Effects of exploration-exploitation strategies in dynamic Forex markets

The use of Reinforcement Learning in Algorithmic Trading

Serban Mihai-Radu, Supervisors: Amin Sharifi Kolarijani, Antonis Papapantoleon, Neil Yorke-Smith

Introduction

- The **foreign exchange (Forex)** market is the largest and most liquid financial market in the world, with daily **trading volumes exceeding $7.5 trillion**. Its decentralized structure, continuous operation, and sensitivity to macroeconomic and geopolitical factors make it **highly volatile** and **difficult to model**. In such environments, **traditional rule-based** or **static statistical strategies** often **fail to adapt** to shifting market dynamics.
- **Reinforcement Learning (RL)** presents a promising alternative by **enabling agents to learn directly from interaction with the market** and **optimize decision-making** over time. However, a **critical challenge** in applying RL to financial domains lies in the design of **exploration-exploitation strategies**—determining how an **agent should balance trying new actions with leveraging past experience**.
- This project investigates **how different exploration mechanisms** affect **learning stability**, **policy quality**, and **trading performance** in non-stationary Forex environments, using a **controlled deep Q-learning framework**.

Background

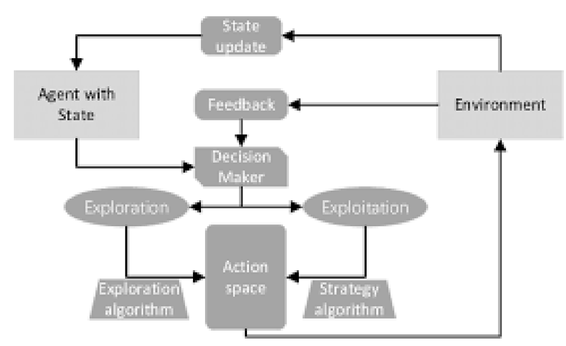
Reinforcement Learning (RL) is a framework for sequential decision-making, where agents learn by **interacting with an environment** and receiving feedback in the form of **rewards**.

- $\mathcal{S}$ is the state space,
- $\mathcal{A}$ is the action space,
- $\mathcal{P}(s'|s,a)$ is the transition probability,
- $\mathcal{R}(s,a)$ is the reward function,
- $\gamma \in [0,1)$ is a discount factor.

The objective is to learn a policy $\pi(a|s)$ that maximizes the expected cumulative discounted reward:

$$J(\pi) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$$

Exploration-exploitation strategies determine how agents balance **trying new actions** versus **using known profitable ones**.



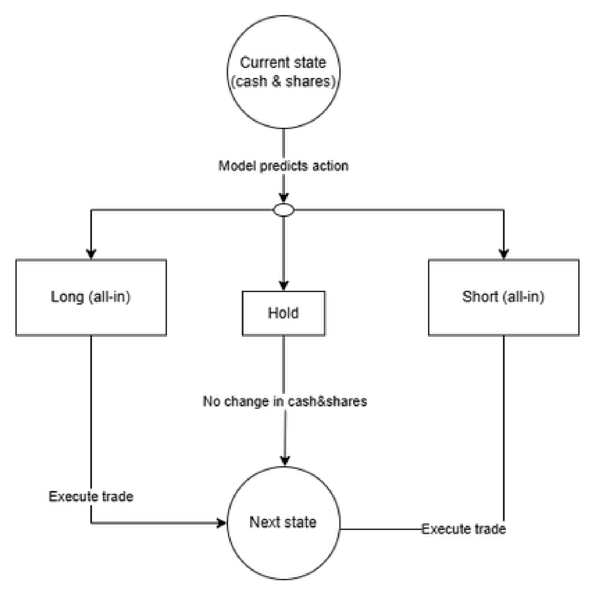
This study focuses on three strategies:

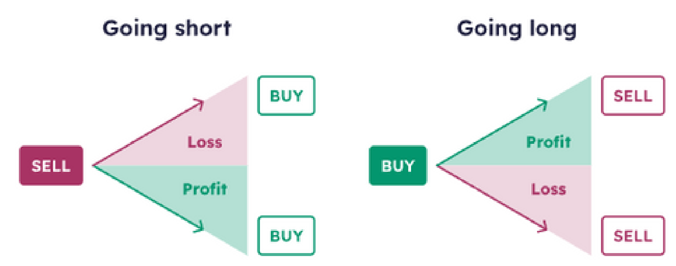
- **Epsilon-Greedy** $a_t = \begin{cases} \text{random action} & \text{with probability } \varepsilon \\ \arg\max_a Q(s_t,a) & \text{with probability } 1-\varepsilon \end{cases}$
- **Boltzmann (Softmax)** $P(a \mid s_t) = \dfrac{\exp(Q(s_t,a)/\tau)}{\sum_{a'} \exp(Q(s_t,a')/\tau)}$
- **Max-Boltzmann (hybrid)** $a_t = \begin{cases} \text{sample from softmax}(Q(s_t,\cdot)) & \text{with probability } \varepsilon \\ \arg\max_a Q(s_t,a) & \text{with probability } 1-\varepsilon \end{cases}$

