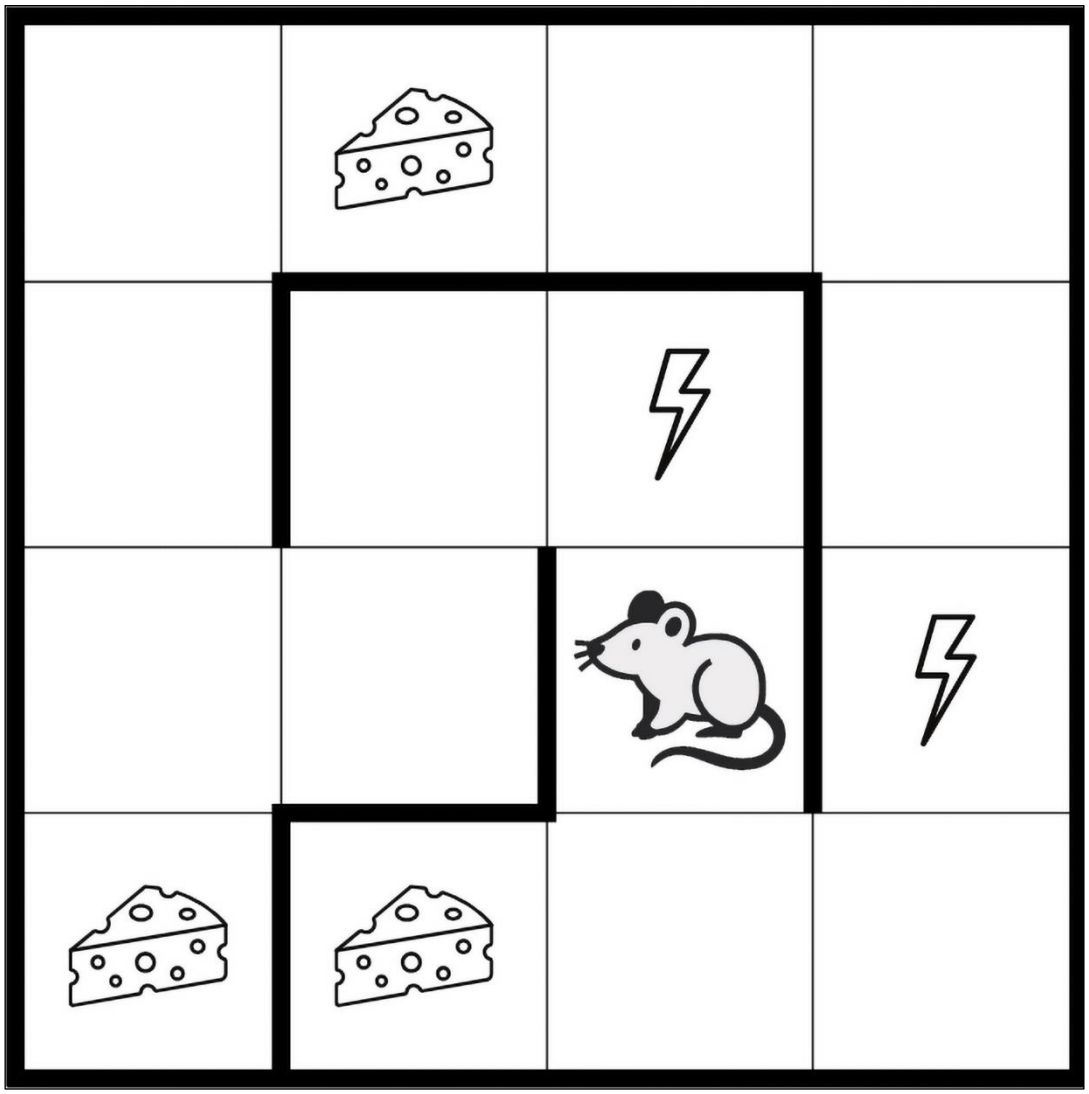
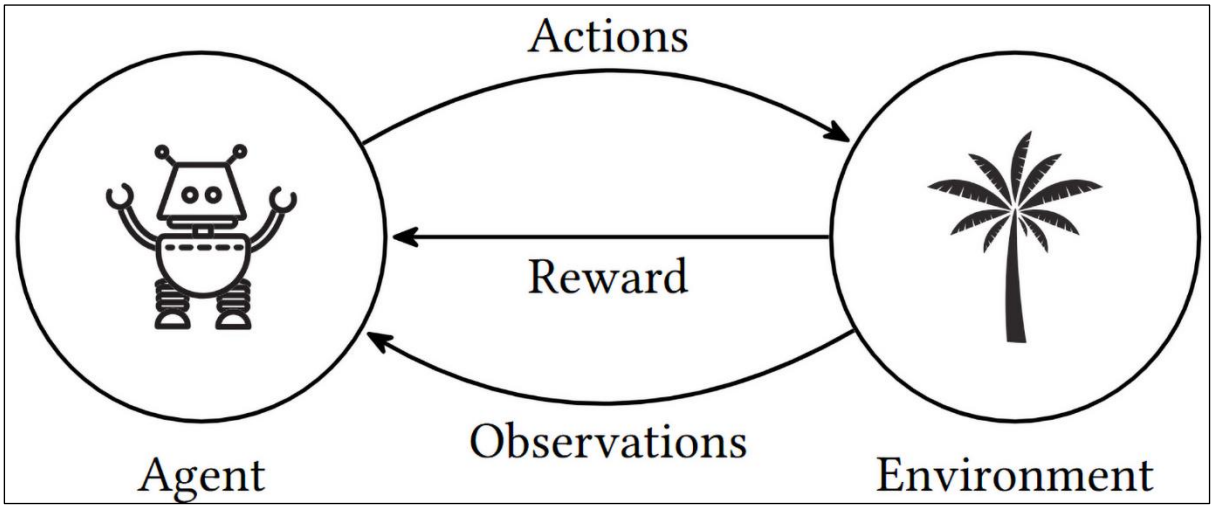
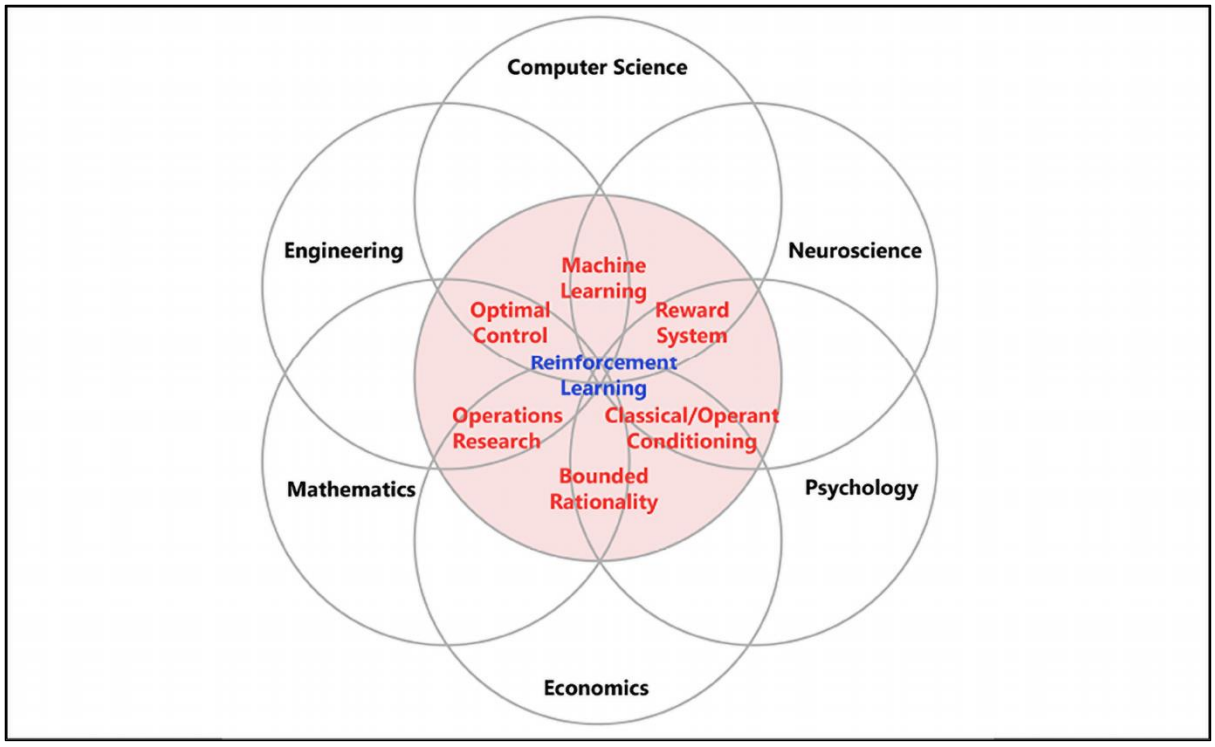
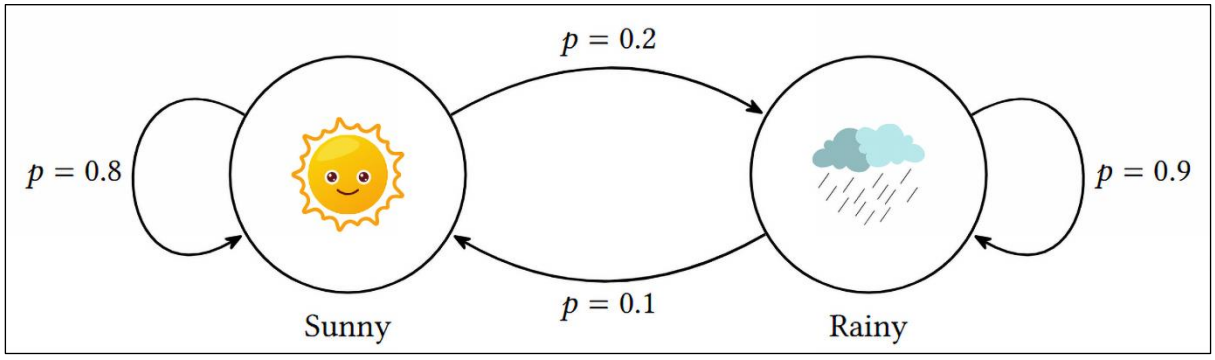
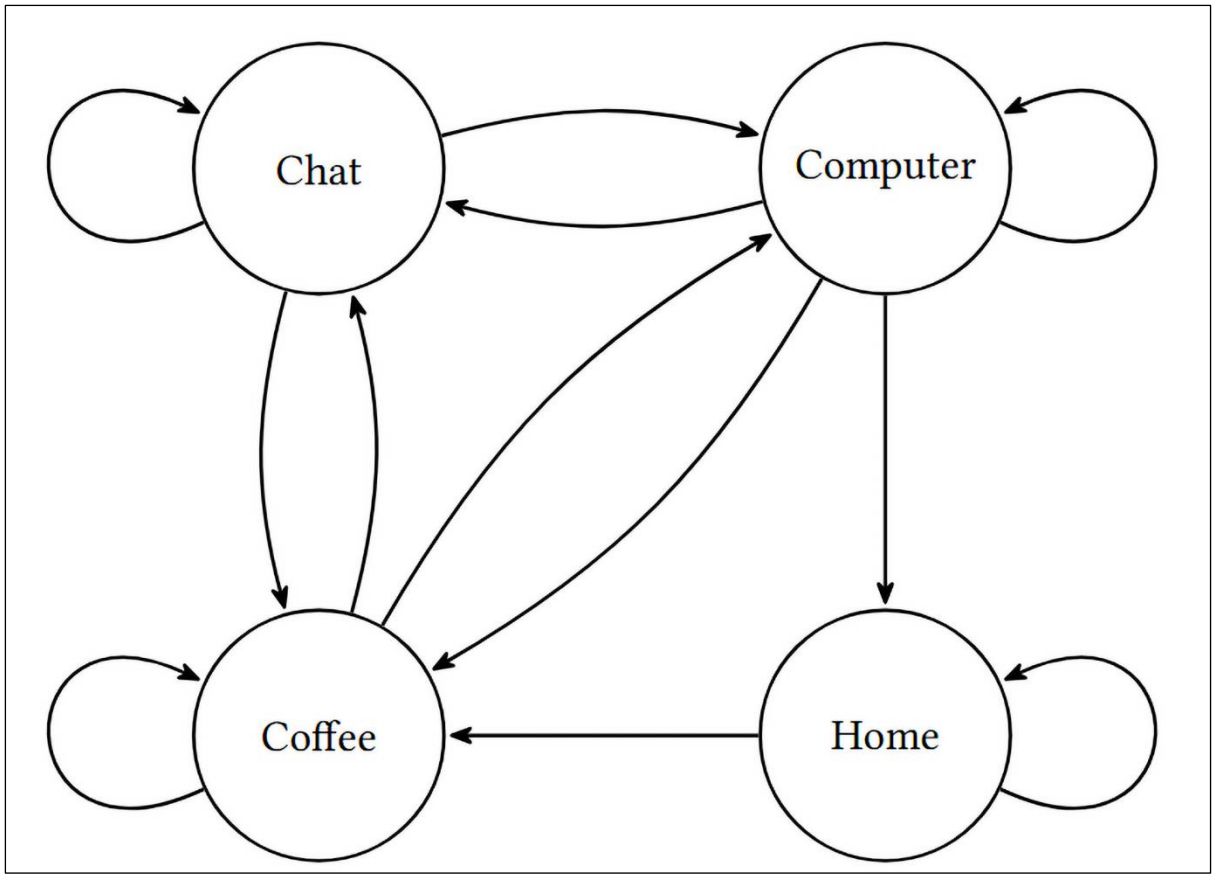
Chapter 1: What Is Reinforcement Learning?

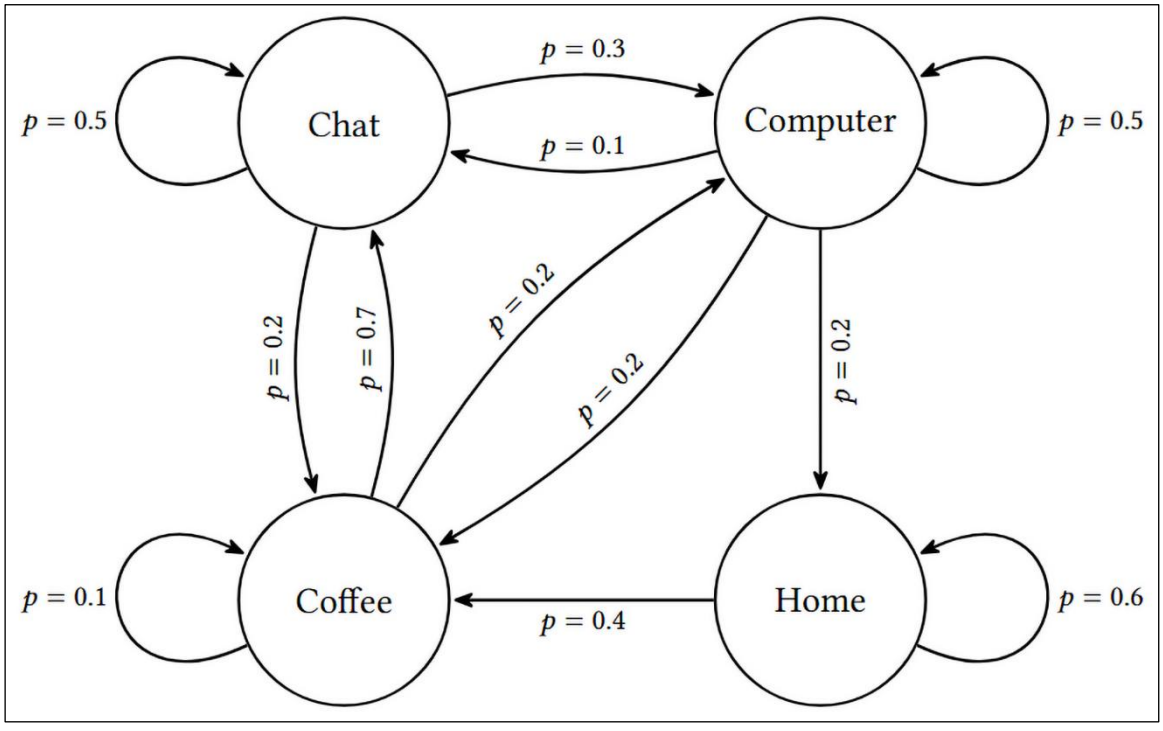


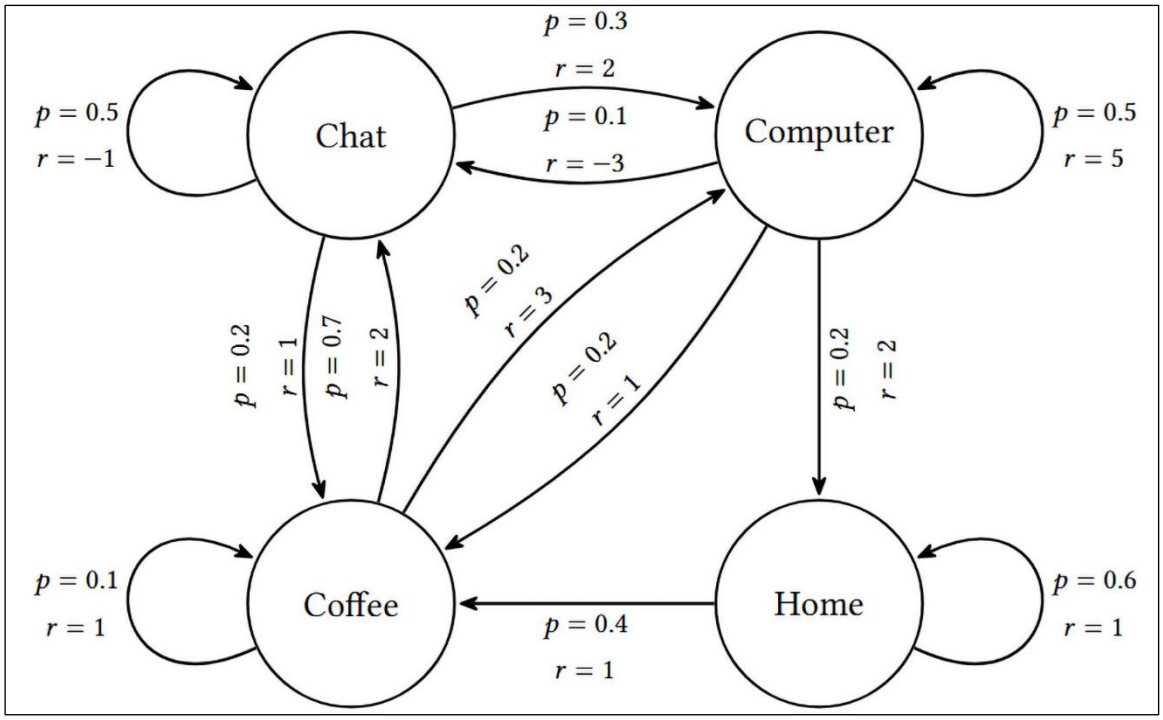


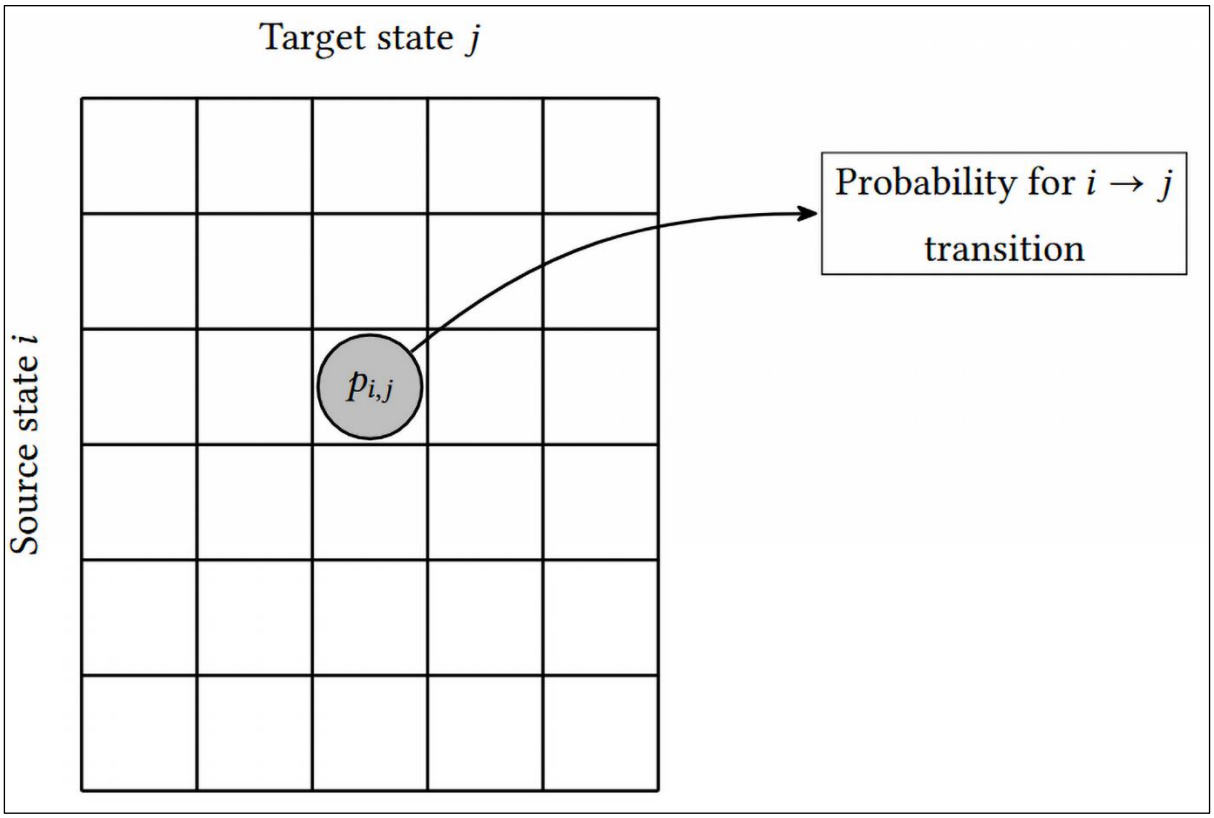

Computer Science
Engineering
Neuroscience
Machine Learning
Optimal Control
Reward System
Reinforcement Learning
Operations Research
Classical/Operant Conditioning
Bounded Rationality
Mathematics
Psychology
Economics

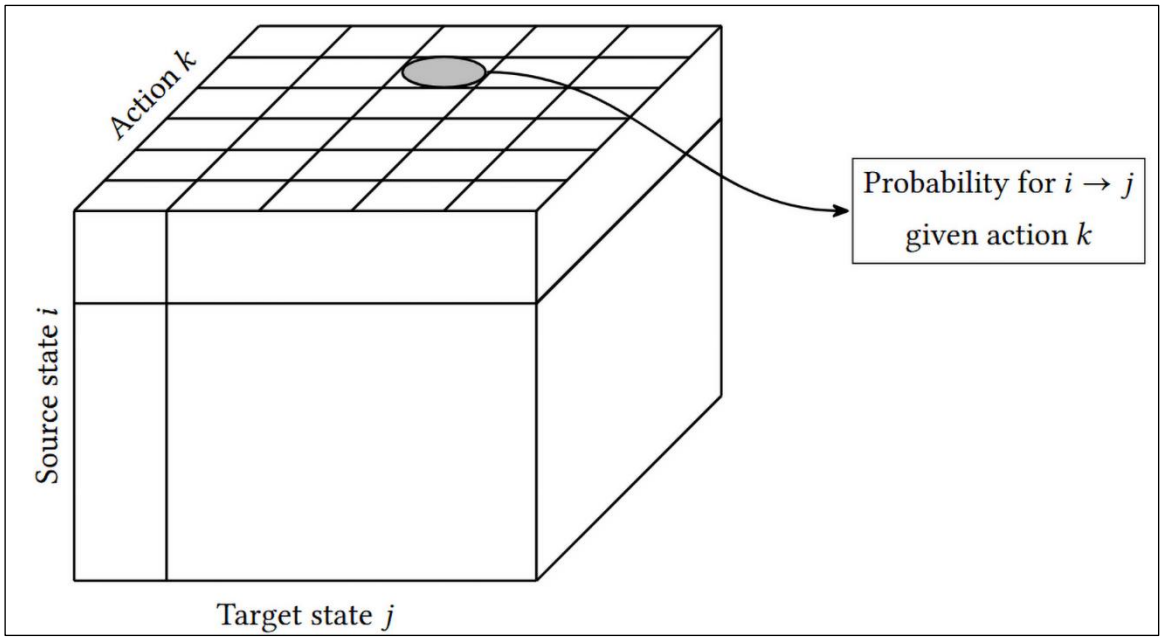

$p = 0.2$
$p = 0.8$
$p = 0.9$
$p = 0.1$
Sunny
Rainy

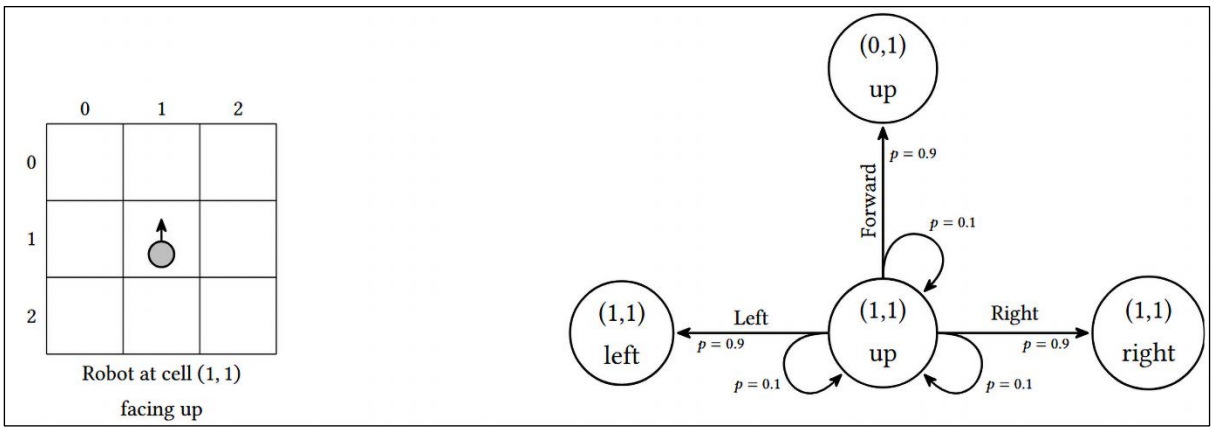

Chat
Computer
Coffee
Home


Chat
Computer
Coffee
Home
p = 0.3
p = 0.1
p = 0.5
p = 0.5
p = 0.2
p = 0.7
p = 0.2
p = 0.2
p = 0.1
p = 0.4
p = 0.6

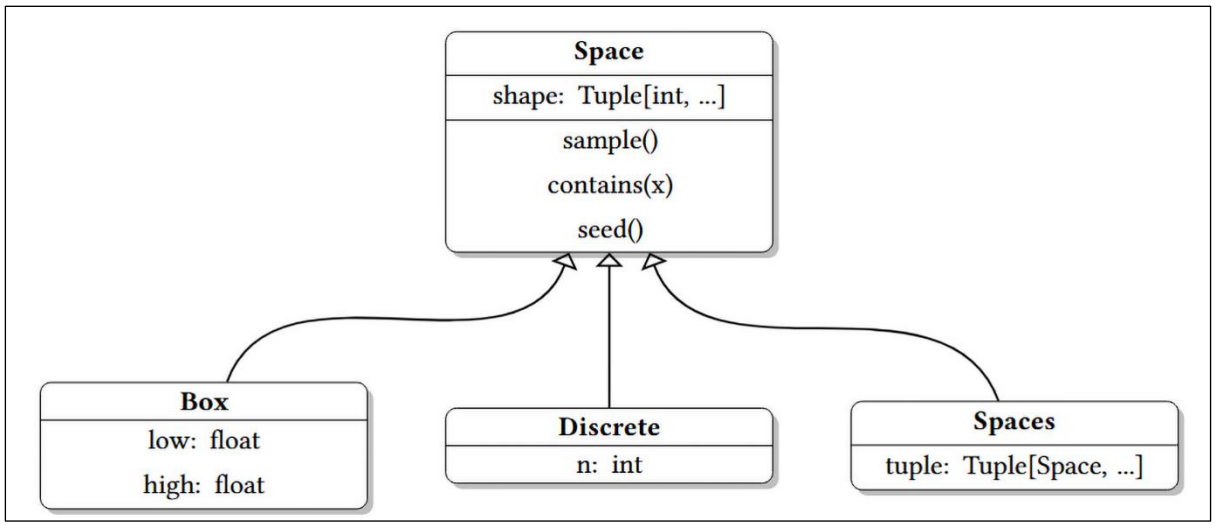

Chat
Computer
Coffee
Home
p = 0.5
r = −1
p = 0.3
r = 2
p = 0.1
r = −3
p = 0.5
r = 5
p = 0.2
r = 1
p = 0.7
r = 2
p = 0.2
r = 3
p = 0.2
r = 1
p = 0.2
r = 2
p = 0.1
r = 1
p = 0.4
r = 1
p = 0.6
r = 1

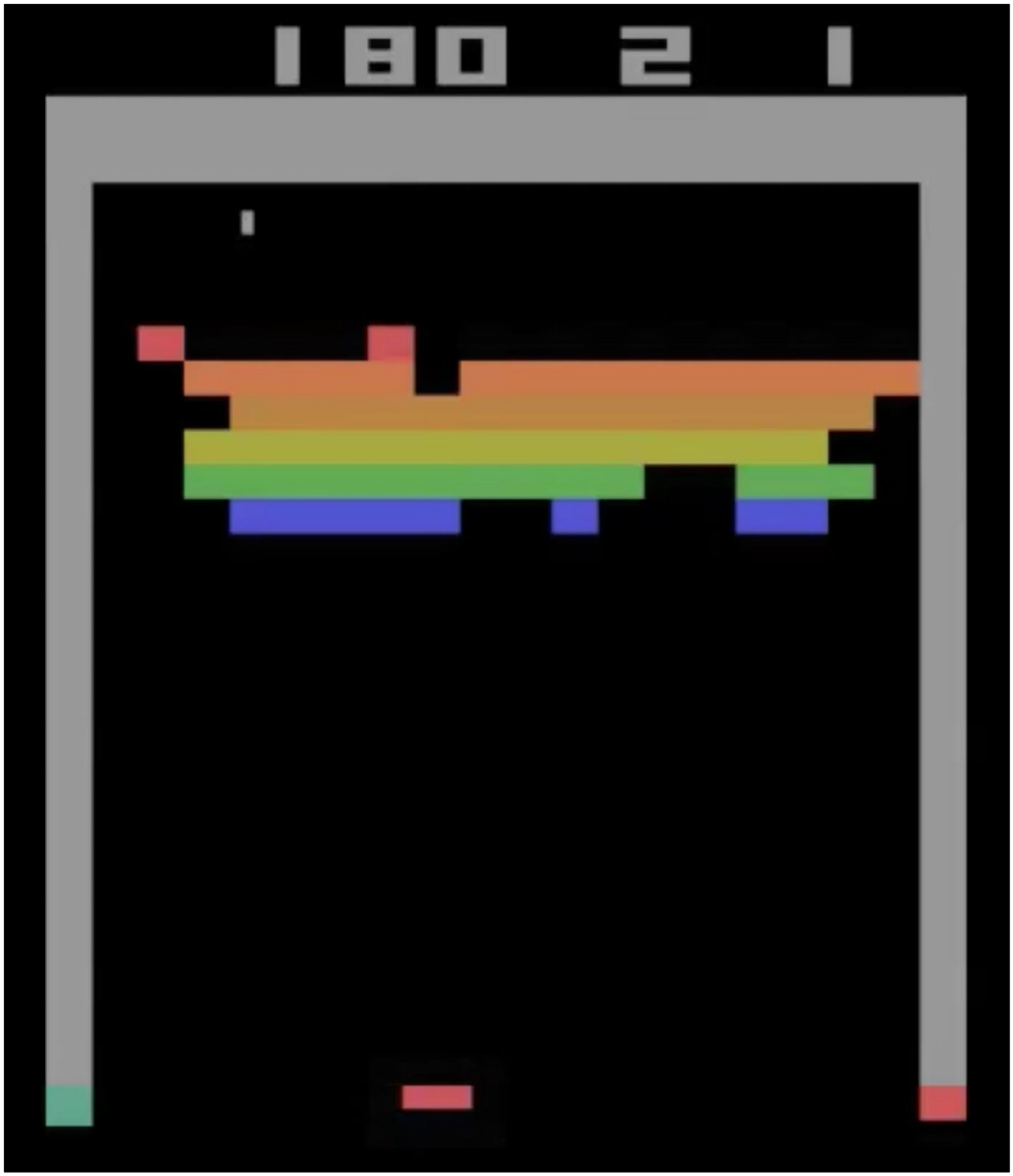

Target state j
Source state i
$p_{i,j}$
Probability for $i \rightarrow j$ transition

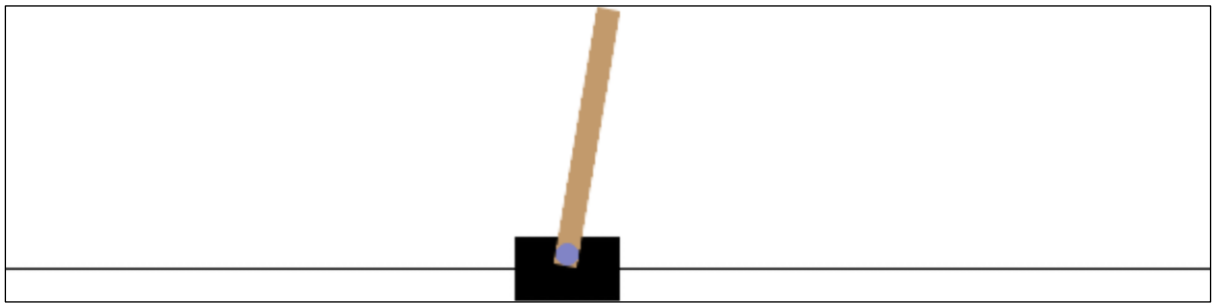

Action k
Source state i
Target state j
Probability for $i \rightarrow j$
given action k

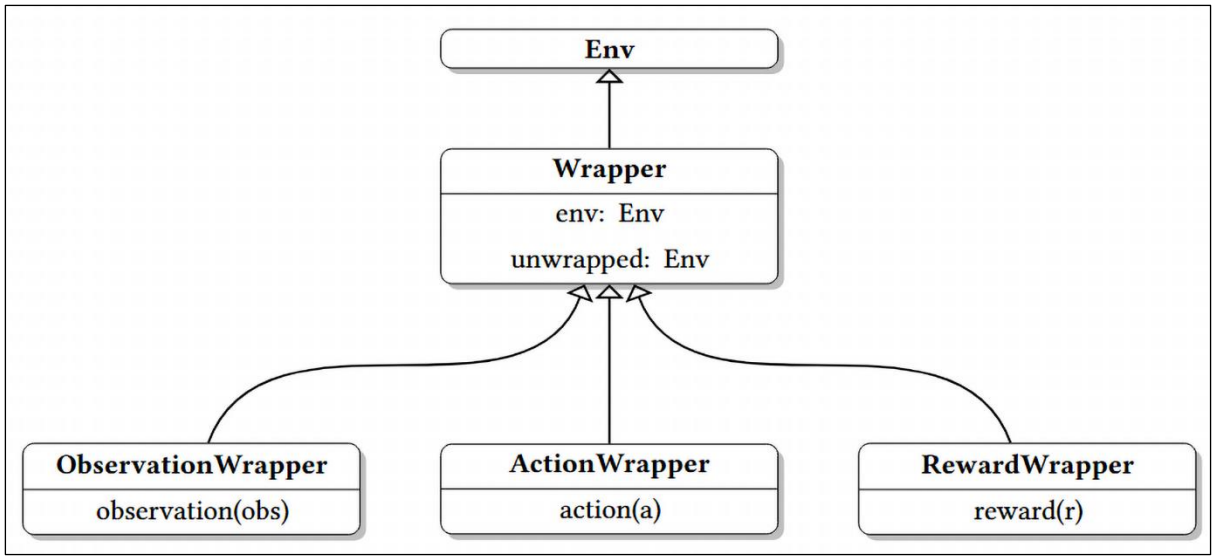

0
1
2
0
1
2
Robot at cell (1, 1)
facing up
(0,1)
up
Forward
$p = 0.9$
$p = 0.1$
(1,1)
left
Left
$p = 0.9$
(1,1)
up
Right
$p = 0.9$
(1,1)
right
$p = 0.1$
$p = 0.1$

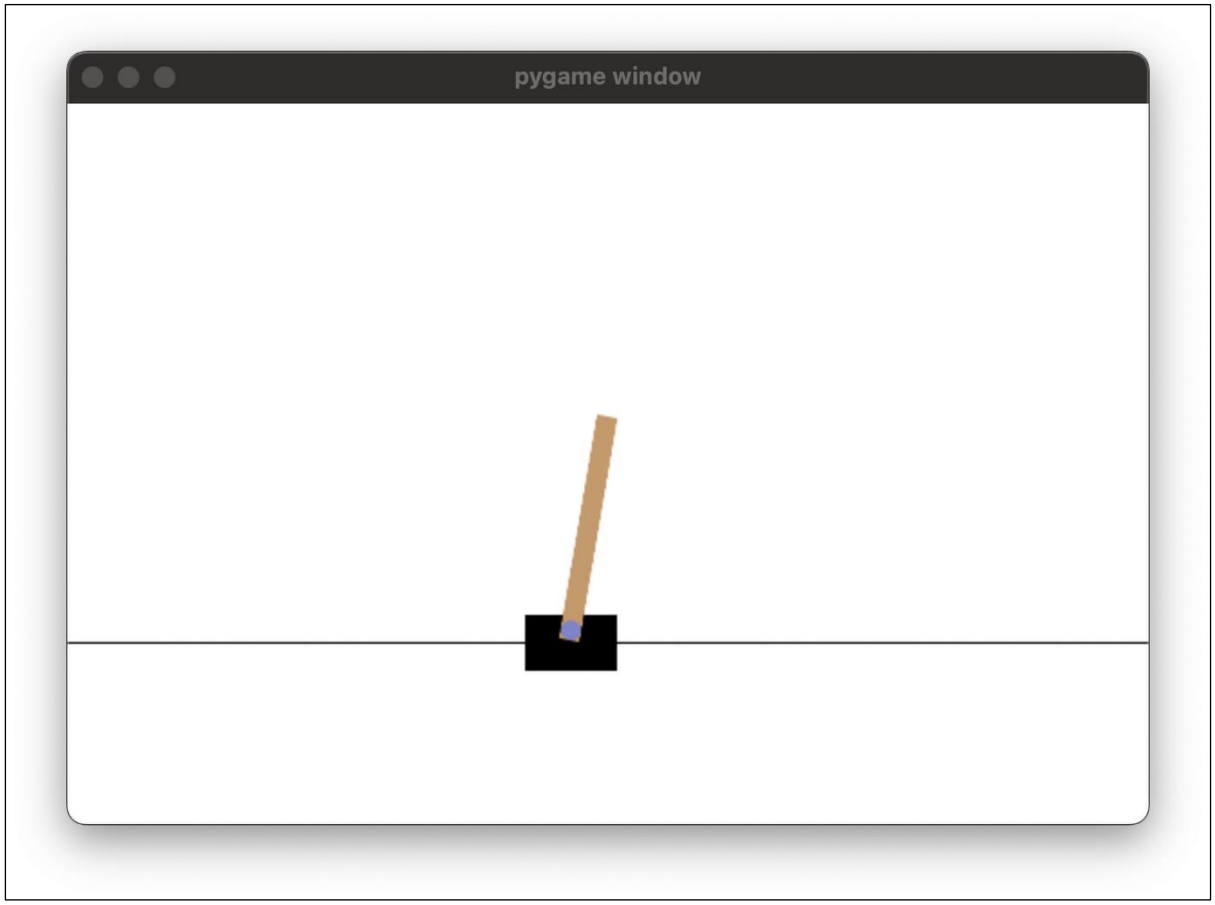
Chapter 2: OpenAI Gym API and Gymansium



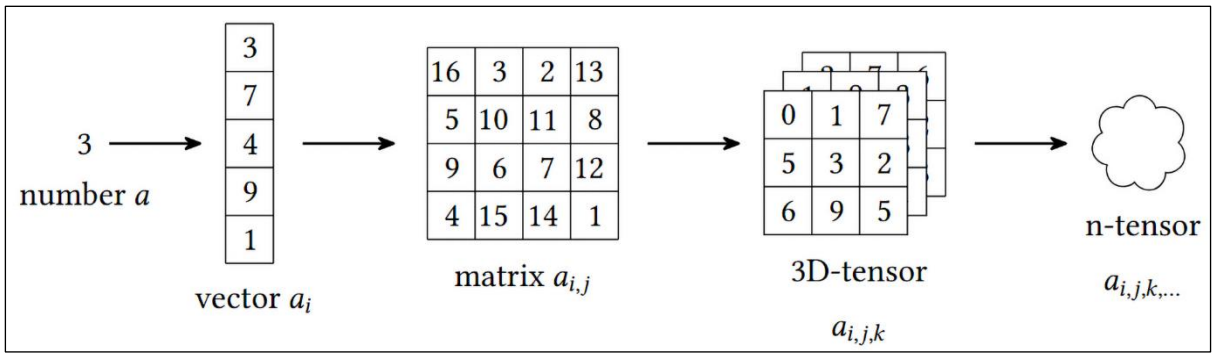

180 2 1

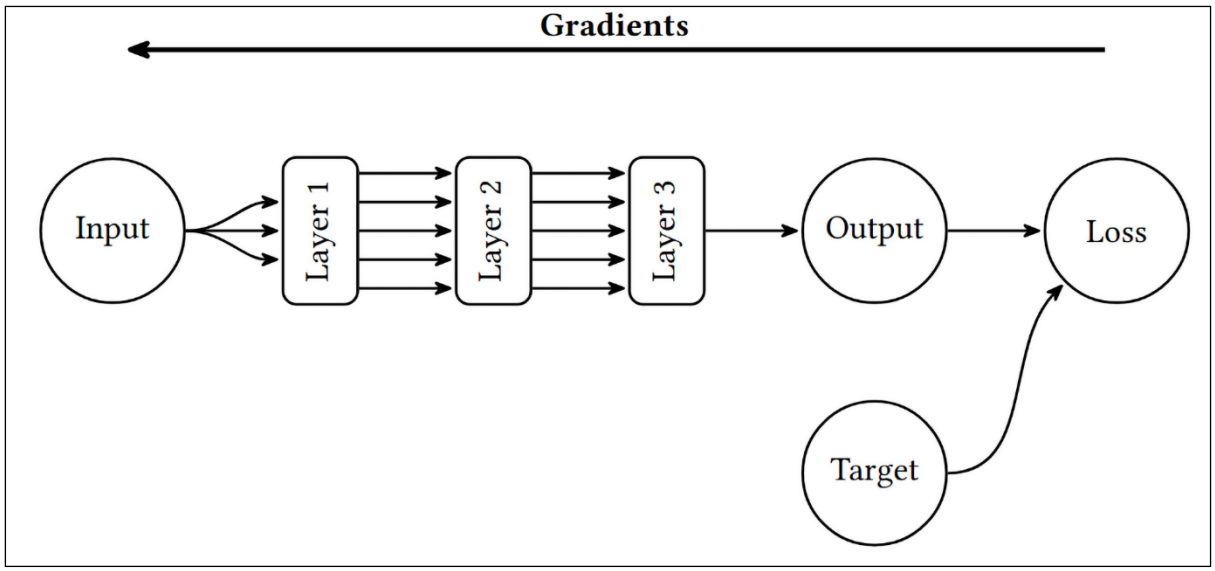


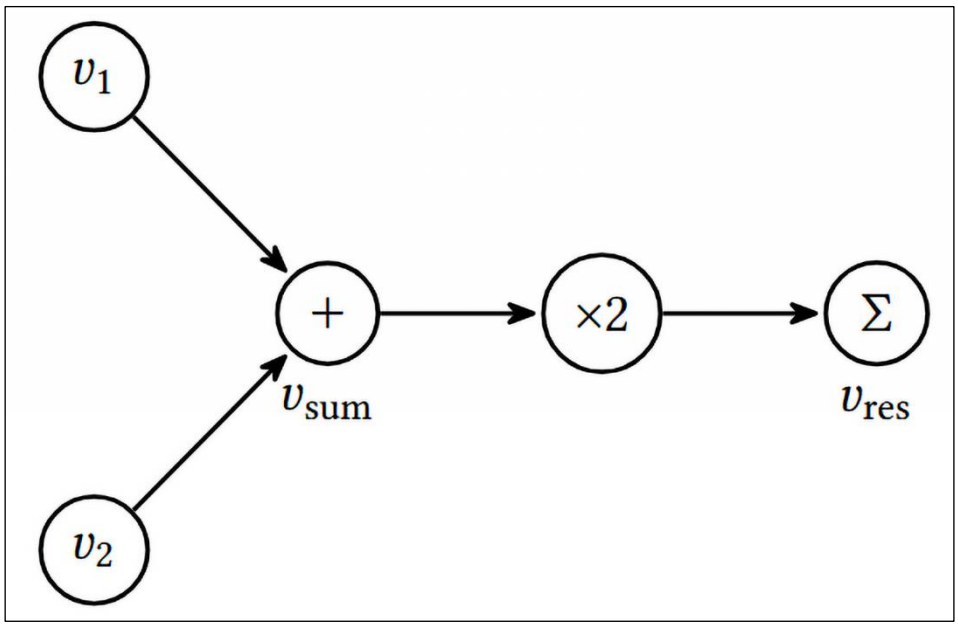

Env
Wrapper
env: Env
unwrapped: Env
ObservationWrapper
observation(obs)
ActionWrapper
action(a)
RewardWrapper
reward(r)

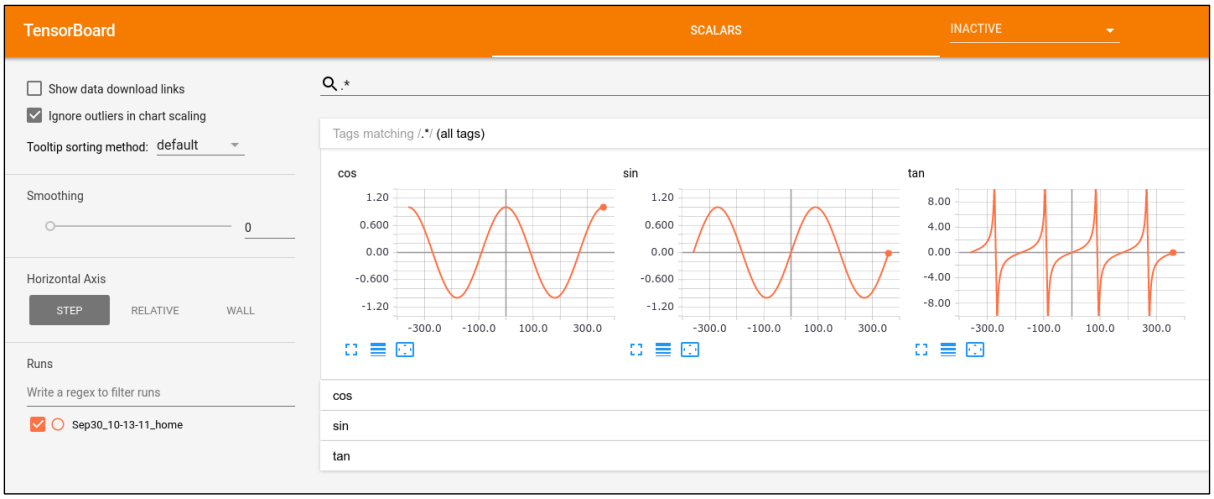

pygame window

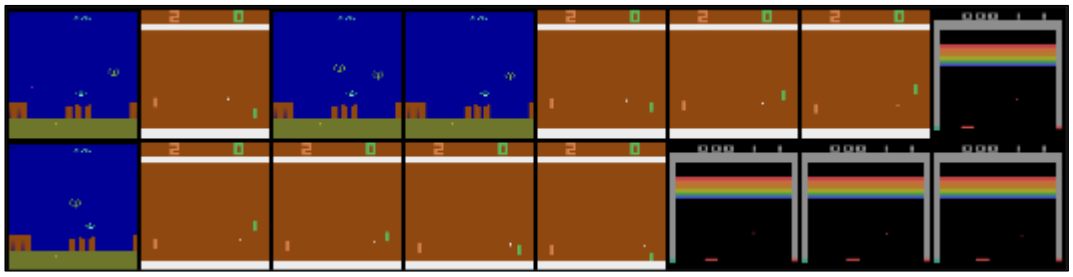
Chapter 3: Deep Learning with PyTorch,

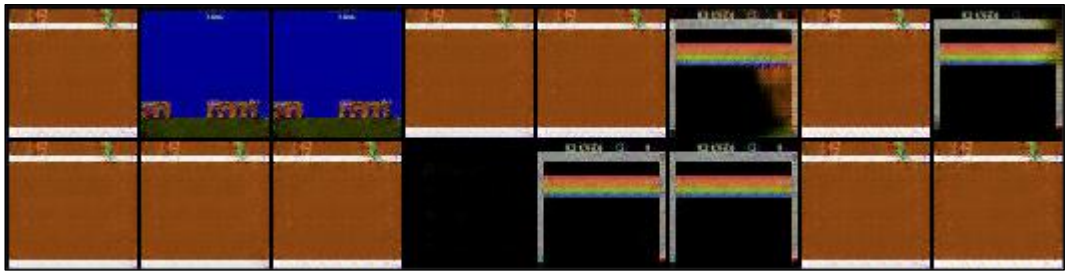




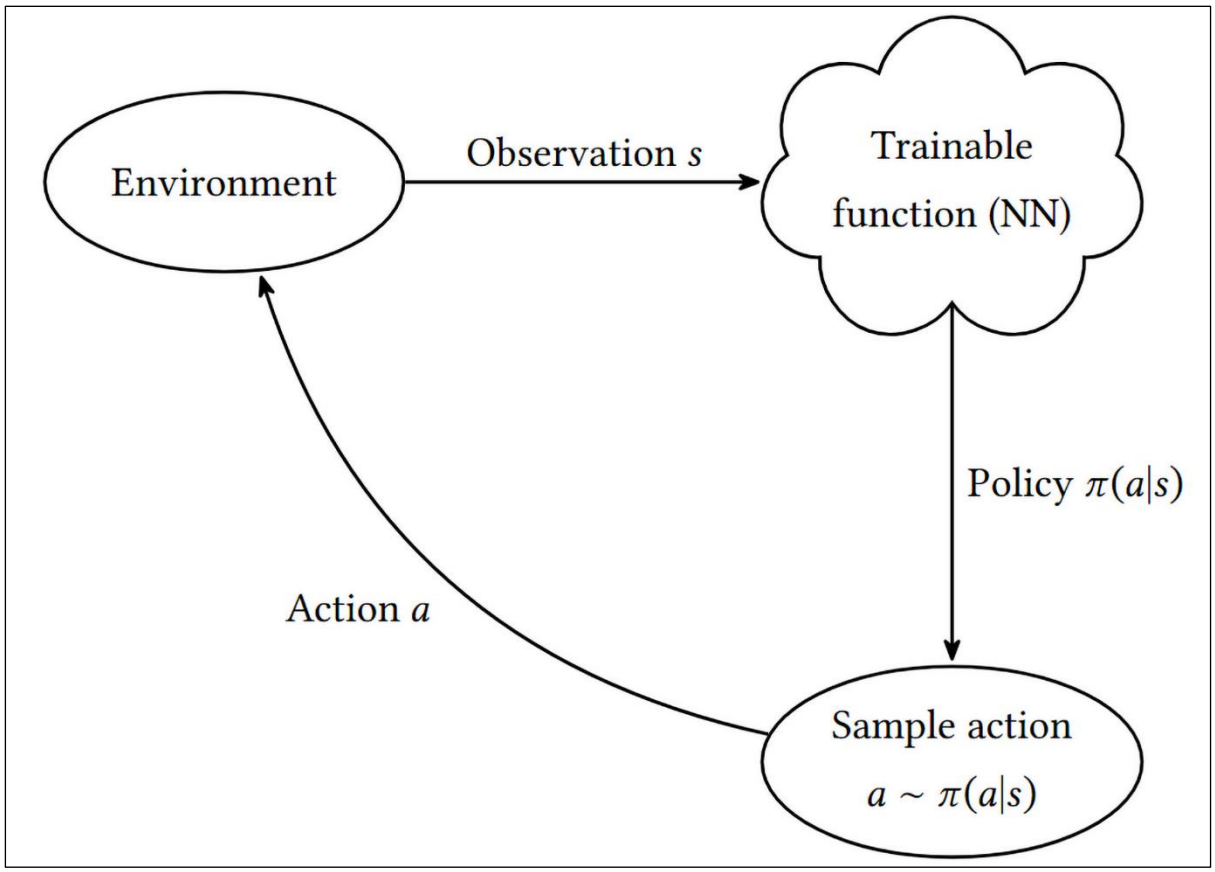



TensorBoard
SCALARS
INACTIVE
Show data download links
Ignore outliers in chart scaling
Tooltip sorting method: default
Smoothing
0
Horizontal Axis
STEP
RELATIVE
WALL
Runs
Write a regex to filter runs
Sep30_10-13-11_home
.*
Tags matching /.*/ (all tags)
cos
sin
tan
1.20
0.600
0.00
-0.600
-1.20
8.00
4.00
-4.00
-8.00
-300.0
-100.0
100.0
300.0

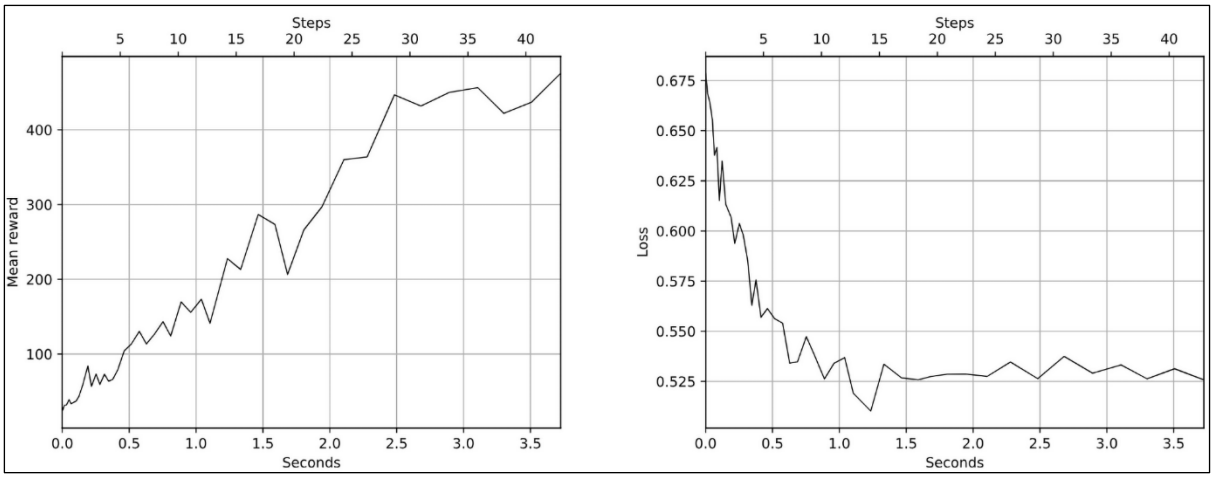


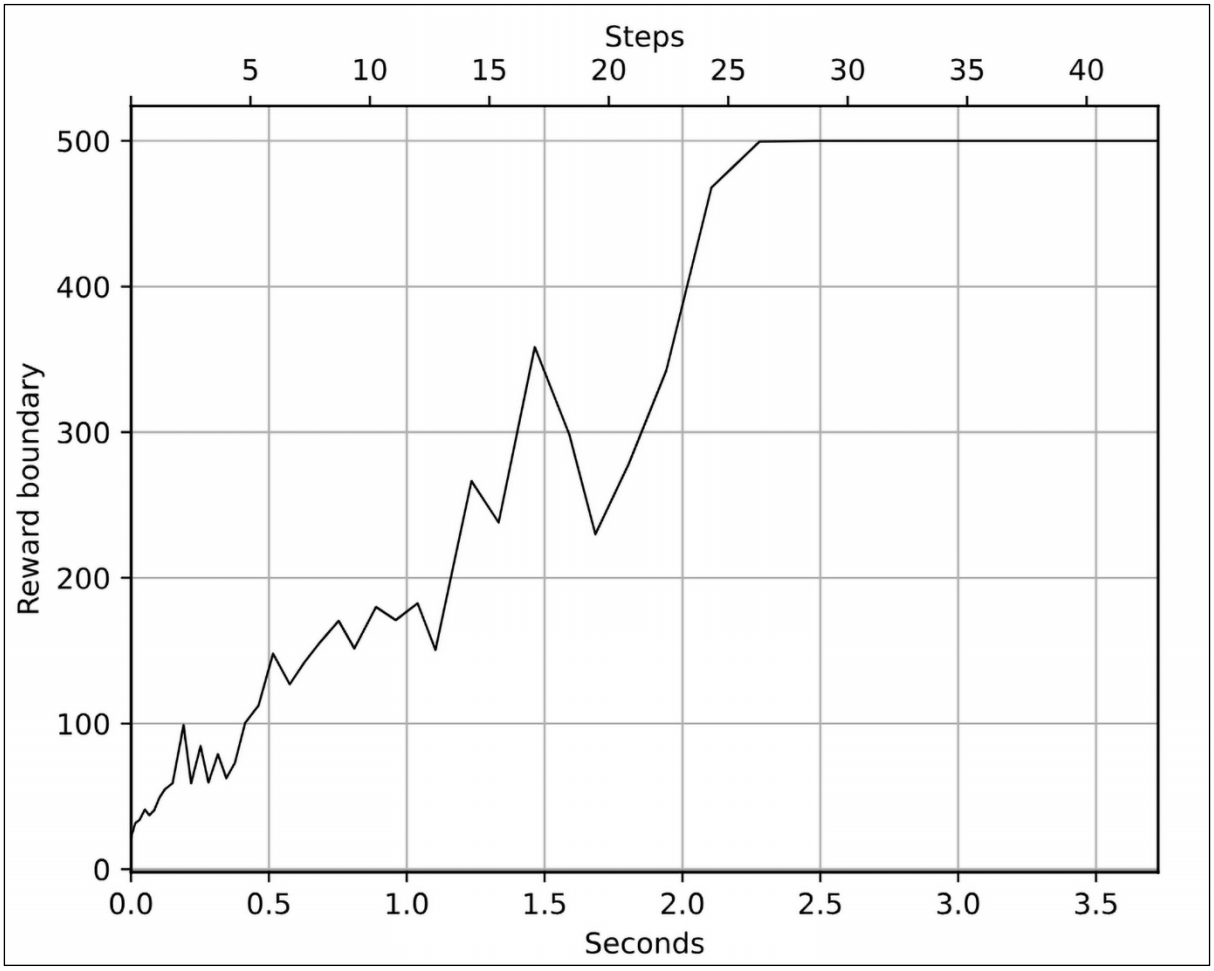


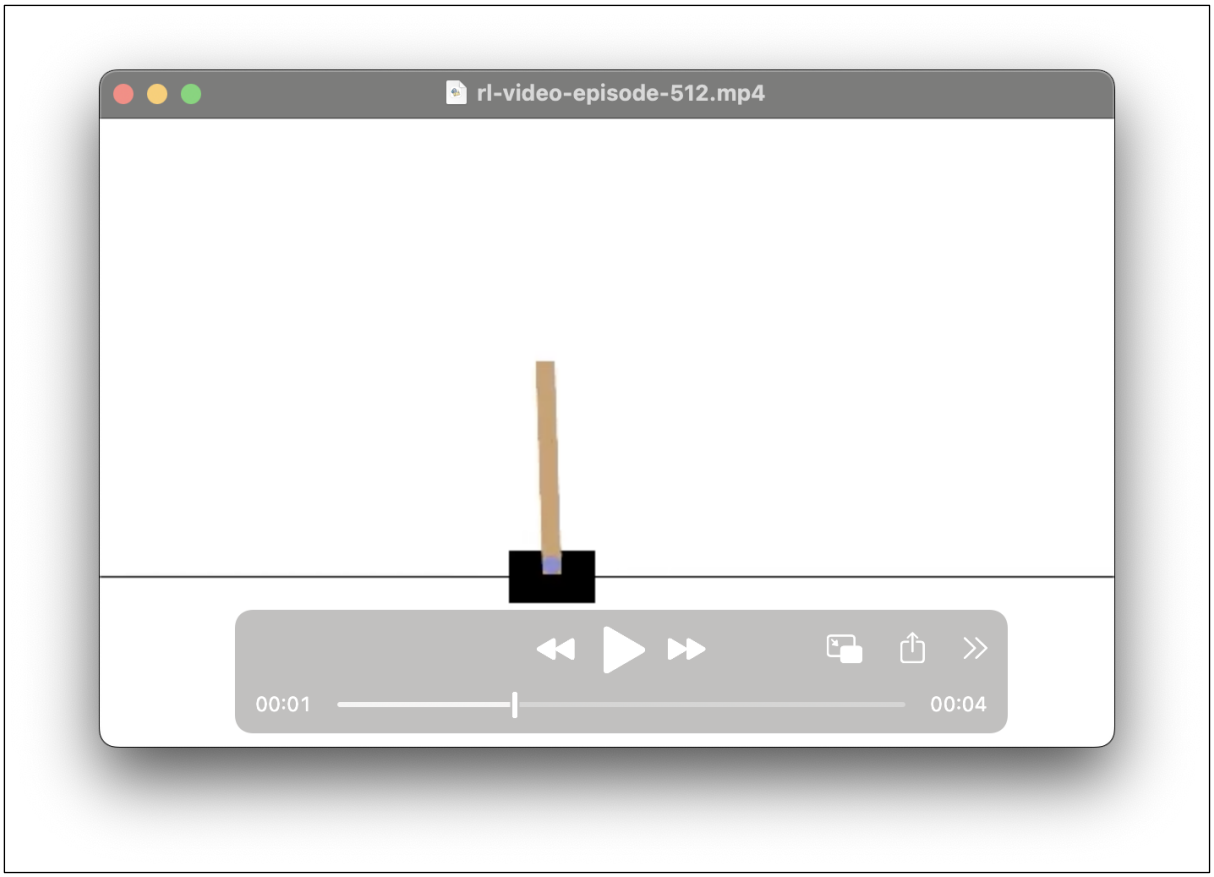
Chapter 4: The Cross-Entropy Method

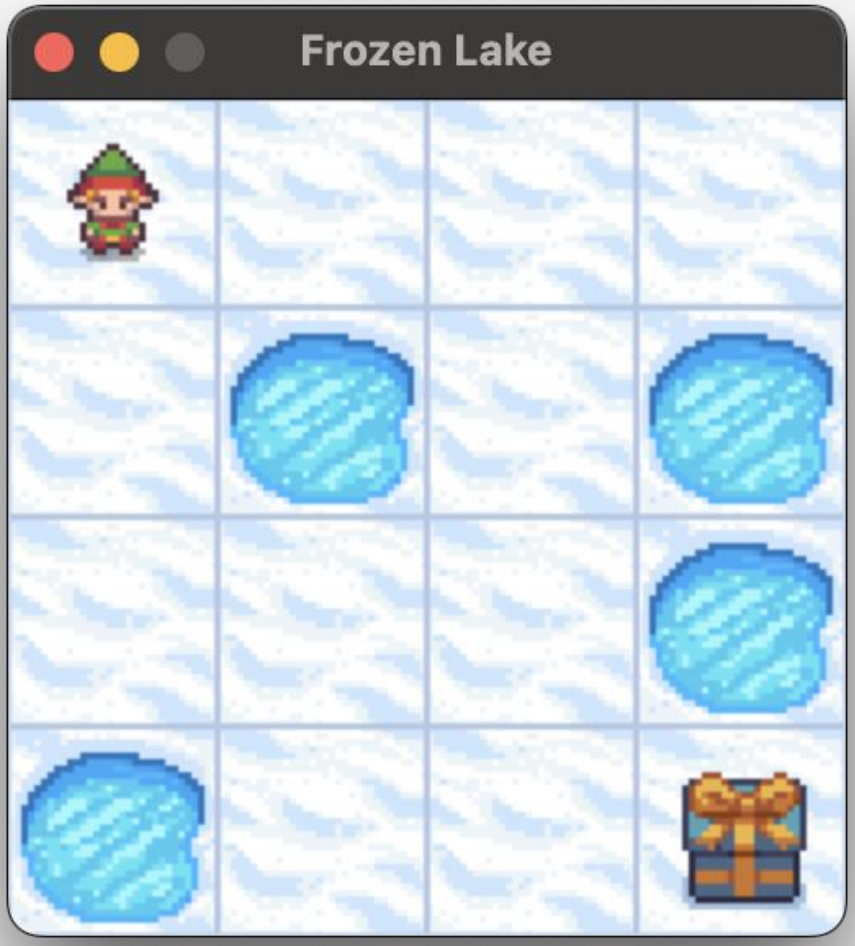


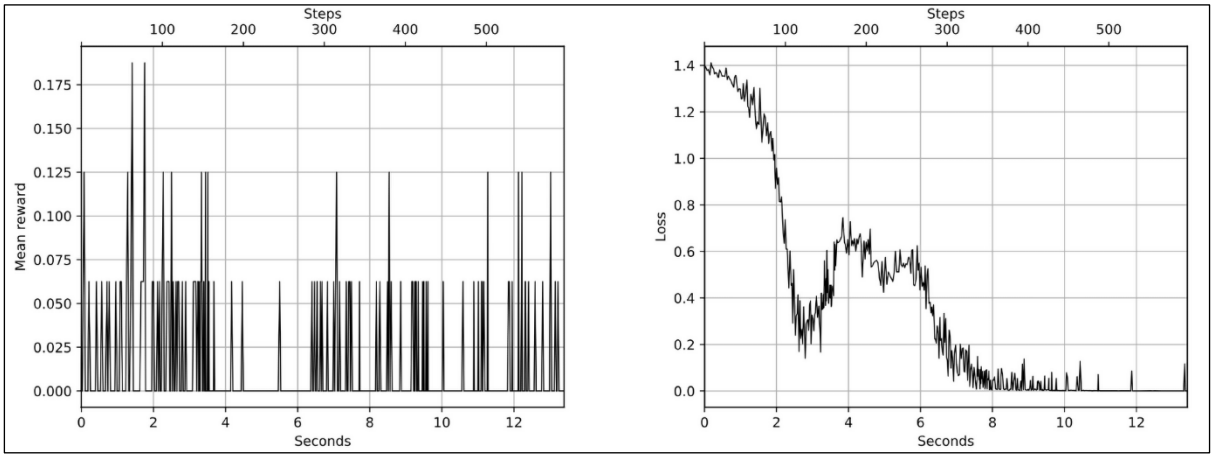
Episode 1:	o_1, a_1, r_1 | o_2, a_2, r_2 | o_3, a_3, r_3 | o_4, a_4, r_4	$R = r_1 + r_2 + r_3 + r_4$
Episode 2:	o_1, a_1, r_1 | o_2, a_2, r_2 | o_3, a_3, r_3	$R = r_1 + r_2 + r_3$
Episode 3:	o_1, a_1, r_1 | o_2, a_2, r_2	$R = r_1 + r_2$
Episode 4:	o_1, a_1, r_1 | o_2, a_2, r_2 | o_3, a_3, r_3	$R = r_1 + r_2 + r_3$

