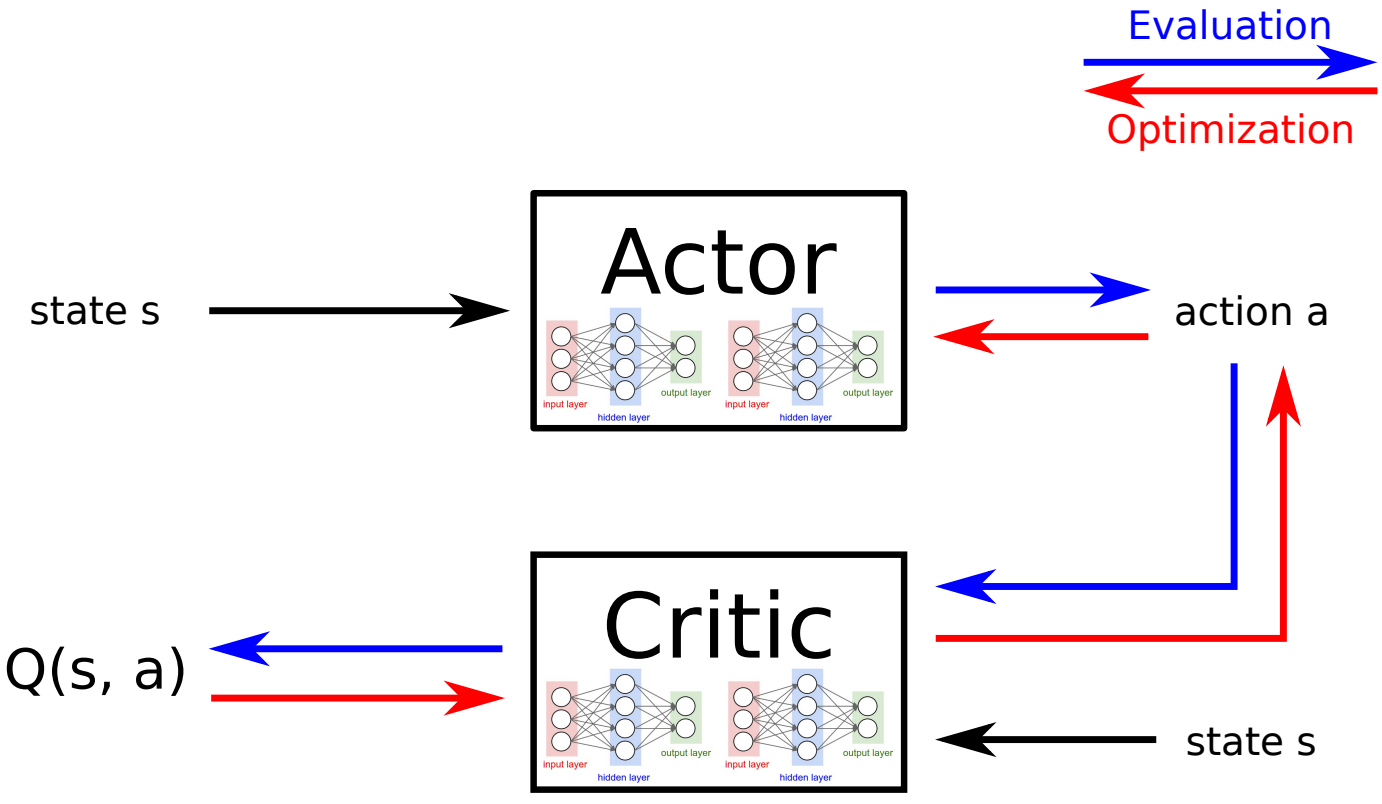
a **greedy** or ϵ-**greedy** policy is used to act
(a.k.a. go to neighboring state with highest (Q-)value)

