Bachelor Informatica

University of Amsterdam

A Controlled Study of Quadratic Risk Aversion in Multi-Agent Reinforcement Learning for Limit-Order-Book Trading

Adam Ru Kun Dong

May 10, 2026

Supervisor(s): Sven Karbach

Signed: Signees

Abstract

This thesis studies quadratic risk aversion in multi-agent reinforcement learning for high-frequency limit-order-book trading. Using JaxMARL-HFT, a market-making agent and an order-execution agent are trained concurrently with independent PPO. Risk sensitivity is introduced through configurable quadratic penalties: a squared-inventory penalty for the market maker and a running squared remaining-quantity penalty for the execution agent, alongside a terminal non-completion penalty for unfinished execution tasks. By varying the corresponding coefficients, ρ_{MM} and ρ_{EX}, the thesis investigates how risk aversion affects learned behaviour, risk–return outcomes, and interaction dynamics. The study uses risk-aversion sweeps, behavioural diagnostics, out-of-sample evaluation, and cross-play between agents trained under different risk levels. The goal is to determine whether quadratic penalties lead to more stable and risk-aware trading behaviour, or mainly alter activity levels and execution patterns.

Contents

CHAPTER 1

Introduction

High-frequency trading (HFT) is commonly organised around limit order books (LOBs), where prices emerge from the continuous submission, cancellation, and matching of orders by many heterogeneous market participants. At this level of market microstructure, two canonical trading tasks are market making and order execution. A market maker provides liquidity by continuously posting bid and ask quotes, aiming to earn the spread while controlling the inventory accumulated through asymmetric fills. An execution trader, by contrast, must buy or sell a target quantity over a fixed horizon, balancing the cost of demanding immediate liquidity against the risk of delaying too much of the order. Both problems have been studied extensively in mathematical finance, where risk aversion is not treated as an optional add-on but as a central part of the objective. The Avellaneda–Stoikov model, for example, derives inventory-sensitive market-making quotes from a risk-averse control problem [6], while the Almgren–Chriss model formulates optimal execution as a trade-off between expected transaction cost and timing risk [2].

Reinforcement learning (RL) provides a flexible alternative to these classical model-based approaches. Instead of specifying a closed-form stochastic model for prices, order arrivals, or market impact, RL agents can learn trading policies directly through interaction with a simulated or data-driven market environment. This is especially attractive in HFT, where limit-order-book dynamics are high-dimensional, noisy, and difficult to model analytically. However, this flexibility also makes reward design central. In particular, it raises the question of how risk aversion should be represented when RL agents are trained for trading tasks that are traditionally defined by explicit risk-return trade-offs.

This question becomes more important in a multi-agent setting. Market-making and execution agents do not act in isolation: the execution agent consumes or supplies liquidity, while the market maker adjusts quotes in response to order flow, inventory, and adverse selection. Studying either agent against a fixed market background can therefore miss important interaction effects. Multi-agent reinforcement learning (MARL) offers a natural framework for this problem because multiple agents can learn concurrently in the same limit-order-book environment. Until recently, however, applying MARL to HFT at realistic scale was computationally difficult. The JaxMARL-HFT framework [26] addresses this limitation by providing a GPU-accelerated, LOBSTER-fed MARL environment for HFT with heterogeneous agents, flexible observation and action spaces, configurable reward functions, and large-scale parallel training. This makes it feasible to run controlled experiments on risk-sensitive agent behaviour.

1.1 Problem

Existing reinforcement-learning approaches to high-frequency trading often include risk-control terms in the reward, but these terms are rarely studied systematically as the main experimental variable. In market making, inventory penalties are used to discourage large positions. In order execution, terminal penalties on unfinished quantity are often used to encourage task completion. However, it remains unclear whether such penalties produce genuinely risk-aware behaviour or whether they mainly change activity patterns. A market maker may reduce inventory exposure by managing inventory more effectively, but it may also do so by trading less. Similarly, an execution agent may reduce unfinished quantity by following a more stable execution schedule, but it may also become overly aggressive and incur higher execution costs.

This problem is especially relevant in a multi-agent setting, where market-making and order-execution agents interact through the same limit order book. The risk aversion of one agent may affect not only its own policy, but also the performance and behaviour of the other agent. A more risk-averse market maker may change the liquidity available to the execution agent, while a more aggressive execution agent may change the order-flow pressure faced by the market maker.

JaxMARL-HFT provides a suitable environment for studying this problem because it supports heterogeneous agents, flexible reward functions, and concurrent training with independent PPO. Its existing experiments include a two-agent market-making and order-execution setup, as well as a fixed quadratic inventory penalty for the market maker and execution-related completion penalties. However, these risk-related terms are not isolated through a controlled study of separate risk-aversion coefficients for both agents.

