Dynamics of Market Making Algorithms in Dealer Markets

Wei Xiong

St Anne's College

University of Oxford

A thesis submitted for

Doctor of Philosophy

Supervisor: Prof. Rama Cont
Trinity 2024

Acknowledgements

Completing this doctoral thesis has been an extraordinary journey, one that has enriched me both intellectually and personally. I am profoundly grateful to the remarkable individuals and institutions who have supported me along the way and made this milestone possible.

First and foremost, I would like to express my deepest gratitude to my supervisor, **Professor Rama Cont**, for his insightful guidance and invaluable support in my work that shaped every page of this thesis. Your brilliance, intellectual rigor, and patience have been a constant source of inspiration. I am especially thankful for your willingness to engage deeply with my ideas, challenging me to ask questions critically and think creatively. I constantly feel that I have much more to learn from your profound wisdom and exceptional intuition in both mathematics and financial markets. This journey would not have been possible without your mentorship.

I owe a great deal to **Hanna Assayag** and **Dr. Alexander Barzykin** from HSBC, whose practical insights and industry expertise have been invaluable. This research grew out of our early conversations, and its development was greatly enriched by our regular and continuous exchanges. Your support has been important in bridging the theory and practice of market making in my work.

I am deeply grateful to my examiners, **Professor Christoph Reisinger** and **Professor Olivier Guéant**, for their thorough review of my work and their thoughtful feedback. I also thank **Professor Reisinger** for organizing and managing the examination, and **Professor Guéant** for traveling from Paris to Oxford for my viva. It was, in fact, **Professor Guéant**'s 2017 course on algorithmic trading in Paris that first sparked my interest in the topics explored in this thesis. Thank you—your insights not only refined the scope and depth of this thesis, but also strengthened my appreciation for academic excellence. Your dedication continues to inspire me.

I wish to honor **Professor Hans Wackernagel**, my mentor at Mines Paris - PSL, who supported me during my application for doctoral position at Oxford in 2019. Though he sadly passed away in January 2025, his legacy of academic achievements and kindness remains a guiding light in my memory.

I will always deeply miss him. I am also profoundly thankful to **Professor Huyên Pham** and **Professor Michel Kern** for their generous support and for writing the recommendation letters that opened doors for me. Your belief in my potential has always been a source of great encouragement.

To my colleagues and friends at the Mathematical Institute at Oxford, thank you for fostering a community where ideas flourish and prosper. Our stimulating discussions and shared moments of curiosity have made this journey intellectually rewarding and enjoyable. While I cannot name everyone individually, I sincerely appreciate your camaraderie and support. A special thanks to my friend, **Yichen Liu**, for our exploration of boutique stores and Chinese restaurants around Canary Wharf and the City of London. Those shared moments brought joy and relief to our days and provided much-needed relaxation during challenging times.

To my family: **my parents**, whose unwavering love and support have laid the foundation for every opportunity I have had; my dearest wife, **Ms. Liang**, for her patience, encouragement, and boundless love. Meeting her has been one of the most wonderful blessings of my life, and together we have built a loving family. I love you more than words can ever express. I am especially grateful for the arrival of our precious daughter, **Mingming (明明)**, in December 2024, who has brought immeasurable joy into our lives. Thank you for being my anchor and my inspiration.

Finally, I would like to acknowledge the financial support I received from the EPSRC Centre for Doctoral Training in Mathematics of Random Systems: Analysis, Modelling and Simulation (UKRI Grant EP/S023925/1). This work would not have been possible without the opportunity to be part of such a dynamic and forward-thinking program.

To all who have supported me on this journey, whether named here or not, I extend my heartfelt thanks.

雄关漫道真如铁，而今迈步从头越。

Statement of Originality

I declare hereby that this thesis is a product of my original research. It does not contain any work that has been previously submitted for any other degree, diploma, certificate, or other qualifications in any university or other tertiary institution. Where contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgment of collaborative research and discussions. This thesis is the product of my own work. Chapter 2, Chapter 3, and Chapter 4 were all carried out in collaboration with my supervisor Professor Rama Cont.

Chapter 2 is based on published work [Xiong and Cont 2021] to which I am the main contributor. Chapter 3 is based on published work [Cont and Xiong 2024] to which I am the main contributor. Chapter 4 is based on my work [Assayag et al. 2024] in collaboration with Hanna Assayag and Dr. Alexander Barzykin from HSBC, to which I am the main contributor. The results presented are part of the research work carried out within the HSBC FX Research Initiative. The views expressed are those of the authors and do not necessarily reflect the views or practices of HSBC.

Any work of other people is fully acknowledged in accordance with the standard referencing practices of the discipline.

Abstract

In the rapidly evolving landscape of electronic over-the-counter markets, the deployment of advanced market making algorithms has led to unprecedented efficiencies but also raised regulatory concerns, particularly regarding the potential for unforeseen effects, such as 'algorithmic collusion' or market instability, resulting from the impact of autonomous trading or market making algorithms. These concerns call for a better understanding of the impact of trading and market making algorithms on market dynamics. In this thesis, we propose a mathematical framework for studying the dynamics of algorithmic markets with a focus on the impact of competition, learning, and heterogeneity in the context of dealer markets. We model strategic interactions and competition among dealers using the framework of game theory—discrete time repeated games, stochastic differential games, and mean field games. The impact of *learning* by agents is introduced in this setting using the concept of reinforcement learning and implemented using decentralized deep reinforcement learning algorithms.

We initiate our investigation by constructing a game-theoretic model where multiple market makers compete for market share, adjusting their pricing spreads in response to evolving market conditions, learned autonomously through market data without direct communication. The learning dynamics are captured through a decentralized multi-agent reinforcement learning approach, revealing a propensity for these algorithms to independently converge to pricing strategies that, while not explicitly collusive, mirror the outcomes of tacit collusion by maintaining supra-competitive price levels.

We extend the analysis to a continuous-time setting, where the interactions of market makers are modelled as a stochastic differential game of intensity control under partial information. Competition among dealers corresponds to a Nash equilibrium, whereas

collusion is described in terms of Pareto optima. This analytical exploration is further enriched by employing the decentralized multi-agent deep reinforcement learning algorithm, which unveils the latent pathways through which learning by market making algorithms can inadvertently lead to tacit collusion, pushing spread levels significantly above competitive equilibrium levels.

The final chapter extends these results to a large population of dealers, whose interactions are modelled as a *mean field game* where a representative dealer interacts with the quotes of other dealers. The benchmark situation representing competition among dealers corresponds to a mean field Nash equilibrium, for which we give conditions for existence and uniqueness. We investigate the influence of learning dynamics in this setting using *mean field deep reinforcement learning*. We show that, in a homogeneous population of dealers, learning can lead to supra-competitive quoting strategies, while the introduction of heterogeneity mitigates this effect.

Our theoretical results and detailed numerical experiments provide interesting perspectives on market dynamics in the age of algorithmic trading and offer insights for market participants, risk managers, regulators, and policy makers on the impact on market behavior of autonomous algorithmic strategies in electronic over-the-counter markets. Market participants may consider autonomous learning algorithms used for market making to generate quoting strategies, but they should remain cautious about potential algorithmic risks that could potentially affect the competitiveness of the market. For risk managers, these market making algorithms are shown to manage inventory risk effectively, as they have learned to adjust spreads based on inventory positions. However, they must be aware of the regulatory risks associated with potential tacit collusion resulting from the interactions of the learning algorithms. Regulators and policy makers are suggested to revisit the existing rules for best execution, enforce mandatory audits on the algorithms, and implement market frameworks that ensure transparency in learning algorithms and encourage competition.

Contents

Funding information

Engineering and Physical Sciences Research Council (EPSRC),
Grant/Award Number: EP/S023925/1

List of Figures

List of Tables

Chapter 1

Algorithmic Market Making in Dealer Markets

1.1 Introduction

The integration of algorithmic trading and market making has revolutionized modern financial markets, particularly in electronic over-the-counter (OTC) markets. Market makers, by quoting both bid and ask prices, play an indispensable role in ensuring market fluidity. However, the rise of algorithmic strategies, while improving efficiency, has inadvertently introduced new complexities and potential risks. Among these is the phenomenon of 'algorithmic collusion', where market makers, without any explicit coordination, might set prices in a way that may generate adverse consequences for market participants.

This thesis explores the dynamics introduced by the interactions of algorithmic market making strategies in electronic OTC markets. Specifically, it investigates how these autonomous algorithms, through their learning dynamics, could potentially lead to market outcomes reminiscent of tacit collusion—a scenario where competitive pricing is undermined without overt coordination among market makers. Such a situation poses significant regulatory challenges, as it blurs the lines between competitive strategy and anti-competitive behavior.

To address these concerns, our research is initiated around a game-theoretic framework complemented by deep reinforcement learning algorithms. In Chapter 2, we construct a model in which multiple market makers compete for market share while autonomously adjusting their pricing strategies based on evolving market conditions. By employing decentralized multi-agent reinforce-

ment learning, we simulate how these market makers learn and adapt to supra-competitive outcomes without direct communication with each other.

In Chapter 3, we analyze the market making process in a dealer market environment through a stochastic differential game of intensity control, under conditions of partial information. Here, we examine how competition among dealers can evolve into collusive behavior purely through algorithmic learning, without any explicit agreement among the participants. Subsequently, in Chapter 4, we extend our analysis to a large population of dealers whose interactions are modelled as a mean field game ([Lasry and Lions 2007; Huang, Caines, and Malhamé 2007]). This allows us to observe the collective behavior of a large number of market makers and examine how their interactions influence overall market dynamics, particularly focusing on the emergence of supra-competitive pricing strategies resulted from the learning algorithms.

By studying these complex interactions and learning processes, this thesis aims to provide deeper insights into the unintended consequences of algorithmic market making. The findings not only contribute to our understanding of market dynamics in the era of algorithmic trading, but also offer valuable perspectives for regulators and market participants to maintain fair and competitive market conditions.

In the following, we will interchangeably use the terms 'market maker', 'dealer', and 'agent'.

1.2 Literature Review

Our work draws on a large body of research literature on market microstructure, game theory, and learning algorithms.

In an insightful study, [Calvano et al. 2020] have shown, in the setting of repeated auctions in a goods market, that competition among pricing algorithms with learning may lead to prices set at levels different from competitive benchmarks, without any explicit exchange of information between market makers, a situation known as 'tacit collusion'.

Tacit collusion is a well-known issue in oligopolistic markets ([Ivaldi et al. 2003; Tirole 1988]), and may emerge in repeated auctions without explicit information sharing ([Han 2021; Skrzypacz and Hopenhayn 2004]). The possibility that tacit collusion may emerge from the interactions of autonomous algorithms learning from market data in a decentralized fashion, a situation sometimes referred to as 'algorithmic collusion' ([Assad et al. 2021; Han 2021]),

has attracted the concerns of market regulators ([Competition & Markets Authority 2021]).

The emergence of tacit collusion from learning has also been studied in various contexts. [Waltman and Kaymak 2008] show how competing producers using a Q-learning algorithm learn to raise prices above the Nash equilibrium price by reducing production in a Cournot competition model. [Calvano et al. 2020] show that Q-learning in a Bertrand competition model may result in prices strictly above competitive levels associated with Nash equilibrium. [Abada and Lambin 2023] apply multi-agent Q-learning to Cournot competition in electricity markets, and conclude that the collusion may result from imperfect exploration. [Asker, Fershtman, and Pakes 2022] compare the pricing outcomes of asynchronous vs. synchronous learning algorithms. [Hettich 2021] shows that deep Q-learning algorithms (DQN) lead to collusive strategies significantly faster. [Han 2021] investigates the effects of experience replay in the learning algorithm, which leads to prices closer to the equilibrium level. This literature focuses primarily on a one-sided goods market where 'producers' fix prices, and may not directly be applicable to a financial market with two-sided order flow.

More recently, [Cartea, Chang, and Penalva 2022] study the impact of discreteness of 'tick size' on algorithmic collusion in market making games. The authors apply stochastic approximation methods to characterize the learning algorithms with a system of ordinary differential equations (ODEs), and show convergence of learning algorithms to pure Nash equilibrium strategy in a 2-player bimatrix market making game. Stochastic approximation techniques are applied to show evidence of tacit collusion that arises in multi-agent setting, where a finer tick size mitigates tacit collusion and promotes competition. [Cartea, Chang, Mroczka, et al. 2022] study tacit collusion in a model with competing liquidity providers whose behavior is modelled using multi-armed bandit algorithms. The study finds that while collusive pricing by algorithms could theoretically occur in multi-agent scenarios, it is highly unlikely without explicit coordination. Compared to [Cartea, Chang, and Penalva 2022; Cartea, Chang, Mroczka, et al. 2022], the thesis demonstrates several key differences in both the model setup, experimental design, and the conclusions. In terms of model setup, the thesis covers a broad range of models specifically tailored for multi-agent market making, including game theory, stochastic differential games, and mean field games. While [Cartea, Chang, and Penalva 2022; Cartea, Chang, Mroczka, et al. 2022] also study game-theoretic models,

the key difference lies in the type of equilibria studied. [Cartea, Chang, and Penalva 2022; Cartea, Chang, Mroczka, et al. 2022] focus primarily on static equilibria, applying multi-armed bandit algorithms and Q-learning, with emphasis on the asymptotic properties of these algorithms through stochastic approximation to ODEs, whereas this thesis investigates the outcomes resulting from multi-agent learning dynamics, which more closely resemble actual market conditions. Both [Cartea, Chang, and Penalva 2022; Cartea, Chang, Mroczka, et al. 2022] and the thesis find evidence of tacit collusion emerging from the interactions of learning algorithms. However, the conclusions are somehow different. [Cartea, Chang, and Penalva 2022] conclude that smaller tick sizes facilitate convergence towards competitive equilibrium levels, whereas Chapter 2 of the thesis finds that tacit collusion can occur even with small tick sizes, while competitive quoting emerges only when tick sizes are unrealistically large. We esteem the different outcomes mainly due to the underlying model structures. [Cartea, Chang, and Penalva 2022] incorporate fill probabilities as functions of quotes, whereas Chapter 2 relies on a 'winner-takes-all' mechanism in which only the market maker quoting the lowest spread can fill the request for quote (RFQ). Furthermore, a key conclusion from [Cartea, Chang, Mroczka, et al. 2022] is that the collusive outcomes resulting from interacting algorithms are unlikely to manifest in real-world market. In contrast, the thesis presents evidence of tacit collusion that persists across multiple model settings, emphasizing the regulatory implications of such outcomes. Overall, the thesis contributes to the growing body of literature on the mechanism and potential for algorithmic collusion in over-the-counter (OTC) markets, offering a perspective on the conditions under which algorithmic collusion may arise.

The models applied in the literature on algorithmic collusion are diverse. More recent literature has also explored algorithmic collusion using classical market microstructure models. [Colliard, Foucault, and Lovo 2022] investigate how algorithmic market makers use Q-learning algorithms to set prices compared to a Nash equilibrium in the Glosten-Milgrom model ([Glosten and Milgrom 1985]) with adverse selection, and find that algorithmic market makers effectively learn to adapt to adverse selection but lead to quoted spreads above competitive levels predicted by Nash equilibrium, particularly when the adverse selection costs decrease. They attribute this deviation to limitations in learning capacity under stochastic environments. [Dou, Goldstein, and Ji 2023] investigate the reinforcement learning algorithm applied by informed traders in the framework of Kyle's model ([Kyle 1985]) consisting of informed traders,

noise traders and market makers. Their findings suggest that AI-driven trading can autonomously form collusive behaviors which lead to reduced liquidity and decreased price informativeness. Thus, it is of interest to investigate whether the tacit collusion exhibited by [Calvano et al. 2020] may also arise in the context of financial markets with market makers competing for a two-sided (buy and sell) order flow. Chapter 2 examines the conditions under which tacit collusion can arise in a two-sided discrete-time dealer market with multiple market makers.

Chapter 3 revisits these issues in more detail in the framework of a market with continuous-time trading with competing market makers who learn from market data, extending the results of [Calvano et al. 2020] and [Xiong and Cont 2021] to a continuous-time setting which better captures some important features of intraday trading in financial markets. In the continuous-time model, the reference asset price is modelled as a Brownian motion. Market makers continuously set ask and bid prices, and the arrival of requests for quotes is modelled using point processes, reflecting the dynamic and stochastic nature of financial markets. In this framework, the objective of each market maker is to maximize cumulative profit and loss over a time interval, rather than focus on profit from a single trade as in Chapter 2. Our modelling framework builds upon the recent literature on continuous-time models for optimal market making in dealer markets: following the pioneering work of [Ho and Stoll 1980; Ho and Stoll 1983] and [Avellaneda and Stoikov 2008], the problem of optimal market making has been formulated as a stochastic control problem where market makers quote ask/bid prices dynamically to maximize their expected profit adjusted for inventory risk over a finite or infinite time horizon ([Avellaneda and Stoikov 2008; Barzykin, Bergault, and Guéant 2023; Bergault and Guéant 2021; Cartea, Jaimungal, and Ricci 2014; Guéant 2017; Guéant and Manziuk 2019]). In contrast to earlier literature in which market makers are assumed to know the market dynamics, recent literature has explored the more realistic case where market makers learn through trial and error, using reinforcement learning ([Vadori et al. 2024; Ardon et al. 2021]). In most of these models, the market maker faces a random environment represented by an order flow represented as a point process, so a natural mathematical modelling framework for the problem is that of *intensity control* of point processes ([Bremaud 1981]). The competition of market makers can be modelled in this setting as a stochastic differential game [Guo and Xu 2019; Cont, Guo, and Xu 2021; Luo and Zheng 2021]. [Cont, Guo, and Xu 2021] model inter-bank

lending as a stochastic differential game of singular control and study Pareto optimal strategies. Competition among market makers is studied by [Luo and Zheng 2021].