RL: Actor-Critic Methods

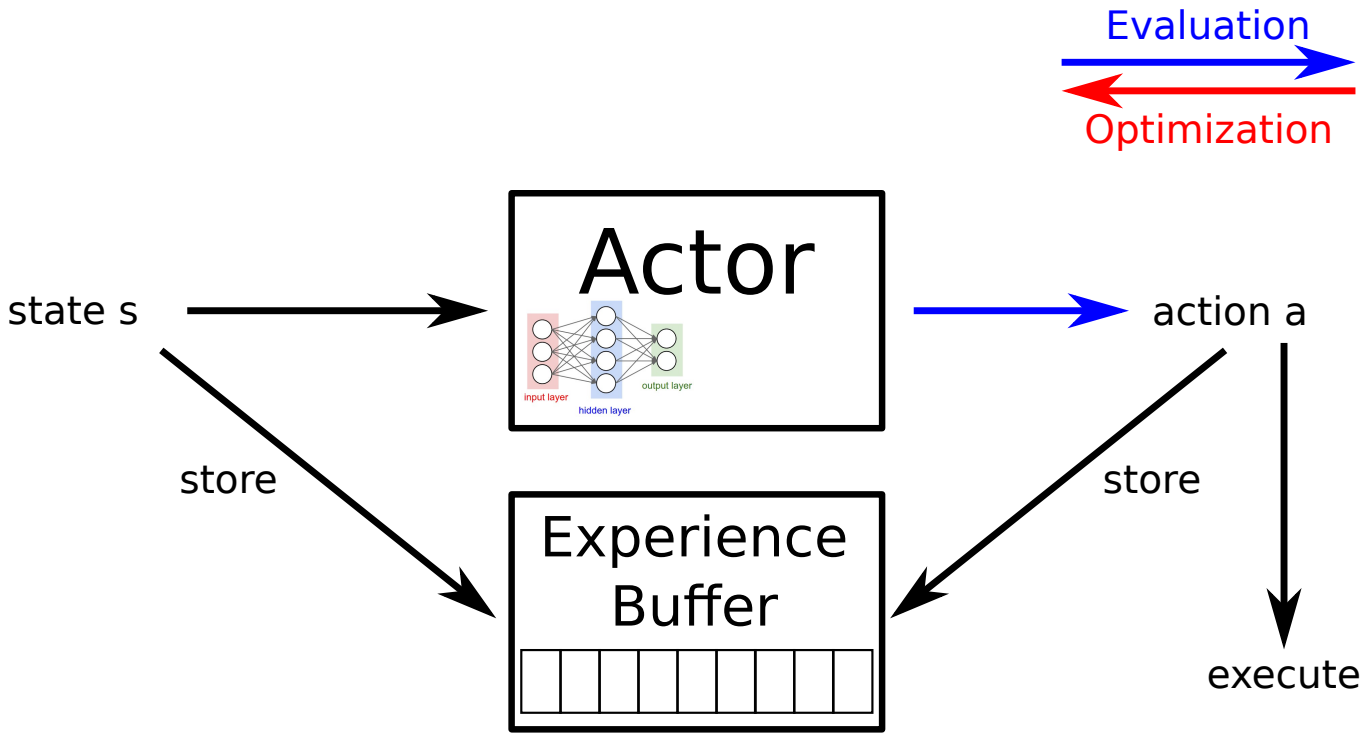


Combination of policy-based and value-based

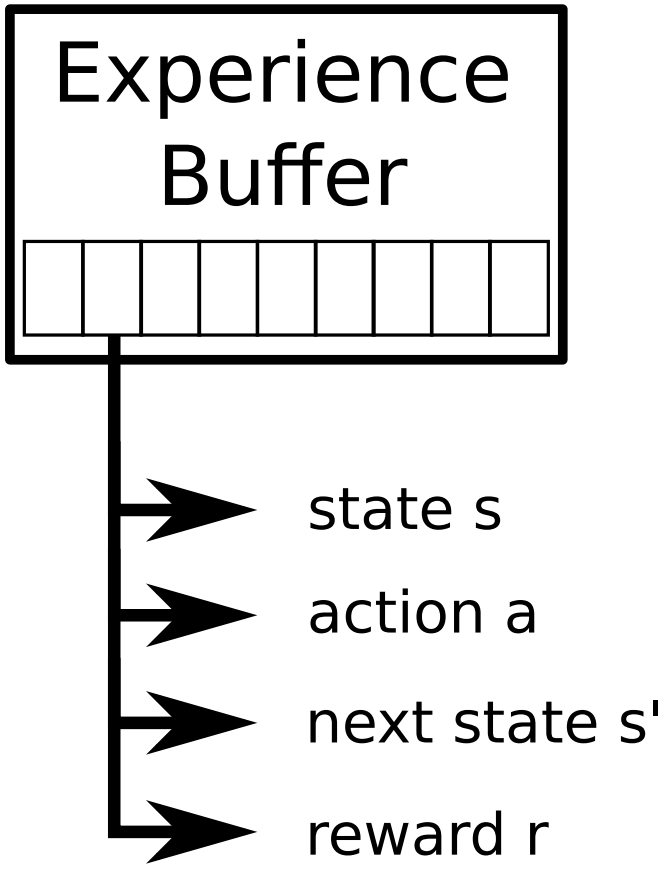
Deep Deterministic Policy Gradient