The concrete problem this thesis addresses is therefore the absence of a controlled empirical study of how separate quadratic risk-aversion coefficients, ρ_{MM} and ρ_{EX}, affect learned behaviour, risk–return trade-offs, and interaction dynamics in a two-agent HFT MARL setting.

1.2 Motivation

Quadratic risk aversion provides an interpretable and feasible way to connect classical trading intuition with reinforcement learning. For the market maker, a squared-inventory penalty targets the risk of accumulating large long or short positions. For the execution agent, a squared remaining-quantity penalty targets the risk of delaying too much of the order until the end of the horizon. Both penalties are additive over time, controlled by scalar risk-aversion coefficients, and compatible with PPO-style policy-gradient training.

This makes quadratic risk aversion a suitable focus for a controlled bachelor-thesis project. Rather than comparing several mathematically different risk-sensitive objectives, this thesis studies one objective family in depth. This narrower scope makes it possible to vary the risk-aversion coefficients systematically, evaluate out-of-sample performance, inspect learned behaviour, and perform cross-play between agents trained under different risk levels.

The motivation is not only to test whether quadratic penalties improve average performance. The central question is behavioural: whether stronger risk aversion leads to genuine inventory control and more stable execution, or whether it mainly produces inactivity, excessive execution aggressiveness, or other unintended effects.

1.3 Research Questions

This thesis addresses the following main research question:

How does quadratic risk aversion affect the learned behaviour, risk–return trade-off, and interaction dynamics of concurrently trained market-making and order-execution agents in a limit-order-book multi-agent reinforcement learning environment?

Quadratic risk aversion is studied through two configurable risk-aversion coefficients: ρ_{MM} for the market maker's squared-inventory penalty and ρ_{EX} for the execution agent's squared remaining-quantity penalty. To answer the main research question, the thesis considers the following sub-questions:

1. **Market-making behaviour:** How does increasing the market maker's quadratic inventory-risk coefficient ρ_{MM} affect inventory exposure, quoting behaviour, and trading activity?
2. **Execution behaviour:** How does increasing the execution agent's quadratic remaining-position coefficient ρ_{EX} affect execution speed, execution stability, and order aggressiveness?
3. **Risk–return trade-off:** How do the risk-aversion coefficients ρ_{MM} and ρ_{EX} affect performance and risk metrics, including final portfolio value, execution cost, variance of outcomes, inventory exposure, unfinished quantity, and drawdown of portfolio value where applicable?
4. **Multi-agent interaction:** How does the risk aversion of one agent affect the behaviour and performance of the other agent?

1.4 Proposed Experimental Setup

The empirical study will be conducted in the JaxMARL-HFT environment, using the two-agent setting with one market-making agent and one order-execution agent trained concurrently with independent PPO. JaxMARL-HFT is suitable for this study because it supports heterogeneous agents with different observation spaces, action spaces, and reward functions, and because its GPU-accelerated design makes repeated training runs and coefficient sweeps feasible [26]. The planned base environment follows the JaxMARL-HFT market-by-order replay setting, using AMZN limit-order-book data subject to data availability and computational constraints.

At a high level, the market-making agent observes limit-order-book and task-specific information, including its own inventory, and selects quoting actions such as posting bid and ask orders at different distances from the spread or choosing not to trade. Its risk-sensitive reward is modified by a quadratic inventory penalty,

$$r_t^{MM} = r_{t,\mathrm{base}}^{MM} - \rho_{\mathrm{MM}} Q_t^2,$$

where Q_t denotes the agent's inventory. The order-execution agent observes market information and private execution information, including the remaining quantity, and chooses execution actions over a finite horizon. Its reward is modified by a running quadratic remaining-position penalty and a terminal non-completion penalty,

$$r_t^{EX} = r_{t,\mathrm{base}}^{EX} - \rho_{\mathrm{EX}} X_t^2 - \kappa |X_T|,$$

where X_t is the remaining quantity and X_T is the unfinished quantity at the end of the episode.

The main experiment is a controlled sweep over the two risk-aversion coefficients. For each agent, the sweep includes a risk-neutral baseline and three positive risk-aversion levels. The final numerical values will be selected after a pilot run that measures the scale of the base rewards and the quadratic penalty terms. For each agent $a \in \{MM, EX\}$, a reference coefficient $\rho_a^\star$ will be chosen such that the quadratic penalty has a comparable magnitude to the base reward on baseline trajectories. The planned sweep is then

$$\rho_a \in \{0,\ 0.25\rho_a^\star,\ \rho_a^\star,\ 4\rho_a^\star\}.$$

This produces separate sweeps for market-maker risk aversion, execution-agent risk aversion, and settings where both agents are risk-sensitive.

