NTNU
Business School

Trading with Deep Reinforcement Learning
Description and Experiments with Synthetic Data

Denis M. Becker

Table of Contents

1 Stable Baselines 3 and Open AI Gym

- We apply deep reinforcement **algorithms** provided by **Stable Baselines 3** (Python library that provides a set of pre-implemented RL algorithms
- Particularly:

Continuous decision space	**Discrete decision space**
---	Deep Q-network (DQN)
Proximal Policy Optimization (PPO)	Proximal Policy Optimization (PPO)
Advantage Actor Critic (A2C)	Advantage Actor Critic (A2C)
Deep Deterministic Policy Gradient (DDPG)	---
Twin Delayed DDPG (TD3)	---
Soft Actor Critic (SAC)	---

2 Environments

- We consider **two** different **RL environments**:
  - **dichotomic** (binary) **decision space** «In» (buy or hold) and «out» (sell or stay-out)
  - **Continues actions** in interval [0,1]
- Both implemented by means of **Open AI Gym** (container that contains many useful helper functions and terminology used in SB3)
- The **RL** agent **maximizes**:
  - either total **wealth** in the end of one trading epoch,
  - or **risk-adjusted performance**.

2.1 Dichotomic Trading Actions

Initialization:

$B_{\text{Out}} = B_0, \;\; S_{\text{Out}} = 0, \;\; t = 0,$ Observe state S_0.

Loop through all periods:

The outgoing values from the previous iteration become the ingoing values:

$B_{\text{In}} = B_{\text{Out}}$

$S_{\text{In}} = S_{\text{Out}}$

Obtain new state S_t and receive the asset's return $r_{\text{Asset},t}$

Dependent on S_t take the binary decison $a \in \{0,1\}$

If $a = 1$ (we go or stay long in stocks):

If $S_{\text{In}} = 0$ (out of risky asset, and long in bonds, i.e., $B_{\text{In}} \geq 0$):

We need to buy risky asset: *

$$S_{\text{Out}} = \frac{B_{\text{In}}}{(1+\tau)} \cdot \left(1 + r_{\text{Asset},t}\right)$$

$$B_{\text{Out}} = 0$$

Else (i.e. $S_{\text{In}} > 0$ and $B_{\text{In}} = 0$ which means long in stocks, out of bonds):

We will stay in risky asset:

$$S_{\text{Out}} = S_{\text{In}} \cdot \left(1 + r_{\text{Asset},t}\right)$$

$$B_{\text{Out}} = 0$$

If $a = 0$ (we sell stocks or stay out):

If $S_{\text{In}} = 0$ (out of stocks, and long in bonds, i.e., $B_{\text{In}} \geq 0$):

$$S_{\text{Out}} = 0$$

$$B_{\text{Out}} = B_{\text{In}} \cdot (1 + r_{\text{B}})$$

Else (i.e., $S_{\text{In}} > 0$ and $B_{\text{In}} = 0$ which means long in stocks, out of bonds):

We need to sell stocks: *

$$B_{\text{Out}} = S_{\text{In}} \cdot (1 - \tau) \cdot (1 + r_{\text{B}})$$

$$S_{\text{Out}} = 0$$

2.2 Continuous Trading Actions

- Proportion/**share of risky asset after trading** and **after transaction costs** is denoted as x.
- Short sales are not allowed, i. e. x is continuous in the interval $x = [0,1]$.
- **Transaction costs** are a **percentage** of value of sold/bought stocks.

Initialization:

$B_{\text{Out}} = B_0,\ S_{\text{Out}} = 0,\ t = 0,$ Observe state S_0.

Loop through all periods:

The outgoing values from the previous iteration become the ingoing values:

$B_{\text{In}} = B_{\text{Out}}, S_{\text{In}} = S_{\text{Out}}$

Obtain new state S_t and receive the asset's return $r_{\text{Asset},t}$.

Dependent on S_t take the continuous decison $a \in [0,1]$

Transform the share in the risky asset, after transaction costs.

$$\Delta S = \frac{B_{\text{In}} - \left(\frac{1}{x} - 1\right) \cdot S_{\text{In}}}{\frac{1}{x} + \tau}$$

Calculate the outgoing values of the risky asset and bank account:

$$B_{\text{Out}} = (B_{\text{In}} - \Delta S \cdot (1 + \tau)) \cdot (1 + r_B)$$

$$S_{\text{Out}} = (S_{\text{In}} + \Delta S) \cdot (1 + r_S)$$

3 Experiments with Synthetic Time Series

- All experiments are run on a virtual Windows machine created in OpenStack with the following specifications:

Operating system:	Windows Server 2022 Standard, version 21H2, OS build: 20348.887
Processor:	Intel Xeon E312xx (Sandy Bridge, IBRS update) 2.60 GHz (2 processors)
Installed RAM:	32.0 GB
System type:	64-bit operating system, x64-based processor

3.1 Experiment #1 - Deterministic, repetitive sequence

- Time series has the following form:

Time	1	2	3	4	5	6	7	8	9	10	11	12	Rep
Return	2 %	-2 %	1 %	2 %	1 %	4 %	-1 %	5 %	-2 %	4 %	-1 %	-2 %	2 % …

- Repeated four times when training: 12 x 4 = 48 timesteps
- Optimal overall return when starting with $B_0 = 10$ => 8.34
- Observation in t consists of only r_t (no information on earned balance, lagged returns, etc.)
- Environment #1: Continues action space in interval [0,1]
- Environment #2: Dichotomic action space: 0 or 1

- Transition probabilities, expected state returns and optimal actions:

		to -2%	to -1 %	to 1 %	to 2 %	to 4 %	to 5 %
from	-2 %			1/3	1/3	1/3	
from	-1 %	1/2					1/2
from	1 %				1/2	1/2	
from	2 %	1/2		1/2			
from	4 %		1				
from	5 %	1					

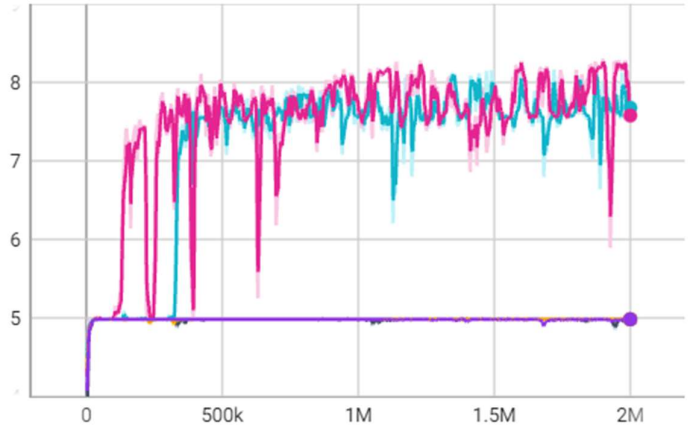
$\mathbb{E}[r_{t+1}]$	Optimal action
>0	Long
>0	Long
>0	Long
<0	Out
<0	Out
<0	Out

- Perfect information is given if states contain information (r_{t-2}, r_{t-1}). It can be represented by the following transition matrix.

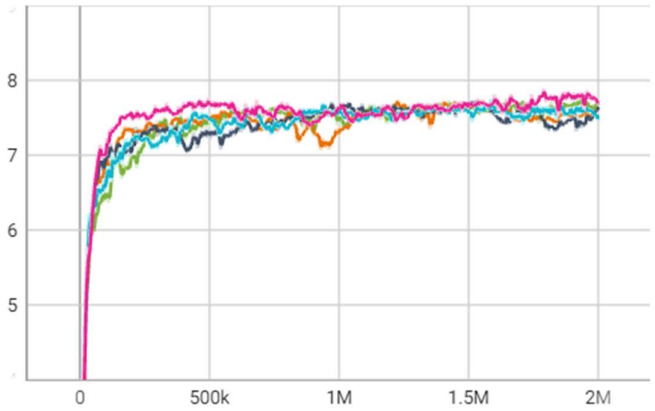
		to 2%, -2%	to -2 %, 1 %	to 1 % 2 %	to 2 % 1 %	to 1 % 4 %	to 4 % -1 %	to -1 % 5 %	to 5 % -2 %	to -2 % 4 %	to 4 % -1 %	to -1 % -2 %	to -2 % 2 %	$\mathbb{E}[r_{t+1}]$	Optimal action
from	2%, -2%		1											1 %	Long
from	-2 %, 1%			1										2 %	Long
from	1 %, 2%				1									1 %	Long
from	2 %, 1 %					1								4 %	Long
from	1 %, 4 %						1							-1 %	Out
from	4 %, -1 %							1						5 %	Long
from	-1 %, 5%								1					-2 %	Out
from	5 %, -2 %									1				4 %	Long
from	-2 %, 4 %										1			-1 %	Out
from	4 %, -1 %											1		-2 %	Out
from	-1 %, -2 %												1	2 %	Long
from	-2 %, 2 %	1												-2 %	Out

Time	1	2	3	4	5	6	7	8	9	10	11	12	Rep
Return	2 %	-2 %	1 %	2 %	1 %	4 %	-1 %	5 %	-2 %	4 %	-1 %	-2 %	2 % …