[Vadori et al. 2024] and [Ardon et al. 2021] build a multi-agent dealer market simulator using reinforcement learning agents and show that reinforcement learning agents replicate some 'stylized facts' of dealer markets.

Multi-agent modelling in stochastic games is studied in recent works by [Cont, Guo, and Xu 2021] and [Guo and Xu 2019]. [Cont, Guo, and Xu 2021] model inter-bank lending activity as a stochastic differential game through singular control and solve the Pareto optimal strategies using a singular stochastic control problem. [Guo and Xu 2019] consider N-player stochastic game of the classical fuel follower problem, in which they explicitly formulate Nash equilibrium strategy. They also study Nash equilibrium strategies of the mean field game extension. The single agent singular stochastic control problem considered in [Cont, Guo, and Xu 2021; Guo and Xu 2019] is introduced and analyzed in [Karatzas 1983] and [Menaldi and Taksar 1989]. The competition and dynamic equilibrium between the market makers are studied by [Luo and Zheng 2021], which considers the Nash equilibrium when the market makers compete for market order flow.

Q-learning is a reinforcement learning algorithm proposed by [Watkins 1989] to handle the Markov decision process by training a state-action value function. Recently, the successful integration of neural networks into the RL algorithms has surpassed human-level performance in several practical learning tasks, such as playing Atari games ([Mnih, Kavukcuoglu, Silver, Graves, et al. 2013; Mnih, Kavukcuoglu, Silver, Rusu, et al. 2015]). We model the learning process of market makers through multi-agent reinforcement learning in Chapter 3. Since in dealer markets, market makers cannot observe competitors' quotes or order flow but only their own prices and order flow plus the market spread, the relevant approach is to apply a decentralized multi-agent learning algorithm.

[Hu and Wellman 1998] prove that under some particular conditions multi-agent Q-learning algorithms converge to Nash equilibrium whenever it is unique. However, our decentralized multi-agent setting where agents do not know each others' states and actions does not satisfy the convergence conditions in [Hu and Wellman 1998]. [Foerster et al. 2018] develop a multi-agent actor-critic method using decentralized actors and a single centralized critic. [Lowe et al. 2017] extend this methodology to the Multi-Agent Deep Deterministic Policy

Gradient (MADDPG) method, where centralized critics are trained for each specific agent.

Mean field games (MFGs), introduced by [Lasry and Lions 2007; Huang, Caines, and Malhamé 2007], provide a more tractable framework for studying strategic interactions in large homogeneous populations. In an MFG, each agent's strategy is influenced by the aggregate effect of the actions of other agents, represented through an aggregate variable ('mean field'). The original works by [Lasry and Lions 2007] study the conditions under which there exist solutions to mean field game problems in continuous time and state space, and have proposed monotonicity conditions for uniqueness. Since then, the existence and uniqueness of solutions to mean field games have been extensively studied in a wide range of settings. For example, the case of discrete state space has been studied by [Gomes, Mohr, and Souza 2013], [Guéant 2015], and [Doncel, Gast, and Gaujal 2019]. Potential mean field game systems are addressed by [Briani and Cardaliaguet 2018]. MFGs with common noise, which can be analyzed with the introduction of the master equation, have been studied by [Carmona, Delarue, and Lacker 2020] and [Cardaliaguet, Delarue, et al. 2019]. Additionally, a significant part of the literature focuses on the social optimum problems in mean field systems, as studied by works such as [Bensoussan, Frehse, and Yam 2013] and [Carmona and Delarue 2013].

In Chapter 4, we extend the study from a finite number of market makers to a scenario with a large population of market makers in a mean field game framework. We use here the framework of extended mean field games with finite state space introduced in [Gomes, Mohr, and Souza 2013] and [Guéant 2015], which are the closest to our problem setting with market makers. [Gomes, Mohr, and Souza 2013] and [Guéant 2015] introduce the mean field games defined on finite state space in continuous time, where the players control the transition matrix, and show the existence of solution to the corresponding mean field game systems. [Gomes, Mohr, and Souza 2013] provide monotonicity conditions to prove the uniqueness, while [Guéant 2015] propose a general uniqueness criterion on the comparison principle. However, in our framework, the representative market maker controls the ask and bid quotes instead of directly monitoring the transition matrix. We propose a fixed-point argument similar to [Gomes, Mohr, and Souza 2013], adapted to our model setting for the existence of a solution to the system of mean field equations. Furthermore, we introduce an algebraic condition inspired by the approach in [Guéant 2015] for the uniqueness of the mean field Nash equilibrium.

Besides the literature contributing to the theoretical aspects, mean field games have seen various applications in financial markets. [Cardaliaguet and Lehalle 2018] formulate the problem of optimal liquidation as a mean field game where market participants' execution jointly affects market impact. [Huang, Jaimungal, and Nourian 2019] considers the optimal execution problem for a major player facing a population of minor agents. [Casgrain and Jaimungal 2020] incorporate agents' heterogeneity as different probability measures in a mean field game of optimal execution problem. [Neuman and Voß 2023] study a multi-agent optimal execution problem and prove its convergence to a mean field game in the large population limit. [Baldacci, Bergault, and Possamaï 2023] consider a major-minor player mean field game where market makers interact with a population of market takers. [Bernasconi-de-Luca et al. 2023] study a mean field game model of market making in dealer markets and study a reinforcement learning algorithm capable of learning the equilibrium. Although our model shares some features of these studies, rather than using deep reinforcement learning as a computational tool to learn (Nash) equilibrium, we focus on the market dynamics resulting from the combination of learning, heterogeneity, and strategic interactions. Indeed, as we will see, the combination of learning and strategic interactions may lead to configurations different from the Nash equilibrium.

1.3 Contributions

In this thesis, we investigate the dynamics induced by competition and learning in dealer markets, using an approach based on multi-player games with (reinforcement) learning. Our analytical results and extensive numerical experiments using state-of-the-art deep reinforcement learning algorithms provide novel insights into the dynamics of the price and the order flow in such markets resulting from the interactions of adaptive market making algorithms used by dealers.

In Chapter 2 we propose a game-theoretic model of a financial market, namely multi-dealer-to-client (MD2C) platforms, in which multiple market makers compete for market share and learn from market data to adjust their spreads under a 'winner-takes-all' market share allocation mechanism. The learning process is modelled via a decentralized Multi-Agent Deep Deterministic Policy Gradient (decentralized MADDPG) algorithm in which market makers cannot observe their competitors' prices.

We show that even in the absence of price information sharing, market prices may converge to levels that are similar to a collusion situation, resulting in 'tacit collusion'. The emergence of tacit collusion depends on the specific mechanism of MD2C platforms through which market makers compete for order flow. The phenomenon of 'tacit collusion' is also linked with the level of competition, mainly reflected by the number of market makers. Inspired by these findings, we discuss some implications for market regulators in dealing with possible tacit collusion induced by the use of learning algorithms by market makers, focusing on the aspect including best execution rules, regulations that encourage competition at adequate level in MD2C platforms, and transparency on market making algorithms.

In Chapter 3, we model the interactions of market makers in a dealer market as a stochastic differential game of intensity control with partial information and study the resulting dynamics of bid-ask spreads resulting from competition among market makers and their learning dynamics.

We first study two benchmark cases: a competitive market, modelled as a Nash equilibrium of the game, and collusion among dealers, modelled as a Pareto optimum of the game. We give conditions for the existence of a Nash equilibrium, which we characterize in terms of a system of coupled Hamilton-Jacobi equations, and exhibit an algorithm based on fictitious play for computing Nash equilibria.

These benchmark cases correspond to hypothetical situations where the dynamics of order flow is known to market makers. In practice, market makers interact with the order flow of the client and learn to adjust their quotes to maximize their profits. We model this learning process by a decentralized multi-agent deep reinforcement learning algorithm ([Hambly, Xu, and Yang 2023]) using a policy gradient method ([Fazel et al. 2018]) to update market makers' strategies, parameterized via neural networks. Our simulation results show that the interactions of market making algorithms through market prices, without any sharing of information, may give rise to tacit collusion, as evidenced by quoted spread levels significantly higher than competitive (Nash) equilibrium.

This emergence of 'tacit collusion' through learning has interesting implications for market design and market regulation. We highlight the impact of correlated behaviors generated from learning even though learning algorithms are refrained from communication, which can lead to the phenomenon of 'tacit

collusion'. Our model, coupled with the use of multi-agent deep reinforcement learning, provides a conceptual framework for studying 'tacit collusion' showing that the latter is a useful tool for exploring these issues.

In Chapter 4, we model competition in dealer markets with a large number of dealers using a Mean Field Game framework, and extend the scope of multi-agent deep reinforcement learning algorithms to investigate learning dynamics in this setting.

We first prove the existence of mean field Nash equilibrium by characterizing it as a solution to the system of coupled Hamilton-Jacobi (HJ) equations and Chapman-Kolmogorov (CK) equations, and give a sufficient condition for the uniqueness of such an equilibrium. We compute the mean field Nash equilibrium using a numerical solution of this PDE system and use it as a benchmark for modelling competition among market makers.

We then investigate the influence of *learning dynamics* in this setting using a novel approach based on *mean field deep reinforcement learning* (MFDRL). This approach allows to model the interactions between a market maker who learns to set ask and bid prices using a deep reinforcement learning algorithm while interacting with the market environment ("mean field") generated by joint actions of other market participants. Our numerical results show that the learning dynamics lead to higher quotes by the market maker when their inventory is at the adverse direction of market order flow, while at inventory levels corresponding to market order flow, the market maker tends to quote more aggressively, with price levels slightly below competitive (Nash) equilibrium levels. This altogether leads to a wider bid-ask spread compared to the mean field Nash equilibrium benchmark, a signature of 'tacit algorithmic collusion' as evidenced by [Cont and Xiong 2024]. We subsequently model the heterogeneity with a learning agent interacting with a mean field equilibrium system, where 'tacit collusion' can be mitigated as the agent learns a quoting strategy closer to mean field equilibrium.

We conclude this chapter by comparing the learning algorithms and human traders in order to demonstrate that the tacitly collusive outcomes observed in our simulations are primarily the result of the interactions of these algorithms. Compared to human traders, deep reinforcement learning algorithms often represent different patterns of decision making processes. These learning algorithms do not simply mimic human behaviors, nor do they follow predefined rule-based strategies. Instead, they learn the quoting strategies by interacting with the market environment through trial and error, with the objective of

maximizing the cumulative reward. The learning process is data-driven and self-interested, without human intervention. These algorithms are unaffected by human emotions or logical thinking that often lead to human traders' rational reasoning-based decision making, especially when competition intensifies or regulatory scrutiny tightens. The different patterns of quoting behavior can lead to market dynamics resulting from interactions of the learning algorithms that differ significantly from those that may arise in markets consisting solely of human traders.

This difference is consistently demonstrated across the models presented in the thesis. Under the settings of these models, rational human traders would not have an incentive to deviate from the associated competitive Nash equilibrium, as doing so often reduces profitability. For example, in the game-theoretic model introduced in Chapter 2, the 'winner-takes-all' mechanism results in an intensely competitive environment where human market makers are keen to sustain the Competitive Nash Equilibrium level. However, DRL algorithms systematically learn to maintain wider spreads above the competitive equilibrium level without any explicit coordination. This is fundamentally different from the behavior of rational human traders, who often understand the competitive norms with the knowledge of their own model parameters.

It is also essential to note that all the learning agents operate in a decentralized manner without explicit communication or coordination among the algorithms, with access only to their own private information. Although collusion between human traders is often characterized by direct communication or explicit agreements, the tacit collusion observed in our simulations arises from decentralized learning. This form of tacit collusion is more challenging to detect and monitor by regulators.

Therefore, we consider the observed tacit collusion in our simulation as a consequence of employing DRL algorithms, highlighting the distinct market dynamics resulting from these algorithms. These outcomes differ from human decision making patterns, creating new regulatory challenges in market making using learning algorithms.

In summary, the thesis provides a theoretical and practical framework for modelling market making in over-the-counter (OTC) markets and sheds light on the role of competition, heterogeneity, and learning in the emergence of tacit collusion in such markets.

Chapter 2

Interactions of Market Making Algorithms: a Study on Perceived Collusion with Deep Reinforcement Learning

This chapter is based on [Xiong and Cont 2021]. In this chapter, we propose a game-theoretic model for a multi-dealer-to-client (MD2C) platform in financial markets in which multiple market makers compete for market share and learn from market data to adjust their spreads. We model this learning process through a decentralized multi-agent reinforcement learning algorithm and show that, even in the absence of price information sharing, under a specific mechanism through which market makers compete for market shares, market prices may converge to levels which are similar to a collusion situation, resulting in 'tacit collusion'. We also briefly discuss the implications of our research for market regulators.

2.1 Modelling the Actions of Market Makers

2.1.1 Market Order Flow and Spread

Let us first consider a market with a single asset and one market maker. The market maker quotes a bid price b at which it is willing to buy the asset, together with an ask price $a > b$ at which it sells the asset. The difference $a - b$ between the ask and bid price is called the 'market spread'.

We introduce market demand/supply order flow functions that represent the average volume of assets traded on the market per unit time interval.

Demand $D(a)$ is a decreasing function of the ask price a measuring the number of assets that investors buy from the market maker per unit period of time. The supply $S(b)$ is an increasing function of the bid price b measuring the number of assets that investors sell to the market maker per unit period of time. In our model, we consider investors as an exogenous factor. As a result, investors modify their demand and supply solely as functions of the ask and bid prices. The number of market makers will not affect the total order flow in the dealer market. Market makers indirectly modulate the order flow through the prices they quote.

The intersection of the demand and supply functions determines the market price of the asset, denoted by v.

$$D(v) = S(v) \tag{2.1.1}$$

Remark 2.1.1. We assume that the market price v is known to market makers, so that they quote ask/bid prices based on the spread. This is an essential difference from the assumption of informed traders in a microstructure model of the security market such as [Glosten and Milgrom 1985].

Several assumptions are imposed on the market demand and supply functions. Due to the monotonicity of demand and supply, the market order flow tightens if the market maker sets a higher spread. As in [Dutta and Madhavan 1997], we first make a symmetry assumption on demand and supply functions to simplify the notation. Moreover, the market maker quotes the spread s instead of quoting the ask and bid prices.

Assumption 2.1.2. Demand and supply functions are symmetric at the market price v.

$$D(v + x) = S(v - x), \forall x \in \mathbb{R} \tag{2.1.2}$$

Meanwhile, the bid and ask prices are symmetric on both sides of the market price v: $a - v = v - b = \frac{s}{2}$.

Generally, market makers would optimize over spread values for maximal revenues; hence we make the second assumption on the analytical property of demand and supply functions so that selling and buying revenues yield unique optima.

Assumption 2.1.3. $D(.)$ and $S(.)$ are C^2 functions and satisfy that, for all $x \geq 0$

- $D(x) \geq 0, \quad S(x) \geq 0$, and there exists $x', x'' \geq 0$ such that $D(x') > 0, \quad S(x'') > 0$
- $D'(x) < 0, \quad S'(x) > 0$
- $\lim_{x\to\infty} D(x) = \lim_{x\to-\infty} S(x) = 0$
- $\sup_{x\geq 0} \frac{D(x)D''(x)}{(D'(x))^2} < 2, \quad \sup_{x\geq 0} \frac{S(x)S''(x)}{(S'(x))^2} < 2$

The market making model developed by [Dutta and Madhavan 1997] also incorporates a stochastic market shock factor in demand/supply functions. However, we assume deterministic models because our primary focus is perceived collusion introduced by pricing algorithms. Also, [Dutta and Madhavan 1997] consider that the market price v is the only competitive price level, and any deviation of ask/bid prices from v is considered as overt or tacit collusion, which cannot explain the ubiquitous existence of nonzero spread in the actual market. To account for the nonzero spread at the competitive price level, we introduce endogenous cost c for the market maker.

Assumption 2.1.4. Market maker has a fixed cost c. The market maker chooses the ask a and bid b ($b < v < a$) to at least cover this fixed cost c.

$$(a - v)D(a) + (v - b)S(b) \geq c \tag{2.1.3}$$

Remark 2.1.5. Note that this cost c is rather a general concept. It can be incorporated with many aspects of market makers' costs, internal or external, such as order processing cost and inventory cost. In some circumstances, market makers also set a targeted revenue level to cover. In that case, the targeted revenue is also included in the cost c.

Remark 2.1.6. The cost c can also be considered as an entry barrier that excludes market makers with too high a cost. Since total market order flow is determined exogenously by demand and supply, the maximum revenue is bounded. Only those market makers with a cost c not higher than the total market revenue will have an incentive to trade. In contrast, if $\forall a, b > 0, (a - v)D(a) + (v - b)S(b) < c$, the maker of the market will not trade, as its profit will be negative.