Out-of-sample evaluation will be performed by freezing trained policies and evaluating them on limit-order-book episodes that were not used for PPO updates. The data will be split chronologically into training, validation, and test periods. Training episodes are used for learning, validation episodes are used for implementation checks and coefficient-scale selection, and the held-out test period is used for final reported metrics. This prevents the evaluation from measuring only performance on the market conditions seen during training.

To study interaction effects, the thesis will perform cross-play between agents trained under different risk-aversion levels. After training, market-making and execution policies are frozen and recombined. The diagonal of the cross-play matrix evaluates agents paired with counterparties from matching risk settings, while the off-diagonal entries evaluate how each agent performs against counterparties trained with different risk preferences. This allows the thesis to test whether the risk aversion of one agent changes the performance and behaviour of the other agent.

The evaluation metrics are chosen to correspond directly to the research questions. Market-making behaviour is evaluated using final portfolio value, variance of portfolio value, maximum absolute inventory, average squared inventory, no-trade frequency, order-submission rate, cancellation rate, quote skewing, and action distributions. Execution behaviour is evaluated using slippage, variance of slippage, unfinished quantity, average squared remaining quantity, completion time, and passive-versus-aggressive order ratios. Risk–return trade-offs are evaluated by comparing performance and risk metrics across the coefficient sweep. Multi-agent interaction is evaluated through cross-play matrices for both market-maker and execution-agent outcomes.

1.5 Contributions

This thesis does not propose a new reinforcement-learning algorithm or a new limit-order-book simulator. Instead, it contributes a controlled empirical study of quadratic risk aversion in a two-agent high-frequency-trading MARL setting. Building on the JaxMARL-HFT framework, the thesis studies how configurable risk penalties affect the behaviour and interaction of a market-making agent and an order-execution agent trained concurrently in the same limit-order-book environment [26].

The main contributions are as follows:

1. **A unified quadratic-risk formulation for two HFT agent types.** The thesis formulates quadratic risk aversion for both agents in a consistent way. For the market-making agent, risk is represented by a squared-inventory penalty controlled by ρ_{MM}. For the execution agent, risk is represented by a squared remaining-quantity penalty controlled by ρ_{EX}.
2. **A controlled risk-aversion experiment design.** The thesis designs systematic risk-aversion sweeps over ρ_{MM} and ρ_{EX} in order to study how different levels of quadratic risk aversion affect learning outcomes. This allows the effect of risk aversion to be analysed as the main experimental variable rather than as an incidental reward-shaping choice.
3. **Risk–return and behavioural evaluation of learned policies.** The thesis evaluates trained agents using both performance metrics and behavioural diagnostics. For the market-making agent, this includes portfolio value, inventory exposure, inventory variance, drawdown, and trading activity. For the execution agent, this includes execution cost, unfinished quantity, completion behaviour, and execution aggressiveness.
4. **Cross-play analysis of multi-agent interaction effects.** The thesis evaluates agents trained under different risk-aversion levels against one another. This cross-play analysis

is used to study whether the risk preference of one agent affects the performance and behaviour of the other agent, thereby addressing the multi-agent nature of the trading problem.

5. **An empirical bridge between classical risk-aware trading theory and MARL.** By focusing on quadratic inventory and remaining-position penalties, the thesis connects classical ideas from risk-aware market making and order execution with modern GPU-accelerated multi-agent reinforcement learning. The resulting study provides evidence on whether simple and interpretable quadratic penalties lead to genuinely risk-aware trading behaviour in a learned multi-agent setting.

1.6 Scope and Limitations

This thesis focuses on quadratic risk aversion as a single interpretable risk-sensitive objective family. It does not compare quadratic penalties against mean–variance objectives, CVaR, entropic utility, or other distributional risk measures. These objectives are relevant in the broader risk-sensitive reinforcement-learning literature, but they introduce additional implementation and evaluation challenges, including different risk-aversion scales, episode-level distributional estimation, and more complex integration with PPO-style training.

The classical Avellaneda–Stoikov and Almgren–Chriss models are used as conceptual references rather than exact optimal benchmarks. They motivate the use of inventory and remaining quantity as risk variables, but the learned agents in this thesis operate in a replay-based limit-order-book MARL environment with discrete action spaces and data-driven background order flow. Therefore, the experiments evaluate empirical behavioural and risk–return effects rather than convergence to closed-form classical optima.

Finally, the thesis does not propose a new MARL algorithm or a new simulator. It builds on JaxMARL-HFT and uses independent PPO as the training method. The main contribution lies in implementing and evaluating a controlled quadratic-risk experiment pipeline, including configurable risk coefficients, coefficient sweeps, behavioural diagnostics, out-of-sample evaluation, and cross-play between trained agents.

1.7 Thesis Structure

The remainder of this thesis is structured as follows.