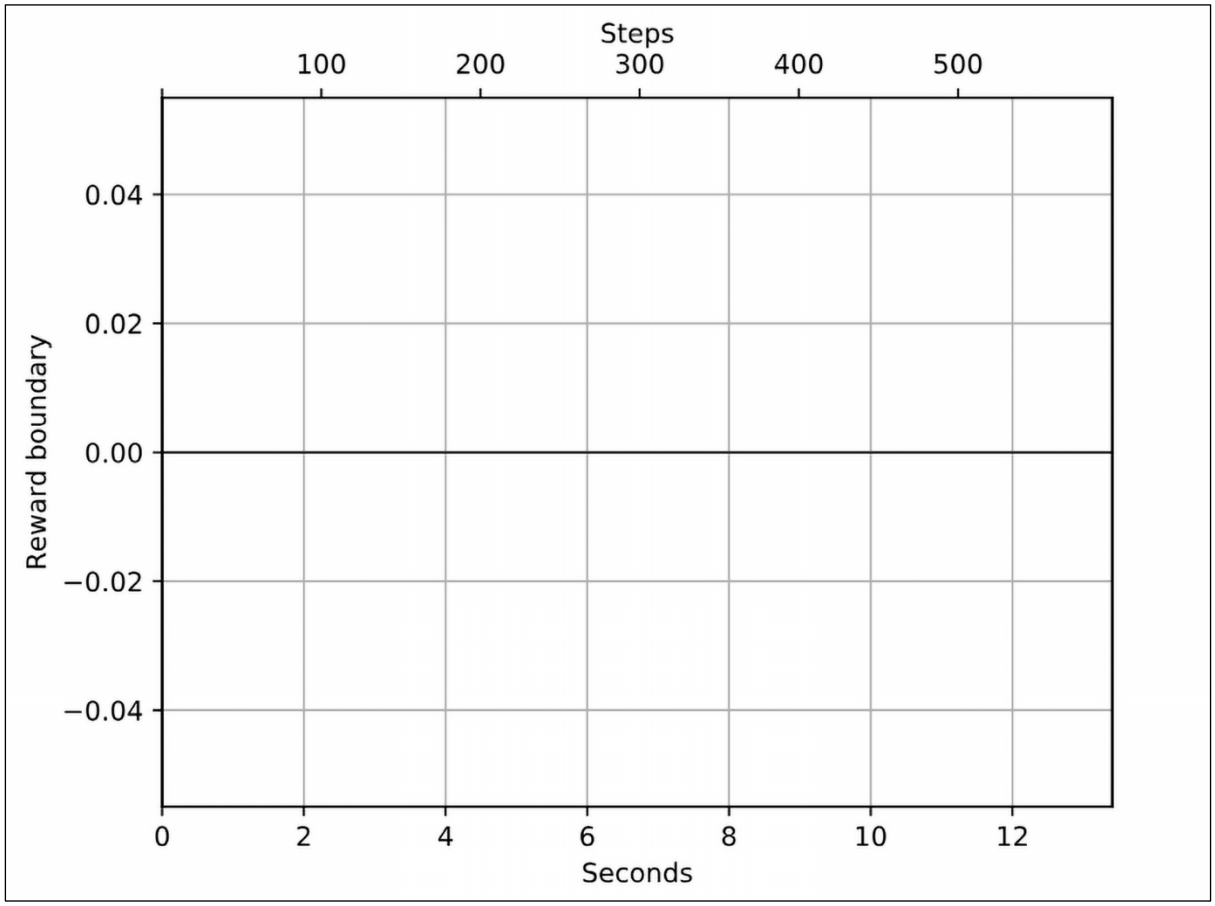

Steps
5 10 15 20 25 30 35 40
Mean reward
100 200 300 400
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
Seconds
Steps
5 10 15 20 25 30 35 40
Loss
0.675 0.650 0.625 0.600 0.575 0.550 0.525
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
Seconds

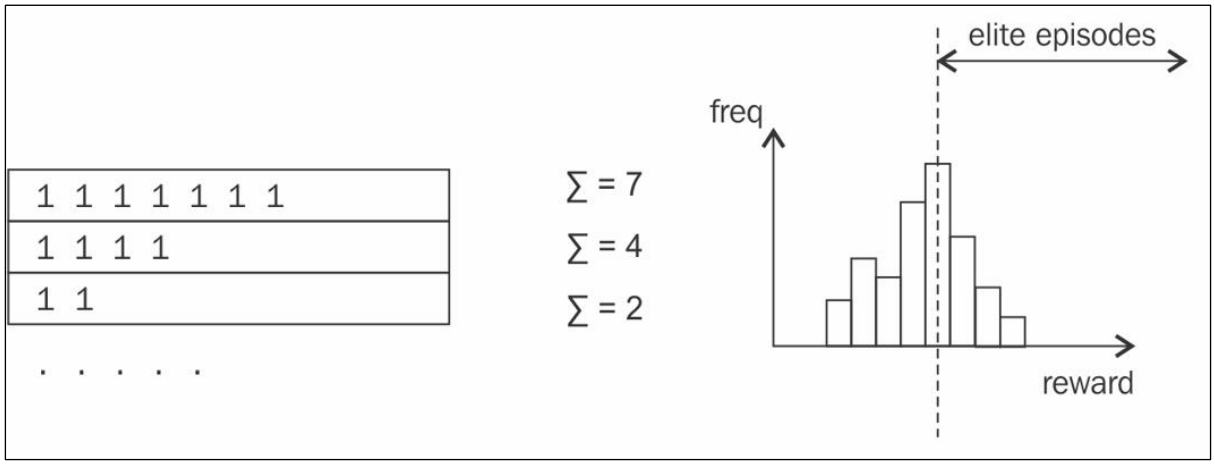

Steps
5 10 15 20 25 30 35 40
Reward boundary
0 100 200 300 400 500
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
Seconds

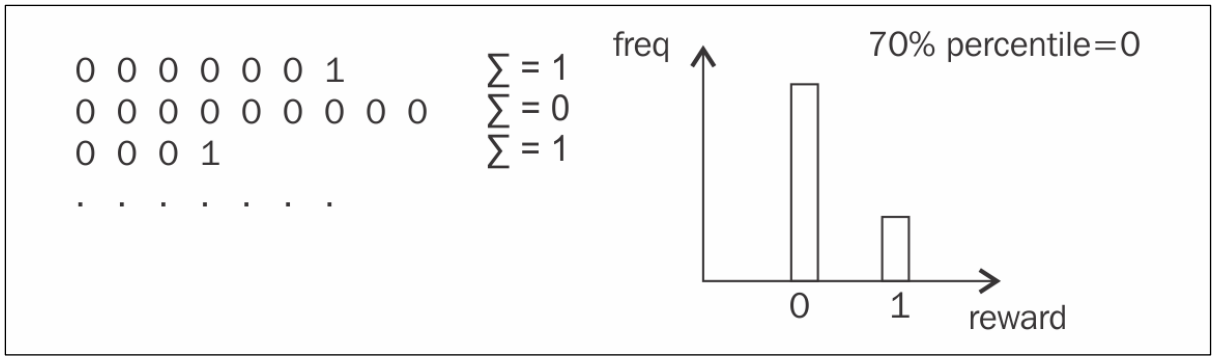

rl-video-episode-512.mp4
00:01
00:04

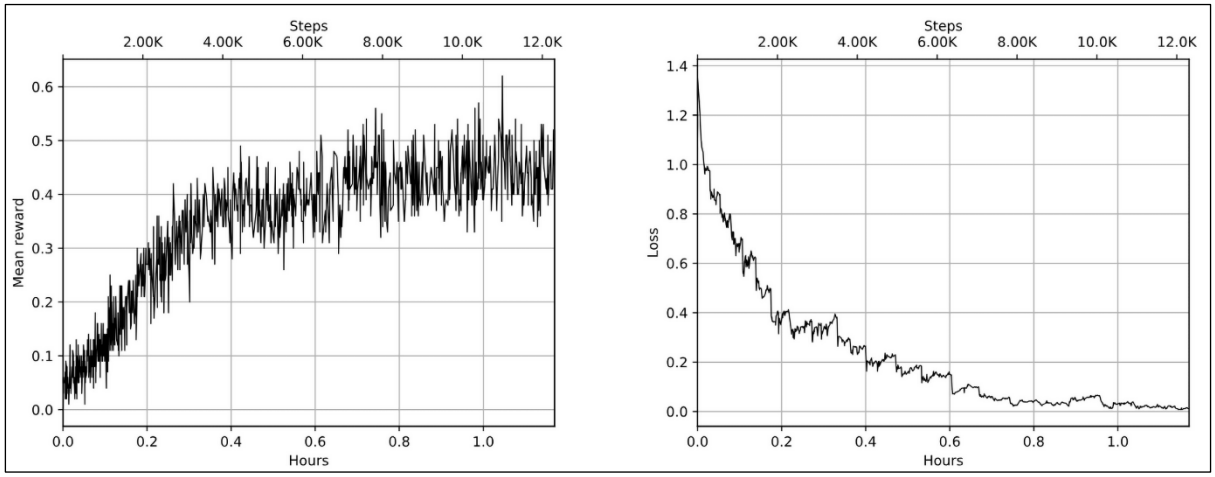

Frozen Lake

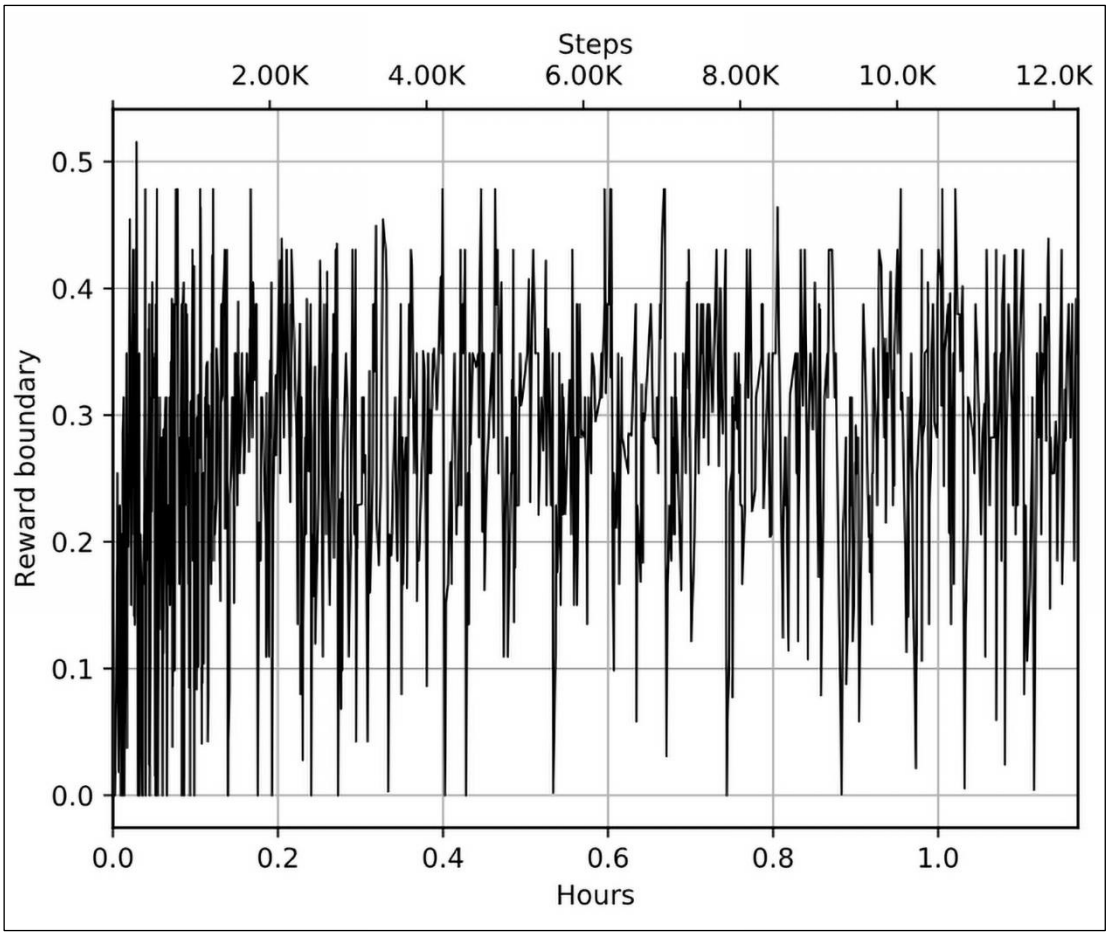

Steps
100
200
300
400
500
Mean reward
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
2
4
6
8
10
12
Seconds
Steps
100
200
300
400
500
Loss
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12
Seconds

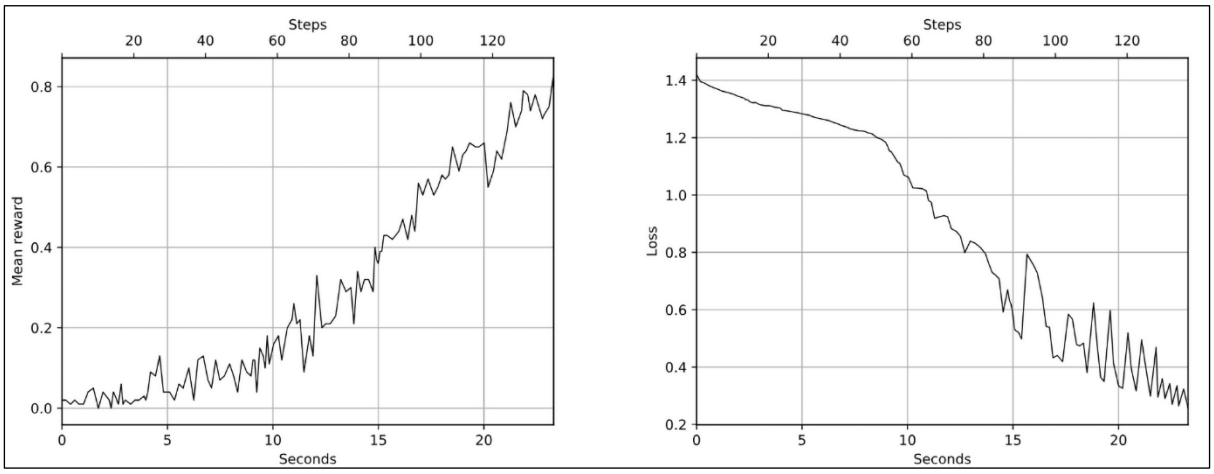

Steps
100
200
300
400
500
0.04
0.02
0.00
−0.02
−0.04
Reward boundary
0
2
4
6
8
10
12
Seconds

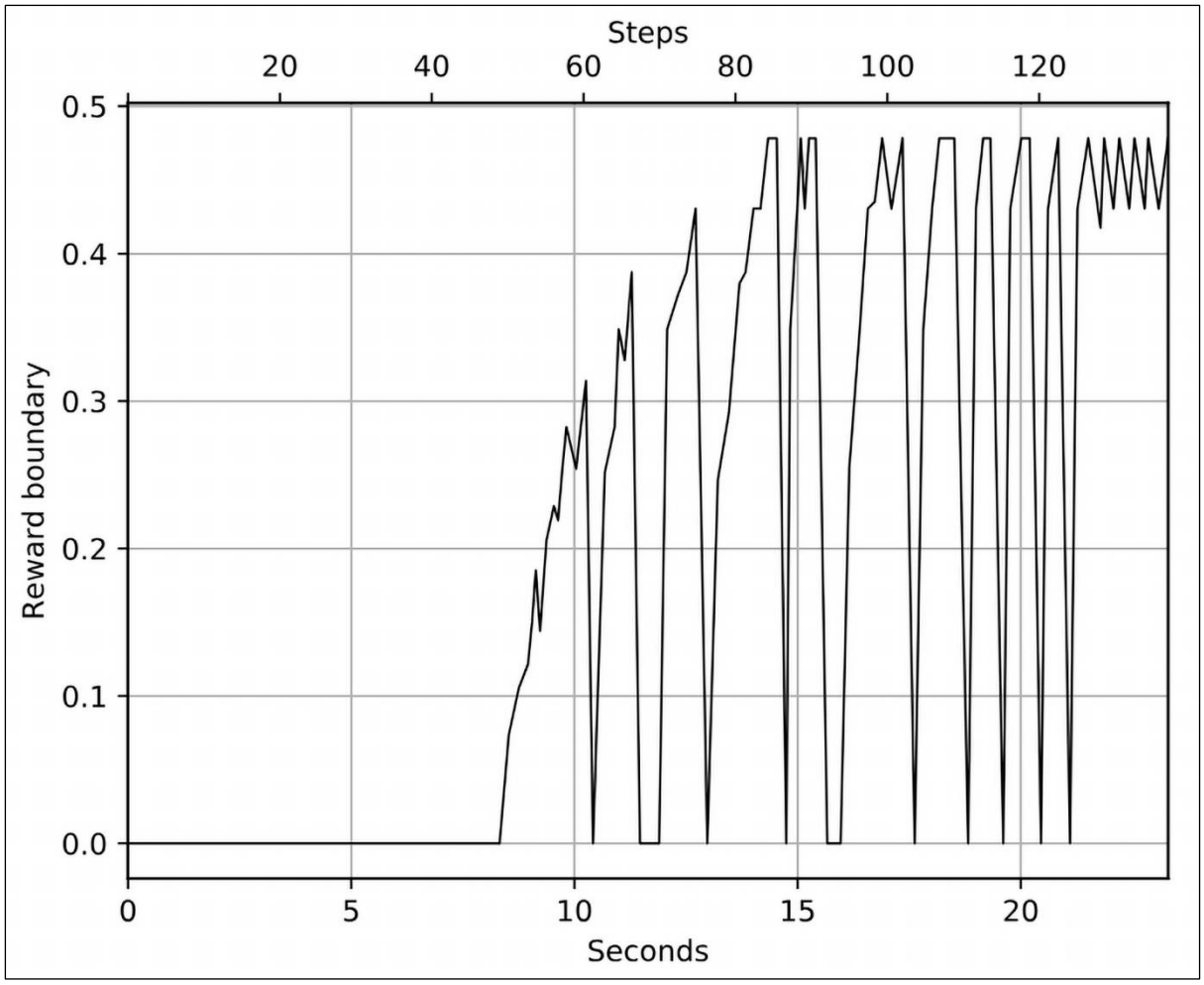

1 1 1 1 1 1 1
1 1 1 1
1 1
. . . . .
∑ = 7
∑ = 4
∑ = 2
elite episodes
freq
reward


0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0 0
0 0 0 1
. . . . . .
∑ = 1
∑ = 0
∑ = 1
freq
70% percentile=0
0
1
reward

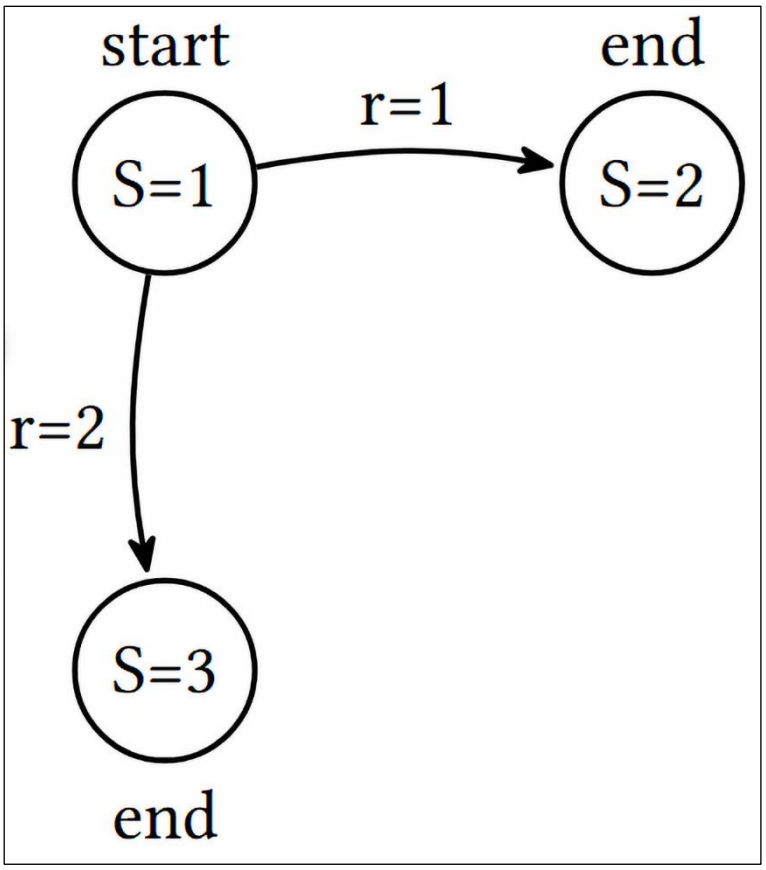

Steps
2.00K
4.00K
6.00K
8.00K
10.0K
12.0K
Mean reward
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Hours
0.0
0.2
0.4
0.6
0.8
1.0
Steps
2.00K
4.00K
6.00K
8.00K
10.0K
12.0K
Loss
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Hours
0.0
0.2
0.4
0.6
0.8
1.0

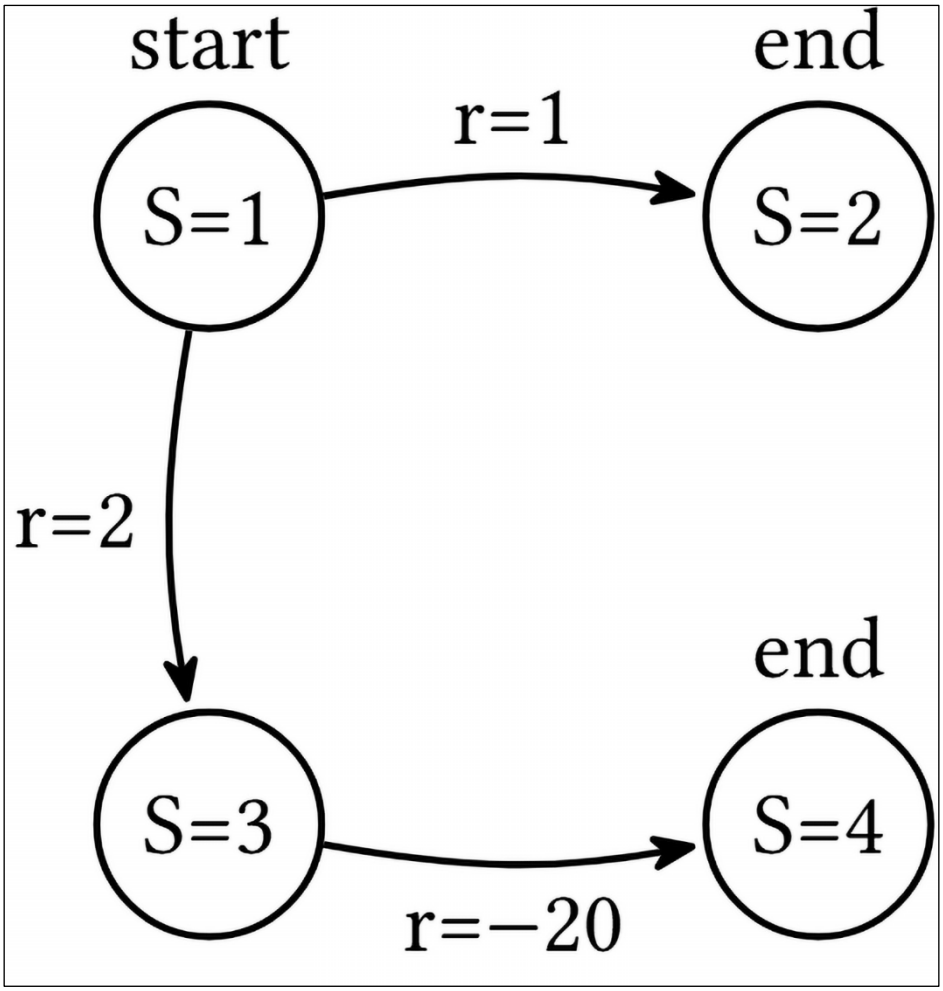

Steps
2.00K
4.00K
6.00K
8.00K
10.0K
12.0K
Reward boundary
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Hours

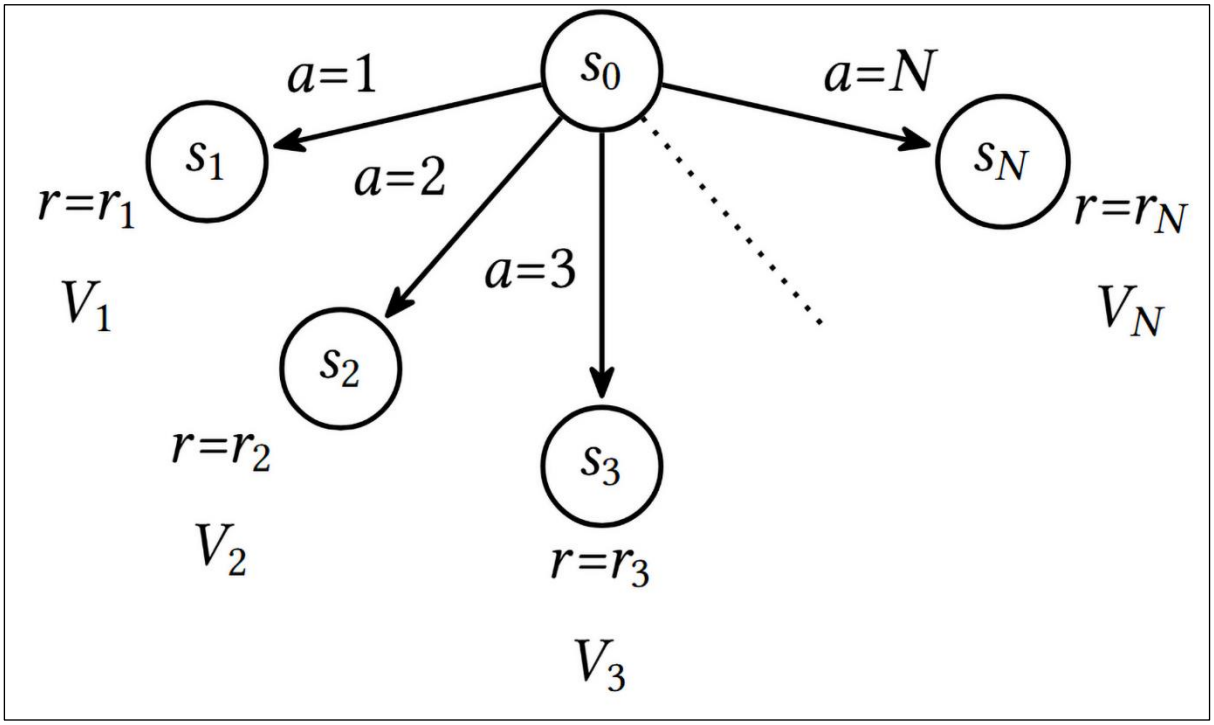

Steps
20
40
60
80
100
120
Mean reward
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
Seconds
Steps
20
40
60
80
100
120
Loss
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0
5
10
15
20
Seconds

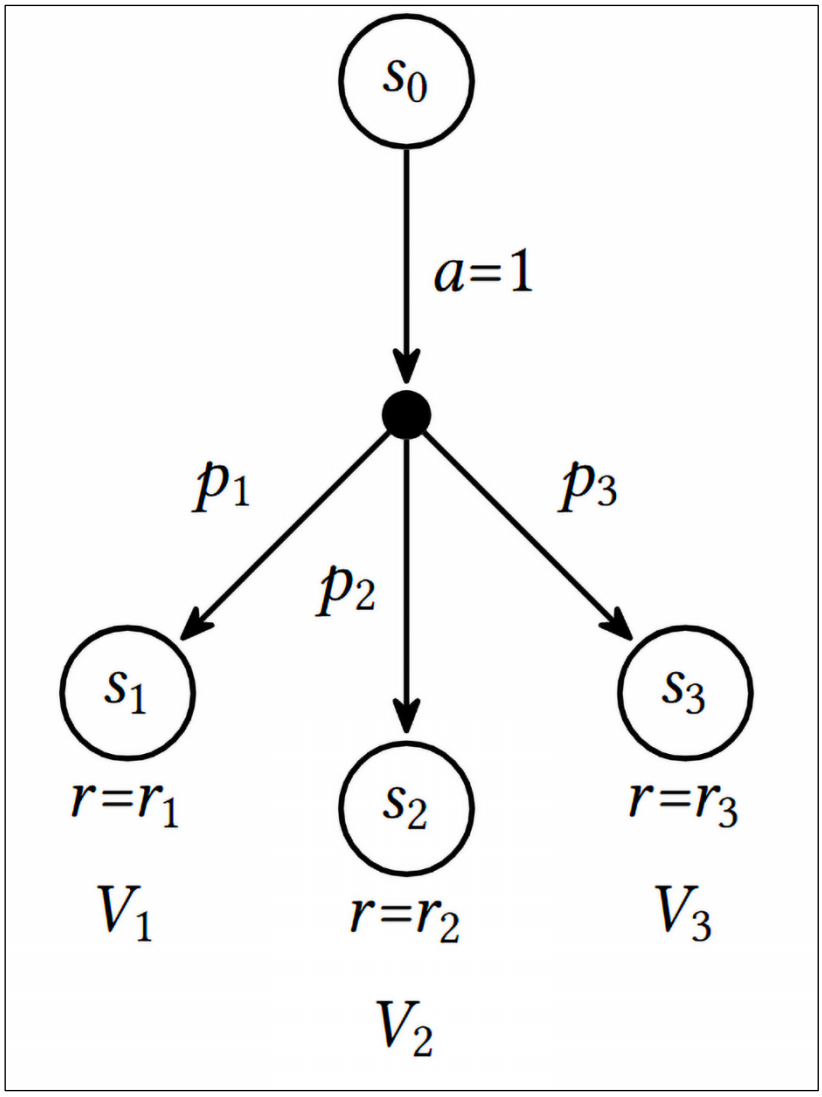

Steps
20
40
60
80
100
120
0.5
0.4
0.3
0.2
0.1
0.0
Reward boundary
0
5
10
15
20
Seconds

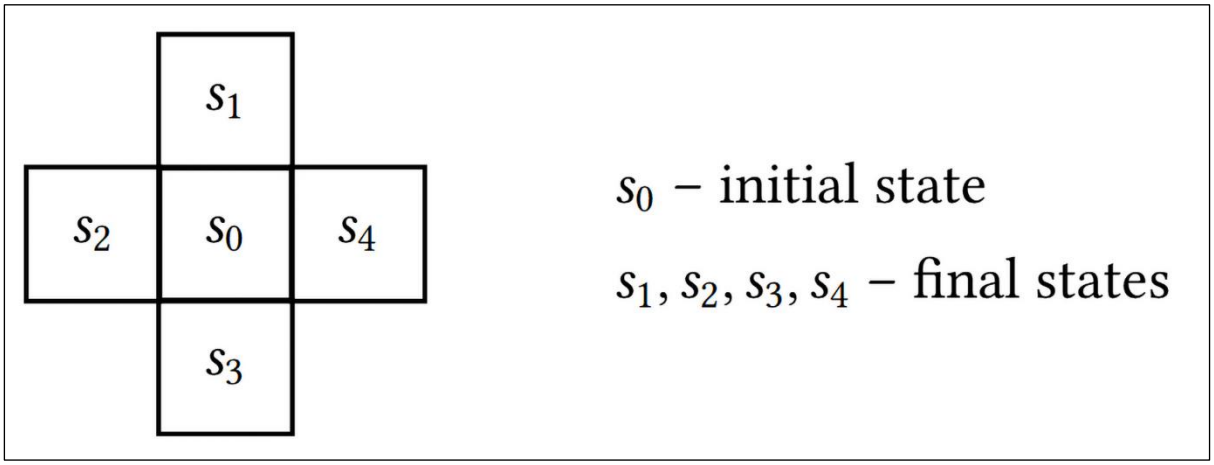
Chapter 5: Tabular Learning and the Bellman Equation

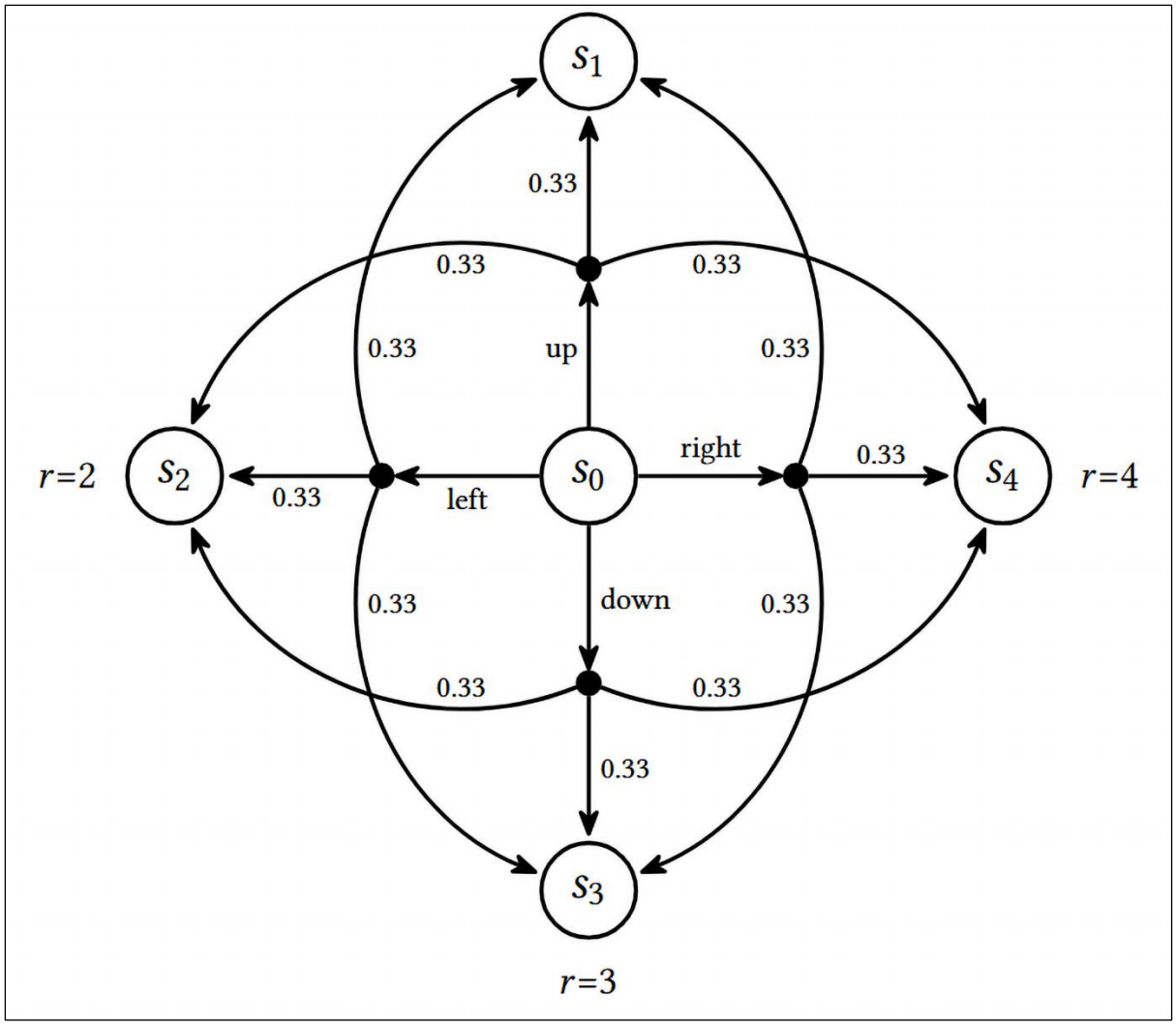


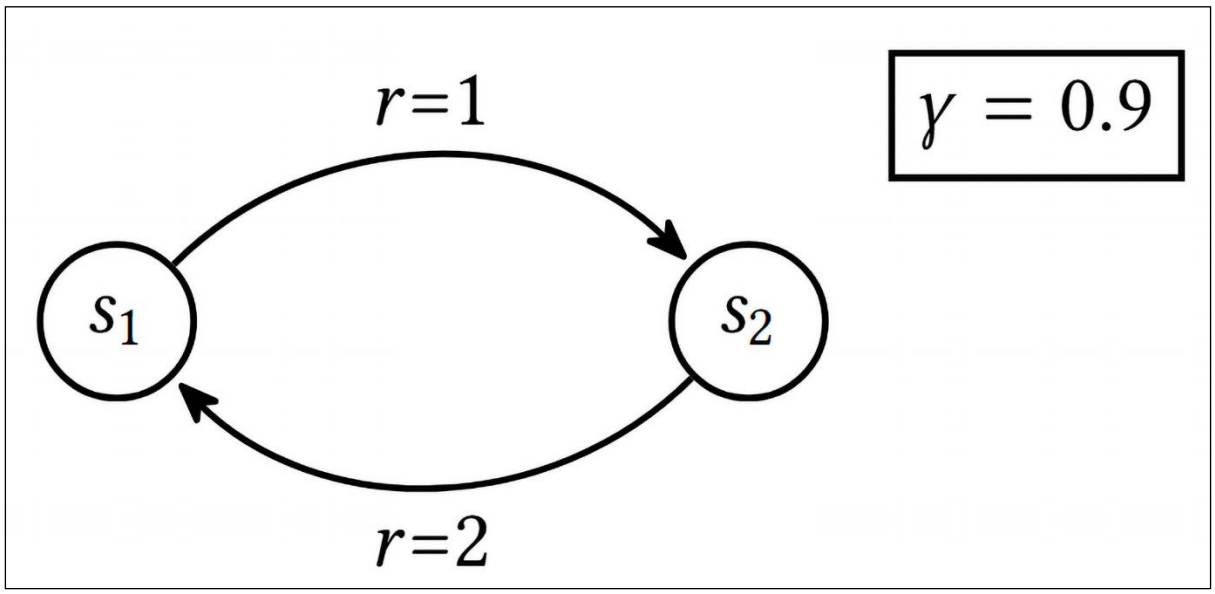

start
end
r=1
S=1
S=2
r=2
end
S=3
S=4
r=−20

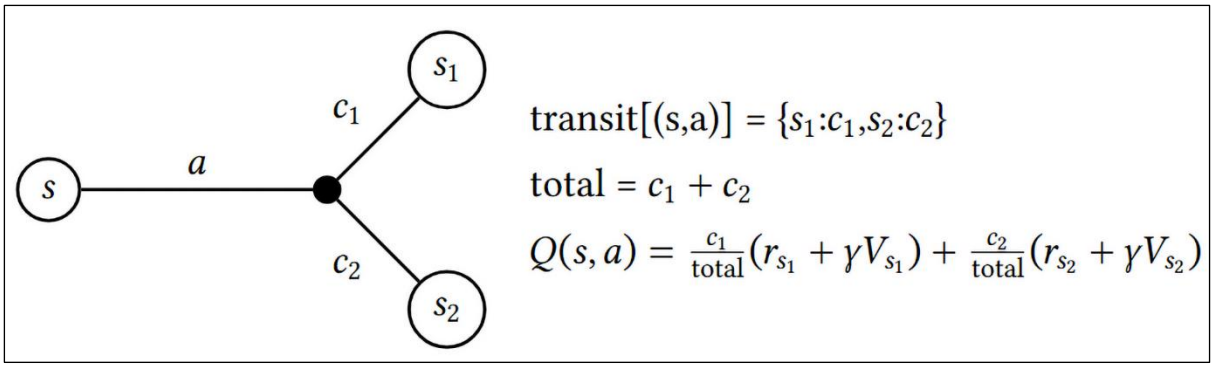

s_0
$a=1$
$a=N$
s_1
$a=2$
s_N
$r=r_1$
$r=r_N$
$a=3$
V_1
V_N
s_2
$r=r_2$
s_3
V_2
$r=r_3$
V_3

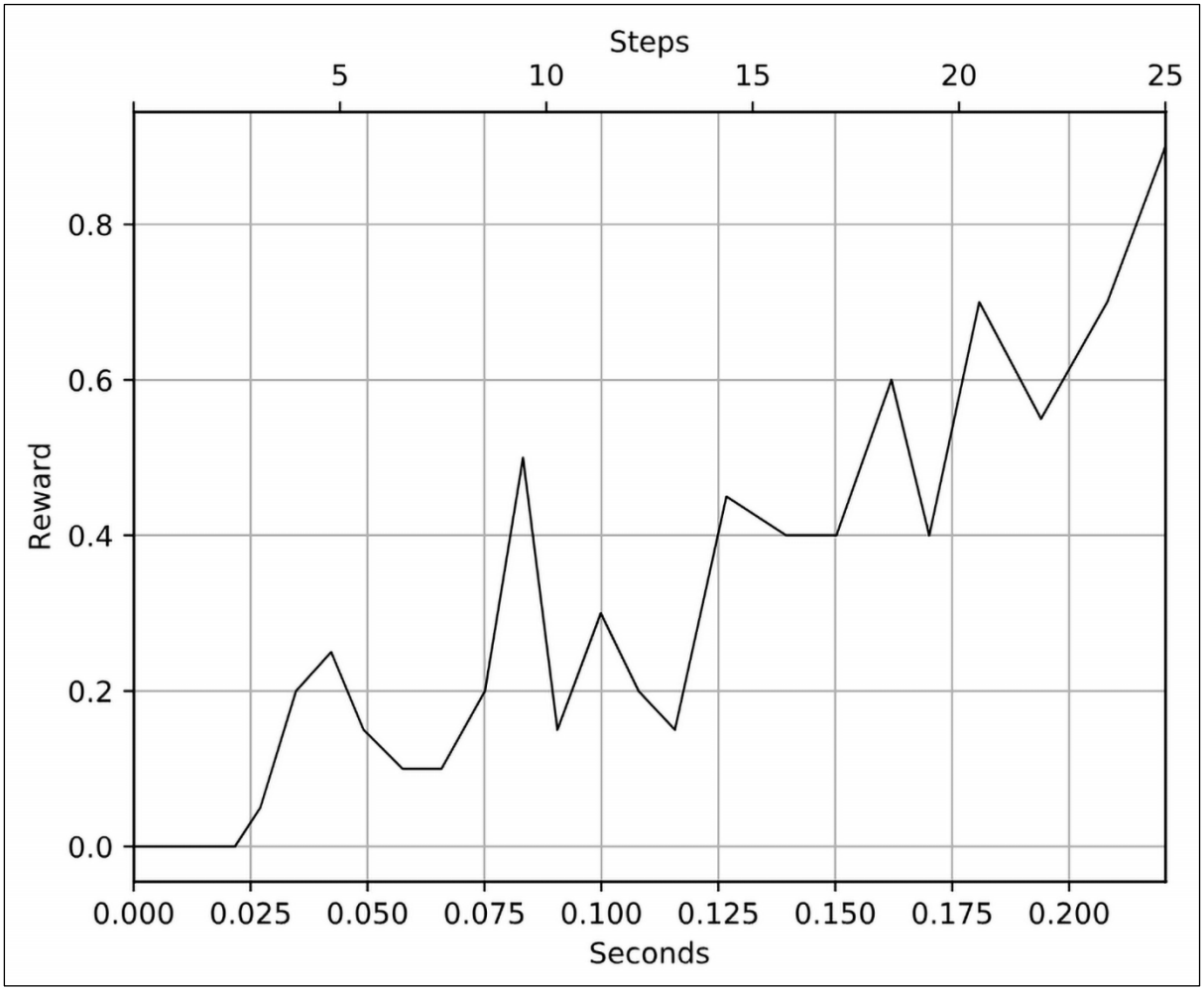

s_0
$a=1$
p_1
p_2
p_3
s_1
s_2
s_3
$r=r_1$
$r=r_2$
$r=r_3$
V_1
V_2
V_3

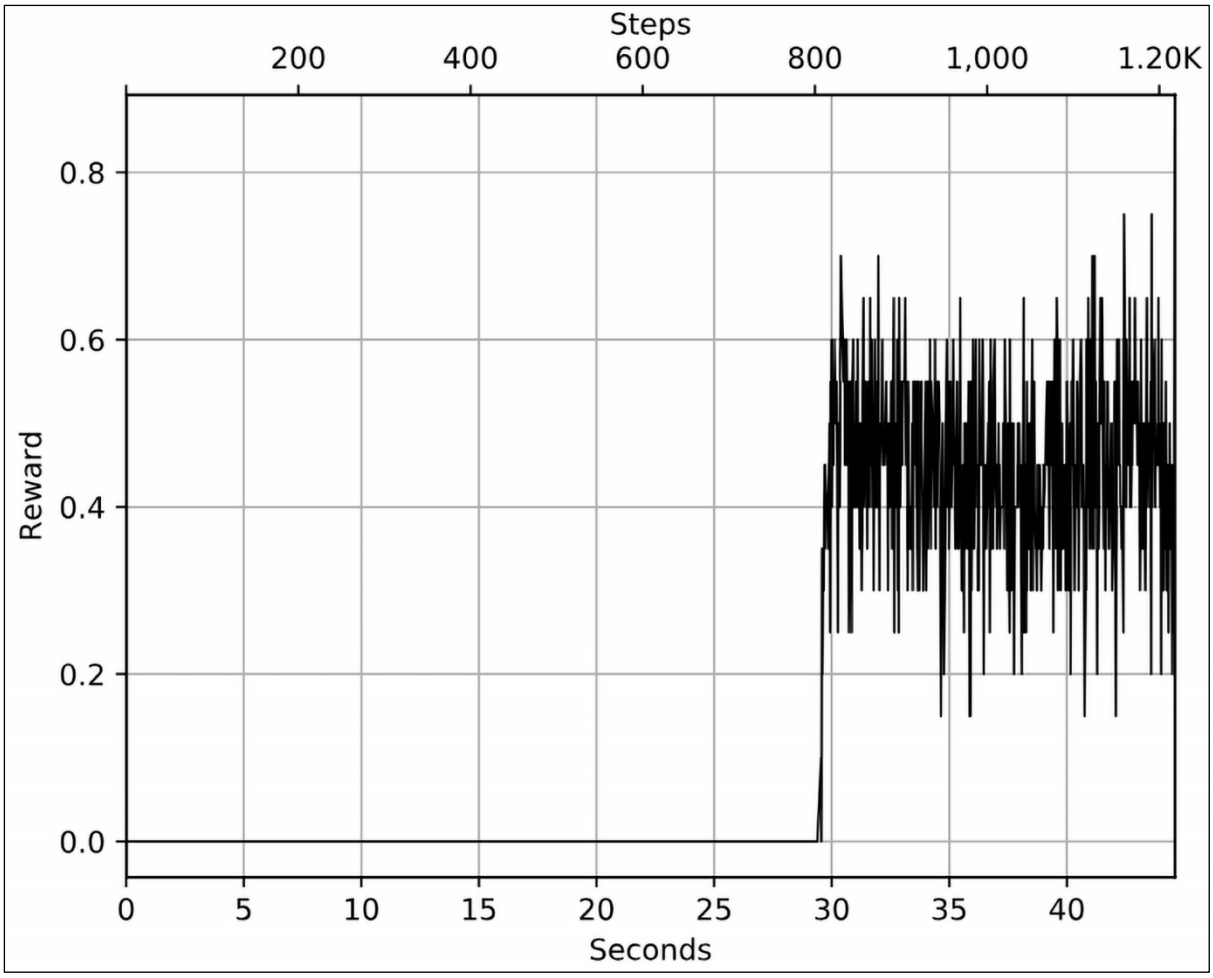

s_1
s_2
s_0
s_4
s_3
s_0 – initial state
s_1, s_2, s_3, s_4 – final states


s_1
0.33
0.33
0.33
0.33
0.33
up
r=2
s_2
0.33
left
s_0
right
0.33
s_4
r=4
0.33
down
0.33
0.33
0.33
0.33
s_3
r=3

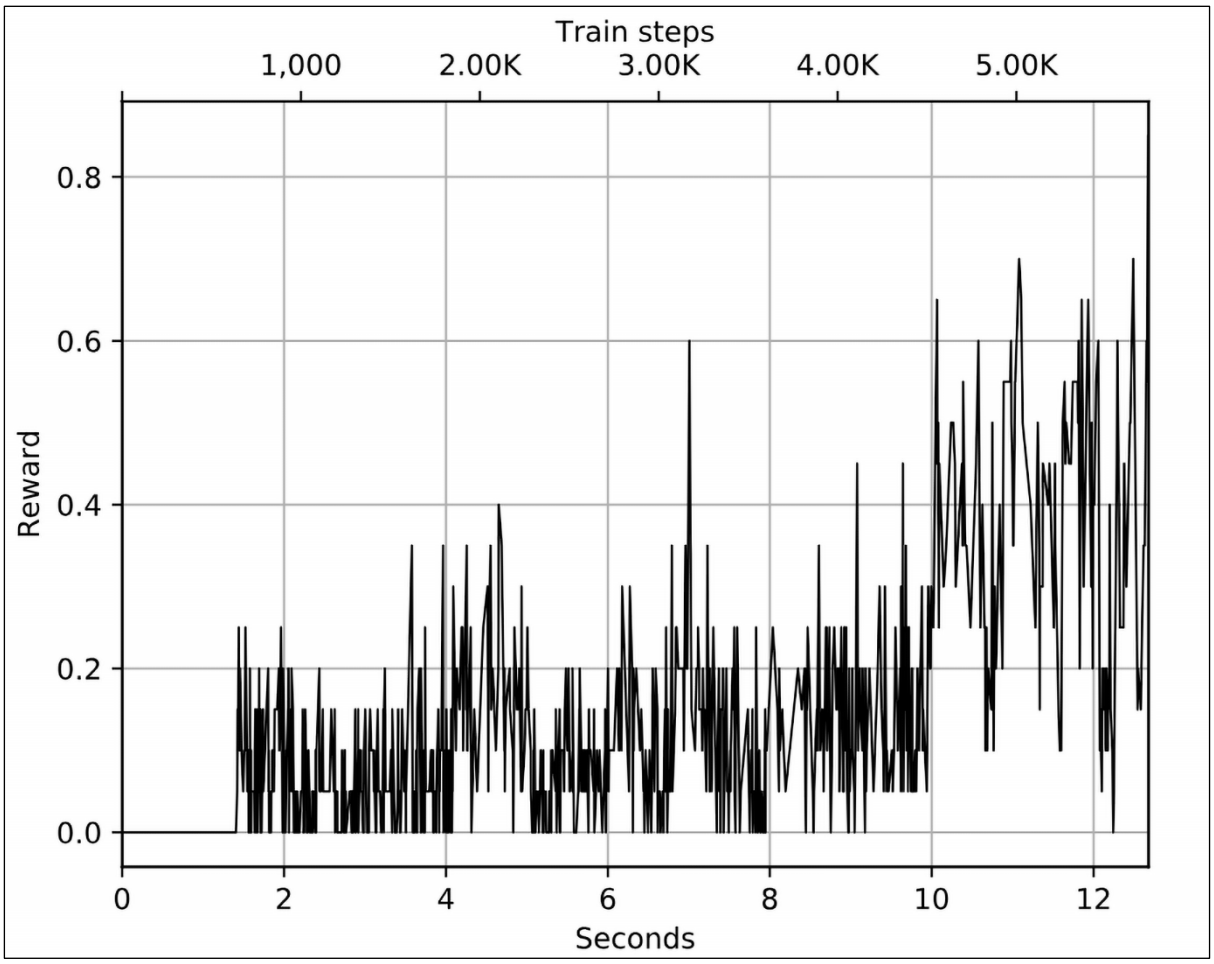

r=1
$\gamma = 0.9$
s_1
s_2
r=2

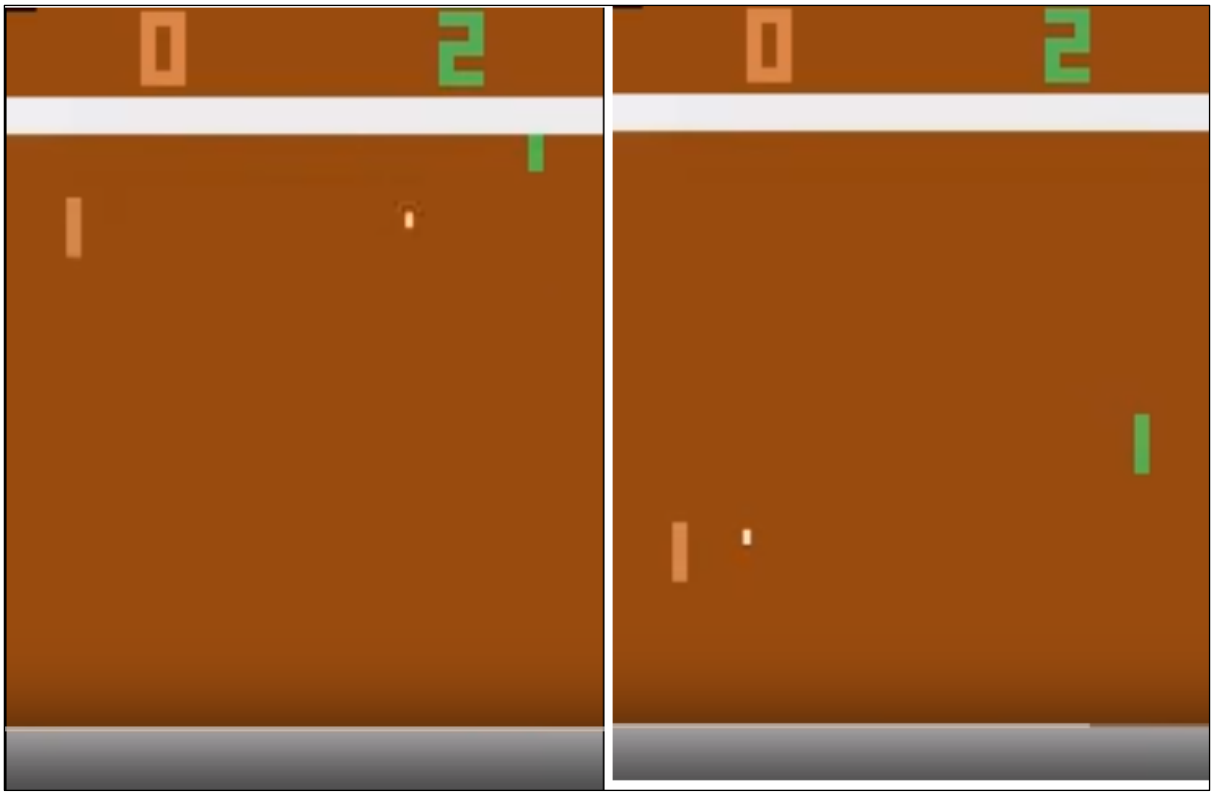

s_1
c_1
s
a
c_2
s_2
transit[(s,a)] = $\{s_1{:}c_1, s_2{:}c_2\}$
total = $c_1 + c_2$
$Q(s,a) = \frac{c_1}{\text{total}}(r_{s_1} + \gamma V_{s_1}) + \frac{c_2}{\text{total}}(r_{s_2} + \gamma V_{s_2})$

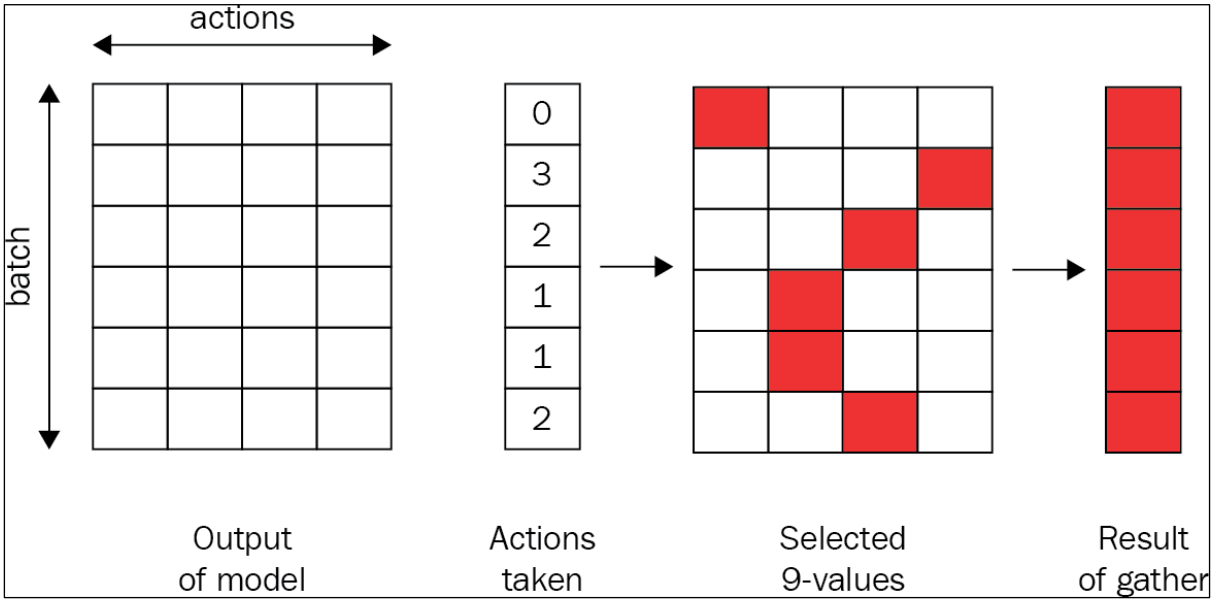

Steps
5
10
15
20
25
0.8
0.6
0.4
0.2
0.0
Reward
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Seconds

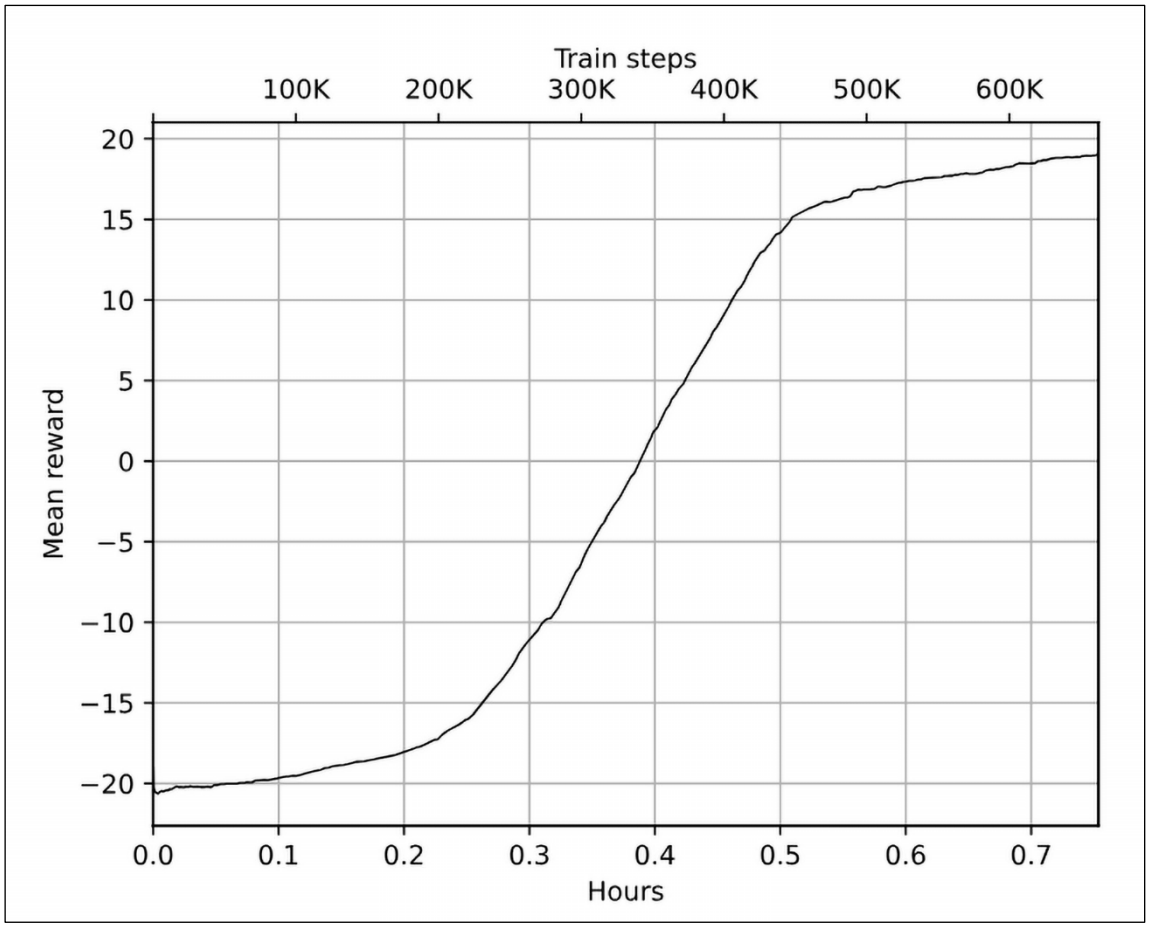

Steps
200
400
600
800
1,000
1.20K
Reward
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
35
40
Seconds

Chapter 6: Deep Q-Networks

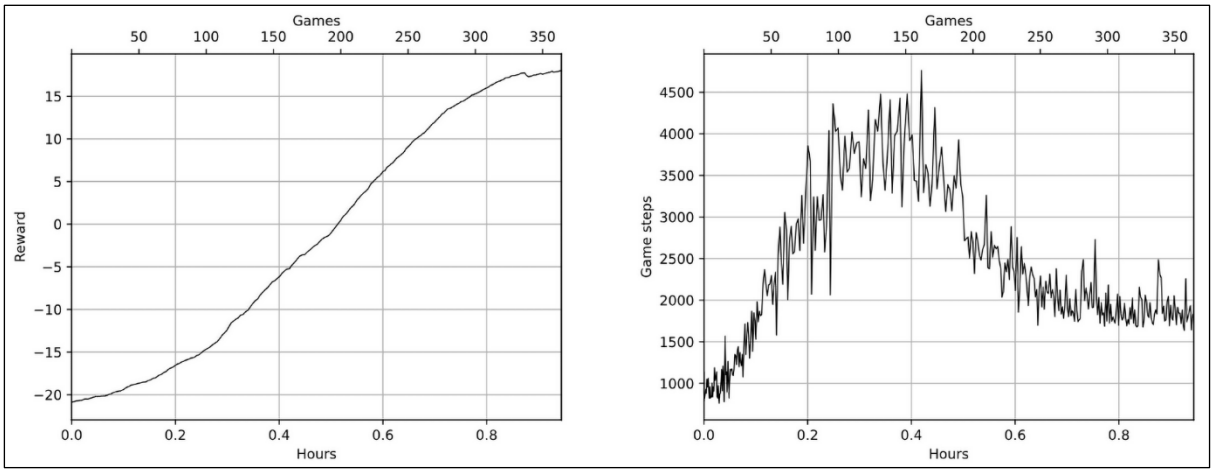


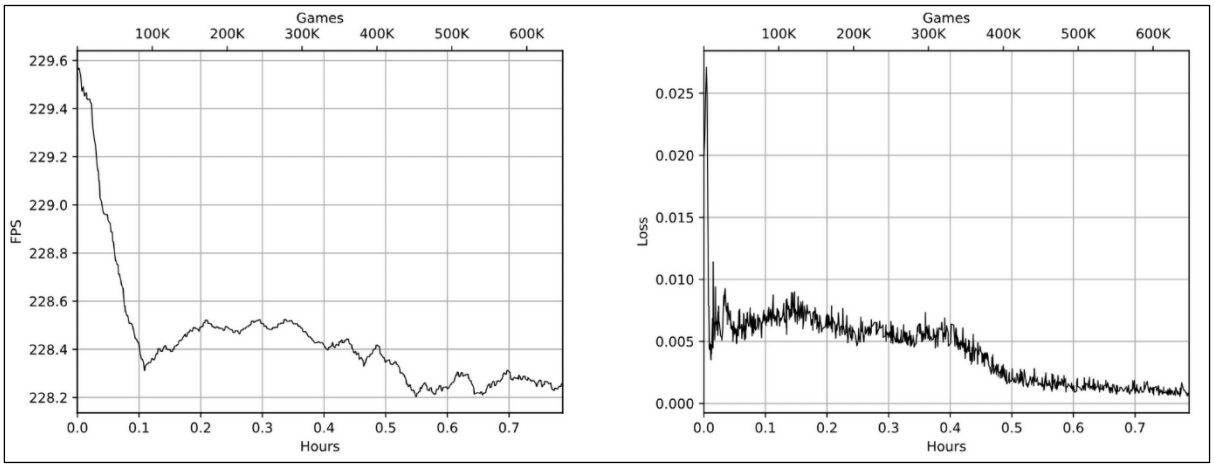


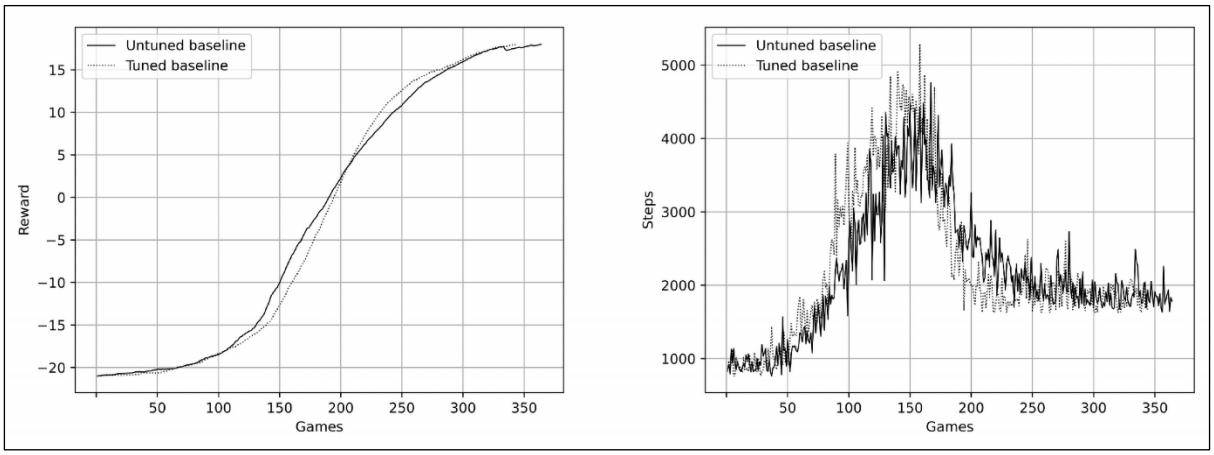

actions
batch
0
3
2
1
1
2
Output of model
Actions taken
Selected 9-values
Result of gather

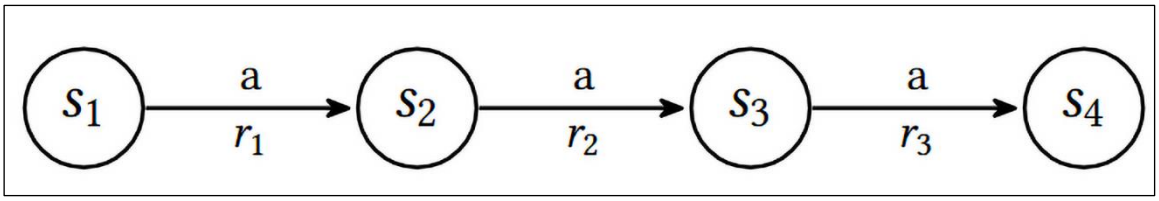

Train steps
100K
200K
300K
400K
500K
600K
20
15
10
5
0
−5
−10
−15
−20
Mean reward
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Hours

