Report on PPO agent for Finance

October 21, 2025

Mathys VINATIER - GitHub Project page
Supervisor:
Pr Kim Tae-Wan

Mathys VINATIER

Contents

Chapter 21

Week 21

Contents

21.1 The PPO Agent Training

This week, we built and trained the model based on the previous recommendations. Several experiments were conducted to fine-tune and identify the optimal configuration of the PPO agent. Different parameter settings were tested across multiple trials, which will be discussed in this report.

As an initial step, we performed small adjustments to intentionally create an overfitting model. This helps verify that the agent can effectively learn from historical data before introducing regularization or exploration strategies. For this purpose, some parameters were fixed as follows:

- Discount factor $\gamma = 0.99$
- Learning rate $\alpha = 3 \times 10^{-4}$
- Generalized Advantage Estimation parameter $\lambda = 0.95$
- Policy clipping coefficient $\epsilon = 0.2$

These parameters play a key role in shaping the agent's learning dynamics. The discount factor γ determines how strongly future rewards influence current decisions—values close to 1 encourage long-term planning, while smaller values emphasize immediate rewards. The learning rate α controls the step size of parameter updates ; higher values speed up learning but risk instability, whereas lower values make learning slower and more stable. The GAE parameter λ balances bias and variance in advantage estimation : larger λ values produce smoother, more consistent updates but can increase variance. Finally, the clipping coefficient ϵ stabilizes policy updates by preventing the new policy from deviating too far from the old one, ensuring more reliable convergence during training.

21.1.1 First training analysis

For the initial training phase, we selected a set of baseline parameters to establish a reference point for the PPO agent's learning behavior. The chosen configuration is summarized below :

- Episodes : 300
- Epochs per update : 10
- Batch size : 128

The goal of this setup was to allow the agent to experience a relatively large number of episodes, providing diverse trajectories for training, while keeping the number of epochs moderate to prevent overfitting on limited data. This configuration encourages the critic network to refine its value estimation across a broader range of states, which in turn supports more stable policy updates by the actor.

The following section presents and discusses the main results obtained from this first round of training :

Figure 21.1: Training 1 - Training of trial 1

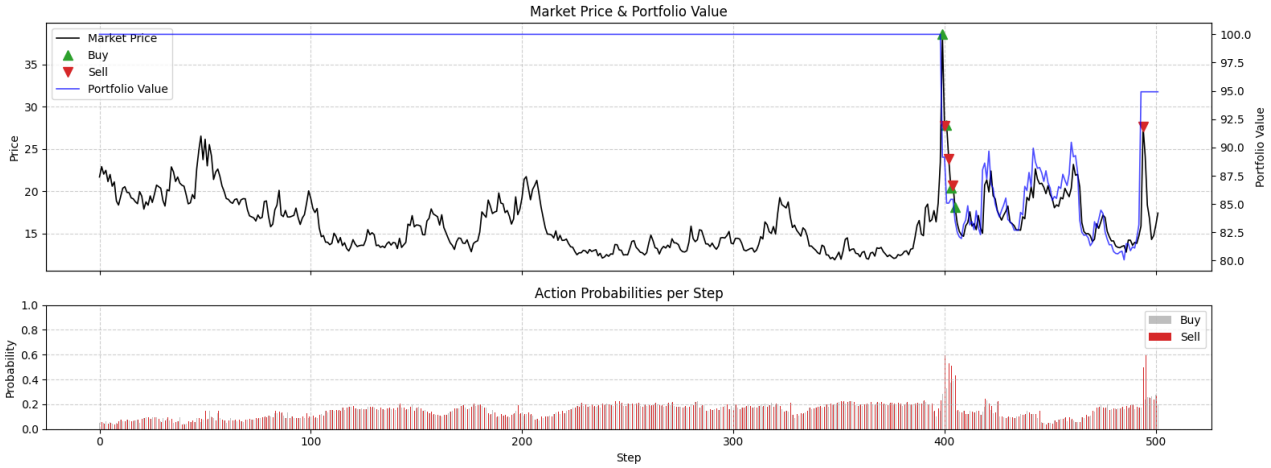

Figure 21.2: Training 1 - Test of trial 1

We can see that the training did not learn well since we have only few trades made, also the test has made very poor actions. We can consider that this first trial is really underfiting.