With symmetry assumption 2.1.2, we can write the revenue of the market maker as a function of the spread s:

$$F(s) = (a - v)D(a) + (v - b)S(b) = s \cdot D(v + \frac{s}{2}) \tag{2.1.4}$$

Proposition 2.1.7. *If Assumption 2.1.2 and Assumption 2.1.3 are satisfied, then the revenue function $F(s)$ can be written as $F(s) = sD(v + \frac{s}{2})$, and has a unique maximum s^* in $\mathbb{R}^+$*

Proof. From Assumption 2.1.2 which leads to (2.1.4), the revenue function $F(s)$ can be written as $F(s) = sD(v+\frac{s}{2})$. From Assumption 2.1.3 that $D'(x) < 0$, we have that $F'(s) = D(v + \frac{s}{2}) + \frac{s}{2}D'(v + \frac{s}{2}) = 0$ if and only if

$$H(s) = \frac{s}{2} + \frac{D(v + \frac{s}{2})}{D'(v + \frac{s}{2})} = 0$$

Differentiating $H(s)$ we obtain

$$H'(s) = \frac{1}{2}\left(2 - \frac{D(v + \frac{s}{2})D''(v + \frac{s}{2})}{(D'(v + \frac{s}{2}))^2}\right)$$

Denote $C = \inf_{s\geq 0} H'(s)$. From Assumption 2.1.3, $H'(s) \geq C > 0, \forall s \geq 0$, since

$$\sup_{x\geq 0} \frac{D(x)D''(x)}{\big(D'(x)\big)^2} < 2$$

Therefore, $H(s)$ increases strictly in $\mathbb{R}^+$, and we have $\lim_{x\to+\infty} H(s) = +\infty$. In fact, by the mean value theorem, for any $s \geq 0$, there exists $\xi_s \in [0, s]$ such that

$$H(s) = H(0) + H'(\xi_s)s \geq H(0) + Cs$$

the right-hand side of the above inequality tends to $+\infty$ as $x \to +\infty$.

Therefore, since $\forall s \geq 0, H(0) < 0$ and $H'(s) > 0$, and $\lim_{x\to+\infty} H(s) = +\infty$, $H(s) = 0$ yields a unique solution in $\mathbb{R}^+$. Hence $F(s)$ has a unique maximum on $\mathbb{R}^+$, denoted by s^*. □

The unique maximum s^* in Proposition 2.1.7 is hence the spread with which a monopolistic market maker gains the maximal revenue. From now on, the market maker quotes the spread s instead of ask a and bid b.

In practice, demand and supply functions can be calibrated using market makers' data on order flow and trades. We will use the following example to illustrate the properties of the model:

$$D(a) = \frac{10}{1 + e^{a-100}} \quad S(b) = \frac{10}{1 + e^{100-b}} \qquad \text{(Example - Order flow)}$$

The market price is then $v = 100$. Consequently the revenue function $F(s)$ takes the form:

$$F(s) = s \cdot D(100 + \frac{s}{2}) = \frac{10s}{1 + e^{\frac{s}{2}}} \qquad \text{(Example - Revenue)}$$

The spread maximizing revenue is $s^* \approx 2.556$. Figure 2.1 shows the shapes of the order flow functions and the revenue function. We will use this example below.

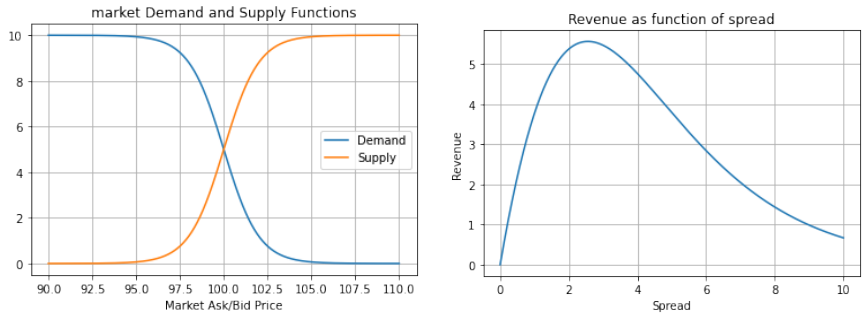


Figure 2.1: An example of demand/supply and revenue functions.

2.1.2 Extension to Multiple Market Makers

We now extend the market making framework in Section 2.1.1 to electronic OTC markets with multiple market makers, namely multi-dealer-to-client (MD-2C) platforms. Over the past decades, multi-dealer-to-client (MD2C) platforms have become increasingly important in the electronification of dealer markets, providing a centralized venue where multiple market makers compete to provide liquidity to institutional clients. The assets traded on MD2C platforms are usually bonds, swaps, and non-deliverable forwards. Typical examples of MD2C platforms include Bloomberg Fixed Income Trading (FIT), Tradeweb, and MarketAxess. These platforms allow clients to send Request for Quotes (RFQs) simultaneously to multiple market makers, thus encouraging competition and improving market efficiency ([Fermanian, Guéant, and Pu 2016]). The client sends an RFQ through the MD2C platform to the selected dealers, and the requested dealers answer the RFQ by quoting ask and bid prices. Dealers on the MD2C platform can observe the number of dealers requested and the composite market spread, for example, the CBBT bid-ask spread on the Bloomberg FIT platform.[1] However, they do not know the prices quoted by other dealers. The client can compare these prices and trade with the dealer that provides the most favorable price.

We also make the assumption that clients always send RFQs to a given number of market makers denoted by N, and that the size of each RFQ is

[1]On some MD2C platforms, the number of requested dealers is not disclosed to the dealers by default, e.g., corporate bonds on MarketAxess ([Wang 2023]). In our framework, we always assume that the market makers know the number of requested dealers.

fixed. The demand/supply order flow functions $D(.)$ and $S(.)$ can therefore represent the flow of RFQs per unit time. Another important assumption in our framework is that the spread quoted by market makers has a tick size $\vartheta > 0$. Hence, the spreads quoted by the market makers are taken from the discrete action space $\mathcal{A}_\vartheta = \{n\vartheta | n \in \mathbb{N}\}$ where $\mathbb{N}$ is the set of natural numbers. We need to assume that the tick size ϑ is small enough for the analysis of the Nash equilibrium, which will be specified in the subsequent section.

From now on, we will use the Greek letter α to represent elements of the discrete space $\mathcal{A}_\vartheta$ to distinguish from the letter s that represents a real number as in the previous section.

Suppose now that the clients send RFQs to N market makers. The cost of the market maker indexed by i is denoted by c_i. Let $\vec{\alpha} = (\alpha_1, ..., \alpha_N) \in (\mathcal{A}_\vartheta)^N$ denote the spreads quoted by the N market makers. The narrowest spread $\alpha^M := \min_{j \in \{1,...N\}} \{\alpha_j\}$ is called the market spread. We assume that market demand and supply take the same format as in the case of a single market maker. The order flow is directed to N market makers by a certain market share mechanism.

Now we specify the routing mechanism of how market order flows are distributed to N market makers. For the market maker i, we introduce a non-negative market share coefficient $\phi_i(\vec{\alpha})$, a function of the joint spread vector $\vec{\alpha}$ of the N market makers. $\phi_i(\vec{\alpha})$ measures proportion of order flow directed to the market maker i. It is an indicator of competitiveness with respect to the quote and endogenous characteristics of the market maker.

Assumption 2.1.8. When there are N market makers on the market, the demand and supply allocated to market maker i are determined by total order flow and market share coefficients:

$$\begin{aligned} D_i(v + \frac{\alpha_i}{2}) &= \frac{\phi_i(\vec{\alpha})}{\sum_{j=1}^{N} \phi_j(\vec{\alpha})} D(v + \frac{\alpha^M}{2}) \\ S_i(v - \frac{\alpha^i}{2}) &= \frac{\phi_i(\vec{\alpha})}{\sum_{j=1}^{N} \phi_j(\vec{\alpha})} S(v - \frac{\alpha^M}{2}) \end{aligned} \tag{2.1.5}$$

where $\alpha^M := \min_j \{\alpha_j\}$ is the market spread.

Remark 2.1.9. Assumption 2.1.8 specifies that the demand order flow of a market maker i is determined by two components: the relative market share

$\frac{\phi_i(\vec{\alpha})}{\sum_{j=1}^{N}\phi_j(\vec{\alpha})}$ and the aggregate order flow $D(v+\frac{\alpha^M}{2})$ as a function of the market spread α^M. Now, with multiple market makers, it is the market spread α^M that affects directly the aggregate order flow, which corresponds to the practical scenarios.

It should be noted that (2.1.5) generally does not guarantee the existence of a Nash equilibrium, up to the format of the market share coefficient $\phi_i(\cdot)$. Taking the example of a hypothetical case where $\phi_i(\vec{\alpha}) \equiv$ Constant, with 2 market makers, the Nash equilibrium would not exist since the revenue of the market maker with the highest spread would be $\frac{\max\{\alpha_1,\alpha_2\}}{2}D(v+\frac{\min\{\alpha_1,\alpha_2\}}{2})$, which can increase to infinity when $\max\{\alpha_1,\alpha_2\} \to \infty$. However, we will study the realistic market share mechanism that corresponds to what happens in MD2C platforms, where the market maker quoting the lowest spread wins the order flow. This market share mechanism leads to a well-defined Nash equilibrium, as we will see later.

In accordance with practical scenarios in MD2C platforms, we define a 'winner-takes-all' (denoted by 'WTA') market share allocation mechanism as our benchmark model for competition.

$$\phi_i(\vec{\alpha}) = \kappa_i \mathbb{I}(\alpha_i = \min_j\{\alpha_j\}), \kappa_i \text{ is constant} \tag{WTA}$$

(WTA) specifies that only those quoting the narrowest spread jointly share the market order flow, which corresponds to the case of MD2C platforms where the market maker providing the narrowest spread wins the RFQ.

We are primarily interested in the market dynamics resulting from the learning algorithms under mechanism **(WTA)**, which will be studied in Section 2.2

We then define the market maker's profit as the objective function. Rewrite $\vec{\alpha} = (\alpha_i, \alpha_{-i})$, where $\alpha_{-i} = (\alpha_1, \ldots, \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_N)$ represents the spreads of other market makers. The market maker i sets the spread α_i by maximizing the profit, denoted by

$$\mathcal{P}_i : \max_{\alpha_i \in \mathcal{A}_\vartheta} \left(\frac{\alpha_i D\Big(v + \frac{\min_j\{\alpha_j\}}{2}\Big)}{\sum_{j=1}^{N}\phi_j(\vec{\alpha})} - \frac{c_i}{\phi_i(\vec{\alpha}) + \mathbb{I}(\phi_i(\vec{\alpha}) = 0)} \right) \phi_i(\vec{\alpha}) \tag{2.1.6}$$

where $\phi_i(\vec{\alpha})$ is defined in (WTA). The multiplication by $\phi_i(\vec{\alpha})$ in (2.1.6) specifies that the cost c_i only incurs during the trade. If the market share $\phi_i(\vec{\alpha})$ is

0, then the market maker i has 0 profit since no trade occurs. The indicator function $\mathbb{I}(\phi_i(\vec{\alpha}) = 0)$ is introduced only to avoid the 0 denominator.

Until now, we have completed defining the game-theoretic model for multiple market makers. Each market maker aims to maximize the objective defined in (2.1.6), which involves interactions between market makers through their quoted spread vector $\vec{\alpha}$. In the subsequent study, we shall focus on the homogeneous case where the N requested market makers share the same parameters, including the costs and market share functions. The setting of homogeneous dealers is common in the literature on the analysis of dealer markets, e.g., [Wang 2023; Riggs et al. 2020]. The homogeneous setting aligns with the situation of many MD2C platforms, where dealers often operate under similar conditions, such as access to the same market information, regulatory frameworks, and comparable cost structures. Specifically, under the 'winner-takes-all' mechanism, the presence of homogeneous market makers with the same cost is actually a consequence of competition. Suppose that the tick size ϑ is small enough and that there are 2 market makers with cost $c_1 < c_2$ under the 'winner-takes-all' mechanism, then the market maker with cost c_1 will unilaterally win all the order flow by quoting a lower spread than the market maker with cost c_2, generating a revenue in the interval $[c_1, c_2)$ which does not cover the cost of the latter market maker. This suggests that with multiple market makers, only those with the same lowest cost are eligible for competition under the (WTA) mechanism, as they are always able to quote a lower spread to exclude the market makers with higher cost from competition. Therefore, we make the following assumption:

Assumption 2.1.10 (Homogeneous case)**.** The N requested market makers have the same fixed cost: $c_i \equiv c$. We consider the following scheme for their respective market shares:

- 'Winner-takes-all': $\phi_i(\vec{\alpha}) \equiv \kappa \mathbb{I}(\alpha_i = \min_j\{\alpha_j\})$, κ is a constant.

2.1.3 Equilibrium and Collusive Spreads

A Nash equilibrium is a situation where no market maker has any incentive to change their quoted spread:

Definition 2.1.11 (Nash equilibrium)**.** A Nash equilibrium for system (2.1.6) is a spread vector $\vec{\alpha}^e = (\alpha_1^e, \ldots, \alpha_N^e) \in (\mathcal{A}_\vartheta)^N$, such that $\forall i \in \{1, \ldots, N\}$,

$$J_i(\alpha_i^e, \alpha_{-i}^e) \geq J_i(\alpha, \alpha_{-i}^e), \forall \alpha \in \mathcal{A}_\vartheta \tag{2.1.7}$$

where $\alpha^e_{-i} = (\alpha^e_1, \ldots, \alpha^e_{i-1}, \alpha^e_{i+1}, \ldots, \alpha^e_N)$ is the spread vector excluding the spread of market maker i.

We are primarily interested in the Nash equilibrium under the 'winner-takes-all' scenario that reflects the application of MD2C platforms in practice. Before studying the Nash equilibrium, we prepare some notations related to the cost c of one market maker.

From Remark 2.1.6 and Proposition 2.1.7 we have

$$s^* D(v + \frac{s^*}{2}) \geq c$$

where s^* is the unique maximum of the revenue function $F(s)$ defined in Proposition 2.1.7. If the cost c is high such that $s^* D(v + \frac{s^*}{2}) = c$, the setting degenerates into the single market maker scenario where s^* is the only possible spread that the single marker maker would quote to cover the cost c. To study Nash equilibrium with multiple market makers, we only consider the case where

$$s^* D(v + \frac{s^*}{2}) > c \tag{2.1.8}$$

Then the following equation admits two distinct solutions.

$$sD(v + \frac{s}{2}) = c \tag{2.1.9}$$

Denote the two distinct solutions by $\tilde{s}_1, \tilde{s}_2$ so that $\tilde{s}_1 < s^* < \tilde{s}_2$. The quoted spread of a market maker is between the interval $[\tilde{s}_1, \tilde{s}_2]$ to ensure a non-negative profit. Figure 2.2 shows an example of the revenue function in (Example - Revenue) with cost $c = 0.5$. The number $\tilde{s}_1$ is essential for representing the Nash equilibrium market spread under the 'winner-takes-all' market share mechanism.

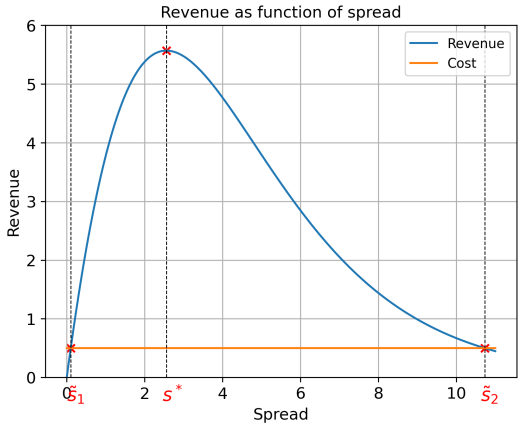


Figure 2.2: When cost $c = 0.5$, the equation (2.1.9) admits two distinct solutions $\tilde{s}_1, \tilde{s}_2$ such that $\tilde{s}_1 < s^* < \tilde{s}_2$.

In the analysis of Nash equilibrium, it is essential to account for the fact that market makers quote spreads from space $\mathcal{A}_\vartheta$ with tick size ϑ. To incorporate this into our analysis, we introduce the projection functions $P_\vartheta^+, P_\vartheta^-$ that map any real value $x \in \mathbb{R}$ to the closest elements in this discrete space $\mathcal{A}_\vartheta$ from above and below, respectively. Formally, for any $x \in \mathbb{R}$, the projection functions are defined as follows:

$$\begin{aligned} P_\vartheta^+(x) &= \min\{z \geq x, z \in \mathcal{A}_\vartheta\} \\ P_\vartheta^-(x) &= \max\{z < x, z \in \mathcal{A}_\vartheta\} \end{aligned} \tag{2.1.10}$$

By definition, the value $P_\vartheta^+(\tilde{s}_1) \in \mathcal{A}_\vartheta$ is the minimum spread that a market maker would possibly quote in the case of a single market maker, given that the revenue $F(P_\vartheta^+(\tilde{s}_1))$ can cover her cost c. If $F(P_\vartheta^+(\tilde{s}_1)) < c$, this implies that the tick size ϑ is not small enough such that the discrete set $\mathcal{A}_\vartheta$ does not contain any value in the interval $[\tilde{s}_1, \tilde{s}_2]$. Then no market maker would ever join the market because the revenue with any quote from $\mathcal{A}_\vartheta$ cannot cover their cost c. Hence we make the assumption on the tick size ϑ for the multi-agent market making problem to be well-defined.