Chapter 2 reviews the literature relevant to this thesis. It discusses classical approaches to market making and order execution, reinforcement-learning methods for trading, multi-agent reinforcement learning in financial markets, and recent GPU-accelerated limit-order-book simulation frameworks. It also positions this thesis relative to existing work on risk-sensitive trading objectives.

Chapter 3 introduces the technical concepts required for the thesis. It explains the structure and mechanics of limit order books, the market-making and order-execution tasks, the reinforcement-learning and multi-agent reinforcement-learning setting, independent PPO, and the quadratic risk-aversion formulation used in this study.

CHAPTER 2

Related Work

This chapter positions the thesis within four related lines of work. First, classical market-making and optimal-execution models motivate the risk variables used in this thesis: inventory for market making and remaining quantity for execution. Second, reinforcement-learning approaches to trading show that learned trading behaviour depends strongly on reward design, especially when risk-control terms are included. Third, multi-agent reinforcement-learning studies show that trading agents cannot be evaluated only in isolation, because the policy of one agent changes the environment faced by another. Finally, recent limit-order-book simulation frameworks, especially JaxMARL-HFT, make controlled high-frequency MARL experiments computationally feasible, but do not yet isolate the effect of separate quadratic risk-aversion coefficients for market-making and execution agents.

2.1 Classical Risk-Aware Trading Models

Classical trading models provide the conceptual foundation for the two risk variables studied in this thesis. In market making, the central risk variable is inventory: a dealer who provides liquidity on both sides of the book may unintentionally accumulate a large long or short position. In order execution, the central risk variable is the remaining quantity to be executed: a trader who delays execution may reduce immediate market impact, but remains exposed to price movements and non-completion risk. Although the classical models considered in this section are not used as exact benchmarks, they clarify why Q_t and X_t are natural variables on which to impose quadratic penalties.

Early dealership models already show that liquidity provision cannot be understood as simply earning the bid–ask spread. Garman's dealership framework [13] models order arrivals as stochastic and highlights that a dealer's survival depends on managing inventory and cash constraints. Amihud and Mendelson [3] and Ho and Stoll [18] further develop this view by showing that bid and ask quotes should depend on the dealer's current inventory and exposure to future price movements. The important implication for this thesis is that inventory is not only an accounting variable after trades have occurred; it is a state variable that should affect future decisions.

Avellaneda and Stoikov [6] provide a modern reference point for high-frequency market making. Their model links inventory, price volatility, order-arrival intensities, and risk aversion in a tractable quoting framework. Guéant, Lehalle, and Fernandez-Tapia [16] extend this line of work by deriving approximations for optimal quotes under inventory constraints. These models are useful for this thesis because they suggest the expected qualitative behaviour of a risk-averse market maker: as inventory becomes large, the agent should either skew quotes to encourage inventory reduction or become less willing to add further inventory. However, these models rely on

stylised assumptions about price dynamics and order-arrival processes, while the present thesis studies learned policies in a replay-based limit-order-book environment. Therefore, Avellaneda–Stoikov-type models are used as conceptual motivation rather than as ground-truth optimal policies.

Classical optimal-execution models motivate the execution side of the thesis. Bertsimas and Lo [8] formulate the execution of a large order as an optimal-control problem, while Almgren and Chriss [2] introduce the canonical trade-off between expected transaction cost and volatility risk. Later models, including Obizhaeva and Wang [29] and Alfonsi, Fruth, and Schied [1], move closer to limit-order-book mechanics by incorporating resilience, book shape, and nonlinear impact. The common insight is that delayed execution has a cost even when it avoids immediate aggressive trading: the trader remains exposed while a position is still unfinished.

Taken together, the classical literature justifies the two quadratic risk terms used in this thesis. The market-making penalty $-\rho_{MM}Q_t^2$ reflects the classical concern that large inventory creates exposure to adverse price movements. The execution penalty $-\rho_{EX}X_t^2$ reflects the classical concern that carrying a large unfinished order creates timing and non-completion risk. The contribution of this thesis is not to re-derive optimal stochastic-control policies, but to test how these classical risk variables affect learned behaviour when they are embedded in a two-agent MARL environment.

2.2 Reward Design in Reinforcement Learning for Trading

Reinforcement learning is attractive for limit-order-book trading because market making and execution can both be formulated as sequential decision problems. In both tasks, however, the reward function is not a neutral implementation detail. It determines whether the agent learns to seek profit, reduce risk, complete a task, avoid trading, or exploit artefacts of the simulator. This is especially important for the present thesis, because the main intervention is a change in the reward through quadratic risk penalties.