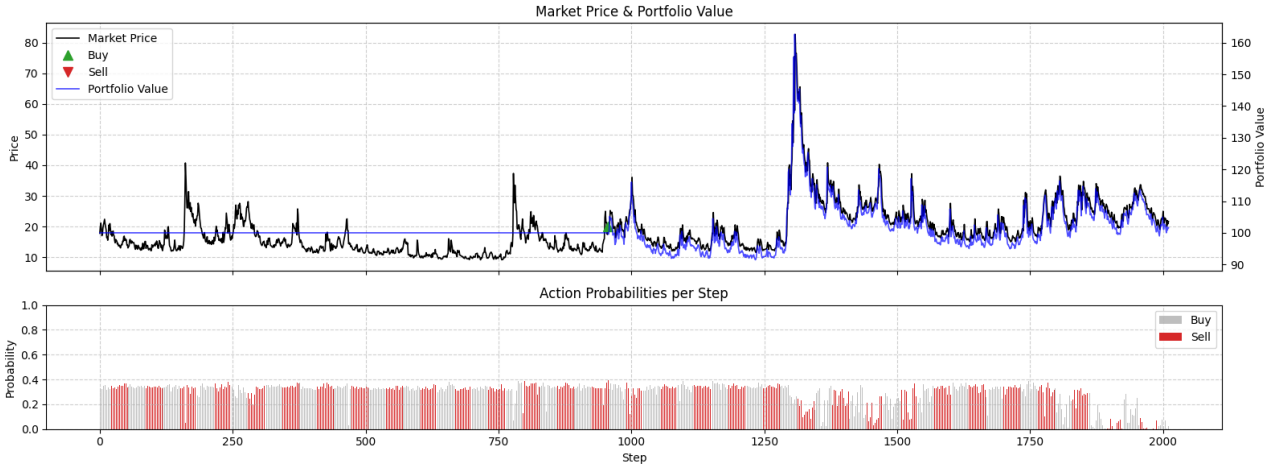

Figure 21.3: Training 1 - Training of trial 2

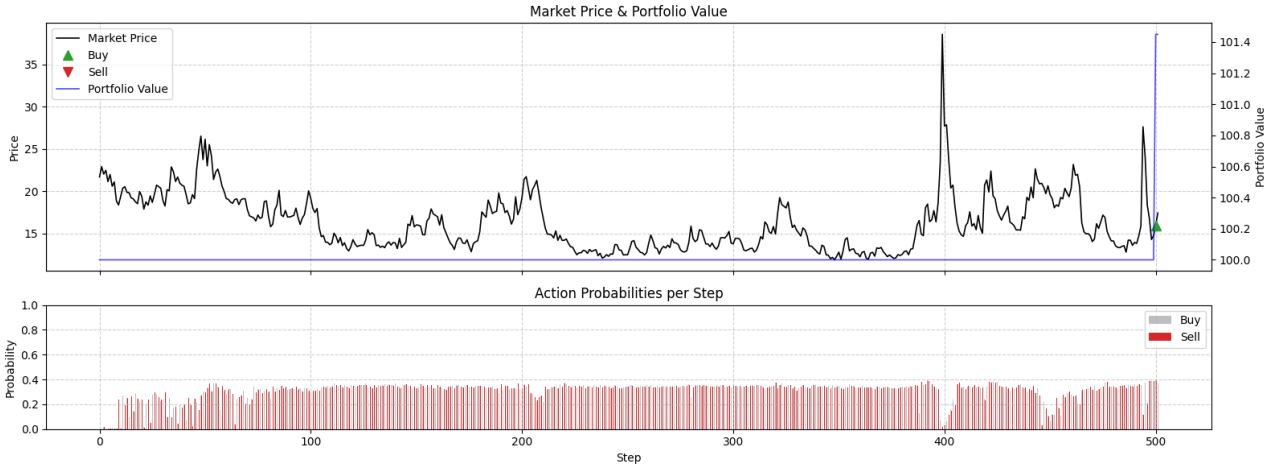

Figure 21.4: Training 1 - Test of trial 2

As previously, we can see that the model is clearly underfiting since it has made only one trade during testing and training which is representing an underfitting model.

Figure 21.5: Training 1 - Training of trial 3

Figure 21.6: Training 1 - Test of trial 3

The model did not take any trade neither on training or testing. We cannot evaluate this model.