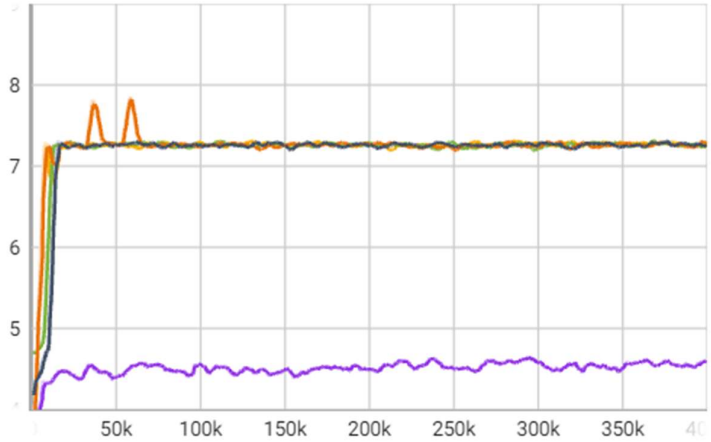
A2C



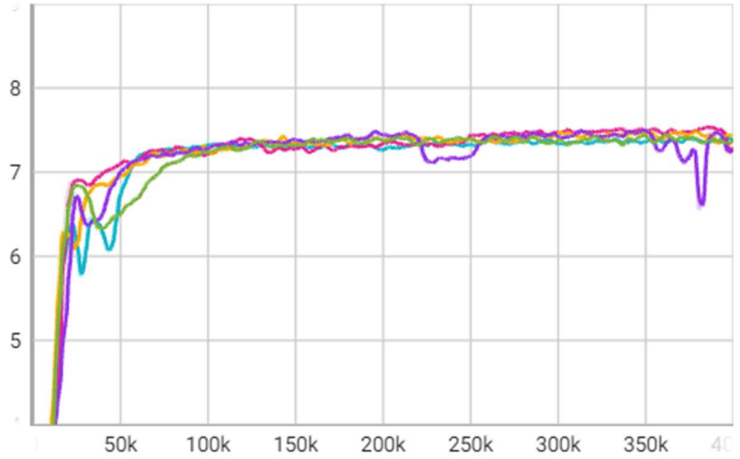
PPO



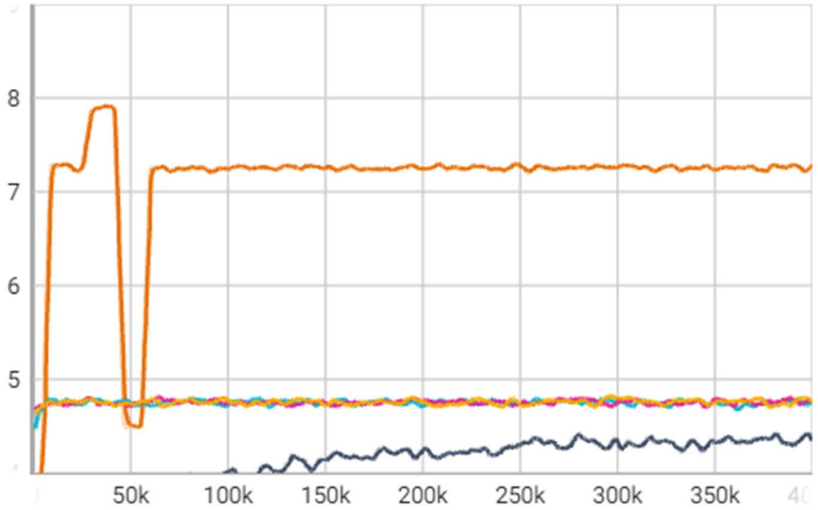
DDPG



SAC



TD3

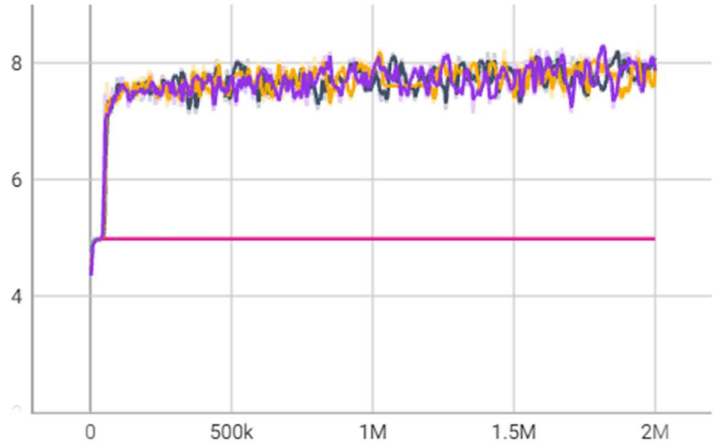


A2C • 2,000,000 time steps • 5 runs (approximately 54 min per run) • 2 out of 5 runs performed well	**PPO** • 2,000,000 time steps • 5 runs (approximately 51 min per run) all runs performed well, but not perfect
DDPG • 400,000 time steps • 5 runs (approximately 126 min per run) • 4 of 5 runs performed well, but not perfect.	**SAC** • 400,000 time steps • 5 runs (approximately 3.5 hours per run) • All runs performed well, but not perfect.

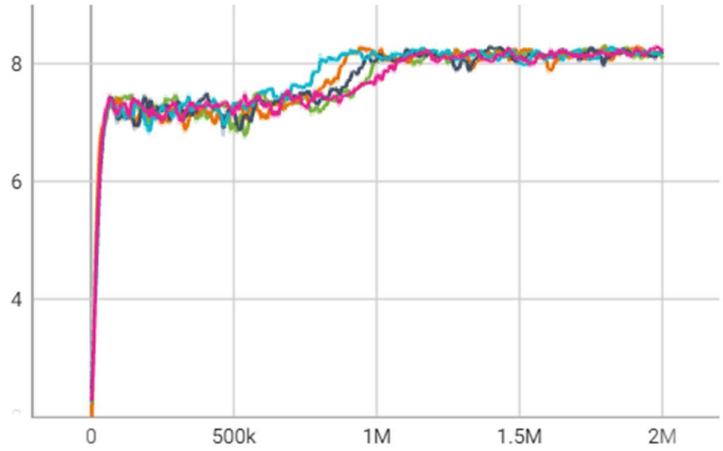
TD3

- 400,000 time steps
- 5 runs (approximately 2.7 hours per run)
- except for one run, all other runs performed badly.

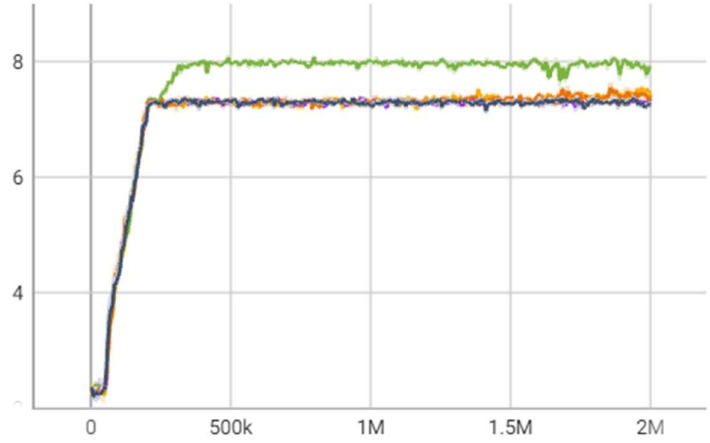
A2C (dichotomic actions)



PPO (dichotomic actions)

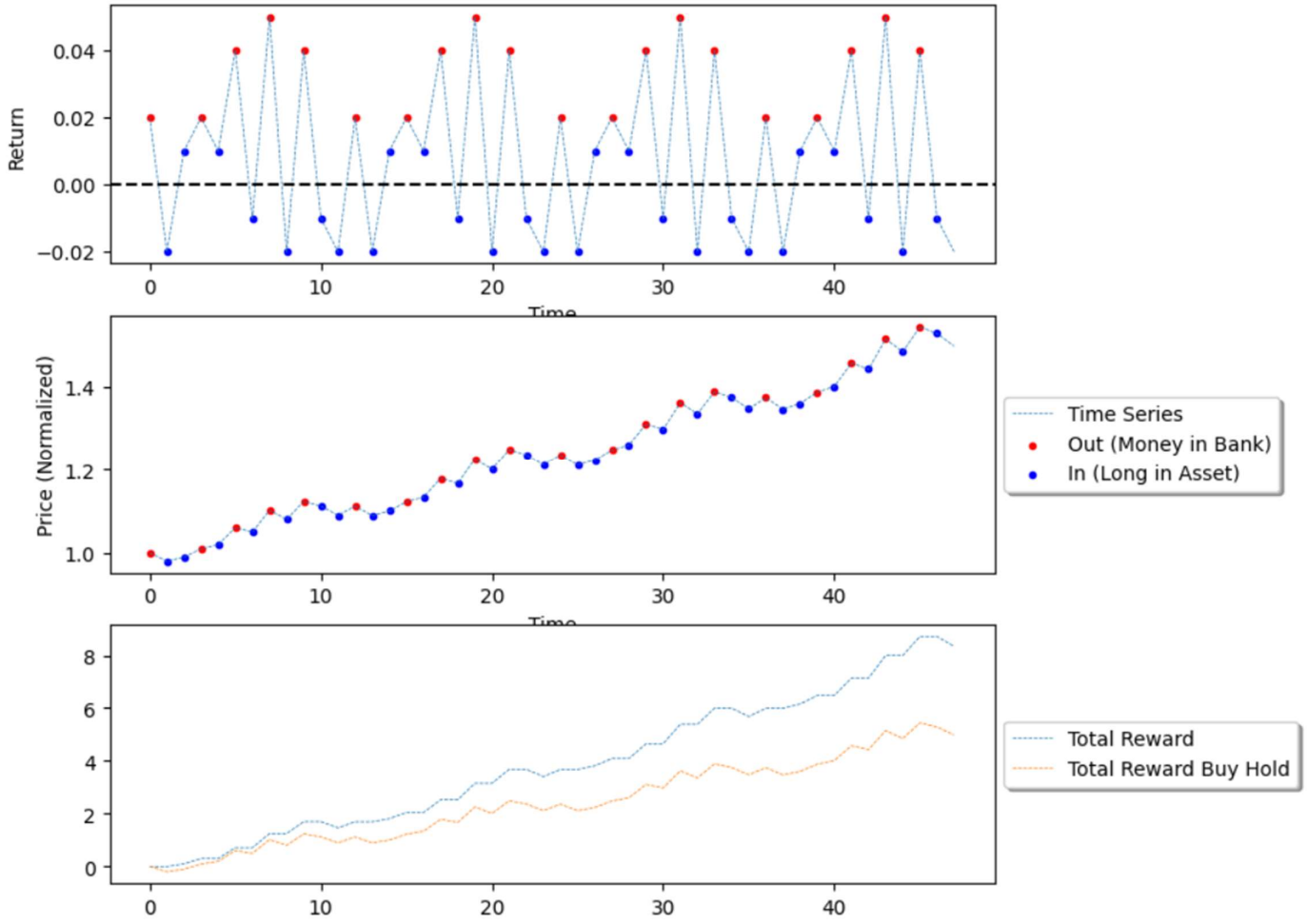


DQN (dichotomic actions)



