Enhancing Pairs Trading with Reinforcement-Learning Constrained Portfolio Optimization

Cristian Petre-Luca [1]

[1]EEMCS, Delft University of Technology | Supervisors: Fenghui Yu, Frans Oliehoek | CSE3000 Research Project

Background & motivation

Pairs trading exploits two cointegrated stocks whose *spread* is mean-reverting — it wanders around a stable value even when each price drifts. With prices x_t, y_t and hedge ratio β,

$$s_t \;=\; x_t \;-\; \beta\, y_t$$

reverts to a long-term mean. Statistical arbitrage trades s_t directly: **short** it when unusually high, **long** it when unusually low. Holding the two legs in proportion $[\,1, -\beta\,]$ makes the book **market-neutral**, profit comes from relative mispricing, which allows **risk hedging**.

Reinforcement learning applications have focused either on adjusting entry/exit rules for pairs trading using the spread formulation, or on portfolio optimization for trend exploitation. We combine the two ideas, recasting pairs trading as a **continuous portfolio-optimisation** problem and train **PPO** agents to size the two legs plus a risk-free asset, rather than just timing fixed-size entries and exits. This lets us test *market neutrality* as a design choice: a **constrained** agent forced into the market-neutral hedge versus a **free** agent that weights the legs independently.

Research questions

1. **Constraint effect.** What performance and behavioural differences does the market-neutral constraint produce, for both cointegrated and non-cointegrated pairs?
2. **Beating the benchmark.** Can the constrained agent outperform a classical z-score entry/exit rule thanks to its position-*sizing* ability?
3. **Cost-aware rewards.** How do transaction-cost-informed rewards reshape the learned policies, returns, and exposure (risk)?

Data and training: pairs and simulations

We tested 11 well-known US equity pairs across industries for cointegration on daily closing prices. We ran the ADF test on the history of the asset, both on the in-sample window (2016–2019) reserved for model training, and the out-of-sample window (2020–2023) reserved for performance testing. **We found that cointegration breaks due to regime shifts**, and the nature of cointegration changes (β drifts...). MCD/YUM and V/MA passed both cointegration tests.

To replicate the mean reversion mechanism, model the **spread** as an AR(1) / Ornstein–Uhlenbeck process $s_t = \theta(1-\phi_t) + \phi_t s_{t-1} + e_t$; one leg is a random walk with drift and apply the cointegration formula to derive the other. Introduced volatility clusterings and markov-chain based **regimes of cointegration breaks** to qualitatively reproduce the market.

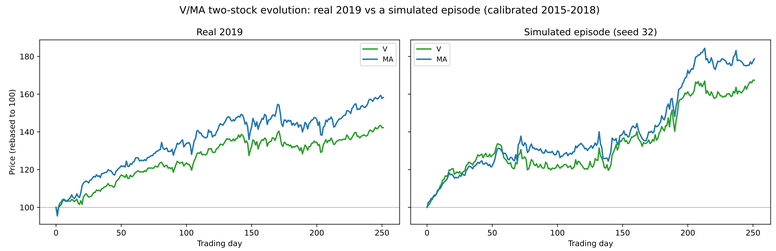

Figure 1. Real (2019) vs. a simulated V/MA episode (calibrated 2015–2018), each rebased to 100. The synthetic legs reproduce the pair's co-movement without sharing its specific realisation.

Methodology: a portfolio-sizing MDP

We frame trading as a Markov decision process and train a **PPO** agent (on-policy actor–critic) to maximise the discounted return $J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\big[\sum_t \gamma^t r_{t+1}\big]$. A single environment step (state, action branch and reward shown below):

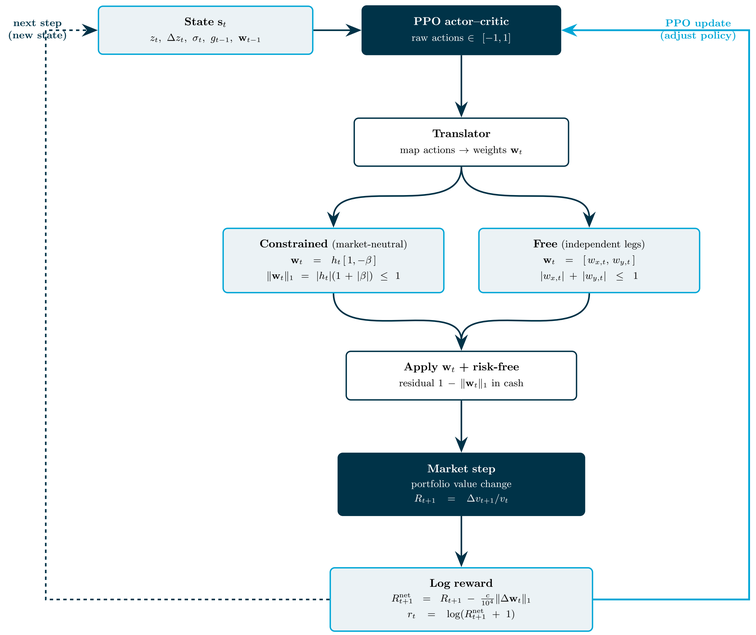

Figure 2. One environment step: state, action, reward, iteration.