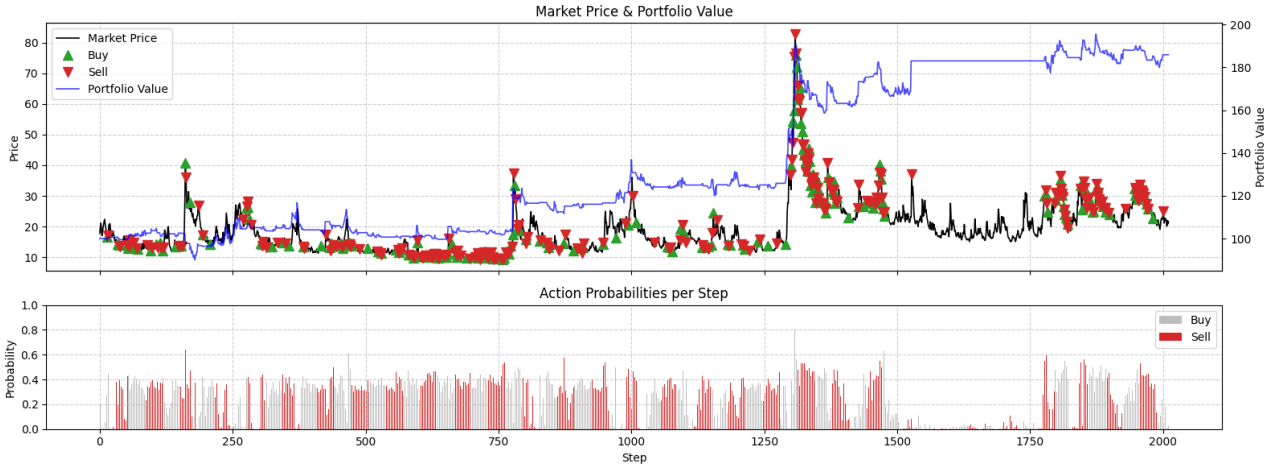

Figure 21.7: Training 1 - Training of trial 4

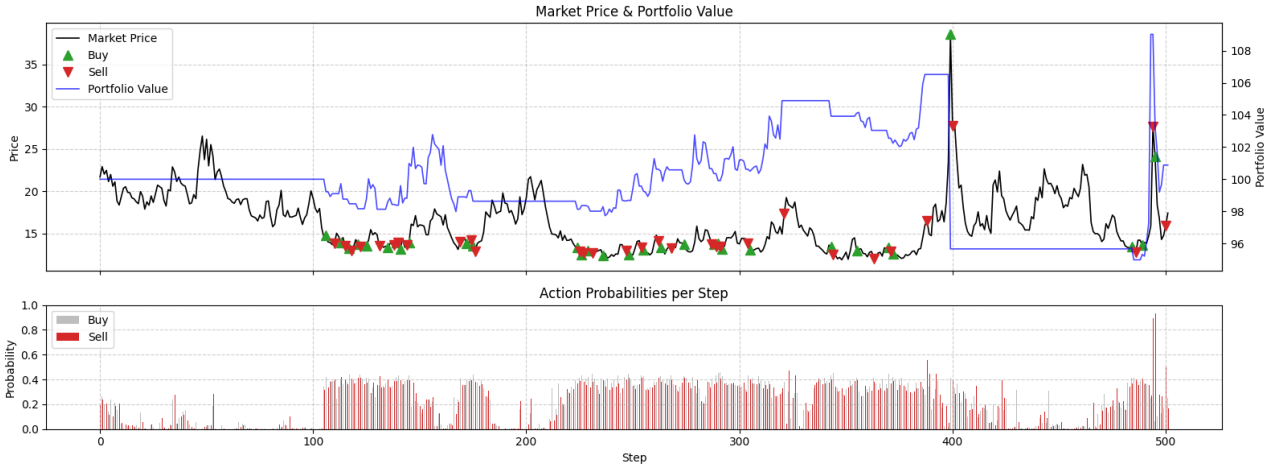

Figure 21.8: Training 1 - Test of trial 4

This model clearly learnt in training. However in testing the model is overfitting since it is not making wise trade.

Figure 21.9: Training 1 - Training of trial 5

Figure 21.10: Training 1 - Test of trial 5

Like the trial 3, we cannot analyse this model since it has made no trade.

Trial	Training Profit	Testing Profit	Annual Return	Annual Volatility	Sharpe Ratio	Max Drawdown
1	−822.5	74	0.17$/year	0.17	−0.07	−0.20
2	4.5	3	0.01$/year	0.01	0.71	0.00
3	0	0	0.00$/year	0.00	0	0.00
4	3930.5	0	0.13$/year	0.13	0.10	−0.11
5	0	0	0.00$/year	0.00	0	0.00

Figure 21.11: Summary of the trials performances

The results from these trials indicate clear signs of underfitting. Most configurations exhibited weak overall performance, and the single trial that achieved relatively good results during training failed to generalize well in the testing phase, suggesting that the model did not effectively capture the underlying dynamics of the environment.

21.1.2 Second training analysis

We would like to get a less overfitting model, for that we will work with more epoch and less episode to get a better trained thru the episodes, we will use the following training parameters :

- Episodes : 200
- Epochs per update : 50
- Batch size : 128

For this training, we are going to get a better look at each episode to better analyse the learning behavior of our model :

Figure 21.12: Training 2 - Training thru the epochs

As observed, the model demonstrates progressively wiser trading behavior across episodes, suggesting that it successfully captures and adapts to the underlying market dynamics during training. The learning curve reveals two distinct phases: initially, the agent learns to execute more strategic and selective trades, improving the efficiency of its decisions; subsequently, the model stabilizes the number of trades while gradually increasing overall profit.

