Deep Q-learning for Limit Order Book Trading in Foreign Exchange Markets

Floris Dobber Adam Keys

Department of Computer Science
University College London

13 September 2024

Introduction

- Application of reinforcement learning (RL) to limit order book (LOB) trading in foreign exchange (FX) markets
- Reinforcement learning
  - Deep Q-Network (DQN) algorithm with double Q-learning, dueling networks, distributional rewards and multi-step returns
  - Ape-X framework for distributed training
- Limit order book simulation
  - Agent based model (ABM) simulator (ABIDES)
  - Simulate entire 10 levels of the LOB based on historical data
- Foreign exchange markets
  - Largest financial market at $7.5 trillion daily volume in 2022 [1]
  - Focus on EUR/USD as most liquid and traded currency pair 22.7% of total volume [1]

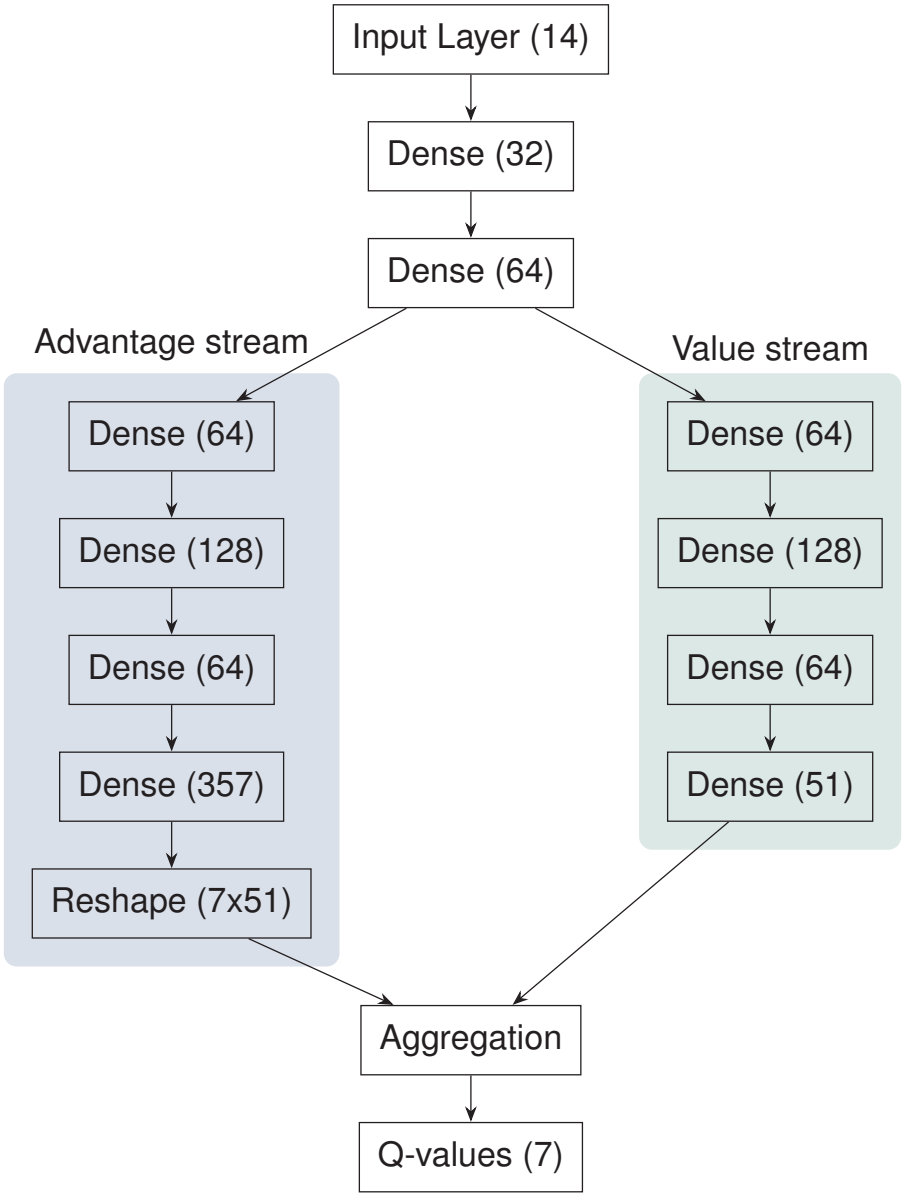
Objectives

- Develop a realistic simulation of the FX market using the ABIDES-Gym framework, focusing on replicating key stylized facts of the EUR/USD market.
- Implement and train deep Q-learning agents to trade in the simulated environment.
- Evaluate the agents' performance across various market regimes, comparing it to both simple and sophisticated baseline strategies.
- Analyze the agents' learned behaviors and trading patterns under different market conditions.