Out-of-sample results (2020–2023)

Cointegrated pairs V/MA and MCD/YUM, against the classical z-rule, hold-both market, and a $+5\%$ risk-free leg.

pair	strategy	Sharpe	ann. ret	turn.	gross
V/MA	constrained	$+0.32\pm0.20$	$+7\%\pm1$	0.21 ± 0.06	0.80 ± 0.03
	free	$+0.23\pm0.21$	$+11\%\pm5$	0.15 ± 0.04	0.76 ± 0.12
	z-rule	$\mathbf{+0.65}$	$+8\%$	0.09	0.43
	market	$+0.19$	$+11\%$	0.00	1.00
	risk-free	0.00	$+5\%$	0.00	0.00
MCD/YUM	constrained	-0.15 ± 0.07	$+4\%\pm1$	0.14 ± 0.03	0.80 ± 0.02
	free	$+0.27\pm0.39$	$+9\%\pm5$	0.37 ± 0.19	0.61 ± 0.07
	z-rule	$\mathbf{+0.43}$	$+8\%$	0.05	0.29
	market	$+0.24$	$+11\%$	0.00	1.00
	risk-free	0.00	$+5\%$	0.00	0.00

(a) No transaction costs.

pair	strategy	Sharpe	ann. ret	turn.	gross
V/MA	constrained	-0.36 ± 0.08	$+3\%\pm0$	0.09 ± 0.01	0.83 ± 0.05
	free	$\mathbf{+0.36\pm0.13}$	$+10\%\pm3$	0.05 ± 0.03	0.29 ± 0.16
	z-rule	$+0.11$	$+5\%$	0.09	0.43
	market	$+0.19$	$+11\%$	0.00	1.00
	risk-free	0.00	$+5\%$	0.00	0.00
MCD/YUM	constrained	-0.39 ± 0.06	$+2\%\pm1$	0.07 ± 0.00	0.84 ± 0.01
	free	$+0.19\pm0.47$	$+6\%\pm3$	0.02 ± 0.02	0.13 ± 0.10
	z-rule	$\mathbf{+0.24}$	$+7\%$	0.05	0.29
	market	$+0.24$	$+11\%$	0.00	1.00
	risk-free	0.00	$+5\%$	0.00	0.00

(b) Net of 10 bps.

Out-of-sample results (continued)

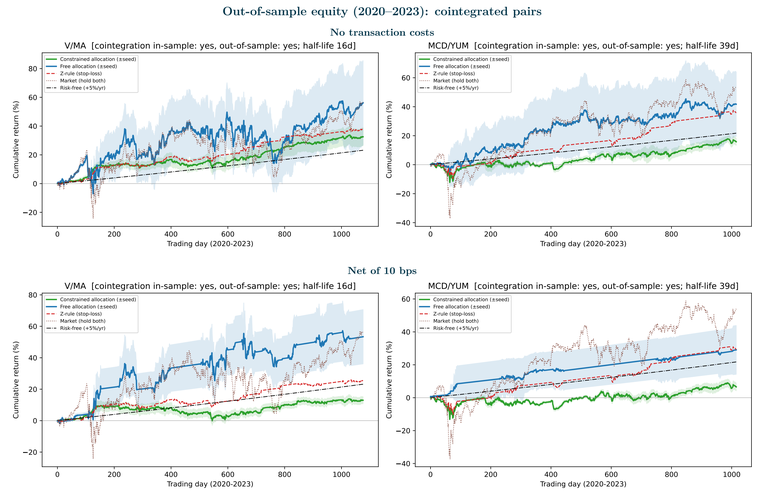

Figure 3. Out-of-sample equity (2020–2023) for V/MA and MCD/YUM, no costs vs. net of 10 bps.

Key findings

- **Constraint works, but does not win.** Obtains market-neutral returns yet over-trades when no arbitrage exists and never learns *when not to trade*, so it loses to the simple z-rule. Its modest returns don't beat the risk free rate on transaction costs, but it shows lesser variance on the returns.
- **Free agent: higher reward potential, more variance.** It can mix directional (β) exposure with some real arbitrage (α), earning higher but far noisier returns. The large action space needs good training to be exploited.
- **Costs reshape policies in opposite ways.** Under 10 bps the free agent sits out (gross $\to 0.13$–0.29) while the constrained agent doubles down on full exposure, with smoother policies and reduced turnovers. Cost-aware training is essential for economic trades.

Conclusion & future work

Position sizing alone does not beat a threshold rule: getting the spread direction right is not enough without knowing *when to stay out*, meaning the agents developped insuficient control. The natural fix and our main hypothesis for future work is to **split the problem**: let a signal decide *when* to enter, and let an RL agent decide only *how much* to size once in a trade. A more robust market simulator, or a real-data-only protocol, could potentially reduce the high seed variance seen here.