Can Artificial Intelligence Trade the Stock Market?

Jedrzej Maskiewicz, Paweł Sakowski

Presentation plan

- Introduction
- Reinforcement Learning
- Research Methodology
- Results

What is algorithmic trading

- **Definition:** Automated execution of trades based on pre-defined rules or algorithms
- Advantages: Speed, precision, and elimination of emotional bias

Introduction

• Context:

- 70% of all equities traded in the US
- 90% of FOREX market
- Key developments:
 - Rise of Deep Reinforcement Learning (DRL) in dynamic decision-making, across fields
 - Potential to surpass classical supervised models
- Focus
 - Test DRL algorithms (e.g., DDQN, PPO) on multiple assets
 - Benchmark against traditional "Buy and Hold"

What is Reinforcement Learning (RL)

Aspect	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Objective	Learn from labelled data to map input to output	Discover hidden patterns in unlabelled data	Learn actions to maximize cumulative reward
Input Data	Labelled examples	Unlabelled examples	Interaction with an environment
Feedback	Direct feedback (correct/incorrect label)	No explicit feedback	Rewards (or penalties)
Examples	Image classification	Clustering	Self driving cars

Reinforcement Learning

Action Position in the market A_t



- Scenario: Teaching a robot to reach a finish line.
 - Goal: Reward of 100 for reaching the finish line.
 - **Problem:** Crediting only the last step ignores the importance of earlier decisions

• Why is This Wrong?

- Earlier actions also contribute to success.
- Rewards should be distributed across all steps leading to the finish line.
- Solution: Use Temporal Difference (TD) to credit actions proportionally.

Temporal Difference (TD) Learning $V(s_t) = r_t + \gamma * r_{t+1} + \gamma^2 * r_{t+2} + \gamma^3 * r_{t+3} \dots$

 $V(s_t) \leftarrow V(s_t) + \alpha \left[r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \right]$

$V(s_t)$ – Value Function

- Represents the expected total reward an agent will accumulate starting from a state s and following a specific policy π
- It evaluates how "good" a state is in terms of future rewards
- γ **Discount factor** determines the importance of future rewards

 α - Learning rate — controls how much the value function is updated with new information

Policy vs Value based approach

- Value-Based Approach
 - Learns to estimate the value function V(s) or Q(s, a)
 - Selects actions that have highest value
- Policy-Based Approach
 - Directly optimizes the policy $\pi(a|s)$, mapping states to actions
 - Learns the probabilities of actions without explicitly learning a value function

Double Deep Q Network

• Q-Learning

 $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma(max_{a'}Q(s_{t+1}, a') - Q(s_t, a_t)]$

- Deep Q-Network (DQN)
 - Replaces the Q-table with a Deep Neural Network to approximate $Q(s_t, a_t)$
 - With $Q(s_t, a_t; \theta)$

Double Deep Q Network

- DQN suffers from **overestimation bias** because the same network is used for both selecting and evaluating actions
- Action Selection: Uses the primary network to choose the best action: $a' = \arg \max_{a'} Q(s_{t+1}, a'; \theta)$
- Action Evaluation: Uses the target network to evaluate the selected action:

 $Q(s_{t+1},a';\theta^-)$

Double Deep Q Network

DQN formula:

 $Q(s_t, a_t; \theta) \leftarrow Q(s_t, a_t; \theta) + \alpha[r_{t+1} + \gamma(\max_{a'}Q(s_{t+1}, a'; \theta^-)) - Q(s_t, a_t; \theta)]$

DDQN formula: $Q(s_t, a_t; \theta) \leftarrow Q(s_t, a_t; \theta) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, \arg \max_{a'} Q(s_{t+1}, a'; \theta); \theta^-) - Q(s_t, a_t; \theta)]$