Assumption 2.1.12. The tick size ϑ satisfies

$$0 < \vartheta < \tilde{s}_2 - \tilde{s}_1 \tag{2.1.11}$$

Assumption 2.1.12 ensures that the quoted spread $P_\vartheta^+(\tilde{s}_1)$ is located in the interval $[\tilde{s}_1, \tilde{s}_2)$:

$$\tilde{s}_1 \leq P_\vartheta^+(\tilde{s}_1) < \tilde{s}_2 \tag{2.1.12}$$

In fact, we shall see later that $P_\vartheta^+(\tilde{s}_1)$ is exactly the lowest possible market spread in Nash equilibrium under the 'winner-takes-all' mechanism. We define a constant K that represents the maximum number of market makers who can simultaneously quote the minimum spread $P_\vartheta^+(\tilde{s}_1)$. Formally,

$$K := \max\{k \in \mathbb{N} | F(P_\vartheta^+(\tilde{s}_1)) \geq k \cdot c\} \tag{2.1.13}$$

From (2.1.12), we see that K is well-defined since $K \geq 1$. K is finite since $F(s) \leq F(s^*)$. This number K will be used to characterize Nash equilibrium under 'winner-takes-all' mechanism. Now we state the following propositions on Nash equilibrium under the 'winner-takes-all' mechanism in Assumption 2.1.10.

Proposition 2.1.13. *Under a 'winner-takes-all' configuration as defined in (WTA), if Assumption 2.1.10 and Assumption 2.1.12 are satisfied, then the Nash equilibrium exists. If a spread vector $\vec{\alpha}^e$ satisfies one of the following conditions depending on the number of market makers N and the constant K defined in (2.1.13).*

- *When $2 \leq N \leq K$, $\alpha_i^e = P_\vartheta^+(\tilde{s}_1), \forall i \in \{1, \ldots, N\}$.*
- *When $N > K$, $\vec{\alpha}^e$ has exactly K coordinates equal to $P_\vartheta^+(\tilde{s}_1)$, and the remaining $(N-K)$ coordinates are strictly higher than $P_\vartheta^+(\tilde{s}_1)$.*
- *When $N > K \geq 2$ and $F(P_\vartheta^+(\tilde{s}_1)) = K \cdot c$, $\vec{\alpha}^e$ has exactly $K-1$ coordinates equal to $P_\vartheta^+(\tilde{s}_1)$, and the remaining $(N-K+1)$ coordinates are strictly higher than $P_\vartheta^+(\tilde{s}_1)$.*

Then $\vec{\alpha}^e$ is a Nash equilibrium.

Proof. We show the existence of Nash equilibrium by examining that the vector $\vec{\alpha}^e$ satisfies Definition 2.1.11 of Nash equilibrium. In all 3 cases, the market spread is $\alpha^M = P_\vartheta^+(\tilde{s}_1)$.

- When $2 \leq N \leq K$, suppose that every market maker quotes the spread $P_\vartheta^+(\tilde{s}_1)$. We consider the market maker indexed by i whose profit is

$$J_i(\alpha_i^e, \alpha_{-i}^e) = \frac{P_\vartheta^+(\tilde{s}_1) D(v + \frac{P_\vartheta^+(\tilde{s}_1)}{2})}{N} - c \geq 0$$

  The inequality holds by the definition of K with $2 \leq N \leq K$. If the market maker i deviates to a spread $\alpha_i \in \mathcal{A}_\vartheta$ such that $\alpha_i < P_\vartheta^+(\tilde{s}_1)$ while the spread of other market makers remains α_{-i}^e, the market maker i wins all order flow but with a profit

$$J_i(\alpha_i, \alpha_{-i}^e) = \alpha_i D(v + \frac{\alpha_i}{2}) - c < 0 \leq J_i(\alpha_i^e, \alpha_{-i}^e)$$

  The profit $J_i(\alpha_i, \alpha_{-i}^e)$ is negative by the definition of $P_\vartheta^+(\tilde{s}_1)$. If the market maker i switches to a spread $\alpha_i \in \mathcal{A}_\vartheta$ such that $\alpha_i > P_\vartheta^+(\tilde{s}_1)$ while the spread of other market makers remains α_{-i}^e, then the market share of i drops to 0, and consequently the profit of the market maker i is 0. In both cases, $J_i(\alpha_i, \alpha_{-i}^e) \leq J_i(\alpha_i^e, \alpha_{-i}^e)$. Hence, when $2 \leq N \leq K$, $\vec{\alpha}^e$ with $\alpha_i = P_\vartheta^+(\tilde{s}_1), \forall i$ is a Nash equilibrium.

- When $N > K$, without loss of generality we assume $\alpha_1^e = \ldots = \alpha_K^e = P_\vartheta^+(\tilde{s}_1)$, and $\alpha_j^e > P_\vartheta^+(\tilde{s}_1), \forall j > K$ and $j \leq N$. The profit of each market maker indexed by $i \in \{1, \ldots, K\}$ is

$$J_i(\alpha_i^e, \alpha_{-i}^e) = \frac{P_\vartheta^+(\tilde{s}_1) D(v + \frac{P_\vartheta^+(\tilde{s}_1)}{2})}{K} - c \geq 0$$

  And the profit of the market maker indexed by $j \in \{K+1, \ldots, N\}$ is $J_j(\alpha_j^e, \alpha_{-j}^e) = 0$. The discussion of the market maker i ($i \in \{1, \ldots, K\}$) is similar to that in the case of $2 \leq N \leq K$. The profit of the market maker i becomes negative if the market maker i lowers the spread from $P_\vartheta^+(\tilde{s}_1)$ and becomes 0 if the market maker i increases the spread from $P_\vartheta^+(\tilde{s}_1)$. Now suppose that the market maker j ($j \in \{K+1, \ldots, N\}$) quoting $\alpha_j^e > P_\vartheta^+(\tilde{s}_1)$ unilaterally deviates to a spread value α_j. The profit of the market maker j has the following 3 cases:

$$J_j(\alpha_j, \alpha_{-j}^e) = \begin{cases} 0, & \text{if } \alpha_j > P_\vartheta^+(\tilde{s}_1) \\ \frac{P_\vartheta^+(\tilde{s}_1) D(v + \frac{P_\vartheta^+(\tilde{s}_1)}{2})}{K+1} - c < 0, & \text{if } \alpha_j = P_\vartheta^+(\tilde{s}_1) \\ \alpha_j D(v + \frac{\alpha_j}{2}) - c < 0, & \text{if } \alpha_j < P_\vartheta^+(\tilde{s}_1) \end{cases} \tag{2.1.14}$$

  Hence, we have $J_j(\alpha_j, \alpha_{-j}^e) \leq J_j(\alpha_j^e, \alpha_{-j}^e) = 0$ for $j \in \{K+1, \ldots, N\}$, and $\vec{\alpha}^e$ is a Nash equilibrium.

- When $N > K \geq 2$ and $F(P_\vartheta^+(\tilde{s}_1)) = K \cdot c$, without loss of generality we assume $\alpha_1^e = \ldots = \alpha_{K-1}^e = P_\vartheta^+(\tilde{s}_1)$, and $\alpha_j^e > P_\vartheta^+(\tilde{s}_1), \forall j > K-1$ and $j \leq N$. The profit of each market maker indexed by $i \in \{1, \ldots, K-1\}$ is

$$J_i(\alpha_i^e, \alpha_{-i}^e) = \frac{P_\vartheta^+(\tilde{s}_1) D(v + \frac{P_\vartheta^+(\tilde{s}_1)}{2})}{K-1} - c > 0$$

  The discussion of the market maker i ($i \in \{1, \ldots, K-1\}$) is similar to previous cases, where they have no incentive to deviate from their quoted spread $P_\vartheta^+(\tilde{s}_1)$. The profit of the market maker indexed by $j \in \{K, \ldots, N\}$ is 0, which can not be improved by unilaterally adjusting their quoted spread. The discussion is similar to (2.1.14) except that $J_j(\alpha_j, \alpha_{-j}^e) = 0$ if $\alpha_j = P_\vartheta^+(\tilde{s}_1)$. Then, we also have $J_j(\alpha_j, \alpha_{-j}^e) \leq J_j(\alpha_j^e, \alpha_{-j}^e) = 0$ for $j \in \{K, \ldots, N\}$, and consequently, $\vec{\alpha}^e$ is a Nash equilibrium.

□

Remark 2.1.14. When $N > K \geq 2$, the Nash equilibrium can take the forms in the second or third point of Proposition 2.1.13.

Remark 2.1.15. For the 'winner-takes-all' mechanism, the form of the Nash equilibrium is, in general, not unique, depending on the tick size ϑ, the cost c, the shape of the revenue function $F(s)$ and the number of requested market makers N. For example, suppose that the tick size ϑ is high enough with relevant assumptions on $F(s)$ and c, such that the intersection of the discrete set of possible quoted spreads $\mathcal{A}_\vartheta$ and the interval $[\tilde{s}_1, \tilde{s}_2]$ only contains two values $\tilde{s}_1$ and s^*. Also, suppose that there are 2 market makers and that

$$\frac{F(s^*)}{2} > F(\tilde{s}_1)$$

Then we can easily check that both $(\tilde{s}_1, \tilde{s}_1)$ and (s^*, s^*) are Nash equilibrium quoting strategies for the 2 market makers, resulting in market spreads of $\tilde{s}_1$ and s^*, respectively.

The spread vector $\vec{\alpha}^e$ in Proposition 2.1.13 results in a market spread of $P_\vartheta^+(\tilde{s}_1)$, which is the minimum value that a market maker is able to quote to possibly make a non-negative profit. We hereby consider this type of Nash equilibrium $\vec{\alpha}^e$ as the competitive benchmark that provides the cheapest market spread to the client, which leads to the following definition of **Competitive Nash Equilibrium**.

Definition 2.1.16 (Competitive Nash Equilibrium)**.** Under a 'winner-takes-all' configuration as defined in (WTA), if Assumption 2.1.10 and Assumption 2.1.12 are satisfied, we call the spread vector $\vec{\alpha}^e$ a Competitive Nash Equilibrium if $\vec{\alpha}^e$ is a Nash equilibrium and

$$\min_{i \in \{1,\dots,N\}} \alpha_i^e = P_\vartheta^+(\tilde{s}_1) \tag{2.1.15}$$

Definition 2.1.16 formalizes the equilibrium in a 'winner-takes-all' market, where market makers compete aggressively, driving market spread down to the lowest feasible level. At this equilibrium, clients receive the best possible pricing, while market makers are incentivized to operate at the margin of profitability. Although the Nash equilibrium from Definition 2.1.11 may not be unique, we state in Proposition 2.1.17 that any Competitive Nash Equilibrium which results in a market spread of $P_\vartheta^+(\tilde{s}_1)$ is characterized by the forms in Proposition 2.1.13.

Proposition 2.1.17. *Under a 'winner-takes-all' configuration as defined in (WTA), if Assumption 2.1.10 and Assumption 2.1.12 are satisfied, and $\vec{\alpha}^e$ is a competitive Nash equilibrium in Definition 2.1.16, then $\vec{\alpha}^e$ satisfies one of the forms specified in Proposition 2.1.13.*

Proof. We discuss the different cases in Proposition 2.1.13.

- When $2 \leq N \leq K$, suppose that there exists $\alpha_j^e > P_\vartheta^+(\tilde{s}_1)$. By the 'winner-takes-all' principle, $\phi_j(\vec{\alpha}^e) = 0$. The market maker j has 0 profit, but can increase the profit to at least

$$\frac{F(P_\vartheta^+(\tilde{s}_1))}{N} - c > 0$$

by unilaterally quoting the spread $P_\vartheta^+(\tilde{s}_1)$, which contradicts the fact that $\vec{\alpha}^e$ is a Nash equilibrium. Therefore, $\forall j \in \{1, \ldots, N\}$, $\alpha_i^e \equiv P_\vartheta^+(\tilde{s}_1)$.

- When $N > K$, let k denote the number of market makers who quote the spread $P_\vartheta^+(\tilde{s}_1)$. We see that $k \geq 1$:

$$k := \sum_{i=1}^{N} \mathbb{I}(\alpha_i^e = P_\vartheta^+(\tilde{s}_1))$$

Without loss of generality, we assume that the market makers indexed by i where $i \in \{1, \ldots, k\}$ quote the spread $P_\vartheta^+(\tilde{s}_1)$, and the remaining market makers quote the spreads higher than $P_\vartheta^+(\tilde{s}_1)$.

If $K < k \leq N$, then by the definition of K, the profit of any of the market makers indexed by 1 to k is

$$\frac{F(P_\vartheta^+(\tilde{s}_1))}{k} - c < 0$$

which can be improved to 0 if the market maker unilaterally increases their quoted spread, which contradicts the fact that $\vec{\alpha}^e$ is a Nash equilibrium. Hence, we have $k \leq K$.

When $K = 1$, we can easily check that $k = 1$ and the proposition is proved since $\vec{\alpha}^e$ satisfies the form specified in the second point of Proposition 2.1.13.

Now we consider the case where $K \geq 2$. If $k \leq K - 2$, then the profit of the market maker $k + 1$ is 0 as she quotes a spread higher than $P_\vartheta^+(\tilde{s}_1)$, but can improve the profit to a strictly positive value

$$\frac{F(P_\vartheta^+(\tilde{s}_1))}{k+1} - c \geq \frac{F(P_\vartheta^+(\tilde{s}_1))}{K-1} - c > 0$$

if the market maker $k + 1$ unilaterally switches to quote $P_\vartheta^+(\tilde{s}_1)$. This also contradicts the Nash equilibrium $\vec{\alpha}^e$. Therefore, we obtain $K - 1 \leq k \leq K$. In this case, k can only take 2 possible values: $K - 1$ or K.

$k = K$ is plausible for the Nash equilibrium $\vec{\alpha}^e$, and a Nash equilibrium with $k = K$ can only take the form specified in the second point of Proposition 2.1.13. When $F(P^+_\vartheta(\tilde{s}_1)) > K \cdot c$, $k = K-1$ contradicts the Nash equilibrium property of $\vec{\alpha}^e$, since the market maker indexed by K can increase her profit from 0 to $\frac{F(P^+_\vartheta(\tilde{s}_1))}{K} - c > 0$ by unilaterally quoting a lower spread $\alpha^e_K = F(P^+_\vartheta(\tilde{s}_1))$. Only when $F(P^+_\vartheta(\tilde{s}_1)) = K \cdot c$, k equal to $K-1$ is consistent with the Nash equilibrium property of $\vec{\alpha}^e$ and in this case any Nash equilibrium with $k = K - 1$ can only take the form specified in the third point of Proposition 2.1.13.

Summarizing the above discussion, we obtain that the Nash equilibrium $\vec{\alpha}^e$ with a market spread of $P^+_\vartheta(\tilde{s}_1)$ takes the forms specified in Proposition 2.1.13. □

Remark 2.1.18. When $N > K$, at a Competitive Nash Equilibrium, the number of market makers quoting the minimum spread $P^+_\vartheta(\tilde{s}_1)$ remains stable due to profit constraints, as Proposition 2.1.17 suggests. However, the specific market makers quoting this spread can change over time. Consider the case where $F(P^+_\vartheta(\tilde{s}_1)) > K \cdot c$. Then there are K market makers that quote $P^+_\vartheta(\tilde{s}_1)$. Consider a market maker currently quoting a higher spread whose profit is 0. To try to make a positive profit, this market maker has the incentive to reduce their spread to $P^+_\vartheta(\tilde{s}_1)$ to capture order flow. However, this increase in competition at $P^+_\vartheta(\tilde{s}_1)$ causes the total profit to be divided among more participants, potentially resulting in negative profits for market makers quoting the minimum spread. As a consequence, one of the market makers quoting $P^+_\vartheta(\tilde{s}_1)$ will eventually increase their spread to avoid further losses or bankruptcy. This restores the equilibrium number of K market makers that quote $P^+_\vartheta(\tilde{s}_1)$, as specified in Proposition 2.1.13. The case where $F(P^+_\vartheta(\tilde{s}_1)) = K \cdot c$ is similar, while the number of market makers quoting $P^+_\vartheta(\tilde{s}_1)$ remains stable, varying between $K-1$ and K over time. Thus, while the identity of the market makers that quote the minimum spread $P^+_\vartheta(\tilde{s}_1)$ may vary, the total number of such market makers remains stable, maintaining the balance of competition and profitability in the market.

So far, we have defined the game-theoretic model (2.1.6) with the market share mechanism (WTA) that corresponds to the stylized features of MD2C platforms. In this framework, market makers compete by quoting bid-ask spreads, with the 'winner-takes-all' mechanism driving intense competition for order flow. The Competitive Nash Equilibrium arises when market makers

quote the lowest possible spread, benefiting clients with minimal transaction costs. However, despite the competitive scenario in such markets, it is possible for market makers to engage in a collusion strategy, where they coordinate to quote higher spreads, by agreeing to maintain supra-competitive spreads for higher profits instead of keeping the spread down to the Competitive Nash Equilibrium spread $P^+_\vartheta(\tilde{s}_1)$. The seminal work [Christie and Schultz 1994] studies the avoidance of odd-eighth quotes by market makers on the NASDAQ and attributes the absence of a $\frac{1}{8}$ quote to the collusive behavior of market makers quoting higher spreads.