Compared to the previous configuration, this version exhibits stronger learning performance, primarily due to the increased number of training epochs. However, the risk of overfitting remains and further evaluation on unseen data is necessary to assess the model's generalization capability and ensure robust performance in different market conditions :

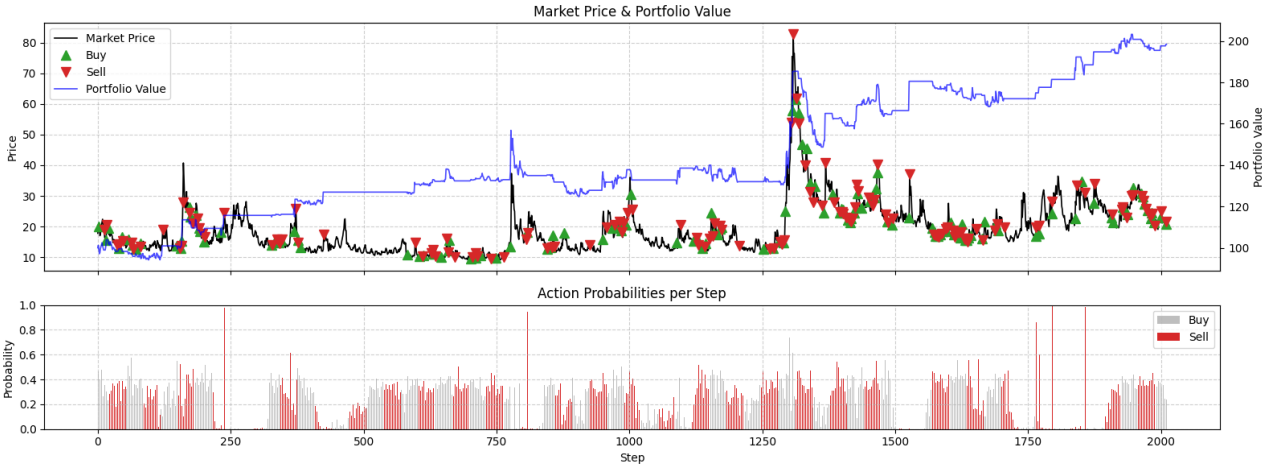

Figure 21.13: Training 2 - Training of trial 1

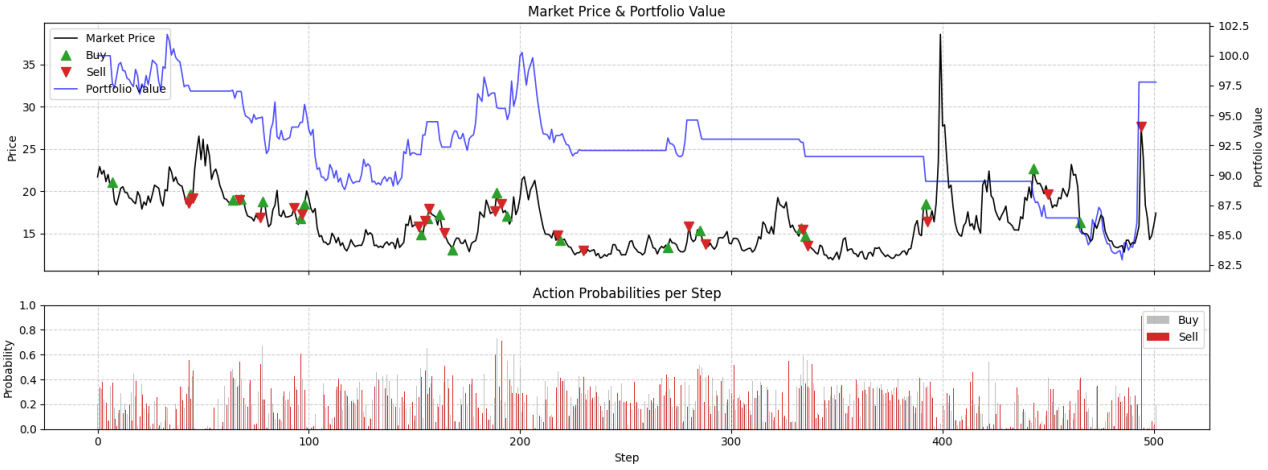

Figure 21.14: Training 2 - Test of trial 1

As we can see, the model has been overfitting. However since it is a stochastic model, we should compute many time our model to determine his usual behavior. Adding other features has input from other related market could be also a solution to our problem (VVIX / S&P500).