A2C (dichotomic actions) • 2,000,000 time steps • 5 runs (approximately 55 min per run) • 3 of 5 runs performed well but not perfect (one state was wrong).	**PPO (dichotomic actions)** • 2,000,000 time steps • 5 runs (approximately 52 min per run) • all runs perform perfect.
DQN (dichotomic actions) • 2,000,000 time steps • 5 runs (approximately 38 min per run) • 1 of 5 worked perfect, the remaining runs performed well (one state was wrong)	

PPO on Training Set



3.2 Experiment #2 – Deterministic Sine-Wave-Returns

- Time series: Deterministic sine-wave of the form:

$$r = 0.01 \cdot \sin(0.1 \cdot t) + 0.002$$

  where $t = 1,2,\ldots$
- Observation in t consists of only r_t (no information on intermediate profits, lagged returns, etc.)
- Environment #1: Continues action space in interval [0,1]
- Environment #3: Dichotomic action space: 0 or 1

3 Experiments

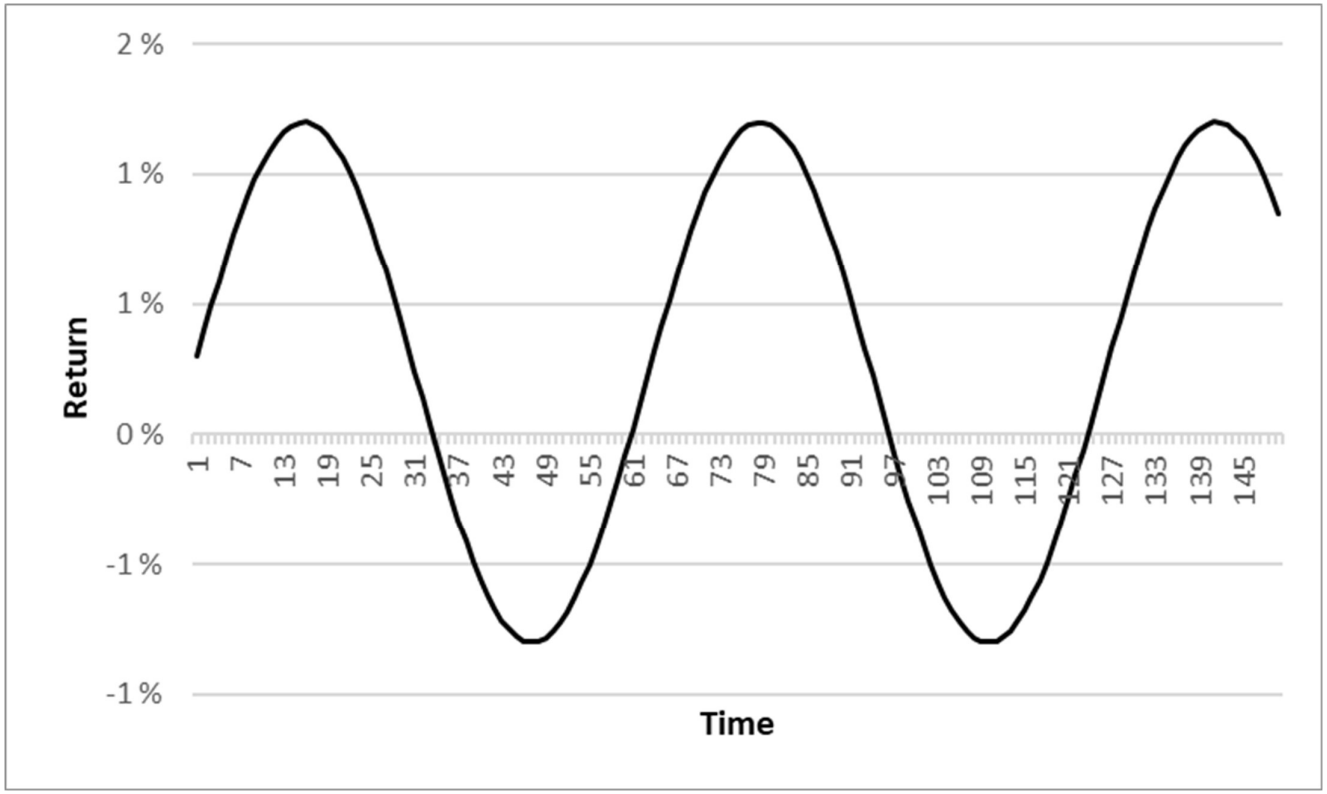


Figure 1: Sine-Wave Return (Deterministic, Discretized)

- **Transition probabilities**, expected state returns and optimal actions (theoretically if there were 100 positive returns, followed by 0%, followed by 100 negative returns, the dataset with discrete points does not have any returns with 0%):

		to r > 0%	*to* r = 0 %	*to* r < 0 %
from	r > 0 %	99 %	1 %	
from	r = 0 %	50 %		50 %
from	r < 0 %		1 %	99 %

$\mathbb{E}[r_{t+1}]$	Optimal action
>0	Long
0	Indifferent
<0	Out

Figure 2: Transition probabilities for (discretized) sine wave

- Training dataset consists of 135 returns (time steps).

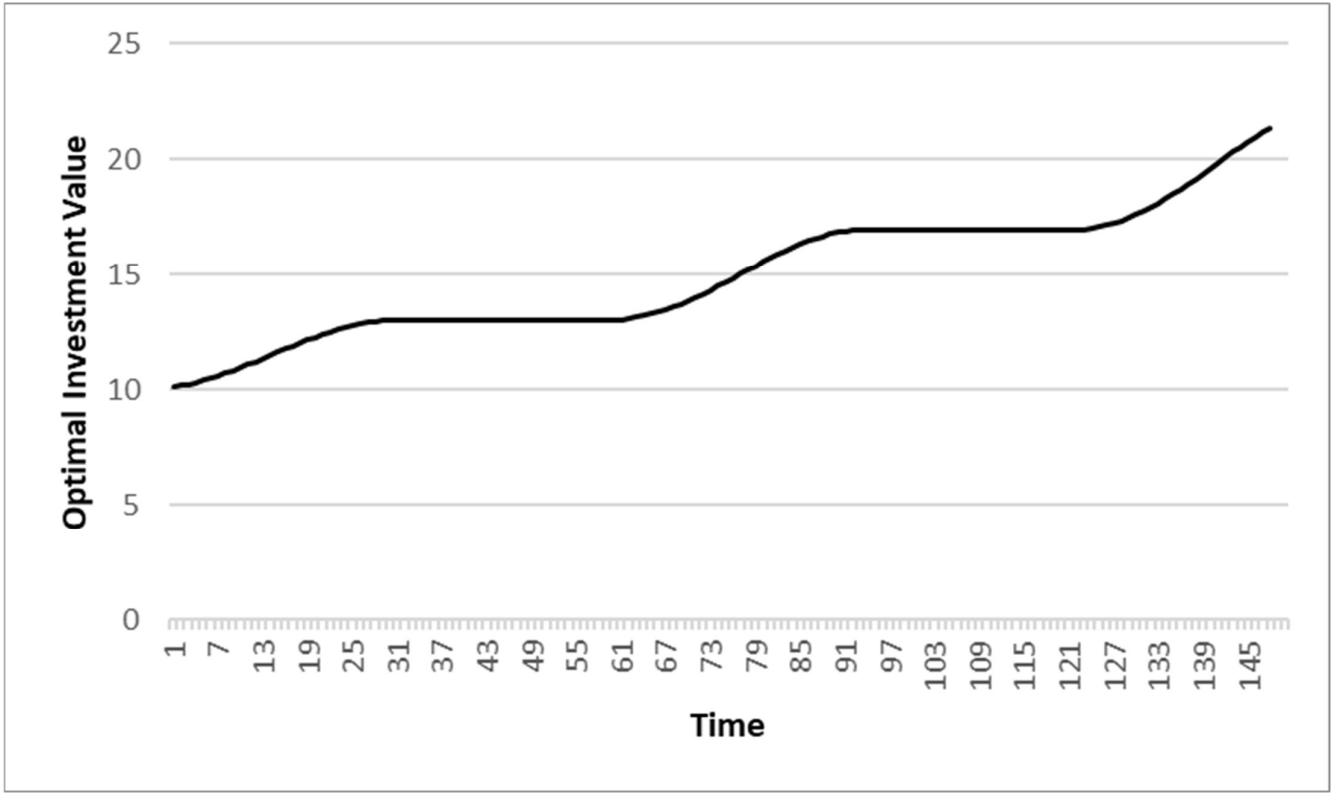
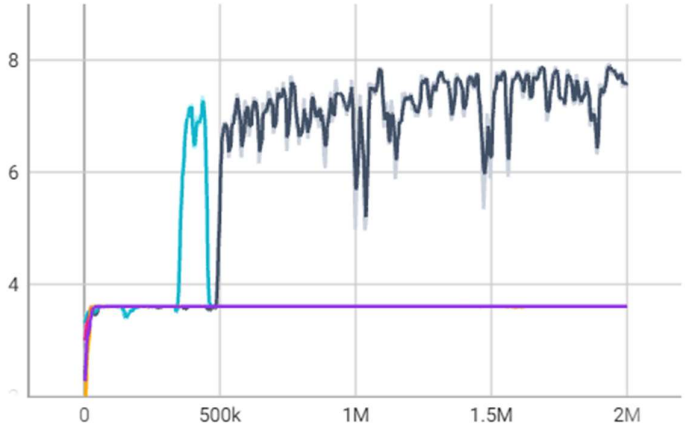