We outline the following key elements of a collusive strategy.

- When market makers enter an agreement of collusion, the resulting market spread is strictly higher than the Competitive Nash Equilibrium market spread $P^+_\vartheta(\tilde{s}_1)$.

- Under a collusive strategy, a reward-punishment mechanism is typically implied ([Tirole 1988]). If one market maker breaks the collusive agreement by quoting a lower spread, the others may retaliate by lowering their spreads respectively, resulting in a price war where everyone earns lower profits. This mechanism helps maintain the collusive arrangement.

In the context of our game-theoretic model with the 'winner-takes-all' market share mechanism, a collusion strategy is formally defined as follows.

Definition 2.1.19 (Collusion strategy)**.** A collusion strategy for system (2.1.6) with N requested market makers is a spread vector $\vec{\alpha}^c = (\alpha^c_i)_{i=1}^N \in (\mathcal{A}_\vartheta)^N$, where $\forall i \in \{1, \ldots, N\}, \alpha^c_i = \alpha^c$, such that

$$\alpha^c > P^+_\vartheta(\tilde{s}_1) \tag{2.1.16}$$

and

$$\frac{F(\alpha^c)}{N} > \frac{F(P^+_\vartheta(\tilde{s}_1))}{K\mathbb{I}(N > K) + N\mathbb{I}(2 \leq N \leq K)} \tag{2.1.17}$$

where $F(\cdot)$ is the revenue function in (2.1.4) and K is the constant defined in (2.1.13)

The market spread α^c under the collusion strategy is called the collusive spread. It is strictly larger than the Competitive Nash Equilibrium market spread $P^+_\vartheta(\tilde{s}_1)$. Since the revenue of the market makers that win the order flow under the Competitive Nash Equilibrium is $\frac{F(P^+_\vartheta(\tilde{s}_1))}{N}$ when $2 \leq N \leq K$

and $\frac{F(P_\vartheta^+(\tilde{s}_1))}{K}$ when $N > K$, as specified in Proposition 2.1.13, (2.1.17) demonstrates that the average revenue per market maker under the collusion strategy $\vec{\alpha}^c$ exceeds the revenue per market maker in the Competitive Nash Equilibrium. This mechanism effectively enforces a reward-punishment scheme, incentivizing the market makers to maintain the collusion strategy $\vec{\alpha}^c$.

To see how the reward-punishment scheme works, suppose that the N requested market makers have agreed to collude and quote the spread α^c. If any market maker deviates by lowering their spread, the other market makers will respond by also lowering their spreads to avoid making 0 profit. This triggers a price war until there are enough market makers that quote the lowest possible spread $P_\vartheta^+(\tilde{s}_1)$ and a Competitive Nash Equilibrium is reached. However, according to the condition (2.1.17, all market makers will make less profit in this scenario compared to the collusion strategy. In other words, market makers are rewarded with higher profits when they maintain the collusion strategy, and are punished with lower profits if any of them breaks the collusion agreement. Hence, they are incentivized to maintain the collusion strategy.

We emphasize that the collusion strategy is prearranged by the N requested market makers in advance. It is important to note that while the collusion strategy represents a coordinated agreement, it may not necessarily constitute a Nash equilibrium. This is because, unlike in a Nash equilibrium where no market maker has an incentive to unilaterally deviate, in a collusion strategy, market makers might still have individual incentives to undercut the agreed collusive spread to capture the entire order flow. However, the threat of collective retaliation through a price war discourages such deviations. Thus, the collusion strategy relies on the implied reward-punishment scheme instead of the inherent stability of a Nash equilibrium to maintain higher spreads and collective profitability.

It is important to note that we do not guarantee the existence of a collusion strategy in Definition 2.1.19. Instead, we formally define what constitutes a collusion strategy should it exist. If such a strategy emerges, we consider the collusive spread as a benchmark for collusion. In fact, we can simply check whether $P_\vartheta^+(s^*)$ and $P_\vartheta^-(s^*)$ satisfy Definition 2.1.19, because either of these two values maximizes the revenue function $F(s)$ in the spread space $\mathcal{A}_\vartheta$ and are most likely to be the agreed spread to quote in such a collusion strategy.

In the algorithmic simulation in Section 2.2, the learning algorithms operating in the market environment do not necessarily converge to give the same

quoted spread. Thus, the standard definition of collusion provided in Definition 2.1.19 may not fully capture the cases where algorithms learn to collude tacitly without directly sharing information. To account for this, we introduce a benchmark for algorithmic collusion from the perspective of the market spread, which directly impacts clients:

Definition 2.1.20 (Perceived collusion strategy)**.** A perceived collusion strategy for system (2.1.6) with N requested market makers is a spread vector $\vec{\alpha}^{pc} = (\alpha_i^{pc})_{i=1}^N \in (\mathcal{A}_\vartheta)^N$, such that

$$\min_{i\in\{1,\dots,N\}} \alpha_i^{pc} = \alpha^c \tag{2.1.18}$$

where α^c is the collusive spread in Definition 2.1.19.

Definition 2.1.20 focuses on the perspective of clients, who directly experience the market spreads. It only requires that the minimum quoted spread among the market makers is equal to the collusive spread α^c, while Definition 2.1.19 requires the actual collusive behavior of all market makers quoting the same collusive spread α^c. Whether there is an explicit agreement to collude does not affect the nature of this new definition: if there is such an agreement, then the market makers quoting α^c are able to redistribute the profits evenly to the colluding participants, which are still higher than the level of Competitive Nash Equilibrium as suggested by (2.1.17). Hence, Definition 2.1.20 provides a more suitable framework for evaluating the potential outcomes of learning algorithms in a market, especially in cases where tacit collusion may occur.

2.2 Emergence of Collusive Behavior from Learning

In this section, we study interactions of market makers when they are all applying multi-agent reinforcement learning algorithms to set their spreads in a repeated game. The motif is that training of RL algorithms only requires data from interactions with the environment. This feature fits into the practical scenarios where a market maker selects a spread at each unit time interval and then gets feedback from market transactions about its profit level with this spread. Based on this spread and profit information, the market maker adjusts its spread in the next step to make a better profit. The automation of these pricing-and-earning steps introduces naturally usage of RL algorithms.

Another stylized feature of market making on MD2C platforms is the decentralization of agents. In the actual dealer market, the market makers are not allowed to communicate. As a result, they only observe partial information concerning their own historical spreads and profits, while the only public information accessible to all is the market spread α^M. Dealers make decisions based on the set of partially observed information. Usually, the spreads have a rather small tick size ϑ, and can be approximately regarded as continuous values in $\mathbb{R}^+$. Therefore, we focus on decentralized multi-agent policy gradient algorithms, which allow continuous state and action spaces.

The algorithm we adapt is the decentralized multi-agent deep-deterministic policy gradient (Decentralized MADDPG), inspired by [Lowe et al. 2017] and [Foerster et al. 2018]. [Lowe et al. 2017] and [Foerster et al. 2018] proposed a multi-agent actor-critic algorithm with decentralized actors and centralized critics, in which the actors are functions of each agent's local observation, and critics are functions of all agents' joint states and actions. In their algorithms, critics are trained in a centralized way, which means certain communication during training steps needs to be allowed. We adapt the MADDPG algorithm to our market making model by constraining the critics to be decentralized as well.

2.2.1 Decentralized Multi-agent Actor-critic Algorithms

In the dealer market that we simulate, there are N requested market makers responding to the order flow. The agents share a continuous state space $\mathcal{X} = [0, B]$, where B is the upper bound of the spread value. We equate the action space with the shared state space: agents choose their spreads from the same continuous action space $\mathcal{A} = \mathcal{X} = [0, B]$. At time step $t = 0, 1, \ldots$, the state $x_t^i \in \mathcal{X}$ of agent i represents its current spread, the action $\alpha_t^i \in \mathcal{A}$ is the spread it sets for the next time step at $t + 1$. In addition to agent i's own spread x_t^i, we also allow i to observe the market spread $x_t^M := \min_j x_t^j$. The market spread x_t^M represents the price at which actual transactions occur at time t, and this information is observable by every market maker after the transaction is complete. Observing local and public information, market makers generate quotes α_t^i for the next time step $t+1$. This setting adapts to the actual MD2C platforms where the market spread is observable for every market maker.

At time t, when agent i at state $x_t^i \in \mathcal{X}$ takes an action $\alpha_t^i \in \mathcal{A}$, it receives a deterministic reward $r_i(x_t^i, x_t^M, \alpha_t^i)$

$$r_i(x_t^i, x_t^M, \alpha_t^i) = J_i(\alpha_t^i, \alpha_t^{-i}) \tag{2.2.1}$$

where $J_i(\alpha_t^i, \alpha_t^{-i})$ is the objective function defined in (2.1.6),

$$\alpha_t^{-i} = (\alpha_t^1, \ldots, \alpha_t^{i-1}, \alpha_t^{i+1}, \ldots, \alpha_t^N)$$

is the joint actions of other agents than i taken at time t. (2.2.1) shows that the reward of agent i is jointly determined by the spreads quoted by each agent.

After taking action α_t^i, the state of agent i in the next time step $t+1$ changes to $x_{t+1}^i = \alpha_t^i$. Mathematically, the transition probability is deterministic:

$$\mathbf{P}\left(x_{t+1}^i | x_t^i, x_t^M, \alpha_t^i\right) = \mathbb{I}\left\{x_{t+1}^i = \alpha_t^i\right\} \tag{2.2.2}$$

With the environment (2.2.1) and (2.2.2) given, at time t market maker i applies certain strategy to pick action based on the partial information (x_t^i, x_t^M):

$$\alpha_t^i = \pi_i(x_t^i, x_t^M) \tag{2.2.3}$$

where $\pi_i(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathcal{A}$ is a deterministic policy mapping the product of state and market spread space $\mathcal{X} \times \mathcal{X}$ to action space $\mathcal{A}$. We restrict the policy to be deterministic because we perceive that in practice, the market makers are more likely to apply a specific pricing algorithm rather than a probabilistic one.

(2.2.1)-(2.2.3) define a Markov decision process with partial observation. The objective of agent i is to find an optimal strategy π_i^*, which maximizes the expected future rewards discounted by a factor $\gamma \in [0, 1)$. The value function $V_i(x, m)$ for a given agent spread x and market spread m is defined in (2.2.4):

$$V_i(x, m) = \max_{\pi_i} \mathbb{E}_{\pi_i}\left[\sum_{t=0}^{\infty} \gamma^t r_i(x_t^i, x_t^M, \alpha_t^i) \middle| x_0^i = x, x_0^M = m\right] \tag{2.2.4}$$

$$\pi_i^*(x, m) = \arg\max_{\alpha} \left\{\sum_{x'} \mathbf{P}(x' | x, m, \alpha)(r_i(x, m, \alpha) + \gamma V_i(x', m'))\right\} \tag{2.2.5}$$

where in (2.2.5) m' is the market spread of the next time step jointly determined by the actions of all agents. In many reinforcement learning algorithms,

a state-action value function $Q_i(x, m, \alpha)$ is introduced instead of the value function to compute the optimal strategy.

$$Q_i(x, m, \alpha) = \max_{\pi_i} \mathbb{E}_{\pi_i} \left[\sum_{t=0}^{\infty} \gamma^t r_i(x_t^i, x_t^M, \alpha_t^i) \middle| x_0^i = x, x_0^M = m, \alpha_0^i = s \right] \quad (2.2.6)$$

$$\pi_i^*(x, m) = \arg\max_{\alpha} Q_i(x, m, \alpha) \quad (2.2.7)$$

Our decentralized multi-agent DDPG algorithm approximates the state-action value functions $Q_i(x, m, \alpha)$ and the optimal policies $\pi_i^*(x, m)$ by neural networks $Q_i(x, m, \alpha|\theta_i^Q)$ and $\pi_i(x, m|\theta_i^\pi)$, for $i \in \{1, \ldots, N\}$. $Q_i(x, m, \alpha|\theta_i^Q)$ is called the critic network that evaluates the performance of the state-action pair (x, m, α), and $\pi_i(x, m|\theta_i^\pi)$ is the actor network that returns an action for the market maker i given its observation (x, m).

In the implementation, the target critic and actor networks are also introduced coupling critics and actors. The target networks are indicated by $\tilde{Q}_i(x, m, \alpha|\tilde{\theta}_i^Q)$ and $\tilde{\pi}_i(x, m|\tilde{\theta}_i^\pi)$. While network parameters θ_i^Q and θ_i^π are updated at every training iteration, the target network parameters $\tilde{\theta}_i^Q$ and $\tilde{\theta}_i^\pi$ are updated slowly to make the training more stationary:

$$\begin{aligned} \tilde{\theta}_i^Q &\leftarrow \tau\theta_i^Q + (1-\tau)\tilde{\theta}_i^Q \\ \tilde{\theta}_i^\pi &\leftarrow \tau\theta_i^\pi + (1-\tau)\tilde{\theta}_i^\pi \end{aligned} \quad (2.2.8)$$

where τ represents the speed of updating the target network parameters. (In our simulation $\tau = 2\%$.)

For parameters critic networks θ_i^Q, we define the loss function:

$$\mathcal{L}_i^Q(\theta_i^Q) = \mathop{\mathbb{E}}_{x,m,\alpha,x',m'} \left[\left(r_i(x, m, \alpha) + \gamma \tilde{Q}_i(x', m', \alpha'|\tilde{\theta}_i^Q) - Q_i(x, m, \alpha|\theta_i^Q) \right)^2 \right] \quad (2.2.9)$$

where the pairs (x, m, α, x', m') are sampled from an experience replay buffer that stores agents' historical states, actions and transitions during the repeated game. The action $\alpha' = \tilde{\pi}_i(x', m'|\tilde{\theta}_i^\pi)$ is given by the target actor network of agent i. Then the stochastic gradient descent can be applied on (2.2.9) to calibrate the parameter θ_i^Q.

To calibrate the parameters of the actor network θ_i^π, we can write the policy gradient for the actor networks:

$$\nabla_{\theta_i^\pi} \mathcal{L}^\pi(\theta^\pi) = - \mathop{\mathbb{E}}_{x,m} \left[\nabla_s Q_i(x, m, \pi_i(x, m|\theta_i^\pi)|\theta_i^Q) \nabla_{\theta_i^\pi} \pi_i(x, m|\theta_i^\pi) \right] \quad (2.2.10)$$

where the historical states (x, m) are also sampled from the experience replay buffer. In the implementation, we can directly define the loss function

$$\mathcal{L}^{\pi}(\theta^{\pi}) = -\mathop{\mathbb{E}}_{x,m}\Big[Q_i(x, m, \pi_i(x, m|\theta_i^{\pi})|\theta_i^{Q})\Big]$$

and gradient descent is automated by PyTorch.

(2.2.8)-(2.2.10) define the scheme of our decentralized MADDPG algorithm for the continuous state and action space. In our model, there is a positive tick size ϑ to be incorporated into the environment. To achieve this, we define the projection function based on (2.1.10), which maps a real value x to the closest element to x in $\mathcal{A}_{\vartheta}$:

$$P_{\vartheta}(x) = P_{\vartheta}^{+}(x)\mathbb{I}\Big(P_{\vartheta}^{+}(x) - x \leq x - P_{\vartheta}^{-}(x)\Big) + P_{\vartheta}^{-}(x)\mathbb{I}\Big(P_{\vartheta}^{+}(x) - x > x - P_{\vartheta}^{-}(x)\Big) \tag{2.2.11}$$

We integrate the projection function P_{ϑ} into the market environment by applying it within the reward function $r(x, m, \alpha)$. Specifically, the continuous values of (x, m, α) produced by the learning algorithm are mapped by P_{ϑ} to the discrete space $\mathcal{A}_{\vartheta}$ before calculating the profits. Although the learning algorithms operate in continuous state and action spaces, the projected spreads ensure that the actual quoted values are in line with the discrete tick size ϑ. This approach provides a consistent framework for different tick sizes, allowing the learning behavior to reflect the realities of dealer markets where quoted spreads are restricted to discrete increments, while still preserving the key dynamics of the learning process.

Both critics Q_i and actors π_i are local functions of agent i's own observation, together with the market spread determined by every agent. The interactions between market makers are realized through the market spread, and the 'winner-takes-all' market share mechanism included in the reward function (2.2.1). In the next section, we present the simulation results of our algorithm applied to the 'winner-takes-all' market share mechanism, which reflects the competitive dynamics observed in MD2C platforms.

2.2.2 When Does Learning Lead to Tacit Collusion?

In this section, we summarize the results of the numerical experiments conducted under the 'winner-takes-all' market share mechanism (WTA). Our goal is to simulate the behavior of market makers using the Decentralized MADDPG algorithm from Section 2.2.1 and analyze the resulting market spreads in various competitive settings.

The market order flow and market revenue as a function of spread are defined by (Example - Order flow) and (Example - Revenue), whose graphs are shown in Figure 2.1. The market price of the asset is $v = 100$, and the unique maximum spread of the revenue function defined in Proposition 2.1.7 is $s^* \approx 2.5569$. The two distinct solutions to equation (2.1.8) are, respectively,

$$\tilde{s}_1 \approx 0.1026, \qquad \tilde{s}_2 \approx 10.7278$$

We fix the upper bound of the action and the state space $B = 10$. That is, the quoted spread value of each market maker ranges between $[0, 10]$.