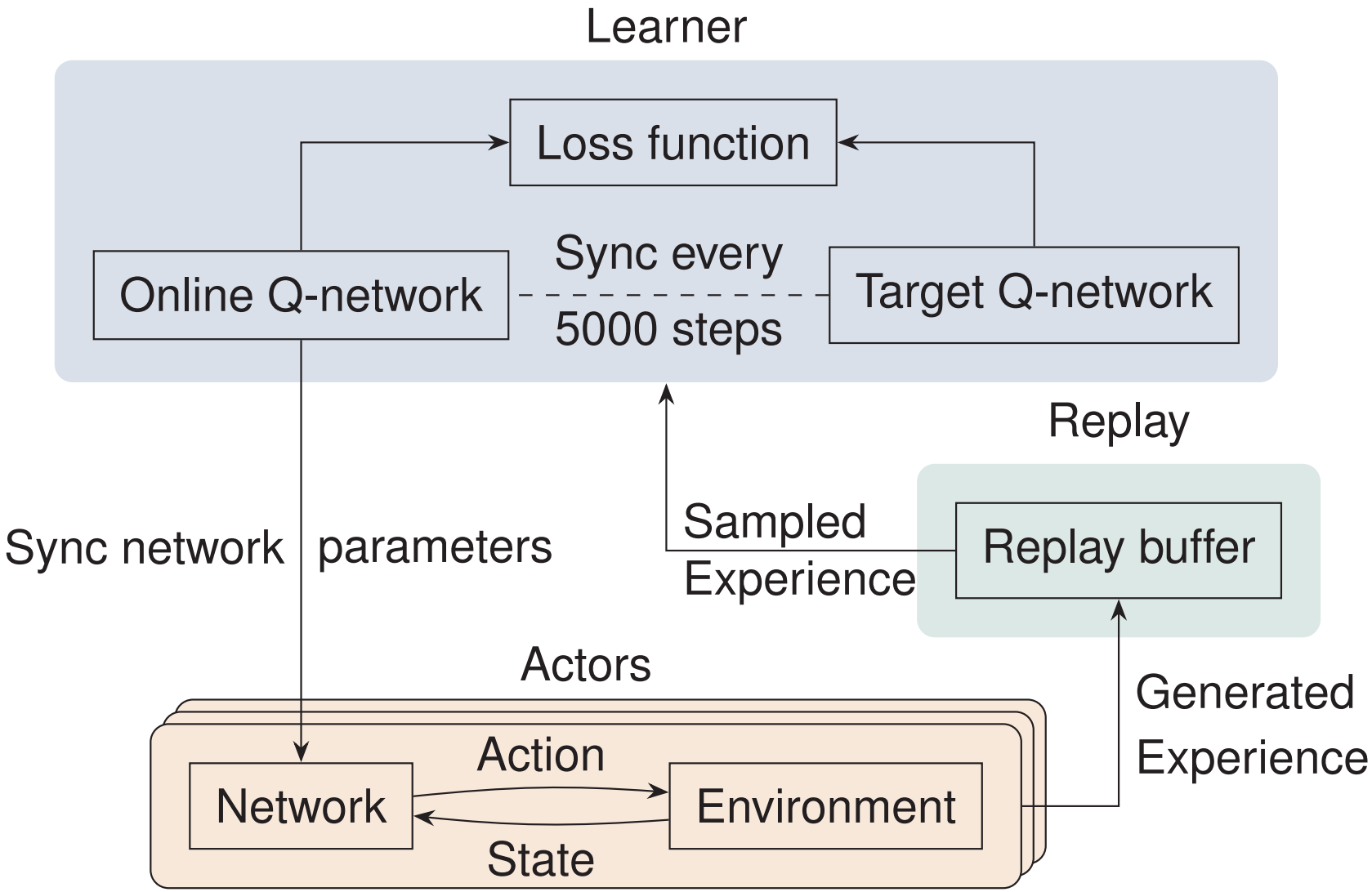
Background: Deep Q-learning

- Combines Q-learning with deep neural networks
- Enhancements
  - Experience replay: Breaks correlations between experiences [2]
  - Double Q-learning: Improved value estimation accuracy [3]
  - Dueling networks: Better generalization across actions [4]
  - Distributional Q-learning: Richer representation of uncertainty [5]
  - Multi-step learning: Faster reward propagation and reduced bias [6]
- Ape-X architecture [7]
  - Distributed reinforcement learning framework
  - Multiple actors generate experiences, central learner updates Q-values
  - Improves learning efficiency and stability

Background: DQN Network Architecture



Background: Ape-X Architecture



Background: Limit Order Book Simulation

- Agent-Based Models (ABMs)
  - Bottom-up approach to market simulation
  - Captures complex market dynamics from agent interactions
- ABIDES Framework [8]
  - Agent-Based Interactive Discrete Event Simulation
  - Simulates high-fidelity market environments
  - Includes Gym interface for RL integration
- Key Features
  - Reproduces heavy-tailed return distributions
  - Captures volatility clustering
  - Models market impact effectively
- Advantages for Research
  - Controllable and reproducible experiments
  - Ability to test strategies without real capital risk
  - Facilitates study of market microstructure

Methodology: Environment

- Episode Structure
  - 1-hour duration, 0.2-second time steps (18,000 steps/episode)
- State Space
  - 14 dimensions including time, cash, inventory, market data
  - Directional signal as probability distribution over price movements
- Action Space
  - 7 discrete actions: buy/sell at 3 price levels + skip
  - Maximum inventory constraint of 10 units
- Reward Function
  - Combination of directional reward and profit-and-loss
  - Curriculum learning approach with shifting weights
- Background Agents
  - Market makers, value agents, momentum agents, noise agents
  - Creates realistic market dynamics and liquidity

Methodology: Synthetic Alpha Signal

$$\hat{p}_t = w\hat{p}_{t-1} + (1-w)d_t, \quad w \in (0,1)$$
$$d_t \sim \text{Dir}(\alpha(\bar{r}_{t+h}))$$
$$\bar{r}_{t+h} = \frac{\bar{x}_{t+h} - x_t}{x_t}, \quad \text{where} \quad \bar{x}_{t+h} = \frac{1}{h}\sum_{i=1}^{h} x_{t+i}$$

Signal parameterized such that probability mass is concentrated on dimension indicating future price move:

$$\alpha(\bar{r}_{t+h}) = \begin{cases} (a^H, a^L, a^L), & \text{if } \bar{r}_{t+h} < -k \\ (a^L, a^H, a^L), & \text{if } |\bar{r}_{t+h}| \leqslant k \\ (a^L, a^L, a^H), & \text{if } \bar{r}_{t+h} > k \end{cases}$$

$$a^H, a^L \in \mathbb{R}^+ \quad \text{where} \quad a^H > a^L$$

Methodology: Reward Function

Portfolio Value	$PV_t = C_t + I_t x_t$
P&L Reward	$r_{t+1}^{P\&L} = PV_t - PV_{t-1}$
Directional Reward	$r_{t+1}^{dir} = \kappa[-1, 0, 1] \cdot \hat{p}_t I_t$
Total Reward	$R_{t+1} = \phi r_{t+1}^{\text{dir}} + (1 - \phi) r_{t+1}^{\text{P\&L}}$

C_t denotes cash, I_t inventory holdings, x_t is the midprice, $\kappa \in \mathbb{R}^+$, and $\hat{p}_t$ is the alpha signal.