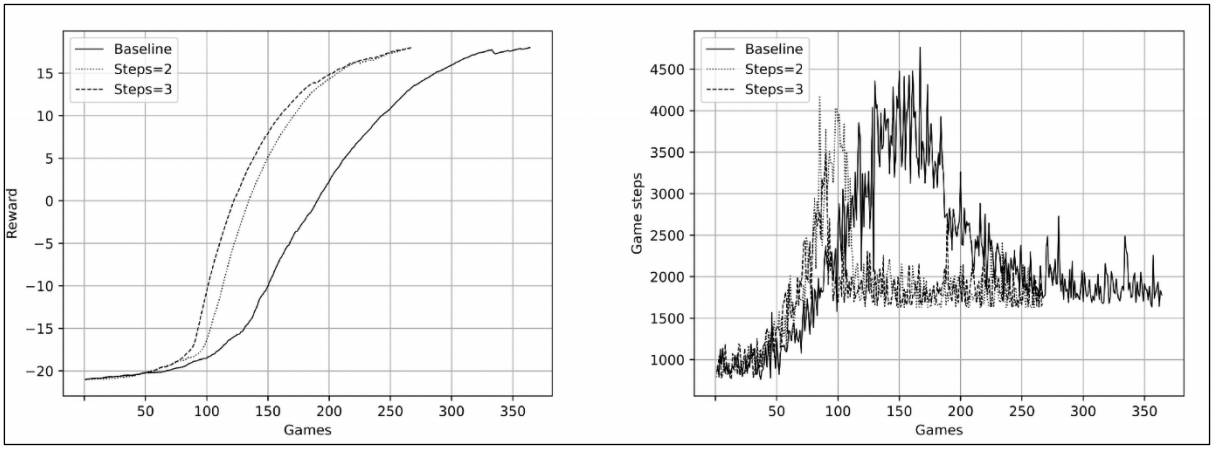
Chapter 8: DQN Extensions

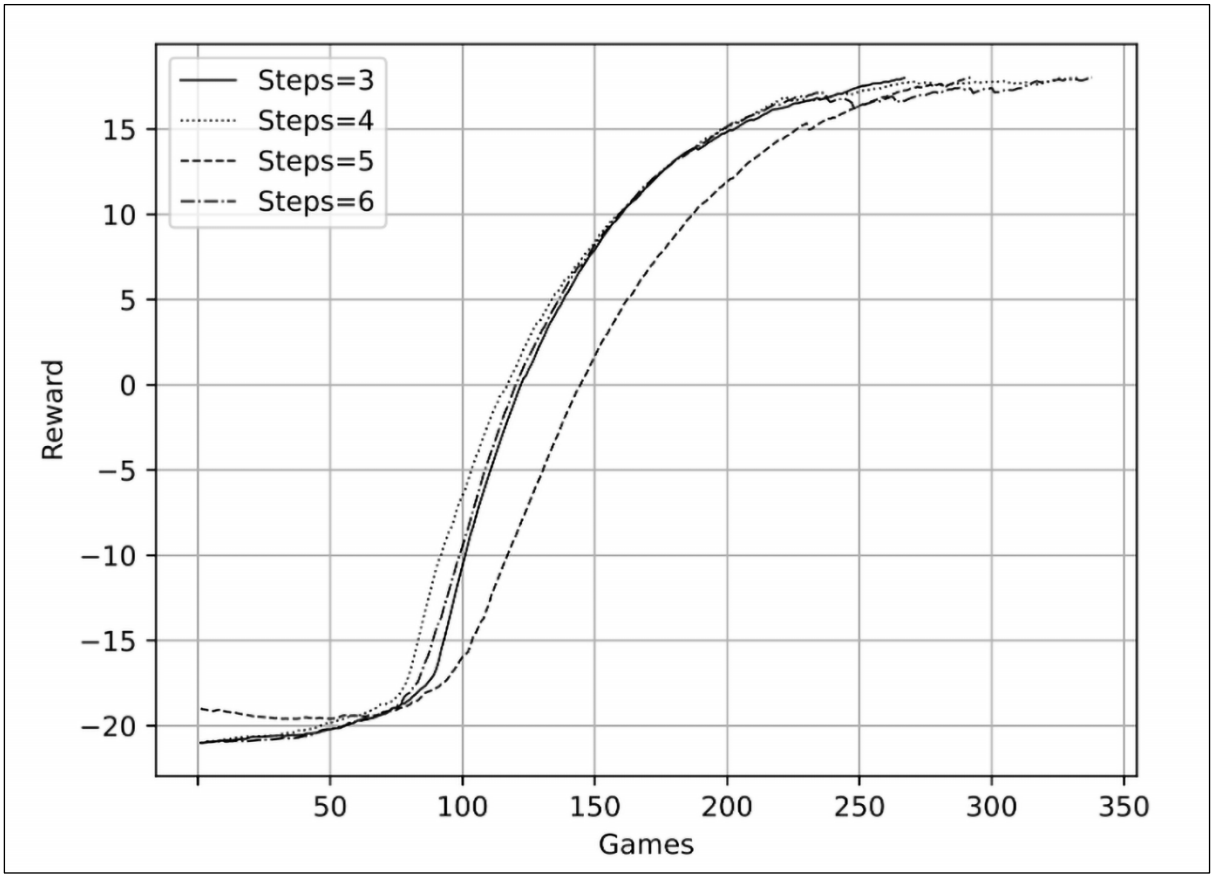


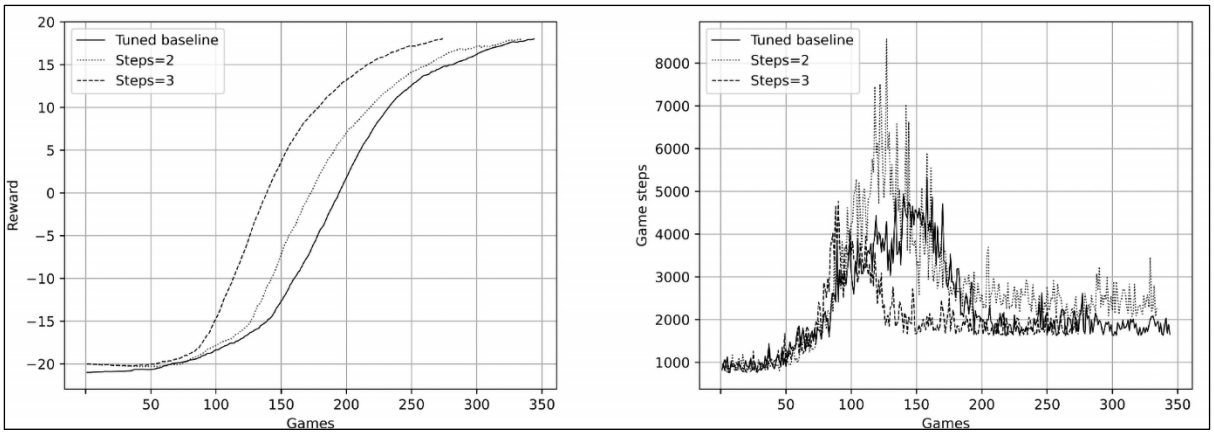


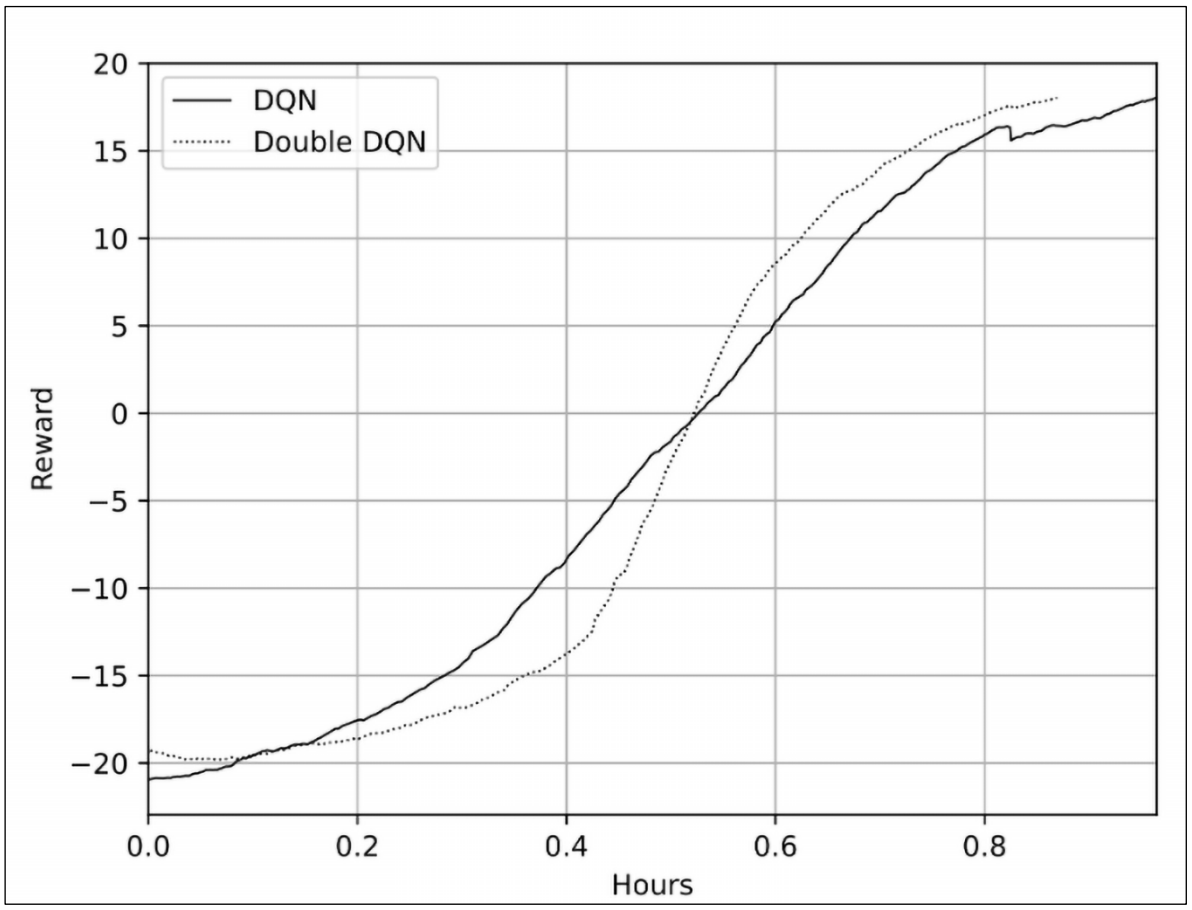


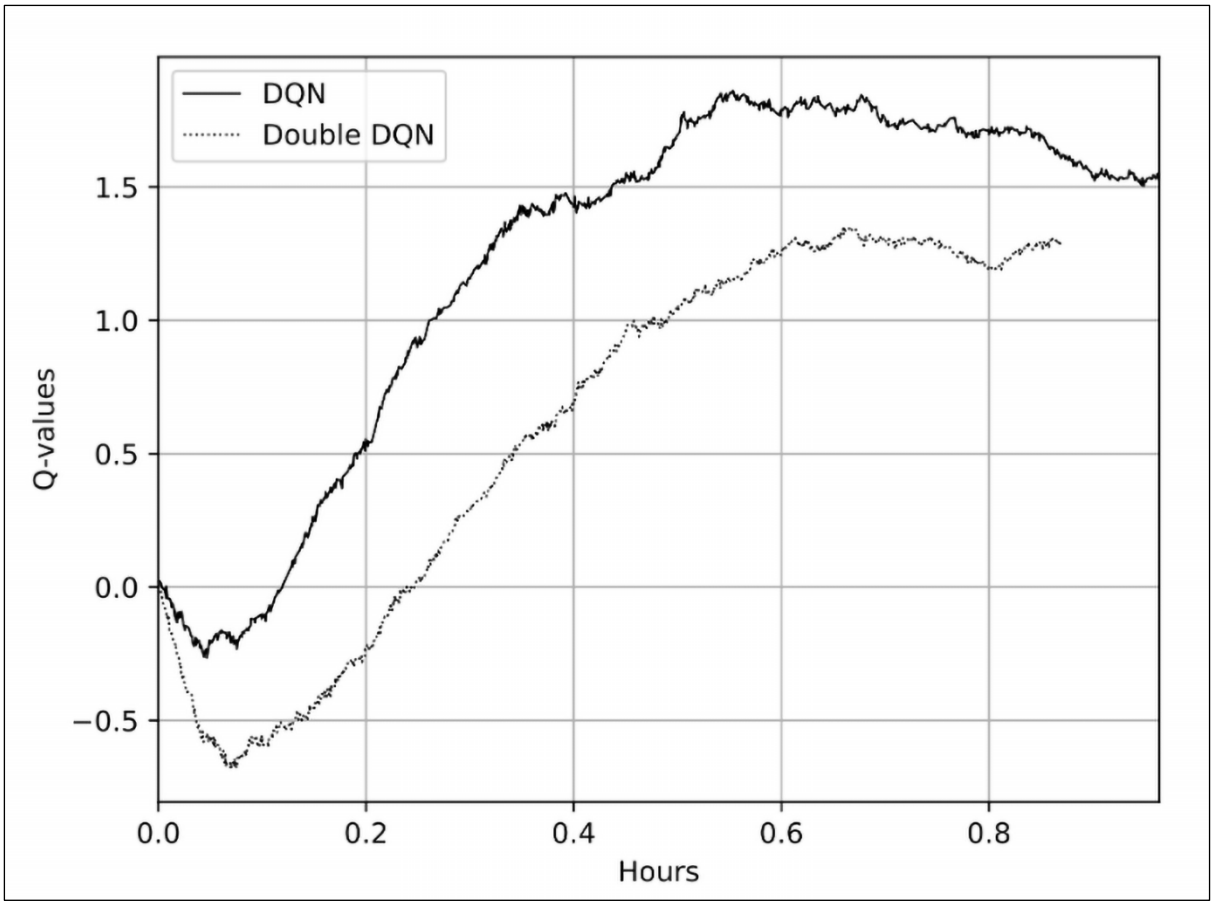

s_1
a
r_1
s_2
a
r_2
s_3
a
r_3
s_4

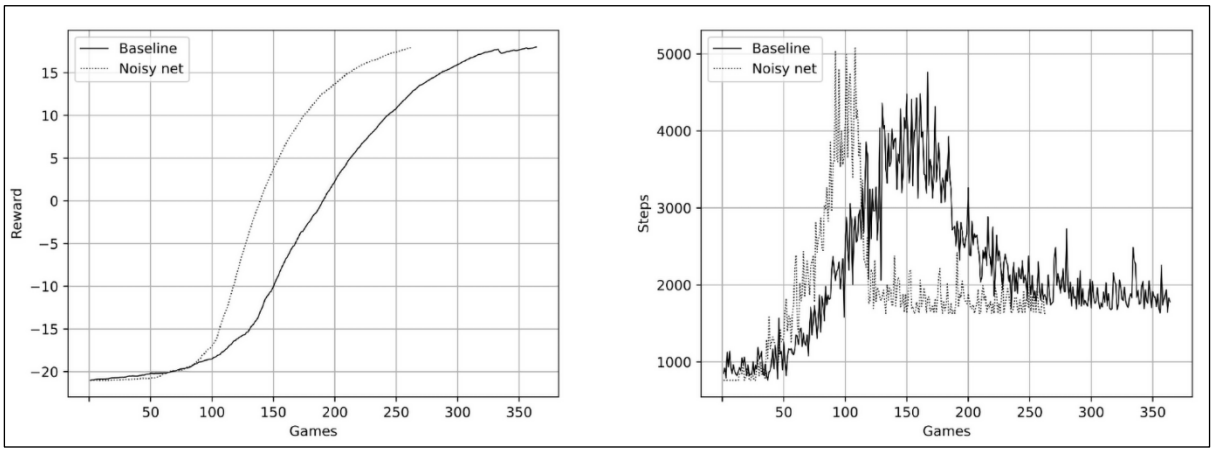

Baseline
Steps=2
Steps=3
Reward
Games
Baseline
Steps=2
Steps=3
Game steps
Games

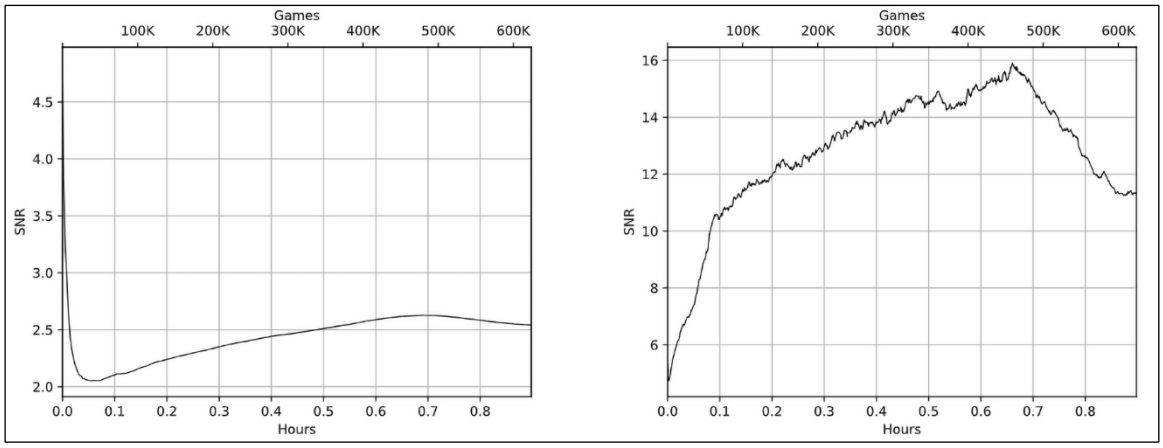

Steps=3
Steps=4
Steps=5
Steps=6
Reward
Games

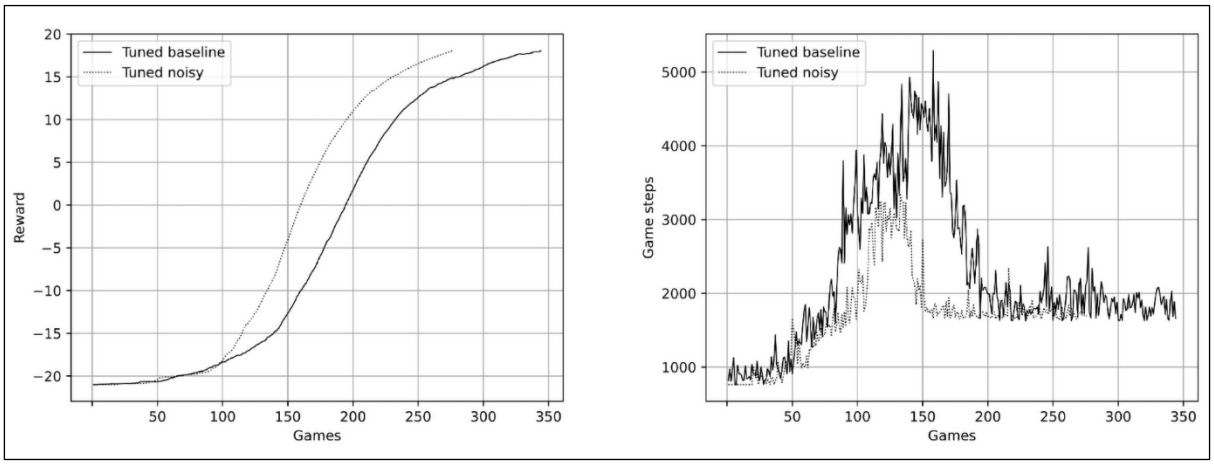

Tuned baseline
Steps=2
Steps=3
Reward
Games
Tuned baseline
Steps=2
Steps=3
Game steps
Games

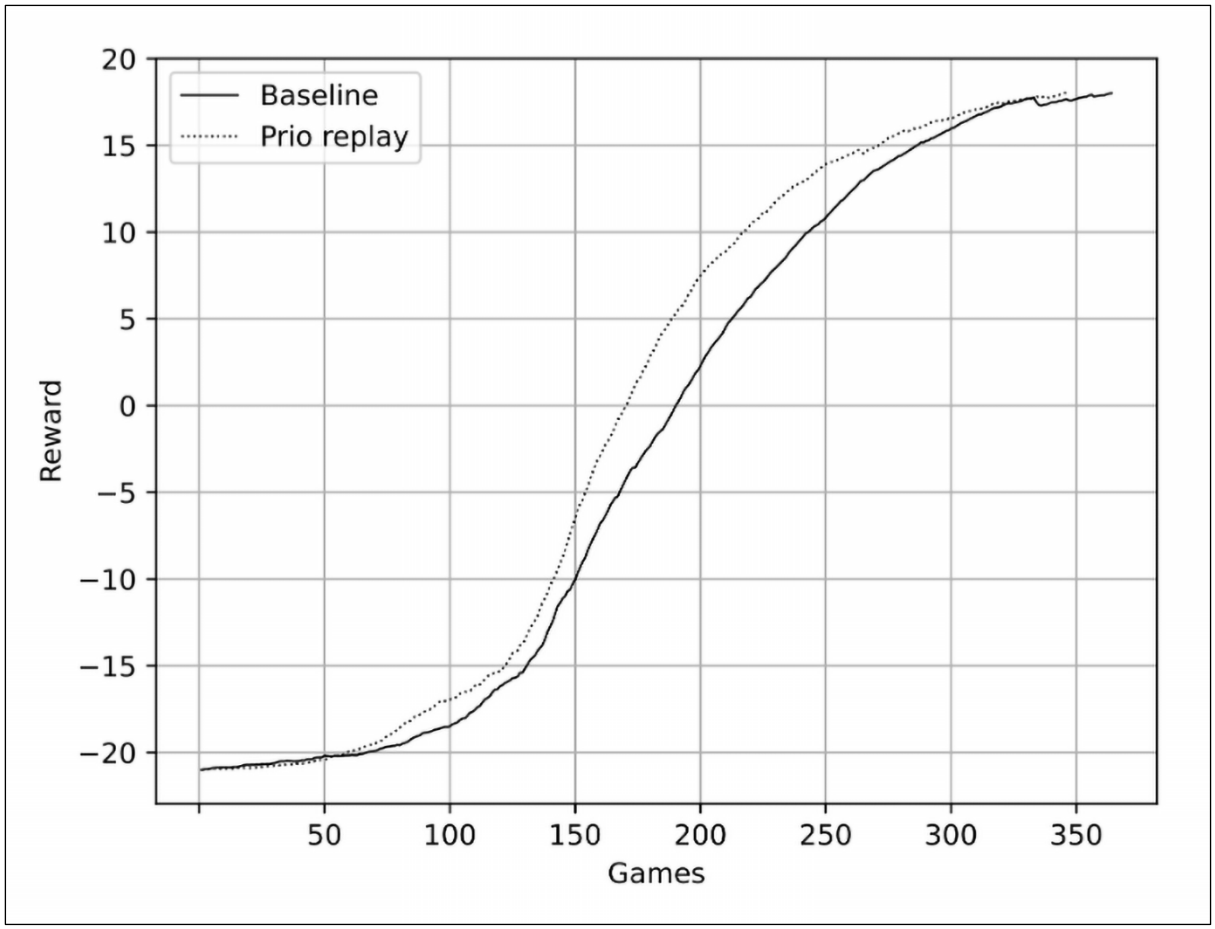

DQN
Double DQN
Reward
Hours

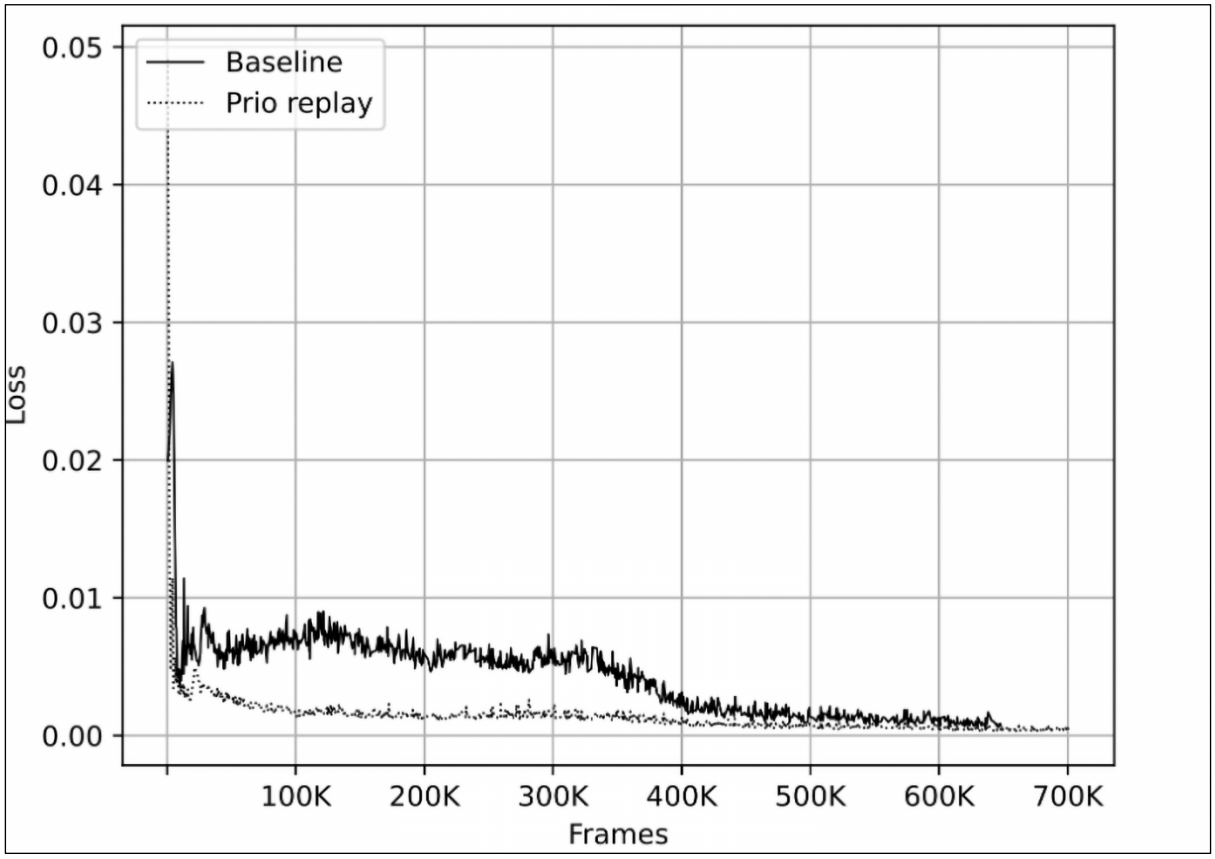

DQN
Double DQN
Q-values
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
Hours

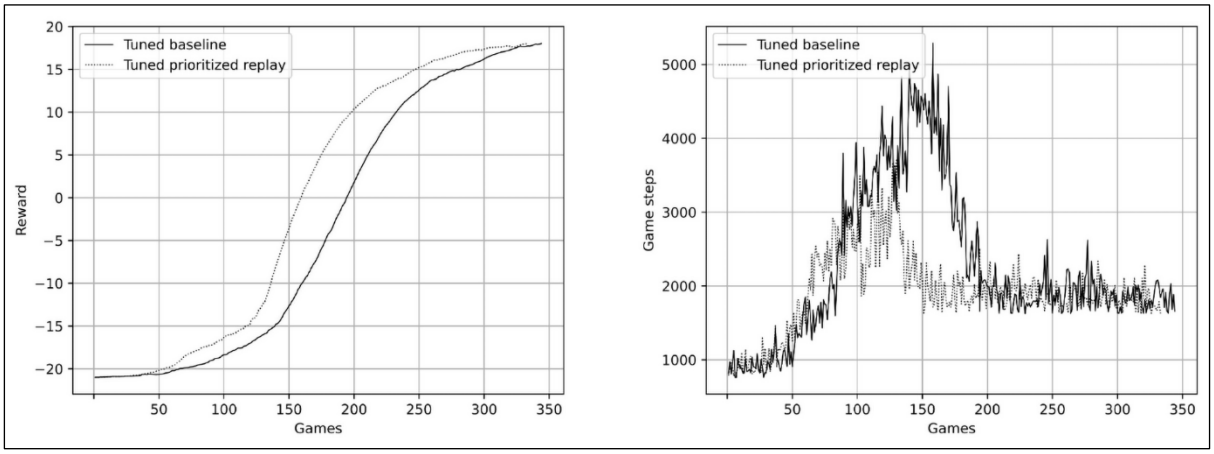

Baseline
Noisy net
Reward
15
10
5
0
−5
−10
−15
−20
50
100
150
200
250
300
350
Games
Baseline
Noisy net
Steps
5000
4000
3000
2000
1000
50
100
150
200
250
300
350
Games

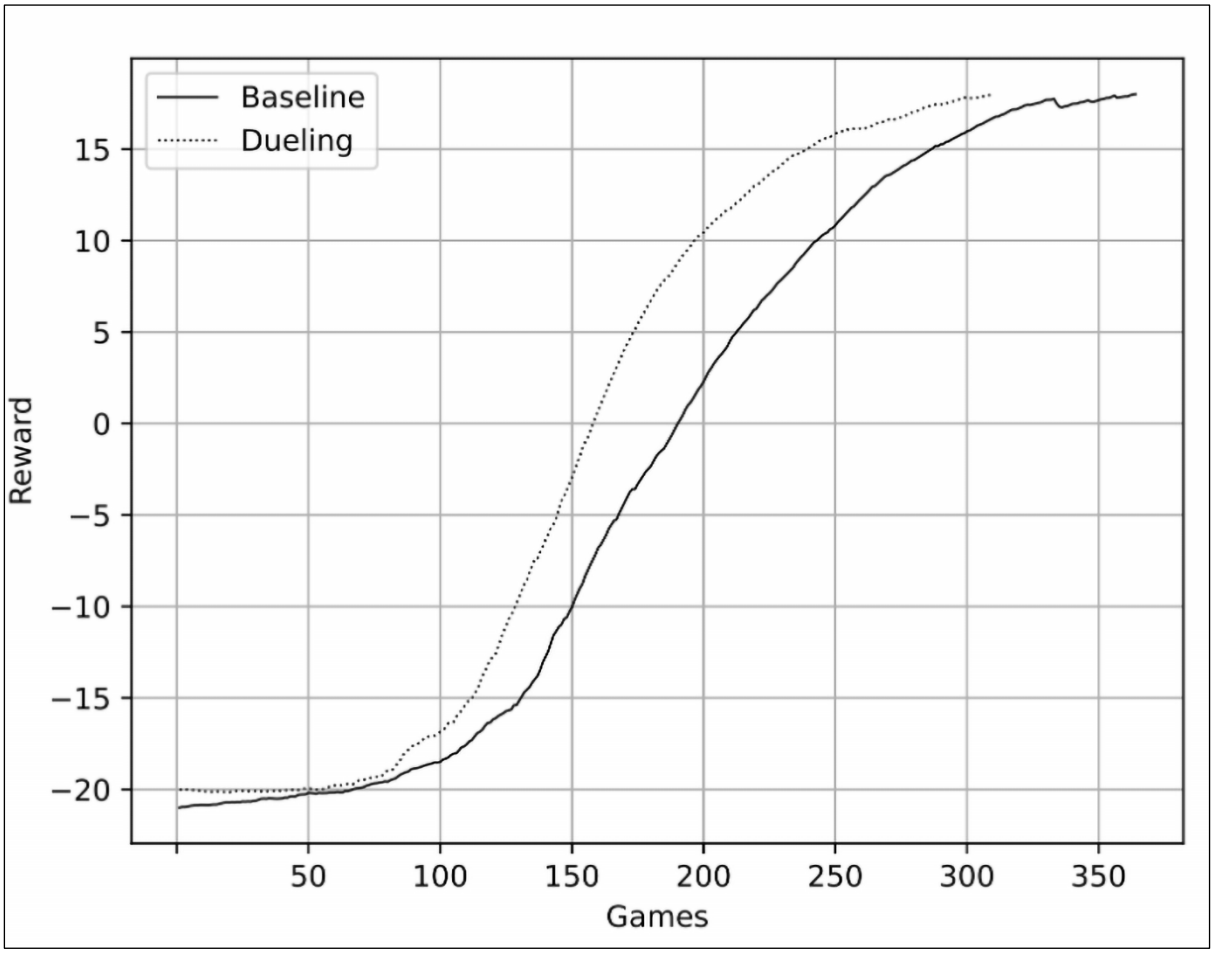

Games
100K
200K
300K
400K
500K
600K
SNR
Hours
Games
100K
200K
300K
400K
500K
600K
SNR
Hours

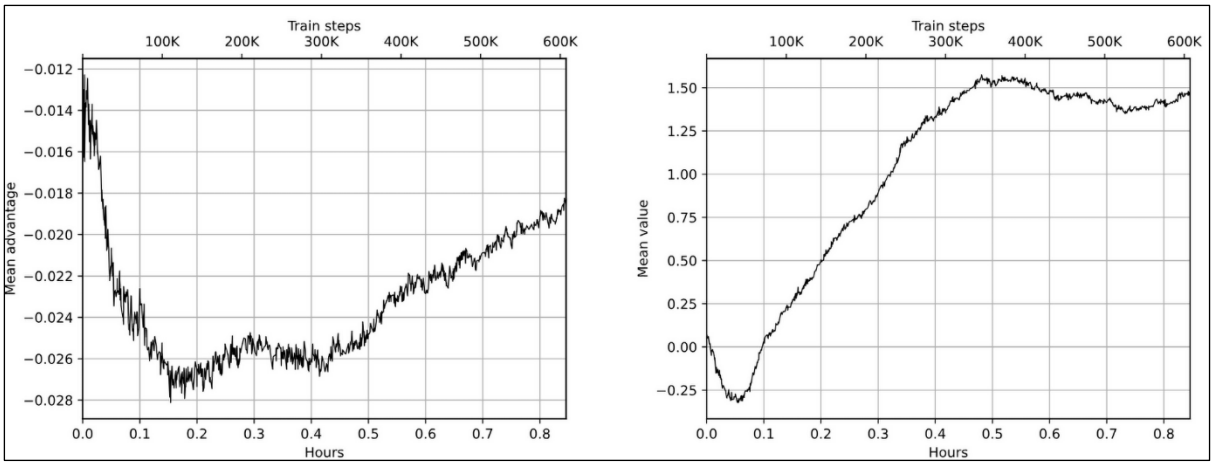

Tuned baseline
Tuned noisy
Reward
Games
Tuned baseline
Tuned noisy
Game steps
Games

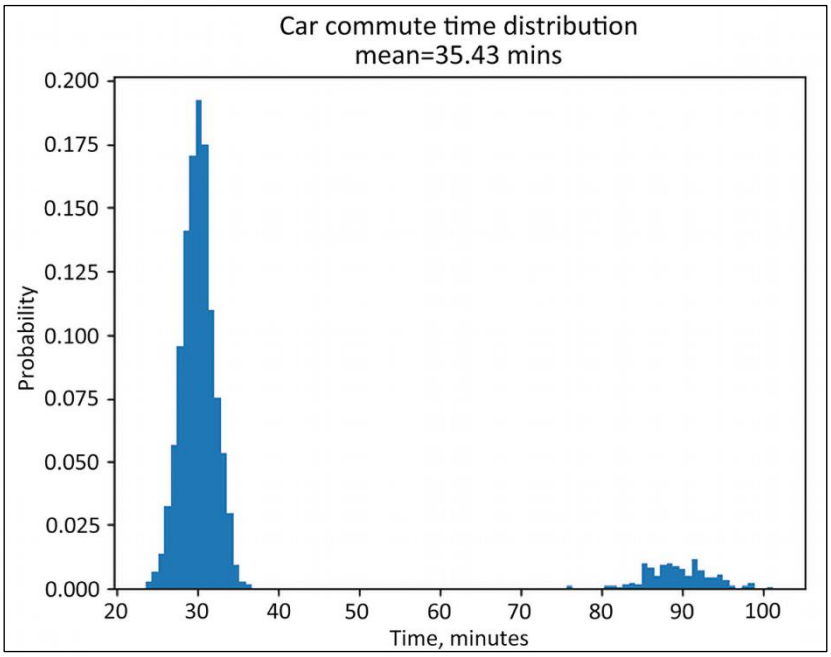

Baseline
Prio replay
Reward
Games
20
15
10
5
0
−5
−10
−15
−20
50
100
150
200
250
300
350

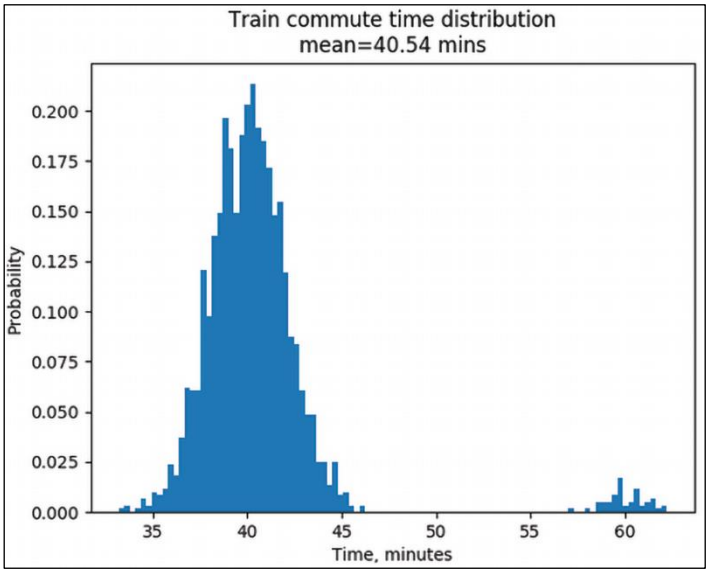

Baseline
Prio replay
Loss
Frames
0.05
0.04
0.03
0.02
0.01
0.00
100K
200K
300K
400K
500K
600K
700K

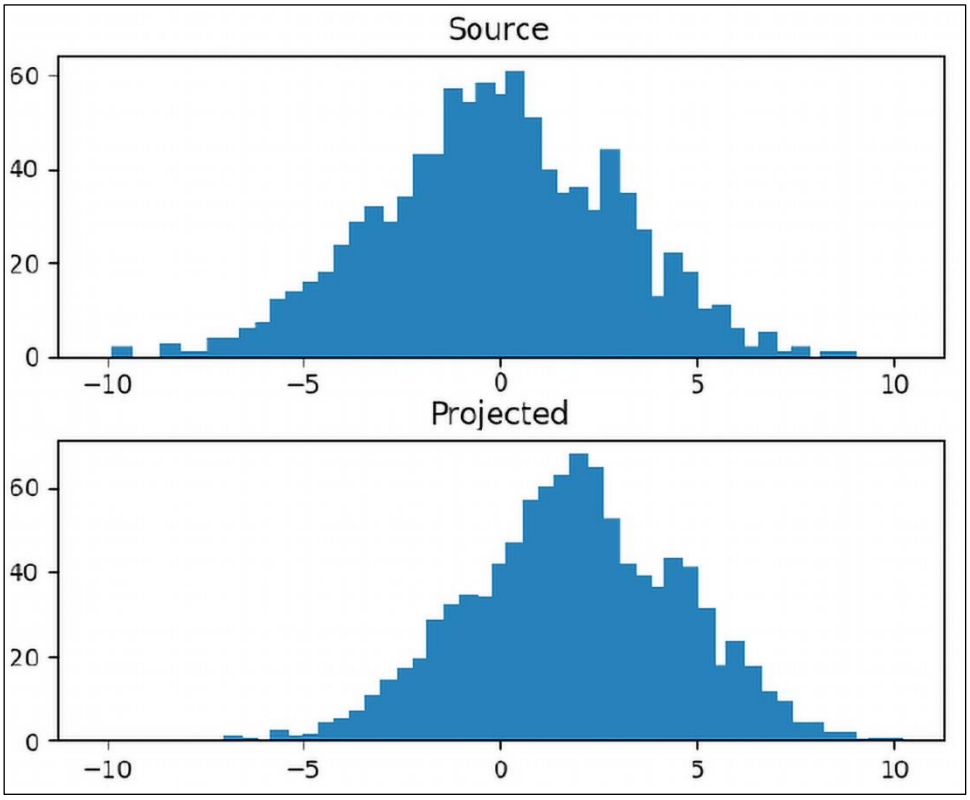

Tuned baseline
Tuned prioritized replay
Reward
Games
Tuned baseline
Tuned prioritized replay
Game steps
Games

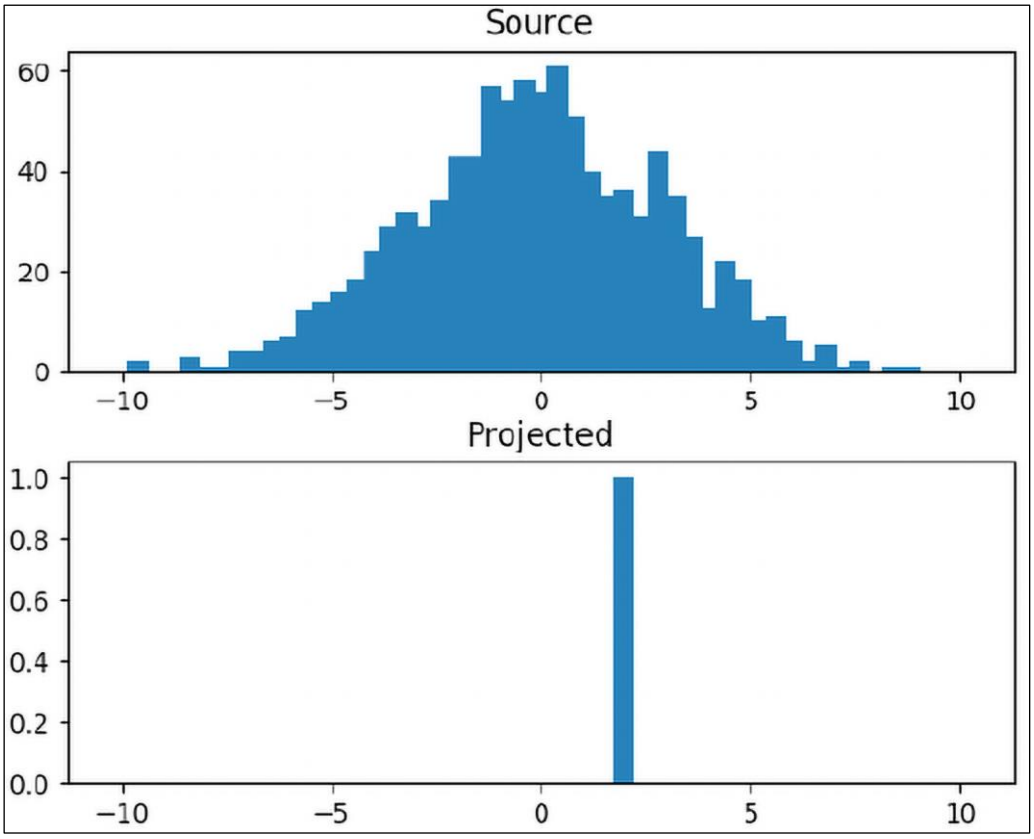

Baseline
Dueling
Reward
Games

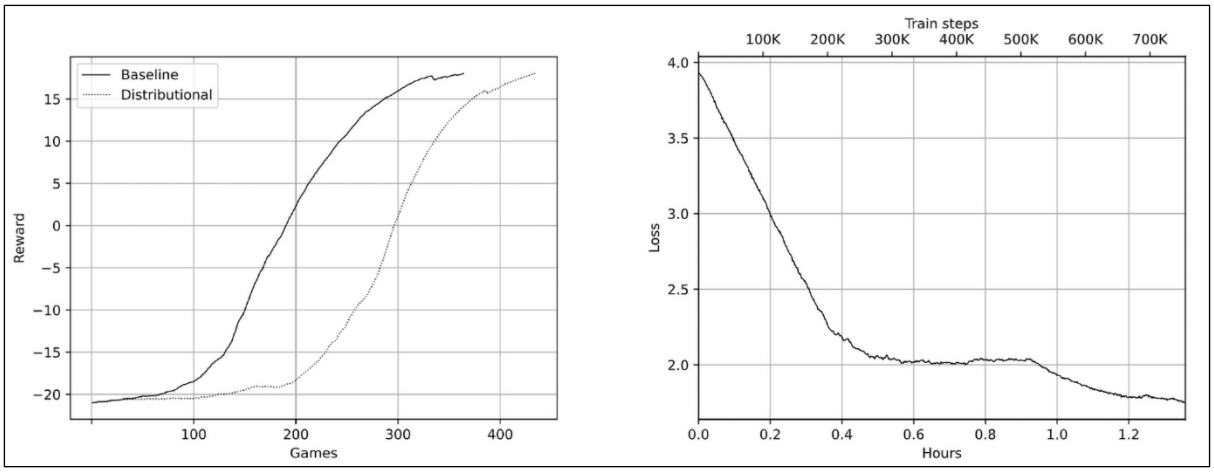

Train steps
100K
200K
300K
400K
500K
600K
Mean advantage
−0.012
−0.014
−0.016
−0.018
−0.020
−0.022
−0.024
−0.026
−0.028
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Hours
Train steps
100K
200K
300K
400K
500K
600K
Mean value
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Hours

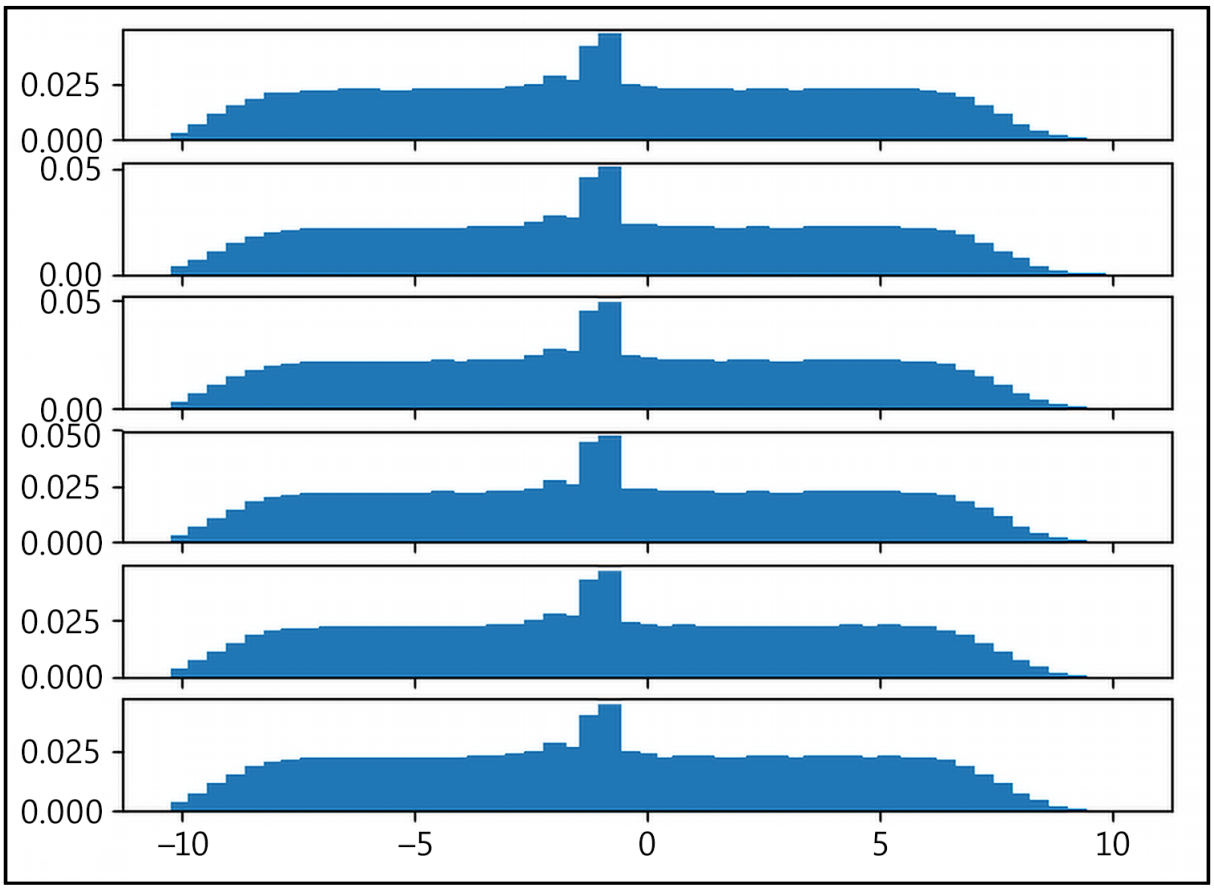

Car commute time distribution
mean=35.43 mins
Probability
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
20
30
40
50
60
70
80
90
100
Time, minutes

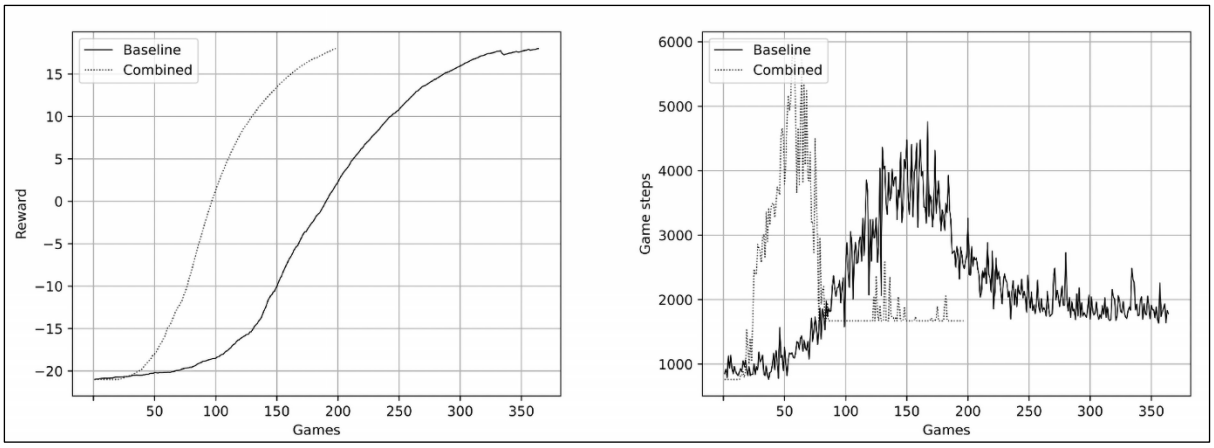

Train commute time distribution
mean=40.54 mins
Probability
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
35
40
45
50
55
60
Time, minutes

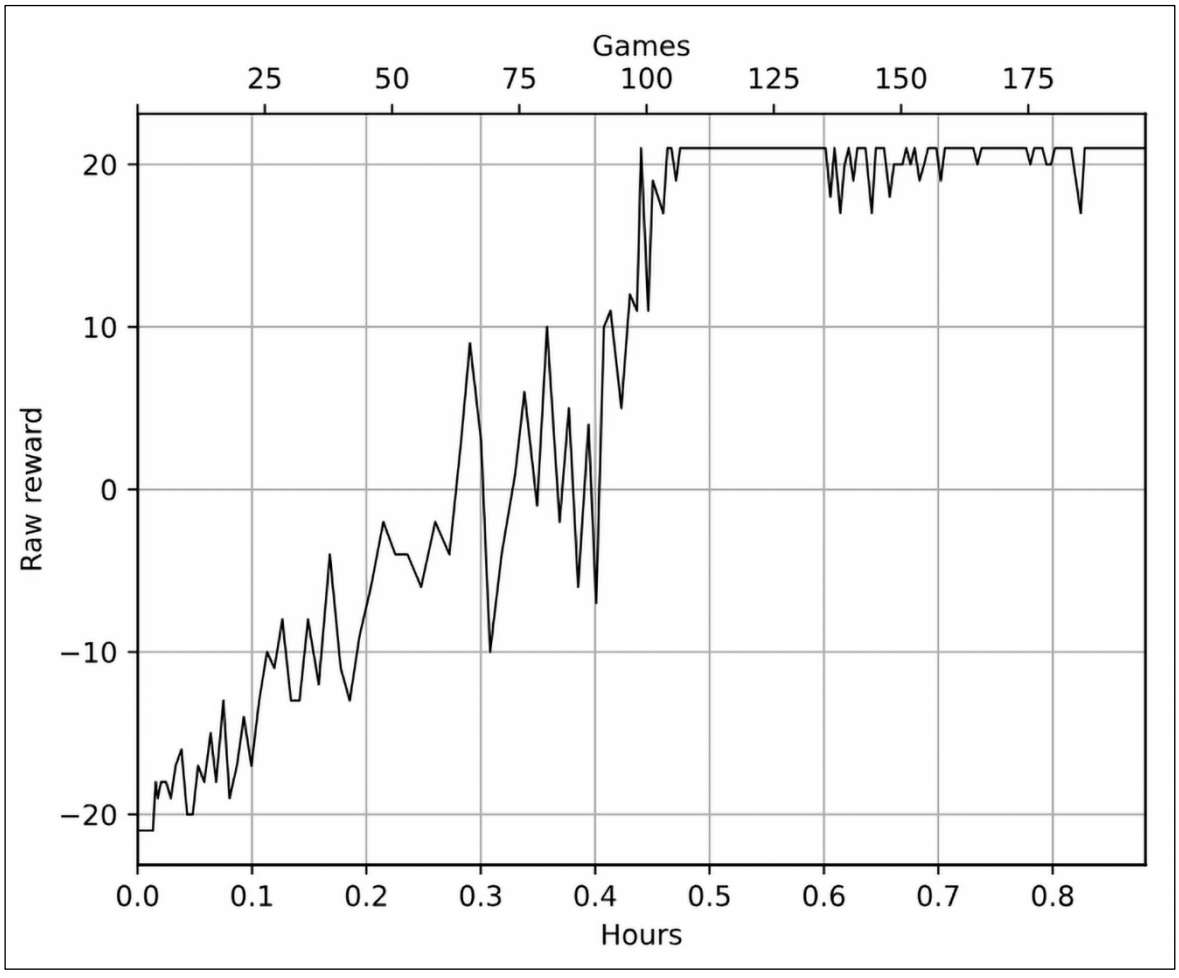

Source
60
40
20
0
−10
−5
0
5
10
Projected
60
40
20
0
−10
−5
0
5
10

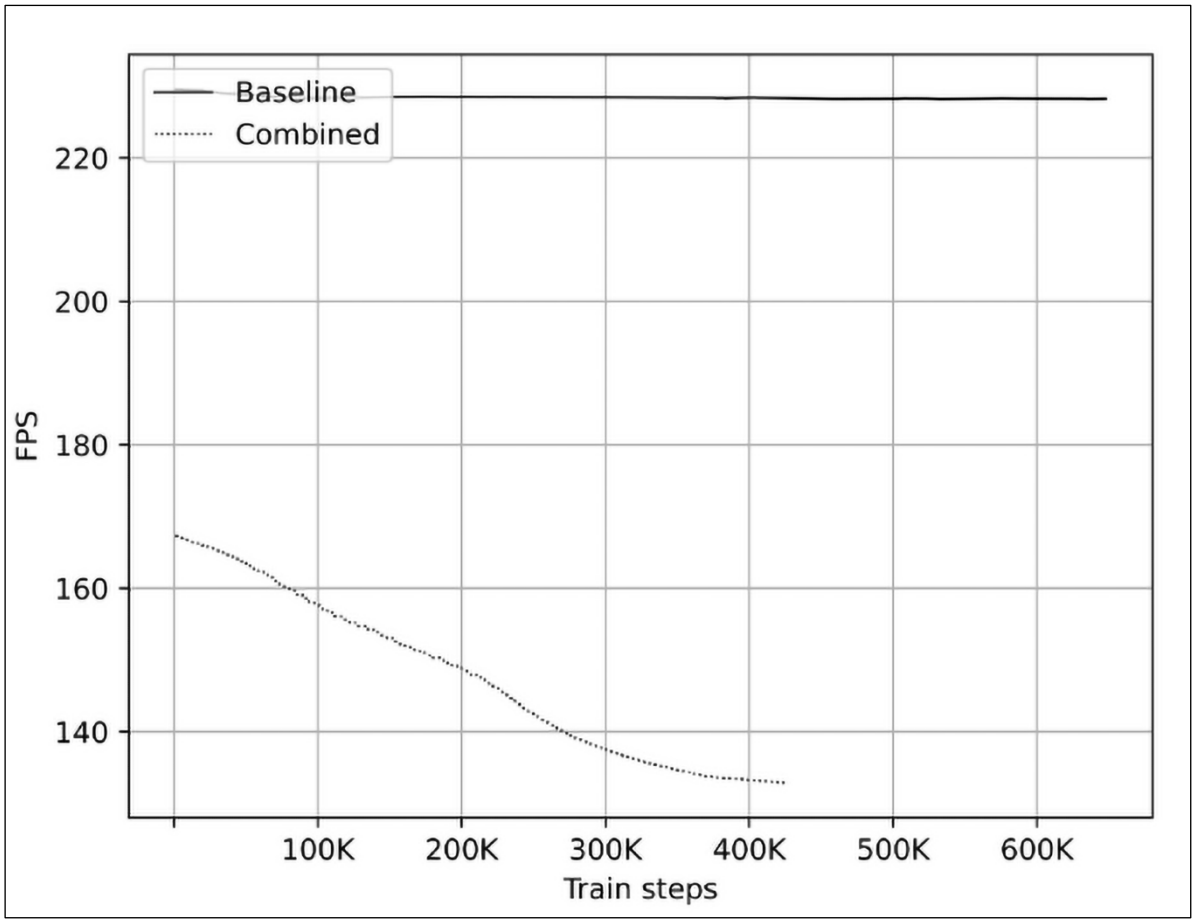

Source
60
40
20
0
−10
−5
0
5
10
Projected
1.0
0.8
0.6
0.4
0.2
0.0
−10
−5
0
5
10

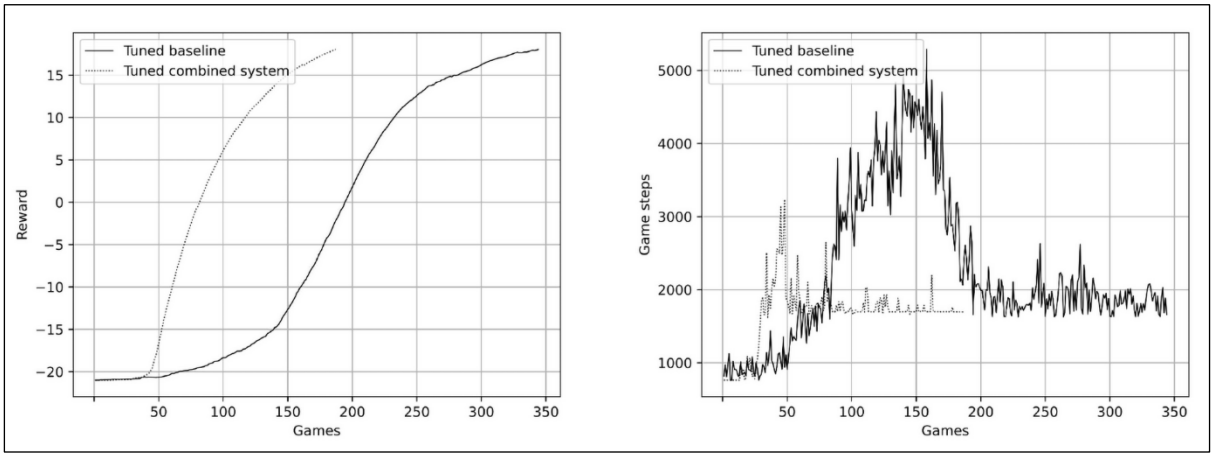

Baseline
Distributional
Reward
Games
Train steps
100K
200K
300K
400K
500K
600K
700K
Loss
Hours

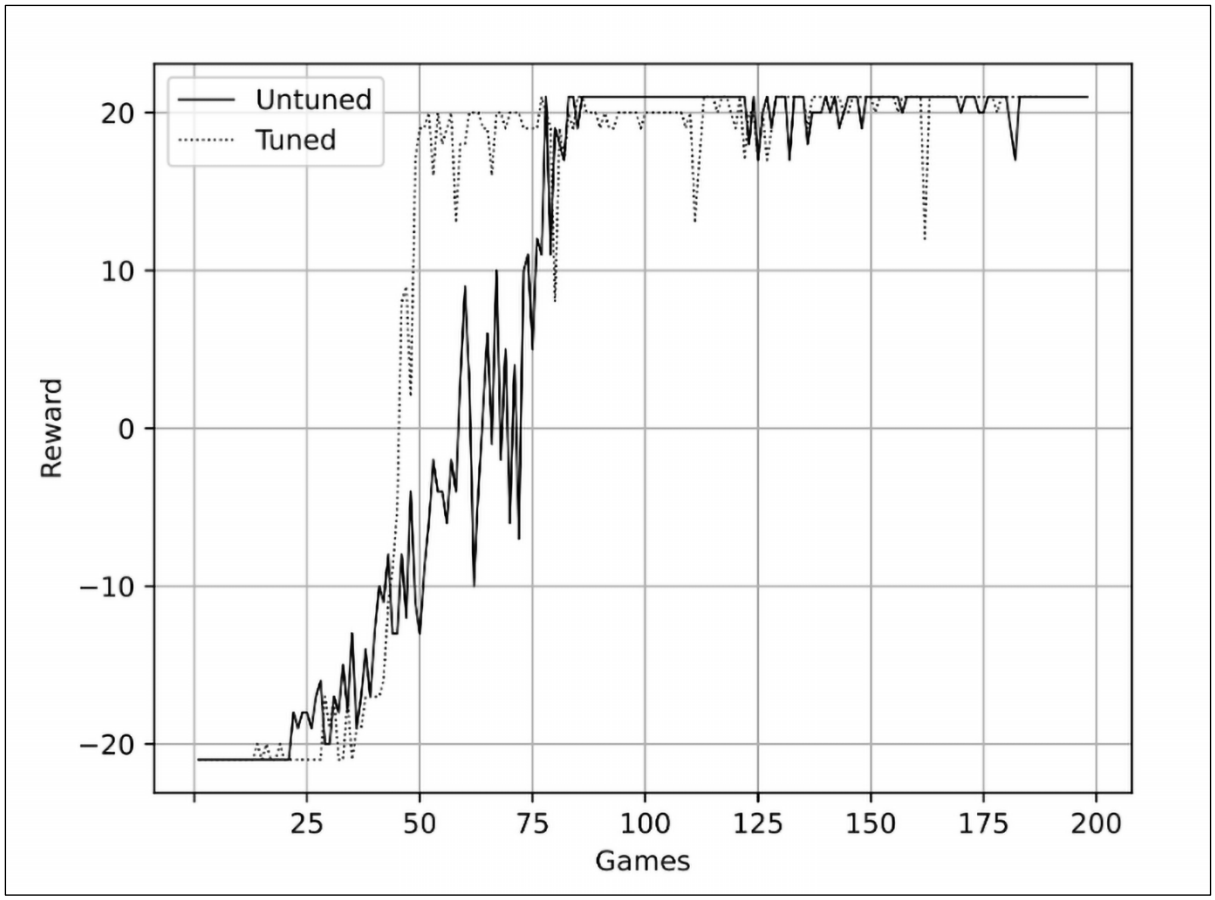



Baseline
Combined
Reward
Games
Baseline
Combined
Game steps
Games

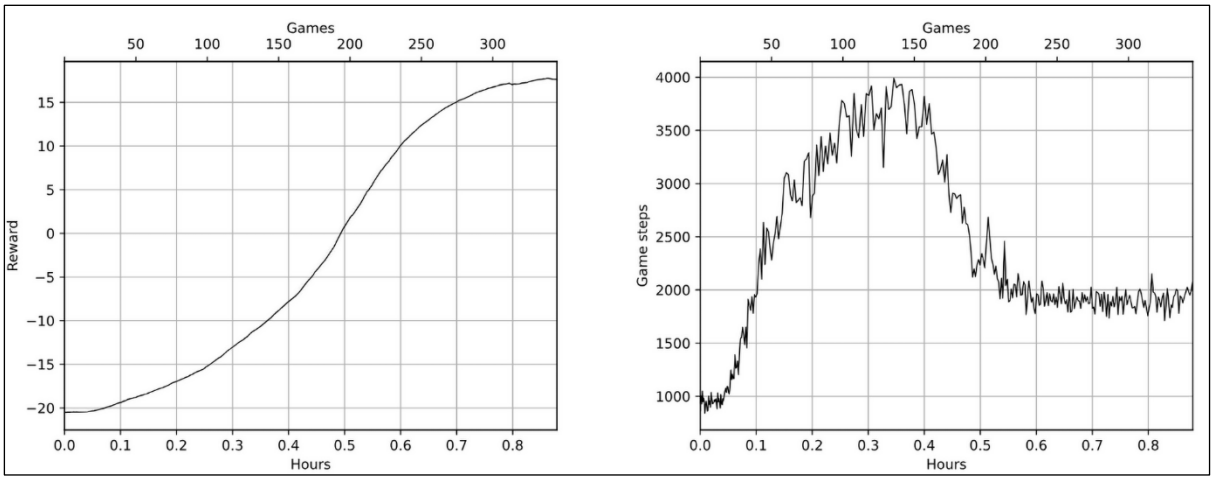

Games
Raw reward
Hours

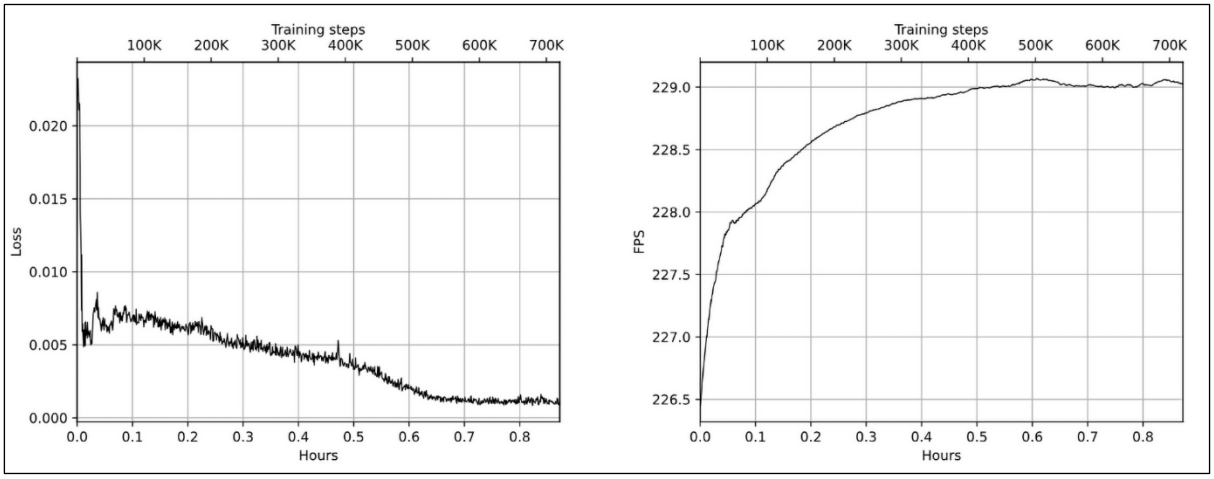

Baseline
Combined
FPS
220
200
180
160
140
100K
200K
300K
400K
500K
600K
Train steps

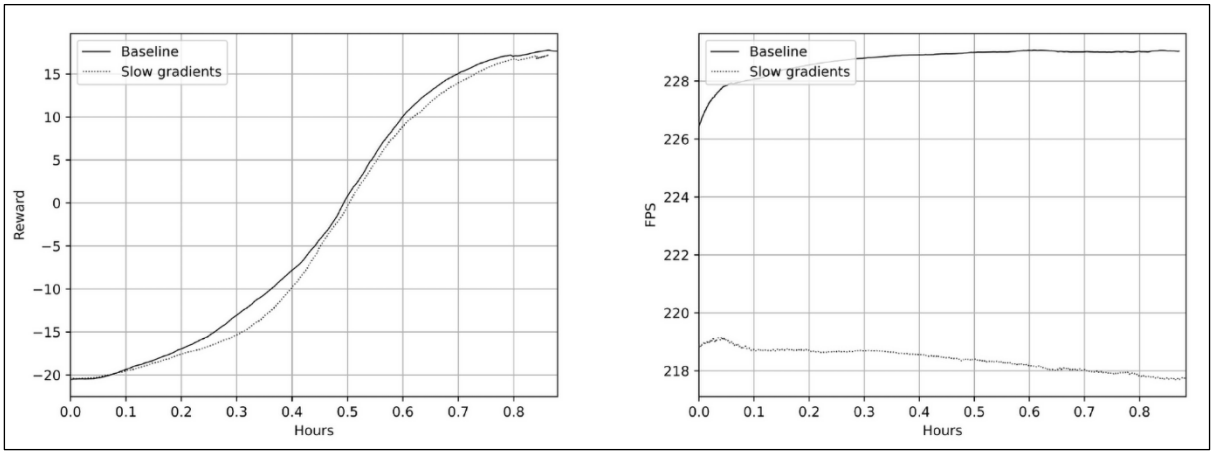

Tuned baseline
Tuned combined system
Reward
15
10
5
0
−5
−10
−15
−20
50
100
150
200
250
300
350
Games
Tuned baseline
Tuned combined system
Game steps
5000
4000
3000
2000
1000
50
100
150
200
250
300
350
Games