In reinforcement-learning market making, inventory control appears repeatedly as a central design problem. Chan and Shelton [10] show that adaptive market-making behaviour can be learned without fully specifying the order-arrival process. Lim and Gorse [21] include inventory and remaining time in the state and use a risk-sensitive terminal reward, showing that the learned policy depends on the selected risk-aversion parameter. Spooner et al. [31] train a market-making agent in a limit-order-book simulator and explicitly design the reward to account for inventory risk. The survey by Gašperov et al. [14] places these methods in a broader inventory-control perspective and emphasizes that state representation, action design, and reward formulation strongly affect learned market-making behaviour.

The implication of this literature is that a market maker can reduce inventory risk in several different ways, not all of which are equally desirable. A learned agent might skew quotes as classical theory suggests, but it might also reduce risk simply by submitting fewer orders or avoiding trades altogether. More recent work reinforces this concern. Deep-learning-based market-making studies such as Kumar [20], Guo et al. [15], and Fernández Vicente et al. [34] show that richer models and inventory-aware rewards can improve stability, but they also make evaluation more complex: high reward or low inventory alone does not reveal whether the agent is actively providing liquidity. This directly motivates the behavioural diagnostics in this thesis, including no-trade frequency, order-submission rate, cancellation rate, inventory trajectories, quote skewing, and action distributions.

The same reward-design issue appears in reinforcement-learning order execution. Nevmyvaka, Feng, and Kearns [27] distinguish between private execution variables, such as time and shares remaining, and market variables derived from order-book activity. Hendricks and Wilcox [17] connect reinforcement learning with the Almgren–Chriss framework by adapting execution schedules to market conditions. More recent deep reinforcement-learning approaches, including Ning,

Lin, and Jaimungal [28], Moallemi and Wang [25], and Lin and Beling [22], show that execution policies can be learned from limit-order-book information and may outperform fixed schedules such as TWAP or VWAP.

For this thesis, the key lesson from the execution literature is that the remaining quantity X_t is not merely part of the environment state; it is also a natural object of risk control. A terminal non-completion penalty encourages the agent to finish the task, but it may not determine how risk is distributed across the episode. A running quadratic penalty $-\rho_{EX} X_t^2$ gives the agent an incentive to reduce unfinished quantity earlier. This can lead to more stable execution schedules, but it can also make the agent overly aggressive and increase slippage. Therefore, execution quality must be evaluated jointly with unfinished quantity, completion time, and the passive-versus-aggressive order ratio.

Karpe et al. [19] are especially relevant because they train an execution agent in the ABIDES limit-order-book simulator and show that learned policies may converge toward TWAP-like behaviour in some scenarios. This suggests that fixed schedules remain useful behavioural references even when the goal is not simply to beat a benchmark. In the present thesis, TWAP-style behaviour is therefore treated as a diagnostic reference point: a quadratic remaining-position penalty may produce earlier execution, but the important question is whether this improves the risk-return trade-off or merely shifts cost from non-completion risk to slippage.

Overall, reinforcement-learning trading papers establish that reward design matters as much as algorithm choice. Prior work has studied inventory-aware market making and adaptive execution, but usually treats them separately or focuses on outperforming a benchmark. This thesis instead studies the controlled effect of one interpretable reward modification across both agents: a quadratic penalty on the relevant risky position variable. The central empirical question is not only whether the modified reward improves average performance, but how the coefficient ρ changes behaviour, variance, risk exposure, and trading activity.

2.3 Multi-Agent Reinforcement Learning for Interacting Trading Agents

Financial markets are multi-agent systems: liquidity, prices, spreads, and execution costs emerge from the interaction of many participants. This creates a limitation for single-agent reinforcement-learning studies. If one agent is trained against a fixed background environment, its learned policy may not remain effective when other strategic agents adapt. This is particularly relevant for the present thesis because market making and order execution are complementary tasks. The market maker provides liquidity by posting quotes, while the execution agent may consume or supply liquidity while completing a target order. Changing the risk aversion of one agent can therefore affect not only its own performance but also the other agent's opportunity set.

Early multi-agent financial-market simulations show that aggregate market behaviour can emerge from interacting agents. Lussange et al. [23] use reinforcement-learning agents in a stock-market simulator and evaluate whether the resulting market reproduces empirical market statistics. Yao et al. [35] similarly study reinforcement-learning-based market simulation and stylized facts. These works are important because they move beyond isolated trading agents and treat market behaviour as an interaction outcome. However, their primary goal is market realism, not the controlled isolation of a specific risk-aversion parameter.

A more directly related line of work studies interaction between liquidity providers and liquidity takers. Ganesh et al. [12] train reinforcement-learning market makers in a multi-agent dealer market and show that agents can learn inventory-dependent pricing behaviour under different competitive settings and reward formulations. Ardon et al. [5] train liquidity-provider and liquidity-taker agents simultaneously in a framework aimed at fully reinforcement-learning-based market simulation. Vadori et al. [32] further develop this idea in an over-the-counter market setting, where agents interact through parameterized reward families and liquidity providers can learn emergent hedging and inventory-dependent skewing.