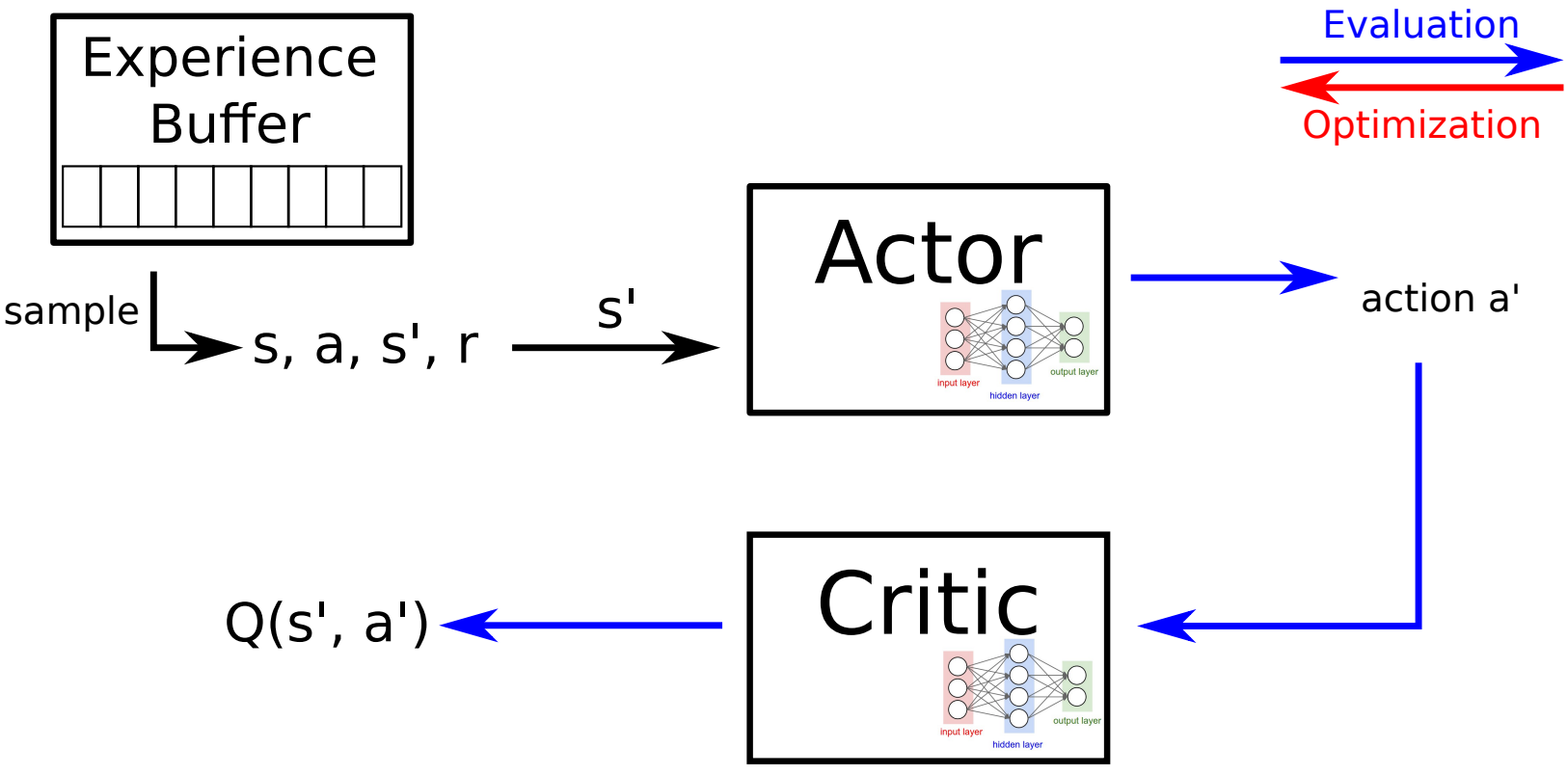
DDPG: Acting


Evaluation
Optimization
Actor
state s
action a
store
store
Experience
Buffer
execute
input layer
hidden layer
output layer