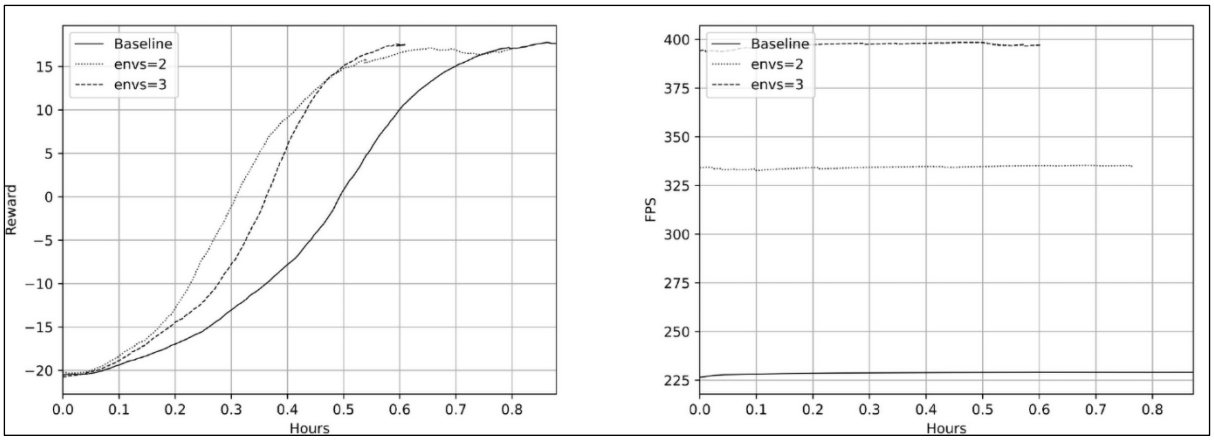

Untuned
Tuned
20
10
0
−10
−20
Reward
25
50
75
100
125
150
175
200
Games

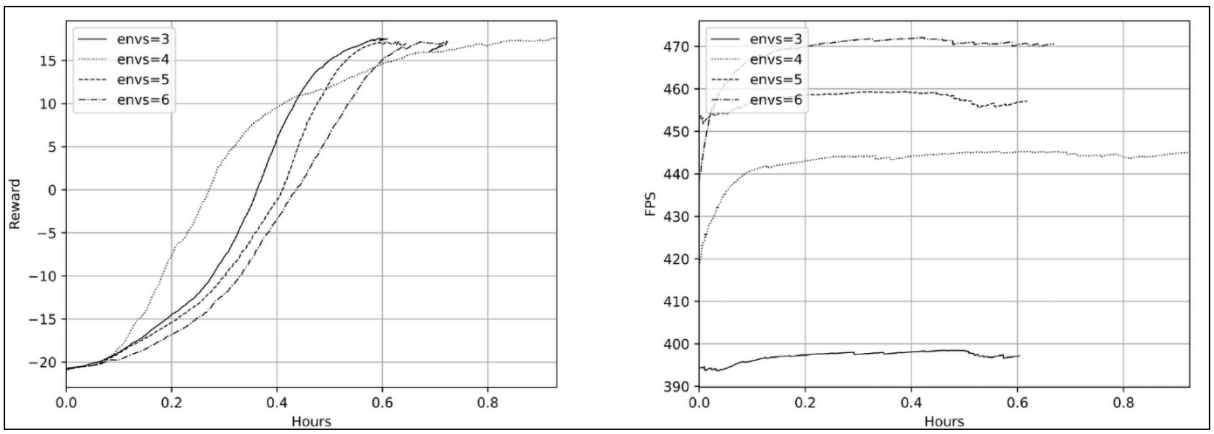
Chapter 9: Ways to Speed up RL Methods

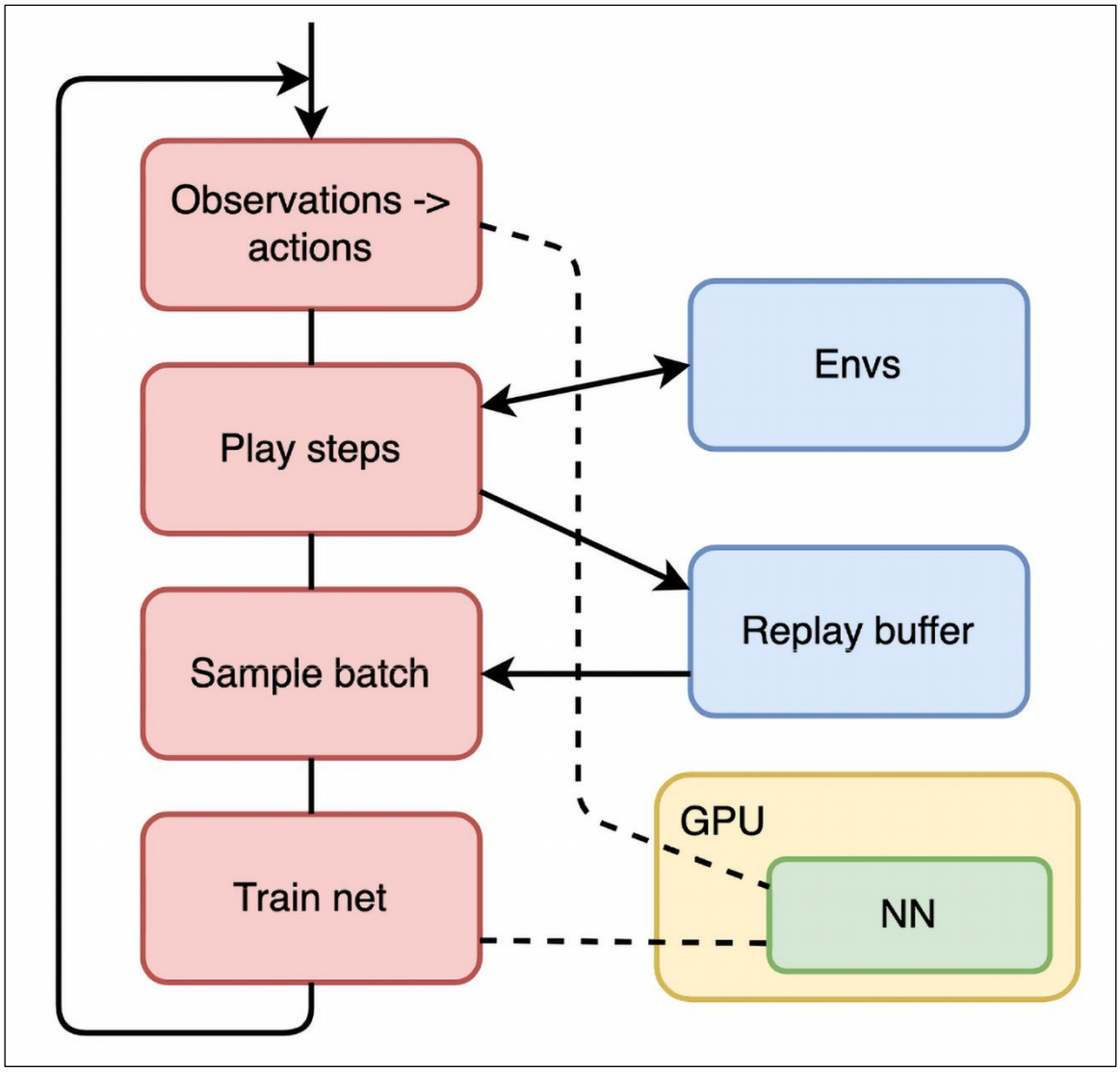


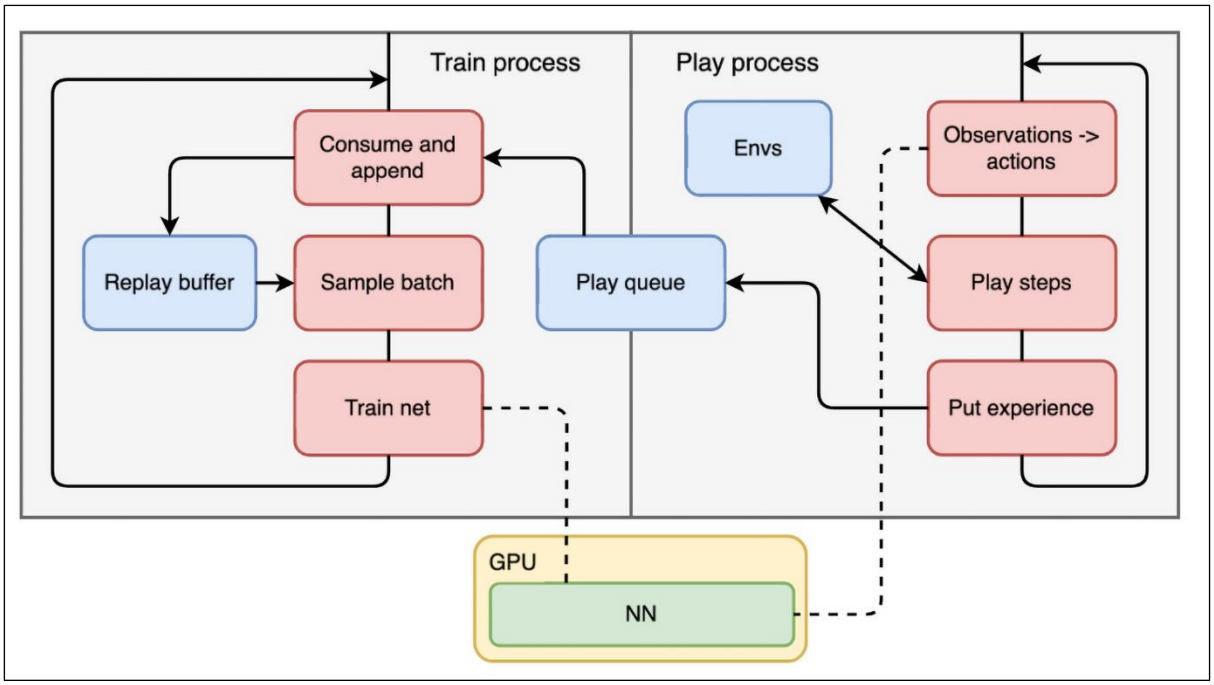


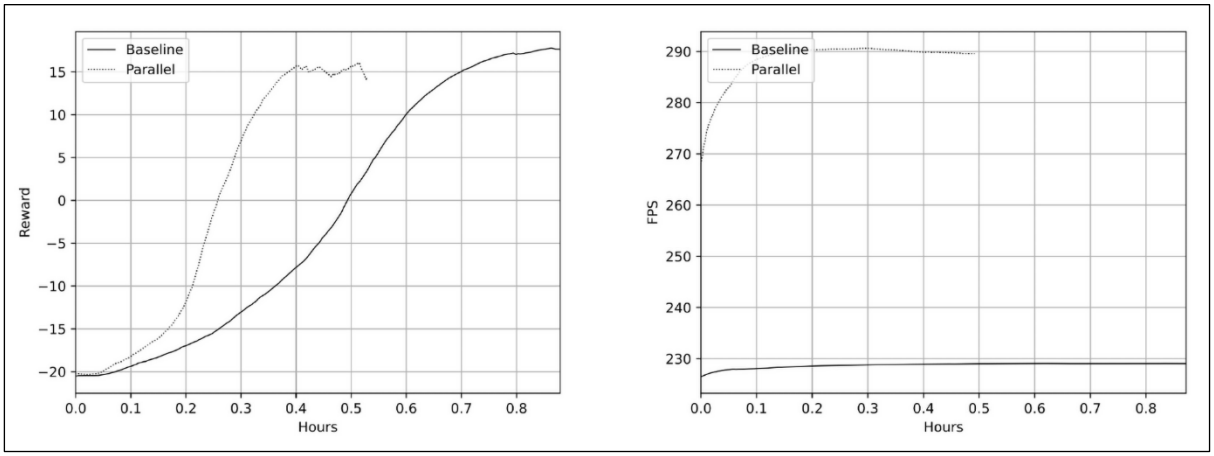


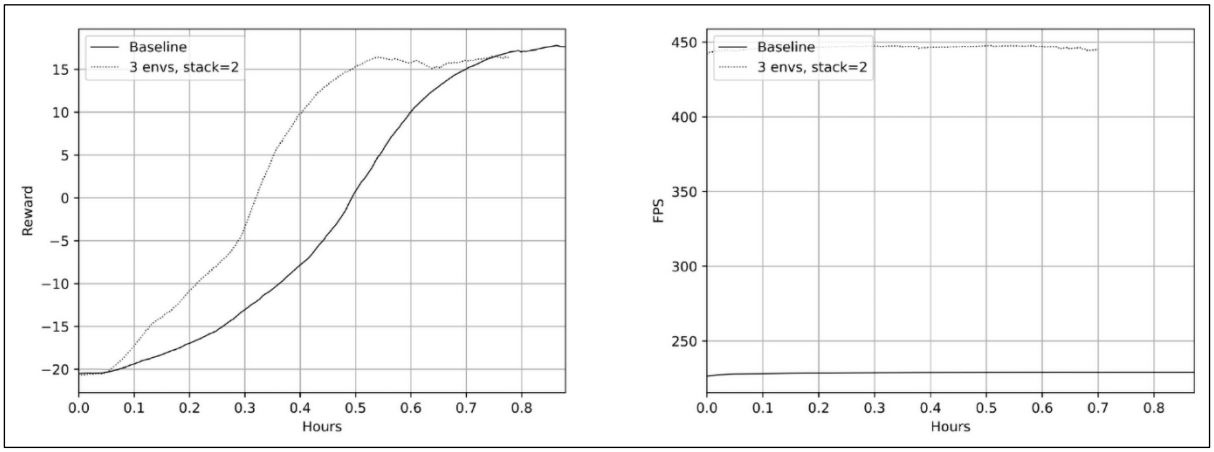

Baseline
envs=2
envs=3
Reward
Hours
FPS

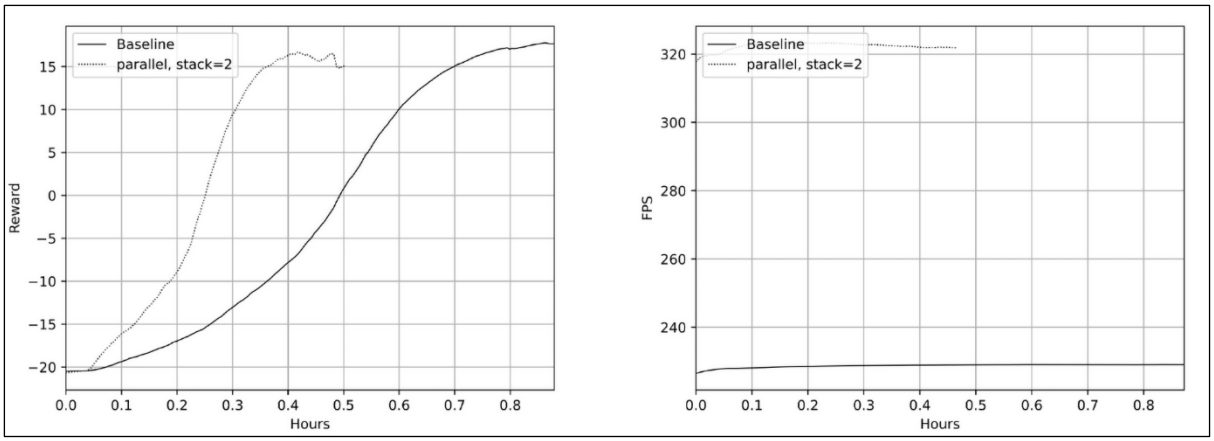

envs=3
envs=4
envs=5
envs=6
Reward
Hours
FPS


Observations -> actions
Play steps
Sample batch
Train net
Envs
Replay buffer
GPU
NN


Train process
Play process
Consume and append
Envs
Observations -> actions
Replay buffer
Sample batch
Play queue
Play steps
Train net
Put experience
GPU
NN

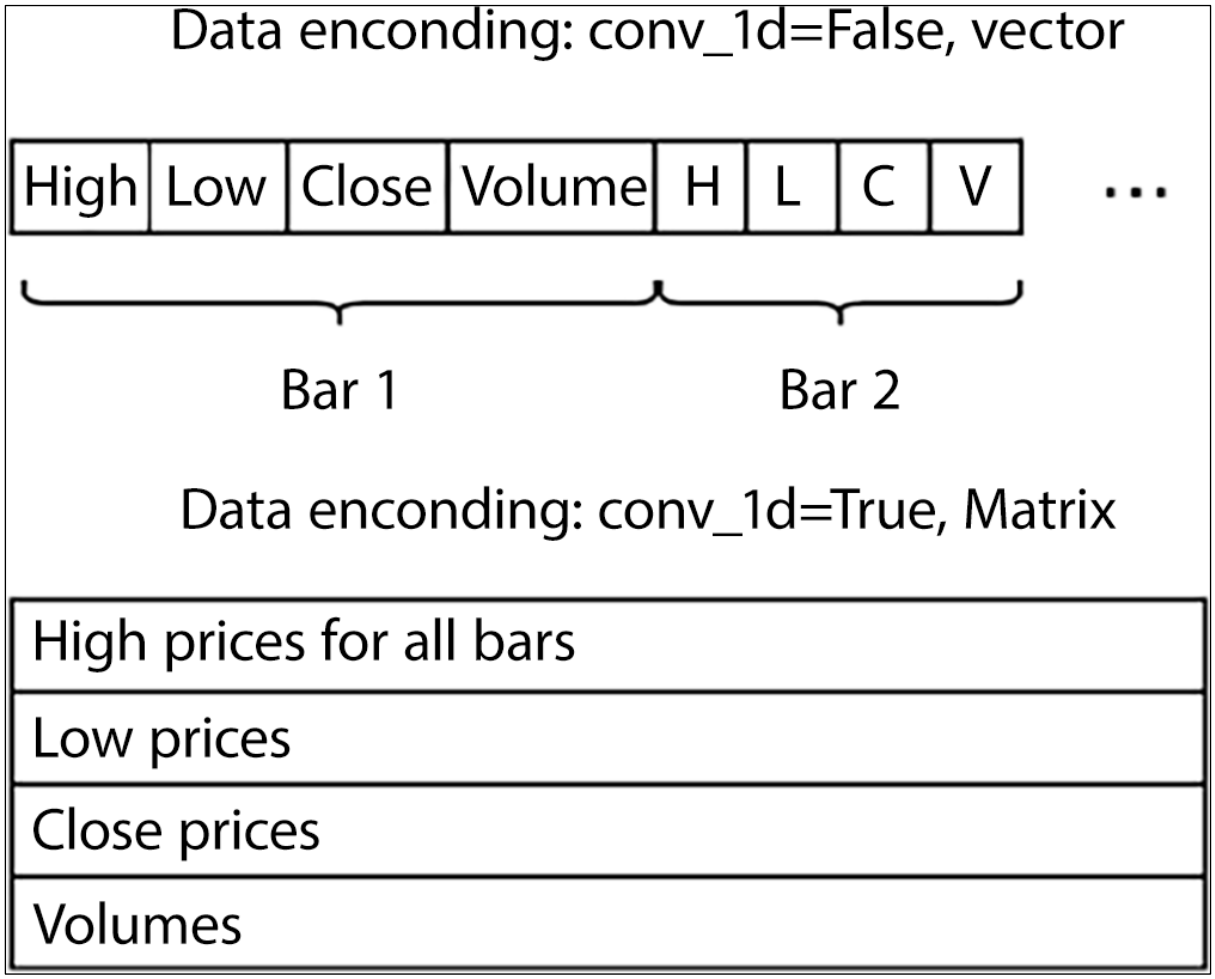

Baseline
Parallel
Reward
Hours
FPS

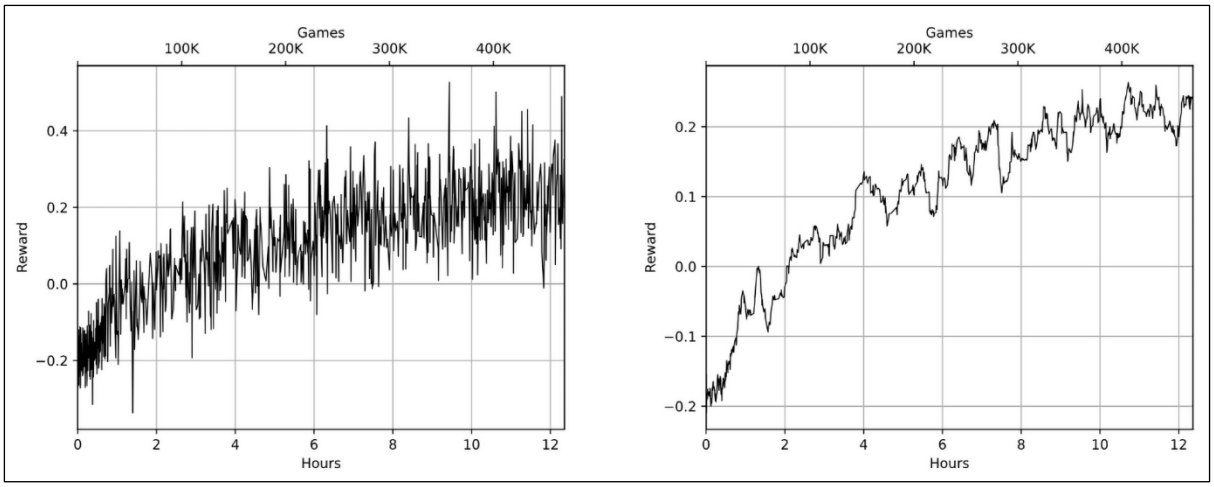

Baseline
3 envs, stack=2
Reward
Hours
FPS

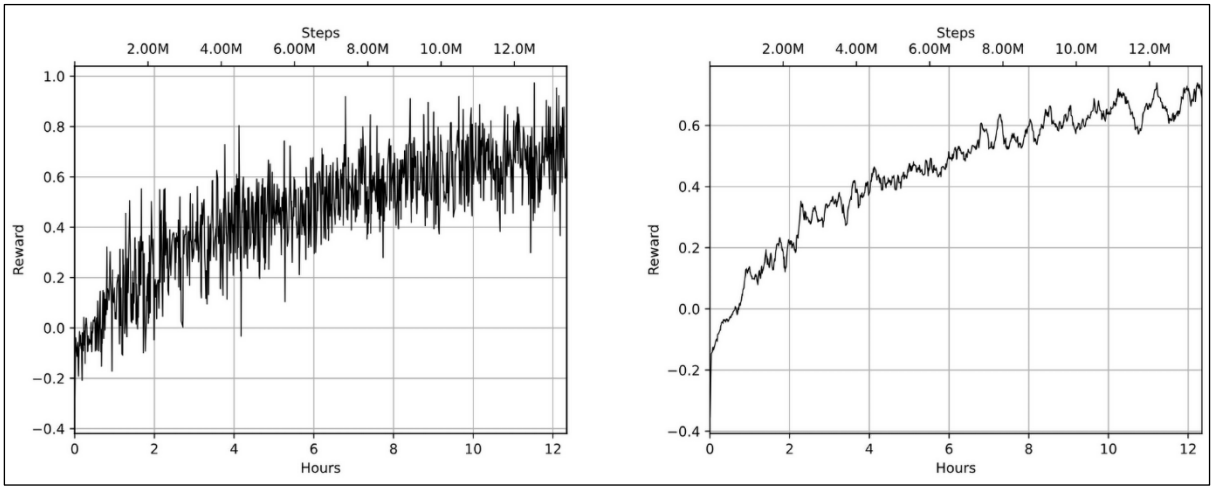

Baseline
parallel, stack=2
Reward
Hours
FPS
Hours

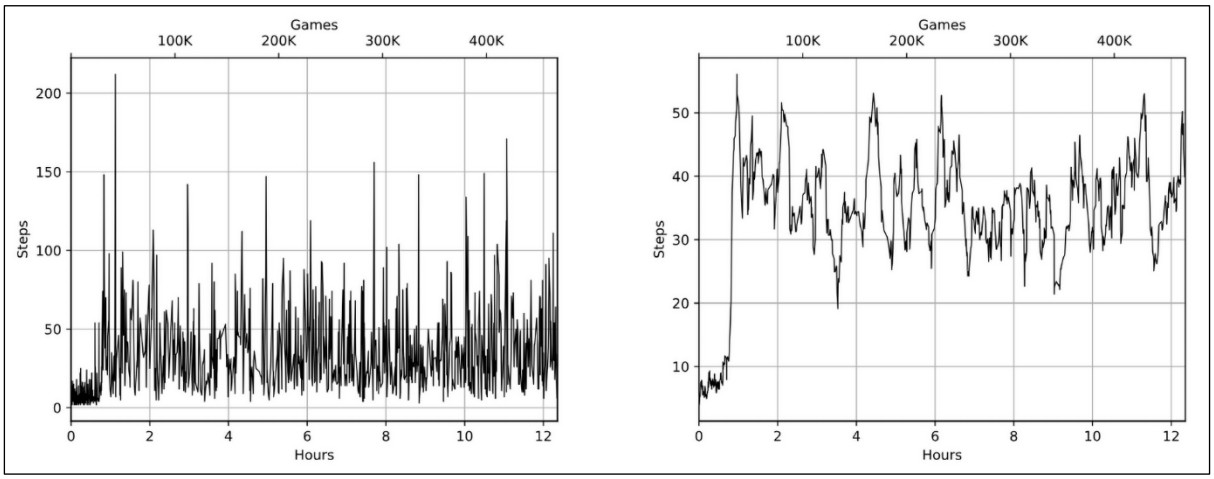
Chapter 10: Stocks Trading Using RL

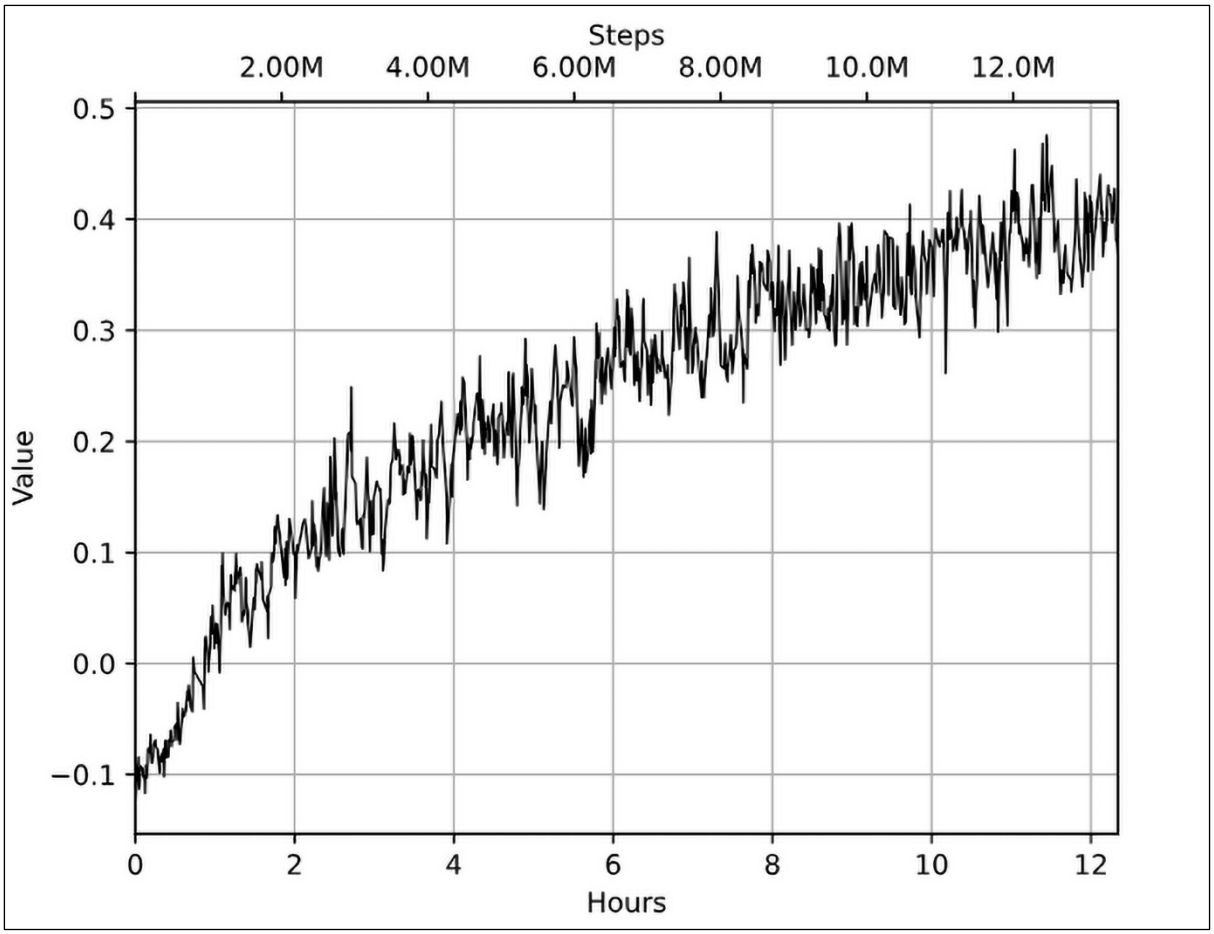


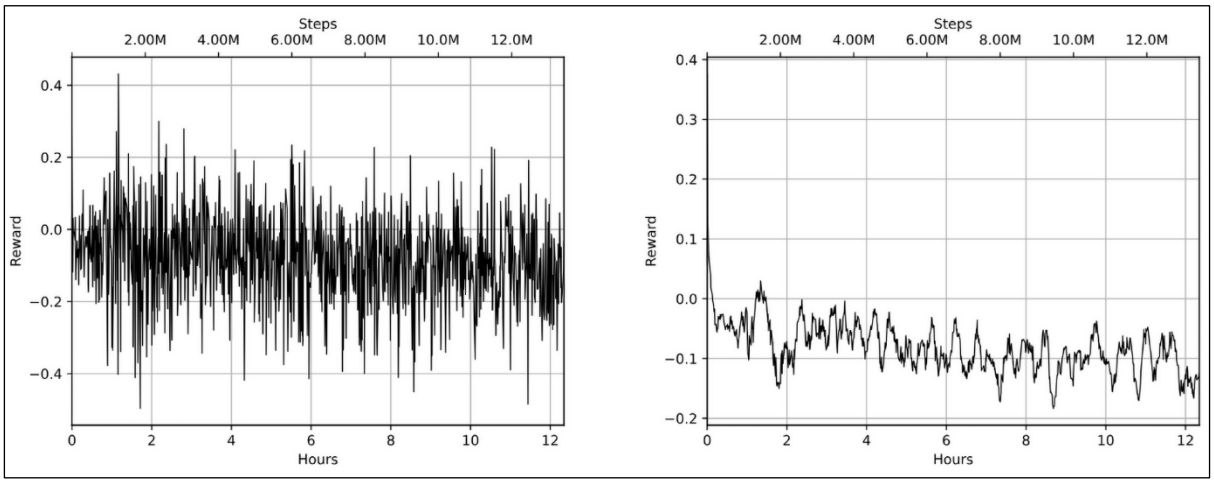


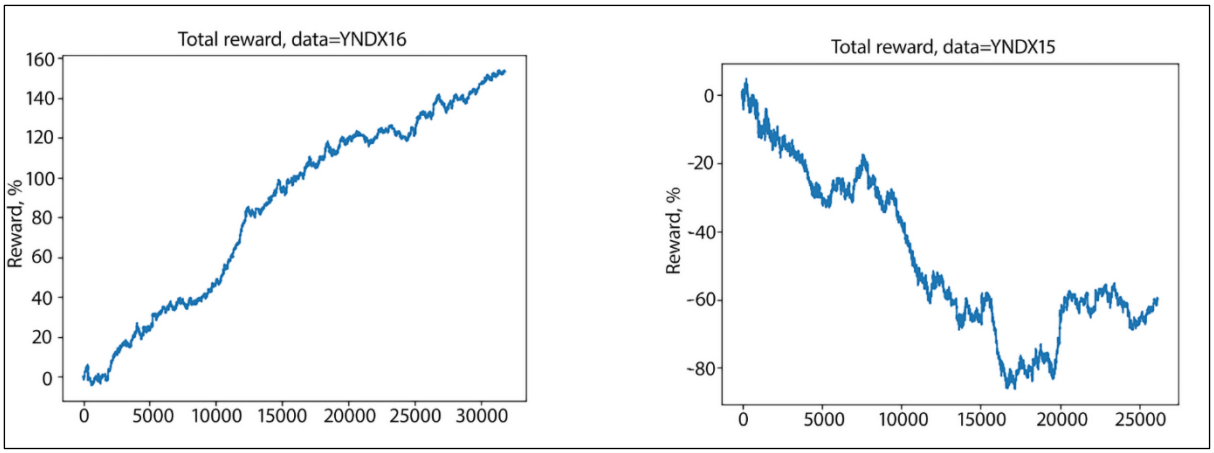

Games
100K
200K
300K
400K
Reward
Hours
Games
100K
200K
300K
400K
Reward
Hours

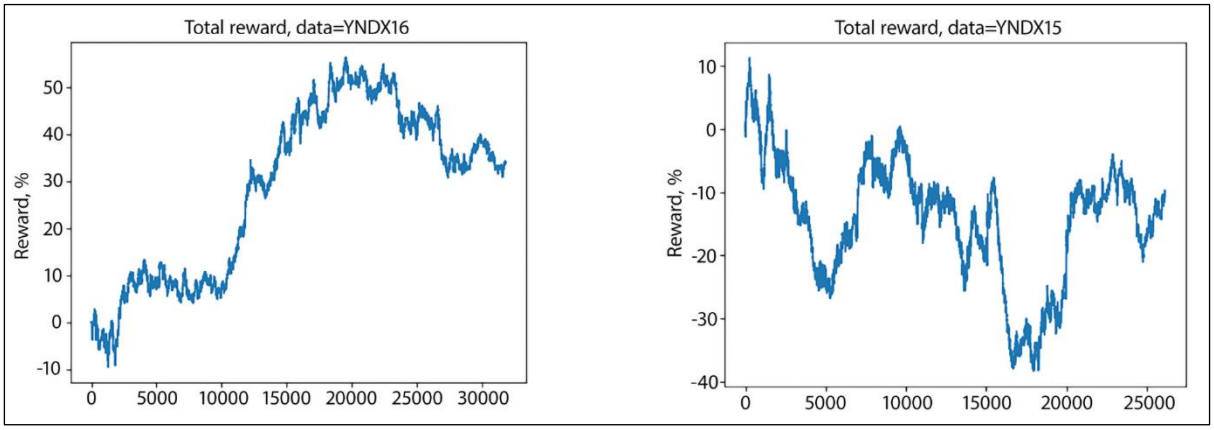

Steps
2.00M
4.00M
6.00M
8.00M
10.0M
12.0M
Reward
Hours
Steps
2.00M
4.00M
6.00M
8.00M
10.0M
12.0M
Reward
Hours

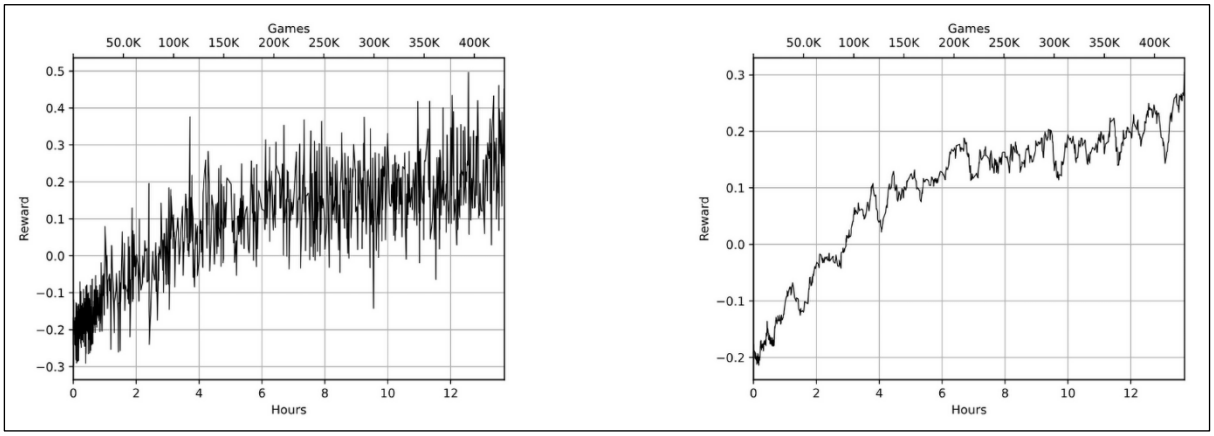

Games
100K
200K
300K
400K
Steps
Hours
Games
100K
200K
300K
400K
Steps
Hours

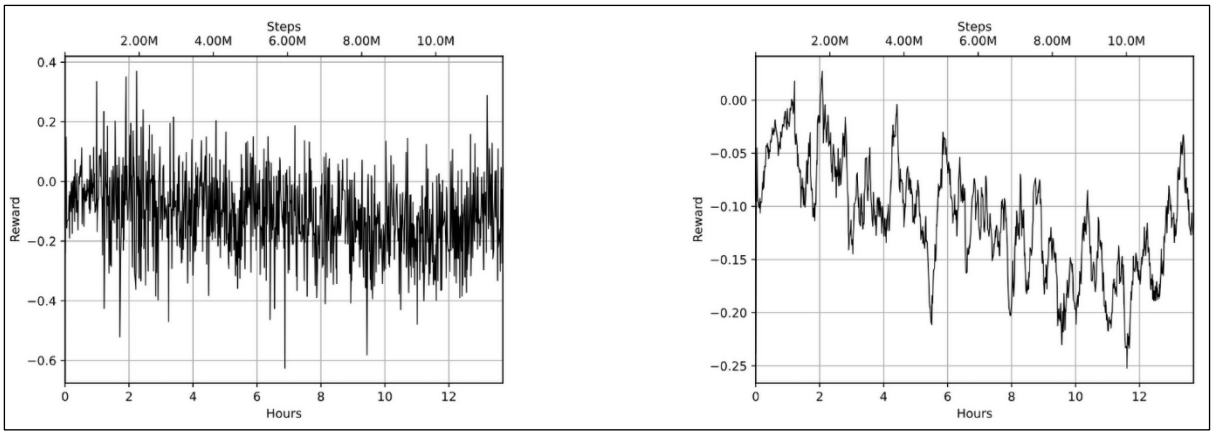

Steps
2.00M
4.00M
6.00M
8.00M
10.0M
12.0M
Value
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
0
2
4
6
8
10
12
Hours


Steps
2.00M
4.00M
6.00M
8.00M
10.0M
12.0M
Reward
0.4
0.2
0.0
−0.2
−0.4
0
2
4
6
8
10
12
Hours
Steps
2.00M
4.00M
6.00M
8.00M
10.0M
12.0M
Reward
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
0
2
4
6
8
10
12
Hours

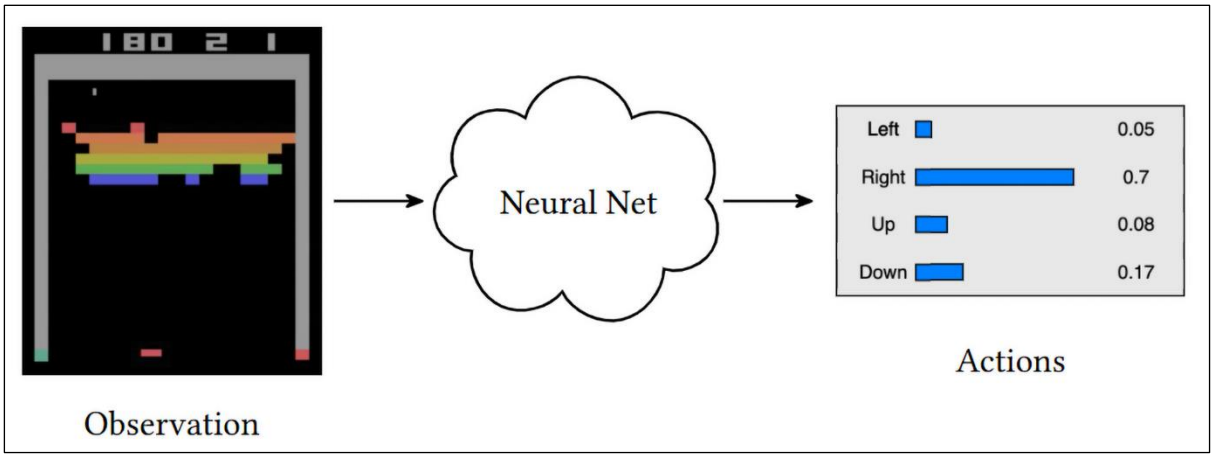

Total reward, data=YNDX16
Reward, %
Total reward, data=YNDX15
Reward, %

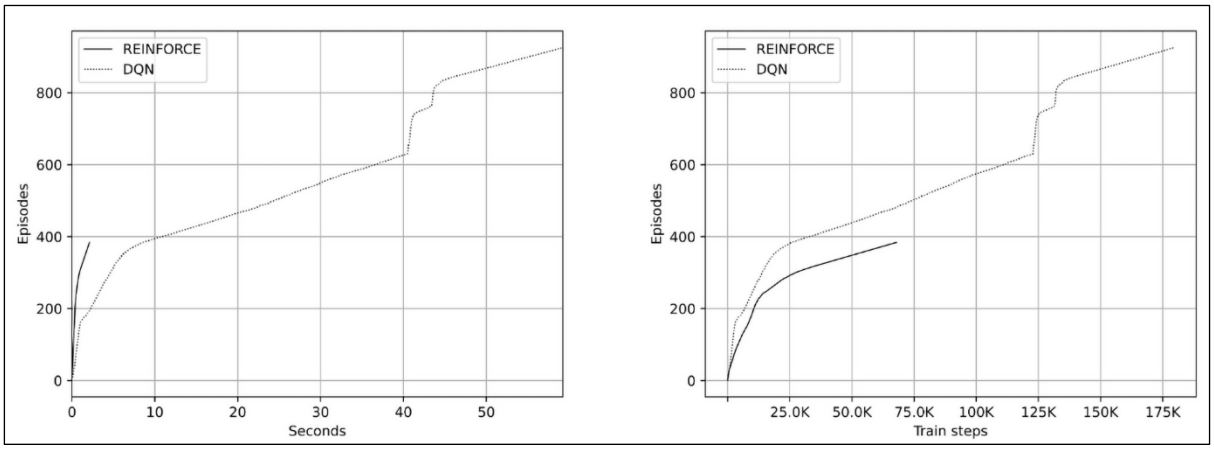

Total reward, data=YNDX16
Reward, %
Total reward, data=YNDX15
Reward, %

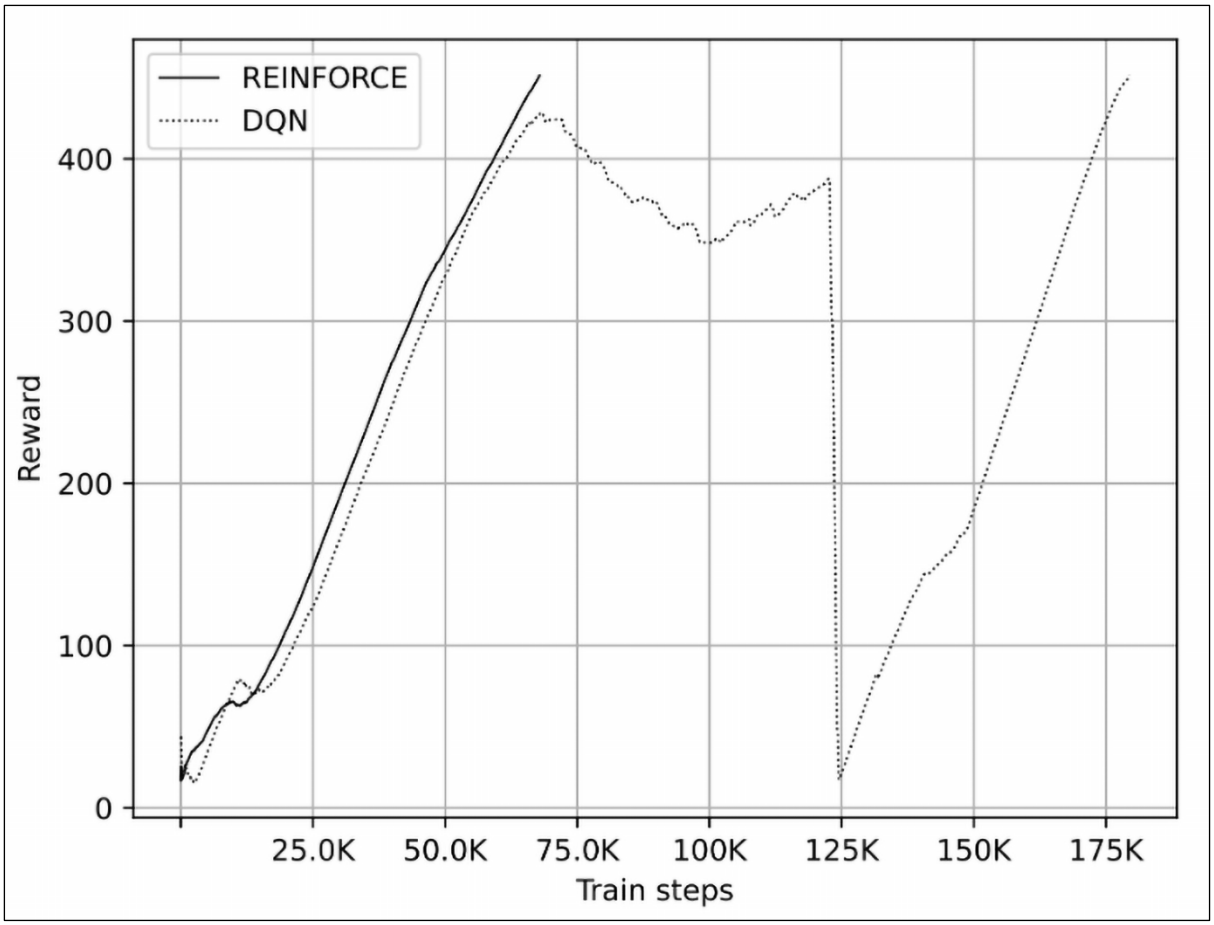

Games
50.0K 100K 150K 200K 250K 300K 350K 400K
Reward
Hours
Games
50.0K 100K 150K 200K 250K 300K 350K 400K
Reward
Hours

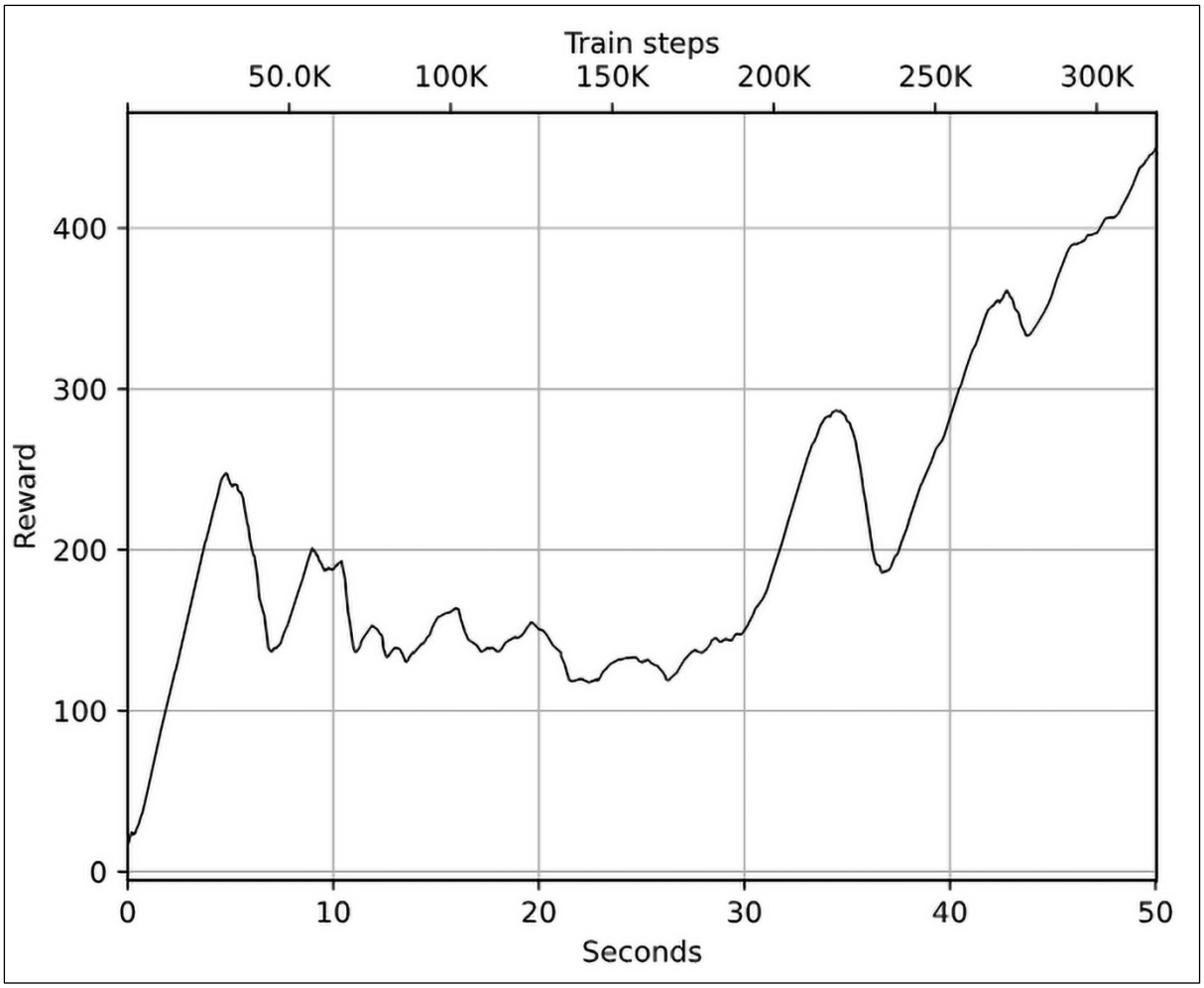

Steps
2.00M
4.00M
6.00M
8.00M
10.0M
Reward
0.4
0.2
0.0
−0.2
−0.4
−0.6
0
2
4
6
8
10
12
Hours
Steps
2.00M
4.00M
6.00M
8.00M
10.0M
Reward
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
0
2
4
6
8
10
12
Hours

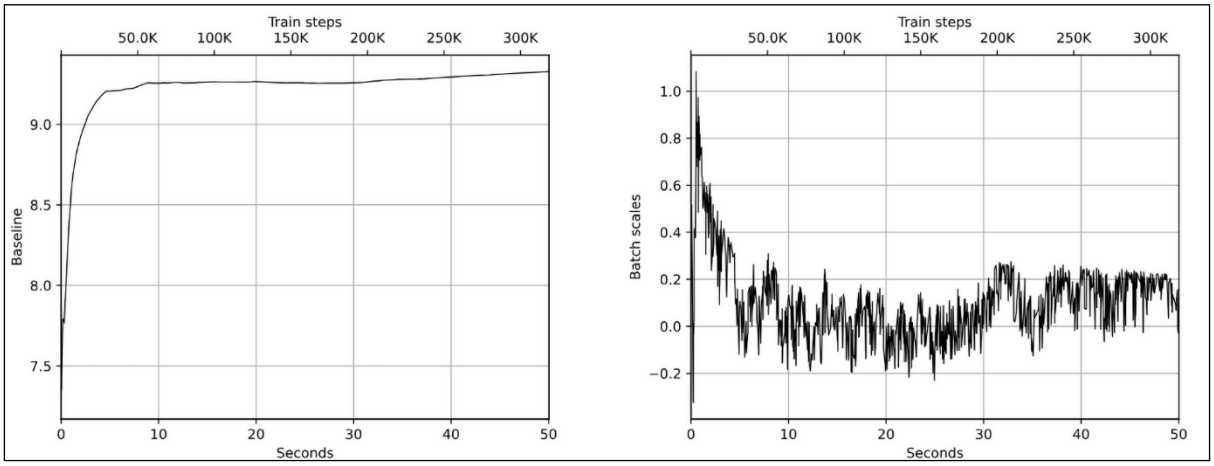
Chapter 11: Policy Gradients

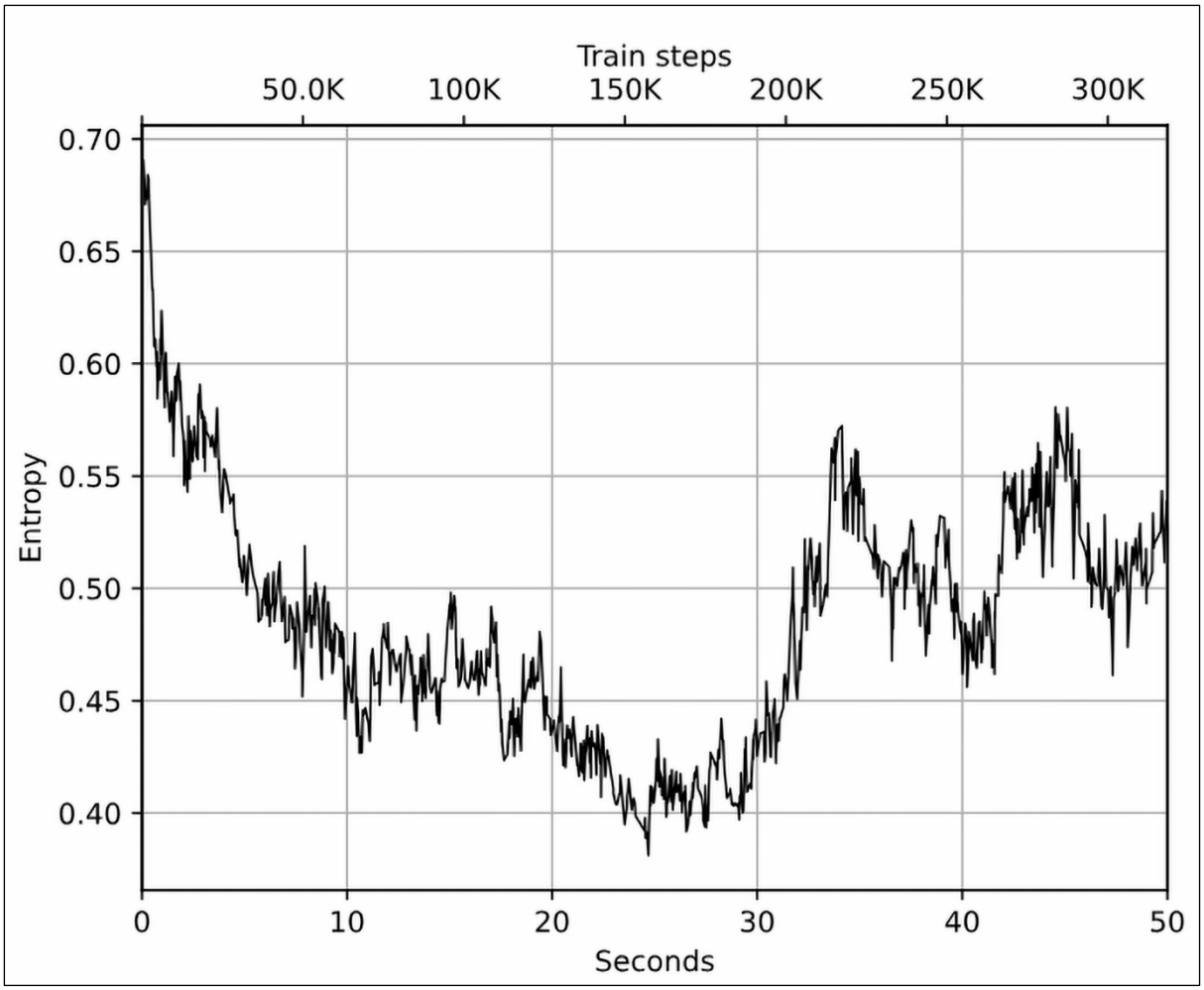


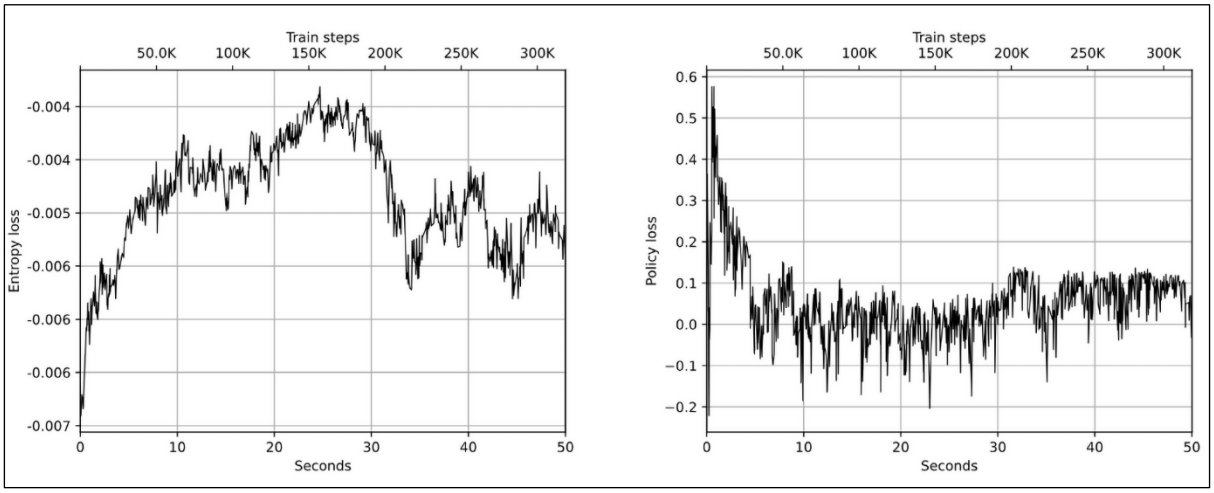


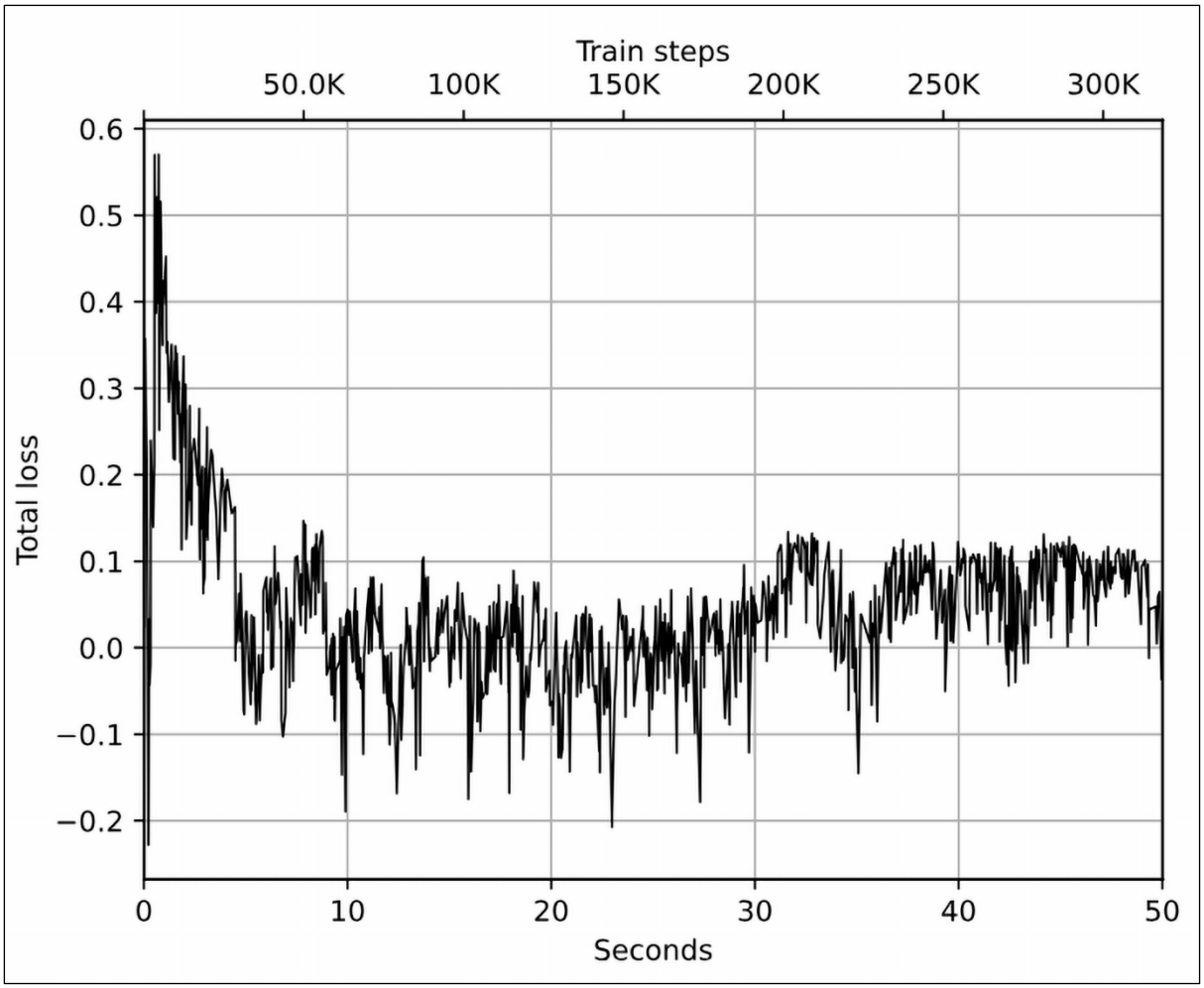

REINFORCE
DQN
400
300
200
100
0
Reward
25.0K
50.0K
75.0K
100K
125K
150K
175K
Train steps

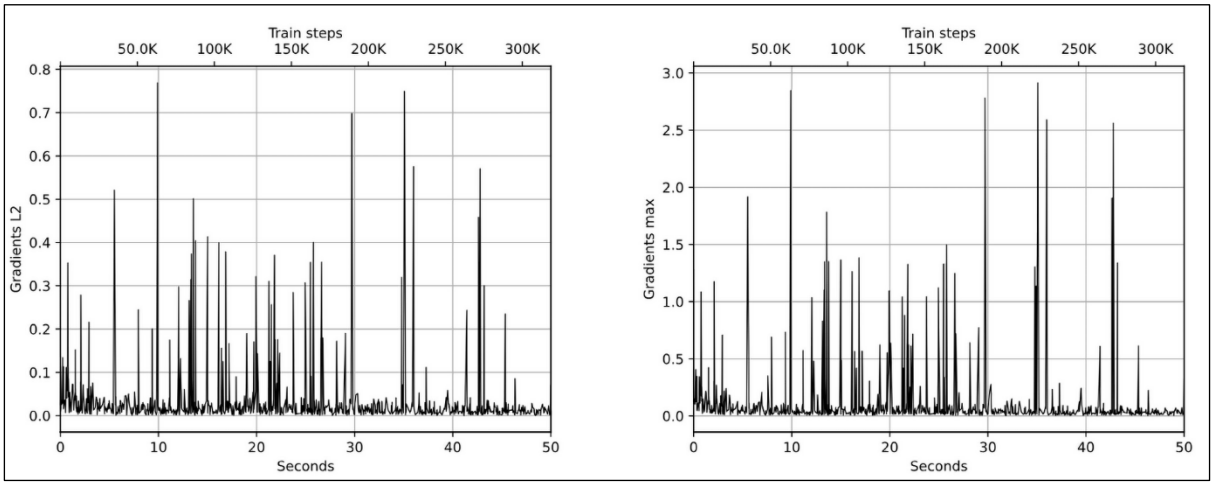

Train steps
50.0K
100K
150K
200K
250K
300K
Reward
0
100
200
300
400
0
10
20
30
40
50
Seconds

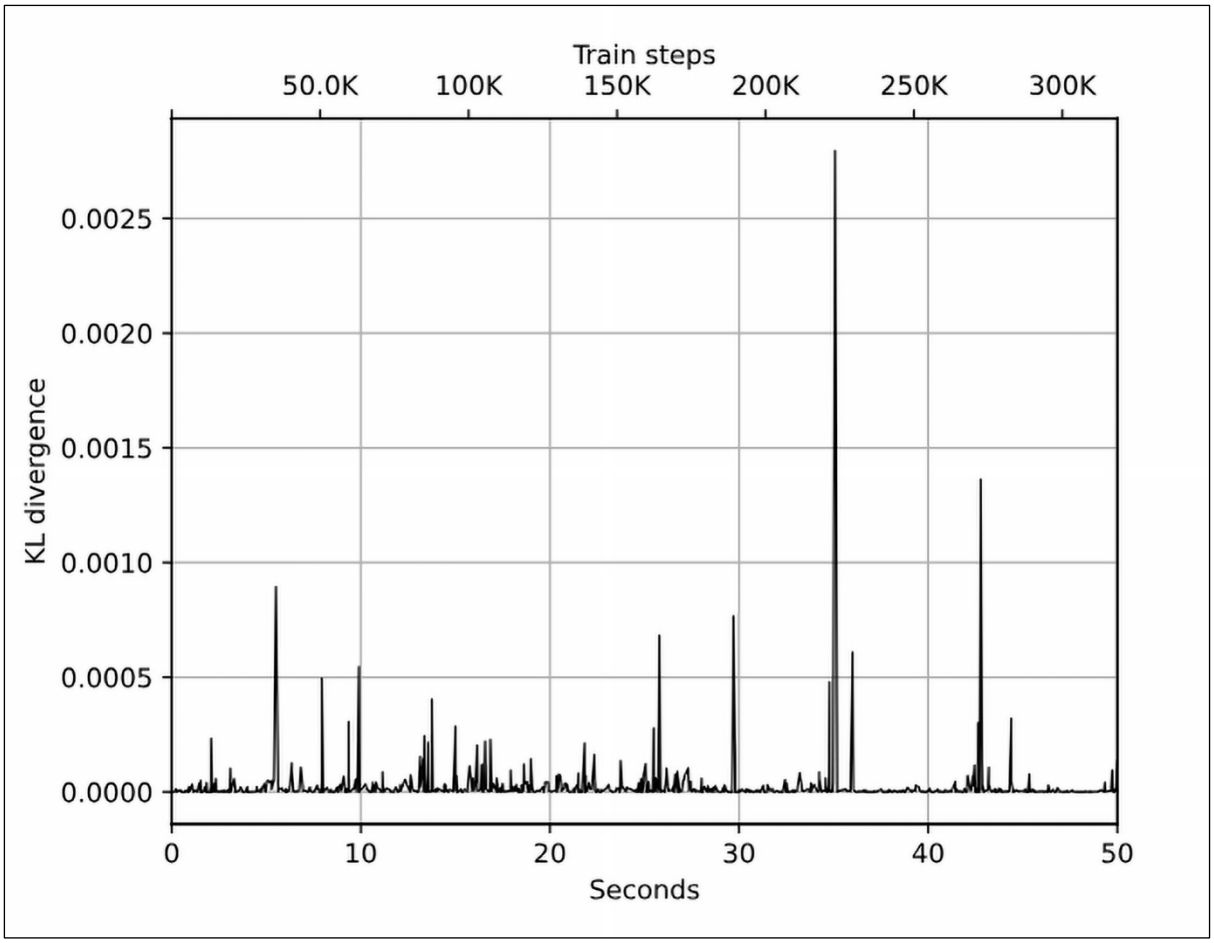

Train steps
50.0K
100K
150K
200K
250K
300K
Baseline
7.5
8.0
8.5
9.0
0
10
20
30
40
50
Seconds
Train steps
50.0K
100K
150K
200K
250K
300K
Batch scales
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
Seconds


Train steps
50.0K
100K
150K
200K
250K
300K
Entropy
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0
10
20
30
40
50
Seconds

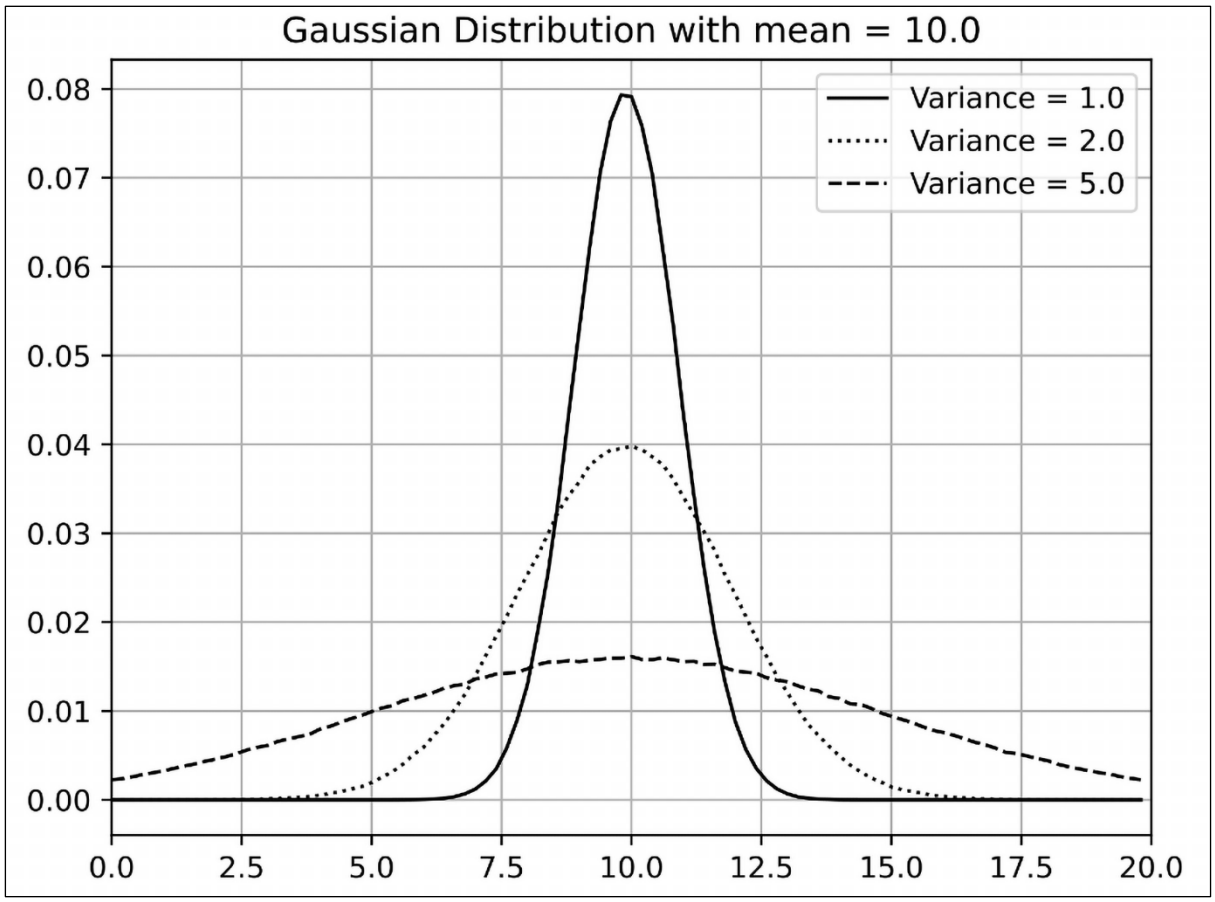

Train steps
50.0K
100K
150K
200K
250K
300K
Entropy loss
-0.004
-0.004
-0.005
-0.006
-0.006
-0.006
-0.007
0
10
20
30
40
50
Seconds
Train steps
50.0K
100K
150K
200K
250K
300K
Policy loss
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
0
10
20
30
40
50
Seconds