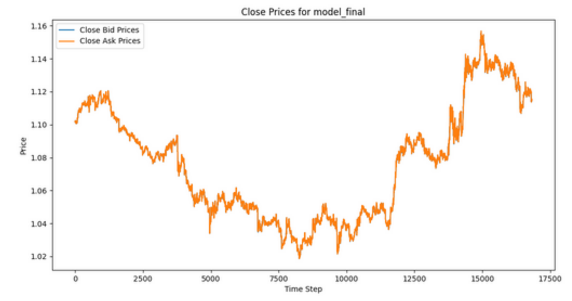
Methods

The agent operates within a **custom Gym-compatible trading environment** that simulates a **realistic foreign exchange** (Forex) setting.

At each timestep, **the agent selects one of three discrete actions**: open a **long position (buy)**, open a **short position (sell)**, or **hold** cash. These actions map to **target exposures of +1, −1, or 0**, respectively, with the entire portfolio allocated accordingly.







Execution is **asymmetric: long trades** are **opened at the ask price** and **closed at the bid**; **short trades** follow the **inverse**. The portfolio **equity is updated after each action**, and the agent receives a **reward equal to the logarithmic change in equity**—capturing relative performance while mitigating scale sensitivity.

$$r_t = \log(E_t) - \log(E_{t-1})$$

A **Deep Q-Network (DQN)** is used to learn the **action-value function Q(s,a)**, enabling the agent to **estimate expected long-term returns** for each action given the current state. The architecture consists of **two hidden layers with ReLU activations**.

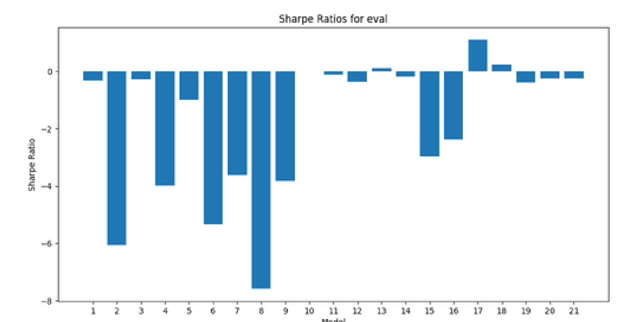
Training is conducted via **experience replay**, with minibatches sampled from a fixed-size buffer, and target network updates occurring periodically to stabilize learning. **All hyperparameters are fixed across experiments** to isolate the effects of the exploration strategy. **Exploration behavior is implemented using custom subclasses of the DQN agent**.

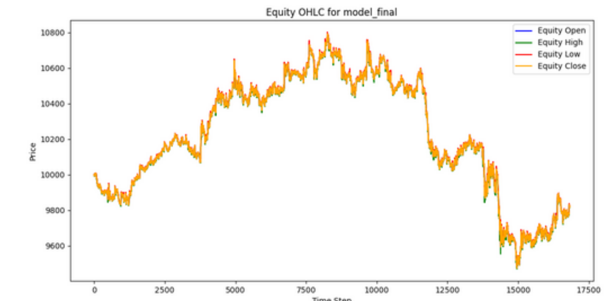
Training is carried out over **20 episodes**, each representing a **complete chronological pass through historical EUR/USD data**. At the **end of every episode**, the **model is checkpointed** and **evaluated on a held-out test set**. During **evaluation**, the **agent acts greedily (without exploration)**, allowing for consistent performance assessment. **Key metrics**—including **Sharpe Ratio, Maximum Drawdown, Profit Factor, Win Rate**, and **total equity gain**—are logged **after each checkpoint** to **track learning progress** and compare the stability and effectiveness of different exploration strategies.

Results

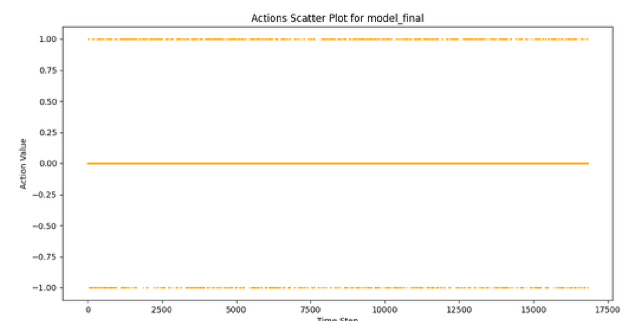
All agents were trained under **identical conditions**, with **only the exploration strategy varied**.