We build up a market environment based on the 'winner-takes-all' mechanism. Each market maker's reward function was computed based on their share of the order flow, governed by the 'winner-takes-all' mechanism (WTA), where a market maker only receives revenue if they quote the narrowest spread. In this setup, if multiple market makers quote the same spread, they share the order flow equally. For a given number of market makers N and tick size ϑ in the configuration, we construct multi-agent actor-critic learners, and train the learners using the decentralized MADDPG algorithm introduced in the previous subsection. Both the actor $\pi_i(x, m|\theta_i^\pi)$ and the critic $Q_i(x, m, s|\theta_i^Q)$ are 3-layer fully connected neural networks, with 128 hidden units at each layer. This network architecture is chosen to balance the model complexity and computational efficiency. The choice of 3 layers has been shown to perform well in many deep reinforcement learning applications ([Mnih, Kavukcuoglu, Silver, Rusu, et al. 2015; Lillicrap et al. 2016]), where moderately deep networks can approximate Q-value functions while ensuring stable training. The architecture is deep enough to capture the non-linearities of the quoting strategies via interactions between the market makers. The 128 neurons per layer provide sufficient capacity to model the quoting strategies without overfitting, with reasonable computational costs.

The actor $\pi_i(x, m|\theta_i^\pi)$ takes market maker i's current spread x and the market spread m, then outputs a value between $[0, 10]$ for market maker i's next spread value. We also apply $\epsilon-$greedy exploration with $\epsilon = 5\%$. The agents have a probability of $\epsilon e^{-\beta \cdot t}$ picking a random action instead of the output of the policy network, where $\beta = 0.001$, t is the number of iterations.

For the optimization algorithm, we apply a standardized Adam optimizer ([Kingma and Ba 2015]) with learning rate 0.0005, momentum decay rates $(0.9, 0.99)$ and non-zero regularization 10^{-6}.

The simulations are carried out with different numbers of market makers N to explore how increased competition influences the pricing dynamics resulting from the interacting algorithms. We will consider the (N, ϑ) combination with $N \in \{2, 4, 8, 16\}$ and $\vartheta = 0.0001$ as the benchmark cases. For each of these combinations, we carried out 50 independent experiments, each of which involved the complete training of N market makers from scratch. An experiment of complete training takes 150 episodes, and in each episode 500 iterations are used. This is done to reduce the variance that may arise from any single round of training, ensuring more robust and reliable results.

The average cumulative reward per episode during the training steps is shown in Figure 2.3 with the confidence interval 95%. The reward curves show that market makers effectively learn to improve their quoting strategies to earn more profits, under different levels of competition. In all the 4 cases, we observe a decreasing trend of episodic reward in the first few episodes, suggesting the impact of competition leading to a decline in their profits in the first few episodes. In the case of 2 market makers, the agents demonstrate steady learning, with rewards stabilizing around $700 - 1000$. With 4 market makers, the reward curves exhibit more variability but show consistent improvement, reaching cumulative rewards per episode around $50 - 400$. In the 8 market maker scenario, the rewards are generally lower and more volatile, fluctuating between 0 and 250 due to increased competition. Finally, with 16 market makers, competition becomes the most intense, leading to highly oscillating patterns, with only a few agents achieving rewards between 100 and 600, while the majority of the market makers remain with much lower profits. This suggests the challenge of profitability in highly competitive environments.

Figure 2.3: Average cumulative reward per episode during training with different numbers of market makers, tick size $\vartheta = 0.0001$, from 50 scenarios

In Figure 2.4, we track the average market spreads during the training steps, which are calculated by averaging the market spreads at each iteration per episode across all 50 experiments. As suggested in Section 2.1.2, the market spread is a direct indicator of the market dynamics, which reflects the interacting behaviors of the learning algorithms over time. The market spread also serves as a proxy for competition. From Definition 2.1.20, it can be considered as a direct metric for perceived collusion.

The patterns observed in Figure 2.4 exhibit a similar trajectory across all configurations with 2, 4, 8 and 16 market makers. Initially, there is a rapid and steep decline in market spreads. This early drop could be attributed to algorithms that attempt to undermine each other as they explore lower spreads to capture order flow in the early stages of learning. However, as training continues, the market spreads begin to rise again, and this increase suggests

that the algorithms start to converge towards a higher and more stable spread level, which is still above the Competitive Nash Equilibrium. An important observation is that as the number of market makers increases from 2 to 16, the overall market spread values decrease, reflecting more intensive competition leading to more competitive market spread levels. However, even with 16 market makers, the final average spreads remain consistently above the theoretical Competitive Nash Equilibrium level 0.1027, indicating that full competition is not achieved. This pattern suggests that despite the increased competition, the learning algorithms converge to supra-competitive spread levels, potentially indicating tacit collusion during the training process. The gradual rise in the spreads, after the initial sharp decline, highlights that market makers might implicitly learn to maintain higher spreads without explicit coordination, a behavior often associated with collusion in repeated games.

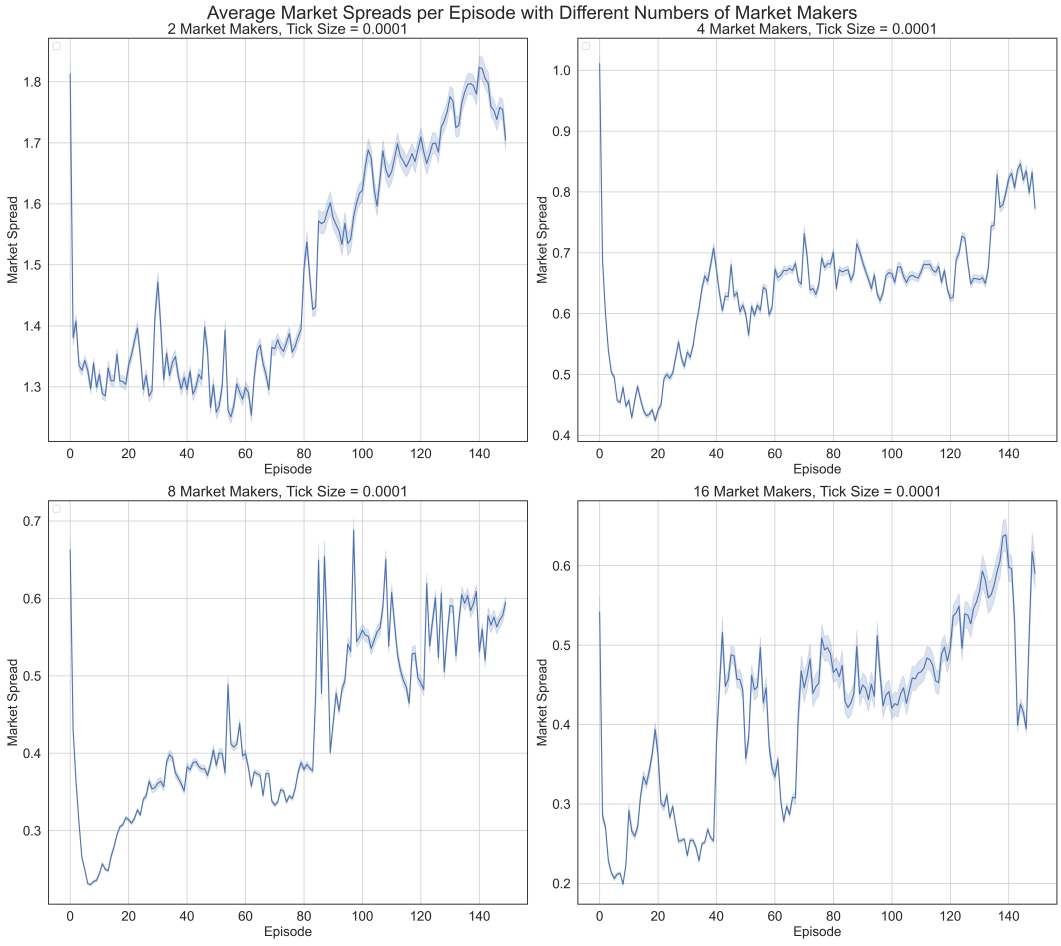


Figure 2.4: Market spreads during the training steps with different numbers of market makers, tick size $\vartheta = 0.0001$

After the training is complete, we visualize the learned quoting strategies

$$\pi_i(x, m|\theta_i^\pi)$$

as a function of the previous-step quoted spread and the previous-step market spread in Figure 2.5. The learned quoting strategy is calculated by averaging the trained actors π_i from 50 independent experiments, on a 2D grid of $[0, 5] \times [0, 5]$.[2] Note that market spread is the minimum of all market makers' quoted spreads; hence, it is the diagonal and lower triangular part of the heat maps that the learned quoting strategies apply to generate next-step quotes.

In the lower triangular regions of the graphs in Figure 2.5, we first observe that market makers tend to quote higher spreads when the market spread is close to 0, as indicated by the lighter vertical bands near the left edges of each graph. This behavior suggests that when the market spread is pretty low close to 0, market makers are less aggressive in lowering their quotes. Rather than reducing their spreads to compete for order flow, they maintain higher spreads for the next step to avoid sacrificing profit margins, as their chances of winning the order flow are already diminished under such conditions with very narrow market spreads. Secondly, we observe that the vertical color changes are subtler than the horizontal color changes, indicating that the learned strategy is more sensitive to changes in the market spread than to its previous-step quoted spread. When the market spread changes, the strategy adjusts the next-step quotes more dramatically. However, when the market spread is fixed, the strategy exhibits less variation in response to changes in the agent's current quoted spread. This suggests that the market conditions affected by competitors play a more significant role in determining the quote adjustments by the algorithm than the agent's prior quoting behavior. Furthermore, the ray-like bands extending from $(0, 0)$ in the graphs indicate that the market makers tend to quote similar values when the previous-step quoted spread and the previous-step market spread change proportionally. This suggests that the learned strategies follow a pattern of consistent adjustments, aligning their quoted spreads when both variables change together. All these patterns highlight the sophisticated nature of the learned quoting strategies, where the algorithms respond to short-term fluctuations in both their own quoting behavior and the market spread, rather than simply lowering spreads to gain immediate market share, especially in such a highly competitive 'winner-takes-all' market configuration.

[2] Although the observation space in learning simulation is $[0, 10]$, spread values beyond 5 is not of our interest because these values are not competitive given the market configuration. Hence, we choose to visualize the quoting strategy with a 2D grid of $[0, 5] \times [0, 5]$

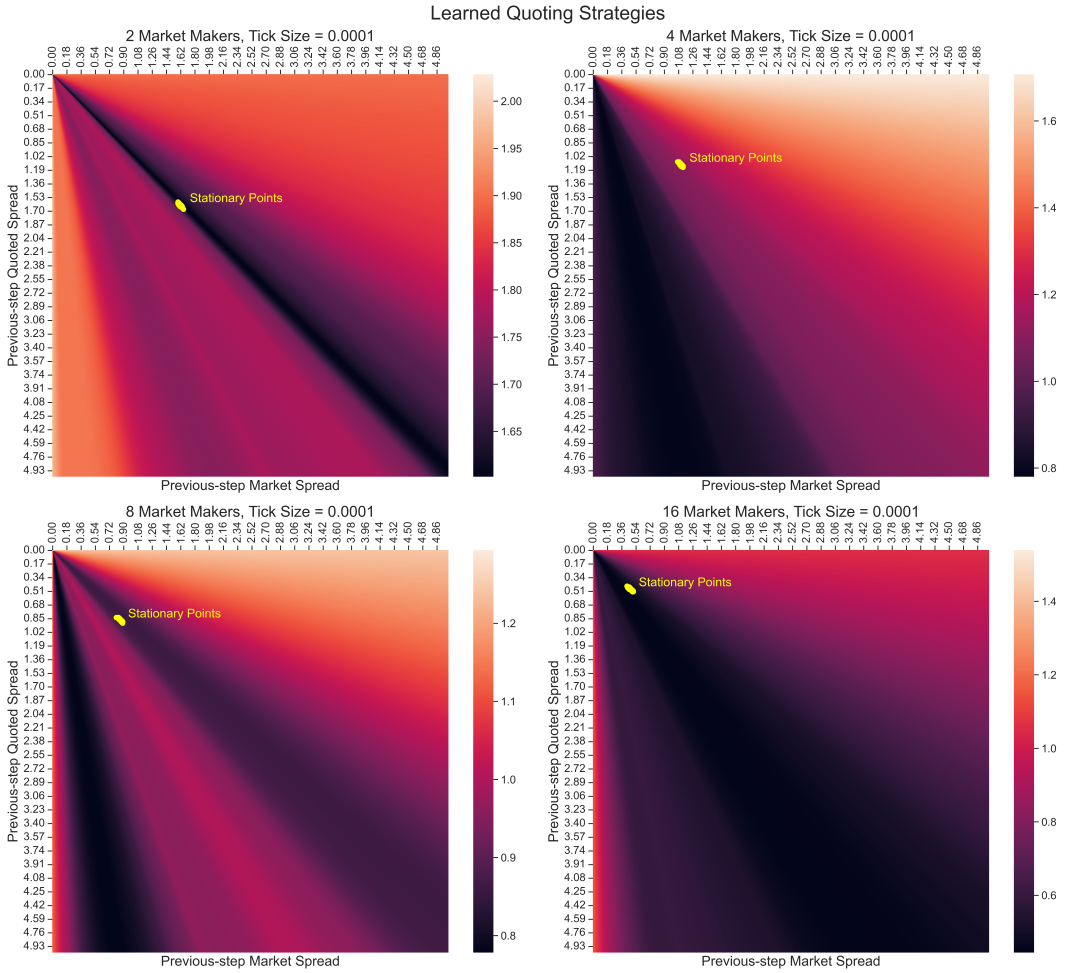


Figure 2.5: Learned strategies with different numbers of market makers.

In addition to the observed patterns above, our results also show that when market makers apply their learned strategies in a market environment through repeated pricing games, the spreads they quote eventually converge to specific stationary points, as shown by the yellow points in Figure 2.5. These stationary points represent stable market spreads that the learned quoting strategies consistently reach. Figure 2.6 shows the results of repeated pricing games with varying numbers of market makers, all of whom start with an initial spread of 1.0000. These results are compared with the Competitive Nash Equilibrium spread $P_\vartheta^+(\tilde{s}_1)$ and the maximum of the revenue function s:

$$P_\vartheta^+(\tilde{s}_1) = 0.1027, \quad s^* \approx 2.5569$$

In the experiment of repeated pricing games, we observe that the stationary spread achieved is independent of the initial spread values selected by the

market makers. This suggests that the learned strategies are robust, leading to universal stationary market spreads regardless of starting conditions. The specific stationary spread values obtained in different scenarios are presented in Table 2.1. Note that these stationary spread values are obtained based on the interactions of the quoting strategies learned after the training is completed, which are different from the market spreads during the training steps shown in Figure 2.4.

Number of agents	2	4	8	16
Stationary market spread	1.6165	1.1187	0.8651	0.4937

Table 2.1: Stationary spreads with different numbers of market makers, tick size $\vartheta = 0.0001$.

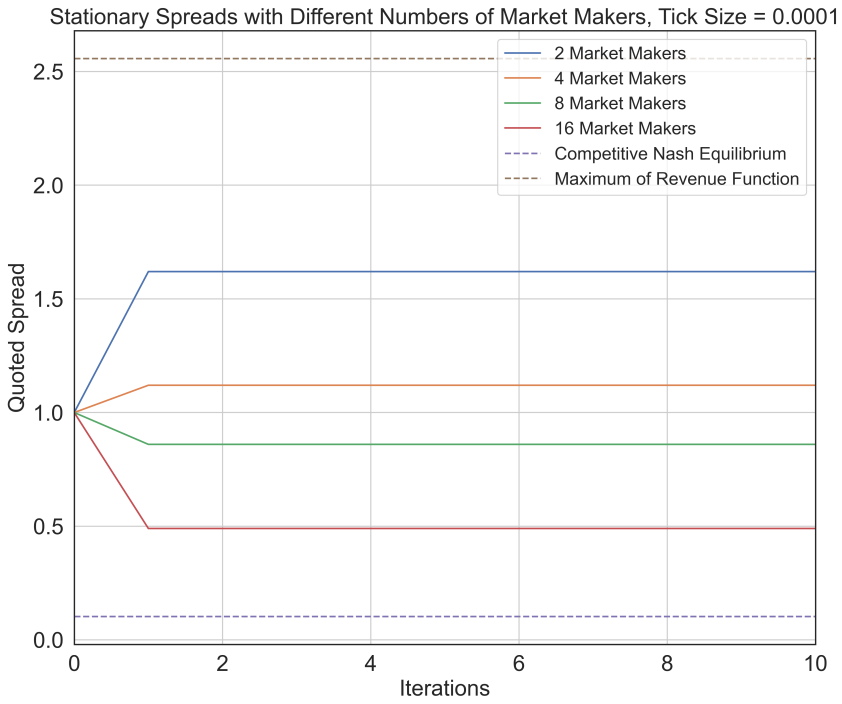


Figure 2.6: Stationary spreads reached by the learned quoting strategies with different numbers of market makers. In this example, the learned quoting strategies all start with initial spreads 1.000.