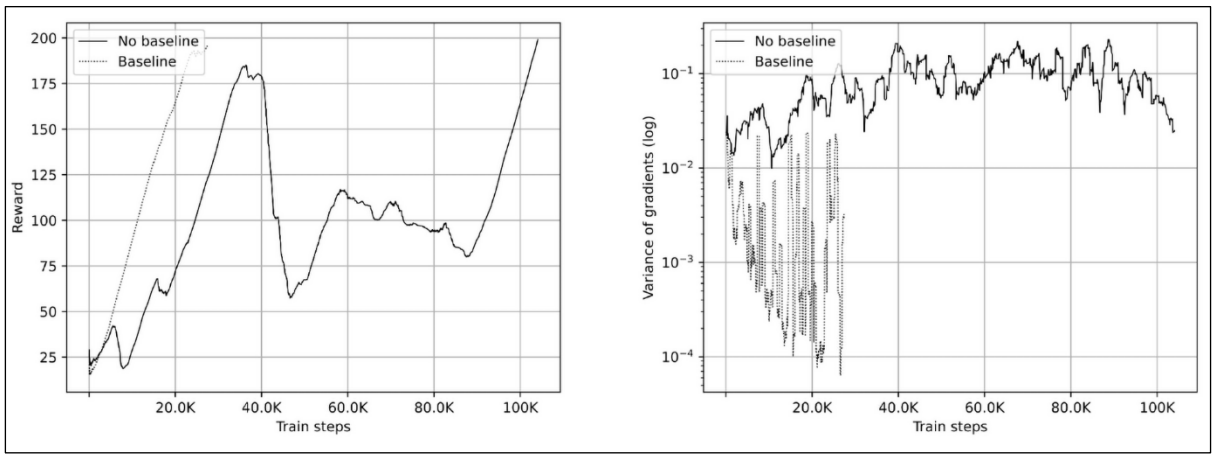

Train steps
50.0K
100K
150K
200K
250K
300K
Total loss
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
0
10
20
30
40
50
Seconds

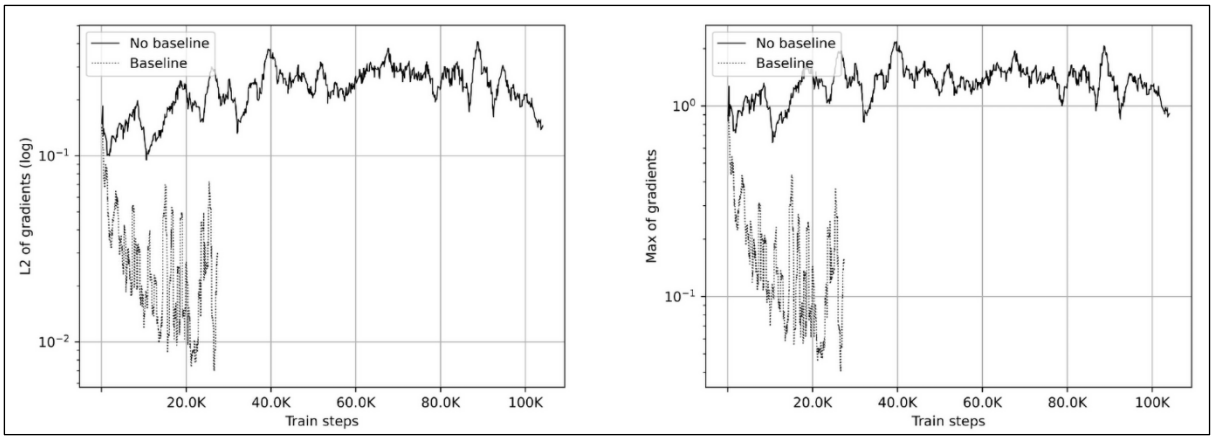

Train steps
50.0K
100K
150K
200K
250K
300K
Gradients L2
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
Seconds
Train steps
50.0K
100K
150K
200K
250K
300K
Gradients max
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
10
20
30
40
50
Seconds

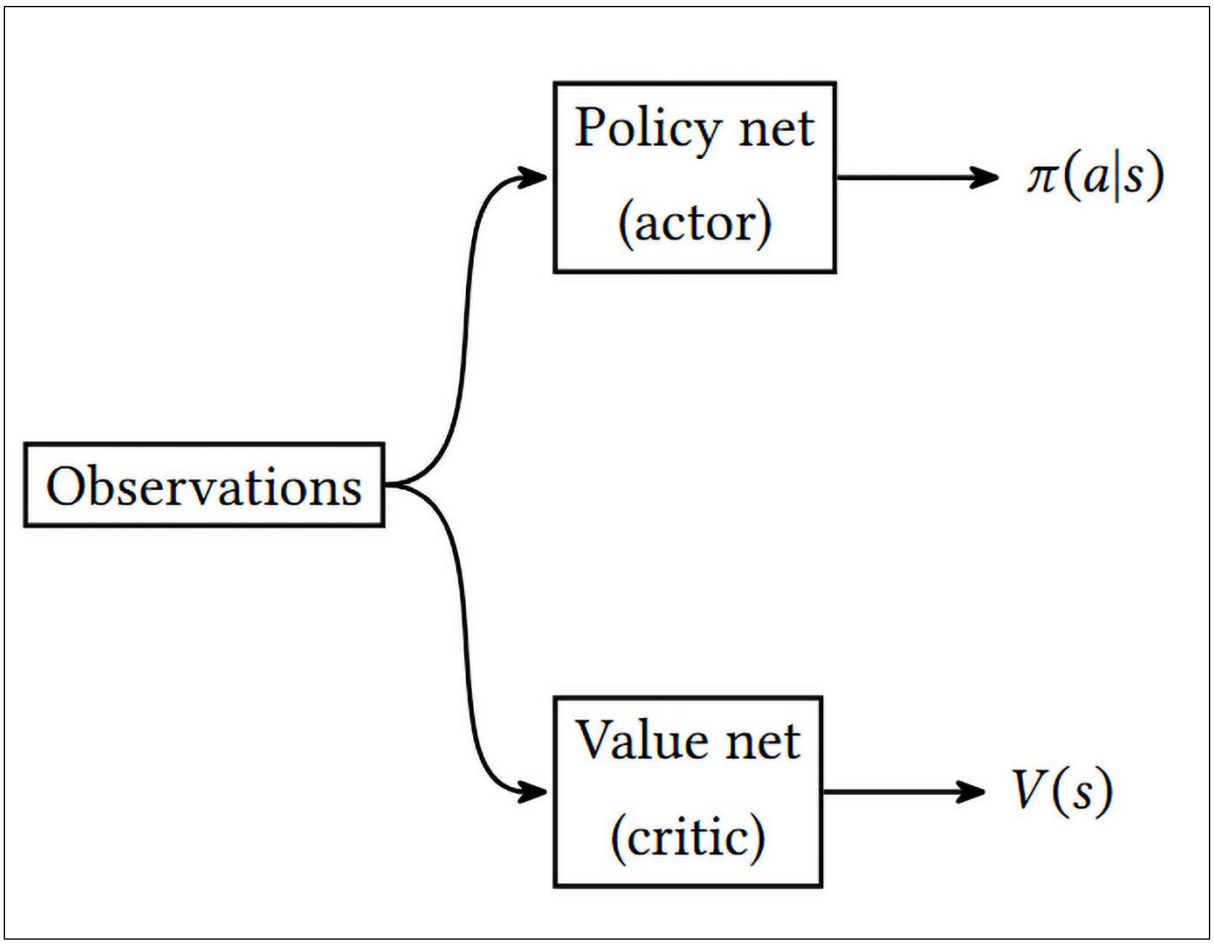

Train steps
50.0K
100K
150K
200K
250K
300K
KL divergence
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
10
20
30
40
50
Seconds

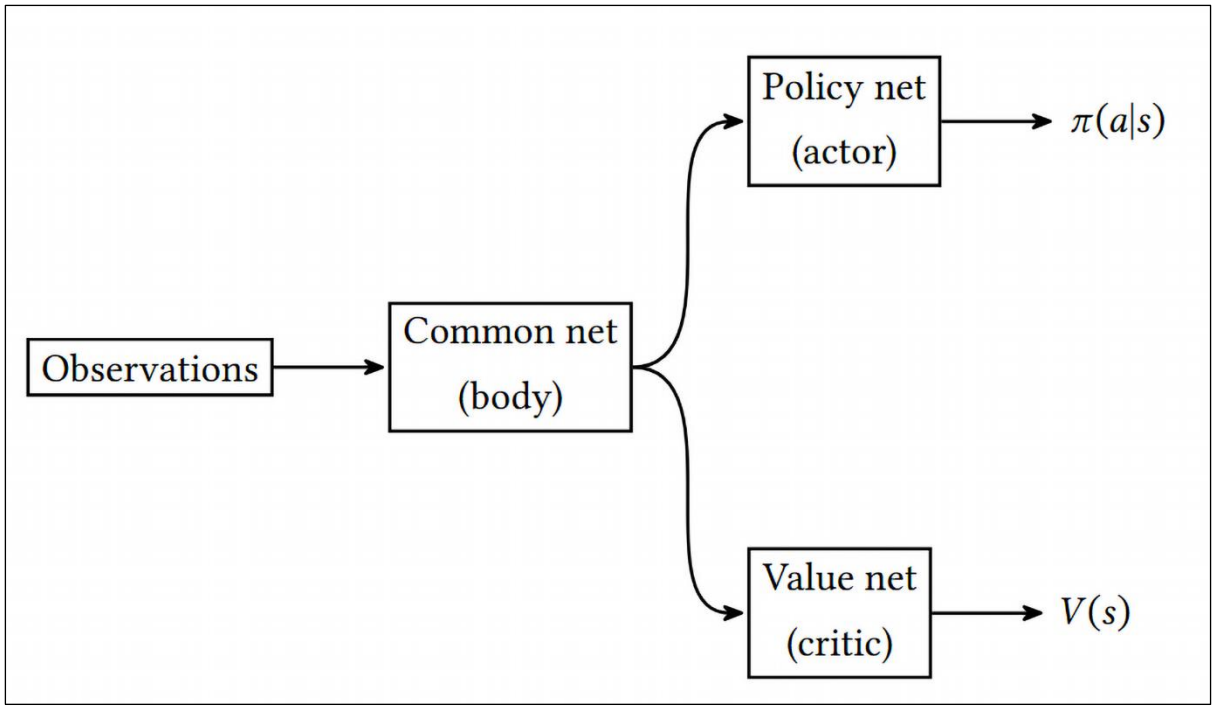
Chapter 12: The Actor-Critic Method: A2C and A3C

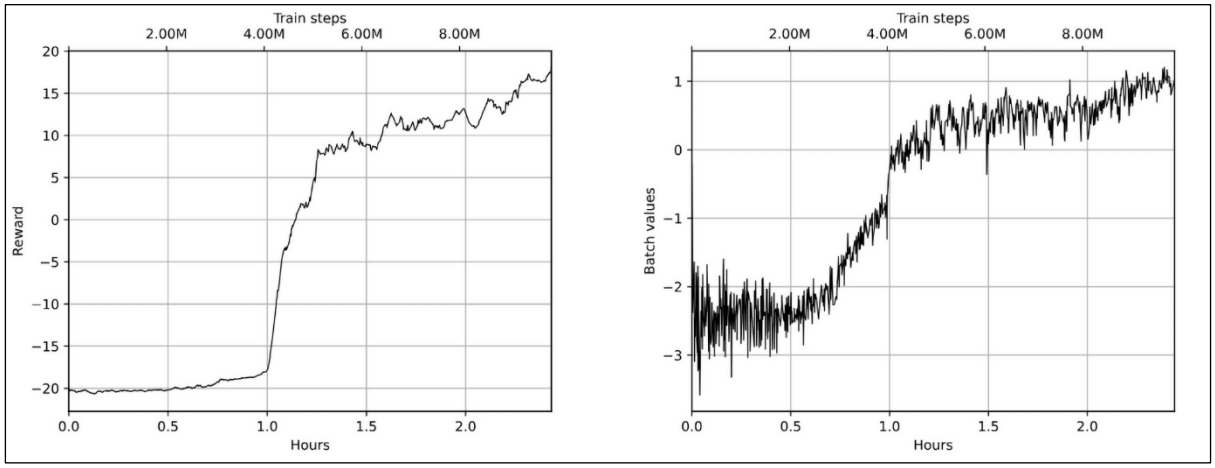


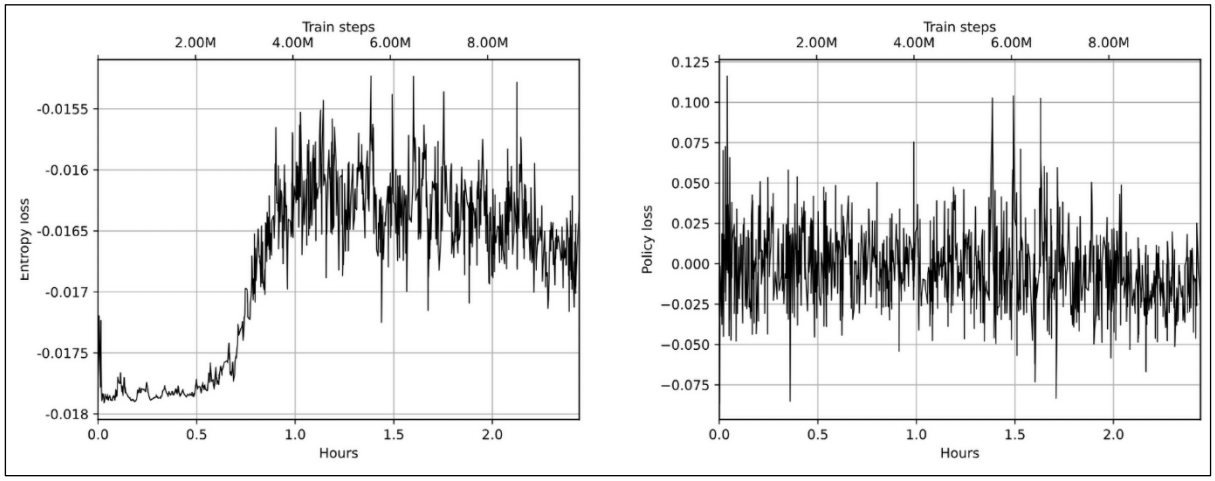


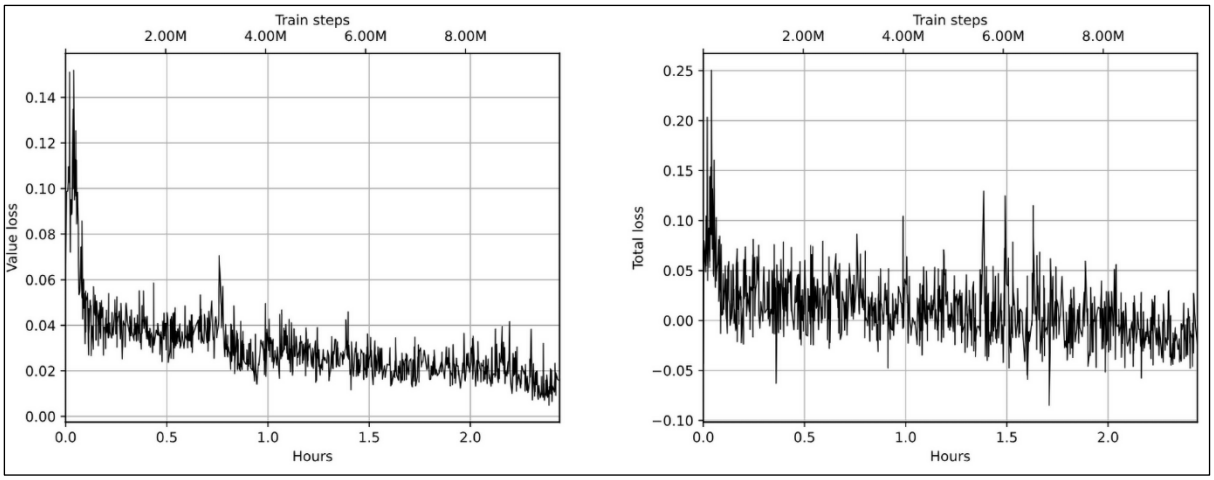

No baseline
Baseline
L2 of gradients (log)
10^{-1}
10^{-2}
20.0K
40.0K
60.0K
80.0K
100K
Train steps
No baseline
Baseline
Max of gradients
10^{0}
10^{-1}
20.0K
40.0K
60.0K
80.0K
100K
Train steps

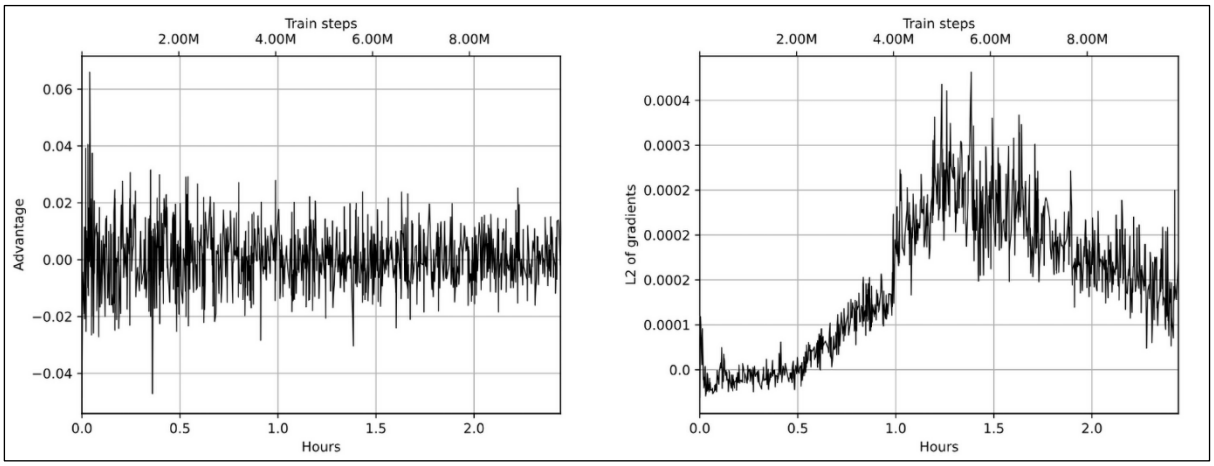

Observations
Policy net
(actor)
$\pi(a|s)$
Value net
(critic)
$V(s)$

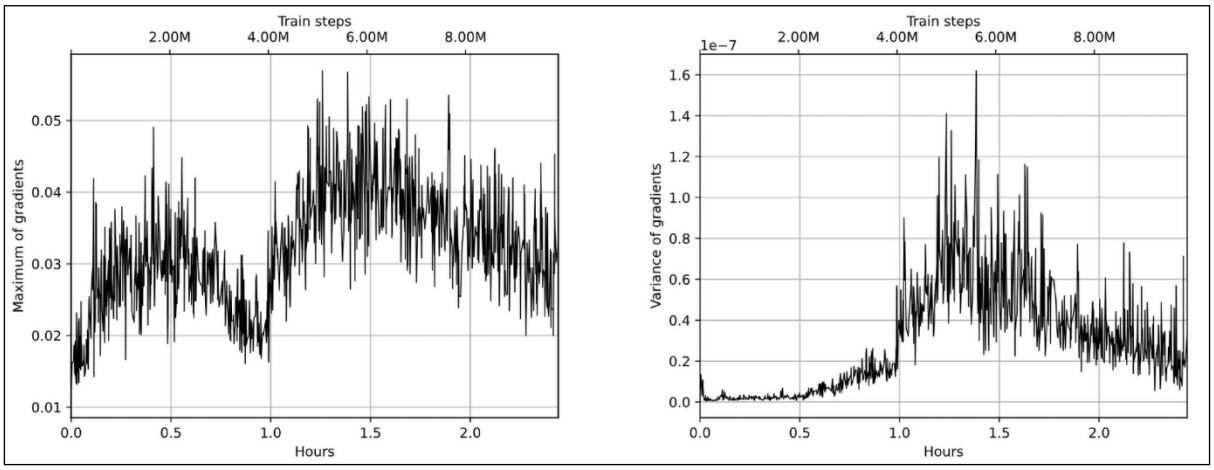

Observations
Common net
(body)
Policy net
(actor)
$\pi(a|s)$
Value net
(critic)
$V(s)$

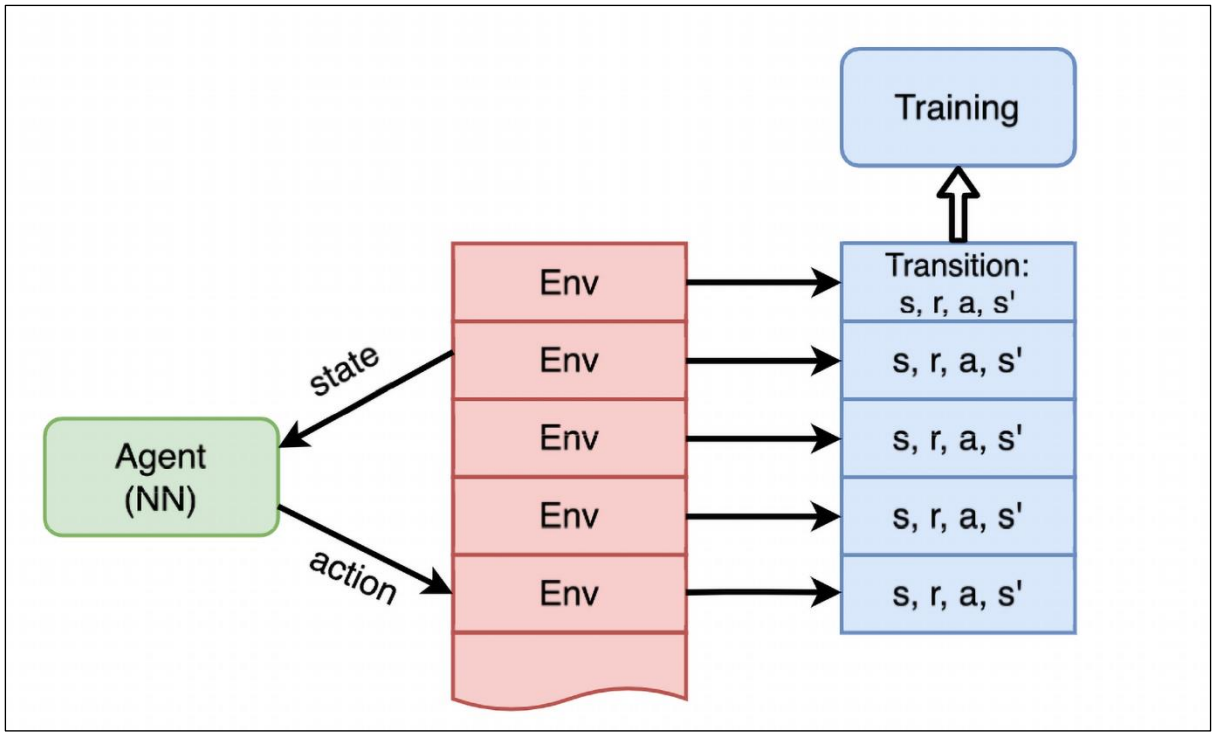

Train steps
2.00M
4.00M
6.00M
8.00M
Reward
Hours
Batch values

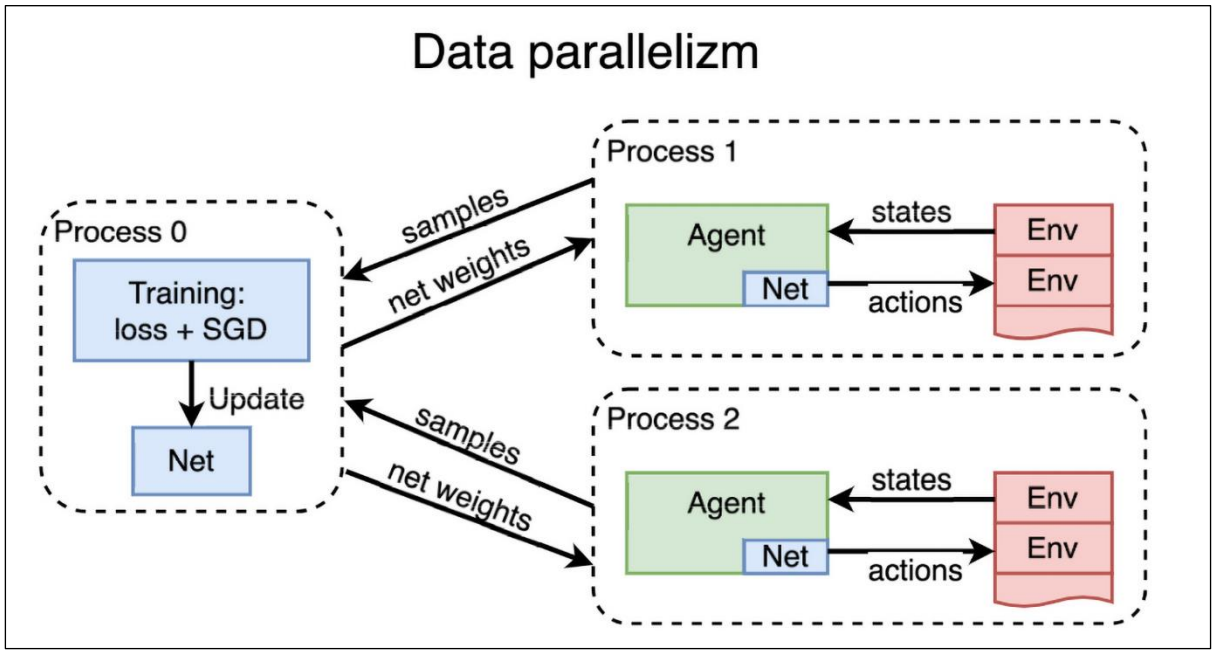

Train steps
2.00M
4.00M
6.00M
8.00M
Entropy loss
Hours
Policy loss

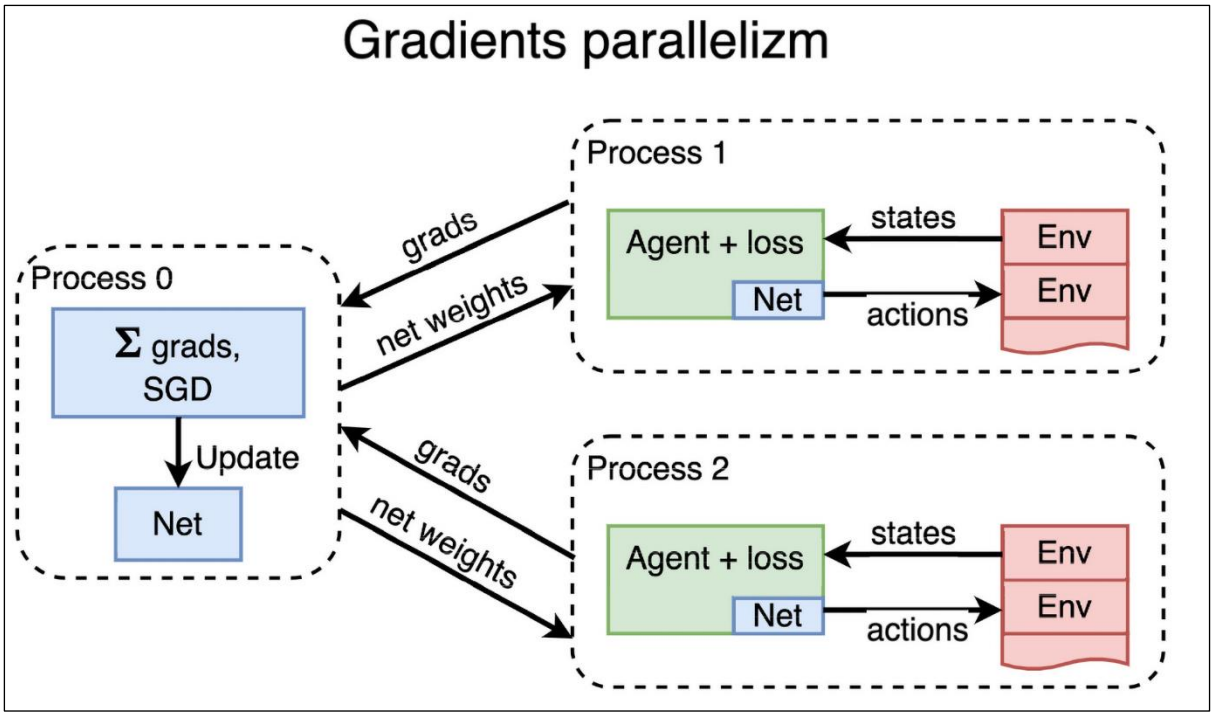

Train steps
2.00M
4.00M
6.00M
8.00M
Value loss
Hours
Total loss

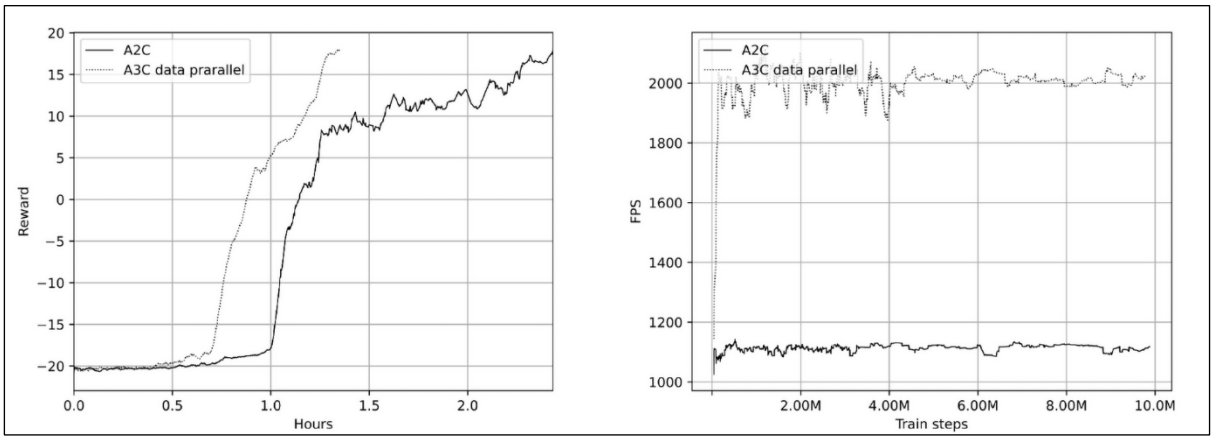

Train steps
Advantage
L2 of gradients
Hours

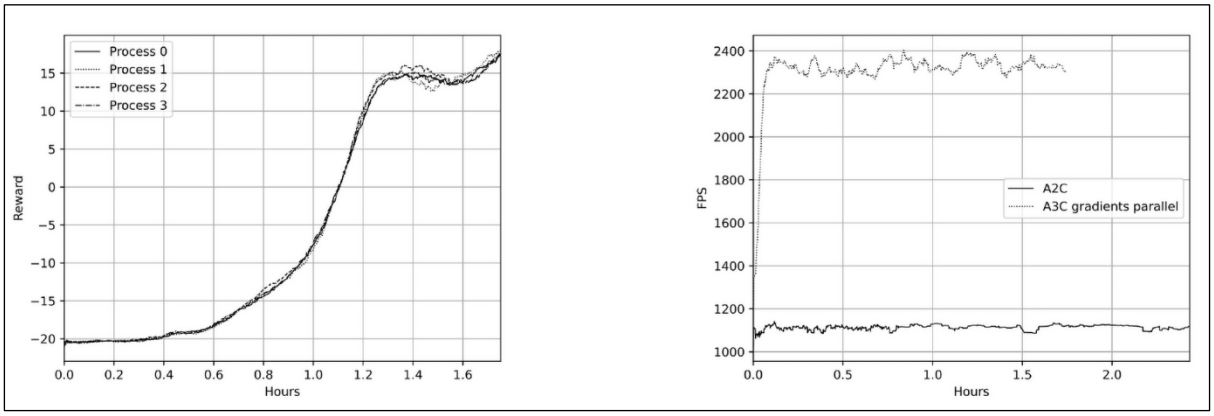

Train steps
Maximum of gradients
1e−7
Variance of gradients
Hours


Training
Env
Env
Env
Env
Env
Transition:
s, r, a, s'
s, r, a, s'
s, r, a, s'
s, r, a, s'
s, r, a, s'
state
action
Agent
(NN)

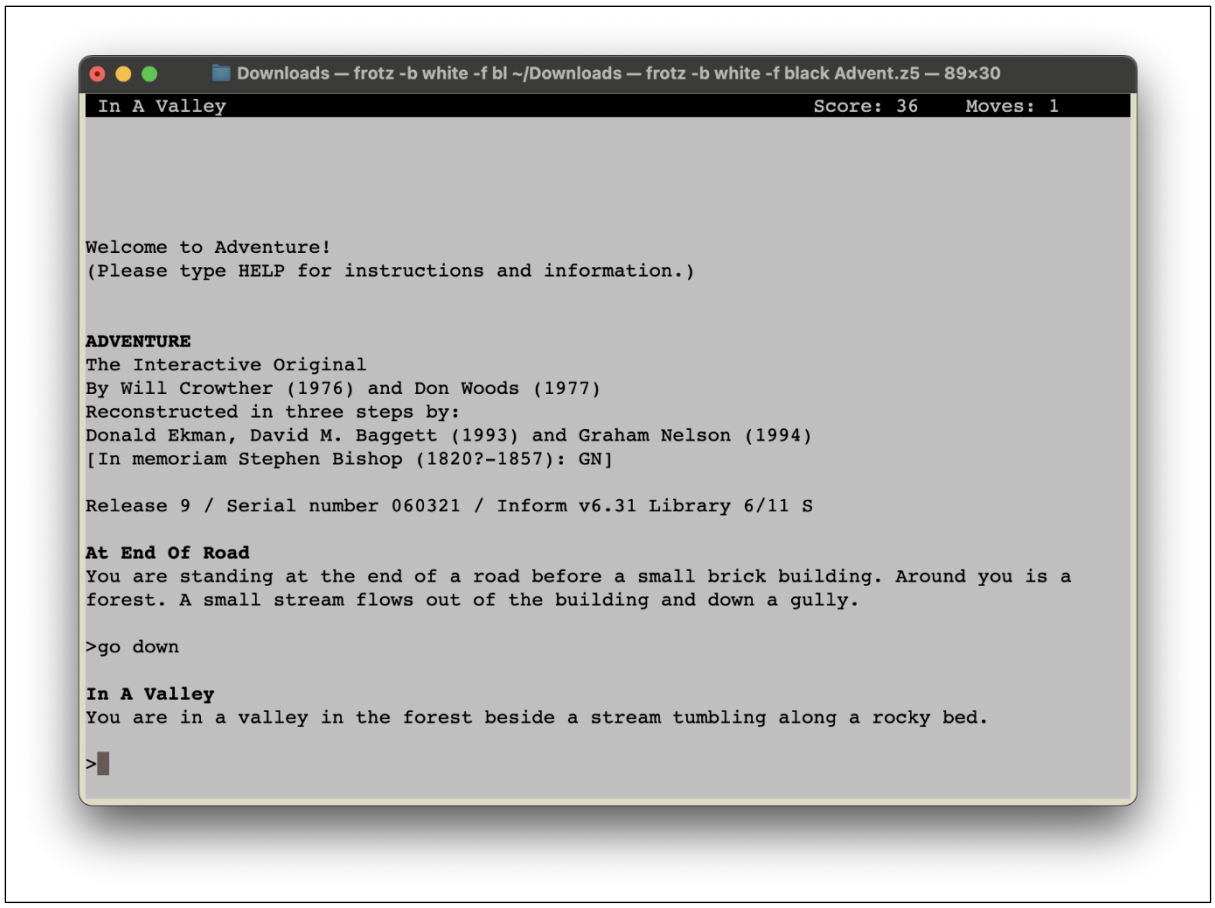

Data parallelizm
Process 0
Training:
loss + SGD
Update
Net
samples
net weights
Process 1
Agent
Net
states
actions
Env
Env
Process 2
Agent
Net
states
actions
Env
Env

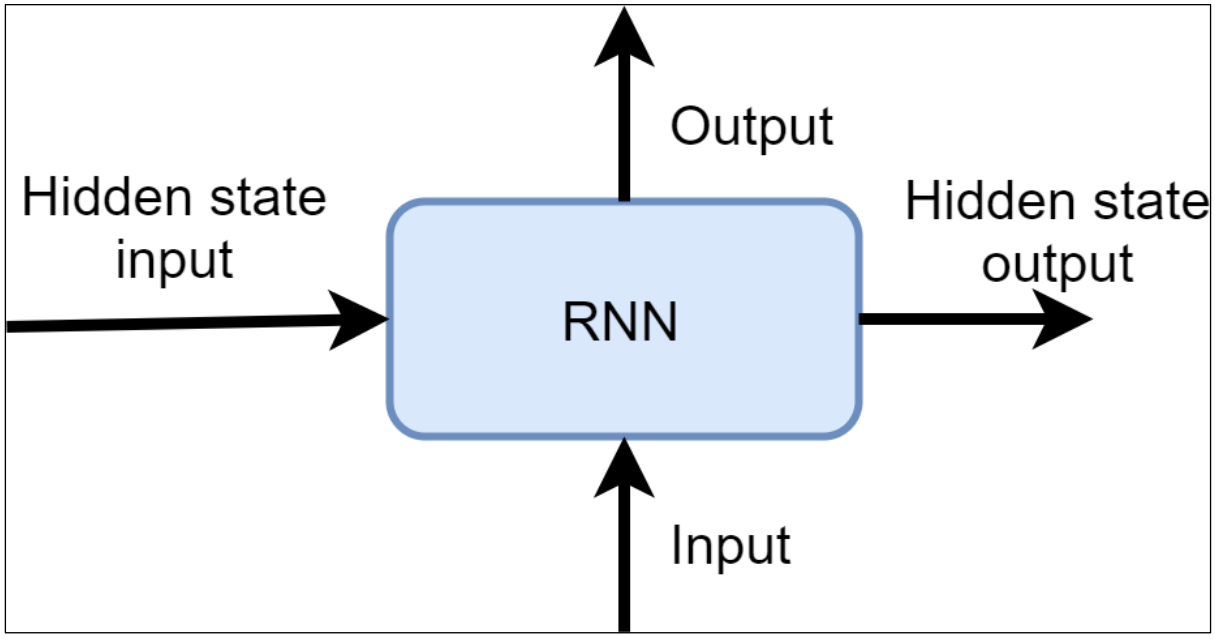

Gradients parallelizm
Process 1
Agent + loss
Net
states
actions
Env
Env
grads
net weights
Process 0
Σ grads,
SGD
Update
Net
grads
net weights
Process 2
Agent + loss
Net
states
actions
Env
Env

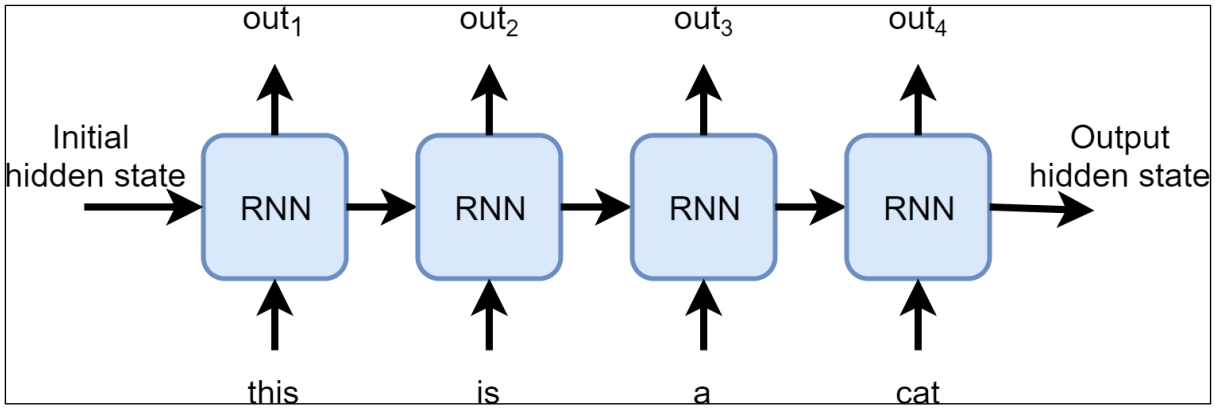

A2C
A3C data prarallel
Reward
Hours
A2C
A3C data parallel
FPS
Train steps

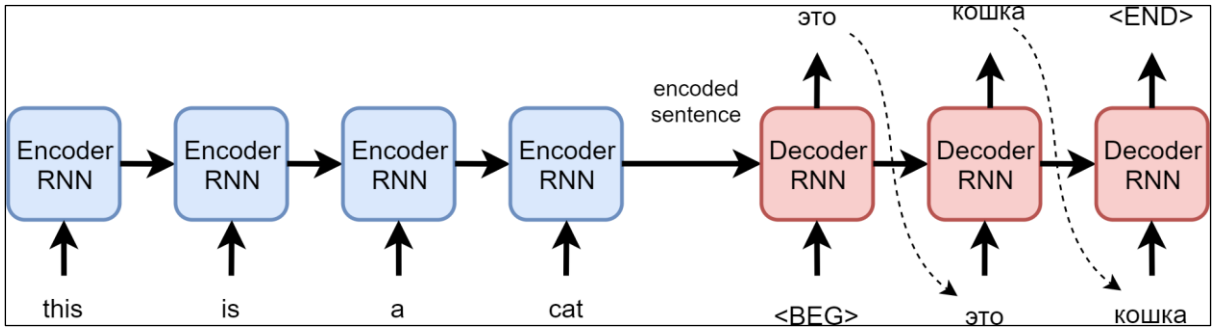

Process 0
Process 1
Process 2
Process 3
Reward
Hours
A2C
A3C gradients parallel
FPS
Hours

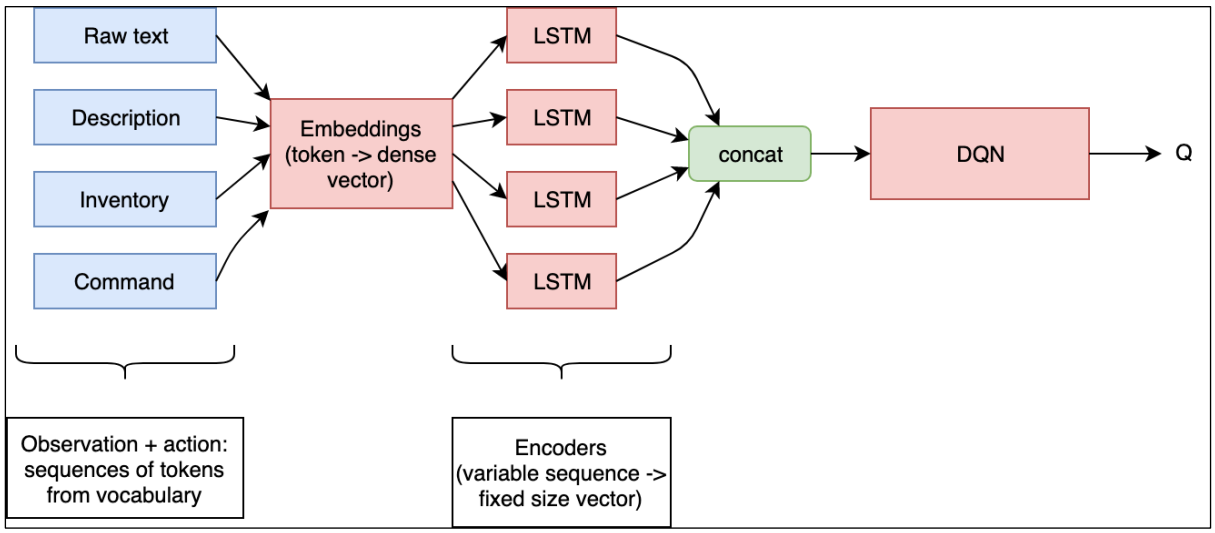
Chapter 13: The TextWorld Environment

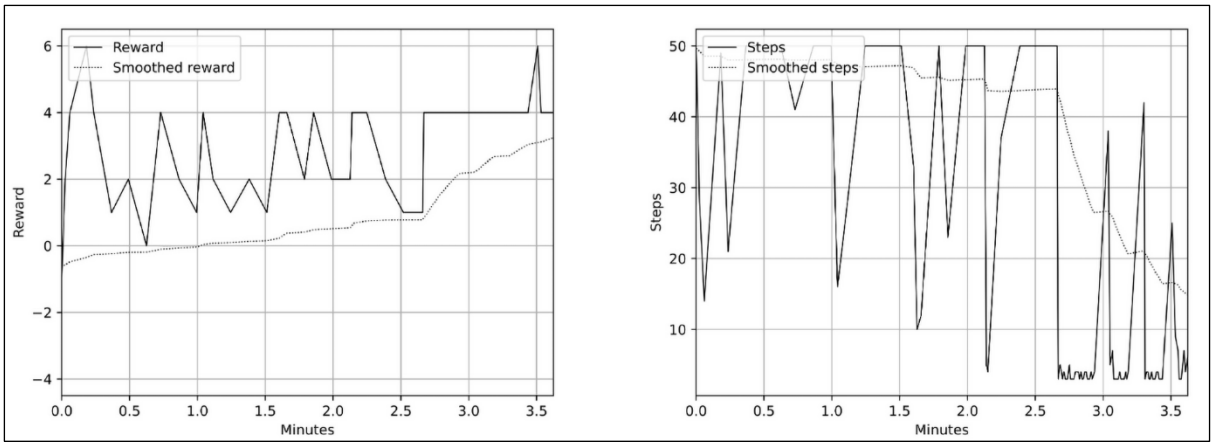


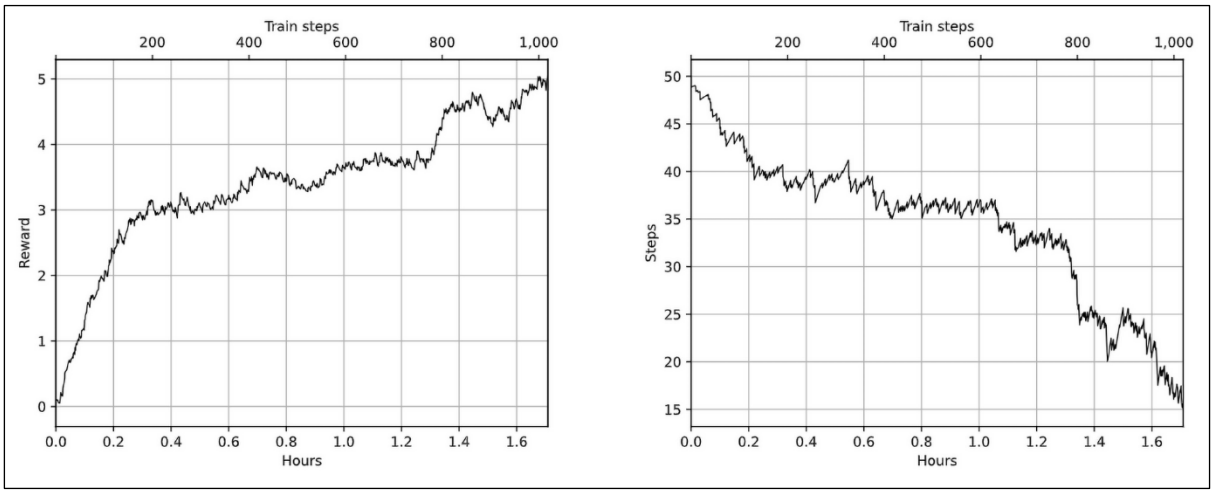


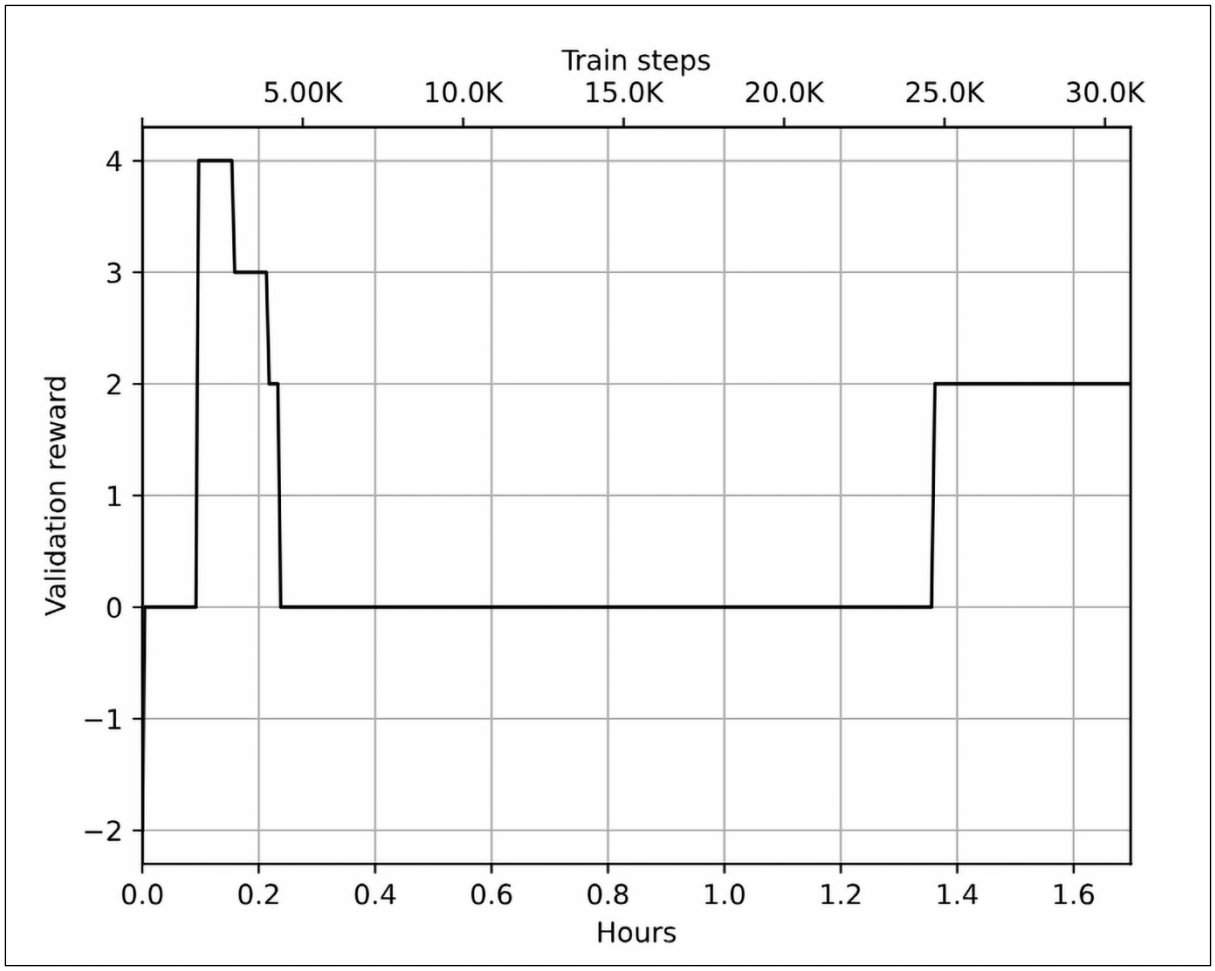

out1
out2
out3
out4
Initial hidden state
RNN
RNN
RNN
RNN
Output hidden state
this
is
a
cat

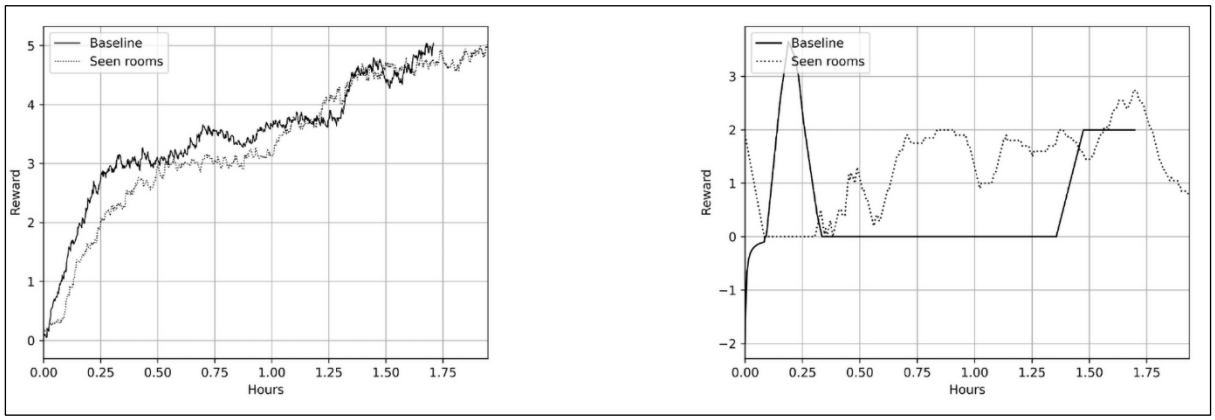

Encoder RNN
Encoder RNN
Encoder RNN
Encoder RNN
encoded sentence
Decoder RNN
Decoder RNN
Decoder RNN
это
кошка
<END>
this
is
a
cat
<BEG>
это
кошка

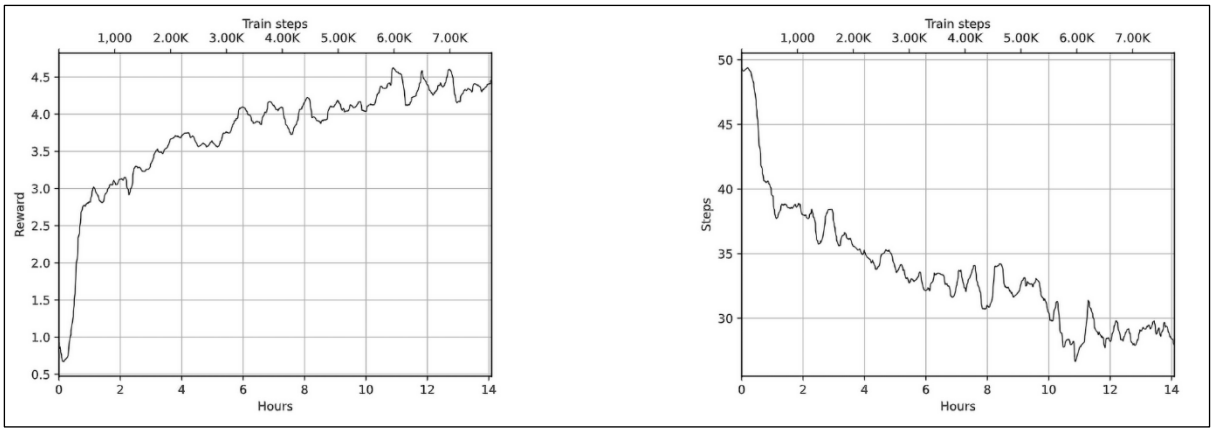

Raw text
Description
Inventory
Command
Embeddings (token -> dense vector)
LSTM
LSTM
LSTM
LSTM
concat
DQN
Q
Observation + action: sequences of tokens from vocabulary
Encoders (variable sequence -> fixed size vector)

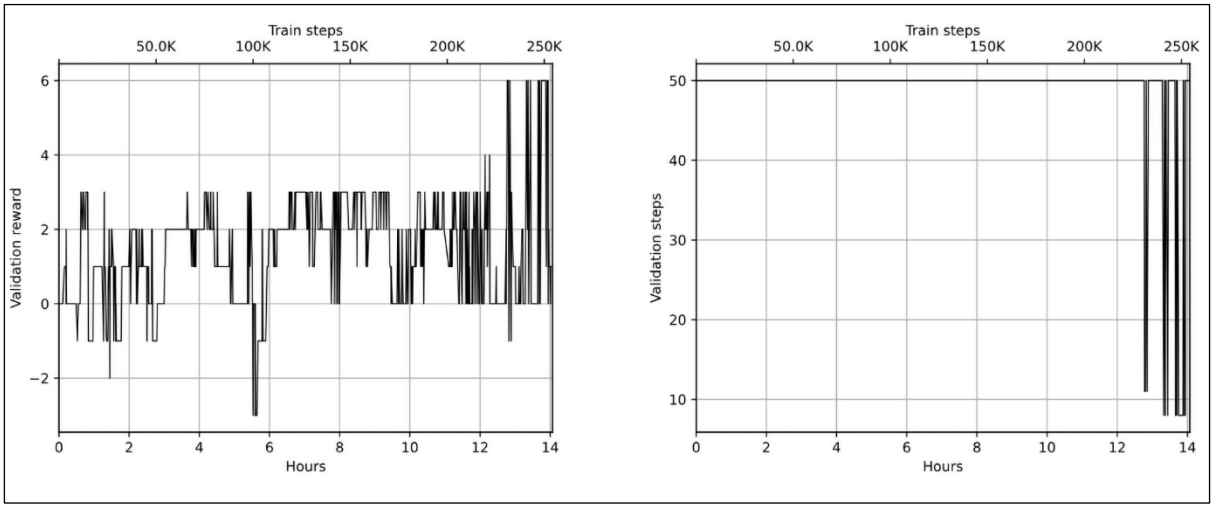

Reward
Smoothed reward
Reward
Minutes
Steps
Smoothed steps
Steps
Minutes

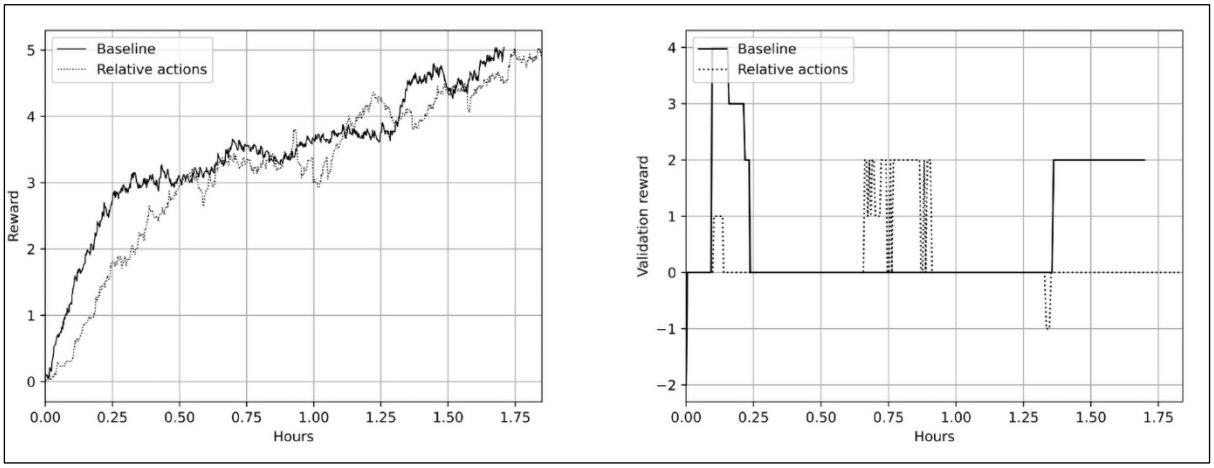

Train steps
Reward
Hours
Train steps
Steps
Hours

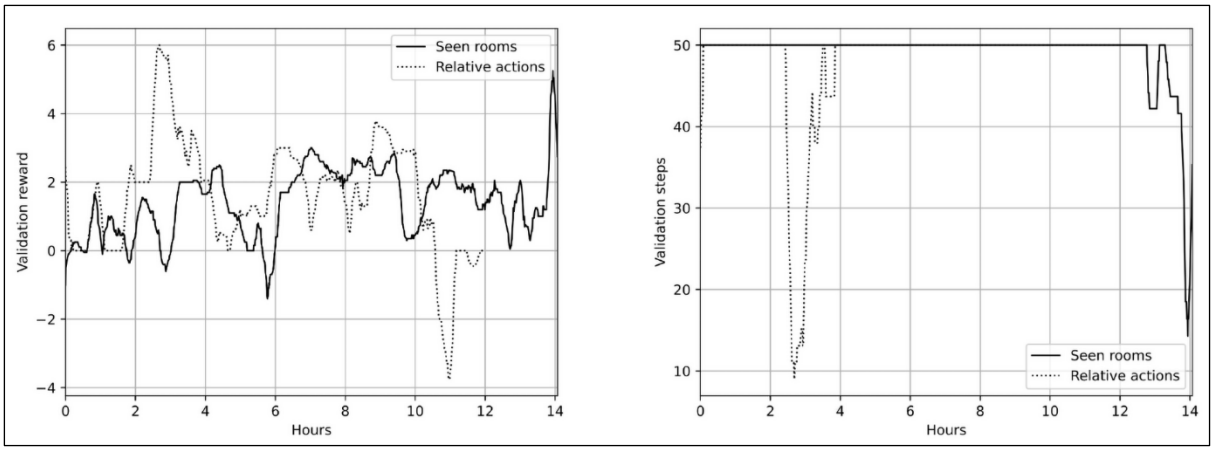

Train steps
5.00K
10.0K
15.0K
20.0K
25.0K
30.0K
Validation reward
4
3
2
1
0
−1
−2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Hours

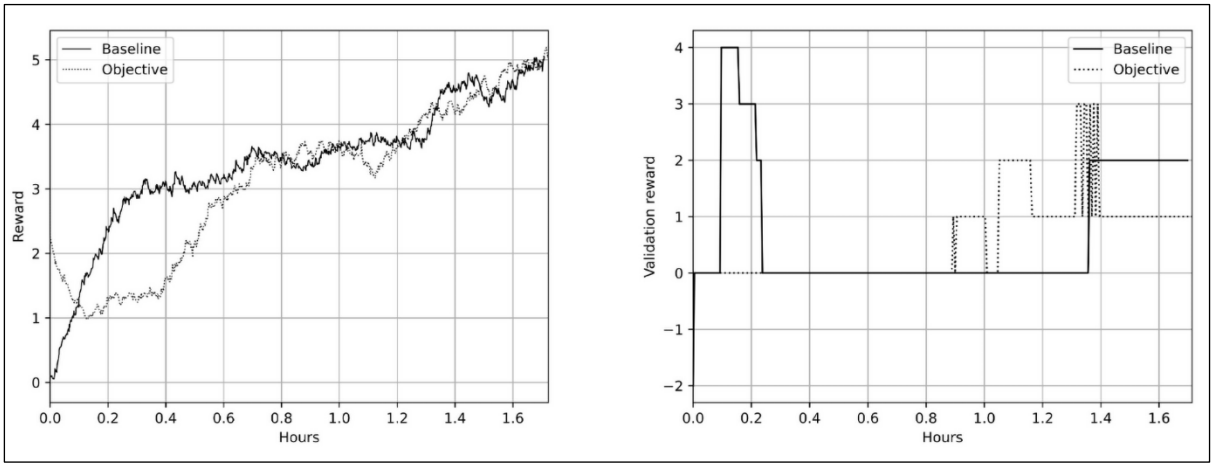

Baseline
Seen rooms
Reward
Hours
Baseline
Seen rooms
Reward
Hours

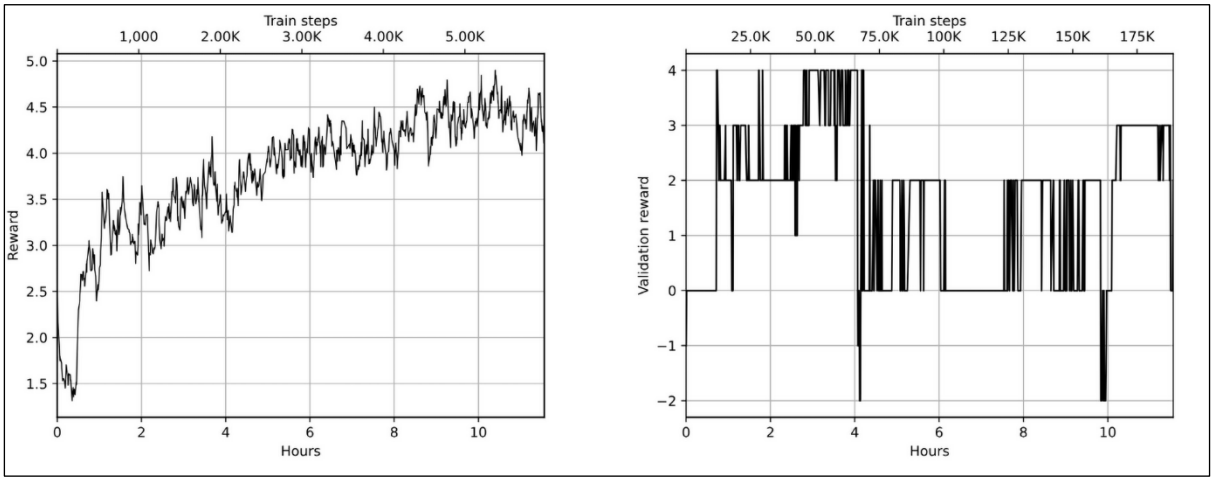

Train steps
1,000
2.00K
3.00K
4.00K
5.00K
6.00K
7.00K
Reward
Hours
Steps

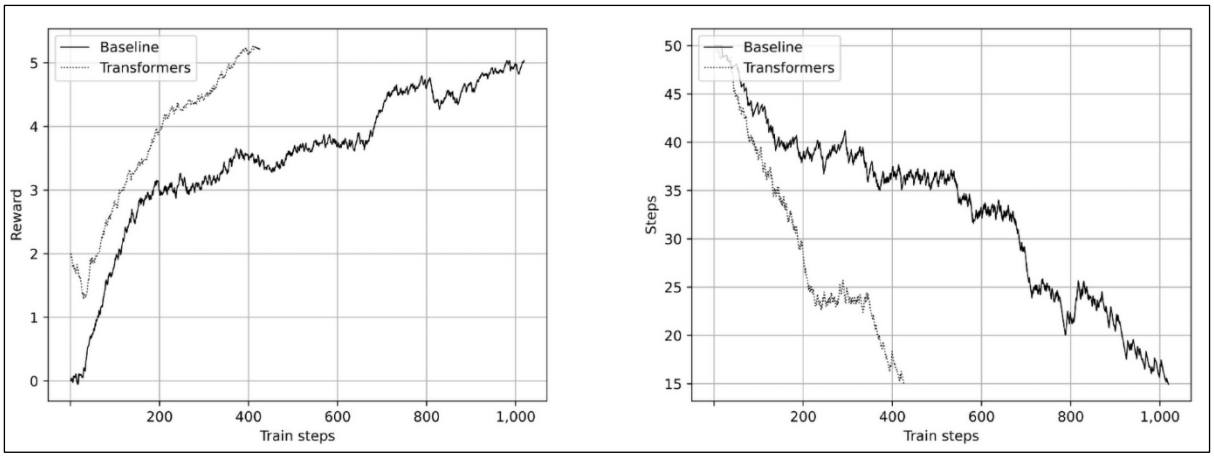

Train steps
50.0K
100K
150K
200K
250K
Validation reward
Validation steps
Hours


Baseline
Relative actions
Reward
Validation reward
Hours

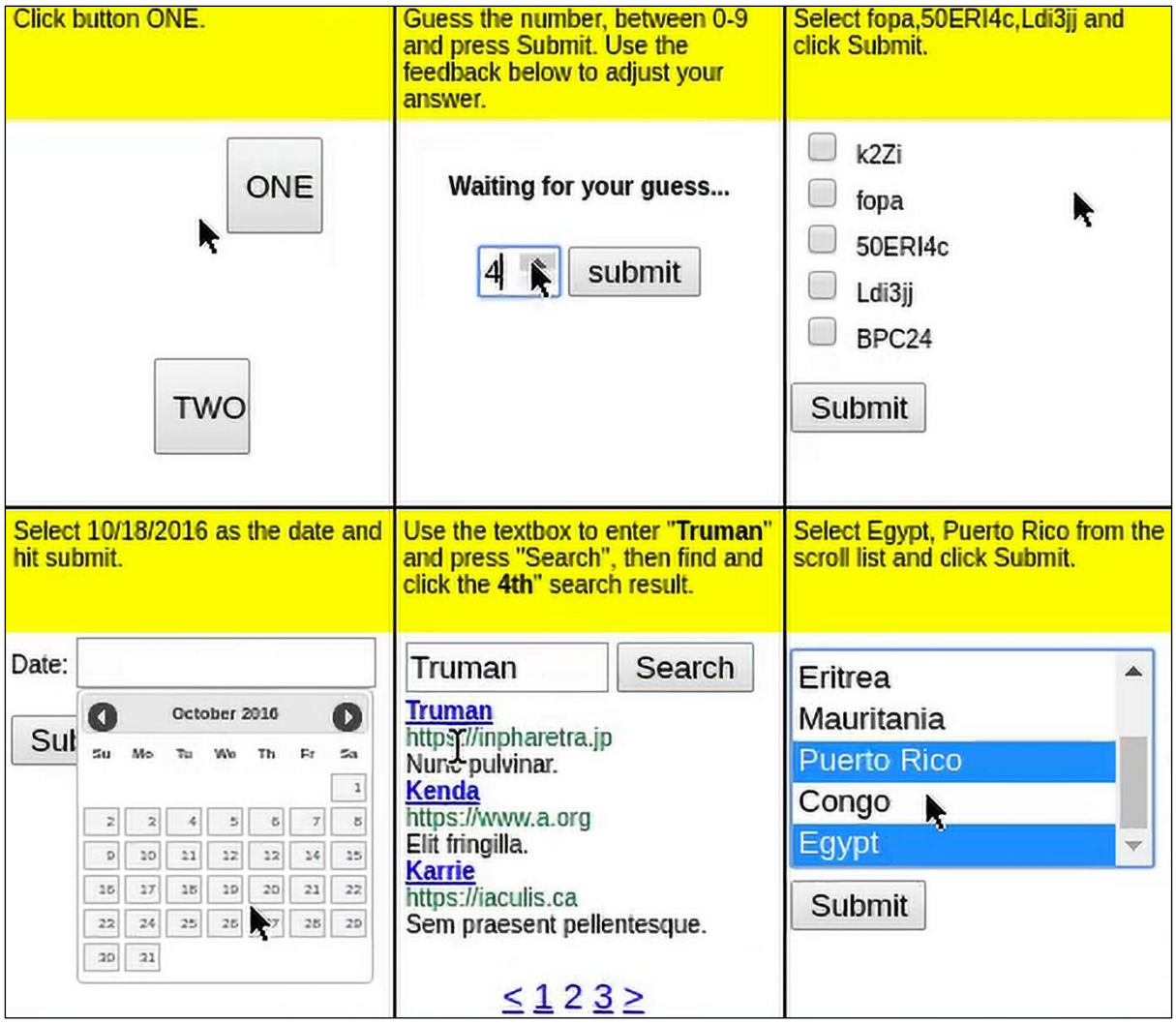

Seen rooms
Relative actions
Validation reward
Hours
Validation steps
Seen rooms
Relative actions
Hours