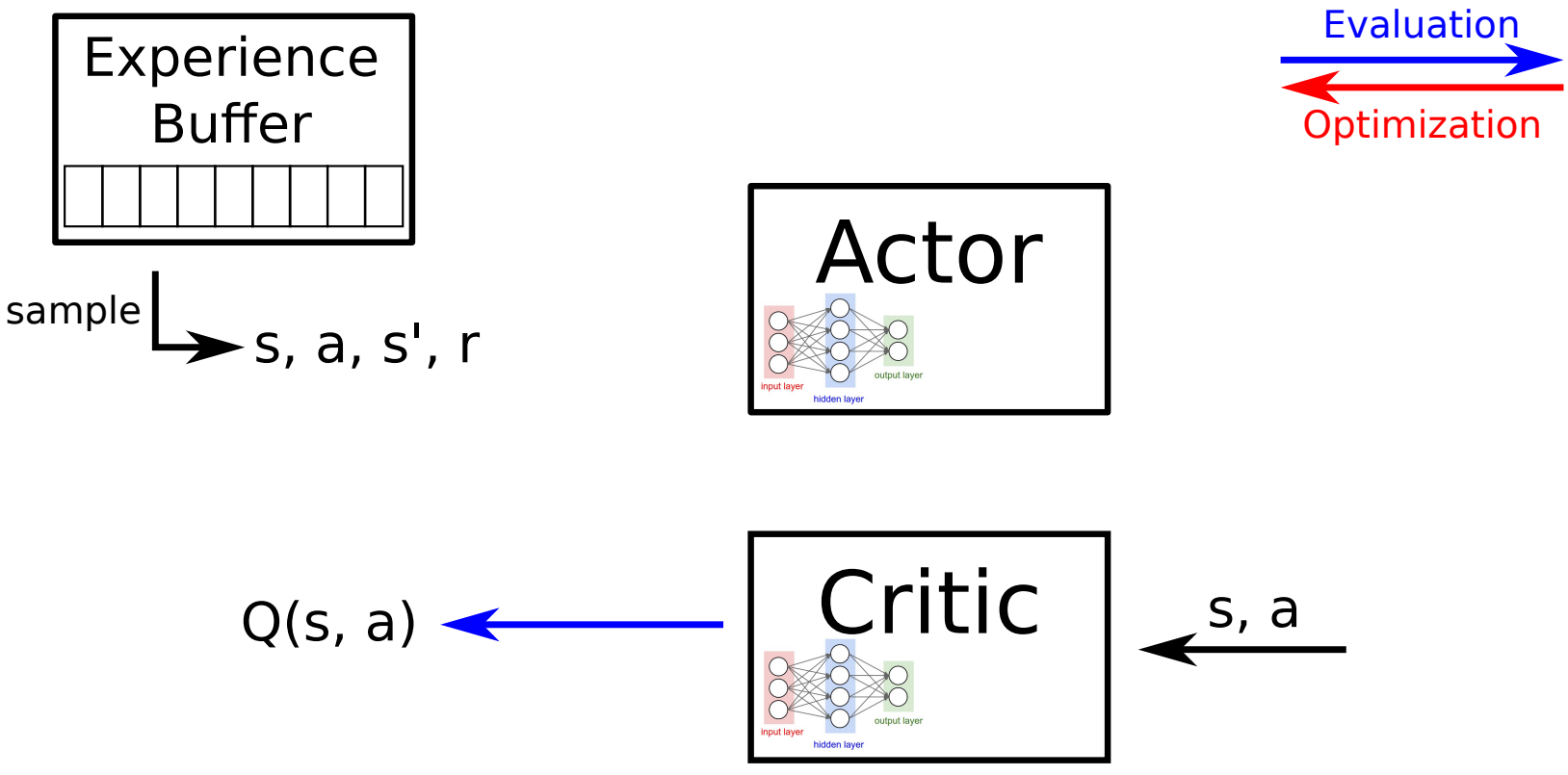
DDPG: Buffer



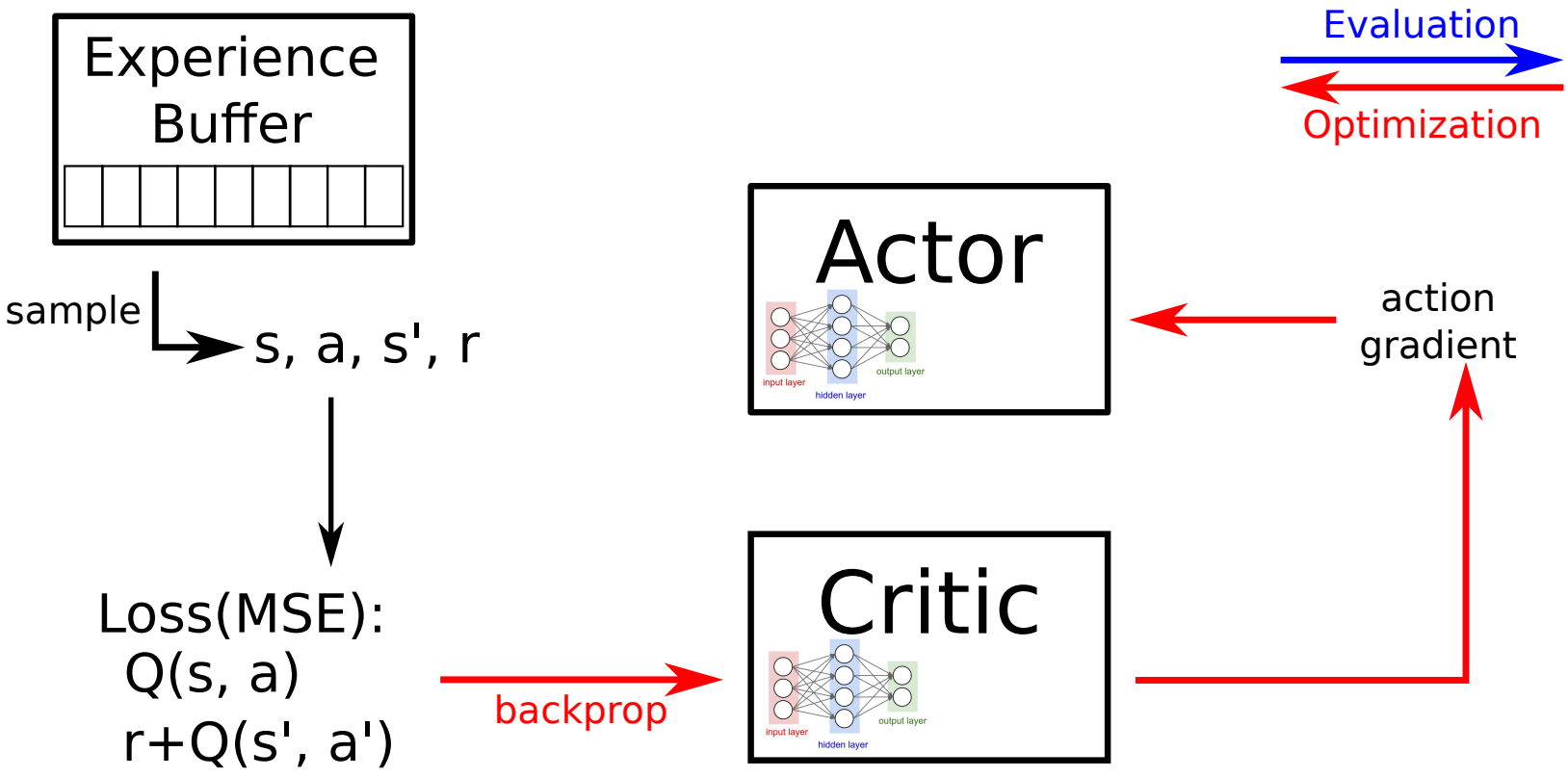
DDPG: Learning - Part 1



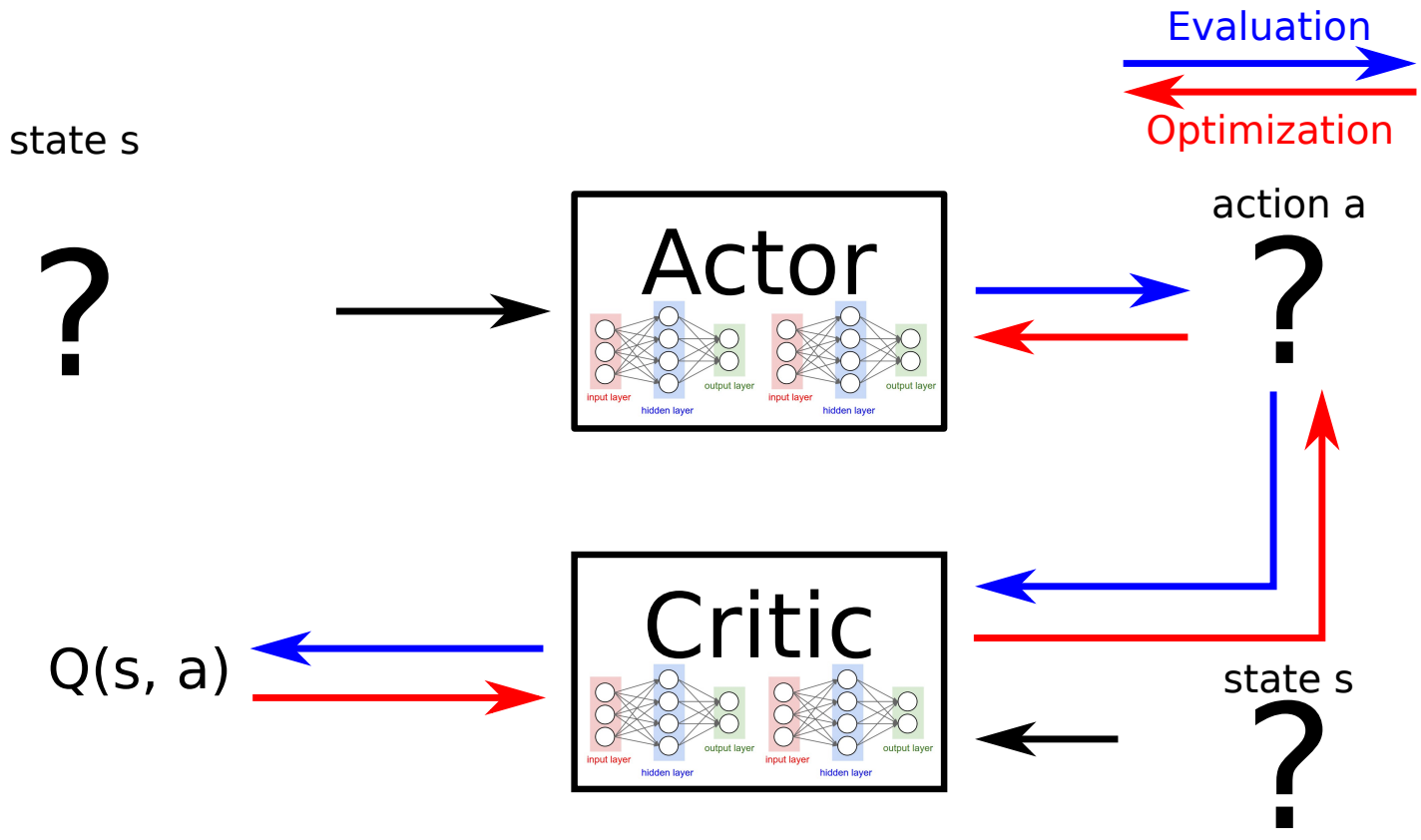
