Regime-Switching Reinforcement Learning for Portfolio Allocation in Pairs Trading

Author: Tsvetelina Ilieva

Responsible Professor: Frans Oliehoek

Supervisor: Fenghui Yu

1. Introduction

- **Pairs trading** is a statistical arbitrage strategy that exploits the **spread** between two historically **cointegrated** stocks, betting that temporary divergences from their long-run equilibrium will revert.
- That equilibrium **isn't stable**. Under different market conditions the spread exhibits different levels of volatility and mean-reverting behavior.
- Classical pairs-trading and RL approaches largely ignore that the spread's dynamics shift across **market regimes**. Existing regime-aware RL targets broad asset-class allocation, not the **cointegration structure specific to pairs trading**.
- We propose a **regime-aware RL agent** that infers the latent market regime and centers its trading signal on the **per-regime** equilibrium rather than a stale rolling mean, making regime information directly actionable **without increasing model complexity**.

2. Research Questions

Does a regime-aware reinforcement learning agent outperform regime-neutral alternatives in portfolio allocation for pairs trading?

- Does a regime-aware Proximal Policy Optimization (PPO) agent outperform a regime-neutral PPO baseline trained on the same data?
- Does a regime-aware PPO agent outperform a classical z-score threshold strategy?
- Does the performance advantage of the regime-aware agent observed in simulation persist when both agents are trained and tested on real historical data?

3. Methodology

Pair Selection
S&P 500 financial sector pairs, Yahoo Finance daily adjusted prices, Pearson p > 0.70, Engle-Granger cointegration, half-life 1-60d

↓

OU Calibration
Fit Ornstein-Uhlenbeck process to spread $z = \log P^A - \beta \log P^B$, estimate global κ (reversion speed), μ (equilibrium), σ (volatility)

↓

HMM Regime Detection
3-state Hidden Markov Model: stable / neutral / crisis
Features: rolling z-score, log-volatility, logit reversion-rate

↓

Per-Regime Parameter Estimation
Estimate κ_r, μ_r, σ_r per regime + transition matrix A

↓

Training episodes setting

SIMULATED: Primary	REAL HISTORICAL: Secondary
Synthetic Episode Generation 3000 episodes, 252 steps, using Markov-chain regimes + per-regime OU Random permutation 80/20 train-eval split	**Sliding-Window Episodes** Re-estimate β over a trailing 252-day window per episode T=252, stride S=5, ≈ 650 episodes, HMM forward filter, 80/20 train-eval split

↓

PPO Agent Training
Observation: regime-adjusted z-score (regime) vs rolling z (baseline), weight, rolling volatility
Action: $w_t \in [-1, 1]$, Reward = $w_t \cdot \Delta z - c \cdot |\Delta w|$
2 agents per setting: regime-aware vs regime-neutral baseline

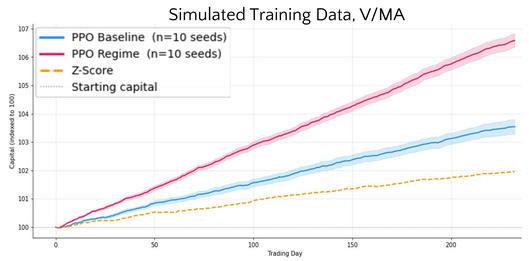
↓

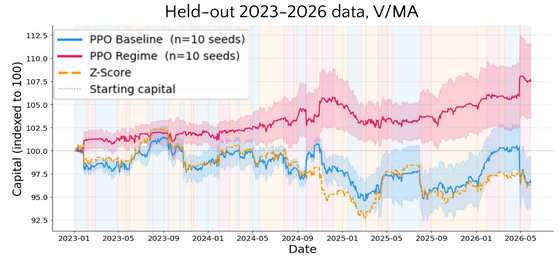
Simulated Test	**Real Out-of-Sample Test**
500 fresh episodes (seed 999) ±20% σ randomization	2023-2026, β from prior 12 months, single path

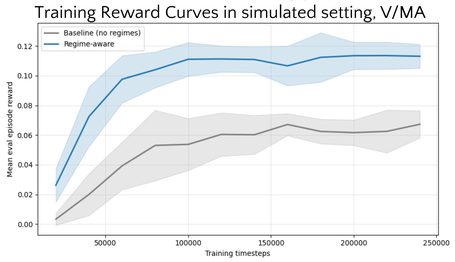
↓

Evaluation
Mean over 10 training seeds
Metrics: Sharpe, return, max drawdown
Regime-aware PPO vs regime-neutral PPO vs classical z-score threshold

4. Results







		Simulation		Real (2023-2026)	
Pair	Strategy	Return (%)	Sharpe	Return (%)	Sharpe
V/MA	PPO Regime	6.58 ± 0.26	1.354 ± 0.064	7.56 ± 1.34	0.567 ± 0.066
	PPO Baseline	3.54 ± 0.27	0.738 ± 0.040	-3.40 ± 0.91	-0.172 ± 0.051
	Z-Score	1.95	0.509	−3.48	−0.221
V/JKHY	PPO Regime	8.32 ± 0.80	1.183 ± 0.084	26.33 ± 6.33	0.609 ± 0.071
	PPO Baseline	3.41 ± 1.16	0.564 ± 0.085	15.72 ± 2.72	0.442 ± 0.066
	Z-Score	−0.37	−0.035	12.37	0.318

- The regime-aware agent **outperforms the baseline and the z-score strategy in both settings**, though with **more variable results** in the real-data setting. The regime-aware agent reaches **higher mean episode reward during training**.
- In **simulation** the regime-aware agent nearly **doubles the baseline's Sharpe** ratio on V/MA **(+0.616)** at a lower drawdown, just as large as on V/JKHY **(+0.619)**.
- The advantage **transfers to real 2023–2026 data**: on V/MA the regime agent earns a **positive Sharpe (0.567)** while both the regime-neutral baseline and the z-score **lose money** (−0.172 and −0.221). On V/JKHY all three strategies profit, yet the regime agent **still leads** (**0.609 vs 0.442**).

5. Discussion

- The **per-regime equilibrium** signal, rather than a stale rolling mean, lets the agent spot high-confidence entries it would otherwise miss, without added model complexity.
- The holds only under a **stable cointegrating relationship**. Both pairs sit in payments/fintech, largely insulated from the 2022–2026 shocks, whereas sectors hit by **structural breaks** would likely degrade the framework regardless of the agent.
- The **sim-to-real gap** can be narrowed by enriching the simulator with **structural breaks** and **volatility clustering**. Replace Gaussian HMM emissions with **Student-t** to make regime detection robust to fat tails.
- Further extension include adding a **cash allocation** to exit non-cointegrated periods, a short-selling **margin constraint** for realistic risk, and **implicit regime learning** that removes the explicit HMM.