 $Q(s_t, a_t; \theta)$ – Value function; estimation of Q-value at state s_t with action a_t calculated by Neural Network θ γ – discount factor

 r_{t+1} - reward after action a_t

 θ^- - target network

DDQN

• Loss function:

 $L(\theta) = E[(r_{t+1} + \gamma \ max_{a'} Q(s', a'; \theta^{-}) - Q(s_t, a_t; \theta))^2]$

Actor – Critic

• Actor:

- Directly optimizes the policy $\pi(a|s; \theta)$, which maps states to actions
- Determines the best action to take in a given state

• Critic:

- Evaluates the Actor's actions by estimating the value function $V(s;\omega)$
- Provides feedback using Temporal Difference (TD) error or Advantage A(s,a)
- A(s, a) = Q(s, a) V(s)

• Workflow:

- Actor proposes an action *a*
- Critic evaluates a by estimating how good it is based on the current policy π
- Actor updates its policy using the feedback from the Critics



Source: Sutton, R. S., & Barto, A. G., Actor-Critic method". Reinforcement Learning: An Introduction. (2018)

Proximal Policy Optimization (PPO)

- Actor Critic family
- Very stable:
 - Ensures smooth policy updates by restricting large changes using the clipping mechanism
 - Reduces the risk of instability during training
- Clipping:

$$L^{CLIP}(\theta) = E[\min(R_t(\theta), CLIP(R_t(\theta), 1 - \varepsilon, 1 + \varepsilon))]$$

 $R_t(\theta) = \frac{\pi(a|s_t;\theta)}{\pi_{OLD}(a|s_t;\theta)}$ probability ratio of the action under current policy π to previous policy π_{OLD}

Methodology Overview

- Assets Analyzed:
 - EUR/USD and S&P 500 Index
- Daily data
- Features
- Reinvest profits after every trade
- Walk forward optimization

Walk forward optimization

Iteration\Years	2005-2016	2017	2018	2019	2020	2021	2022	2023	1
First optimisation	train validatio			test					
Second optimisation		train		validation	test				
Third optimisation		t	rain		validation	test			
Fourth optimisation			train	validation	test				
Fifth final optimisation			t		validation	test			

Evaluation Metrics

- Sharpe Ratio = $\frac{R}{\sigma} * \sqrt{N}$
 - R return of the strategy
 - σ standard deviation of strategy
 - N annualization factor (number of intervals in the year)
- Benchmarks:
 - Buy and Hold
 - "Perfect annualized" Agent

Results EUR/USD

EURUSD	Final balance	CAGR	Sharpe ratio	Sortino Ratio	Maximum Drawdown	Win Rate	In Long	In Short	Out of the market
DDQN_NN	14 444.74	7.63%	1.842	2.043	-4.40%	58.6%	17.7%	26.2%	56%
DDQN_T	12 512.91	4.58%	1.035	0.853	-11.30%	55.4%	2.5%	30.7%	66.7%
PPO_NN	10 360.83	0.71%	0.108	0.136	-14.70%	50.3%	49.5%	30.1%	20.3%
PPO_T	13 271.26	5.82%	0.847	1.197	-11.60%	48.5%	40.4%	45.6%	13.9%
				Benchma	arks:				
Buy and hold	9 641.81	-0.84%	-0.108	-0.159	-22.50%	100%	100%	0%	0%
"perfect" annual strategy	13 150.82	6.61%	0.935	1.396	-9.20%	100%	40%	60%	0%

Results EUR/USD

DDQN_Agent_NN_1D_EURUSD DDQN_Agent_T_1D_EURUSD PPO_Agent_T_1D_EURUSD PPO_Agent_NN_1D_EURUSD 14000 ---- Perfect_Agent — Buy_and_hold 13000 12000 Balances 11000 10000 9000 2019 2021 2023 2024 2020 2022 Date