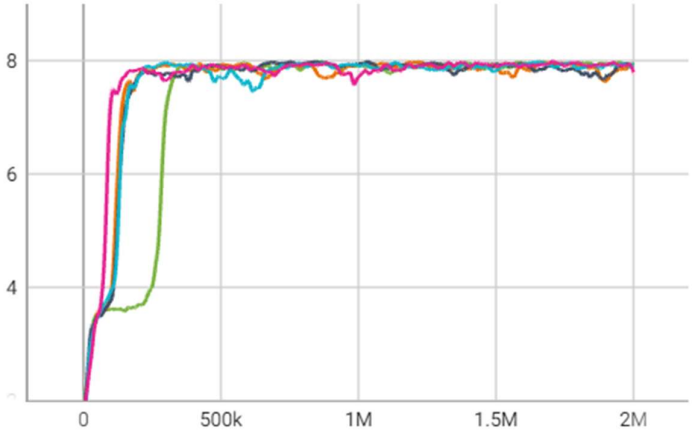
Figure 3: Optimal Trajectory of investment for sine-wave returns

- Optimal overall return when starting with $B_0 = 10$ => 8.061

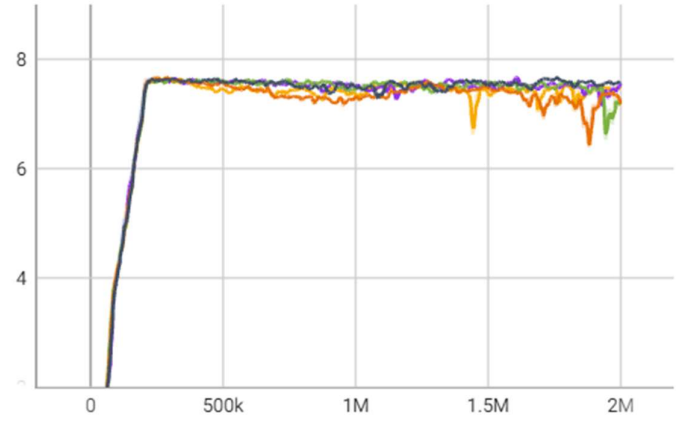
A2C (dichotomic actions)



PPO (dichotomic actions)

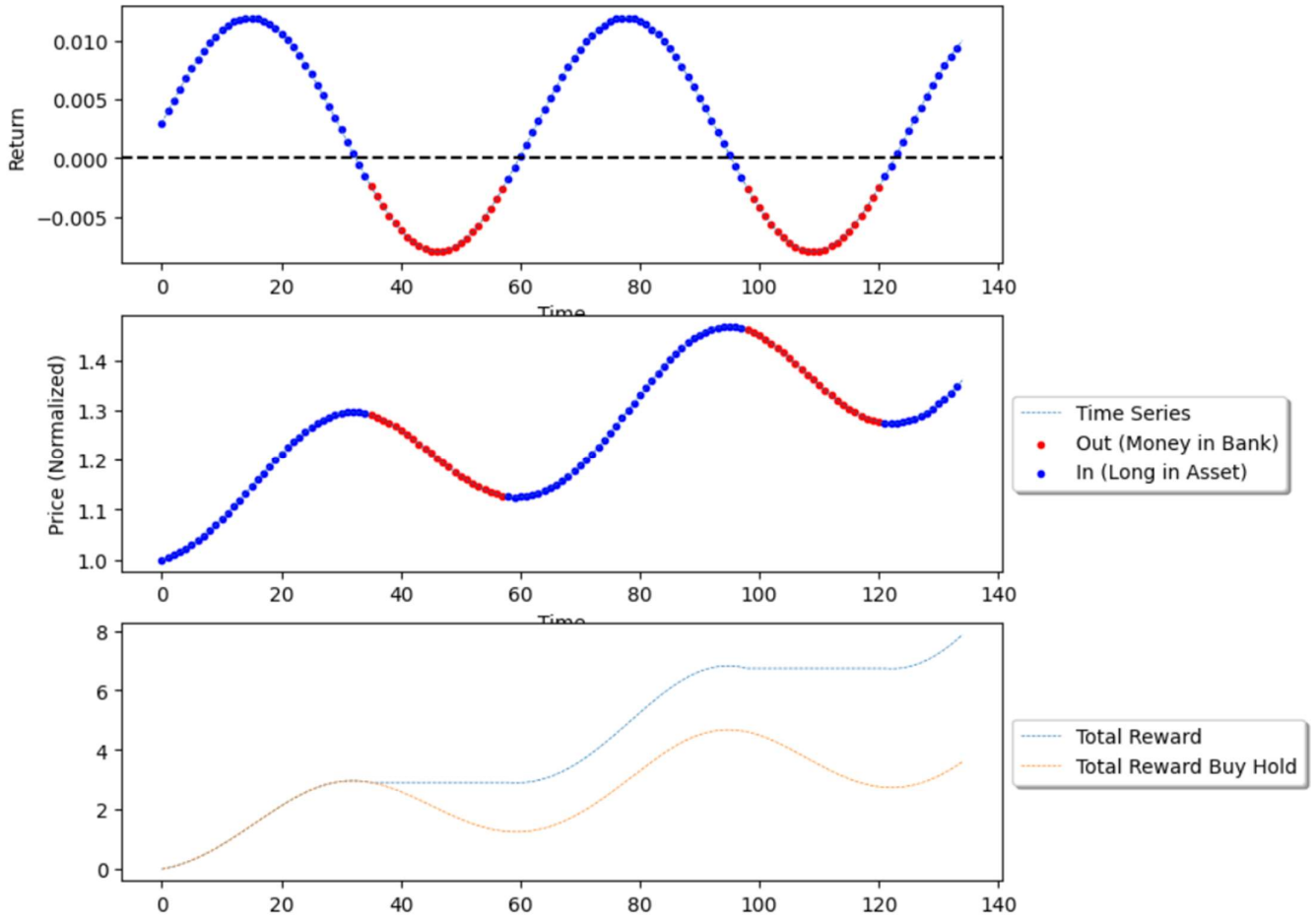


DQN (dichotomic actions)

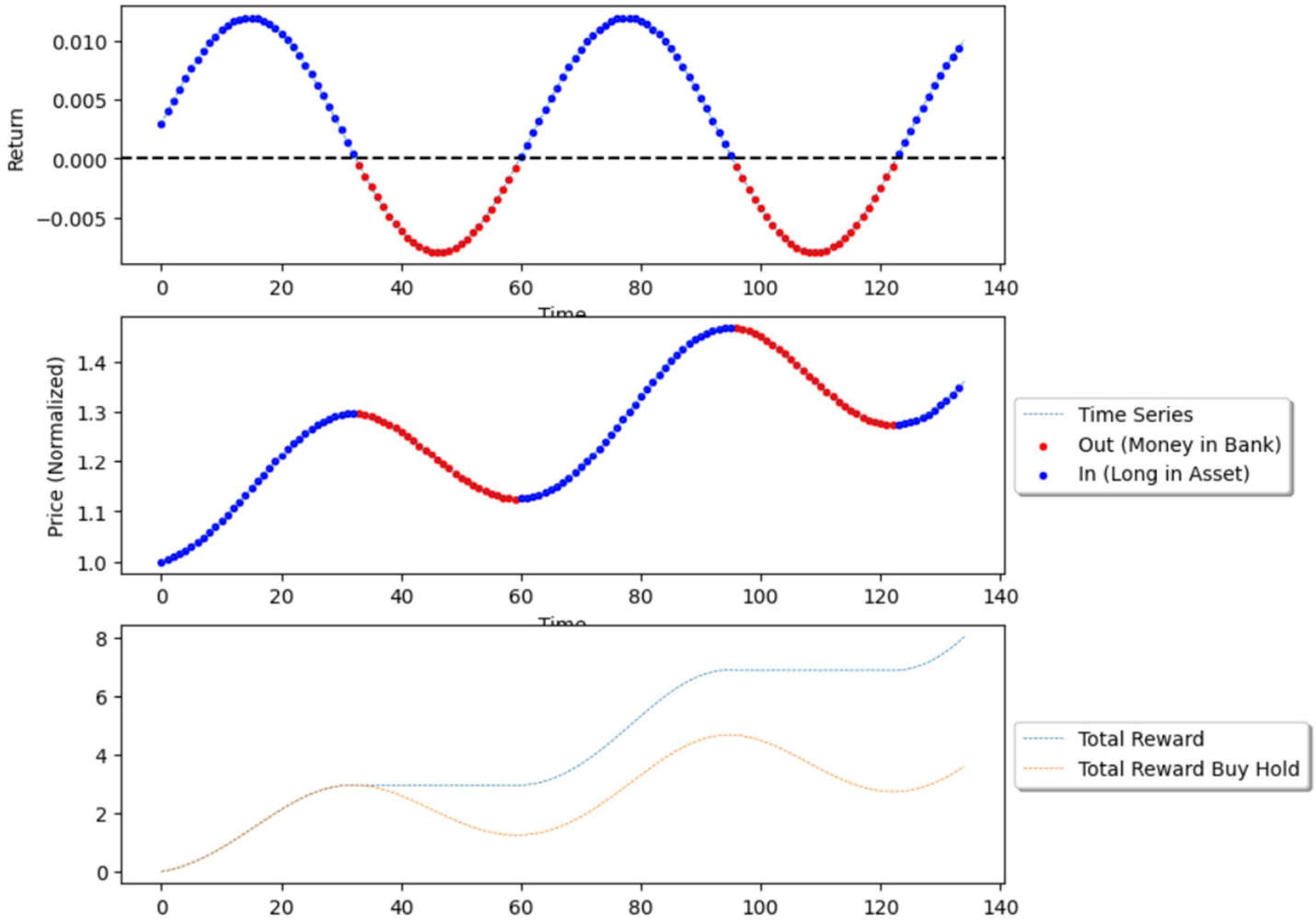


A2C (dichotomic actions)	**PPO (dichotomic actions)**
• 2,000,000 time steps • 5 runs (ca. 53 min per run) • 1 of 5 runs performed well but not perfect (one state wrong). The best model's total return is 7.594.	• 2,000,000 time steps • 5 runs (ca. 51 min per run) • All runs performed well with the best model giving a total return of 7.97.
DQN (dichotomic actions) • 2,000,000 time steps • 5 runs (ca. 33.5 min per run) • All runs performed well, the best model achieved 7.596.	