The implication for this thesis is that market-maker risk aversion cannot be evaluated only by looking at the market maker's final portfolio value, and execution-agent risk aversion cannot be evaluated only by looking at the execution agent's slippage. If a more risk-averse market maker trades less, the execution agent may face a different liquidity environment. Conversely, if a more risk-averse execution agent trades earlier or more aggressively, the market maker may face different order-flow pressure and inventory risk. This motivates cross-play evaluation: agents trained under one pair of risk coefficients should be evaluated against agents trained under other coefficients.

Other studies bring multi-agent trading closer to limit-order-book simulation. Karpe et al. [19] use ABIDES to train an execution agent in the presence of other simulated market participants, while Fernández Vicente et al. [33] compare deep Q-learning market makers in non-competitive and competitive simulated markets. These studies show that interaction effects matter for learned trading behaviour. However, they do not systematically vary separate market-making and execution risk-aversion coefficients in a concurrent two-agent high-frequency setting. This is the specific gap addressed by the present thesis.

2.4 LOB Simulation Frameworks and JaxMARL-HFT

Controlled reinforcement-learning research in high-frequency trading requires a simulator that captures important limit-order-book mechanisms while still allowing repeated training and evaluation. Order submission, cancellation, matching, price-time priority, and interaction with background order flow all affect the behaviour learned by trading agents. The choice of simulator is therefore not only an implementation detail; it determines which research questions can be studied.

ABIDES [9] is one of the most influential open-source agent-based market simulation frameworks. It models the market as a discrete-event system in which trading agents interact with an exchange agent through messages, and it supports configurable latencies between agents and the exchange. ABIDES-Gym [4] adds a Gym-style reinforcement-learning interface and has been used for market-making and execution experiments. MAXE [7] and PyMarketSim [24] also provide tools for agent-based limit-order-book simulation and controlled studies of trading agents.

These frameworks establish that simulated limit-order-book environments are useful for reinforcement-learning trading research, but they also reveal a trade-off. High-fidelity simulation can be computationally expensive, especially when many agents, many seeds, and many hyperparameter settings are required. This matters for the present thesis because a controlled ρ-sweep is not a single training run. It requires repeated training across risk coefficients, seeds, and evaluation settings. A framework that is too slow would force the study either to use very few coefficients or to omit cross-play and out-of-sample evaluation.

JAX-LOB [11] addresses part of this computational problem by moving limit-order-book simulation into the JAX ecosystem, enabling GPU-accelerated processing of many books in parallel. JaxMARL [30] provides a complementary JAX-based multi-agent reinforcement-learning library. JaxMARL-HFT [26] combines these directions by extending JaxMARL and building on JAX-LOB to support GPU-accelerated multi-agent reinforcement learning for high-frequency trading. It supports heterogeneous agents with different observation spaces, action spaces, and reward functions, and demonstrates a two-agent setup in which a market-making agent and an order-execution agent are trained concurrently using independent PPO.

JaxMARL-HFT is therefore the most natural starting point for this thesis. It provides exactly the type of setting needed to study interaction between a liquidity-providing market maker and a liquidity-demanding execution agent. At the same time, it also shows why the contribution of this thesis must go beyond simply adding an inventory penalty. JaxMARL-HFT already includes selected experiments with a quadratic market-maker inventory penalty and reports that learned market makers may trade very infrequently, with the no-trade tendency exacerbated by

the inventory penalty [26]. This observation turns the inventory penalty from a straightforward reward modification into an empirical question: does higher ρ_{MM} create genuinely safer inventory management, or does it mainly suppress trading?

The same framework also leaves room for a more systematic execution-risk study. JaxMARL-HFT includes an execution agent and terminal non-completion penalties, but it does not isolate a running quadratic remaining-position penalty $-\rho_{EX} X_t^2$ across controlled coefficient sweeps. Nor does it evaluate the interaction of separately trained market-making and execution agents across a cross-play matrix. These omissions define the experimental contribution of this thesis: to use JaxMARL-HFT not only as a simulator, but as the basis for a controlled quadratic-risk experiment pipeline.

2.5 Positioning of This Thesis

The reviewed literature leads to a specific research gap. Classical trading models identify inventory and remaining quantity as central risk variables, but they do so under stylised stochastic-control assumptions. Reinforcement-learning trading papers show that market-making and execution policies can be learned from limit-order-book information, but they also show that reward design can strongly shape behaviour and may produce unintended inactivity or aggressiveness. Multi-agent reinforcement-learning papers demonstrate that interaction effects matter, but generally focus on market realism, competitive behaviour, or broad agent-based simulation rather than on isolating risk-aversion coefficients. Simulation frameworks make controlled trading-agent experiments possible, and JaxMARL-HFT provides a suitable GPU-accelerated heterogeneous two-agent environment, but it does not yet provide a systematic study of separate quadratic risk aversion for both market making and order execution.