Balances of All Agents and Benchmarks Over Time for EURUSD

Results SP500

S&P 500	Final balance	CAGR	Sharpe ratio	Sortino Ratio	Maximum Drawdown	Win Rate	In Long	In Short	Out of the market
DDQN_NN	20 009.45	22.43%	0.993	1.225	-15%	51.3%	51.7%	33.5%	14.7%
DDQN_T	30 565.62	38.41%	1.6	2.218	-25.4%	51.7%	62.6%	22%	15.4%
PPO_NN	11 567.96	4.3%	0.183	0.193	-41.6%	54.7%	59.8%	13.4%	26.7%
PPO_T	40 010.83	49.76%	2.158	2.651	-26.1%	58.6%	73.9%	16.9%	9.1%
				Benchma	arks:				
Buy and hold	19 482.77	21.37%	0.834	1.014	-33.9%	100%	100%	0%	0%
"perfect" annual strategy	28 982.19	36.23%	1.505	1.791	-33.9%	100%	80%	20%	0%

Results SP500



Balances of All Agents and Benchmarks Over Time for SPXUSD

Results SP500 – test sample from 1/7/2024

Strategy Tester Visualization : test	t_us500 on U5500.pro,Daily from 202	4.07.01 to 2024.11.29								- 0 X
File View Charts Tester	Help									
11 00 M Q Q U	🙂 谢 📲	B) Skip to 2024.11.29 0	10:00							
Symbol Bid Ask	100,049y 10,124 10,040 10,192 10,240									
- US500.pro 5523.0 5523.7 547.0	\$40 ± \$10.4									
										5012
										5662
										500.2
										5512
										500.2
										540
										011
										6041
		_ الا ي ي الله								100.17
										40112
Symbols Ticks 2Apr 2	924 4 Apr 2524 8 Apr 2524 10 Apr	1024 12 Apr 2024 15 Apr 2024 18 Apr 20	104 22 Apr 2004 24 Apr 2004 25 Apr 2004	30 Apr 2024 2 May 2024 4 May 2024	1 May 2024 10 May 2024 14 May 2024 16 May 2	124 20 May 2024 22 May 2024 24 May 2024	28 May 2024 Ja May 2024 Jan 2024 53	An 2024 5 An 2024 13 An 2024 13 An 2024 17 An 2024	1936 824 236 824 23 36 824 23 36 824 1 36 284	
* Time	Ticket	Symbol	Action	Туре	Volume	Price	S/L	T/P		Comment
2024.07.01 0000100	2	U\$500.pro	market	BUY	0.03	5535.4				
Trade History Operations	Journal									
For Help, press F1	1									
4 62450570 - OANDATMS-MT5: De	emo Account - Hedge - OANDA TMS	Brokers S.A [US500.pro,Daily]								- 0 ×
00 00 File View Insert Ch	narts Tools Window Help					a sector and the				X - 0 ×
Q+++74	≓ T 88 • M1 M5 M	15 M30 H1 H4 D1 W1 MN	A 🗠 • 😅 • 🐻 📧 🚨	🕫 🙅 🍓 🗖 Algo Trading	New Order 1 ³ 00 A ⁿ 0	< ⊞ !!!! !@				۵ 🖤 🐒 🗖
						ran anton terretariation terr				10096
										10081
										10067
										10052
										10038
										10023
										1000
										1979
										9965
										9901
										0.0%
atter .										D.OTH
2024.07.00										0.0%
2024-07-06 Overview Settings Inputs	Graph Agents Journal									00:00:01 / 00:01:51

Methodological Challenges in Applying DRL to Stock Market Trading

- In financial markets, the agent's actions (e.g., buy, sell, out of the market) do not influence the broader market state
- TD algorithms attribute a discounted premium of future rewards to present actions
 - Example:

In a sequence like *long*, *long*, *short*, *short*, future rewards from short positions are incorrectly attributed to earlier long positions

Recomendations

Sutton, R. S., & Barto, A. G. Reinforcement Learning: An Introduction (2018)

Thank you for your attention