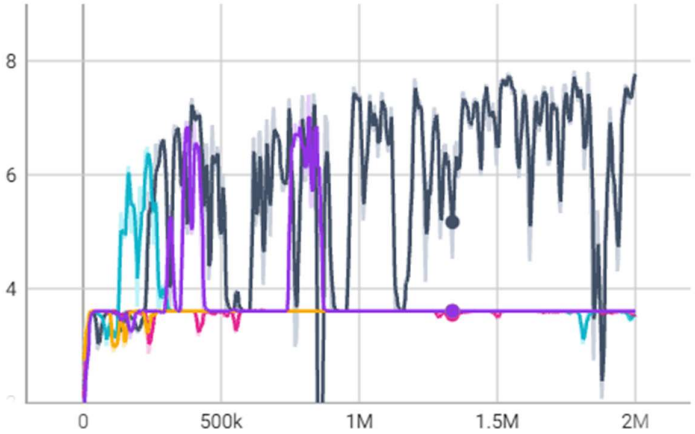
DQN on Trainingsset



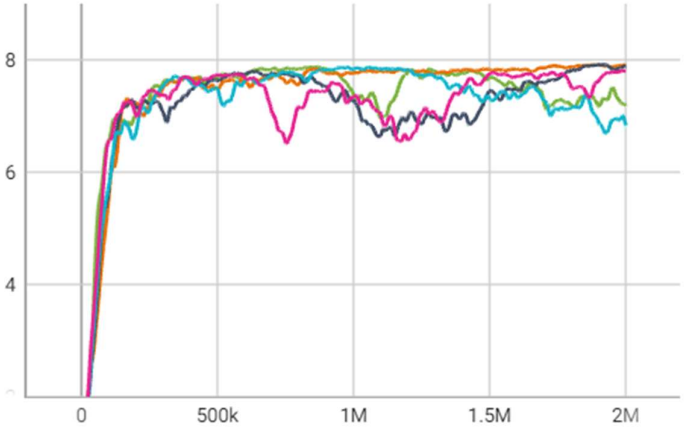
PPO on Trainingsset



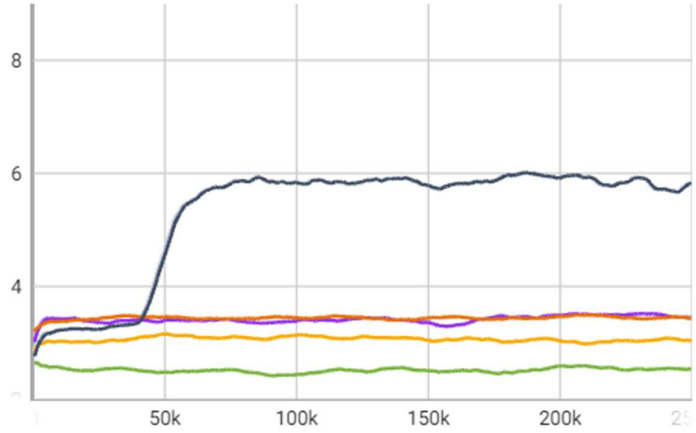
A2C (Continuous)



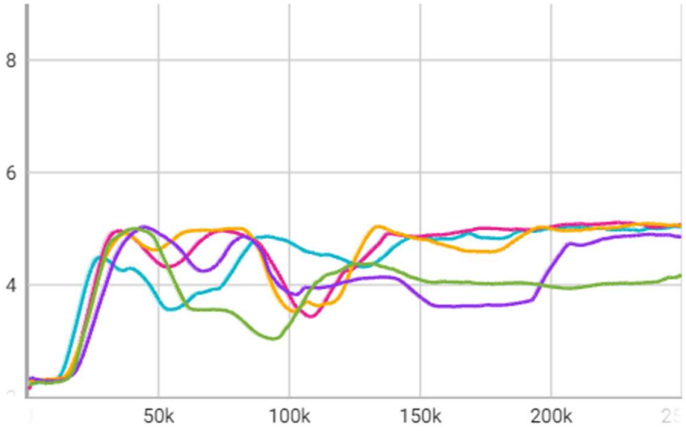
PPO (Continuous)



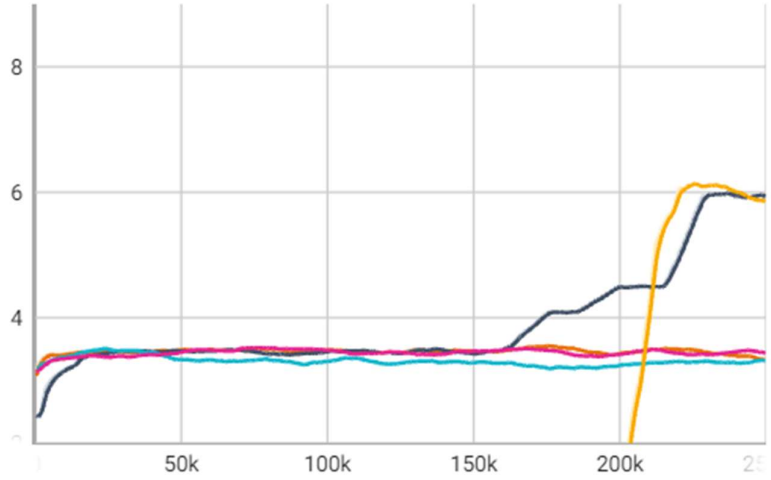
DDPG (Continuous)



SAC (Continuous)



TD3 (Continuous)

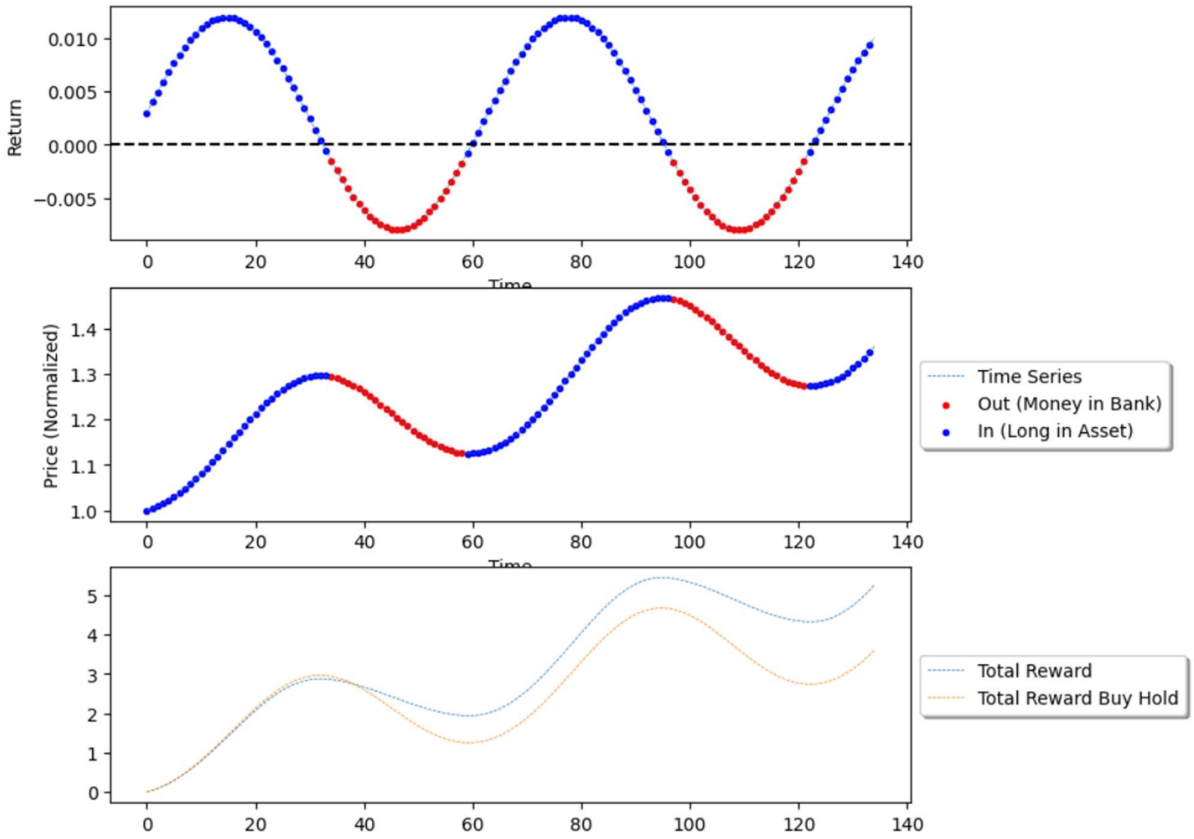


A2C (Continuous) • 2,000,000 time steps • 5 runs (ca. 54 min per run) • 1 of 5 runs performed well but not perfect. The best model's total return is 7.719.	**PPO (Continuous)** • 2,000,000 time steps • 5 runs (ca. 48 min per run) 5 runs performed well but not perfect. The best model's total return is 7.904.
DDPG (Continuous) • 250,000 time steps • 5 runs (ca. 1.18 hours per run) • All runs not satisfying. The best model's total return is 5.819.	**SAC (Continuous)** • 250,000 time steps • 5 runs (ca. 1.9 hours per run) • All runs not very satisfying. The best model's total return is 5.069.

TD3 (Continuous)

- 250,000 time steps
- 5 runs (ca. 1.3 hours per run)
- All runs not satisfying. The best model's total return is 5.912.