This thesis addresses that gap by studying one risk-sensitive objective family in a controlled way. The market-making agent receives a quadratic inventory penalty $-\rho_{MM} Q_t^2$, and the execution agent receives a quadratic remaining-position penalty $-\rho_{EX} X_t^2$. By varying ρ_{MM} and ρ_{EX}, the thesis evaluates how risk aversion changes learned behaviour, risk-return trade-offs, and interaction dynamics. The evaluation goes beyond average reward by measuring final portfolio value, slippage, variance, inventory exposure, unfinished quantity, trading activity, execution timing, action distributions, and cross-play performance.

The thesis deliberately does not compare many different risk-sensitive objectives such as mean-variance, CVaR, or entropic utility. Those objectives are important in the broader risk-sensitive reinforcement-learning literature, but they introduce additional methodological complications, including different risk-aversion scales, episode-level distributional estimation, and more complex integration with PPO-style training. Focusing on quadratic risk aversion makes the experiment interpretable, additive over time, and feasible for a controlled bachelor-thesis study. The resulting contribution is therefore not a new MARL algorithm or a new market simulator, but a systematic empirical analysis of how quadratic risk aversion affects two interacting high-frequency trading agents.

CHAPTER 3

Theoretical background

Bibliography

[1] Aurélien Alfonsi, Antje Fruth, and Alexander Schied. Optimal execution strategies in limit order books with general shape functions. *Quantitative Finance*, 10(2):143–157, February 2010. _eprint: https://doi.org/10.1080/14697680802595700.

[2] Robert Almgren and Neil Chriss. Optimal execution of portfolio transactions. *The Journal of Risk*, 3(2):5–39, January 2001.

[3] Yakov Amihud and Haim Mendelson. Dealership market: Market-making with inventory. *Journal of Financial Economics*, 8(1):31–53, March 1980.

[4] Selim Amrouni, Aymeric Moulin, Jared Vann, Svitlana Vyetrenko, Tucker Balch, and Manuela Veloso. ABIDES-Gym: Gym Environments for Multi-Agent Discrete Event Simulation and Application to Financial Markets, October 2021.

[5] Leo Ardon, Nelson Vadori, Thomas Spooner, Mengda Xu, Jared Vann, and Sumitra Ganesh. Towards a fully RL-based Market Simulator. In *Proceedings of the Second ACM International Conference on AI in Finance*, pages 1–9, November 2021. arXiv:2110.06829 [cs].

[6] Marco Avellaneda and Sasha Stoikov. High-frequency trading in a limit order book. *Quantitative Finance*, 8(3):217–224, April 2008.

[7] Peter Belcak, Jan-Peter Calliess, and Stefan Zohren. Fast Agent-Based Simulation Framework with Applications to Reinforcement Learning and the Study of Trading Latency Effects, August 2020.

[8] Dimitris Bertsimas and Andrew W. Lo. Optimal control of execution costs. *Journal of Financial Markets*, 1(1):1–50, April 1998.

[9] David Byrd, Maria Hybinette, and Tucker Hybinette Balch. ABIDES: TOWARDS HIGH-FIDELITY MARKET SIMULATION FOR AI RESEARCH.

[10] Nicholas Tung Chan and Christian Shelton. An Electronic Market-Maker. April 2001. Accepted: 2004-10-20T20:50:09Z.

[11] Sascha Frey, Kang Li, Peer Nagy, Silvia Sapora, Chris Lu, Stefan Zohren, Jakob Foerster, and Anisoara Calinescu. JAX-LOB: A GPU-Accelerated limit order book simulator to unlock large scale reinforcement learning for trading, August 2023.

[12] Sumitra Ganesh, Nelson Vadori, Mengda Xu, Hua Zheng, Prashant Reddy, and Manuela Veloso. Reinforcement Learning for Market Making in a Multi-agent Dealer Market, November 2019. arXiv:1911.05892 [q-fin].

[13] Mark B. Garman. Market microstructure. *Journal of Financial Economics*, 3(3):257–275, June 1976.

[14] Bruno Gašperov, Stjepan Begušić, Petra Posedel Šimović, and Zvonko Kostanjčar. Reinforcement Learning Approaches to Optimal Market Making. *Mathematics*, 9(21):2689, January 2021.

[15] Hong Guo, Jianwu Lin, and Fanlin Huang. Market Making with Deep Reinforcement Learning from Limit Order Books, May 2023. arXiv:2305.15821 [q-fin].

[16] Olivier Guéant, Charles-Albert Lehalle, and Joaquin Fernandez-Tapia. Dealing with the inventory risk: a solution to the market making problem. *Mathematics and Financial Economics*, 7(4):477–507, September 2013.