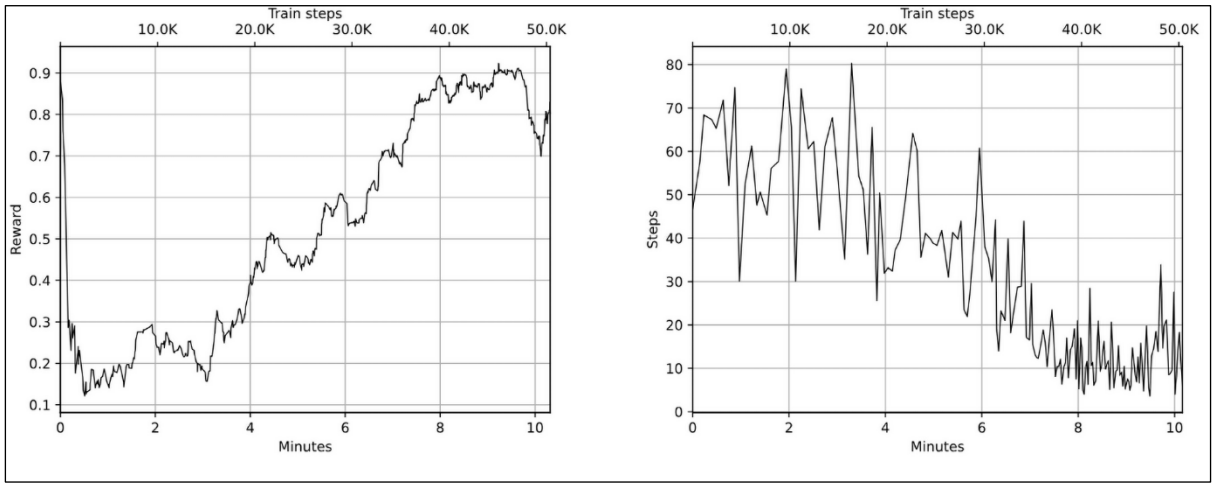

Baseline
Objective
Reward
Hours
Baseline
Objective
Validation reward
Hours

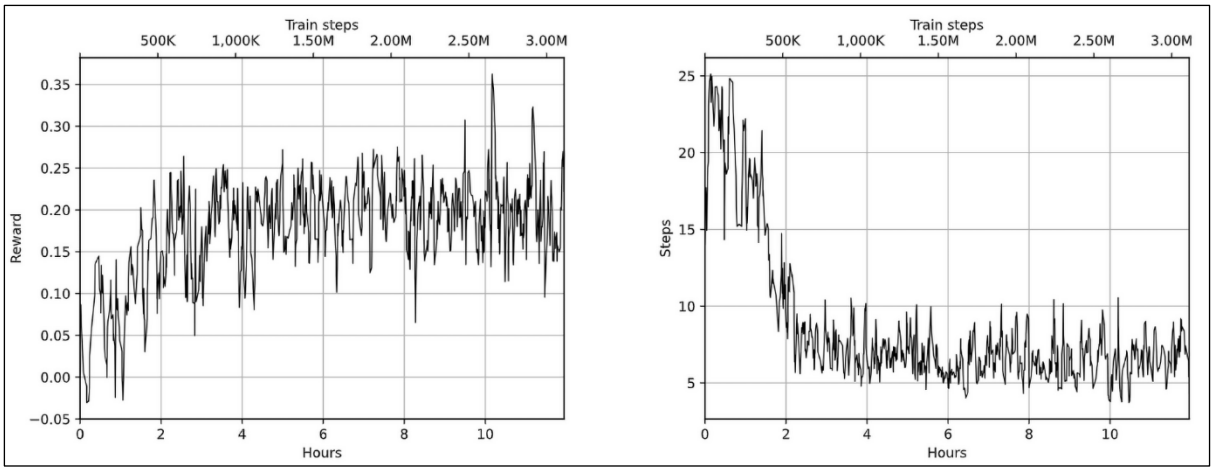

Train steps
1,000
2.00K
3.00K
4.00K
5.00K
Reward
Hours
Train steps
25.0K
50.0K
75.0K
100K
125K
150K
175K
Validation reward
Hours

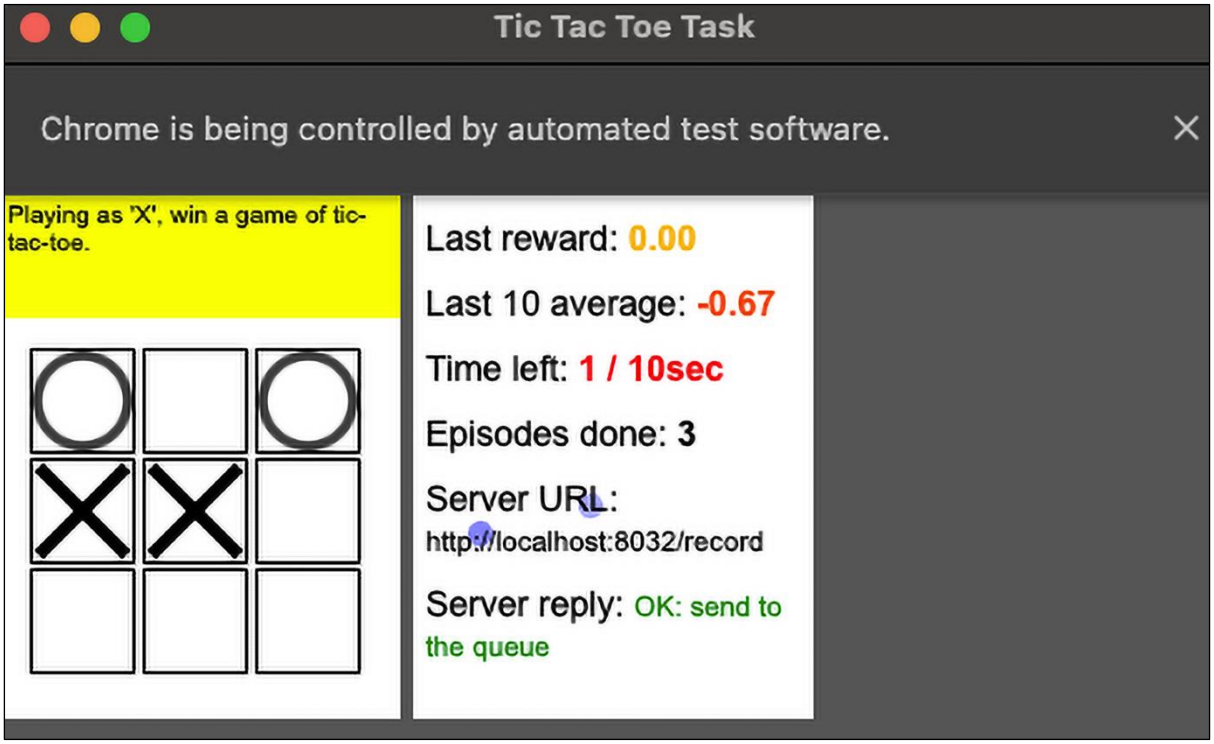

Baseline
Transformers
Reward
Train steps
0
1
2
3
4
5
200
400
600
800
1,000
Baseline
Transformers
Steps
Train steps
15
20
25
30
35
40
45
50
200
400
600
800
1,000

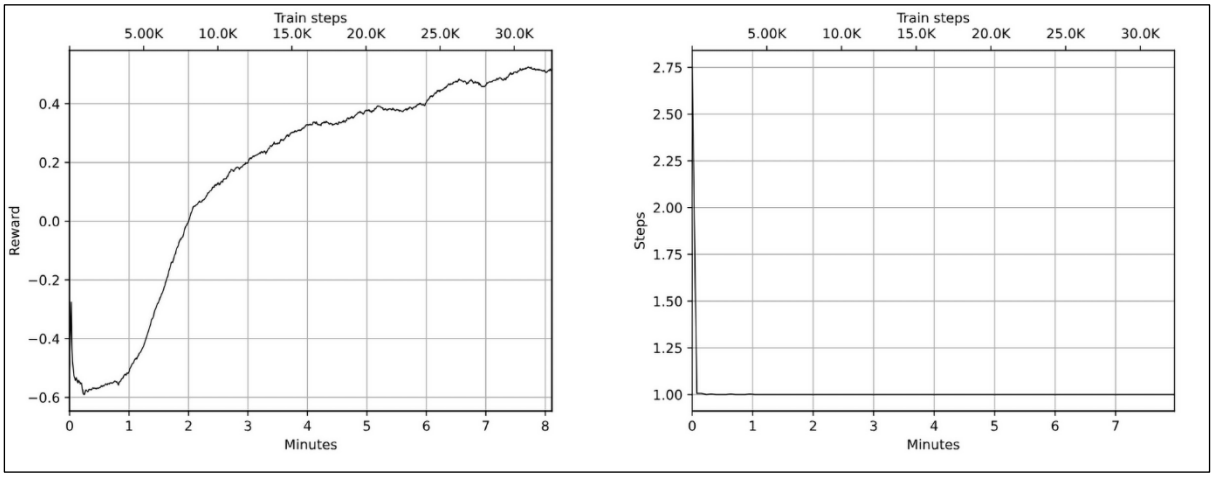
Chapter 14: Web Navigation

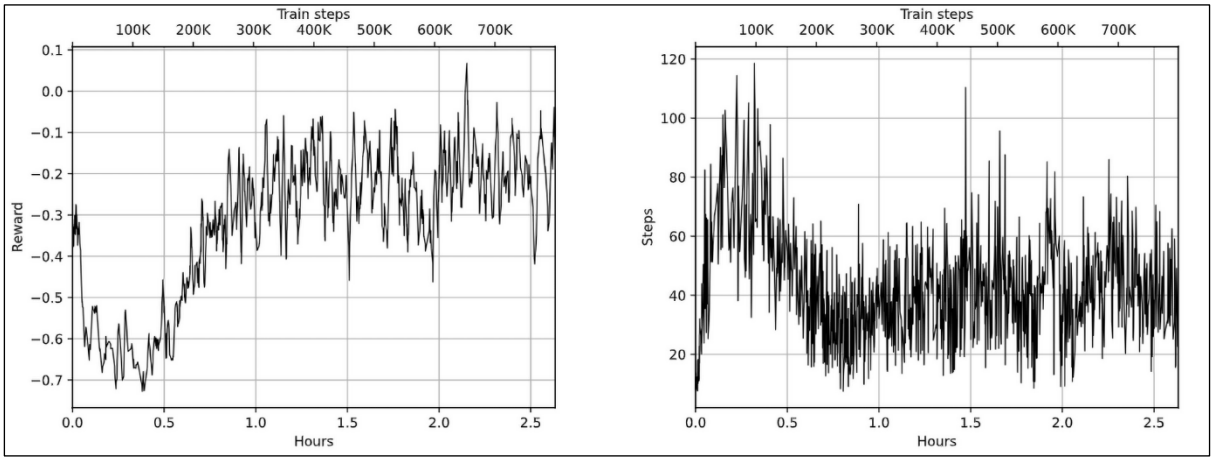




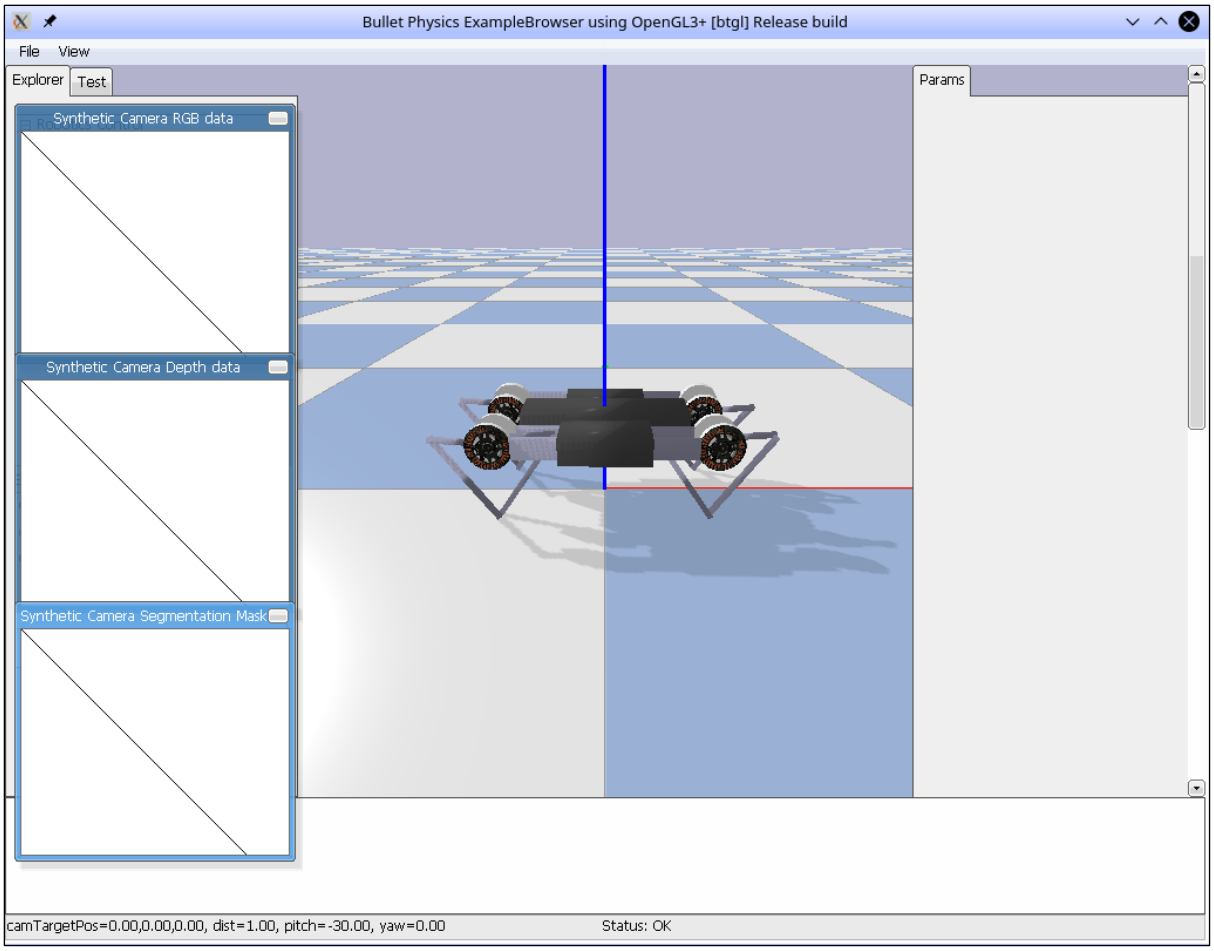

Train steps
500K
1,000K
1.50M
2.00M
2.50M
3.00M
Reward
Hours
Steps

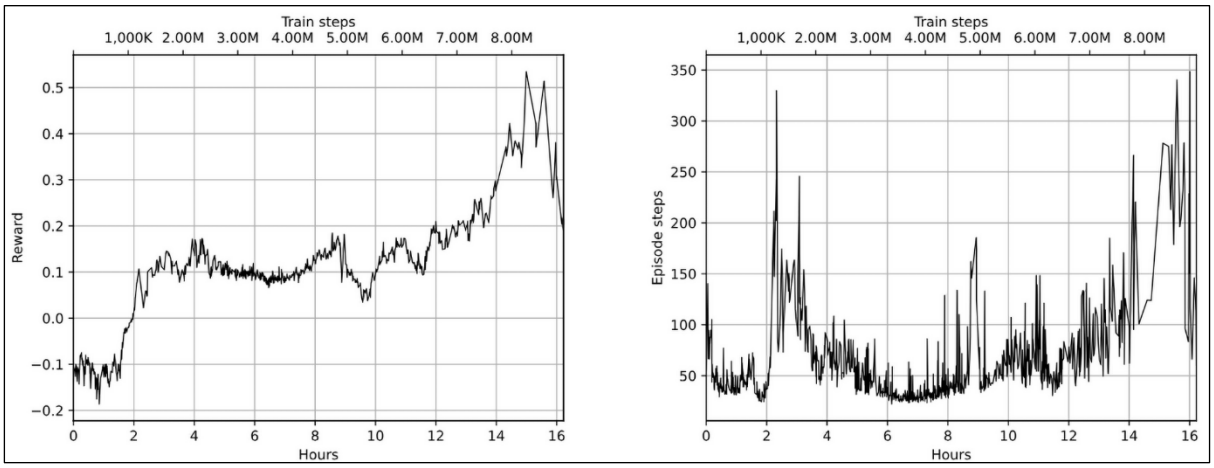

Tic Tac Toe Task
Chrome is being controlled by automated test software.
Playing as 'X', win a game of tic-tac-toe.
Last reward: 0.00
Last 10 average: -0.67
Time left: 1 / 10sec
Episodes done: 3
Server URL:
http://localhost:8032/record
Server reply: OK: send to the queue

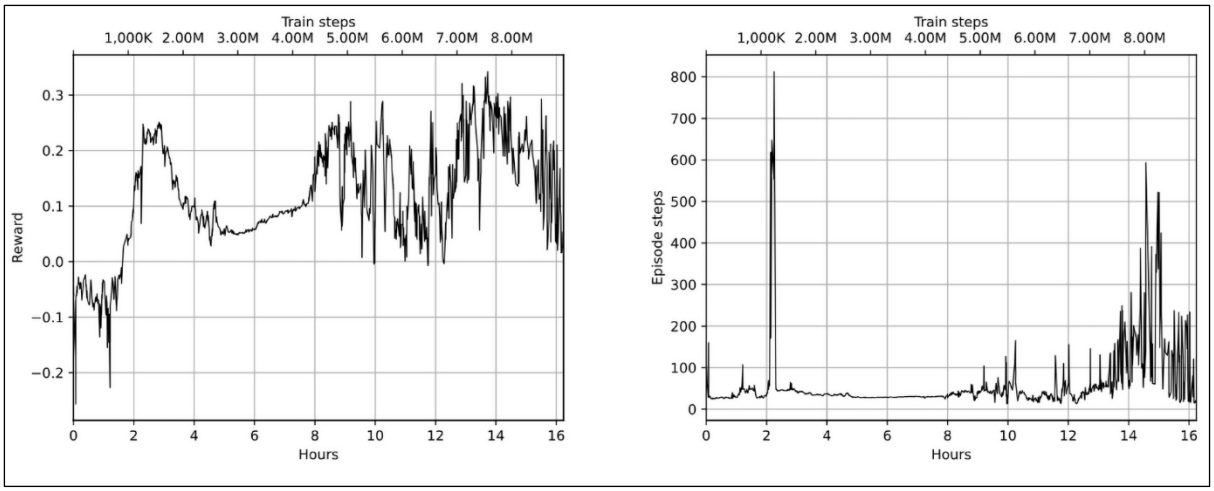

Train steps
5.00K
10.0K
15.0K
20.0K
25.0K
30.0K
Reward
Minutes
Steps

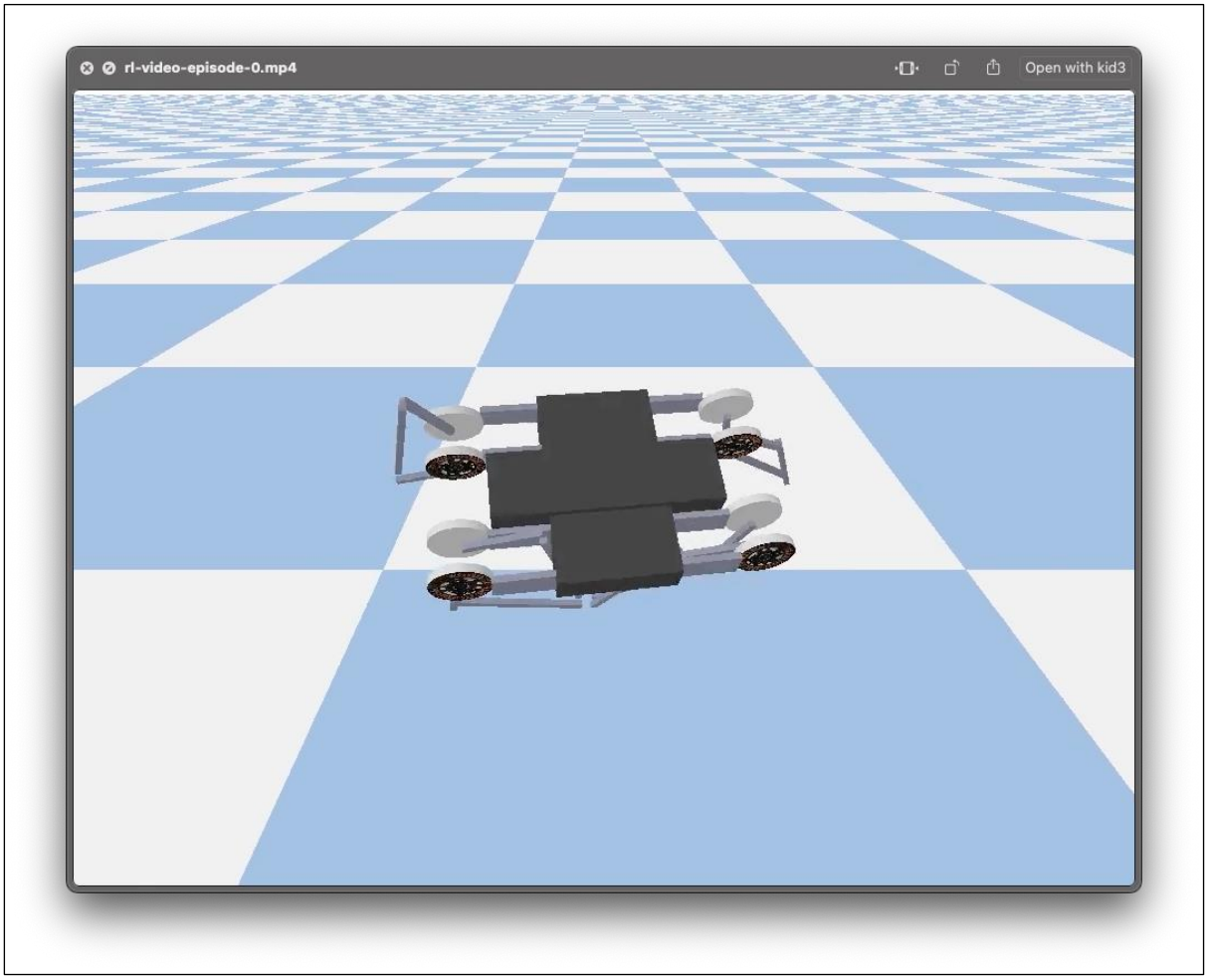

Train steps
100K
200K
300K
400K
500K
600K
700K
Reward
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
−0.7
0.0
0.5
1.0
1.5
2.0
2.5
Hours
Train steps
100K
200K
300K
400K
500K
600K
700K
Steps
120
100
80
60
40
20
0.0
0.5
1.0
1.5
2.0
2.5
Hours

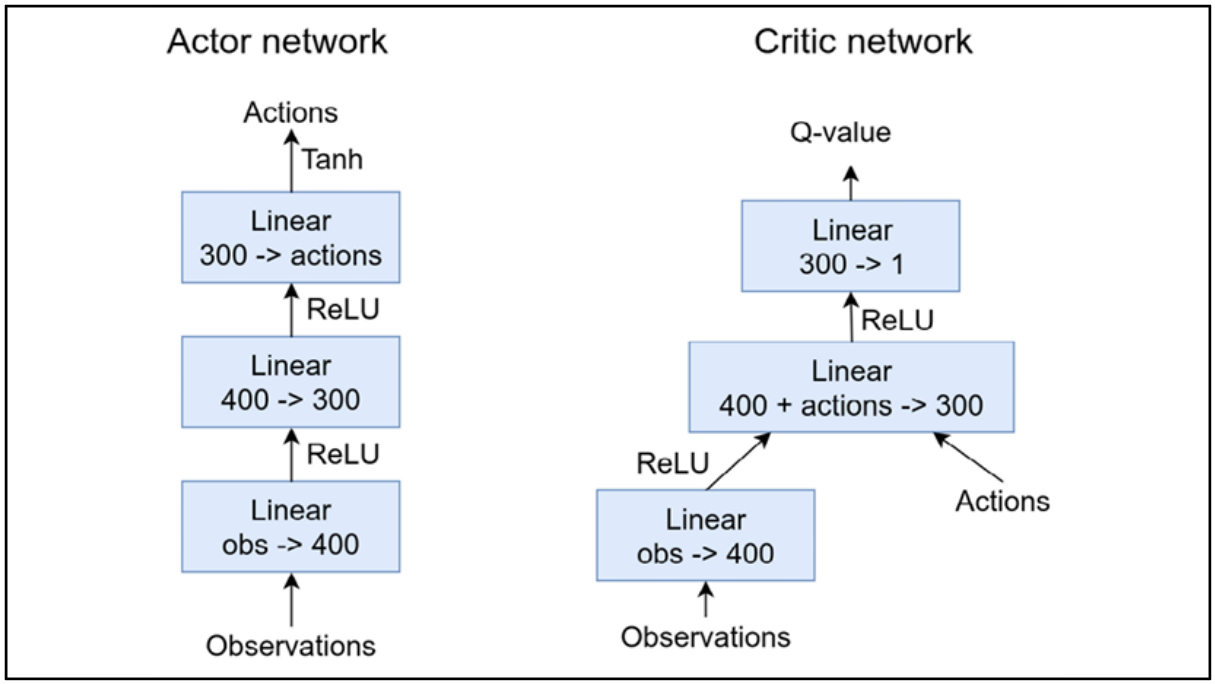
Chapter 15: Continuous Action Space

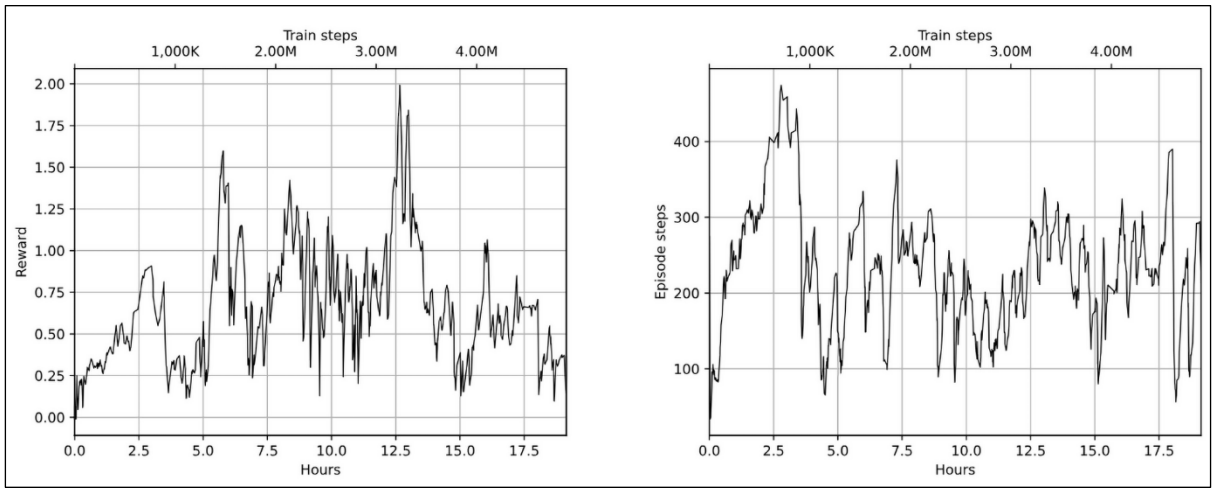


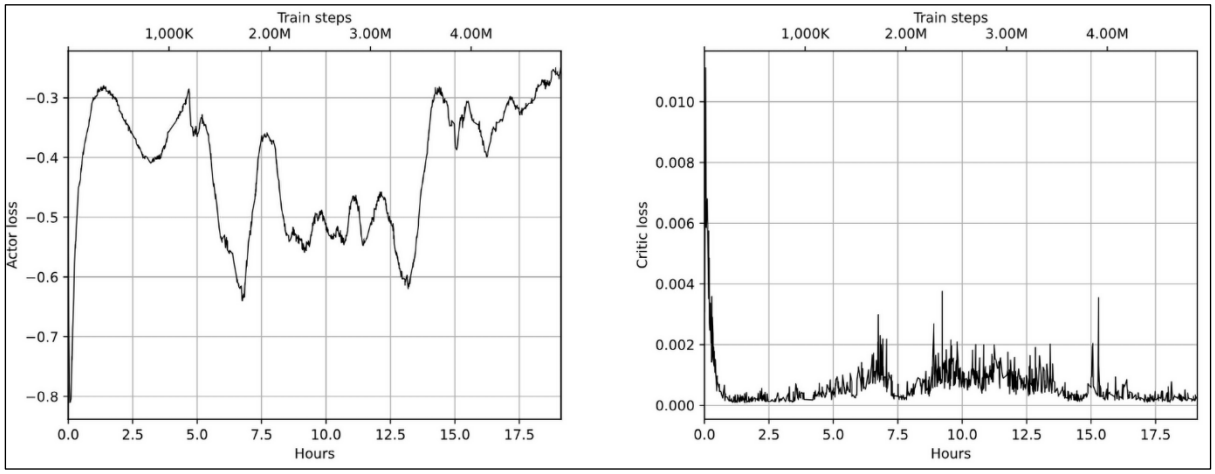


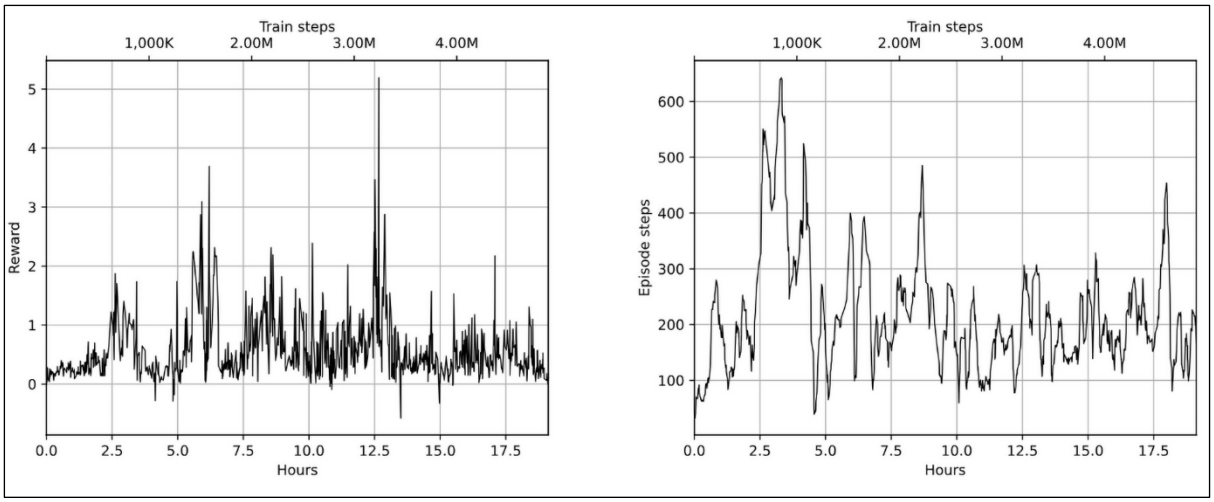

Train steps
1,000K 2.00M 3.00M 4.00M 5.00M 6.00M 7.00M 8.00M
Reward
0.3
0.2
0.1
0.0
−0.1
−0.2
0 2 4 6 8 10 12 14 16
Hours
Train steps
1,000K 2.00M 3.00M 4.00M 5.00M 6.00M 7.00M 8.00M
Episode steps
800
700
600
500
400
300
200
100
0
0 2 4 6 8 10 12 14 16
Hours

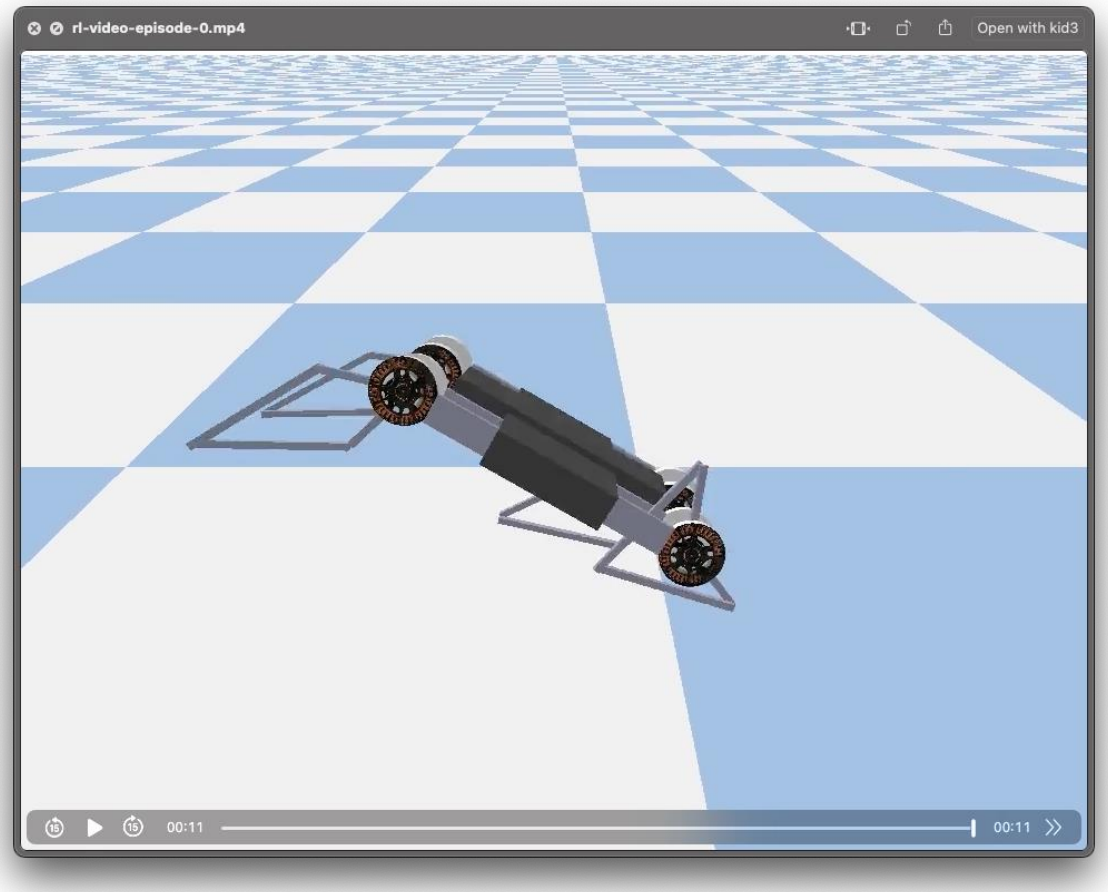

rl-video-episode-0.mp4
Open with kid3

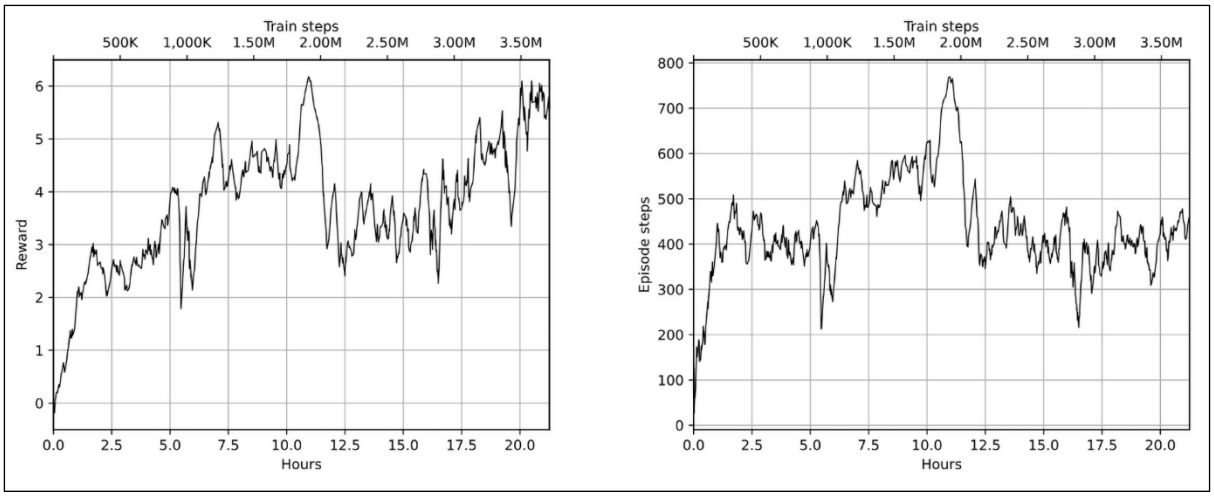

Actor network
Actions
Tanh
Linear
300 -> actions
ReLU
Linear
400 -> 300
ReLU
Linear
obs -> 400
Observations
Critic network
Q-value
Linear
300 -> 1
ReLU
Linear
400 + actions -> 300
ReLU
Actions
Linear
obs -> 400
Observations

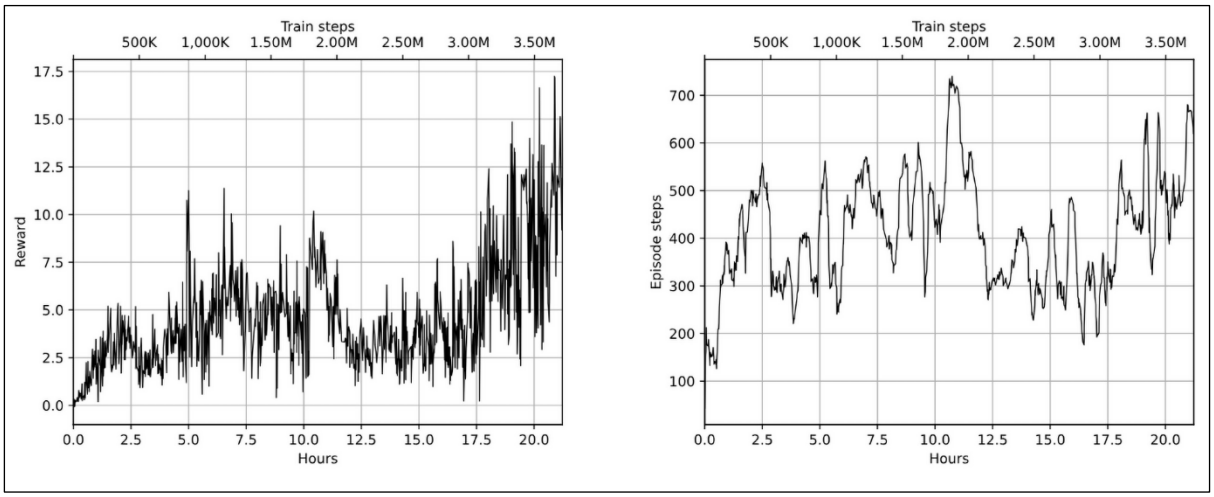

Train steps
1,000K
2.00M
3.00M
4.00M
Reward
Hours
Episode steps

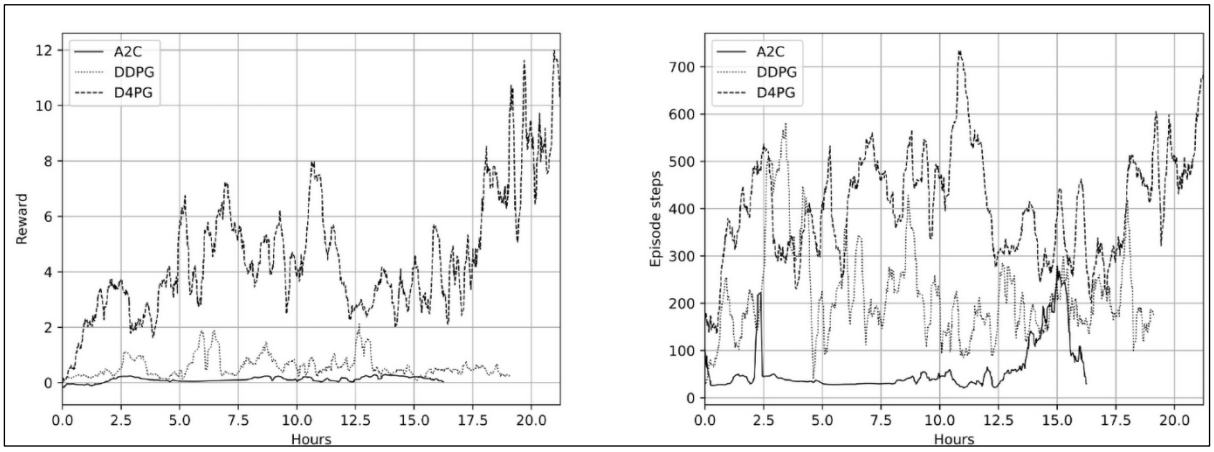

Train steps
1,000K
2.00M
3.00M
4.00M
Actor loss
Critic loss
Hours


Train steps
1,000K
2.00M
3.00M
4.00M
Reward
Hours
Train steps
1,000K
2.00M
3.00M
4.00M
Episode steps
Hours

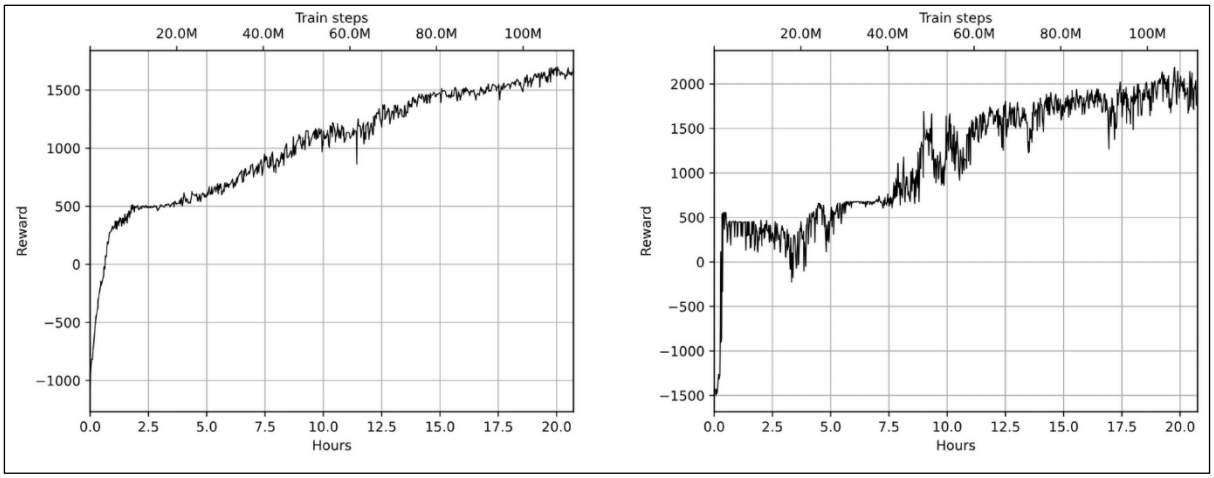

rl-video-episode-0.mp4
Open with kid3
00:11
00:11

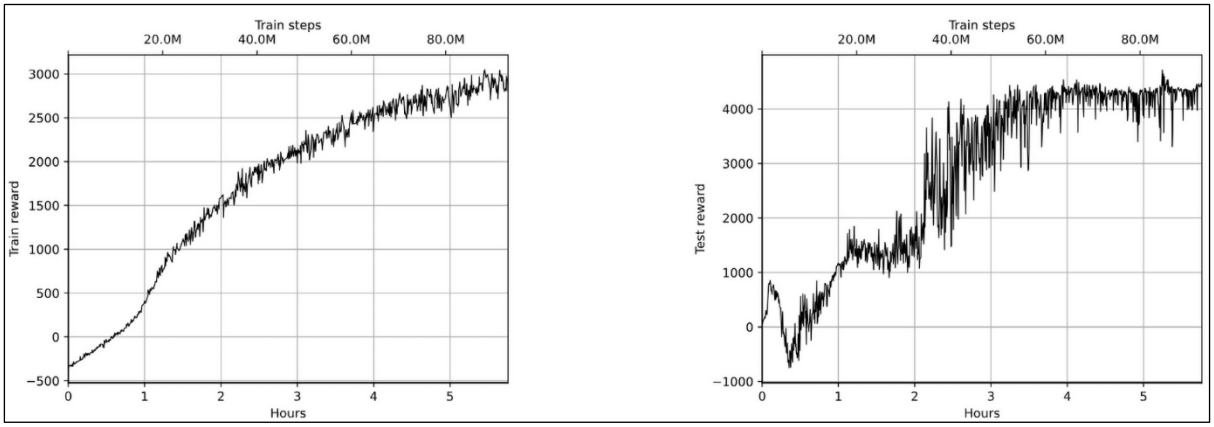

Train steps
500K 1,000K 1.50M 2.00M 2.50M 3.00M 3.50M
Reward
Hours
Train steps
500K 1,000K 1.50M 2.00M 2.50M 3.00M 3.50M
Episode steps
Hours

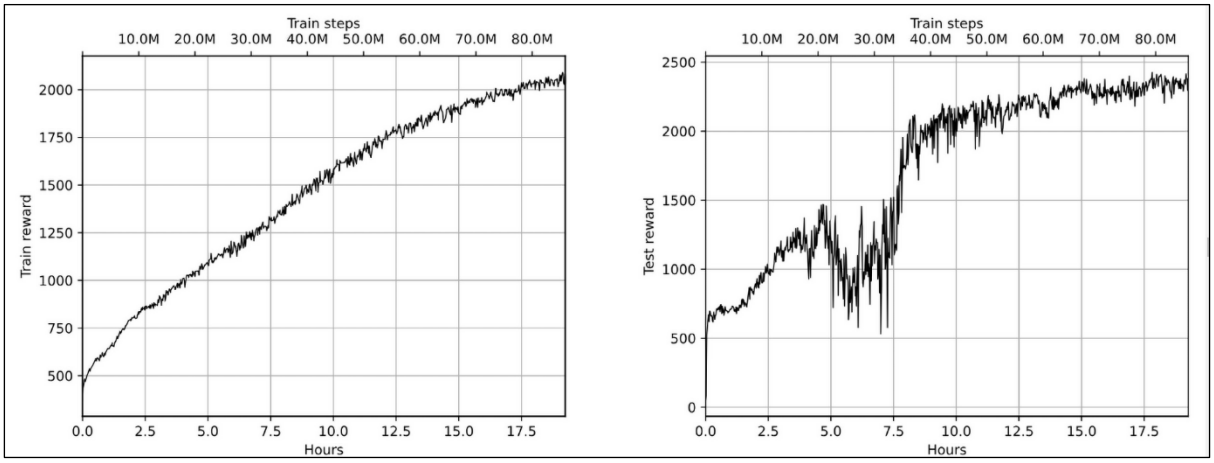

Train steps
500K 1,000K 1.50M 2.00M 2.50M 3.00M 3.50M
Reward
Hours
Train steps
500K 1,000K 1.50M 2.00M 2.50M 3.00M 3.50M
Episode steps
Hours

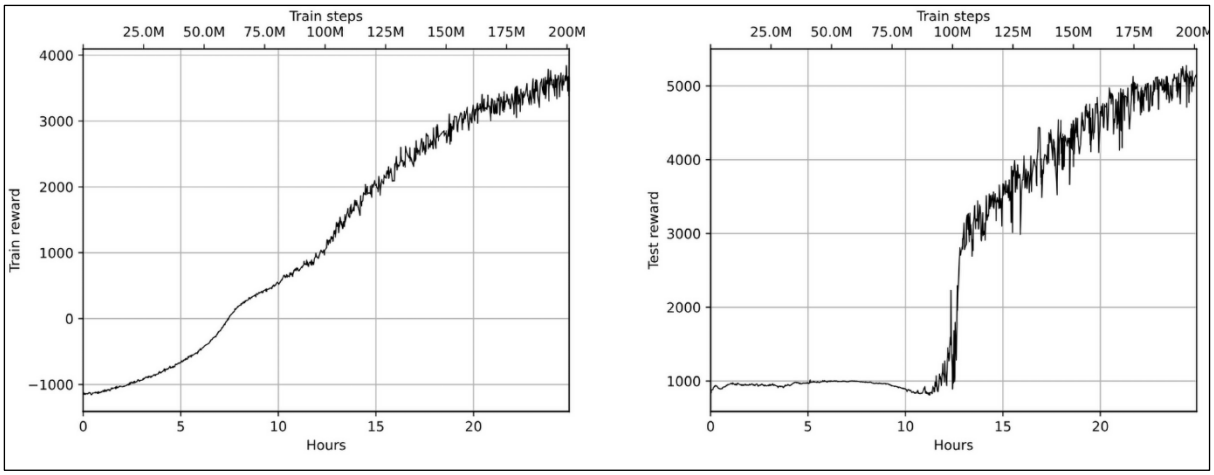

A2C
DDPG
D4PG
Reward
Hours
A2C
DDPG
D4PG
Episode steps
Hours

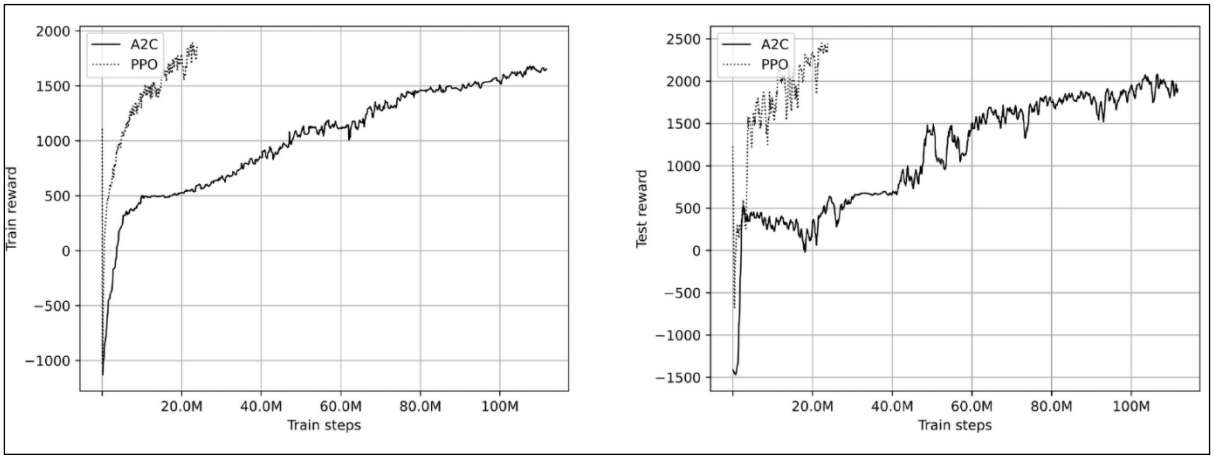
Chapter 16: Trust Regions

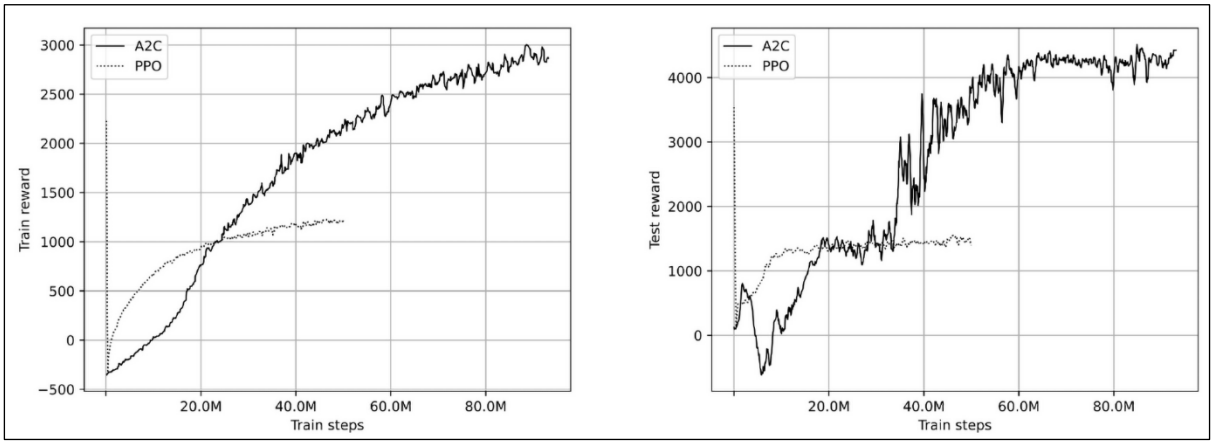


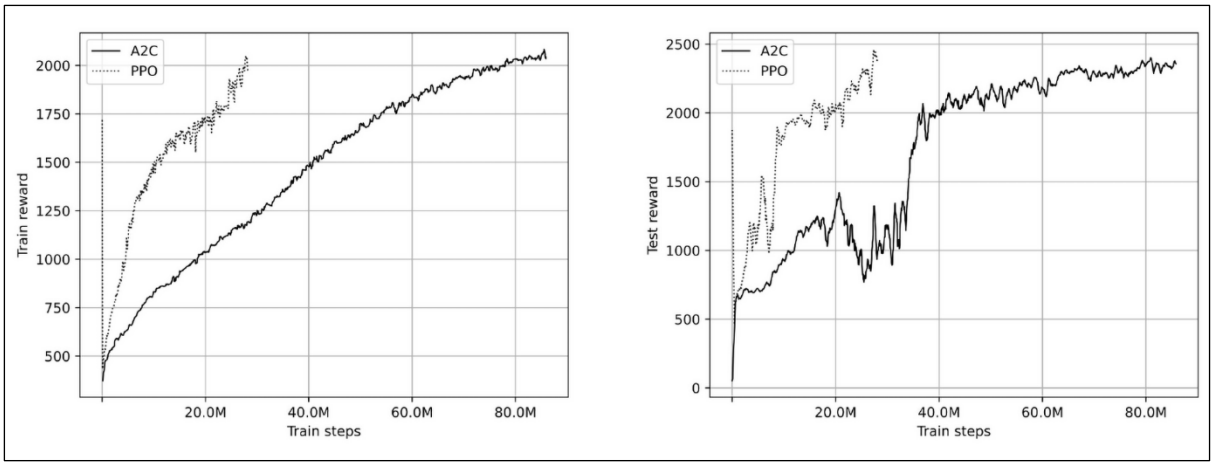


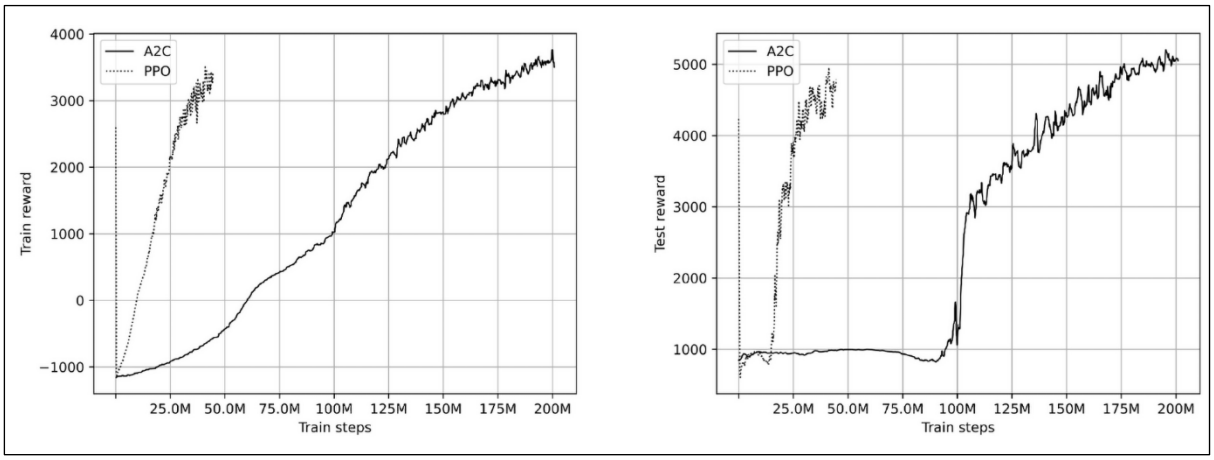


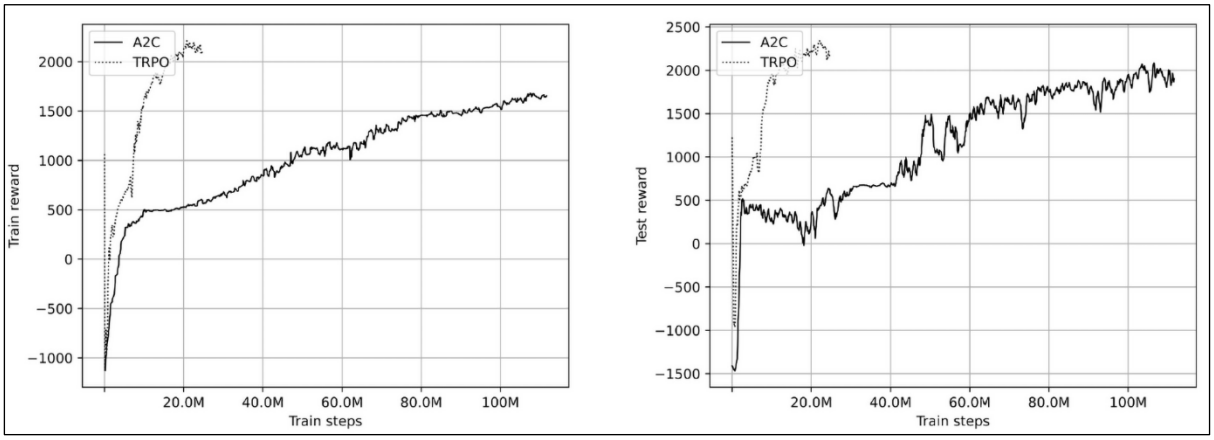

Train steps
25.0M 50.0M 75.0M 100M 125M 150M 175M 200M
Train reward
Hours
Test reward

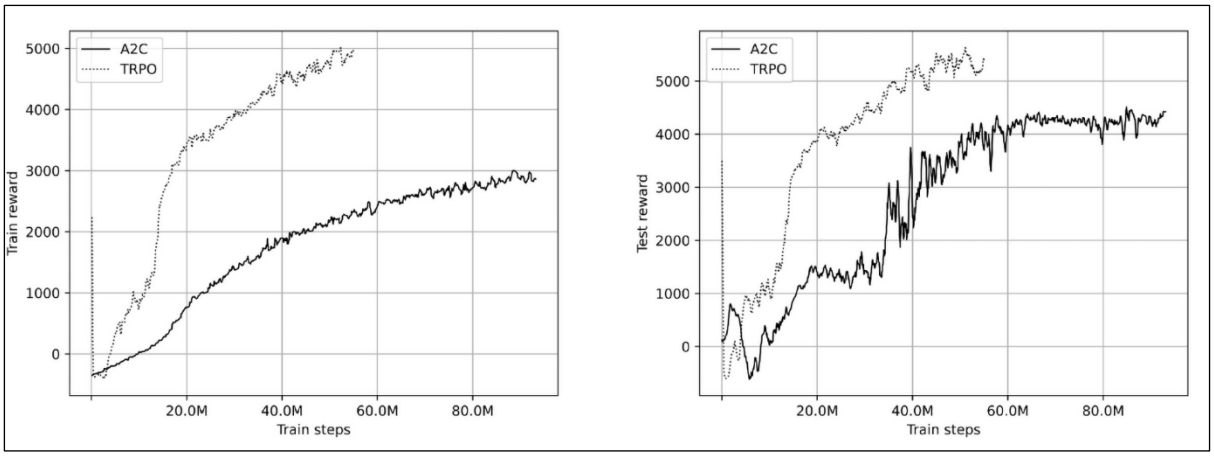

A2C
PPO
Train reward
Train steps
Test reward

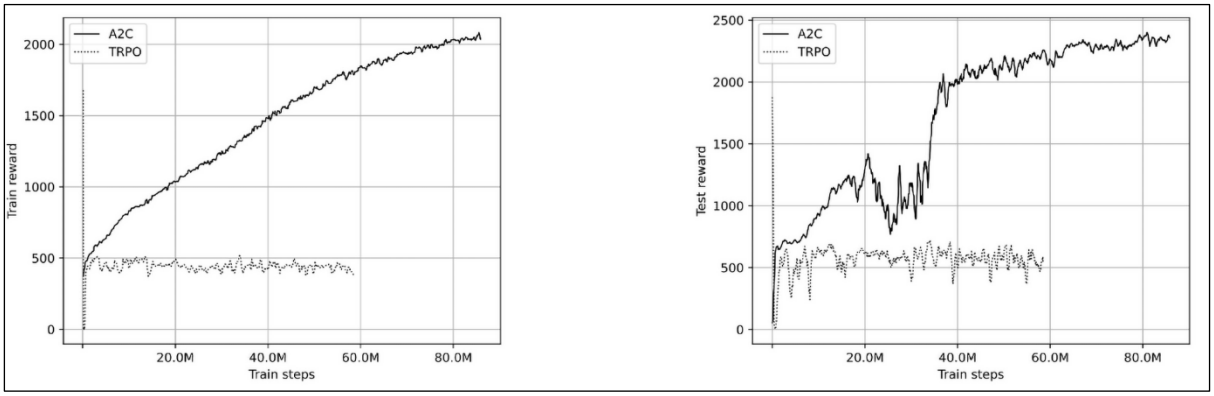

A2C
PPO
Train reward
Test reward
Train steps

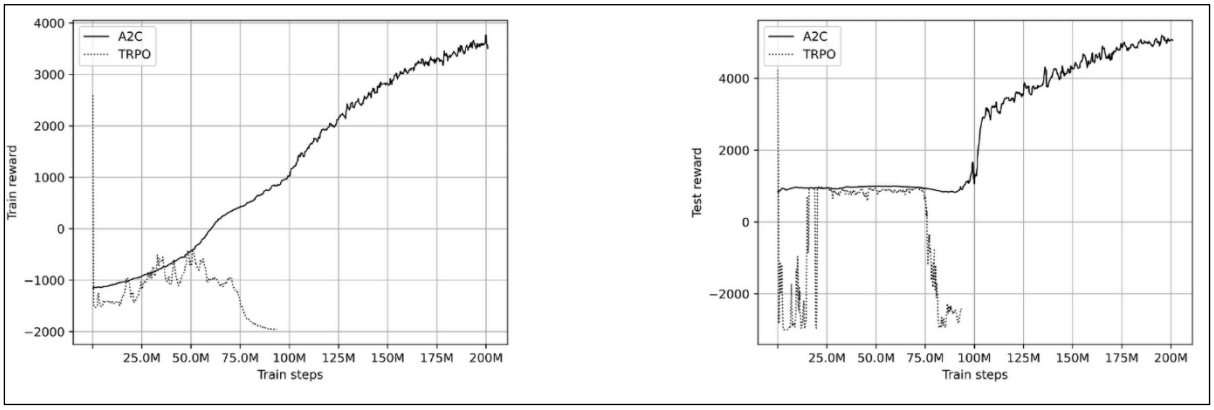

A2C
PPO
Train reward
Train steps
20.0M
40.0M
60.0M
80.0M
Test reward

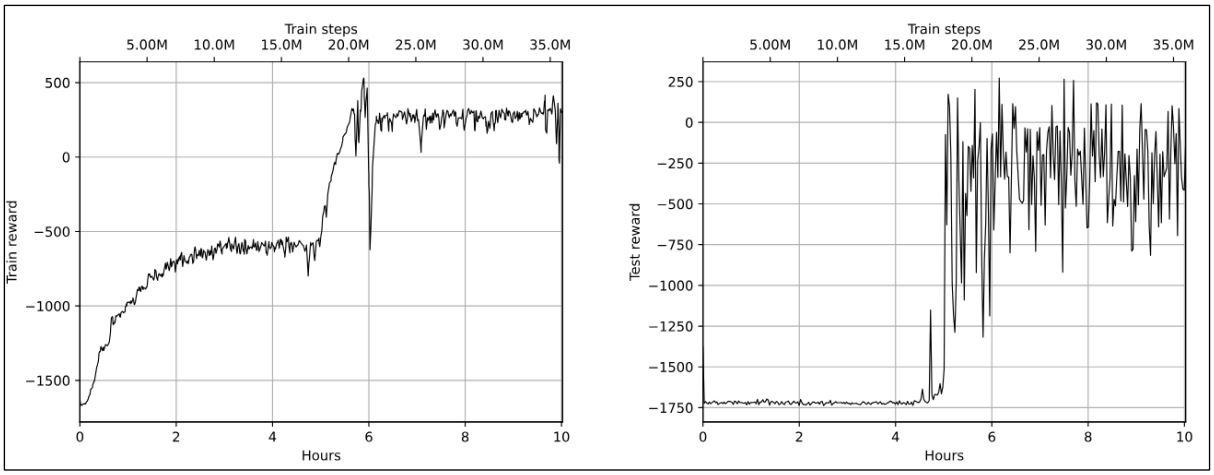

A2C
PPO
Train reward
Train steps
25.0M
50.0M
75.0M
100M
125M
150M
175M
200M
Test reward

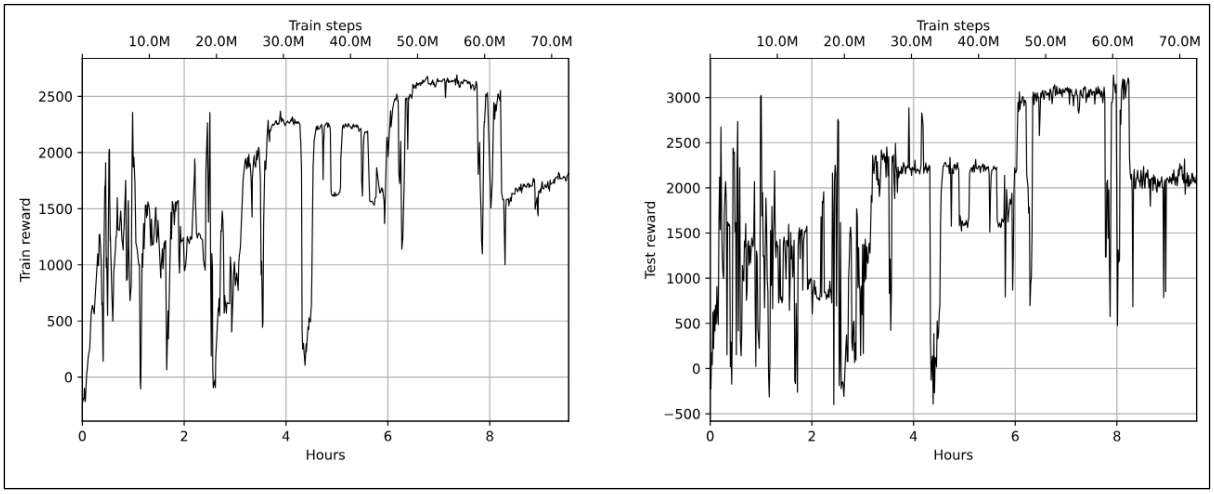

A2C
TRPO
Train reward
Train steps
20.0M
40.0M
60.0M
80.0M
100M
Test reward

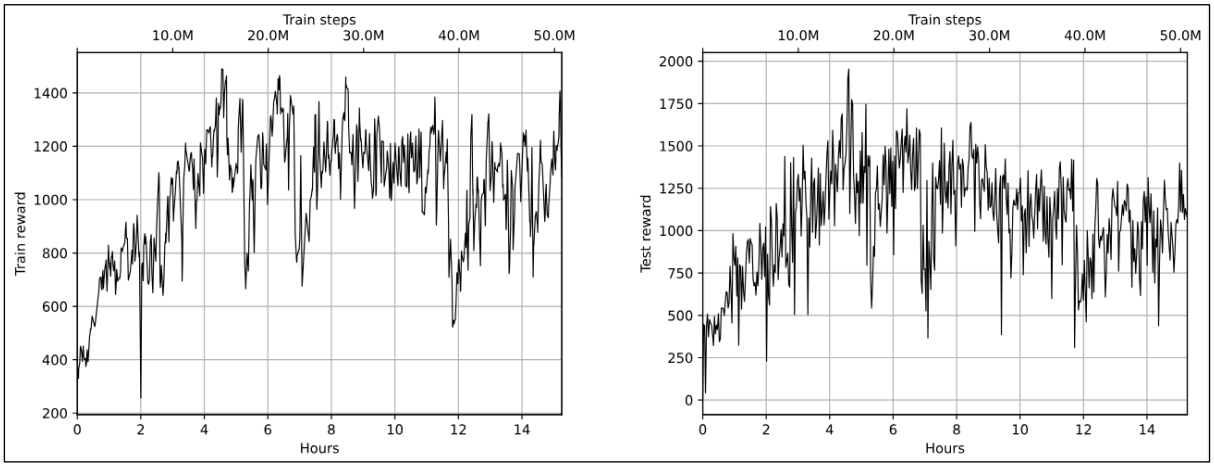

A2C
TRPO
Train reward
Train steps
Test reward
5000
4000
3000
2000
1000
0
20.0M
40.0M
60.0M
80.0M