After each step, the directional weight ($\phi_0 = 1$) reduces by some scalar factor:

$$\phi \leftarrow \gamma\phi, \qquad \gamma \in (0, 1)$$

Methodology: Agent

- Ape-X Distributed Architecture
  - 4 actors, each running 4 parallel environments
- Training Parameters
  - Total training steps: 3,000,000
  - Buffer size: 200,000 experiences
  - Learning rate: 5e-6 (static)
  - Discount factor (gamma): 0.99
  - Target network update frequency: 5000 steps
  - Multi-step learning with n=10

Methodology: Baseline Agents

- Random Agent
  - Selects actions uniformly at random
  - Establishes lower performance bound
- Buy-and-Hold Agent
  - Buys fixed amount at start, holds until end
  - Provides simple profit baseline
- Aggressive Agent
  - Crosses spread to place limit orders
  - Uses directional signal to inform trading decisions
- Common Constraints
  - Maximum inventory size of 10 units

Motivations

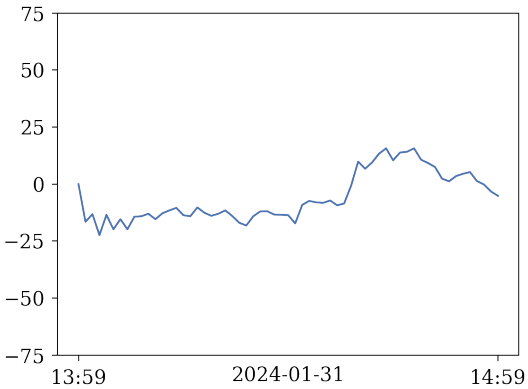
Research Questions

- Can a deep Q-learning agent learn effective trading strategies in a simulated FX market environment?
- To what extent can the RL agents adapt their strategy to various market conditions, such as high volatility, trends, and price jumps?

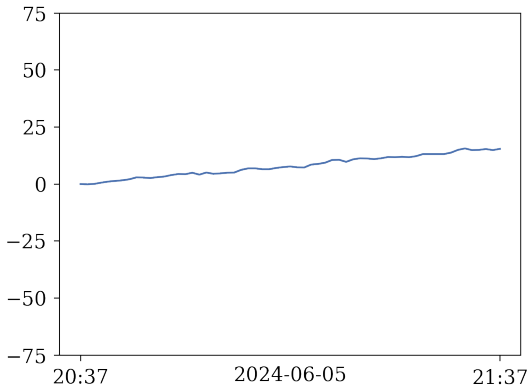
Experimental Setup

- 8 distinct market regimes identified
  - High volatility, Low volatility
  - Upward trend, Downward trend, No trend
  - Upward jump, Downward jump
  - Mixed (combination of all regimes)
- Training Process
  - One agent trained per regime
  - 35 one-hour episodes per regime for training
- Evaluation
  - 7 one-hour episodes per regime for testing
  - Each agent tested across all 8 regimes
- Performance Metrics
  - Returns and excess returns
  - Sharpe ratio
  - Maximum drawdown
  - Action frequency and trading style analysis