A key observation from Table 2.1 and Figure 2.6 is that as the number of market makers increases, the stationary market spread decreases. When the number of market makers increases to 16, the stationary market spread is the closest to the Competitive Nash Equilibrium level. This suggests that

increased competition with more market makers encourages the learning algorithms to quote lower spreads, as they learn to compete with each other to win the order flow. With more market makers, the learning algorithms lead to more competitive outcomes compared to the case of 2 market makers. However, the stationary spreads achieved by the learning algorithms are all above the Competitive Nash Equilibrium level despite the decreased values with increasing numbers of market makers. In particular, the stationary spreads for 2 and 4 market makers are significantly higher than the competitive outcome. In other words, seemingly collusive results have been achieved by the learning algorithms without sharing information. This phenomenon is called tacit collusion, featured by pricing algorithms maintaining supra-competitive outcomes. Overall, the learning algorithms can lead to outcomes that suggest tacit collusion, especially when the number of market makers is small. As competition increases with more market makers, the learning algorithms converge to more competitive levels. However, even with 16 market makers, the spread remains above the Competitive Nash Equilibrium level, indicating that full competition may not be entirely achieved.

We provide further evidence of tacit collusion by illustrating a mechanism that resembles a reward-punishment scheme, where deviations from supra-competitive spread levels are implicitly discouraged, leading to the reestablishment of supra-competitive spreads. After allowing the learning algorithms to converge to a stationary market spread in a repeated game setup starting from randomized initial spreads, we introduce a downward perturbation to the quoted spread of the market maker indexed by 1 while holding the quoted spreads of other market makers constant. This perturbation forces the market spread to decrease temporarily. Figure 2.7 shows the behavior of the market makers' quoting strategies with this perturbation. In the case of 4 and 8 market makers, the other market makers lower their quoted spreads following this perturbation, after which the system reverts back to the stationary market spread levels. The collective behavior of the other market makers' simultaneously lowering their spreads can be interpreted as a mechanism of punishment where they decrease their spreads and launch a price war. In the case of 2 and 16 market makers, although the other market makers initially react by slightly increasing their quoted spreads, the learned strategies anticipate a rise in the market spread at the next step, thereby reducing their quoted spreads to win order flow. Consequently, the overall system also returns to the original stationary spread level. Therefore, this dynamic following a downward

perturbation of one market maker mimics a tacit reward-punishment scheme. Although there is no explicit punitive action in some cases, such as aggressive lowering of spreads by the competitors, the system's dynamics discourage deviations and reinforce the collusive equilibrium.

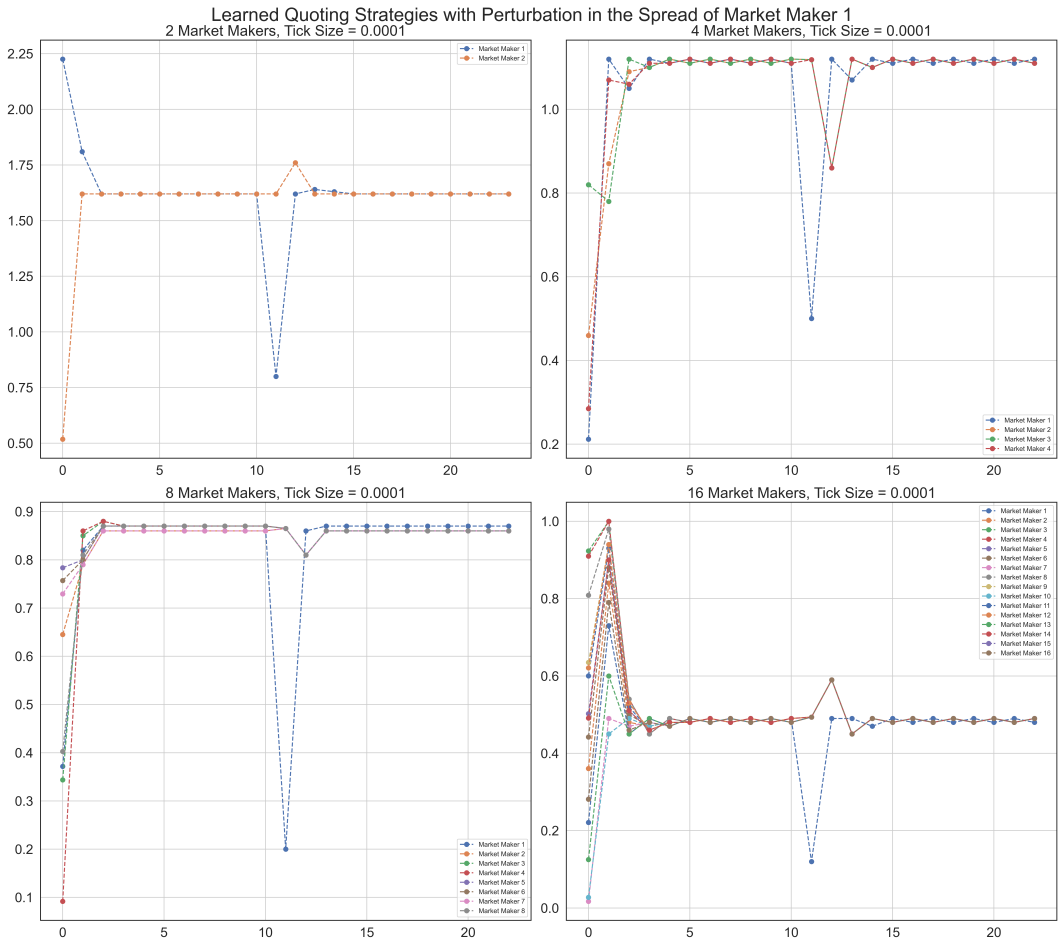


Figure 2.7: Stationary spreads reached by the learned quoting strategies with different numbers of market makers. In this example, the learned quoting strategies all start with initial spreads 1.000.

In summary, our results demonstrate that the spreads quoted by the learning algorithms converge to stationary levels when the algorithms are applied for pricing. These stationary spreads decrease with increasing competition brought by more market makers, approaching but not fully reaching the Competitive Nash Equilibrium. Notably, when fewer market makers are present, the spreads remain significantly above the competitive level, indicating supra-competitive outcomes that mimic tacit collusion, in which a mechanism similar

to the reward-punishment scheme is present in the interactions of the pricing algorithms.

2.2.3 Experiment with Increased Number of Market Makers

We have seen that increasing the number of market makers resulted in lower stationary spreads, gradually approaching the Competitive Nash Equilibrium level. In this section, we extended our analysis to 32 market makers to explore whether further increasing competition would lead to more competitive outcomes, as observed in our previous experiments with 2, 4, 8 and 16 market makers. We carry out one experiment of learning simulation with 32 market makers and present the reward curves and the learned quoting strategy in Figure 2.8. Interestingly, as shown in Figure 2.8, the results show that increasing the number of market makers does not lead to a significantly more competitive market. In fact, the stationary market spread with 32 participants is at a closer level as observed with only 2 market makers, as shown in Table 2.2.

Number of agents	32
Stationary market spread	1.4758

Table 2.2: Stationary spreads with 32 market makers, tick size $\vartheta = 0.0001$.

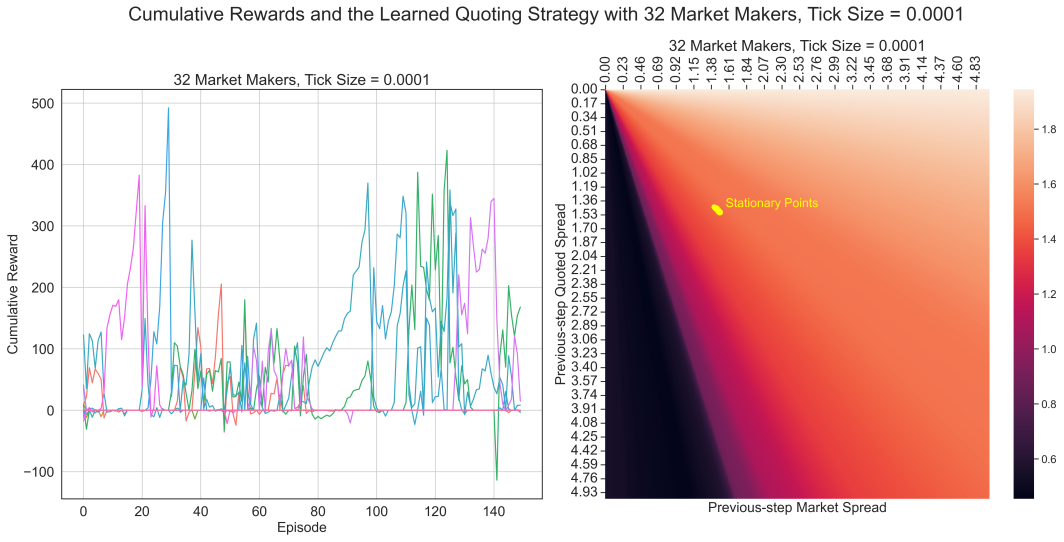


Figure 2.8: Cumulative rewards per episode and the learned quoting strategy with 32 market makers, tick size $\vartheta = 0.0001$, from 1 experiment.

The cumulative reward curve shows that, despite the large number of market makers, only a few effectively learn and achieve non-negative rewards. The majority of the market makers do not secure a consistent share of the order flow, as shown by their negative or near-zero cumulative rewards. This indicates that the learning process in a highly competitive environment becomes more challenging, with only a small subset of market makers able to successfully adapt their strategies and earn positive rewards. The 'winner-takes-all' nature of the market intensifies as more market makers enter, resulting in many algorithms failing to learn effective quoting strategies.

This outcome creates a dynamic similar to the scenario with fewer market makers. In this case, only a few participants dominate the market and achieve sustained positive rewards, while the rest either fail to compete effectively or are driven out by the dominant players. This limited number of effective learners contributes to maintaining supra-competitive spreads.

The results suggest that having more market makers does not necessarily drive the market toward fully competitive outcomes. With 32 participants, the competition becomes so intense that only a handful of market makers can effectively learn and compete, while the rest are left behind. This creates a scenario where, despite the high number of competitors, the market remains concentrated in the hands of a few effective players.

Under this 'winner-takes-all' mechanism, the market does not benefit from the theoretical advantages of increased competition. Instead, the failure of most market makers to learn efficient strategies limits the overall competitiveness of the market, leading to a spread level that is close to the case with only 2 market makers. This outcome challenges the expectation that more algorithmic participants naturally lead to better pricing outcomes, demonstrating that intense competition between algorithms can result in a concentrated, non-competitive market when the learning process is constrained by the 'winner-takes-all' mechanism.

2.2.4 Experiments with Other Tick Sizes

In this section, we test the impact of larger tick sizes on the behavior of learning algorithms, to compare with previous experiments in which the tick size is set at 0.0001. Specifically, we analyze the experiments with 2 market makers in various tick sizes, including $0.01, 0.1, 1$, and with 4 market makers in a tick size of 1.

The cumulative reward curves show that as the tick size increases, the fluctuations in the reward values reduce significantly. This is because with larger tick sizes, the room for fine-tuning the spreads by the learning algorithms becomes limited, and the agents quickly converge to stable spread levels. For example, when the tick size is set to 1, both the 2 and 4 market makers' learning algorithms stabilize at a stationary market spread of 1, which coincides with the corresponding Competitive Nash Equilibrium.

The larger tick size limits the flexibility of learning algorithms to adjust their quoted spreads, hence constraining their ability to deviate from equilibrium behavior. We consider this experiment with larger tick sizes as a validation test for the learning algorithms, demonstrating that when the space for adjustment is restricted, the algorithms can maintain the spreads near the corresponding Competitive Nash Equilibria instead of augmenting their spreads to higher levels, which could have brought them higher profits. It is worth noting that in real MD2C platforms such large tick sizes are rare, as most assets are traded with more granular tick sizes. Our previous experiments with a tick size of 0.0001 are more representative of real-world scenarios, where the quasi-continuous nature of the smaller tick size offers more flexibility to adjust spreads. Compared to previous experiments with tick size $\vartheta = 0.0001$, we conclude that tacit collusion is more likely to occur in real-world scenarios when tick sizes are relatively small, as increased flexibility makes learning algorithms more able to fine-tune their spreads and maintain supra-competitive pricing strategies. We attribute the competitive outcome which arises from the settings with larger tick size to the competition model that has a higher Competitive Nash Equilibrium driven by the larger tick sizes. We believe that it is the nature of the less realistic underlying model, not the learning algorithms, that seemingly mitigates the tacit collusion found in the more realistic contexts with smaller tick sizes.

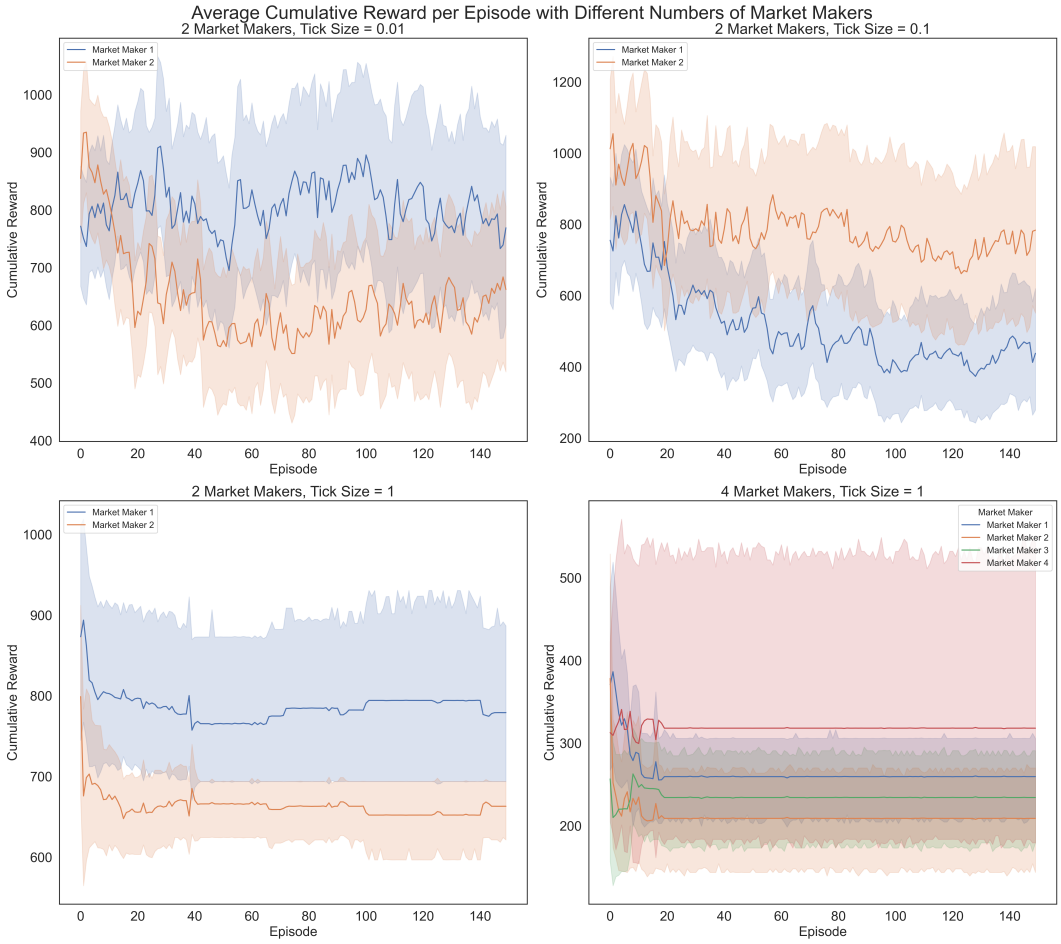


Figure 2.9: Average cumulative reward per episode during training with different numbers of market makers, with larger tick sizes, from 50 scenarios

Number of agents	2	2	2	4
Tick size	0.01	0.1	1	1
Stationary market spread	1.52	1.3	1	1
Competitive Nash Equilibrium	0.11	0.2	1	1

Table 2.3: Stationary spreads with different numbers of market makers and other tick sizes.

2.3 Implications for Market Regulation

The results of our experiments provide several important insights for regulators and policy makers concerned with the possibility of the emergence of tacit

collusion from learning algorithms applied in market making. As demonstrated in our numerical simulations, learning algorithms frequently sustain supra-competitive spread levels, even when the market structure should encourage competition with more competing market makers.

An important regulatory implication involves the concept of best execution, which currently focuses on due diligence at the client-dealer level, ensuring that brokers and dealers seek the best possible terms for their clients ([Financial Industry Regulatory Authority 2014; Securities and Exchange Commission 2023; European Securities and Markets Authority 2018]). This requirement typically involves monitoring the quality of trade executions and comparing competing market venues to ensure optimal pricing for client orders. However, these rules are currently applied to human decision making by brokers rather than to the algorithms used by market makers. Although brokers are required to ensure the best execution for client orders, these regulations lack specific guidance for best execution by the market making algorithms, especially the learning algorithms that can potentially be used by market makers. This could allow tacit collusion or supra-competitive spreads to emerge without regulatory supervision. In our experiments, the best execution is linked to the spread set by the learning algorithms, where the best price for clients should ideally reflect the Competitive Nash Equilibrium $P^{+}_{\vartheta}(\tilde{s}_1)$. However, our results show that this equilibrium is rarely achieved. If regulators aim to address potential tacit collusion in market making algorithms, specific best execution rules should extend beyond client-dealer interactions and include clear guidelines for determining the best possible market spreads. This would require precise reporting of costs (c) by market makers, and platforms should monitor order flow functions to accurately calculate and enforce spreads that align with the Competitive Nash Equilibrium.