SAC Continuous – Training Set



Note: Red points indicate less thatn 50% long; blue points indicate more than 50% long. Downward trend indicates some long position (despite indicated red) that loses value.

3.3 Experiment #3 – Geometric Brownian Motion and Regime Shifts

- Geometric Brownian Motion in Discrete Time:

$$\Delta S_t = \mu \cdot S_t \cdot \Delta t + \sigma \cdot S_t \cdot \Delta W \quad \text{where} \quad \Delta W \sim \mathcal{N}(0, \Delta t)$$

- Transition matrix for regime shifts:

	h = 0	h = 1
h = 0	0,99	
h = 1		0,97

where 1 indicates downward trend, and 0 indicates upward trend.

$$\mu = (1 - h) \cdot \mu_{\text{Up}} + h \cdot \mu_{\text{Down}}$$
$$\sigma = (1 - h) \cdot \sigma_{\text{Up}} + h \cdot \sigma_{\text{Down}}$$

- Theoretically shifts are not predictable from the r_t, because shifts do not depend on it.

- Shifts may be identified in the hindsight, when the average return becomes visible.

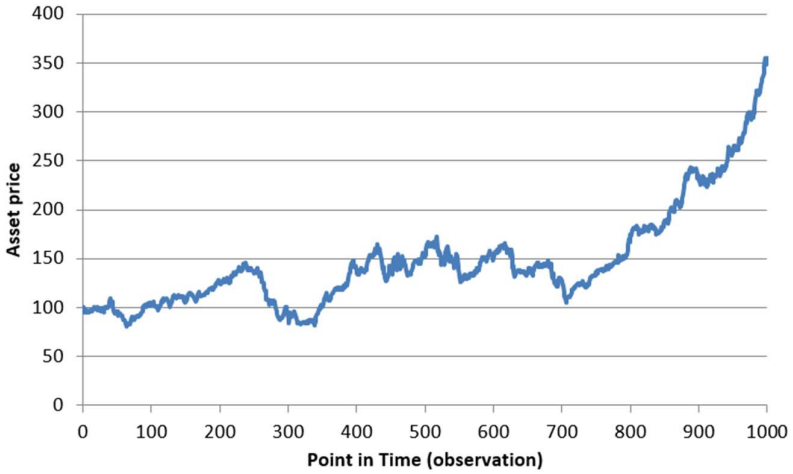


Figure 4: Observed trajectory for GBM with regime shifts

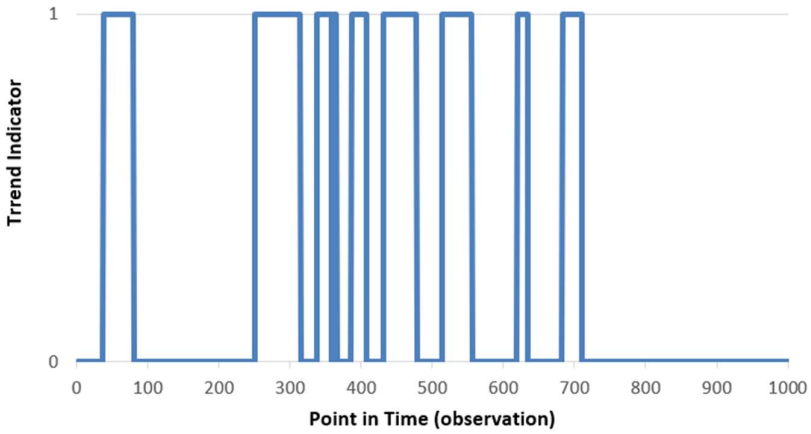


Figure 5: Observed Regime Shifts

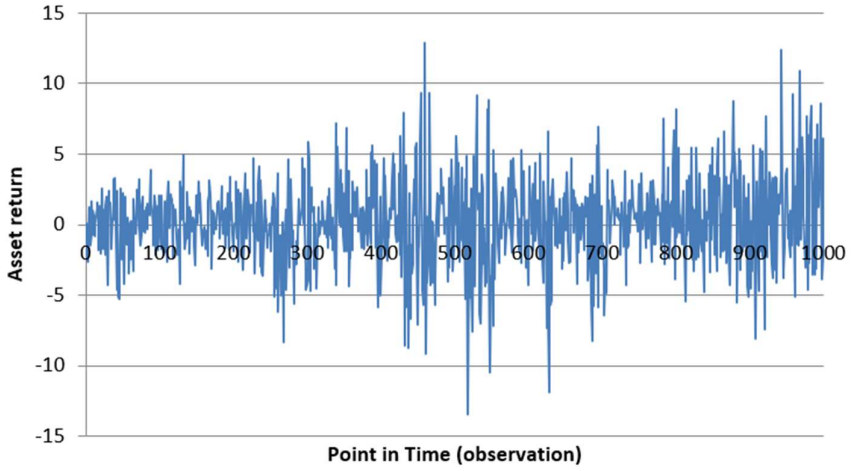


Figure 6: Observed returns for GBM with regime shifts

State space = $\{\boldsymbol{r_t}\}$

- The trend cannot be spotted (identified) with r_t as the only information.
- r_{t+1} cannot be predicted.
- The optimal decision of reward maximizer: always in if whole series trends upwards, always out if whole series trends downwards.

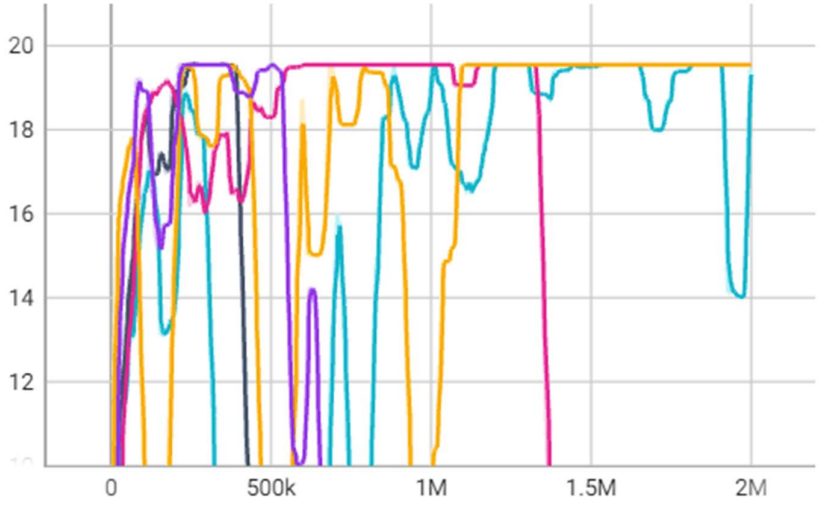
State space = $\{\boldsymbol{r_t}, \boldsymbol{h_t}\}$

- Perfect information concerning the trend.
- r_{t+1} cannot be predicted.
- The optimal decision of reward maximizer: always in if $h = 0$, always out if $h = 1$.

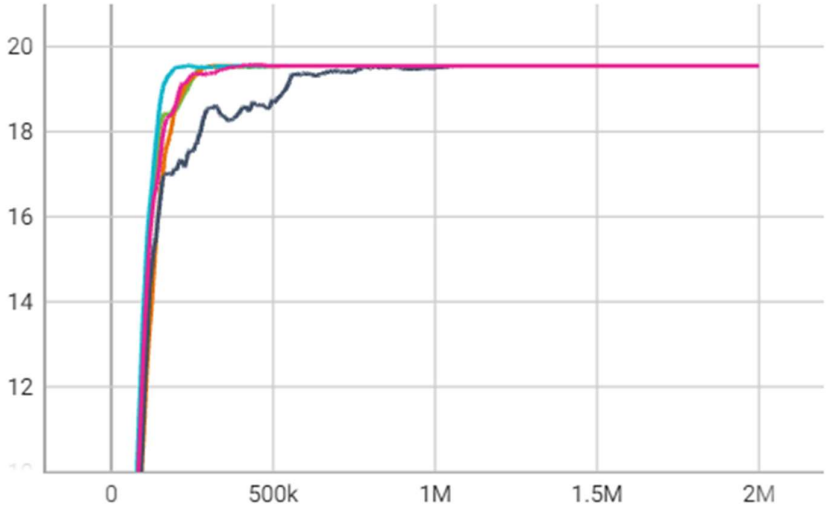
State space = $\{\boldsymbol{r_t}, \boldsymbol{r_{t-1}}, \boldsymbol{r_{t-2}}\}$

- The trend cannot be predicted (in $t-1$) nor will it be identified immediately (in t).
- However, the sequence of lagged returns may give some information on the trend.
- This information will depend on length of lags and σ.
- Expected behavior: more in than out if $h = 0$, more out than in if $h = 1$.