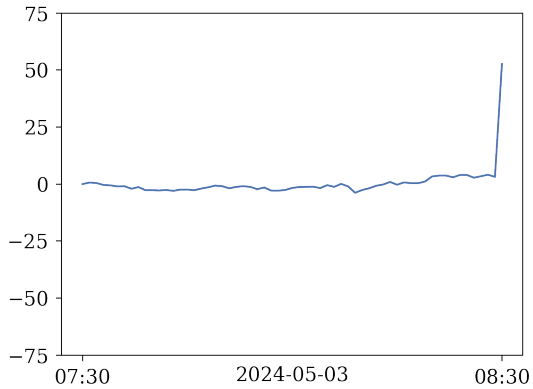
Results: Selected Training Periods



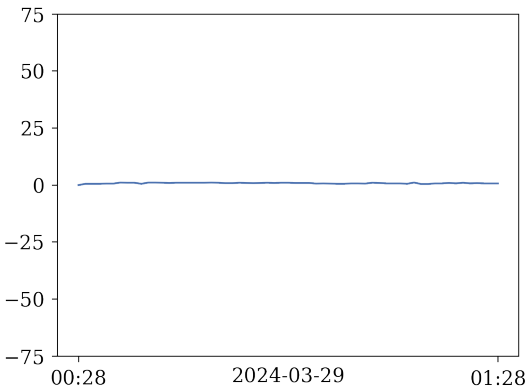
(a) High Volatility



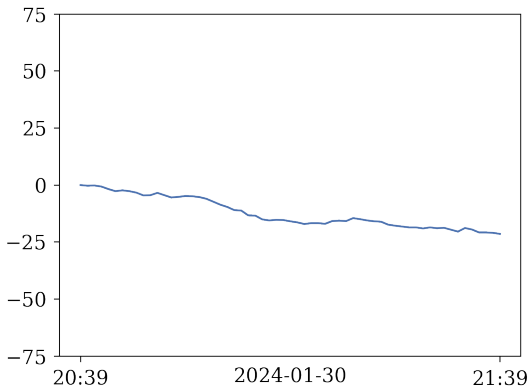
(b) Upward Trend



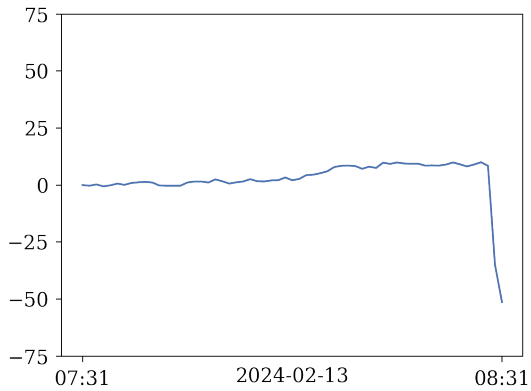
(c) Upward Jump



(d) Low Volatility
No Trend



(e) Downward Trend



(f) Downward Jump

Results: Environment Analysis

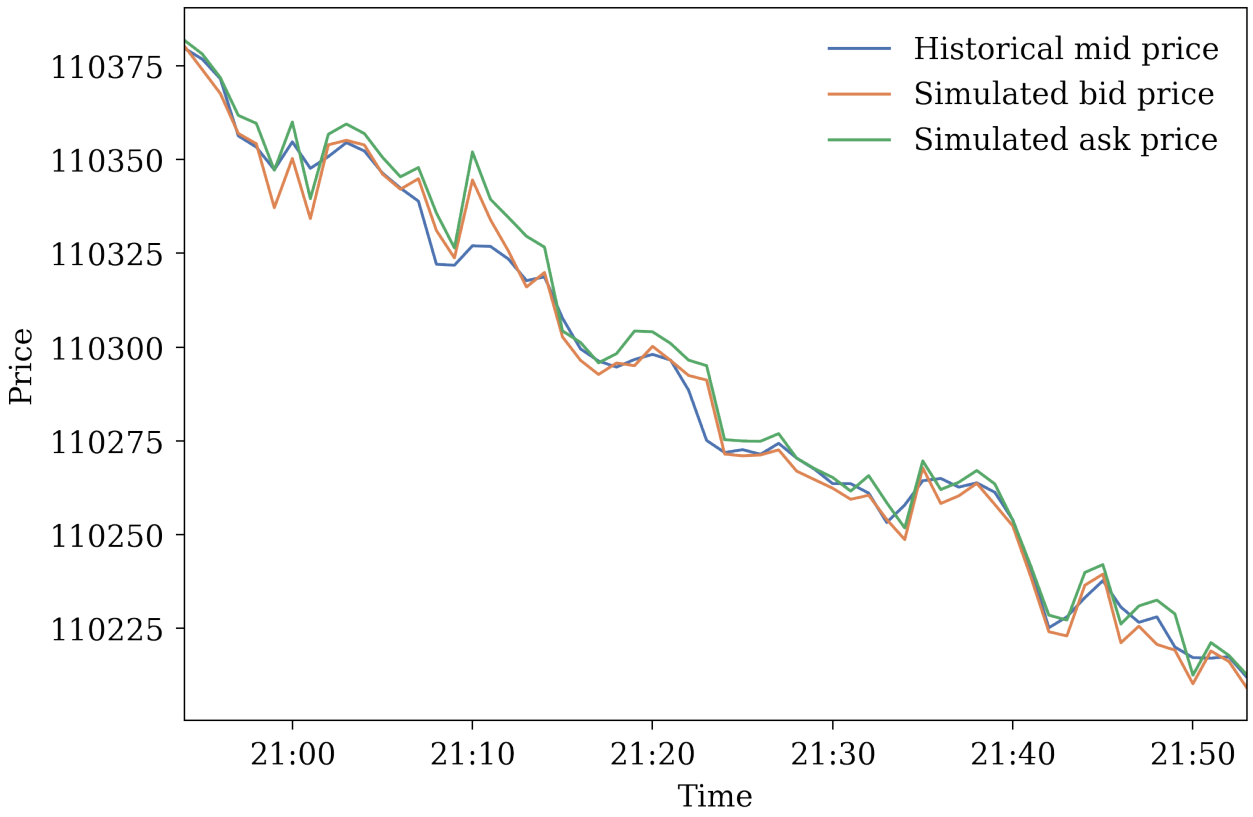
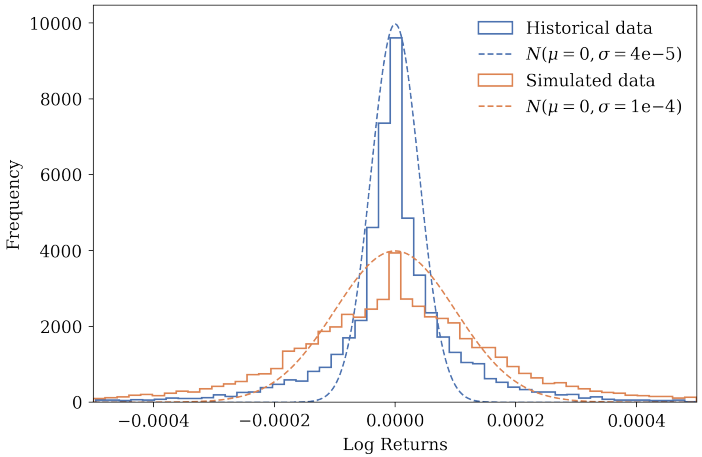


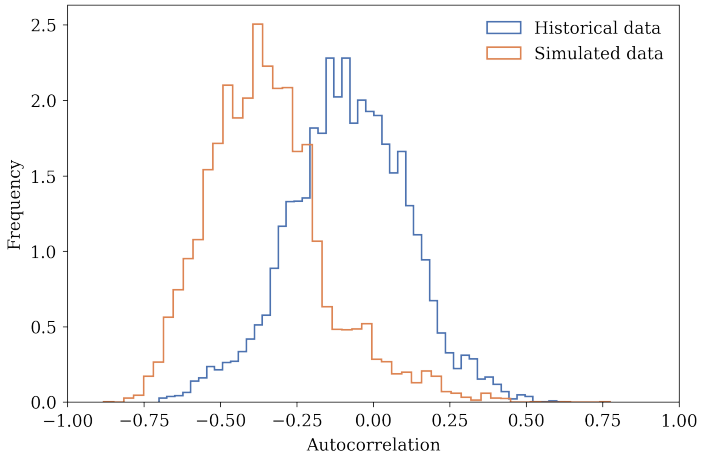
Figure: Comparison of historical and simulated prices.