DDPG: Learning - Part 2



DDPG: Learning - Part 3

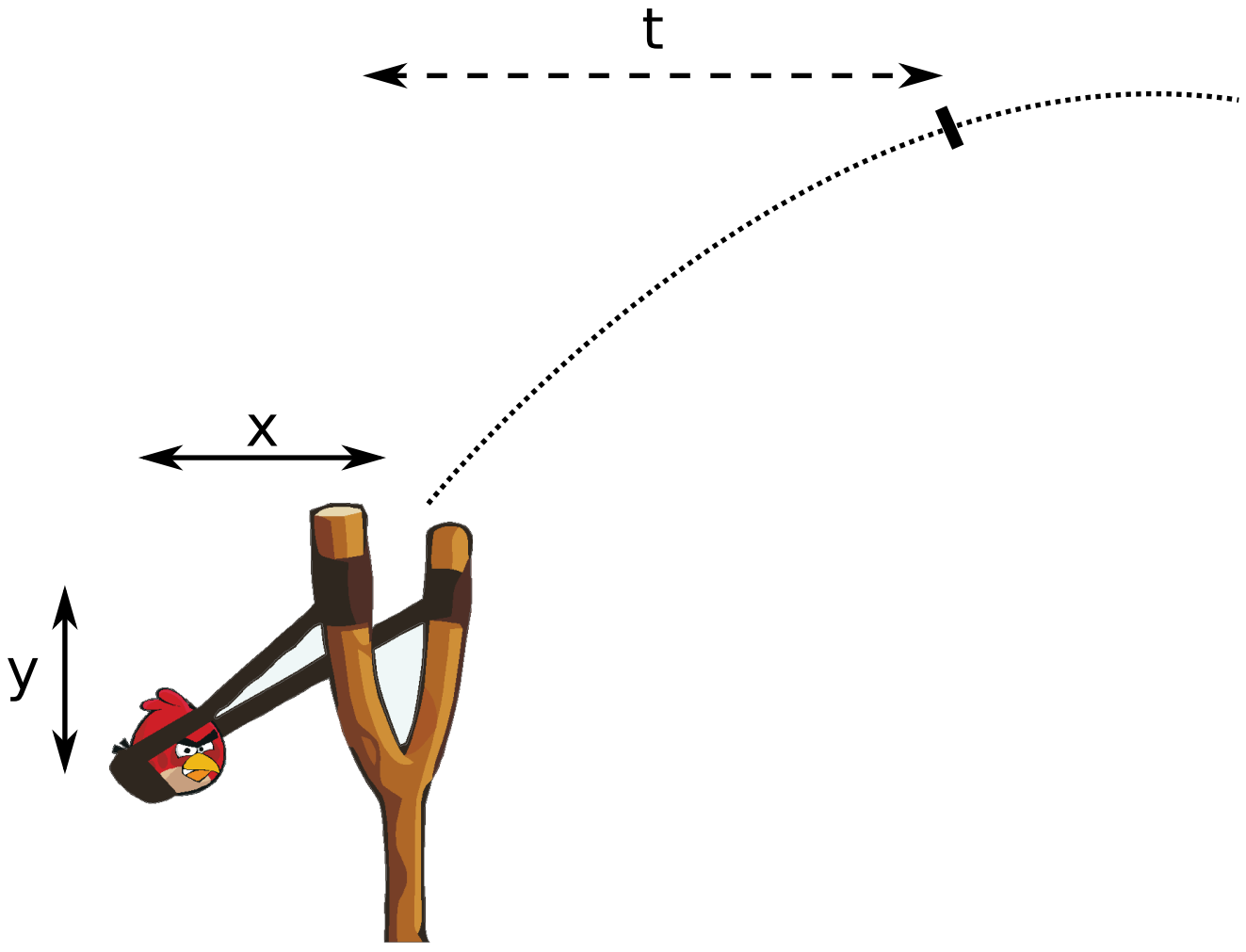


Optimization: Backpropagation using chain rule across the two networks