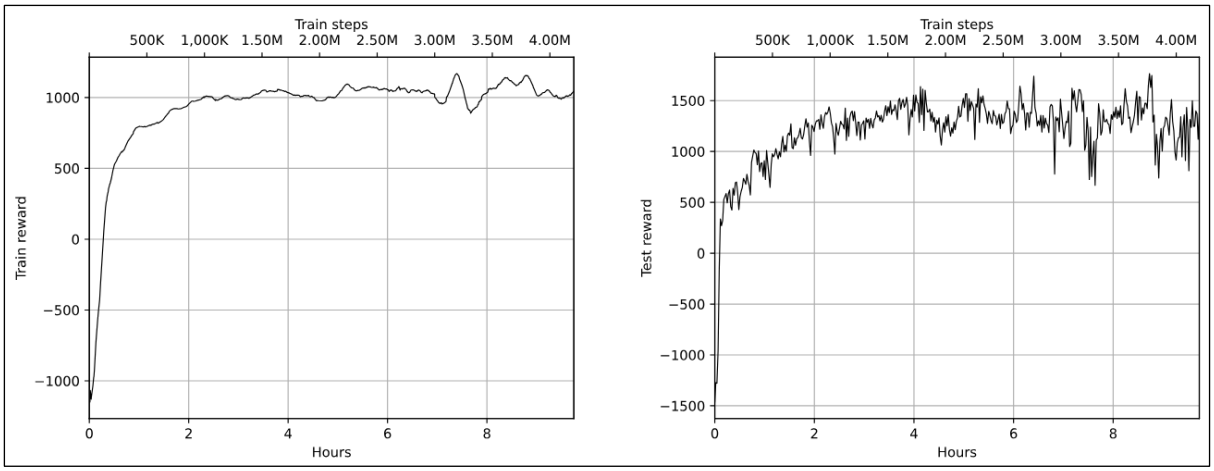

A2C
TRPO
Train reward
Train steps
Test reward
2500
2000
1500
1000
500
0
20.0M
40.0M
60.0M
80.0M

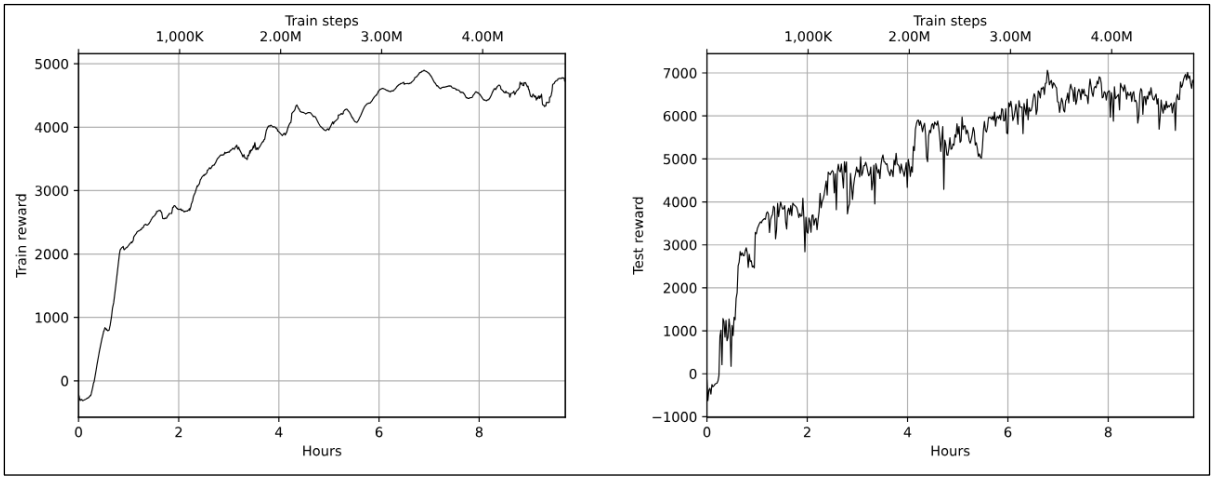

A2C
TRPO
Train reward
Train steps
Test reward
4000
3000
2000
1000
0
−1000
−2000
25.0M
50.0M
75.0M
100M
125M
150M
175M
200M

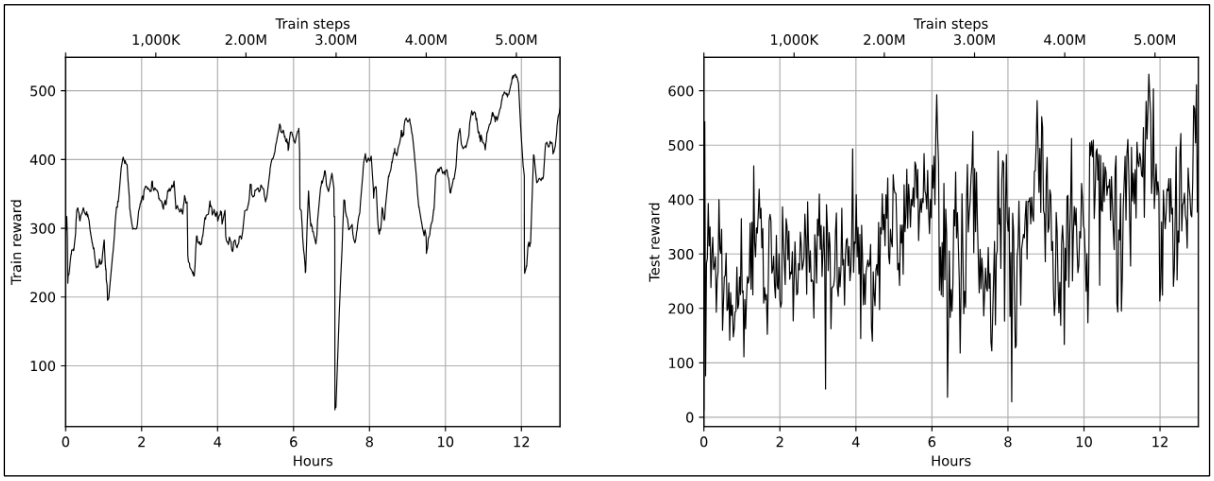

Train steps
5.00M
10.0M
15.0M
20.0M
25.0M
30.0M
35.0M
Train reward
Hours
Test reward

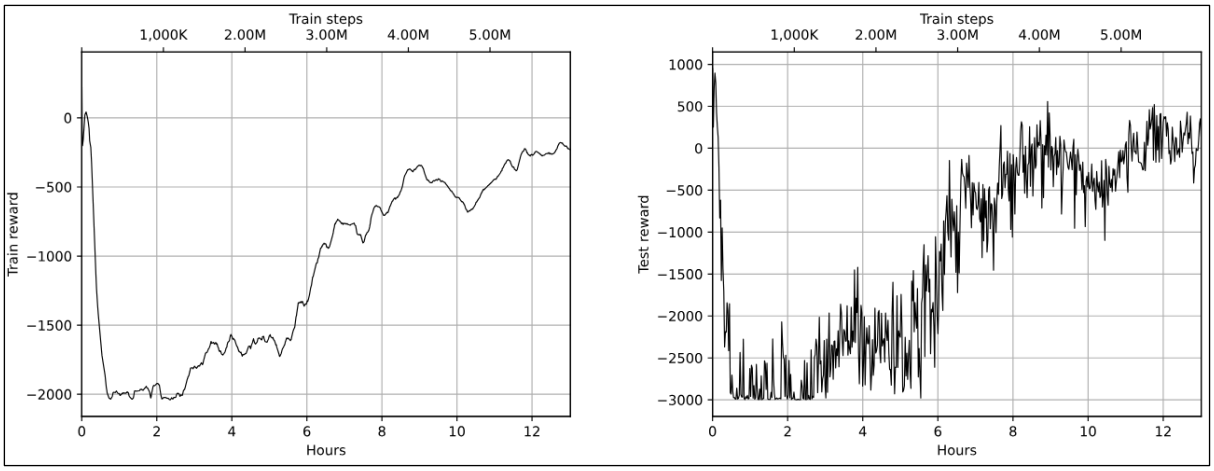

Train steps
10.0M
20.0M
30.0M
40.0M
50.0M
60.0M
70.0M
Train reward
Hours
Test reward


Train steps
10.0M
20.0M
30.0M
40.0M
50.0M
Train reward
Hours
Test reward

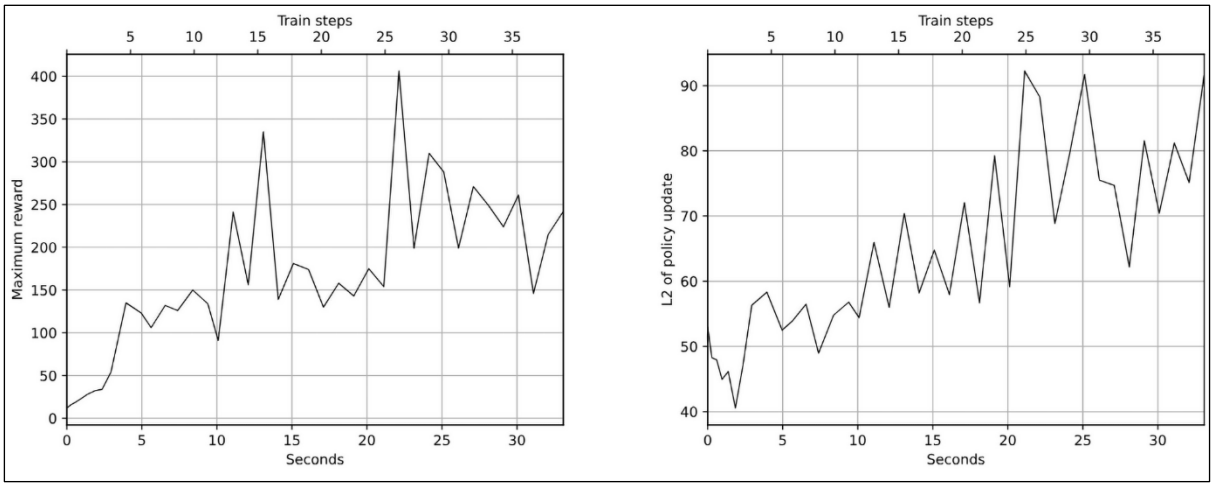

Train steps
500K 1,000K 1.50M 2.00M 2.50M 3.00M 3.50M 4.00M
Train reward
Hours
Train steps
500K 1,000K 1.50M 2.00M 2.50M 3.00M 3.50M 4.00M
Test reward
Hours

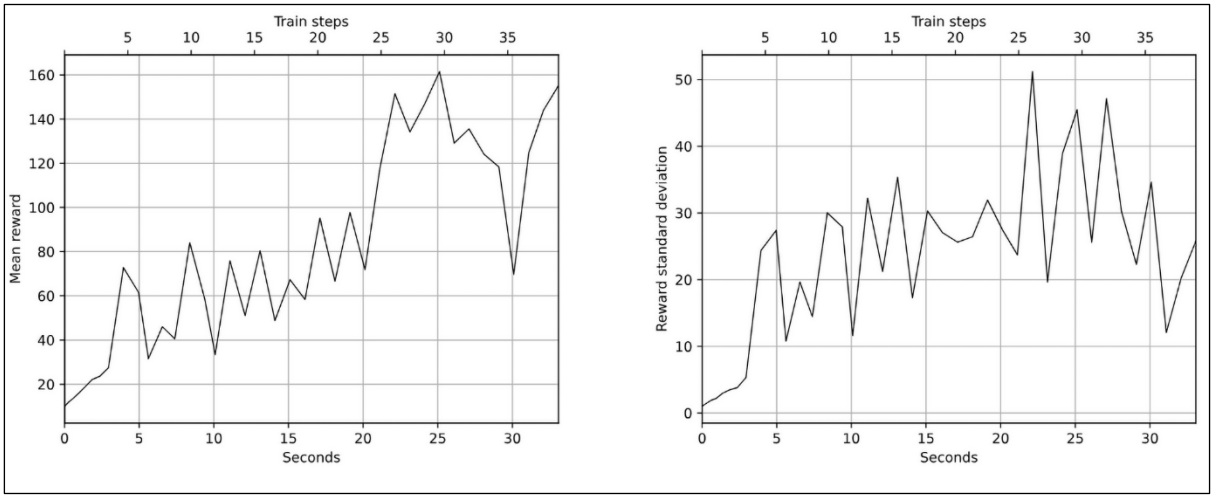

Train steps
1,000K 2.00M 3.00M 4.00M
Train reward
Hours
Train steps
1,000K 2.00M 3.00M 4.00M
Test reward
Hours

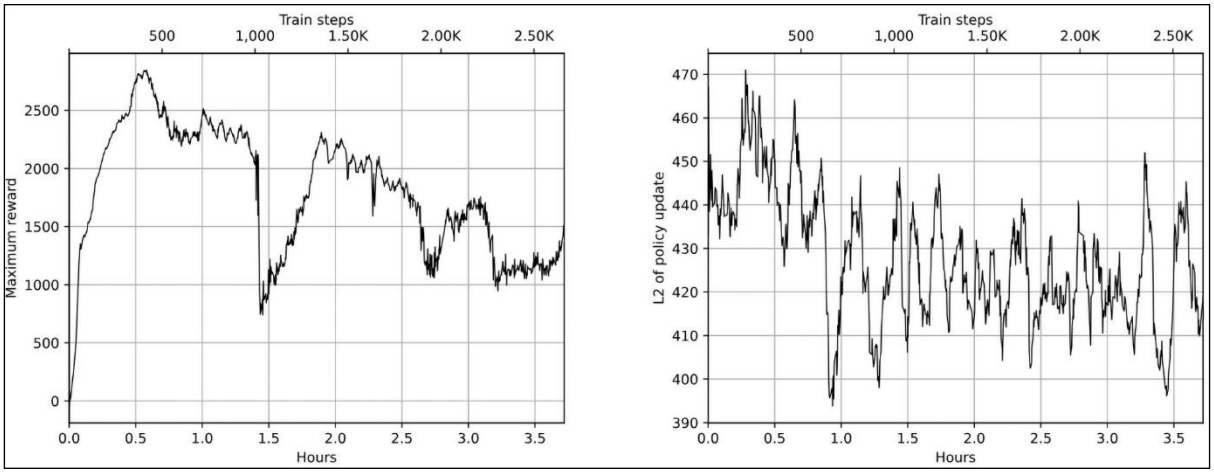

Train steps
1,000K 2.00M 3.00M 4.00M 5.00M
Train reward
Hours
Train steps
1,000K 2.00M 3.00M 4.00M 5.00M
Test reward
Hours

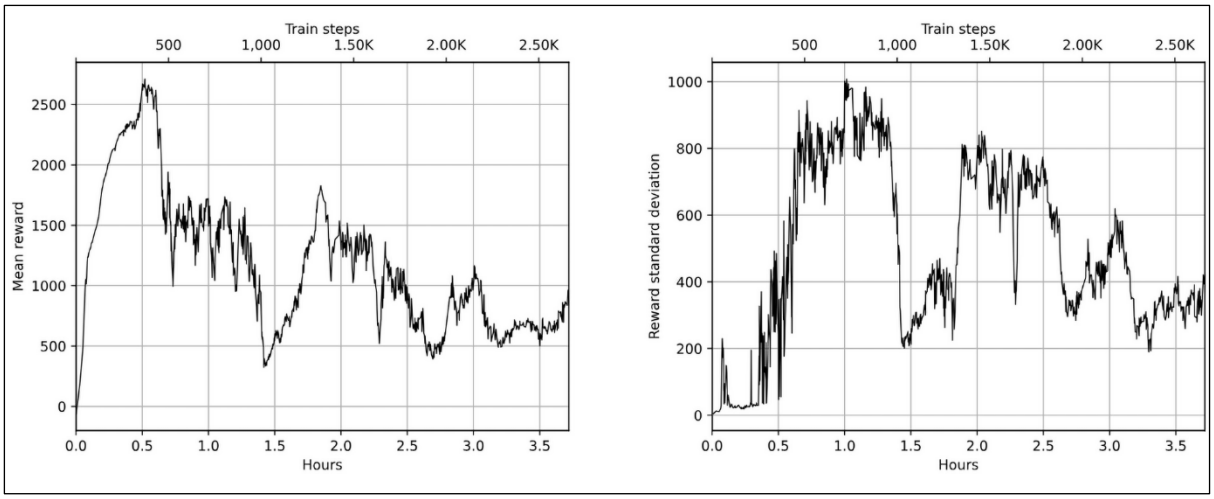

Train steps
1,000K
2.00M
3.00M
4.00M
5.00M
Train reward
0
−500
−1000
−1500
−2000
0
2
4
6
8
10
12
Hours
Train steps
1,000K
2.00M
3.00M
4.00M
5.00M
Test reward
1000
500
0
−500
−1000
−1500
−2000
−2500
−3000
0
2
4
6
8
10
12
Hours

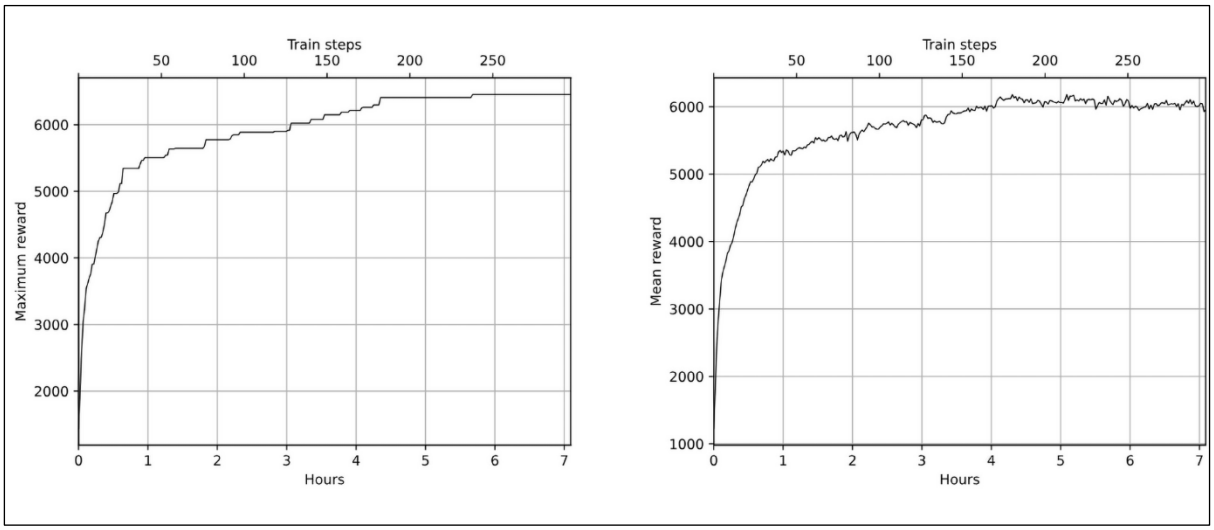
Chapter 17: Black-Box Optimization in RL

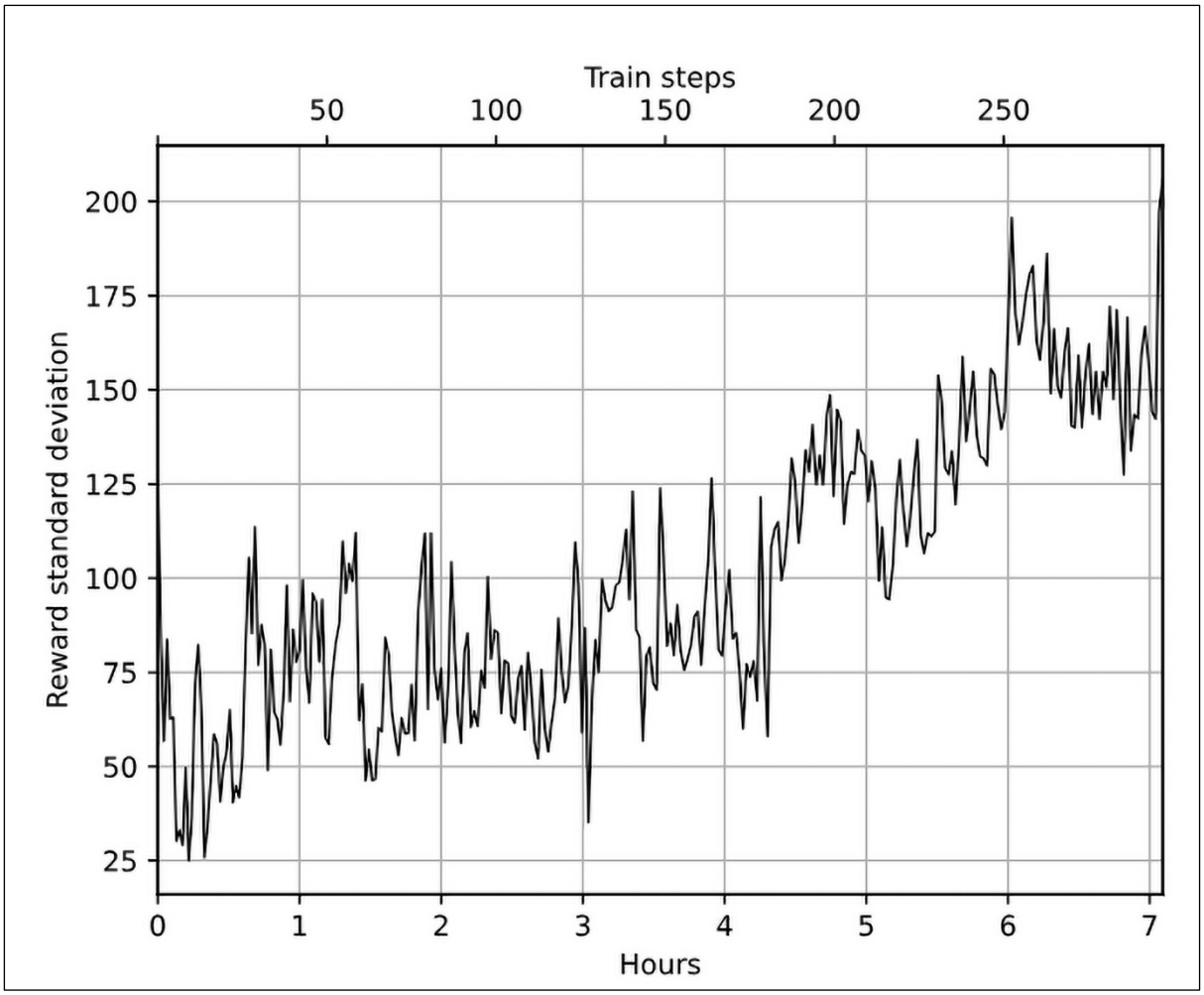




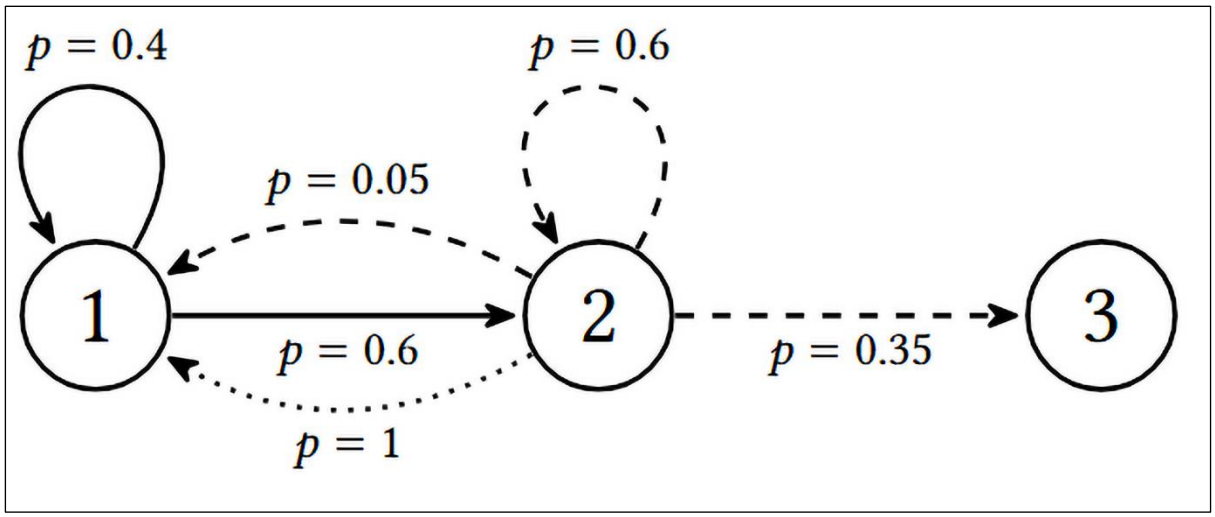


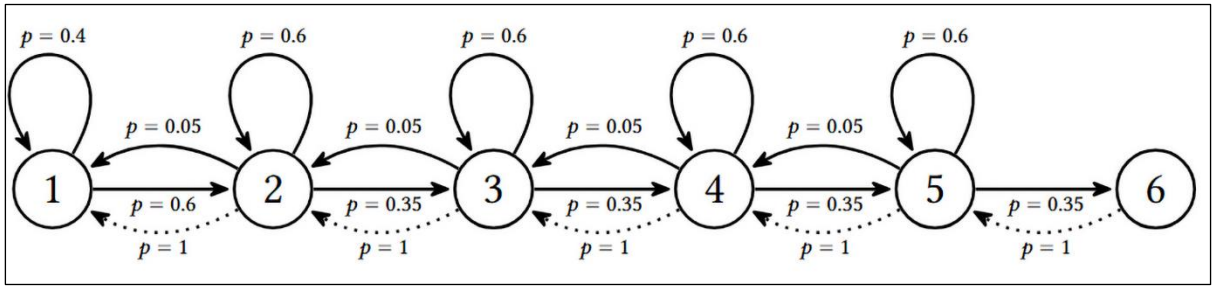

Train steps
500
1,000
1.50K
2.00K
2.50K
Mean reward
2500
2000
1500
1000
500
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Hours
Train steps
500
1,000
1.50K
2.00K
2.50K
Reward standard deviation
1000
800
600
400
200
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Hours

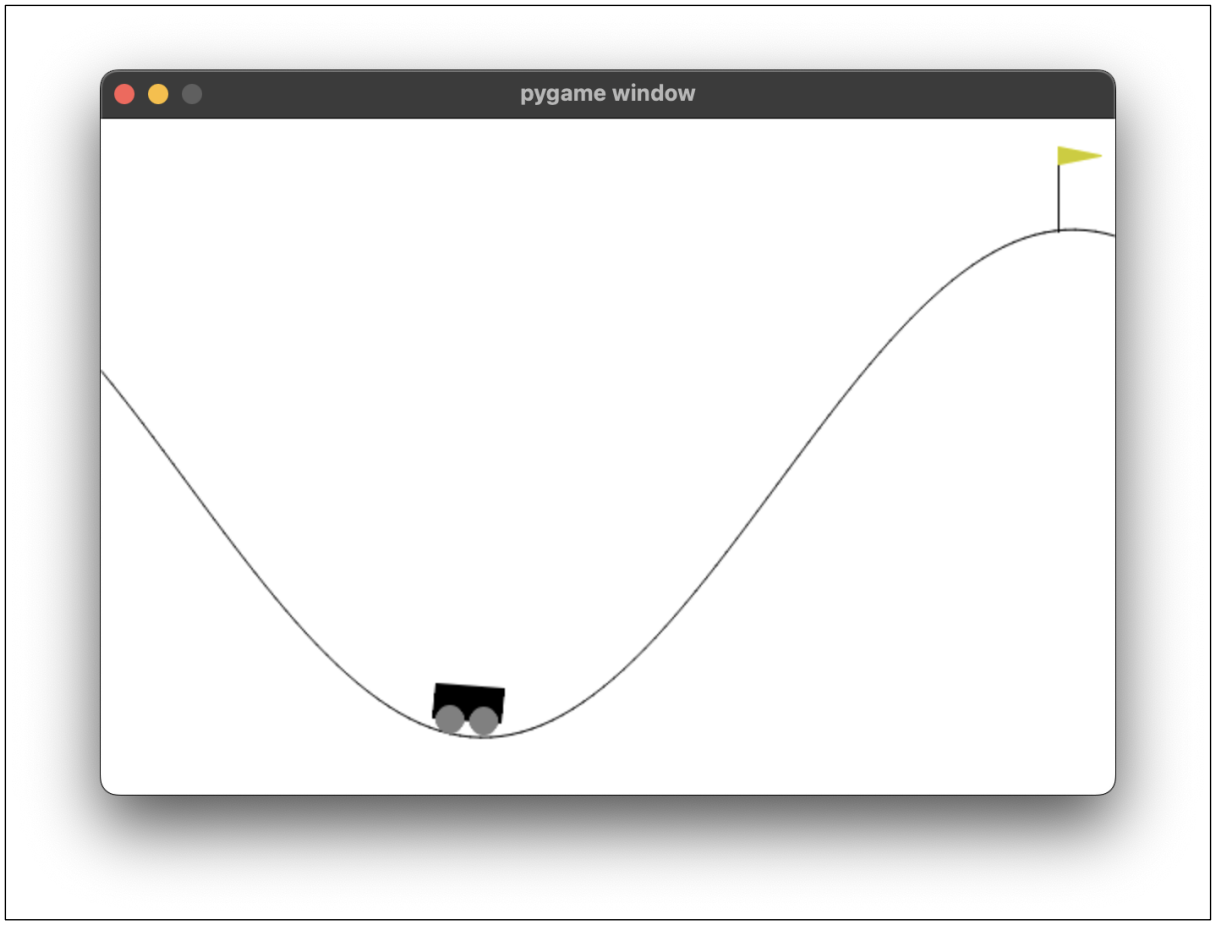

Train steps
50
100
150
200
250
Maximum reward
6000
5000
4000
3000
2000
0
1
2
3
4
5
6
7
Hours
Train steps
50
100
150
200
250
Mean reward
6000
5000
4000
3000
2000
1000
0
1
2
3
4
5
6
7
Hours

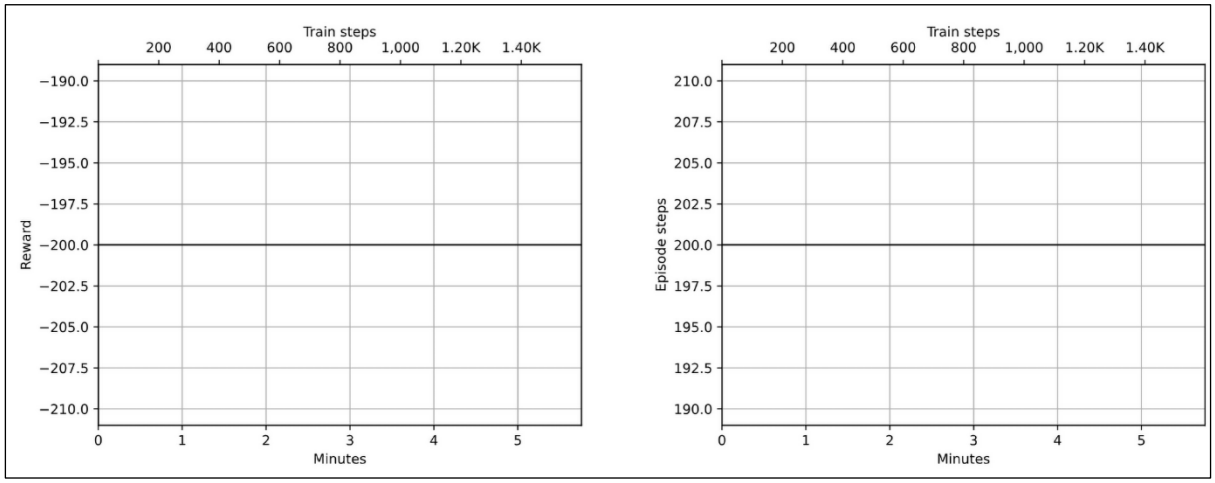

Train steps
50
100
150
200
250
Reward standard deviation
200
175
150
125
100
75
50
25
0
1
2
3
4
5
6
7
Hours

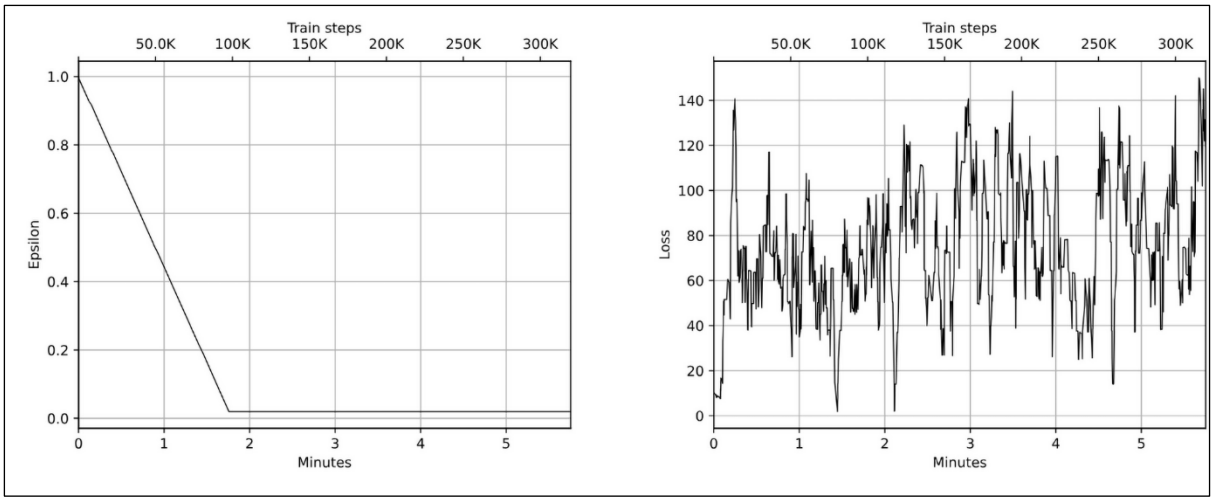
Chapter 18: Advanced Exploration

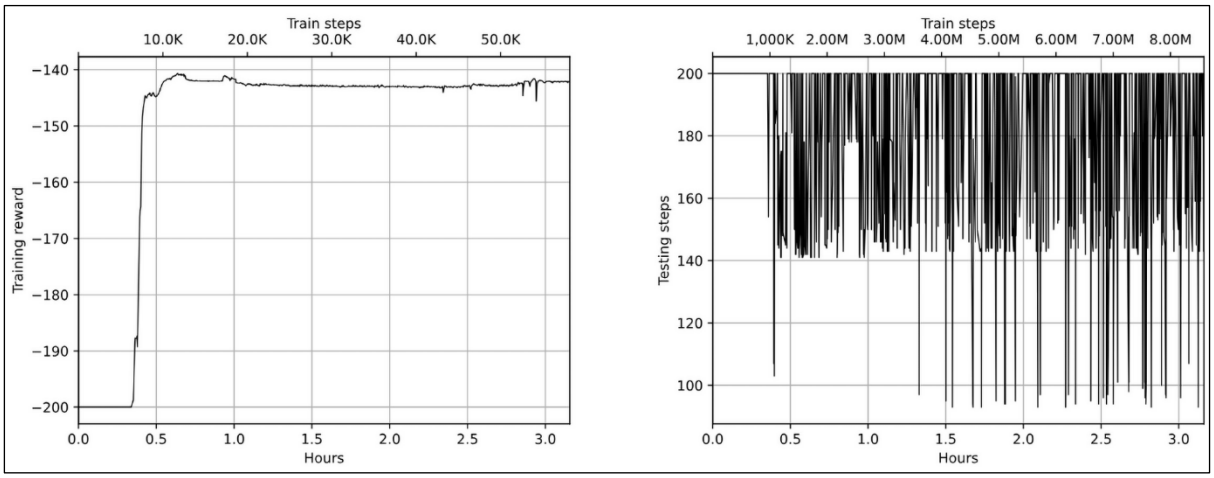


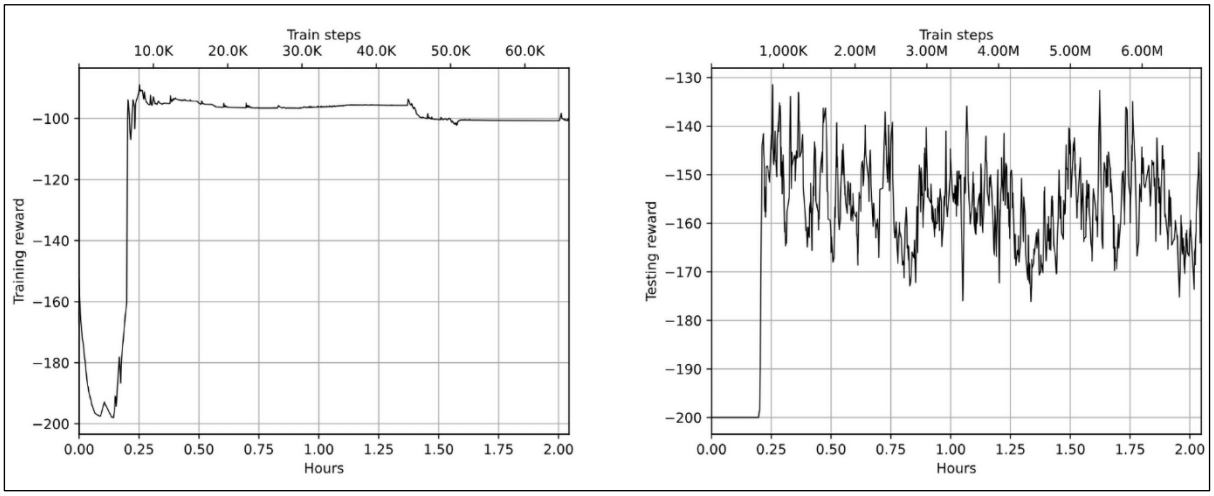


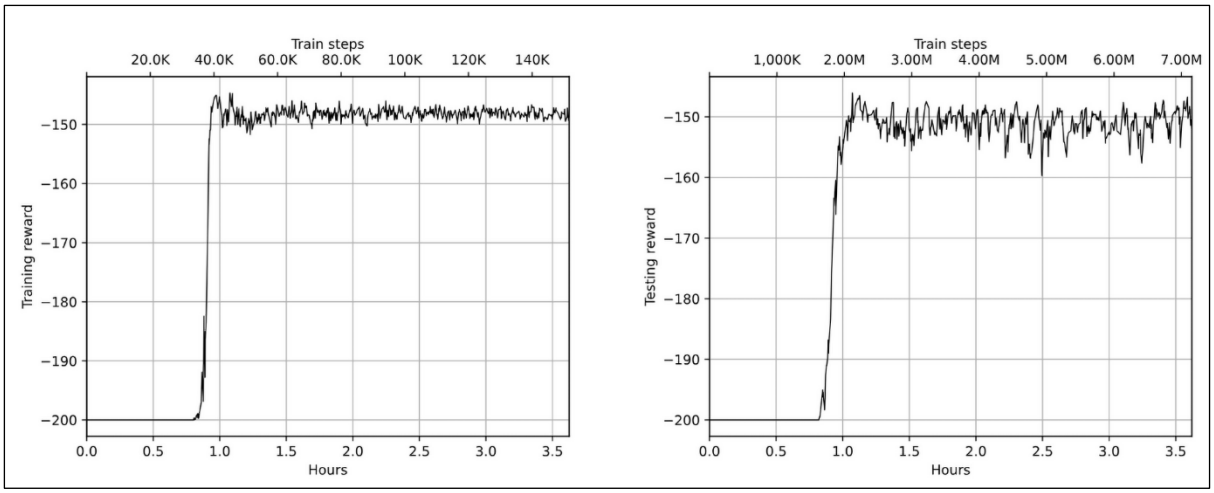

pygame window

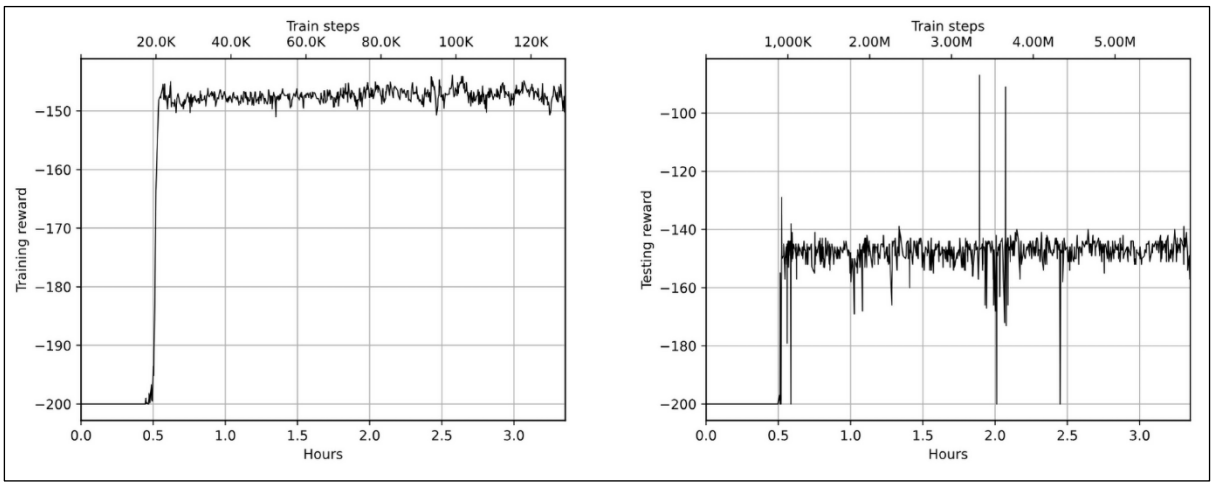

Train steps
200
400
600
800
1,000
1.20K
1.40K
−190.0
−192.5
−195.0
−197.5
−200.0
−202.5
−205.0
−207.5
−210.0
Reward
0
1
2
3
4
5
Minutes
Train steps
200
400
600
800
1,000
1.20K
1.40K
210.0
207.5
205.0
202.5
200.0
197.5
195.0
192.5
190.0
Episode steps
0
1
2
3
4
5
Minutes

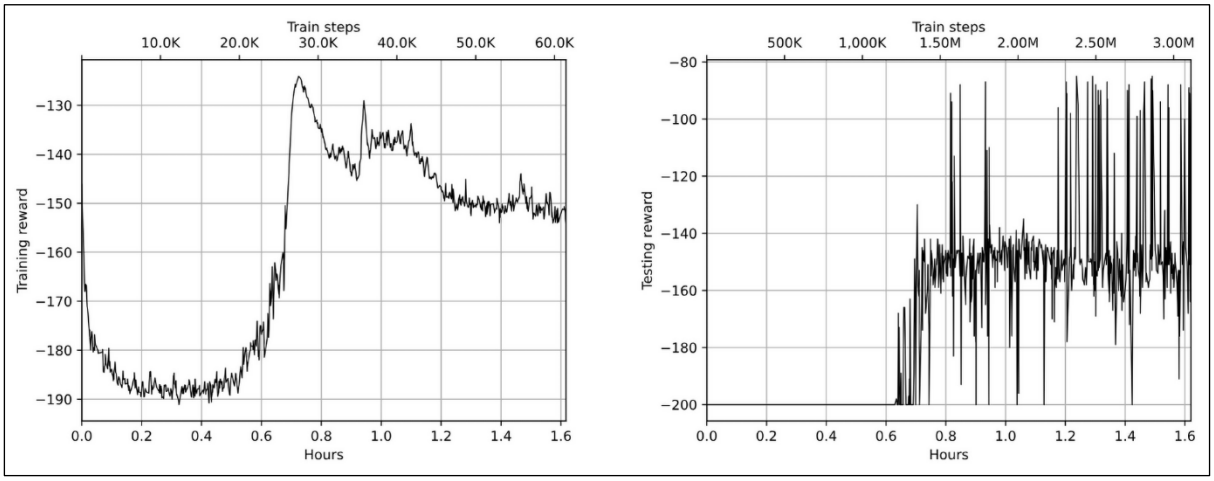

Train steps
50.0K
100K
150K
200K
250K
300K
Epsilon
Minutes
Loss

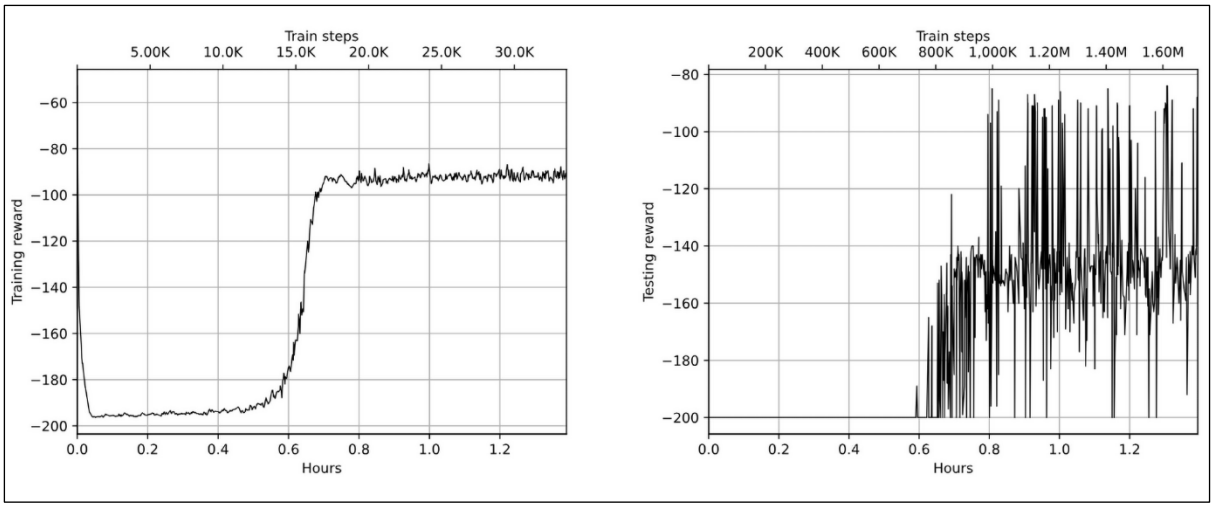

Train steps
10.0K
20.0K
30.0K
40.0K
50.0K
Training reward
Hours
1,000K
2.00M
3.00M
4.00M
5.00M
6.00M
7.00M
8.00M
Testing steps

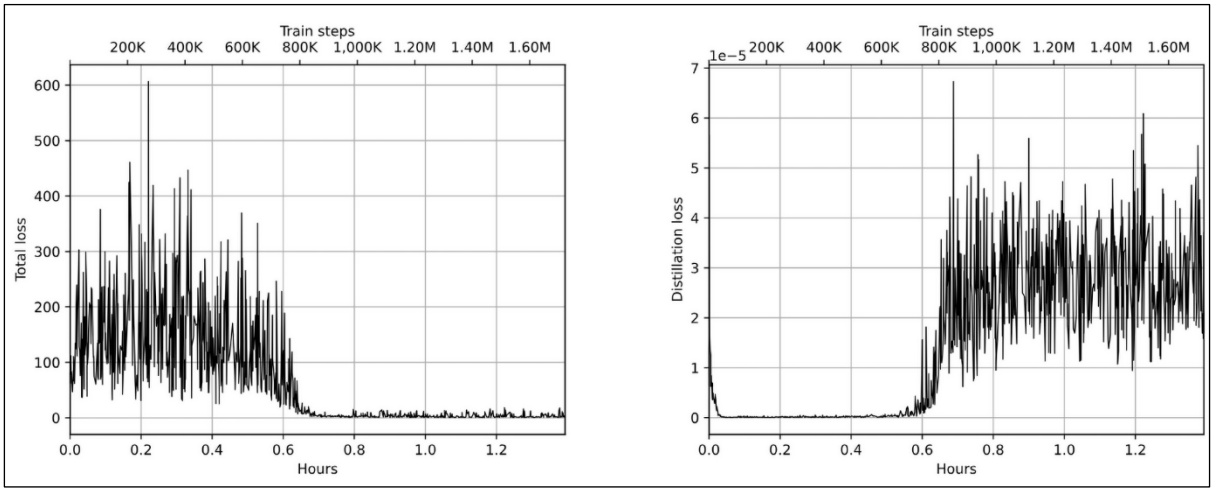

Train steps
10.0K
20.0K
30.0K
40.0K
50.0K
60.0K
Training reward
Hours
1,000K
2.00M
3.00M
4.00M
5.00M
6.00M
Testing reward

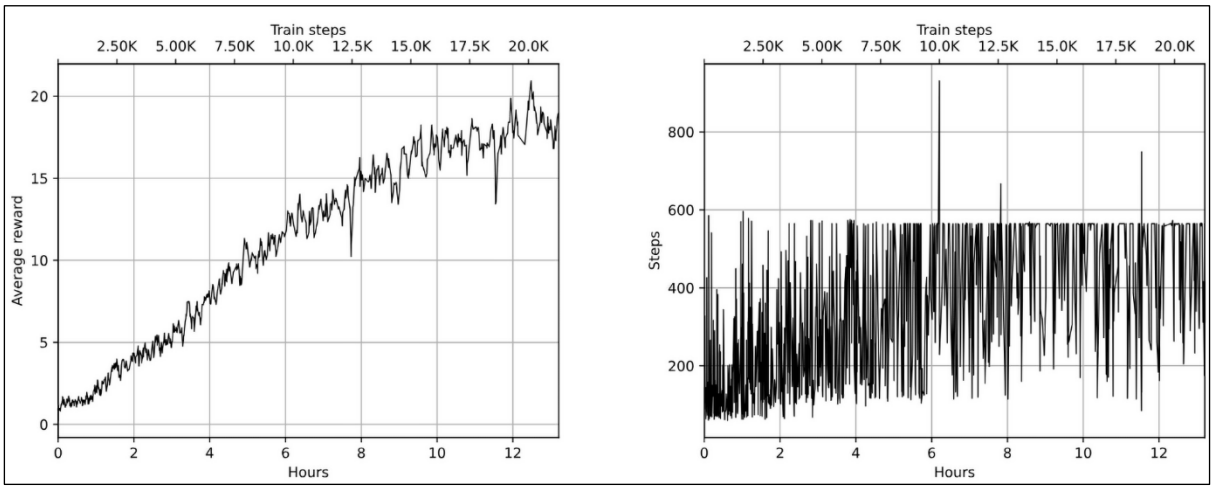

Train steps
20.0K 40.0K 60.0K 80.0K 100K 120K 140K
Training reward
−150 −160 −170 −180 −190 −200
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
Hours
Train steps
1,000K 2.00M 3.00M 4.00M 5.00M 6.00M 7.00M
Testing reward
−150 −160 −170 −180 −190 −200
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
Hours

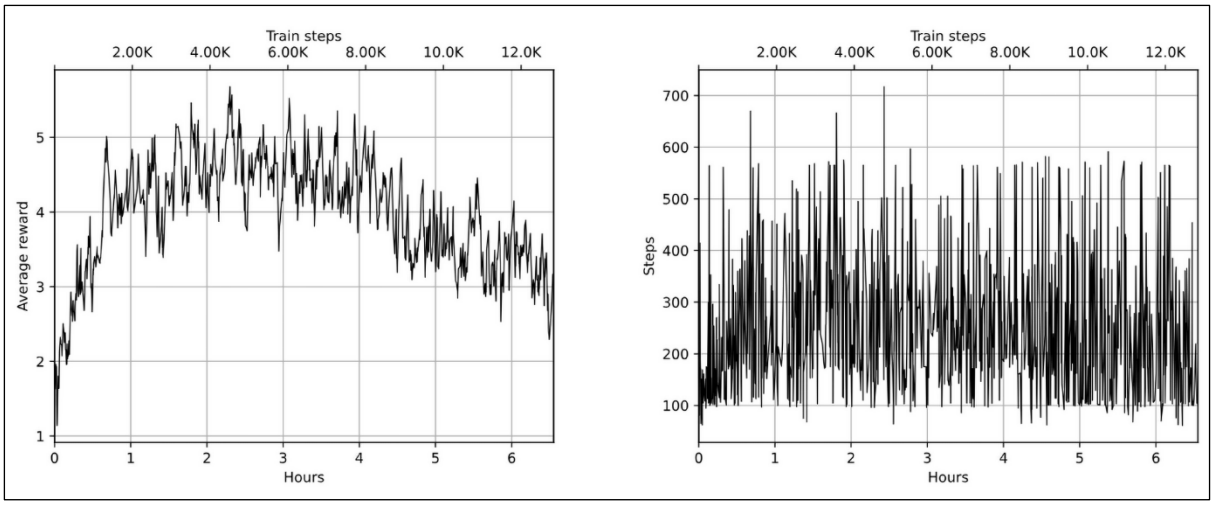

Train steps
20.0K 40.0K 60.0K 80.0K 100K 120K
Training reward
−150 −160 −170 −180 −190 −200
0.0 0.5 1.0 1.5 2.0 2.5 3.0
Hours
Train steps
1,000K 2.00M 3.00M 4.00M 5.00M
Testing reward
−100 −120 −140 −160 −180 −200
0.0 0.5 1.0 1.5 2.0 2.5 3.0
Hours


Train steps
10.0K 20.0K 30.0K 40.0K 50.0K 60.0K
Training reward
−130 −140 −150 −160 −170 −180 −190
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6
Hours
Train steps
500K 1,000K 1.50M 2.00M 2.50M 3.00M
Testing reward
−80 −100 −120 −140 −160 −180 −200
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6
Hours

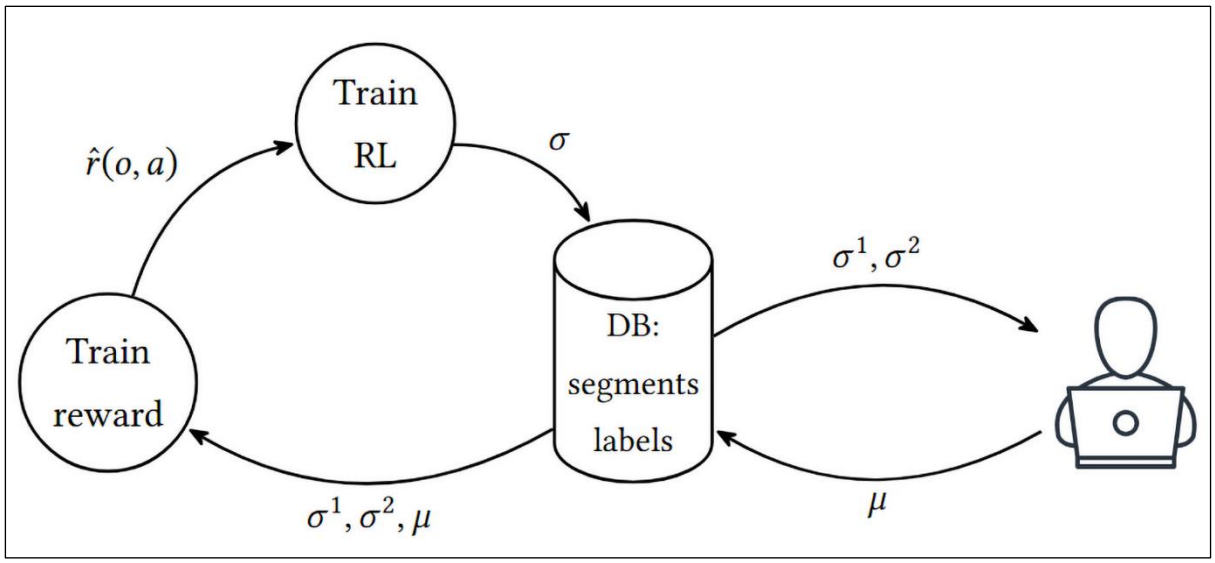

Train steps
5.00K 10.0K 15.0K 20.0K 25.0K 30.0K
Training reward
−60 −80 −100 −120 −140 −160 −180 −200
0.0 0.2 0.4 0.6 0.8 1.0 1.2
Hours
Train steps
200K 400K 600K 800K 1,000K 1.20M 1.40M 1.60M
Testing reward
−80 −100 −120 −140 −160 −180 −200
0.0 0.2 0.4 0.6 0.8 1.0 1.2
Hours

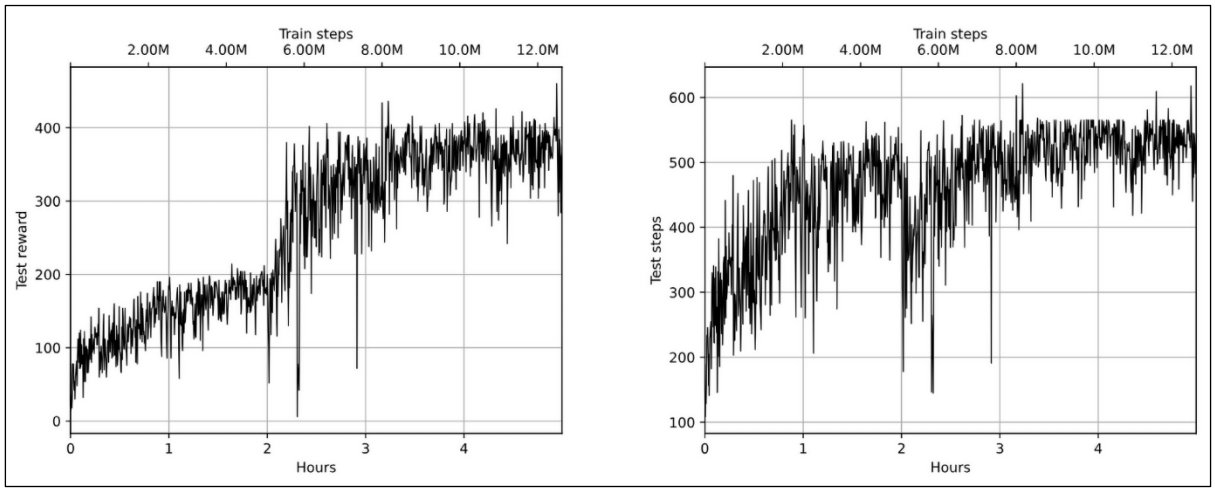

Train steps
200K 400K 600K 800K 1,000K 1.20M 1.40M 1.60M
Total loss
600 500 400 300 200 100 0
0.0 0.2 0.4 0.6 0.8 1.0 1.2
Hours
Train steps
1e−5
200K 400K 600K 800K 1,000K 1.20M 1.40M 1.60M
Distillation loss
7 6 5 4 3 2 1 0
0.0 0.2 0.4 0.6 0.8 1.0 1.2
Hours

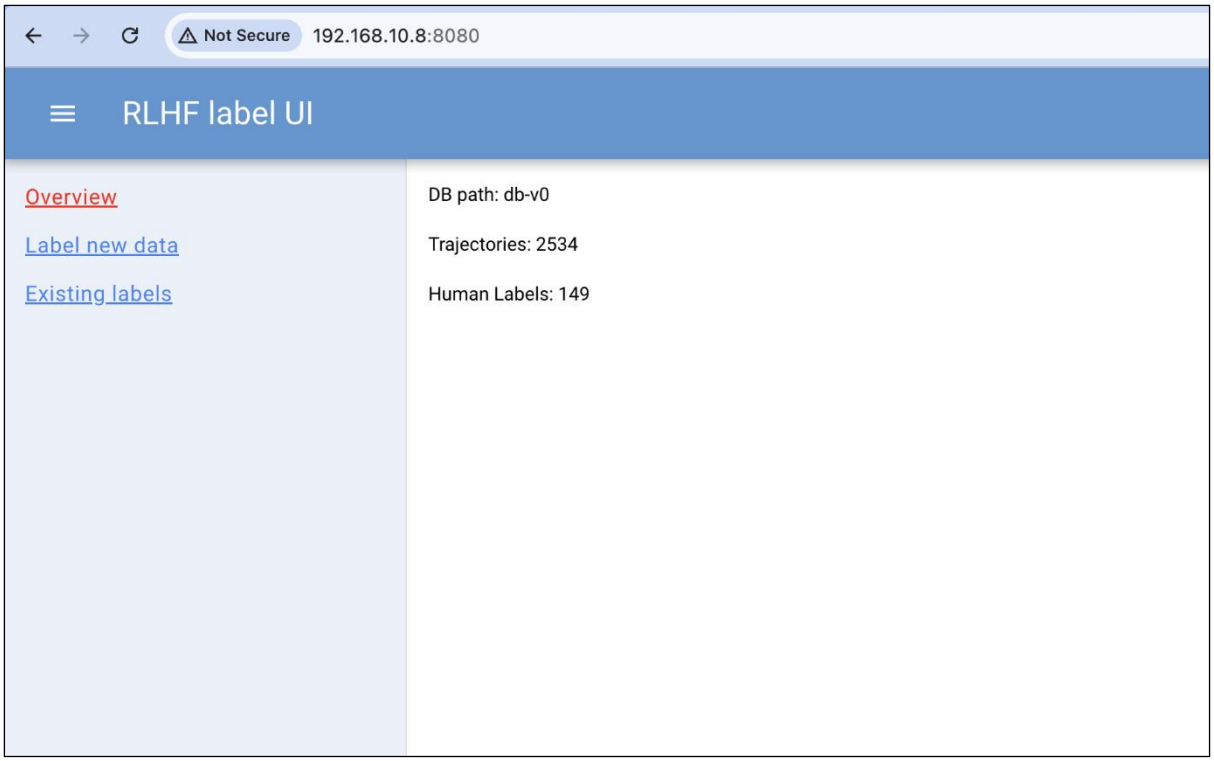

Train steps
2.50K 5.00K 7.50K 10.0K 12.5K 15.0K 17.5K 20.0K
Average reward
20 15 10 5 0
0 2 4 6 8 10 12
Hours
Train steps
2.50K 5.00K 7.50K 10.0K 12.5K 15.0K 17.5K 20.0K
Steps
800 600 400 200
0 2 4 6 8 10 12
Hours

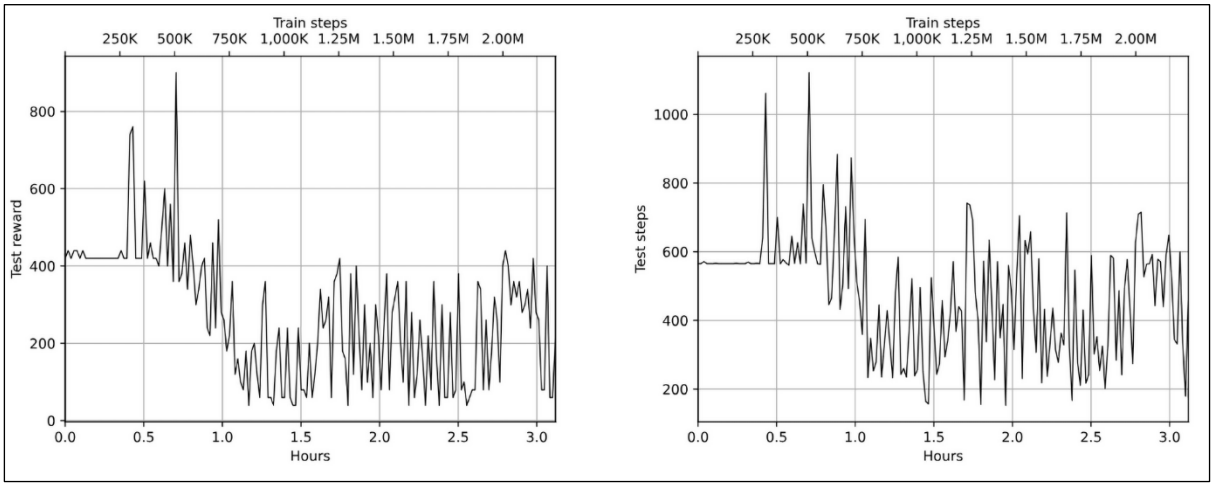

Train steps
2.00K
4.00K
6.00K
8.00K
10.0K
12.0K
Average reward
Hours
Train steps
2.00K
4.00K
6.00K
8.00K
10.0K
12.0K
Steps
Hours

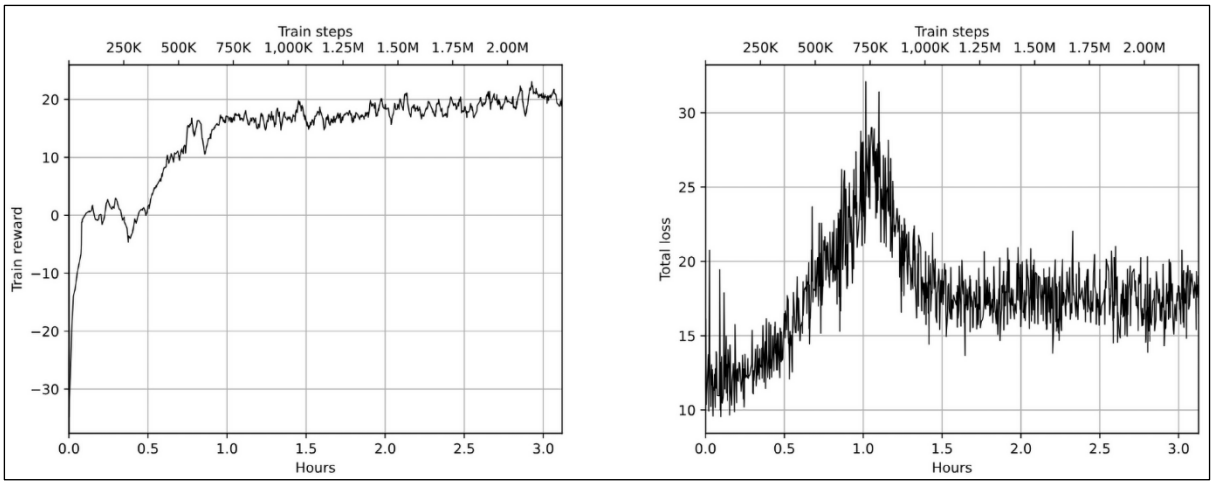
Chapter 19: Reinforcement Learning with Human Feedback