- Simulated prices closely match historical data

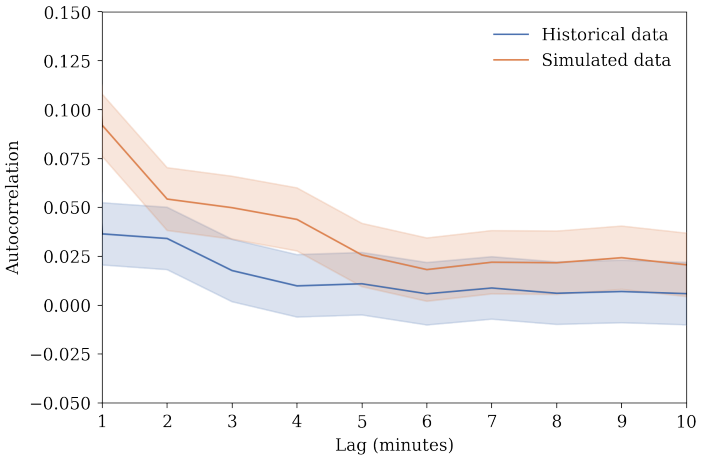
Results: Environment Analysis II



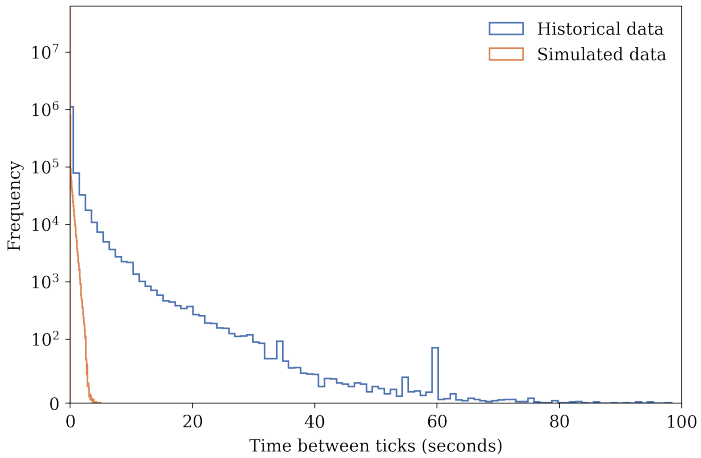
(a) Log Return Distribution



(b) Return Autocorrelation

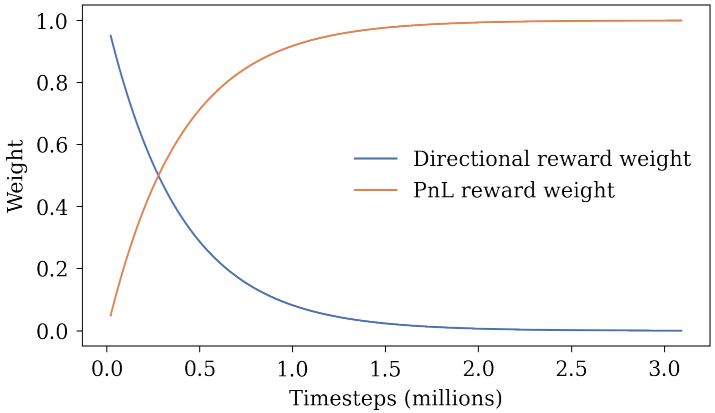


(c) Volatility Autocorrelation

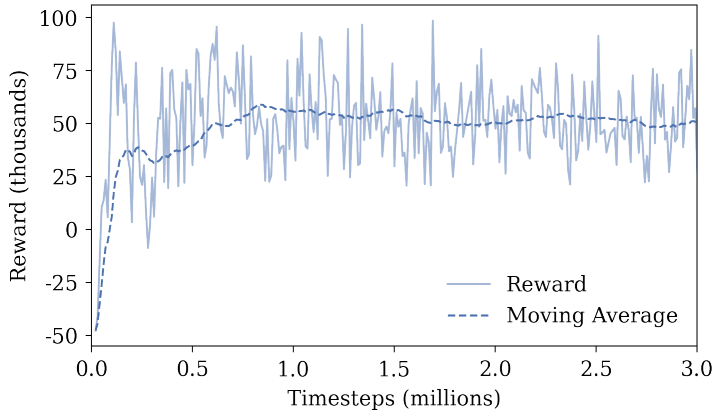


(d) Inter-arrival Times

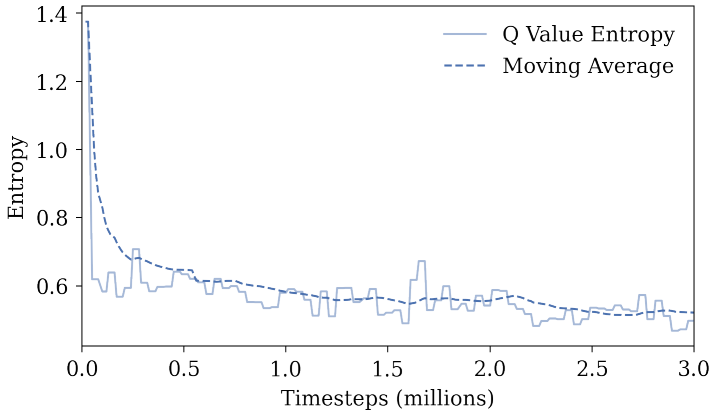
Results: Training Performance (High Volatility)



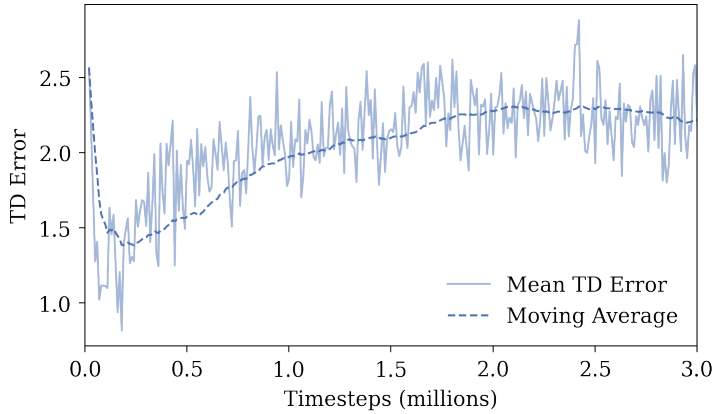
(a) Reward Weight



(b) Reward



(c) Entropy



(d) TD Error / Loss

- Transition from directional to profit-based reward

Results: Baseline Strategies Returns

Table: Average return per episode for the baseline strategies across market regimes. The values represent the average percentage return per episode (one hour) in the testing set.

Agent	Market Type								Agent Avg.
	HV	LV	UT	DT	NT	UJ	DJ	MX	
Random	-0.08	-0.05	-0.13	-0.05	-0.05	-0.14	-0.02	-0.10	-0.08
Buy and Hold	0.05	-0.01	0.09	-0.11	-0.01	0.07	-0.08	0.01	0.00
Aggressive	**0.86**	**0.38**	**0.35**	**0.46**	**0.36**	**0.45**	**0.51**	**0.65**	**0.50**
Market Avg.	0.28	0.11	0.10	0.10	0.10	0.12	0.14	0.19	0.14

- Environments are challenging for simple strategies
- Aggressive agent outperforms other baselines in all market regimes

Results: Returns

Table: Average return per episode for each agent in different market conditions. The values represent the average percentage return per episode (one hour) in the testing set.

Agent	Market Type								Agent Avg.
	HV	LV	UT	DT	NT	UJ	DJ	MX	
High volatility	0.42	0.16	0.19	0.19	0.17	0.23	**0.24**	**0.31**	0.24
Low volatility	0.17	0.11	0.11	0.12	0.10	0.10	0.11	0.11	0.12
Upward trend	0.21	0.12	0.13	0.14	0.13	0.14	0.12	0.14	0.14
Downward trend	0.33	0.17	0.19	0.18	0.18	0.20	0.19	0.28	0.21
No trend	0.39	**0.26**	**0.21**	**0.23**	**0.22**	0.21	**0.24**	0.30	**0.26**
Upward jump	0.41	0.19	0.18	**0.23**	0.19	**0.25**	0.18	0.25	0.24
Downward jump	**0.48**	0.10	0.11	0.11	0.10	0.13	0.12	0.20	0.17
Mixed	0.21	0.11	0.10	0.12	0.08	0.12	0.16	0.16	0.13
Market Avg.	0.33	0.15	0.15	0.16	0.15	0.17	0.17	0.22	0.19

- All agents consistently achieve positive returns
- High volatility is the most profitable regime, rest are similar
- No agent beat the aggressive baseline return

Results: Sharpe Ratio

Table: Annualized Sharpe ratio for each agent in different market regimes. The values represent the annualized Sharpe ratio assuming a risk-free rate of zero.

