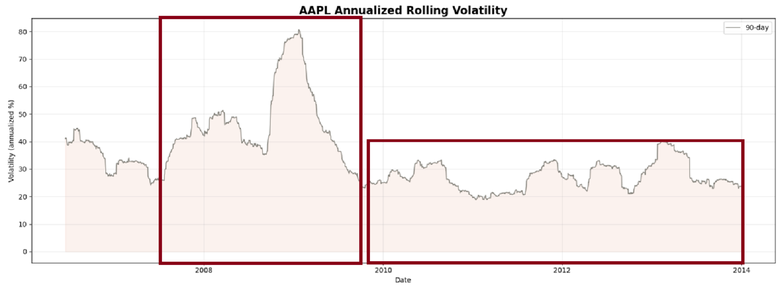
Reinforcement Learning for Regime-Dependent Optimal Stopping in Pairs Trading

author: **Mihail Bankov**, supervisor: **Fenghui Yu**, responsible professor: **Frans Oliehoek**

1. Background

Pairs trading is the action of trading two historically cointegrated stocks at the same time. If the stock prices diverge, and we expect them to converge, we can buy the cheaper one and sell the more expensive one, to make a profit on their spread (difference). The behaviour of stocks, also known as the market regime, needs to be considered when deciding how to trade. A visual example of regimes can be seen in the following graphic:



Optimal stopping deals with when to buy/sell. In our experiments, we trade with the spread of two correlated stocks: $spread = A - \beta \cdot B$. At any point in time, we can hold 1/0/-1 times the spread.

2. Motivation

- Optimal stopping, meaning knowing the right time to buy/sell, is very important in making a profit
- Current Theoretical solutions require known parameters [2]
- Current Learning-based solutions are either regime-agnostic or do not deal with optimal stopping [1] [3]

3. Research question

We want to examine whether and how being aware of the regime parameters improves the optimal stopping for pairs trading, when considering a Reinforcement Learning (RL) agent. This includes answering the following questions:

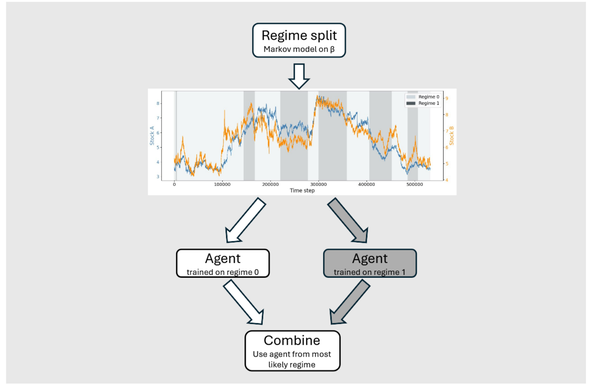
- SQ1: How can we find and incorporate regimes into a Reinforcement Learning model?
- SQ2: Does a regime-aware model perform better than a regime-agnostic model on stock data with regime switches?
- SQ3: Does a regime-aware model perform similarly to a regime-agnostic model on stock data without regime switches?

4. Methodology

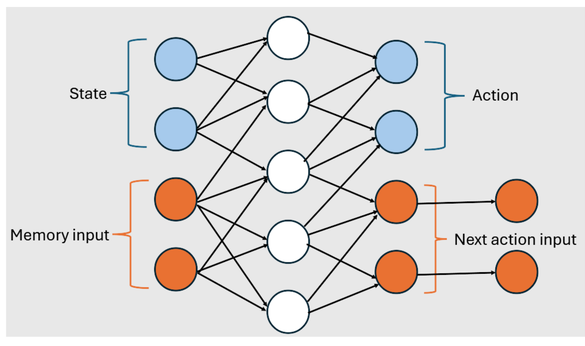
- Compare different algorithms on real data with multiple regimes, and on generated data with one regime
- The following pipeline is used for each agent and data type:

- Different models include a non-RL baseline algorithm, a baseline Double Deep-Q network (DQN), a Markov DQN, and a Recurrent DQN (DRQN). The Markov DQN is explicitly regime-aware, and the DRQN infers regimes
- At each point in time, algorithms get as input the market state, and output their position (1/0/-1)
- Multiple iterations are done for each algorithm, because of high variance

Markov DQN



Recurrent DQN



5. Results

- The algorithms are compared on Wheat/Corn futures data, as well as generated data similar to the stocks MPC/PSX
- Baseline DQN outperforms the regime-aware models on single-regime data
- The Markov DQN outperforms the DQN on more iterations (p=0.033); The used setup slightly differs from Experiment 1; susceptible to outliers.
- The results hold both for the profit, and risk-weighted profit (Sharpe ratio)

Experiment 1: 5 iterations each

Algorithm	Data	Returns, percentage	Sharpe Ratio
DRQN	Real	10.90 ± 12.43	0.63 ± 0.65
	Generated	488.47 ± 376.93	6.57 ± 2.77
Markov DQN	Real	8.71 ± 9.99	0.64 ± 0.68
	Generated	668.76 ± 238.84	6.47 ± 0.94
DQN	Real	8.03 ± 3.89	1.12 ± 1.01
	Generated	1775.85 ± 1388.81	8.78 ± 2.63
Bollinger bands	Real	40.77	1.69
	Generated	18.64	0.64
4% annual compound interest	Real	6.01	–
	Generated	11.80	–

Experiment 2: 20 iterations each

Agent	Data	Returns, percentage	Sharpe Ratio
DQN	Real	0.80 ± 9.28	0.13 ± 0.52
Markov DQN	Real	8.74 ± 16.11	0.41 ± 0.60

6. Discussion

- The regime-aware Markov DQN outperforms the baseline DQN on real data, confirming SQ 1 and 2
- Both the DRQN and Markov DQN do not outperform the baseline DQN on generated data; This does not necessarily confirm SQ3
- Still, we have to be cautious with the conclusions, because of the high result variance and existing outliers

References

[1] Álvaro Cartea, Sebastian Jaimungal, and Leandro Sánchez-Betancourt. Deep reinforcement learning for algorithmic trading, 2021.

[2] Emily Crawford Das, Phong Thanh Luu, Jingzhi Tie, and Qing Zhang. Pairs trading under a mean reversion model with regime switching, 2024.

[3] Andrea Macrì, Sebastian Jaimungal, and Fabrizio Lillo. Deep reinforcement learning for optimal trading with partial information, 2025.