A2C (Continuous)

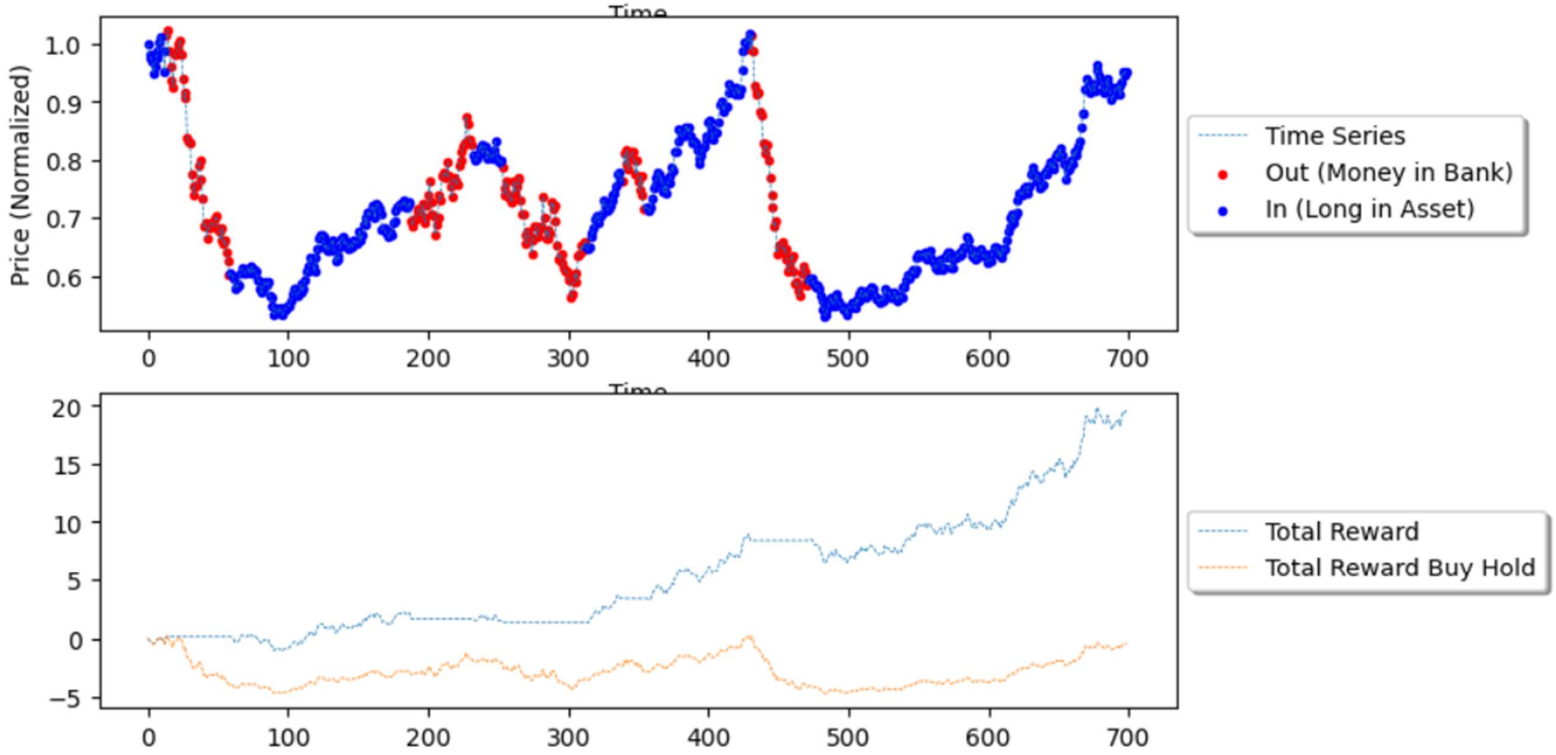


PPO (Continuous)

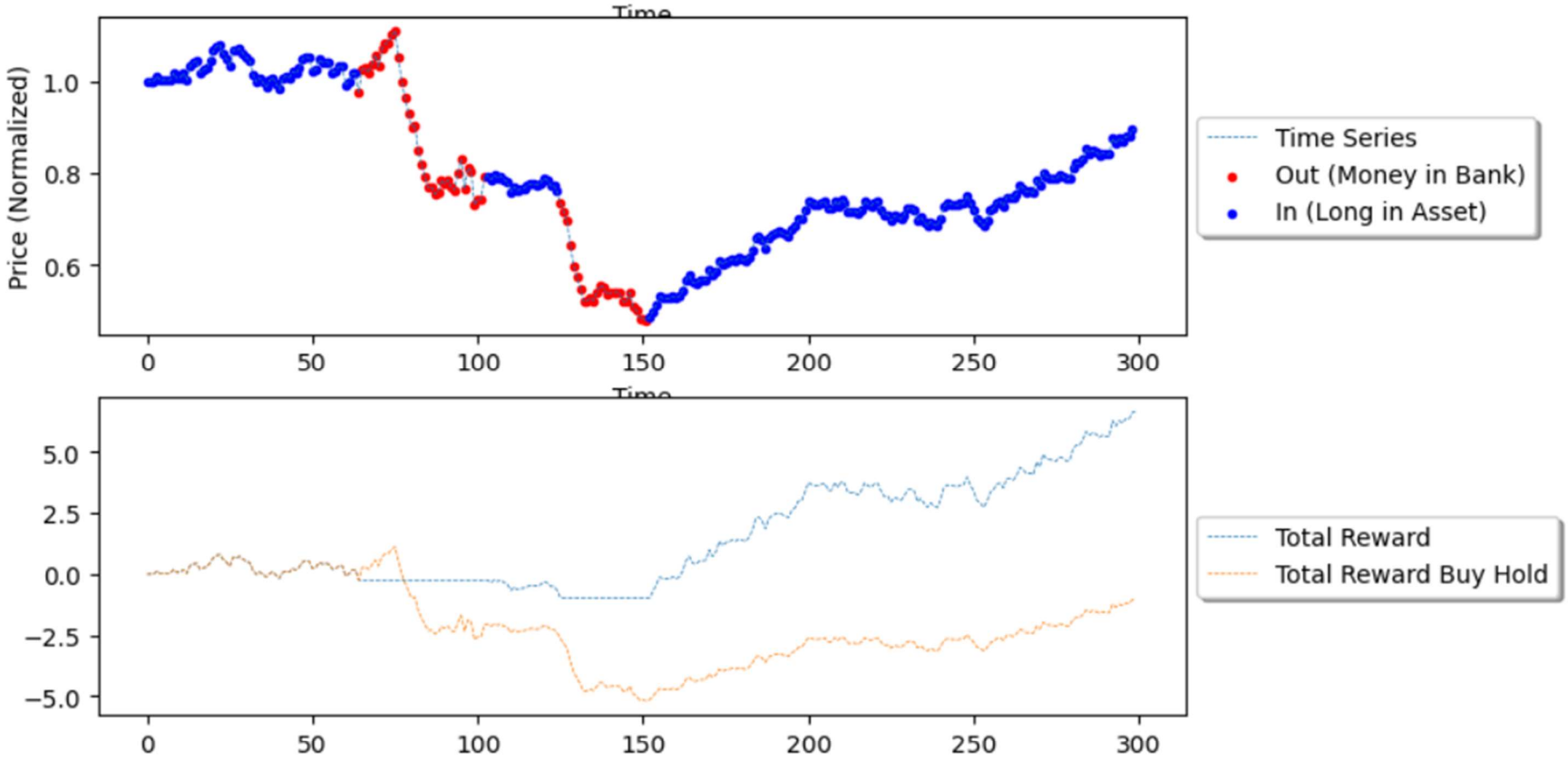


A2C (Continuous)	**PPO (Continuous)**
• 2,000,000 time steps • 5 runs (approximately 52 minutes per run) • Very unstable, but all five models performed well at some point. • All runs not satisfying. The best model's total return is 19.55.	• 2,000,000 time steps • 5 runs (approximately 52 minutes per run) • All runs converged to a total return of 19.55.

Best PPO on Training Set



Best PPO – Test Set



DDPG (Continuous)

- 250,000 time steps
- 5 runs (approximately 1,32 hours per run)
- All runs performed badly, no total return over 11.11

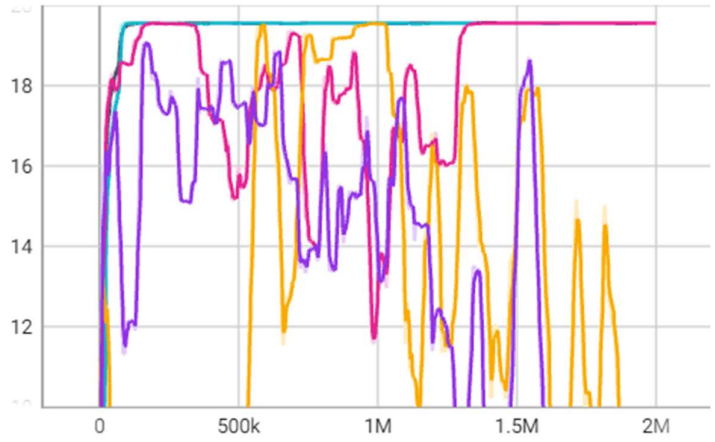
DDPG (Continuous)

- 250,000 time steps
- 5 runs (approximately 2,1 hours per run)
- All runs performed badly, all total returns below 9.

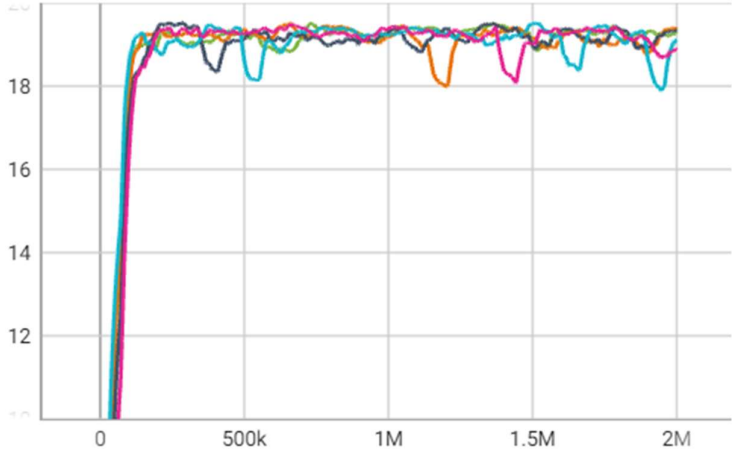
TD3 (Continuous)

- 250,000 time steps
- 5 runs (approximately 1.47 hours per run)
- All runs performed badly, all total returns below 11.