Agent	Market Type								Agent Avg.
	HV	LV	UT	DT	NT	UJ	DJ	MX	
High volatility	**3.48**	2.21	2.79	2.55	2.34	2.64	3.05	**3.34**	2.80
Low volatility	0.99	**3.65**	2.96	3.31	**3.20**	**3.23**	3.03	2.65	2.88
Upward trend	1.76	2.17	2.76	2.78	2.67	2.88	2.26	2.57	2.48
Downward trend	1.39	2.57	2.16	2.33	2.64	2.41	2.36	2.42	2.29
No trend	1.95	2.64	2.46	2.20	2.43	2.14	2.47	2.45	2.34
Upward jump	3.08	2.60	2.95	2.70	2.94	2.79	2.37	3.10	2.82
Downward jump	1.82	3.24	**3.66**	**3.84**	2.70	2.74	**3.49**	2.67	**3.02**
Mixed	2.72	2.69	2.78	2.94	2.30	2.76	2.95	2.74	2.73
Market Avg.	2.15	2.72	2.82	2.83	2.65	2.70	2.75	2.74	2.67

- All agents surpassed the aggressive baseline Sharpe ratio of 1.43
- Demonstrates robust risk management capabilities in all agents

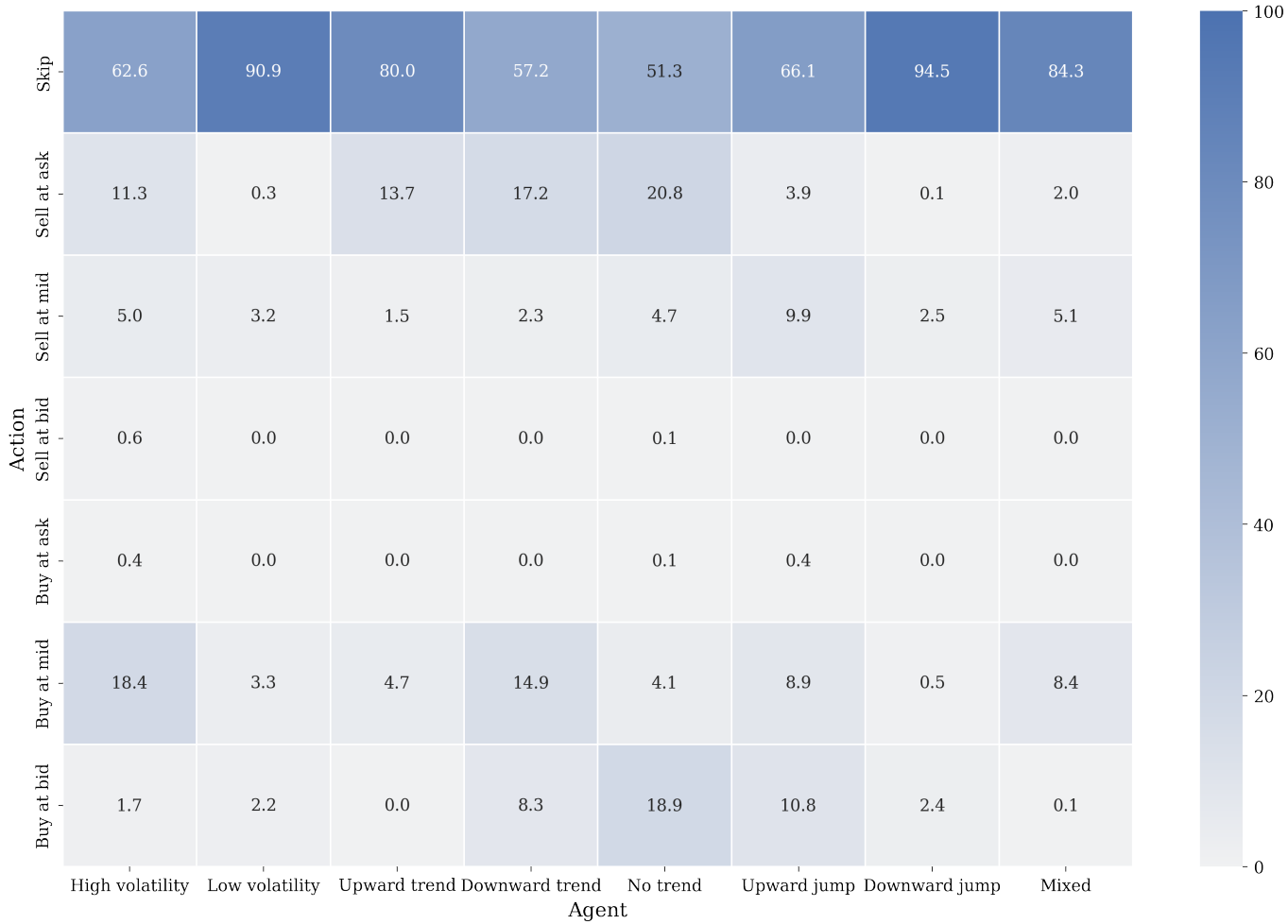
Results: Excess Returns

Table: Average excess return per episode for each agent in different market regimes. The values represent the average percentage excess return per episode (one hour) in the testing set.

Agent	Market Type								Agent Avg.
	HV	LV	UT	DT	NT	UJ	DJ	MX	
High volatility	0.32	0.17	0.02	0.38	0.18	0.11	**0.38**	**0.27**	0.23
Low volatility	0.07	0.13	-0.06	0.31	0.11	-0.03	0.20	0.04	0.10
Upward trend	0.11	0.14	-0.05	0.31	0.13	0.02	0.25	0.11	0.13
Downward trend	0.23	0.19	0.02	0.37	0.19	0.09	0.32	0.24	0.21
No trend	0.29	**0.27**	**0.04**	0.42	**0.22**	0.08	0.37	0.26	**0.24**
Upward jump	0.32	0.20	0.02	**0.43**	0.20	**0.13**	0.32	0.22	0.23
Downward jump	**0.38**	0.12	-0.05	0.30	0.11	0.00	0.26	0.18	0.16
Mixed	0.10	0.13	-0.07	0.31	0.09	0.00	0.30	0.13	0.12
Market Avg.	0.23	0.17	-0.02	0.35	0.16	0.05	0.30	0.18	0.18

- Correlation between returns and market returns is 0.178
- The aggressive baseline had 0.17% excess return in UT