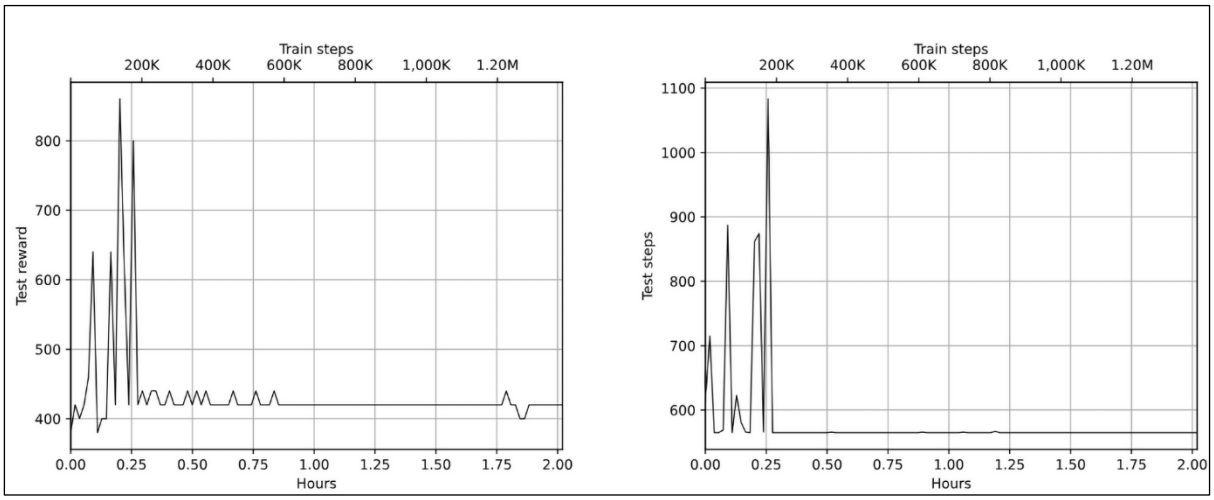


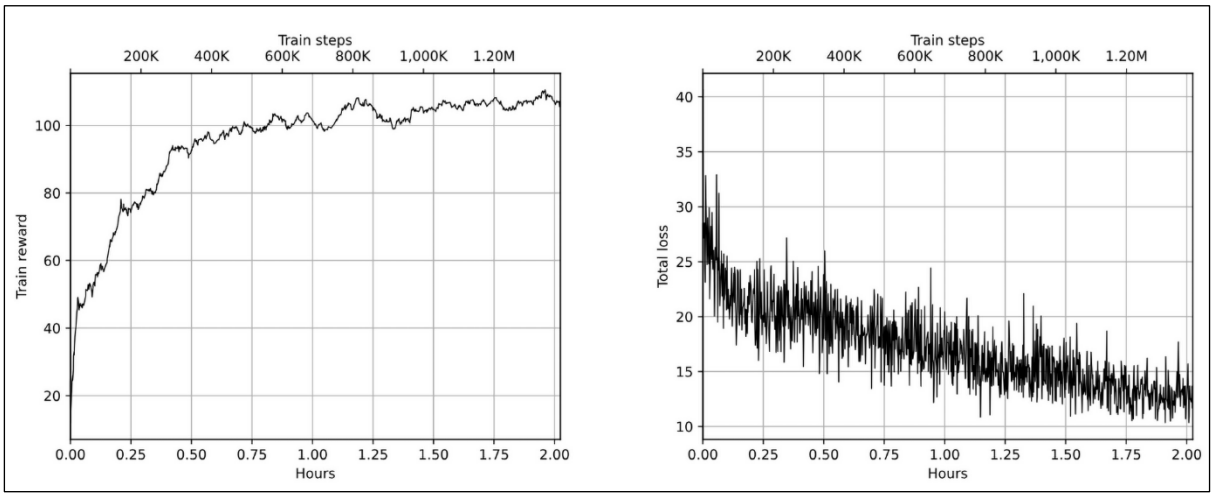


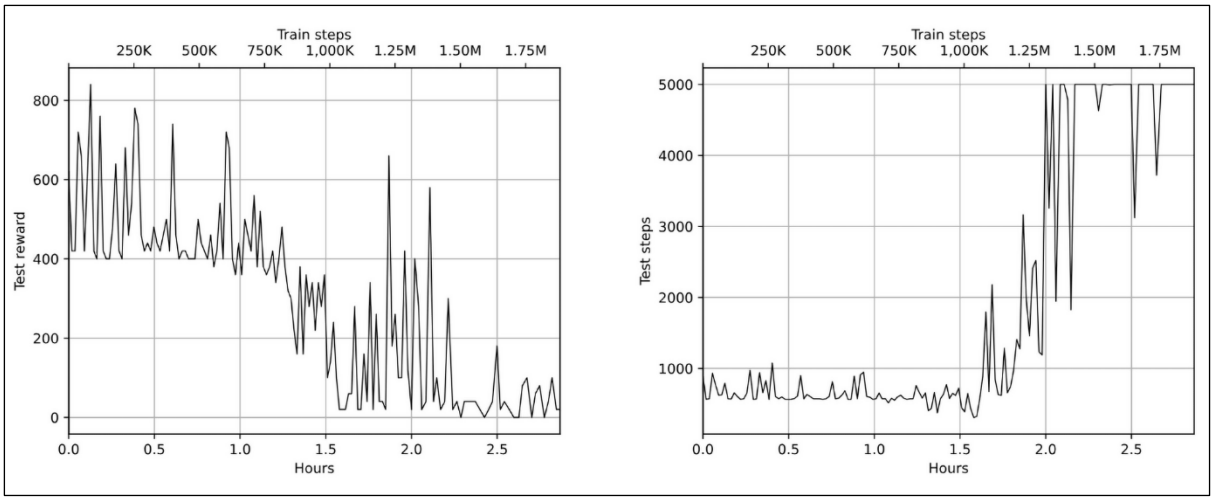

Not Secure
192.168.10.8:8080
RLHF label UI
Overview
Label new data
Existing labels
DB path: db-v0
Trajectories: 2534
Human Labels: 149

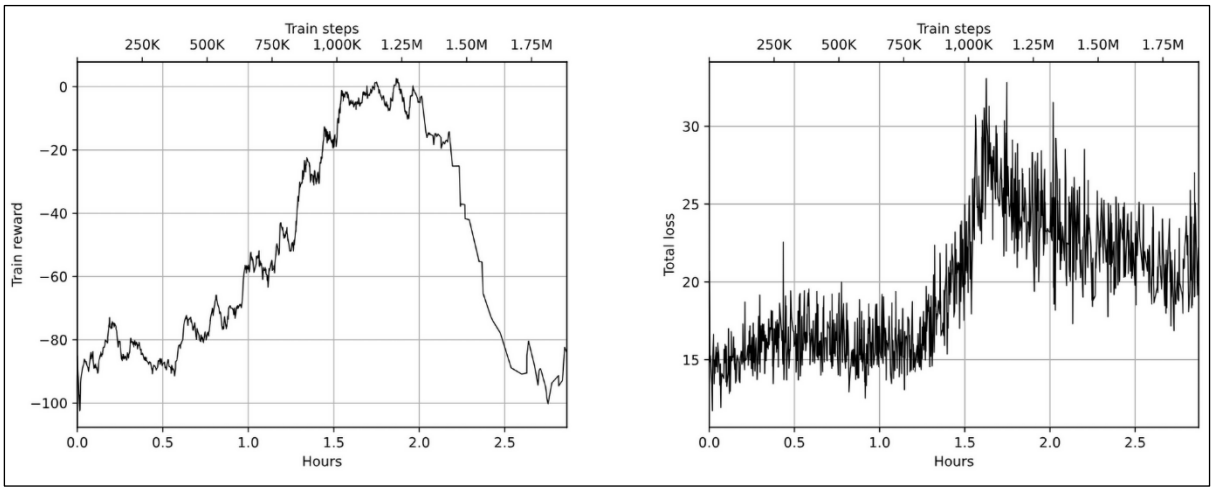

Train steps
250K 500K 750K 1,000K 1.25M 1.50M 1.75M 2.00M
Test reward
Hours
Train steps
250K 500K 750K 1,000K 1.25M 1.50M 1.75M 2.00M
Test steps
Hours


Train steps
250K 500K 750K 1,000K 1.25M 1.50M 1.75M 2.00M
Train reward
Hours
Train steps
250K 500K 750K 1,000K 1.25M 1.50M 1.75M 2.00M
Total loss
Hours

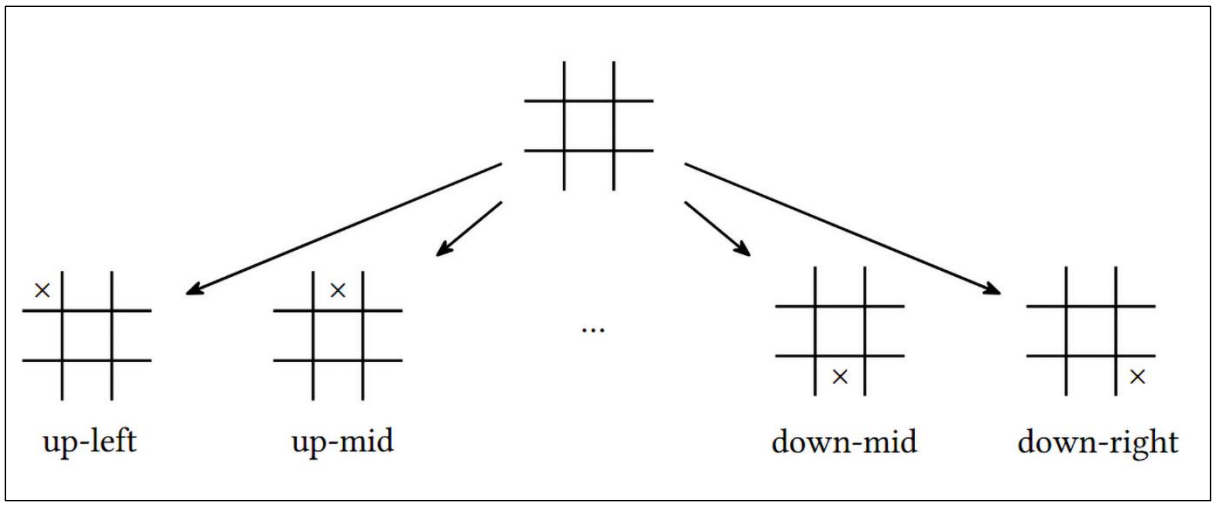

Train steps
200K 400K 600K 800K 1,000K 1.20M
Test reward
Hours
Train steps
200K 400K 600K 800K 1,000K 1.20M
Test steps
Hours

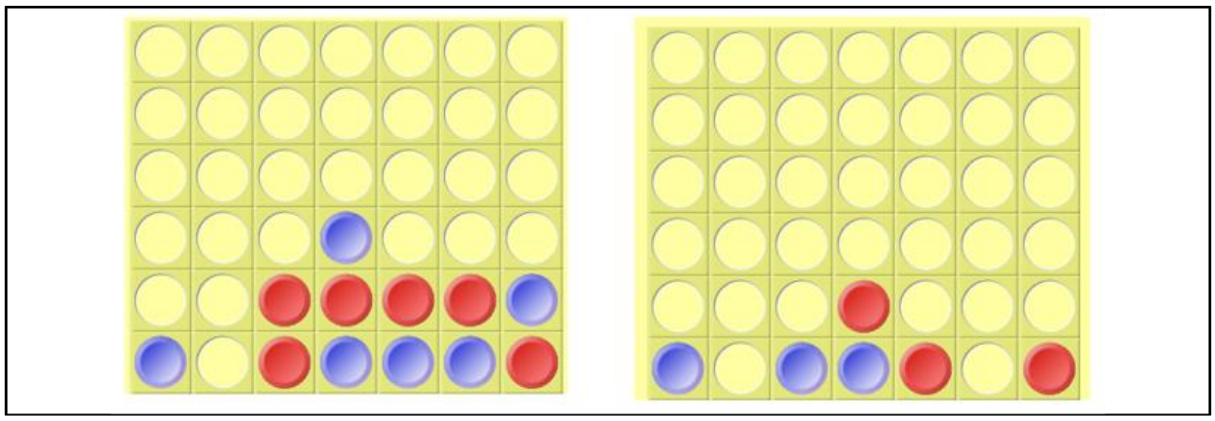

Train steps
200K 400K 600K 800K 1,000K 1.20M
Train reward
Hours
Train steps
200K 400K 600K 800K 1,000K 1.20M
Total loss
Hours

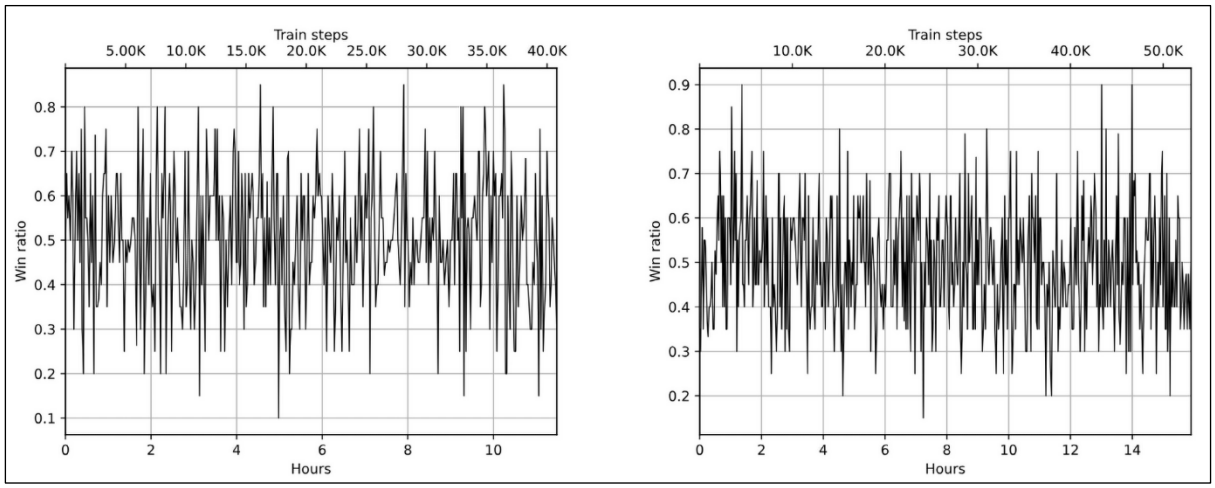

Train steps
250K
500K
750K
1,000K
1.25M
1.50M
1.75M
Test reward
Hours
Test steps

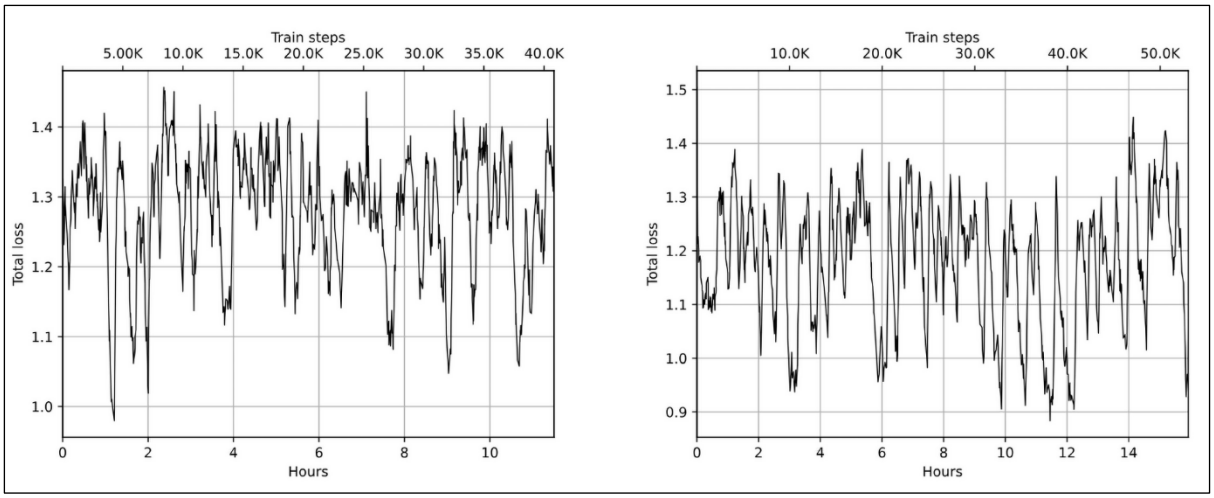

Train steps
250K
500K
750K
1,000K
1.25M
1.50M
1.75M
Train reward
Hours
Total loss

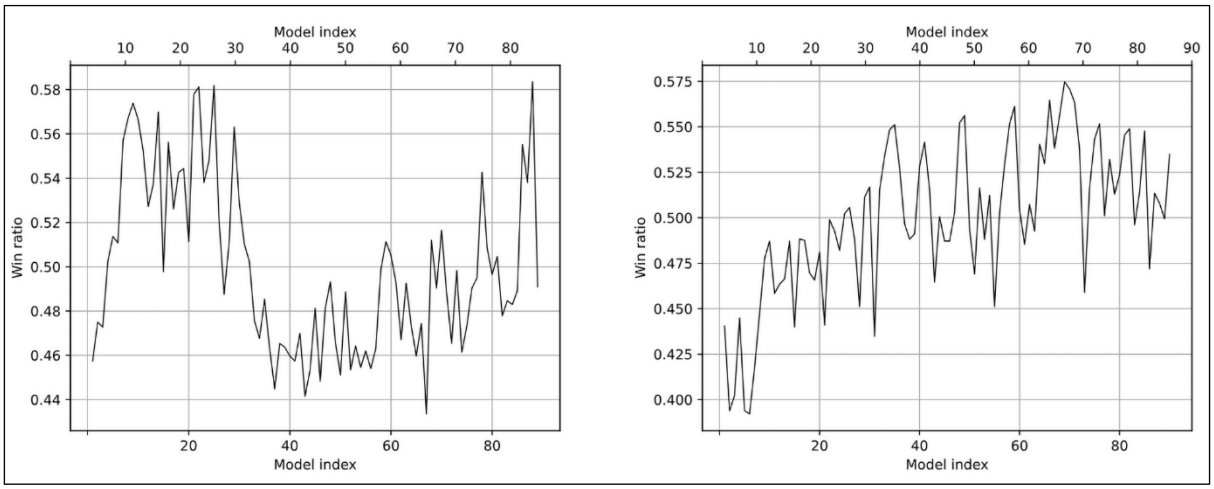
Chapter 20: AlphaGo Zero and MuZero

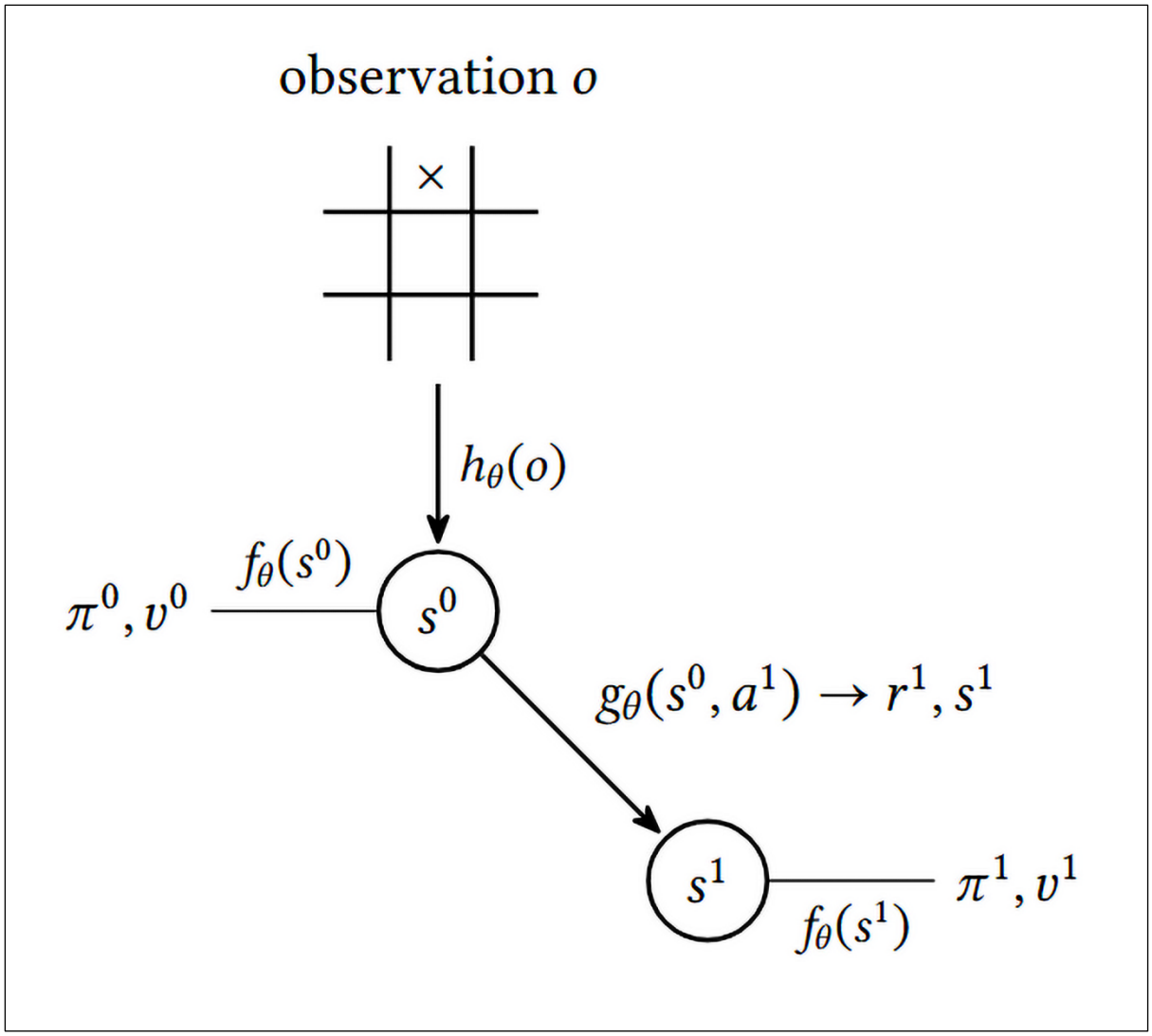


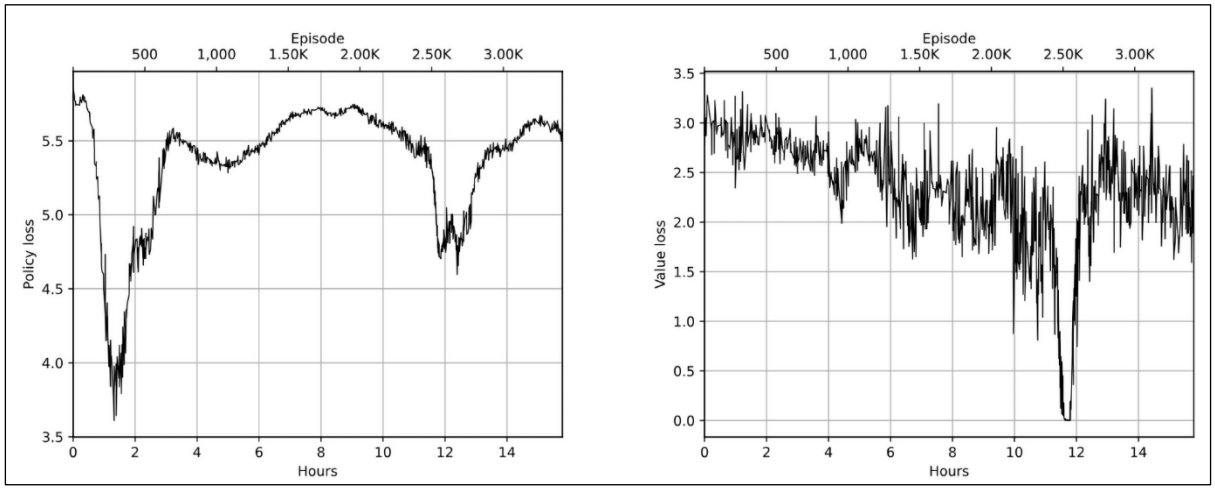


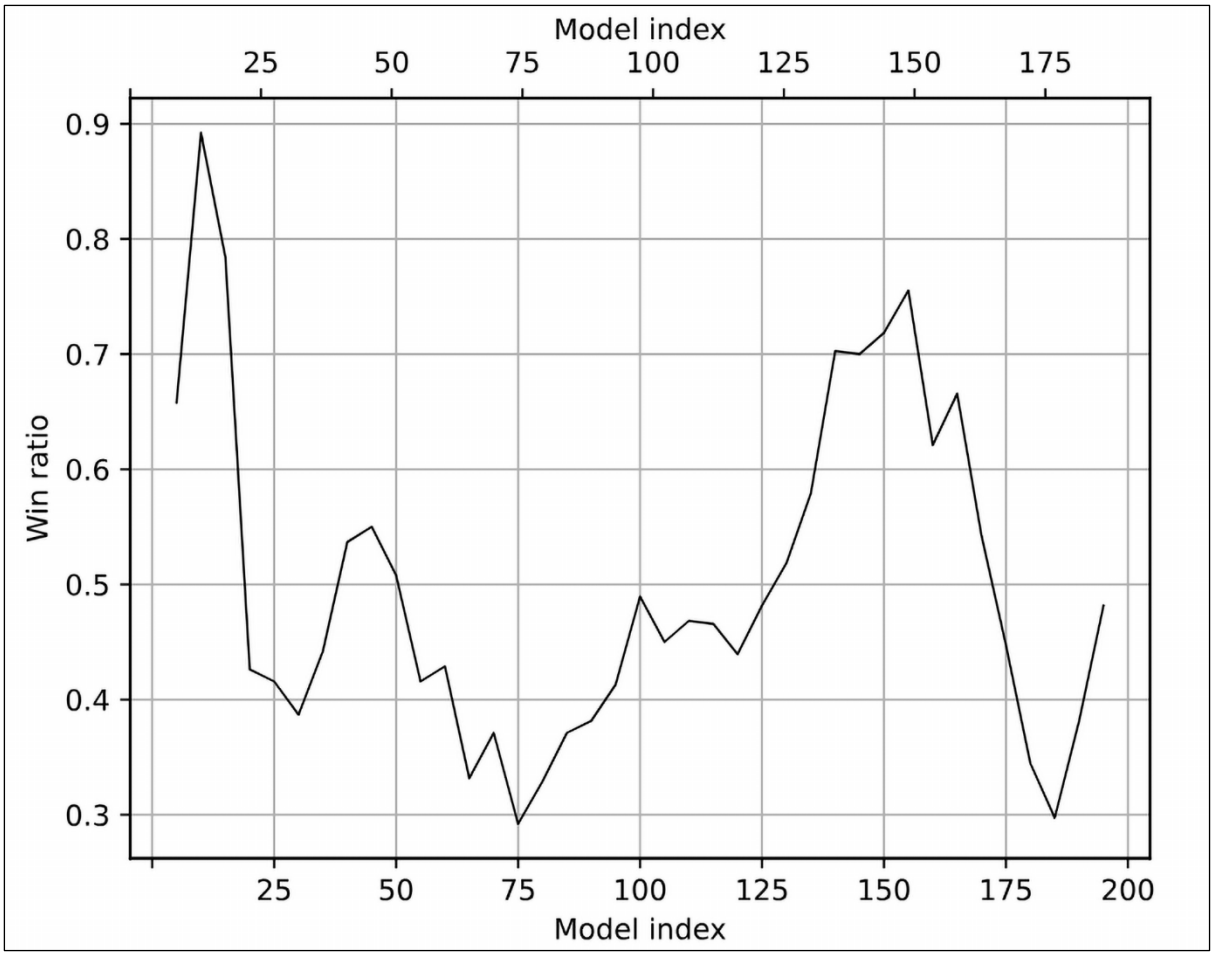



Train steps
5.00K
10.0K
15.0K
20.0K
25.0K
30.0K
35.0K
40.0K
Total loss
Hours
Train steps
10.0K
20.0K
30.0K
40.0K
50.0K
Total loss
Hours

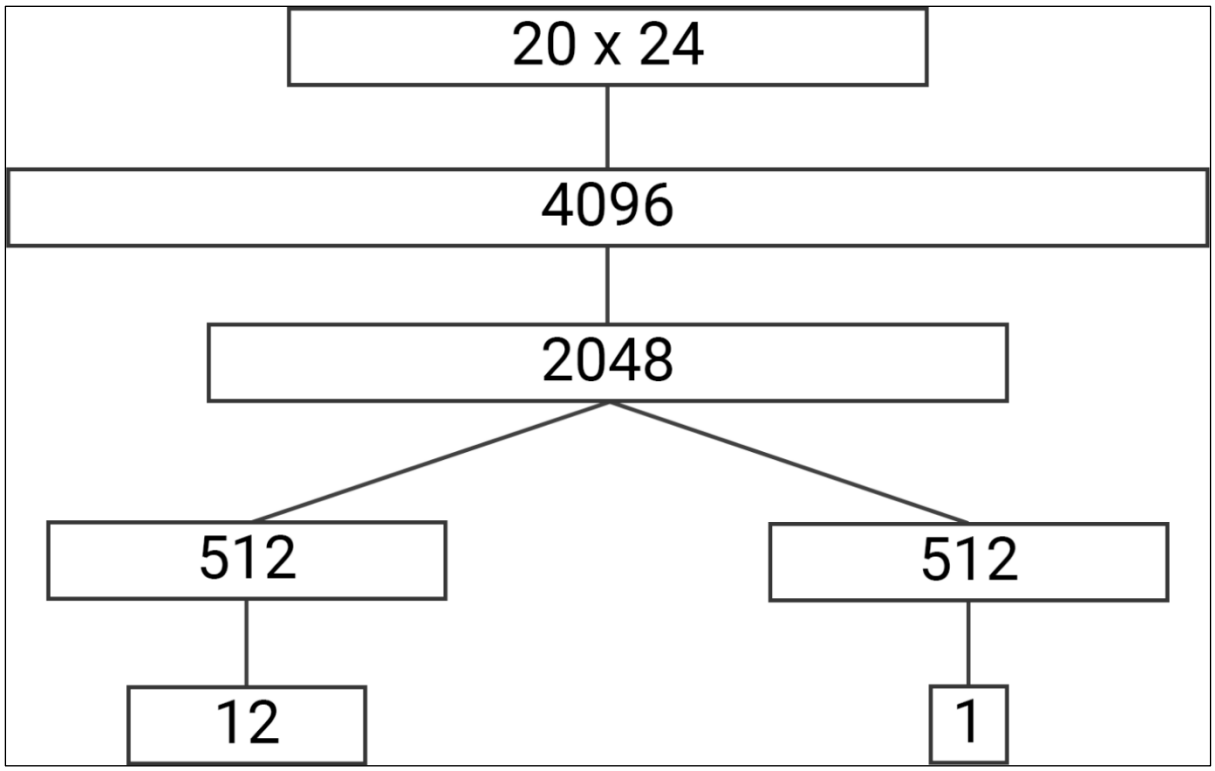

Model index
Win ratio
Model index
Model index
Win ratio
Model index

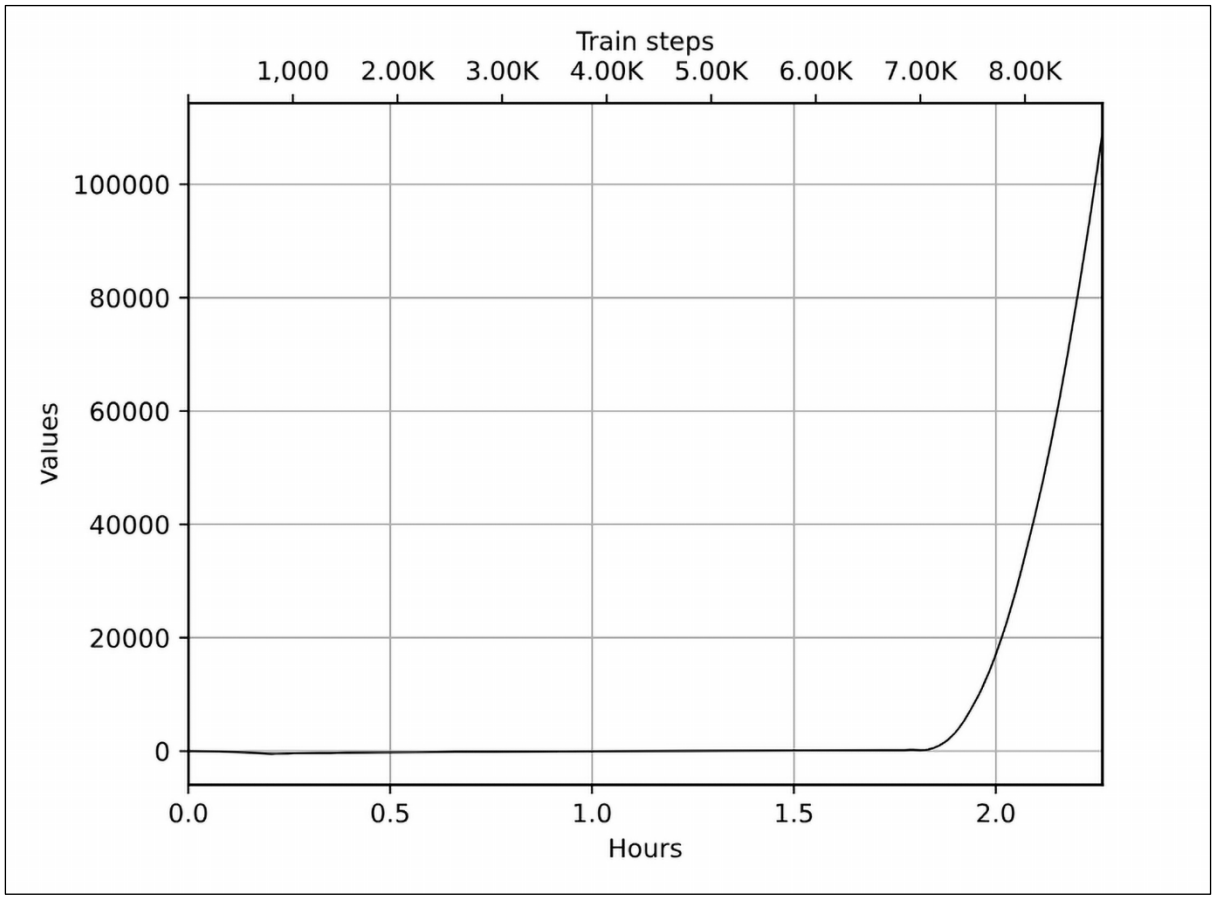

observation o
$h_\theta(o)$
$f_\theta(s^0)$
π^0, v^0
s^0
$g_\theta(s^0, a^1) \rightarrow r^1, s^1$
s^1
π^1, v^1
$f_\theta(s^1)$

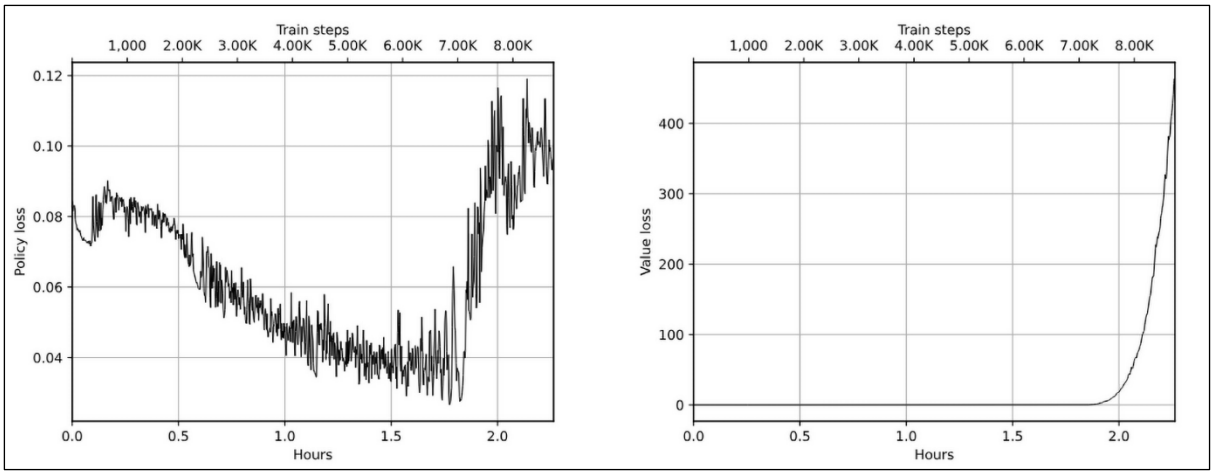

Episode
500
1,000
1.50K
2.00K
2.50K
3.00K
Policy loss
5.5
5.0
4.5
4.0
3.5
0
2
4
6
8
10
12
14
Hours
Episode
500
1,000
1.50K
2.00K
2.50K
3.00K
Value loss
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
2
4
6
8
10
12
14
Hours

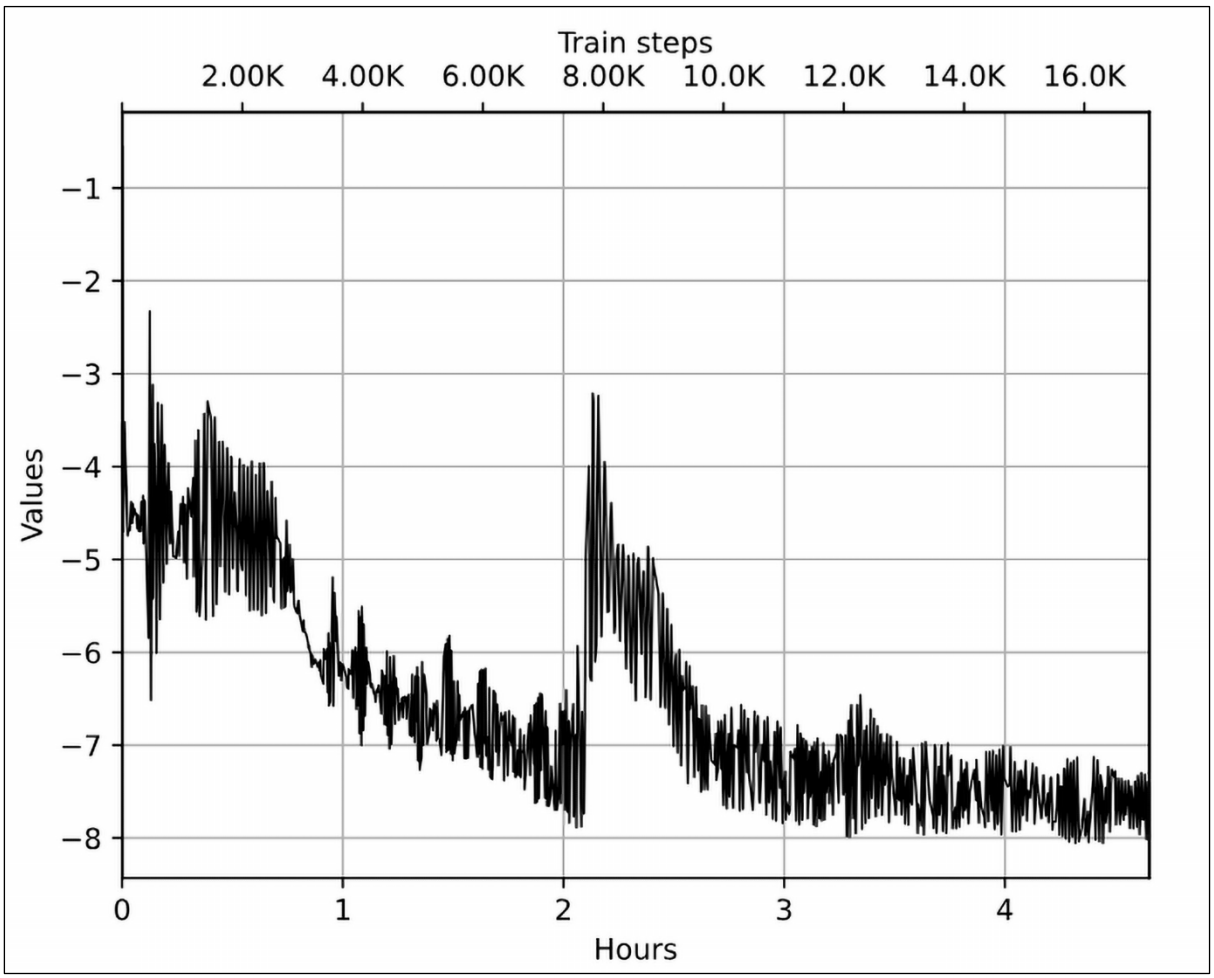

Model index
25
50
75
100
125
150
175
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Win ratio
25
50
75
100
125
150
175
200
Model index

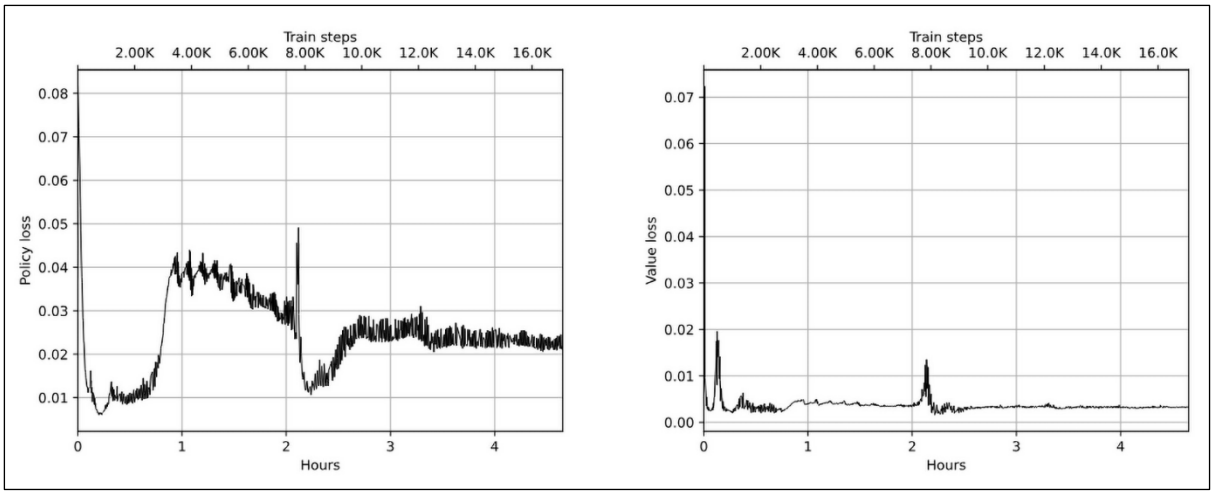
Chapter 21: RL in Discrete Optimization

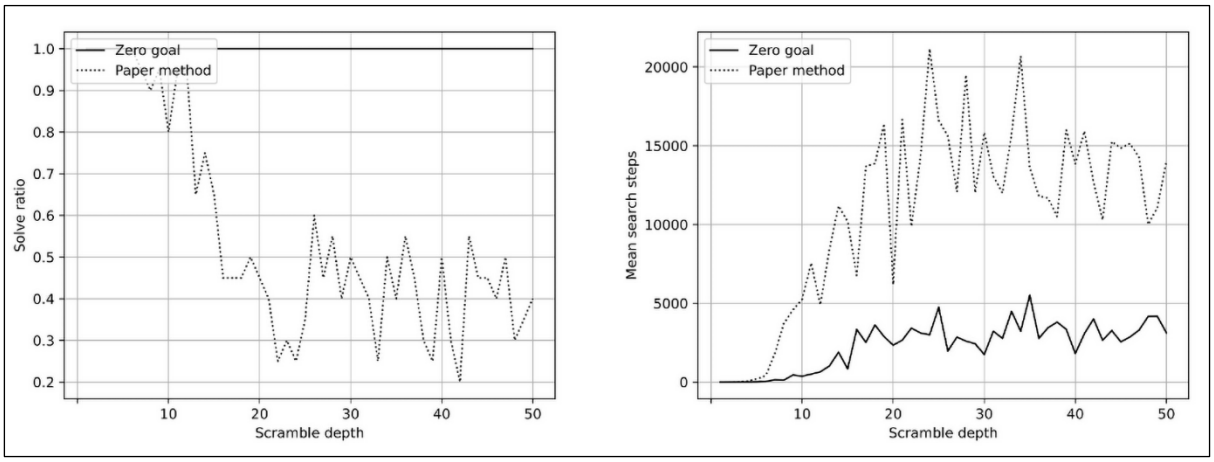


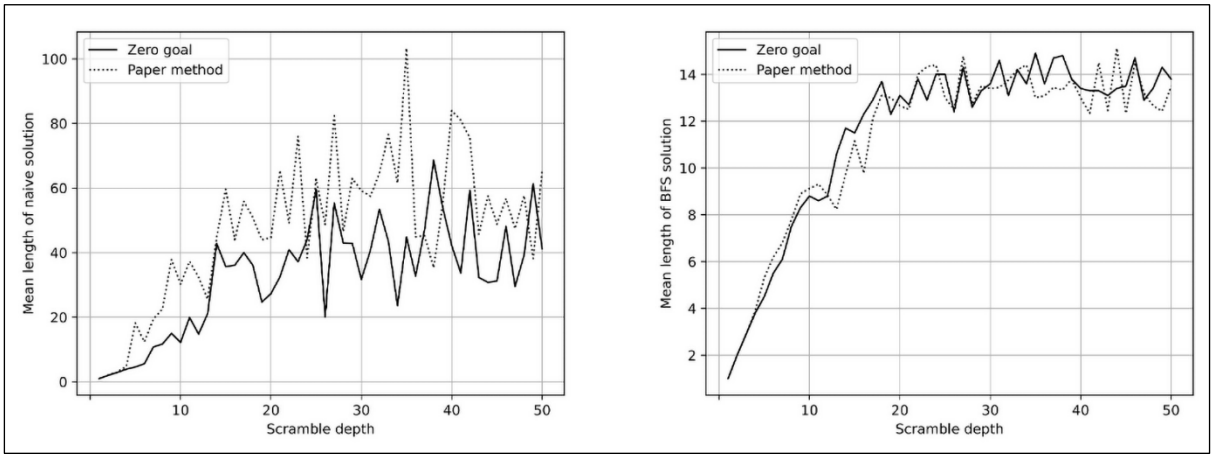


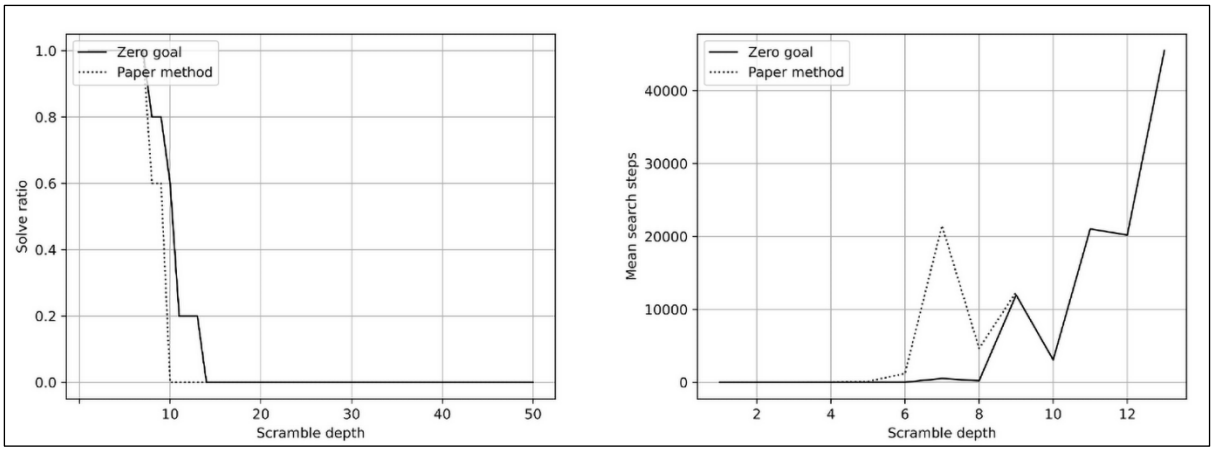

Train steps
1,000 2.00K 3.00K 4.00K 5.00K 6.00K 7.00K 8.00K
Policy loss
0.12
0.10
0.08
0.06
0.04
0.0 0.5 1.0 1.5 2.0
Hours
Train steps
1,000 2.00K 3.00K 4.00K 5.00K 6.00K 7.00K 8.00K
Value loss
400
300
200
100
0
0.0 0.5 1.0 1.5 2.0
Hours

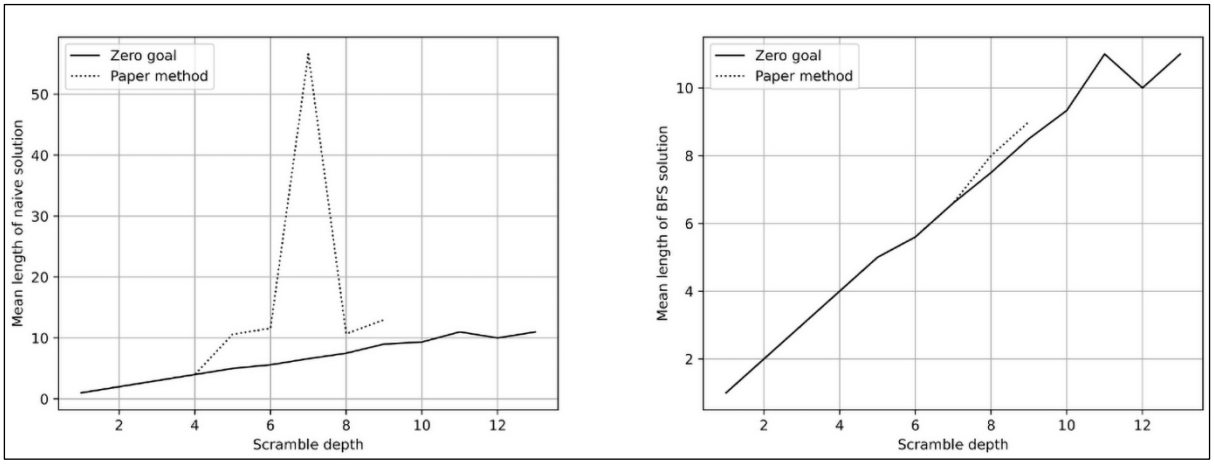

Train steps
2.00K 4.00K 6.00K 8.00K 10.0K 12.0K 14.0K 16.0K
Values
−1
−2
−3
−4
−5
−6
−7
−8
0 1 2 3 4
Hours


Train steps
2.00K 4.00K 6.00K 8.00K 10.0K 12.0K 14.0K 16.0K
Policy loss
Hours
Train steps
2.00K 4.00K 6.00K 8.00K 10.0K 12.0K 14.0K 16.0K
Value loss
Hours

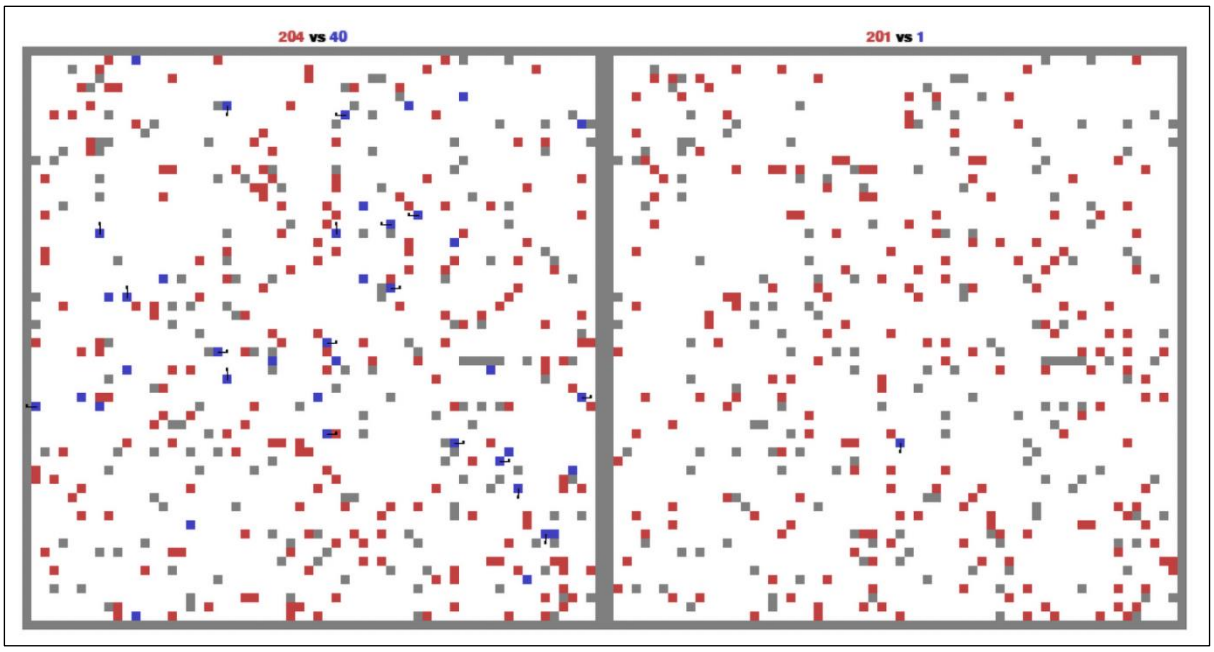

Zero goal
Paper method
Solve ratio
Scramble depth
Zero goal
Paper method
Mean search steps
Scramble depth

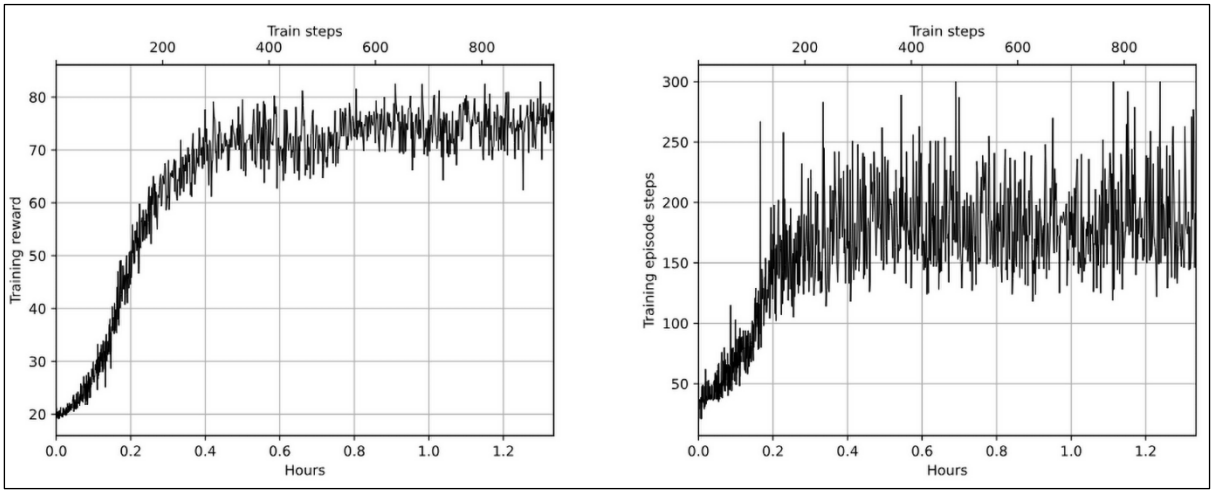

Zero goal
Paper method
Mean length of naive solution
Scramble depth
Zero goal
Paper method
Mean length of BFS solution
Scramble depth

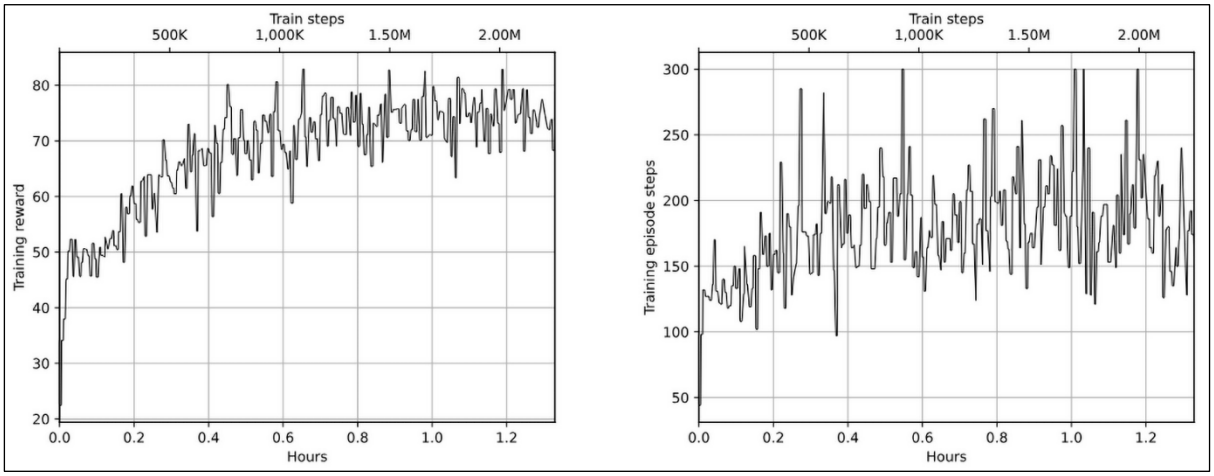

Zero goal
Paper method
Solve ratio
Scramble depth
Mean search steps

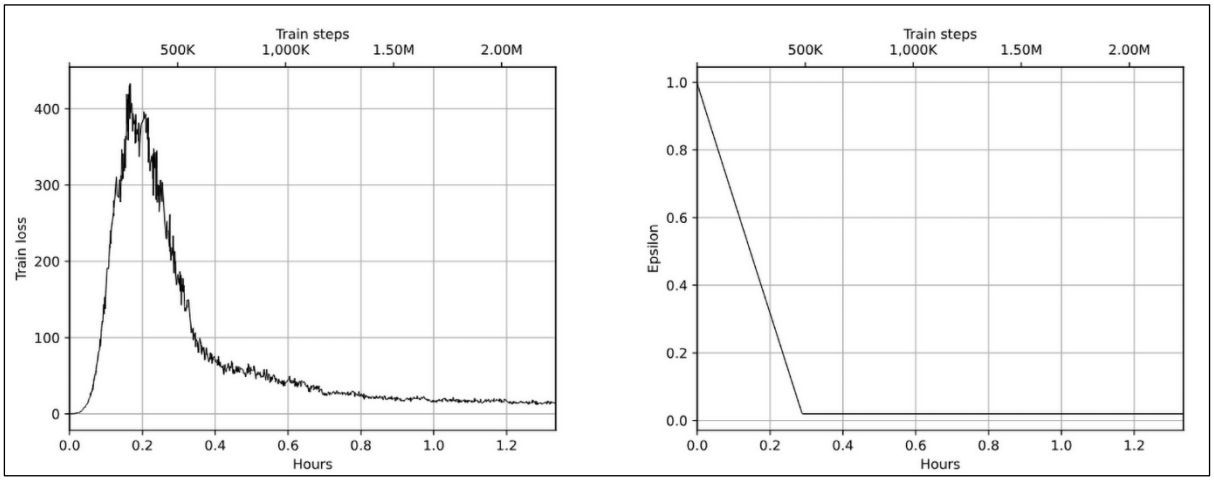

Zero goal
Paper method
Mean length of naive solution
Mean length of BFS solution
Scramble depth

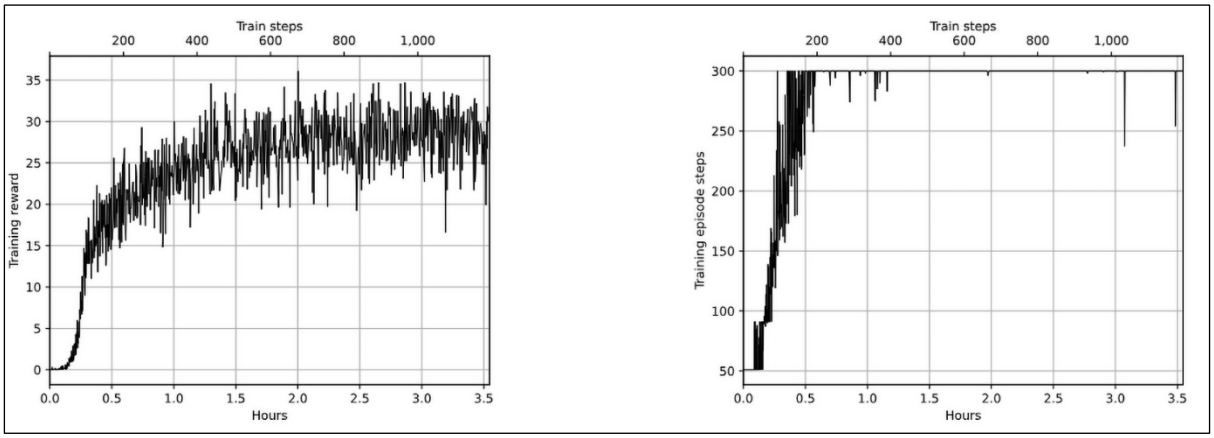
Chapter 22: Multi-Agent RL

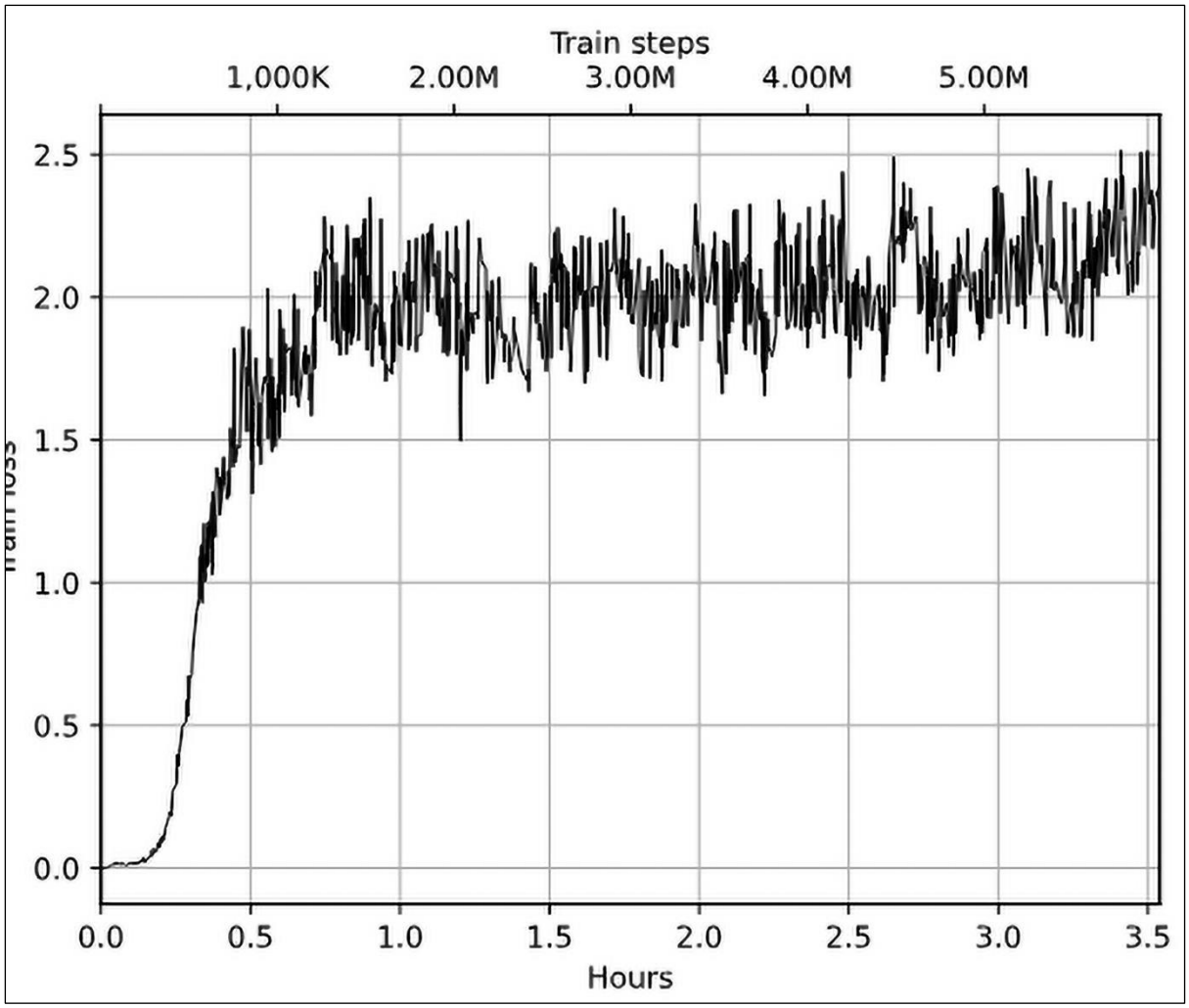


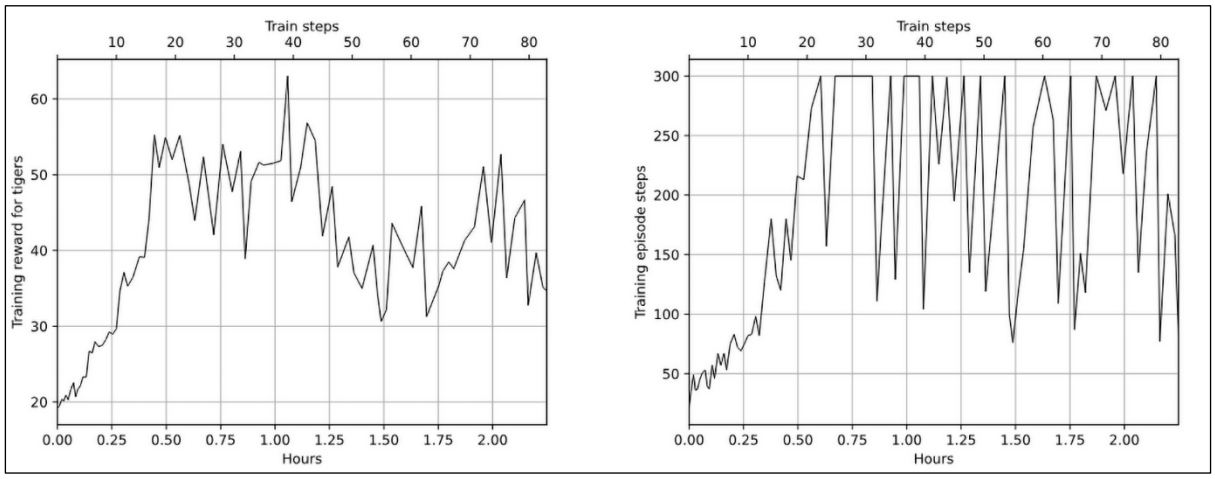


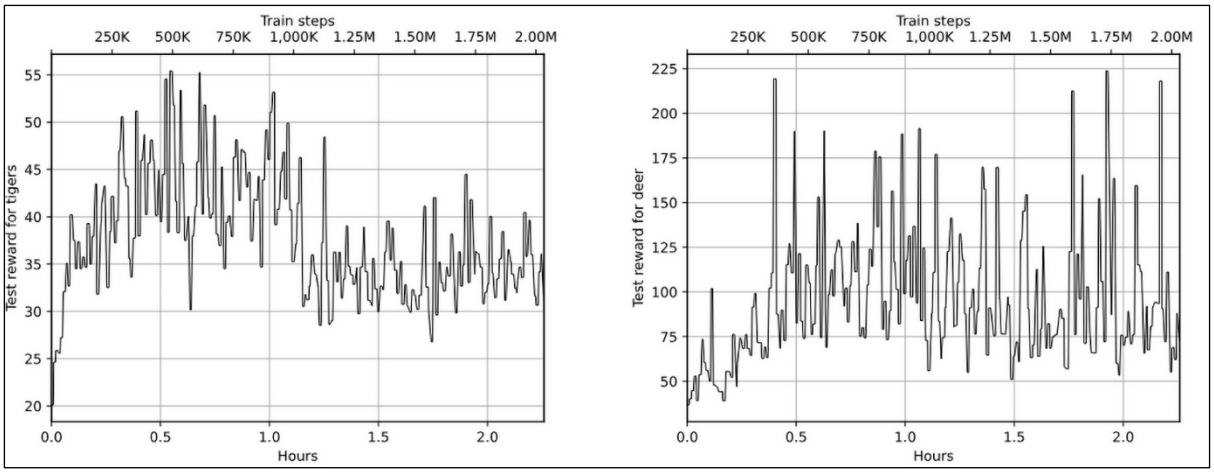


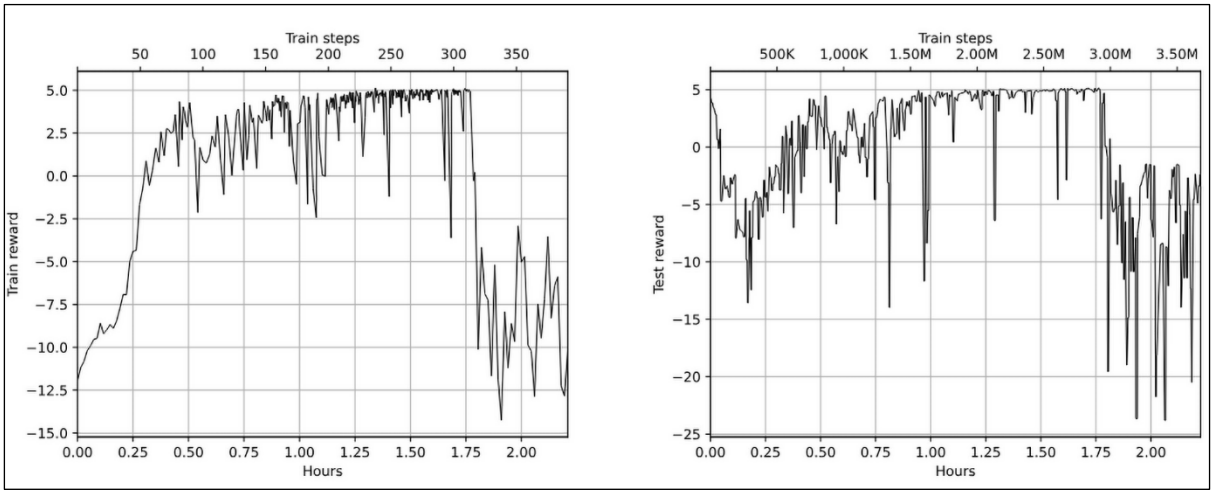

Train steps
500K
1,000K
1.50M
2.00M
Train loss
Hours
Epsilon


Train steps
200
400
600
800
1,000
Training reward
Training episode steps
Hours


Train steps
1,000K
2.00M
3.00M
4.00M
5.00M
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Hours


Train steps
10
20
30
40
50
60
70
80
Training reward for tigers
60
50
40
30
20
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Hours
Train steps
10
20
30
40
50
60
70
80
Training episode steps
300
250
200
150
100
50
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Hours


Train steps
250K 500K 750K 1,000K 1.25M 1.50M 1.75M 2.00M
Test reward for tigers
Hours
Train steps
250K 500K 750K 1,000K 1.25M 1.50M 1.75M 2.00M
Test reward for deer
Hours


Train steps
50 100 150 200 250 300 350
Train reward
Hours
Train steps
500K 1,000K 1.50M 2.00M 2.50M 3.00M 3.50M
Test reward
Hours