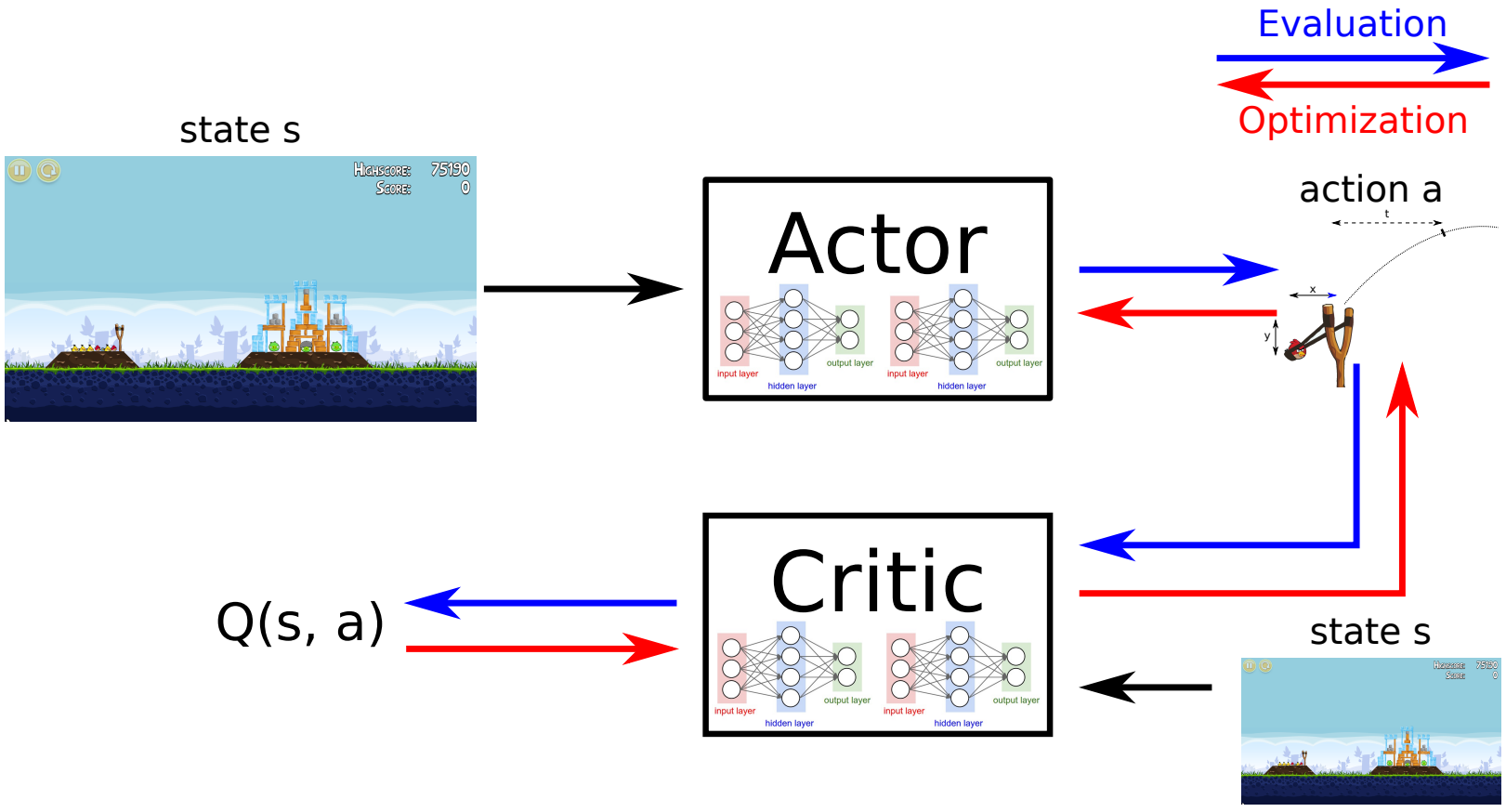

Evaluation
Optimization
state s
?
Actor
input layer
hidden layer
output layer
action a
?
Critic
Q(s, a)
state s
?