Results: Trading Style

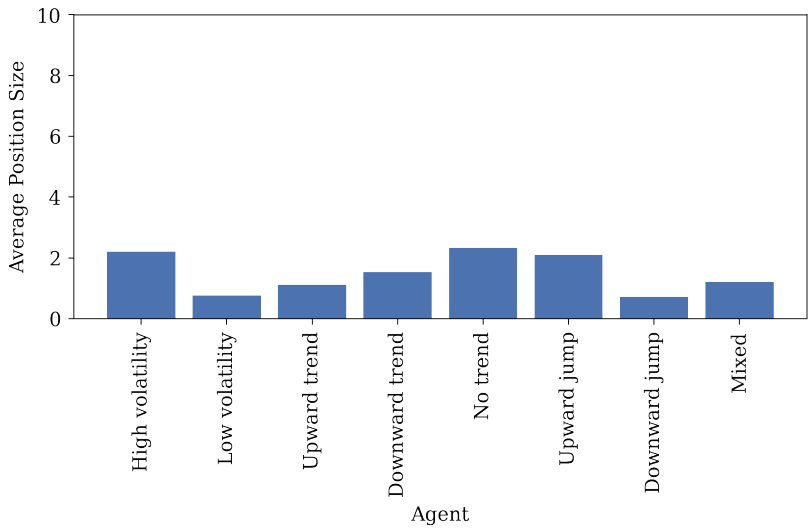


- Agents exhibit a passive trading style, skip most frequently
- Some variation in trading style across agents

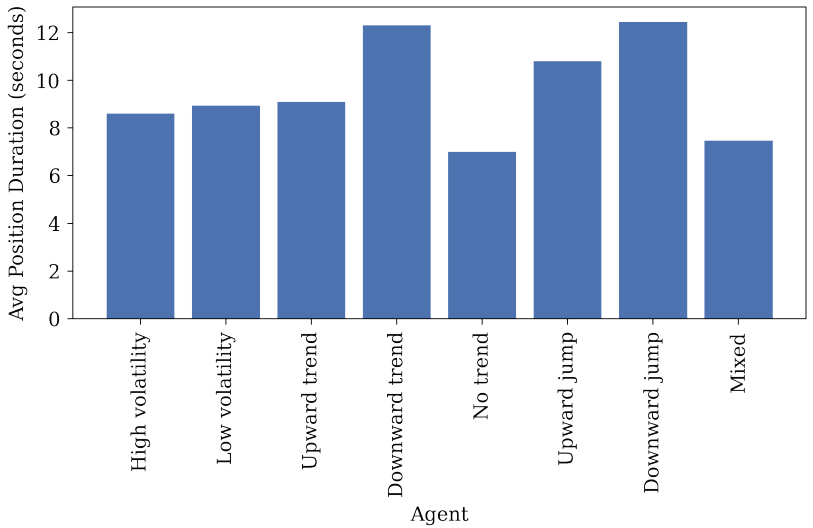
Results: Trading Style II

- Minor differences in trading style across market regimes
- Same pattern can be seen for individual agents

Results: Position Analysis



(a) Average Position Size



(b) Average Position Duration

- Agents maintain small position sizes and short holding periods
- Conservative trading style with limited exposure
- Aggressive agent: 9.11 average size and 60 sec average duration
- Pearson correlation between position size and returns is 0.52

Conclusions and Future Work

Conclusions

- Deep Q-learning agents can learn profitable, uncorrelated trading strategies in simulated FX markets
- Agents demonstrate superior risk-adjusted returns compared to baselines
- Conservative trading style with short holding periods and small positions
- Limited adaptability to changing market conditions during execution

Future Work

- Implement more sophisticated curriculum learning approaches
- Develop automated methods to improve simulation fidelity
- Expand action space to allow deeper order book interactions

Acknowledgements

- Dr Paris Pennesi
- Dr Pawel Chilinski
- Dr Riccardo Della Vecchia

References I

[1] Monetary and Economic Department, “Triennial central bank survey: Otc foreign exchange turnover in april 2022,” Bank for International Settlements, October 2022. [Online]. Available: http://www.bis.org/statistics/rpfx22.htm

[2] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, pp. 529–533, 2015. [Online]. Available: https://doi.org/10.1038/nature14236

[3] H. van Hasselt, A. Guez, and D. Silver, “Deep reinforcement learning with double q-learning,” 2015.

References II

[4] Z. Wang, T. Schaul, M. Hessel, H. van Hasselt, M. Lanctot, and N. de Freitas, “Dueling network architectures for deep reinforcement learning,” 2016.

[5] M. G. Bellemare, W. Dabney, and R. Munos, “A distributional perspective on reinforcement learning,” 2017. [Online]. Available: https://arxiv.org/abs/1707.06887

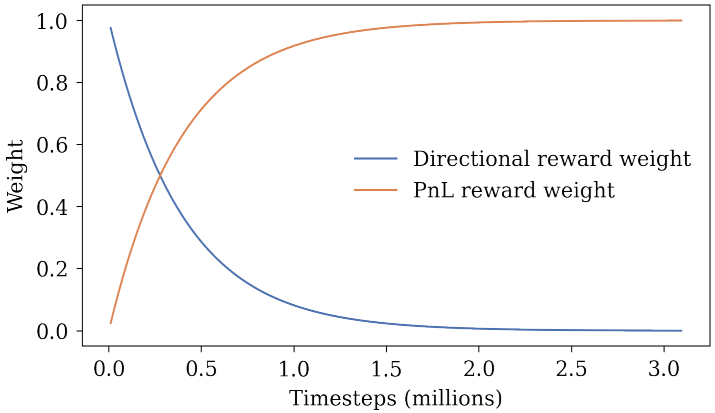
[6] R. Sutton and A. Barto, “Reinforcement learning: An introduction,” *IEEE Transactions on Neural Networks*, vol. 9, no. 5, pp. 1054–1054, 1998.

[7] D. Horgan, J. Quan, D. Budden, G. Barth-Maron, M. Hessel, H. van Hasselt, and D. Silver, “Distributed prioritized experience replay,” 2018.