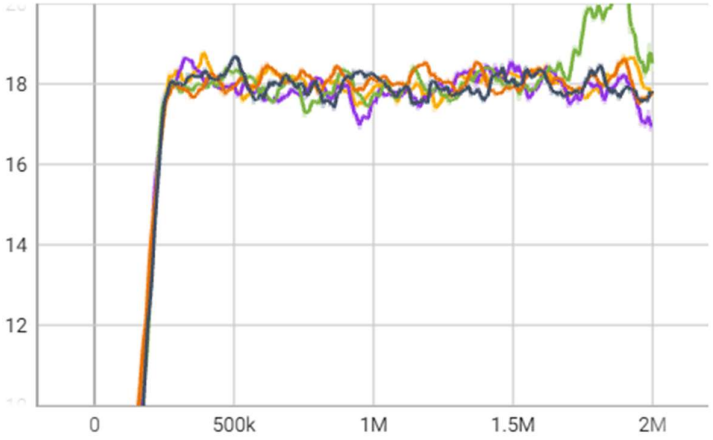
A2C (binary)



PPO (binary)

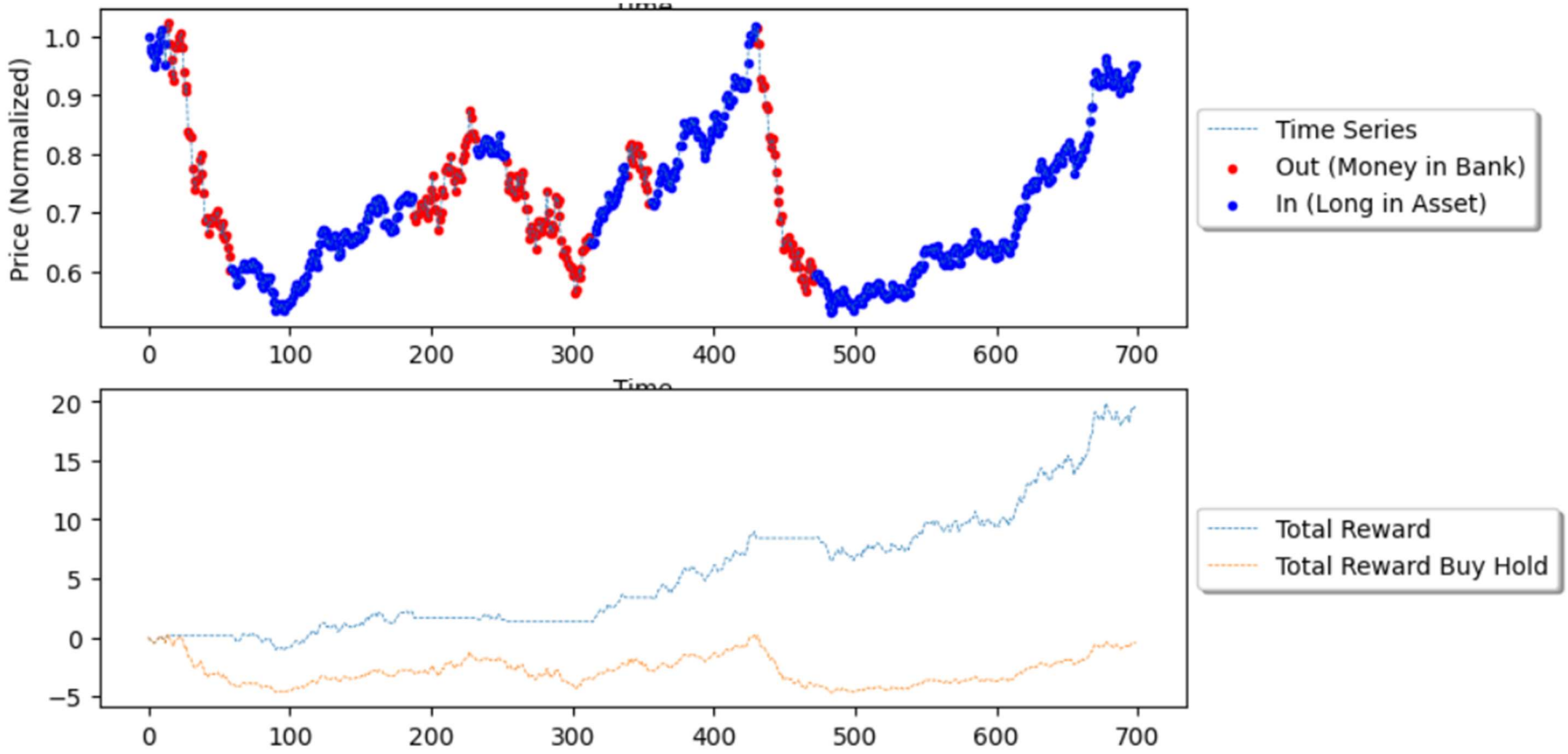


DQN (binary)

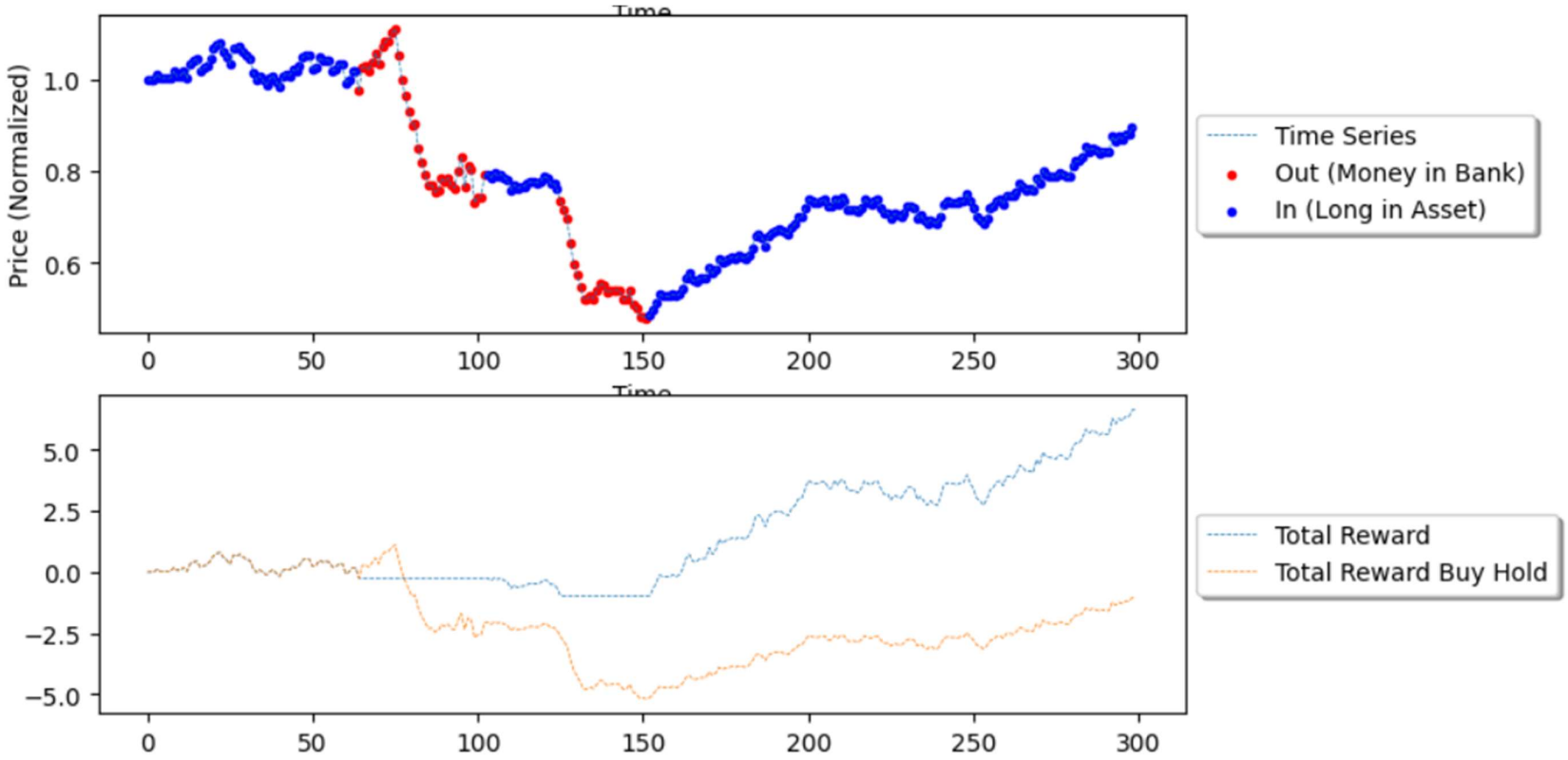


A2C (binary) • 2,000,000 time steps • 5 runs (ca. 53 minutes per run) • Unstable, four of five models reached 19:55 at some point)	**PPO (binary)** • 2,000,000 time steps • 5 runs (ca. 51 minutes per run) • All runs performed quite well, best model in the end 19.37 (at some point 19.54)
DQN (Discrete) • 2,000,000 time steps • 5 runs (ca. 34 minutes per run) • All runs performed quite well, best model at some point 20.53; at end point 18.3)	

Best PPO on Training Set



Best PPO – Test Set



Obs (all)	Obs (test)	State	Training	Test
0-13		**0**	**Nice**	
14-59		**1**	**Nice**	
60-188		**0**	**Nice**	
189-232		**1**	**Nice**	
233-253		**0**	**Nice**	
254-313		**1**	**Nice**	
314-338		**0**	**Nice**	
339-356		**1**	**Nice**	
357-430		**0**	**Nice**	
431-472		**1**	**Nice**	
473-764	*xx-64*	**0**	**Nice**	**Nice**
765-803	*65-103*	**1**		**Nice**
804-825	*104-125*	**0**		**Nice**
826-852	*126-152*	**1**		**Nice**
853-1000	*153-300*	**0**		**Nice**

4 Take-aways, To-Do's and Future research

Take-aways:

- Some RL methods in SB3 deliver "bad" performance even for simple deterministic time series.
- PPO seems to perform best for a variety of time series, and state spaces.
- The action space (discrete versus continuous) affects the solution even if the underlying problem and its solution should be the same for both action spaces.
- Problem is quickly overfitted: Problematic with respect to the features that can be added to the state space.

4 Summary

To Do:

- Train/Test with more data.
- Test on more test sets.
- Implement on-training cross validation (not really available by default in SB3)
- Implement early stopping, best model selection (A2C).
- Implement the risk-return trade-off