[17] Dieter Hendricks and Diane Wilcox. A reinforcement learning extension to the Almgren-Chriss model for optimal trade execution. In *2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, pages 457–464, March 2014. arXiv:1403.2229 [q-fin].

[18] Thomas Ho and Hans R. Stoll. Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial Economics*, 9(1):47–73, March 1981.

[19] Michaël Karpe, Jin Fang, Zhongyao Ma, and Chen Wang. Multi-Agent Reinforcement Learning in a Realistic Limit Order Book Market Simulation. In *Proceedings of the First ACM International Conference on AI in Finance*, pages 1–7, October 2020. arXiv:2006.05574 [q-fin].

[20] Pankaj Kumar. Deep Reinforcement Learning for Market Making. *New Zealand*, 2020.

[21] Ye-Sheen Lim and Denise Gorse. Reinforcement Learning for High-Frequency Market Making.

[22] Siyu Lin and Peter A. Beling. An End-to-End Optimal Trade Execution Framework based on Proximal Policy Optimization. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 4548–4554, Yokohama, Japan, July 2020. International Joint Conferences on Artificial Intelligence Organization.

[23] Johann Lussange, Ivan Lazarevich, Sacha Bourgeois-Gironde, Stefano Palminteri, and Boris Gutkin. Modelling Stock Markets by Multi-agent Reinforcement Learning. *Computational Economics*, 57(1):113–147, January 2021.

[24] Chris Mascioli, Anri Gu, Yongzhao Wang, Mithun Chakraborty, and Michael Wellman. A Financial Market Simulation Environment for Trading Agents Using Deep Reinforcement Learning. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 117–125, Brooklyn NY USA, November 2024. ACM.

[25] Ciamac C. Moallemi and Muye Wang. A Reinforcement Learning Approach to Optimal Execution, November 2021.

[26] Valentin Mohl, Sascha Frey, Reuben Leyland, Kang Li, George Nigmatulin, Mihai Cucuringu, Stefan Zohren, Jakob Foerster, and Anisoara Calinescu. JaxMARL-HFT: GPU-Accelerated Large-Scale Multi-Agent Reinforcement Learning for High-Frequency Trading, November 2025. arXiv:2511.02136 [q-fin].

[27] Yuriy Nevmyvaka, Yi Feng, and Michael Kearns. Reinforcement learning for optimized trade execution. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*, pages 673–680, Pittsburgh, Pennsylvania, 2006. ACM Press.

[28] Brian Ning, Franco Ho Ting Lin, and Sebastian Jaimungal. Double Deep Q-Learning for Optimal Execution, June 2020. arXiv:1812.06600 [q-fin].

[29] Anna Obizhaeva and Jiang Wang. Optimal Trading Strategy and Supply/Demand Dynamics, June 2005.

[30] Alexander Rutherford, Benjamin Ellis, Matteo Gallici, Jonathan Cook, Andrei Lupu, Gardar Ingvarsson, Timon Willi, Ravi Hammond, Akbir Khan, Christian Schroeder de Witt, Alexandra Souly, Saptarashmi Bandyopadhyay, Mikayel Samvelyan, Minqi Jiang, Robert Tjarko Lange, Shimon Whiteson, Bruno Lacerda, Nick Hawes, Tim Rocktaschel, Chris Lu, and Jakob Nicolaus Foerster. JaxMARL: Multi-Agent RL Environments and Algorithms in JAX, November 2023.

[31] Thomas Spooner, John Fearnley, Rahul Savani, and Andreas Koukorinis. Market Making via Reinforcement Learning, April 2018. arXiv:1804.04216 [cs].

[32] Nelson Vadori, Leo Ardon, Sumitra Ganesh, Thomas Spooner, Selim Amrouni, Jared Vann, Mengda Xu, Zeyu Zheng, Tucker Balch, and Manuela Veloso. Towards Multi-Agent Reinforcement Learning driven Over-The-Counter Market Simulations, August 2023. arXiv:2210.07184 [cs].

[33] Oscar Fernández Vicente, Fernando Fernández Rebollo, and Francisco Javier García Polo. Deep Q-Learning Market Makers in a Multi-Agent Simulated Stock Market. In *Proceedings of the Second ACM International Conference on AI in Finance*, pages 1–9, November 2021. arXiv:2112.04494 [cs].

[34] Óscar Fernández Vicente, Fernando Fernández, and Javier García. Automated market maker inventory management with deep reinforcement learning. *Applied Intelligence*, 53(19):22249–22266, October 2023.

[35] Zhiyuan Yao, Zheng Li, Matthew Thomas, and Ionut Florescu. Reinforcement Learning in Agent-Based Market Simulation: Unveiling Realistic Stylized Facts and Behavior, March 2024. arXiv:2403.19781 [q-fin].