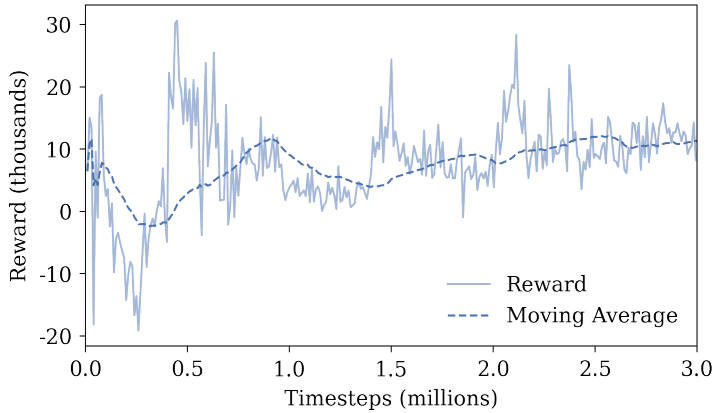
References III

[8] D. Byrd, M. Hybinette, and T. H. Balch, “Abides: Towards high-fidelity multi-agent market simulation,” in *Proceedings of the 2020 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, ser. SIGSIM-PADS ’20. New York, NY, USA: Association for Computing Machinery, 2020, paper https://arxiv.org/pdf/1904.12066, p. 11–22. [Online]. Available: https://doi.org/10.1145/3384441.3395986

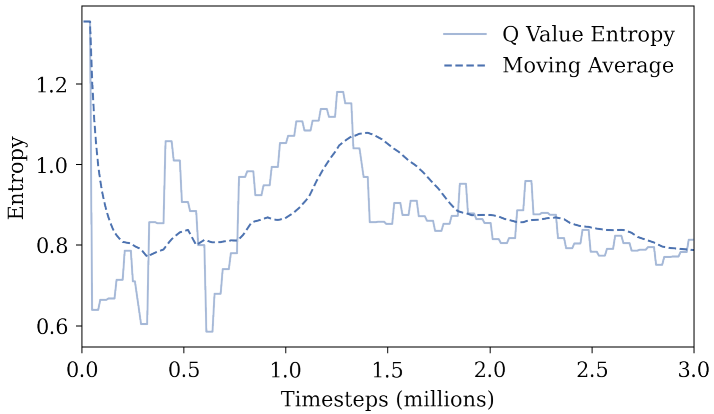
Appendix: Training Performance (Upward Trend)



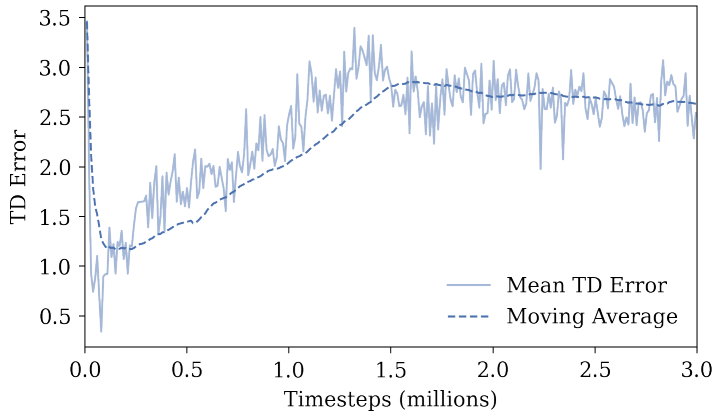
(a) Reward Weight



(b) Reward



(c) Entropy



(d) TD Error / Loss

Appendix: Baseline Strategies Excess Returns

Table: Average excess return for the baseline strategies across market regimes. The values represent the average percentage excess return in the testing set.

Agent	Market Type								Agent Avg.
	HV	LV	UT	DT	NT	UJ	DJ	MX	
Random	-0.18	-0.03	-0.30	0.14	-0.03	-0.26	0.12	-0.14	-0.09
Buy and Hold	-0.05	0.00	-0.08	0.08	0.00	-0.06	0.06	-0.02	-0.01
Aggressive	0.76	0.38	0.17	0.66	0.37	0.32	0.66	0.60	0.49
Market Avg.	0.18	0.12	-0.07	0.29	0.11	0.00	0.28	0.15	0.13

Appendix: Baseline Strategies Sharpe Ratios

Table: Average Sharpe ratio for the baseline strategies across market regimes. The values represent the average Sharpe ratio in the testing set.

Agent	Market Type								Agent Avg.
	HV	LV	UT	DT	NT	UJ	DJ	MX	
Random	-0.13	-0.24	-0.62	-0.26	-0.25	-0.68	-0.10	-0.34	-0.33
Buy and Hold	0.06	-0.03	0.27	-0.34	-0.02	0.18	-0.19	0.04	0.00
Aggressive	1.94	1.02	1.05	1.39	1.24	1.30	1.61	1.89	1.43
Market Avg.	0.63	0.25	0.24	0.26	0.32	0.27	0.44	0.53	0.37

Appendix: Baseline Strategies Max Drawdown

Table: Average maximum drawdown for the baseline strategies across market regimes. The values represent the average maximum drawdown in the testing set.

Agent	Market Type								Agent Avg.
	HV	LV	UT	DT	NT	UJ	DJ	MX	
Random	-0.22	-0.09	-0.16	-0.11	-0.10	-0.17	-0.11	-0.17	-0.14
Buy and Hold	-0.19	-0.06	-0.06	-0.13	-0.08	-0.10	-0.15	-0.14	-0.11
Aggressive	-0.14	-0.07	-0.06	-0.06	-0.05	-0.07	-0.06	-0.07	-0.07
Market Avg.	-0.18	-0.07	-0.09	-0.10	-0.07	-0.11	-0.10	-0.13	-0.11

Appendix: Max Drawdown

Table: Maximum drawdown for each agent in different market regimes. The maximum drawdown is the largest loss an agent would have experienced during a single episode (one hour). The values represent the maximum drawdown in percentage points.

Agent	Market Type								Agent Avg.
	HV	LV	UT	DT	NT	UJ	DJ	MX	
High volatility	-0.04	-0.01	-0.02	-0.01	-0.01	-0.01	-0.02	-0.02	-0.02
Low volatility	-0.07	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02	-0.02	-0.02
Upward trend	-0.05	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02	-0.02	-0.02
Downward trend	**-0.13**	-0.01	**-0.03**	-0.02	-0.01	**-0.03**	-0.03	-0.03	**-0.04**
No trend	-0.07	**-0.02**	-0.01	-0.02	-0.01	-0.03	-0.03	**-0.04**	-0.03
Upward jump	-0.04	-0.01	-0.01	-0.02	-0.01	-0.01	**-0.04**	-0.02	-0.02
Downward jump	-0.08	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	-0.03	-0.02
Mixed	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Market Avg.	-0.06	-0.01	-0.01	-0.01	-0.01	-0.02	-0.02	-0.02	-0.02