Dr. L. Bird: State

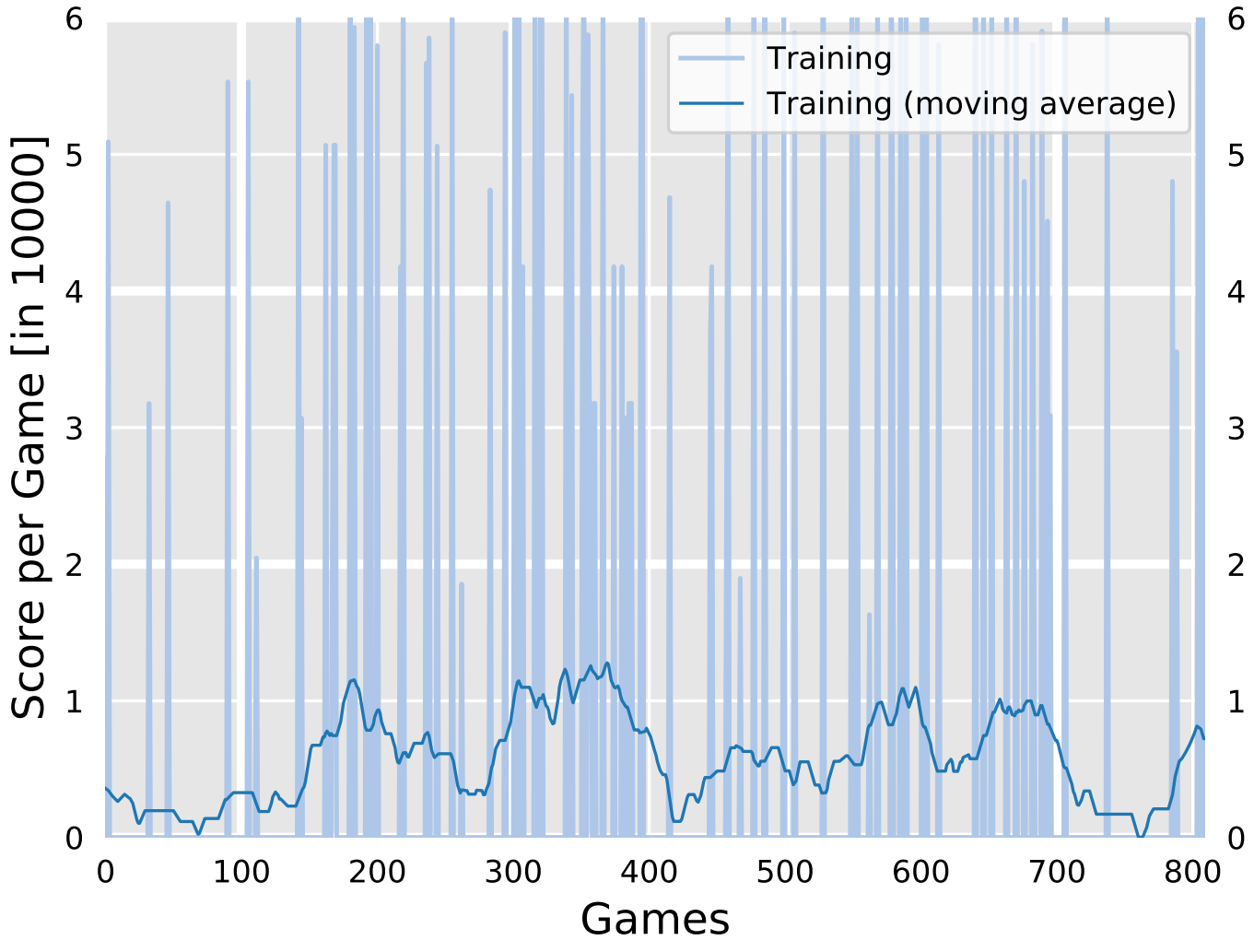
Dr. L. Bird: Action



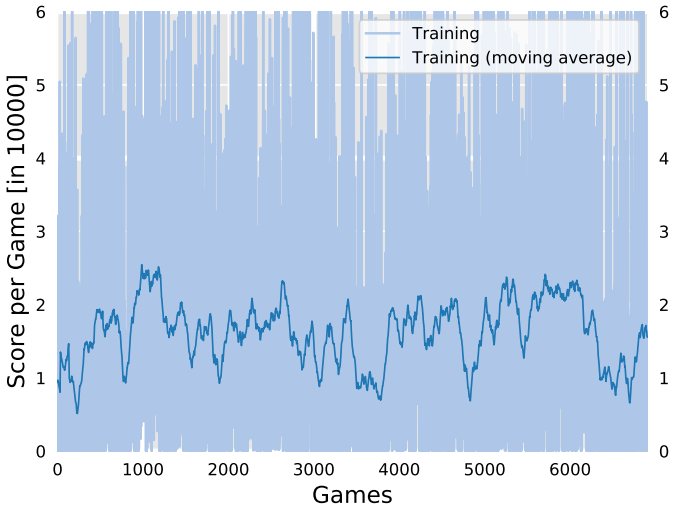
Dr. L. Bird - Agent

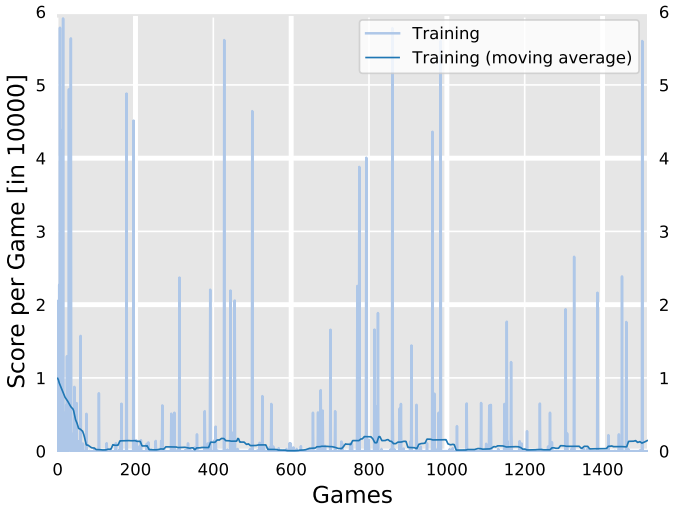


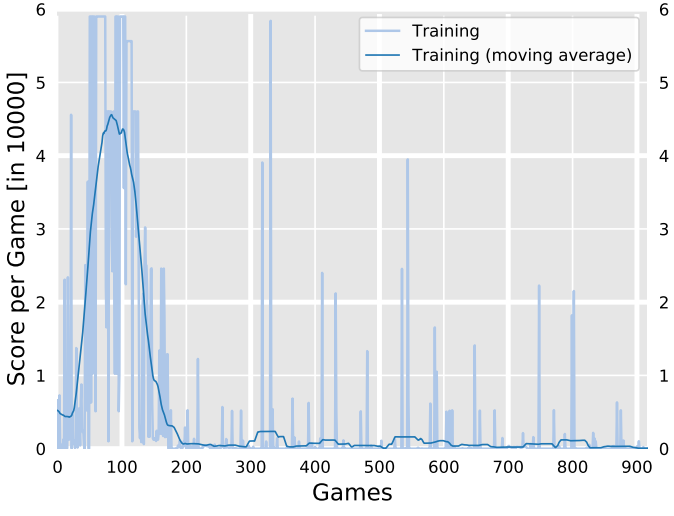
Results

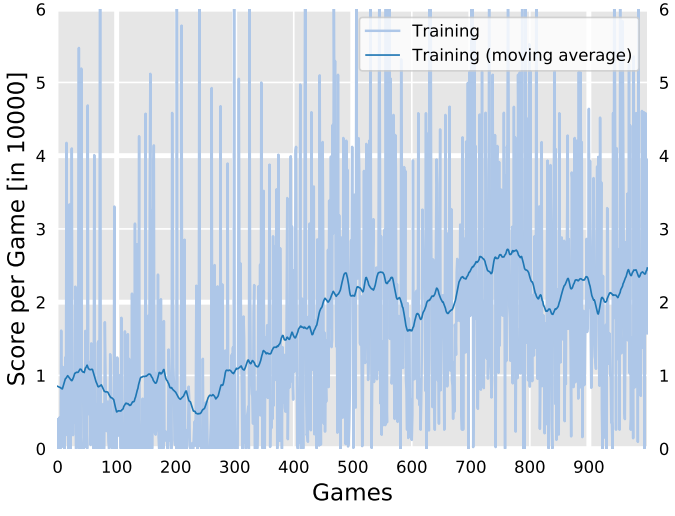

Training
Training (moving average)
Score per Game [in 10000]
Games
0
1
2
3
4
5
6
0
100
200
300
400
500
600
700
800