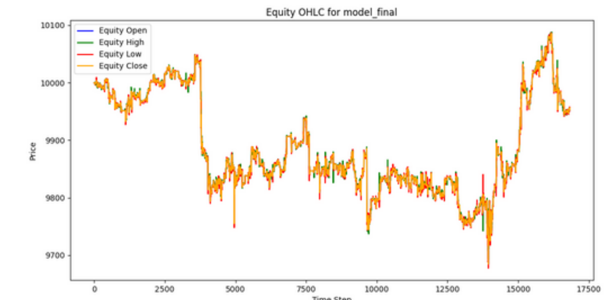
The **Epsilon-Greedy** agent **showed unstable learning behavior** and **frequent policy collapse**. It often **converged to near-inactive strategies**, executing very few trades **during evaluation**. Its final performance yielded a **negative Sharpe Ratio** and the **lowest equity return of all strategies.**



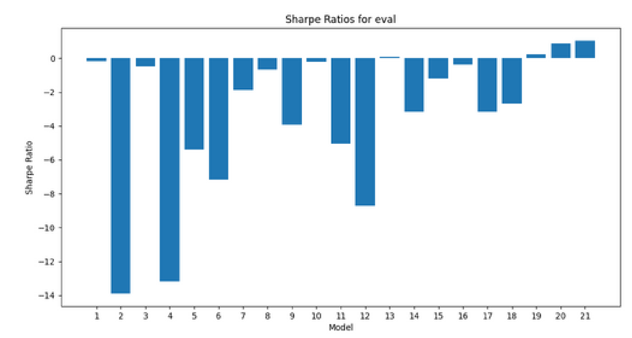


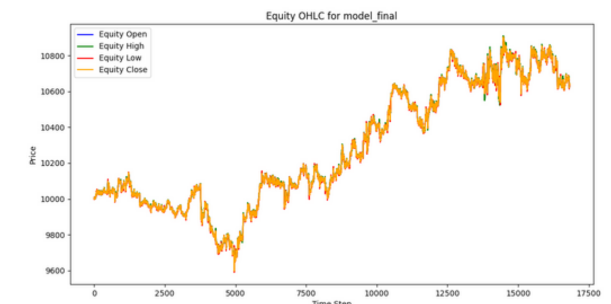
The **Boltzmann agent** demonstrated a **more balanced action distribution** and **slightly improved risk-adjusted returns**. However, its **performance remained volatile** across checkpoints, and it **failed to consistently outperform random baselines** or consolidate gains in later training stages.





The **Max-Boltzmann agent** achieved **the strongest and most stable results across all metrics.** It showed a **clear upward trend in Sharpe Ratio** and **equity growth** throughout training, along with **moderate drawdown and consistent trading activity**. This agent **effectively balanced exploration and exploitation**, **avoiding premature convergence** while **limiting exposure to high-risk actions**.





Evaluation metrics—such as Sharpe Ratio, Profit Factor, Maximum Drawdown, and total equity gain—**consistently favored Max-Boltzmann**, highlighting its robustness in non-stationary market conditions. Performance trends across checkpoints further supported its stability and learning efficiency over time.

Discussion and Conclusions

Exploration strategy had a **decisive impact** on both the **learning dynamics** and **ultimate performance** of RL agents.

Epsilon-Greedy, despite being widely used, **frequently resulted in unstable behavior or complete policy collapse** due to its **uniform, value-agnostic sampling**. It **lacked the ability to distinguish between moderately good and clearly poor actions**, leading to **ineffective or overly cautious policies**.

Boltzmann exploration offered **some improvement** by **biasing action selection toward higher-value options**, but its **lack of structured exploitation** often **prevented the agent from consolidating gains**.

In contrast, the **Max-Boltzmann** strategy demonstrated a **clear and sustained advantage**, by **combining stochastic, value-weighted exploration with fallback greedy actions.** This approach proved **especially valuable in financial environments**, where single missteps can cause compounding losses. The **Max-Boltzmann** agent **achieved the highest Sharpe Ratios, profit factors, and equity growth**, while maintaining smooth, convergent learning curves. **These findings suggest that hybrid, value-aware exploration strategies are not merely preferable—they may be essential in high-risk, non-stationary RL domains like trading.**

Future research could explore **curiosity-driven methods** and **more nuanced reward functions** to further enhance adaptability and robustness.

References

[1] Bank for International Settlements. Triennial Central Bank Survey: Foreign Exchange Turnover in April 2022. Bank for International Settlements, October 2022.
[2] Richard S. Sutton and Andrew G. Barto. Reinforcement Learning: An Introduction. MIT Press, 2 edition, 2018.
[3] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey, 1996.
[4] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al. Human-level control through deep reinforcement learning. Nature, 518(7540):529–533, Feb 2015.
[5] Chien-Yi Huang. Financial trading as a game: A deep reinforcement learning approach. arXiv preprint arXiv:1807.02787, 2018.
[6] Valentina Zangirolami and Matteo Borrotti. Dealing with uncertainty: balancing exploration and exploitation in deep recurrent reinforcement learning, 2024.