Results


Training
Training (moving average)
Score per Game [in 10000]
Games
0
1000
2000
3000
4000
5000
6000

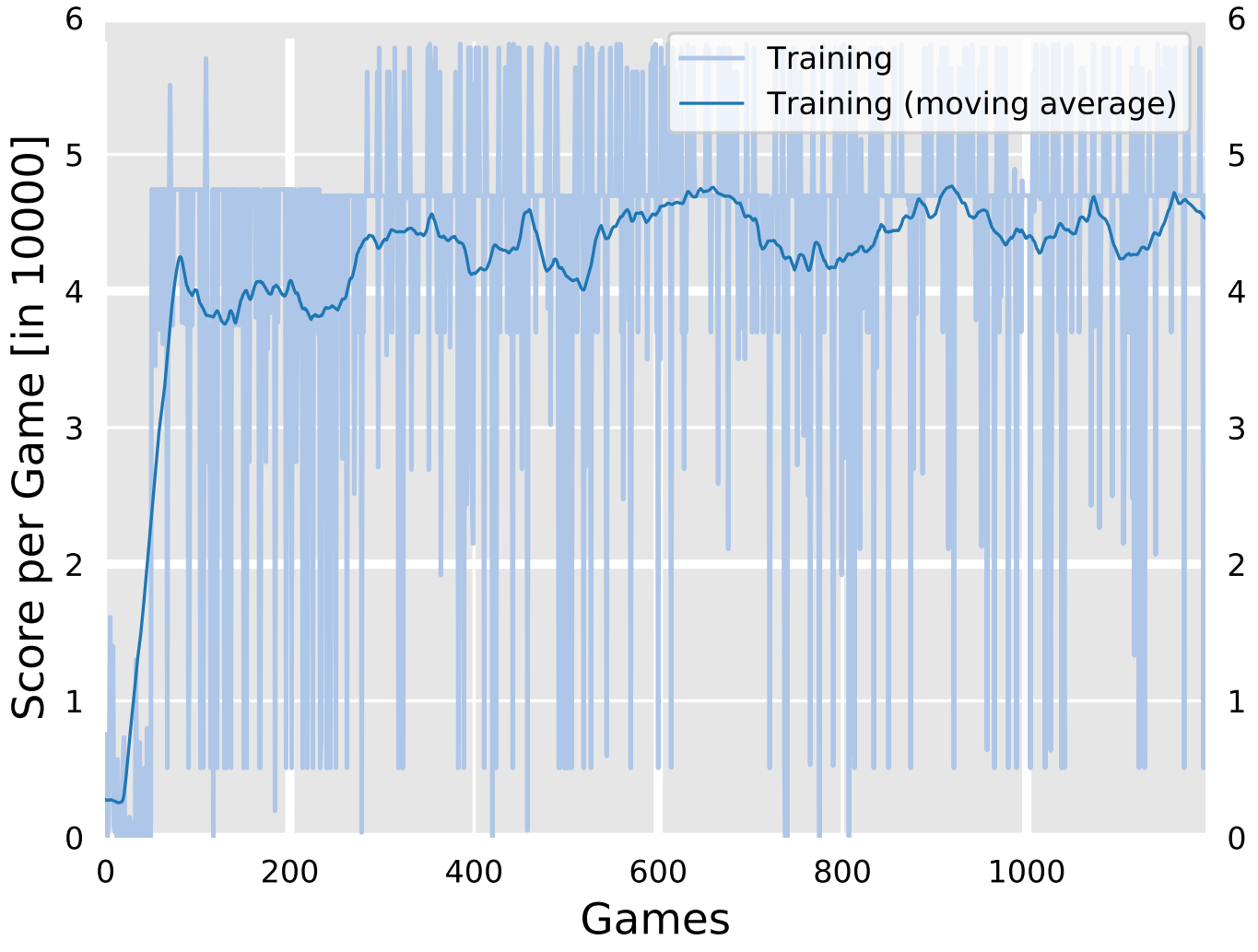

Training
Training (moving average)
Score per Game [in 10000]
Games
0
200
400
600
800
1000
1200
1400


Training
Training (moving average)
Score per Game [in 10000]
Games
0
100
200
300
400
500
600
700
800
900


Training
Training (moving average)
Score per Game [in 10000]
Games
0
100
200
300
400
500
600
700
800
900

One Good Run

INTERNATIONAL CENTER FOR COMPUTATIONAL LOGIC



Tested Versions

INTERNATIONAL CENTER
FOR COMPUTATIONAL LOGIC

- DDPG loss function:
  - Q-Learning-based (off-policy)
  - Sarsa-based (on-policy)
  - TD-based (estimated value)
  - Monte-Carlo-based (cumulative return)
- Stochastic Policy Gradient:
  - policy-based
  - uses statistics of probability distribution
  - output (sampled for action):
    - mean
    - variance
- A3C (asynchronous advantage actor-critic)
  - actor-critic
  - stochastic policy
  - parallel (asynchronous) execution of multiple agents
  - advantage instead of Q-value (relative value of actions)

$\rightarrow$ no success so far

Sources neural network schematics:

- http://www.asimovinstitute.org/neural-network-zoo/
- http://cs231n.github.io/neural-networks-1/

Source policy-based method example:
http://www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching.html

Discussion