The implications extend beyond simply setting price benchmarks for best execution. In our experiments, as we increase the number of market makers from 2 to 16, the market spread consistently decreases, suggesting that more intense competition between the learning algorithms leads to more competitive pricing. However, once we increase the number of market makers to 32, we observe that market spreads rise again, closely resembling the levels seen with 2 market makers. This outcome suggests that, while moderate competition leads to more competitive outcomes, having too many market makers interacting through learning algorithms may reduce the effectiveness of competition. From a regulatory perspective, this implies that MD2C platforms

should consider policies that encourage interactions between a sufficient number of market makers but avoid overcrowding, where inefficiencies could arise. For example, platforms could be required to send requests for quotes to a defined number of market makers for each client trade. Forcing multiple simultaneous requests to a reasonable number of market makers (for example, 8 to 16 in our simulation) could drive the market spreads closer to Competitive Nash Equilibrium levels, as observed in our experiments. However, regulators should avoid pushing this number too high, as observed in the case of 32 market makers, where learning algorithms maintain higher market spread levels. Thus, establishing an upper bound on the number of active algorithmic market makers who can simultaneously receive requests for quotes could discourage the learning algorithms from inefficient or collusive behaviors while ensuring sufficient competition for more favorable prices to clients.

Furthermore, regulators could also impose stricter transparency rules that require disclosure of algorithmic market making strategies and their impact on spreads, which would make it easier for both market participants and regulators to identify patterns that may indicate anti-competitive behavior. Additionally, regulators could consider implementing mandatory auditing of market making algorithms, particularly those using reinforcement learning or other AI-driven techniques. The audit could involve routinely testing the algorithms under various market conditions and tick sizes to evaluate whether the interactions of algorithms adheres to competitive principles. Moreover, platforms could be required to monitor real-time market activity and identify patterns of unusually stable spreads, which could potentially signal anti-competitive equilibrium (similar to the absence of odd-eighth spreads in NASDAQ suggested by [Christie and Schultz 1994]). This ensures that the platforms are held accountable for the behaviors of market makers to promote a competitive environment.

2.4 Summary

We have proposed a game-theoretic model for a dealer market with multiple market makers, where the process by which pricing algorithms learn from market data is modelled through a decentralized multi-agent reinforcement learning algorithm. Theoretically, we outline the existence of a best possible market spread at the Competitive Nash Equilibrium. Our model allows to identify configurations in which tacit collusion among algorithms emerges, the

main insight being that this depends on the level of competition among market makers.

Through our experiments with varying numbers of market makers and small tick size $\vartheta = 0.0001$, we have observed that as the number of market makers increases, the market spreads generally decrease, but still remain above the Competitive Nash Equilibrium level. This indicates that more intense competition derives more competitive pricing, but does not lead to fully competitive outcomes. However, when the number of market makers reaches 32, the market spreads unexpectedly increase again, highlighting the potential inefficiencies of learning that can arise in overcrowded markets.

These results may be extended in various directions. Although the results presented above assume a symmetric configuration with identical market makers, these results are readily extendable to the case where market makers differ in market share functions. One may also consider more sophisticated mechanisms for the allocation of market share, as well as the impact of regulatory intervention. Finally, there is scope to consider a dynamic version of this model in the framework of stochastic differential games, as in [Cont, Guo, and Xu 2021]. We address this extension in Chapter 3.

Chapter 3

Dynamics of Market Making Algorithms: Learning and Tacit Collusion

This chapter is based on [Cont and Xiong 2024]. Section 3.1 describes our model setting for a continuous-time dealer market with competition among market makers, formulated as a stochastic differential game of intensity control. Section 3.2 describes the competition among dealers in terms of a Nash equilibrium. Section 3.3 describes collusion among dealers and establishes the connection between collusion and Pareto optima. In Section 3.4, we describe a fictitious play algorithm to numerically calculate Nash equilibria. In Section 3.5 we describe how the learning dynamics of market makers can be modelled using decentralized multi-agent deep reinforcement learning and its implementation using a Decentralized Multi-Agent Deep Deterministic Policy Gradient (Decentralized MADDPG) algorithm and present simulation evidence for tacit collusion.

3.1 A Continuous-time Dealer Market with Multiple Market Makers

We propose a continuous-time dynamic model of a dealer market with competing market makers, formulated as a stochastic differential game of *intensity control* ([Bremaud 1981]) with partial information. In the following, we will interchangeably use the terms 'market maker', 'dealer', and 'agent'.

We consider a market with a single asset and N market makers. The

market price of the asset is modelled by a Brownian motion

$$S_t = S_0 + \sigma W_t \tag{3.1.1}$$

where $(W_t)_{t\geq 0}$ is a standard Brownian motion on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ satisfying usual conditions, representing the flow of market information.

The order flow arrives in the form of a Request for Quotes (RFQ). Clients send RFQs to all market makers who respond by proposing bid/ask quotes around the market price. Market maker i quotes an ask price $S_t^{a,i}$ and a bid price $S_t^{b,i}$ where

$$S_t^{a,i} = S_t - \varrho_t^{a,i} \qquad S_t^{b,i} = S_t + \varrho_t^{b,i} \tag{3.1.2}$$

We refer to $\varrho_t^{a,i}$ and $\varrho_t^{b,i}$ as the centered ask and bid quotes.

We model the number of buy/sell order flows as *càdlàg* point processes $N^a(dt)$ and $N^b(dt)$ with constant arrival intensity λ^a and λ^b, and each order has a constant order size Δ.[1]

The market makers' quotes depend on their inventory, which we denote by q_t^i. As orders arrive with aggregate size multiples of Δ within a given time interval, the inventory q_t^i takes discrete values, multiples of Δ and is generally subject to limits. We therefore assume that the inventory of the market maker i takes values in $\mathcal{Q}_i = \{-Z_i, -Z_i + \Delta, \ldots, Z_i - \Delta, Z_i\}$. Therefore, there are $1+2Z_i/\Delta$ possible values for q_t^i. Note that we impose inventory limits not only for considering practical market making scenarios, but also for mathematical reasons of boundedness ([Guéant 2017; Guéant, Lehalle, and Fernandez-Tapia 2013]). Inventory limits are essential to prove the boundedness of the objective function in Proposition 3.2.2, which will be used in the proof of Theorem 3.2.6. It would be mathematically challenging to obtain the existence of the Nash equilibrium if the inventory were unbounded.

[1]Note that the constant order size assumption could be relaxed with an order size distribution. In this case the system of coupled Hamilton-Jacobi equations will be incorporated with an integral term over the order size space. [Bergault and Guéant 2021] and [Barzykin, Bergault, and Guéant 2023] study such an extension and prove the existence and uniqueness of a classical and viscosity solution to the HJB equation. But for our work the extension will impose challenges on numerical simulation and the design of reinforcement learning algorithm, with higher approximation error due to numerical integral and much longer time for the learning algorithm to explore on order size space to converge. We shall leave this extension for future research and focus on current constant order size setting throughout the paper.

Throughout this chapter, we consider only stationary problems in an infinite time horizon to align with the settings used in the learning algorithm simulation. Therefore, the *quoting strategy* for the market maker i can then be represented as a map $\delta^i : \mathcal{Q}_i \mapsto \mathbb{R}$ that we can represent as a vector $\boldsymbol{\delta}^i = (\vec{\delta}^{a,i}, \vec{\delta}^{b,i})$ with components indexed by inventory levels in $\mathcal{Q}_i$:

$$\vec{\delta}^{a,i} = \left(\delta^{a,i}_{q_i}\right)_{q_i \in \mathcal{Q}_i}, \quad \vec{\delta}^{b,i} = \left(\delta^{b,i}_{q_i}\right)_{q_i \in \mathcal{Q}_i} \tag{3.1.3}$$

We use the symbol δ to represent the quoting strategy vector, whereas in (3.1.2), the symbol ϱ represents the process of ask and bid quotes. To avoid confusion, we clarify that ϱ_t will always refer to the stochastic process of quoting strategies, and δ_q will denote the coordinate of the quoting strategy vector $\vec{\delta}$ indexed by inventory q. For clarity, we use bold symbols to represent the collection of vectors, where each vector is a quoting strategy of a market maker.

- The centered ask and bid quotes $\vec{\delta}^{a,i}, \vec{\delta}^{b,i}$ from the market maker i are vectors in $\mathbb{R}^{2\frac{Z_i}{\Delta}+1}$ with each coordinate corresponding to the quote at a specific inventory level $q_i \in \{-Z_i, -Z_i + \Delta, \ldots, Z_i - \Delta, Z_i\}$. For convenience, we directly index the coordinates of these vectors by inventory levels. For example $\delta^{a,i}_{q_i}, \delta^{b,i}_{q_i}$ are, respectively, $(\frac{Z_i+q_i}{\Delta}+1)-th$ coordinates of $\vec{\delta}^{a,i}, \vec{\delta}^{b,i}$, where they represent centered ask/bid quotes by the market maker i when her inventory level is q_i.

- $\boldsymbol{\delta}^i = (\vec{\delta}^{a,i}, \vec{\delta}^{b,i})$ denote the quoting strategy of market maker i. Hence $\boldsymbol{\delta}^i \in \mathbb{R}^{2\frac{Z_i}{\Delta}+1} \times \mathbb{R}^{2\frac{Z_i}{\Delta}+1}$.

- $\boldsymbol{\delta}^{a,-i}, \boldsymbol{\delta}^{b,-i}$ are collections of ask and bid quoting strategies of all market makers *excluding* i. They include $N-1$ vectors, each $\mathbb{R}^{2\frac{Z_j}{\Delta}+1}$ valued corresponding to the quoting strategy of market maker $j \neq i$. These vectors are sorted by market maker's index. Namely $\boldsymbol{\delta}^{a,-i} = (\vec{\delta}^{a,j})_{j=1,\ldots,i-1,i+1,\ldots,N} \in \prod_{j\neq i} \mathbb{R}^{2\frac{Z_j}{\Delta}+1}$, for $\boldsymbol{\delta}^{b,-i}$ and vice versa. We denote by $\boldsymbol{\delta}^{-i} = (\boldsymbol{\delta}^{a,-i}, \boldsymbol{\delta}^{b,-i})$ the quoting strategies of the competitors of the market maker i'. For simplicity of notation, we use $\vec{\delta}^{-i}$ to denote either $\boldsymbol{\delta}^{a,-i}$ or $\boldsymbol{\delta}^{b,-i}$ as variables in functions f^i_a and f^i_b below.

The probability that the RFQ gets executed against a market maker depends on their quotes and those of other competing market makers. We model

this probability through a pair of functions f_a^i and f_b^i representing the dependence of execution probabilities for market maker i on market makers' quotes:[2]

$$f_a^i : \mathbb{R} \times \prod_{j \neq i} \mathbb{R}^{2\frac{Z_j}{\Delta}+1} \to \mathbb{R}^+, \qquad f_b^i : \mathbb{R} \times \prod_{j \neq i} \mathbb{R}^{2\frac{Z_j}{\Delta}+1} \to \mathbb{R}^+$$

satisfying

$$\sum_{i=1}^{N} f_a^i(\delta, \vec{\delta}^{-i}) \leq 1, \qquad \sum_{i=1}^{N} f_b^i(\delta, \vec{\delta}^{-i}) \leq 1.$$

The inequality corresponds to the possibility that an RFQ is not executed by any market maker with probability $(1 - \sum_{i=1}^{N} f_a^i)$ (resp. $(1 - \sum_{i=1}^{N} f_b^i)$), because the clients are not satisfied with the quotes or do not have any intention to trade. We will specify assumptions on execution probabilities in more detail in the following.

We consider stationary feedback quoting strategies $\vec{\delta}^{a,i}, \vec{\delta}^{b,i}$: $\vec{\delta}^{a,i}, \vec{\delta}^{b,i}$ represented as $\mathbb{R}^{2\frac{Z_i}{\Delta}+1}$-valued vectors, in which each coordinate $\delta_{q_i}^{a,i}$ corresponds to the centered quote at inventory level $q_i \in \mathcal{Q}_i$. Thus, at time t the market maker i will quote

$$\varrho_t^{a,i} = \delta_{q_{t-}^i}^{a,i}, \quad \varrho_t^{b,i} = \delta_{q_{t-}^i}^{b,i}$$

The market maker i only quotes prices when her inventory does not exceed the limits $\pm Z_i$. The (ask/bid) order flow executed by the market maker i may then be represented as a pair of point processes $N^{a,i}(dt)$ and $N^{b,i}(dt)$ with intensity

$$\begin{aligned} \nu_t^{a,i} &= \lambda^a f_a^i\Big(\delta_{q_{t-}^i}^{a,i}, \boldsymbol{\delta}^{a,-i}\Big)\mathbb{I}(q_{t-}^i > -Z_i) \\ \nu_t^{b,i} &= \lambda^b f_b^i\Big(\delta_{q_{t-}^i}^{b,i}, \boldsymbol{\delta}^{b,-i}\Big)\mathbb{I}(q_{t-}^i < Z_i) \end{aligned} \tag{3.1.4}$$

The ask price $S_t^{a,i}$ is always above the bid price $S_t^{b,i}$. However, there are circumstances where either $\delta^{a,i}$ or $\delta^{b,i}$ could be negative, usually when the market maker i quotes very aggressively so that her quoted ask price stays below the market mid price, or her quoted bid price is above the mid price. This happens when the market maker holds a non-zero inventory and is eager to attract the order flow in the opposite direction of the inventory to reduce her inventory risk. We assume that the centered quotes are constrained by

[2]The notation $\vec{\delta}^{-i}$, used in the execution probabilities f_a^i and f_b^i, represents the quoting strategies of other market makers: $\vec{\delta}^{-i} \in \prod_{j \neq i} \mathbb{R}^{2\frac{Z_j}{\Delta}+1}$.

a limit, that is $\delta^{a,i} \geq -\delta_\infty, \delta^{b,i} \geq -\delta_\infty$ where $\delta_\infty > 0$ is a given constant.[3] Therefore, the admissible strategy space of each market maker i is

$$\mathcal{A}_i = \Bigg\{ \boldsymbol{\varrho}_t^i = \left(\delta^{a,i}_{q_{t-}^i}, \delta^{b,i}_{q_{t-}^i}\right) \Big| \delta^{a,i}, \delta^{b,i} \in \mathbb{R}^{2^{\frac{Z_i}{\Delta}+1}}; \\ \delta^{a,i}_{q_i} \geq -\delta_\infty, \delta^{b,i}_{q_i} \geq -\delta_\infty, \forall q_i \in \{-Z_i, \ldots, Z_i\} \Bigg\} \tag{3.1.5}$$

We let I_δ denote the interval $[-\delta_\infty, \infty)$. $\mathcal{A}_i$ contains stochastic processes $\boldsymbol{\varrho}_t^i = \left(\delta^{a,i}_{q_{t-}^i}, \delta^{b,i}_{q_{t-}^i}\right)$ that are feedback strategies depending on the inventory q_{t-}^i, while $(I_\delta)^{2^{\frac{Z_i}{\Delta}+1}} \times (I_\delta)^{2^{\frac{Z_i}{\Delta}+1}}$ is the space of possible values of market makers' quoting strategies.

We use the following notations for partial derivatives: for $i \in \{1, \ldots, N\}$ and $j \neq i$ we denote

$$\begin{aligned}
\partial_1 f_a^i &= \frac{\partial f_a^i}{\partial \delta}(\delta, \vec{\delta}^{-i}), & \partial_1 f_b^i &= \frac{\partial f_b^i}{\partial \delta}(\delta, \vec{\delta}^{-i}) \\
\partial_{11}^2 f_a^i &= \frac{\partial^2 f_a^i}{\partial \delta^2}(\delta, \vec{\delta}^{-i}), & \partial_{11}^2 f_b^i &= \frac{\partial^2 f_b^i}{\partial \delta^2}(\delta, \vec{\delta}^{-i}) \\
\partial_{j,q_j} f_a^i &= \frac{\partial f_a^i}{\partial \delta^j_{q_j}}(\delta, \vec{\delta}^{-i}), & \partial_{j,q_j} f_b^i &= \frac{\partial f_b^i}{\partial \delta^j_{q_j}}(\delta, \vec{\delta}^{-i}) \\
\partial_{j,q_j}\partial_1 f_a^i &= \frac{\partial^2 f_a^i}{\partial \delta^j_{q_j}\partial \delta}(\delta, \vec{\delta}^{-i}), & \partial_{j,q_j}\partial_1 f_b^i &= \frac{\partial^2 f_b^i}{\partial \delta^j_{q_j}\partial \delta}(\delta, \vec{\delta}^{-i})
\end{aligned} \tag{3.1.6}$$

Note that the symbol ∂_{j,q_j} represents the first-order derivative with respect to the coordinate $\delta^j_{q_j}$ in $\vec{\delta}^{-i}$.

We make the following assumptions on f_a^i and f_b^i.

Assumption 3.1.1. f_a^i, f_b^i are twice continuously differentiable on $\mathbb{R} \times \prod_{j \neq i} \mathbb{R}^{2^{\frac{Z_j}{\Delta}+1}}$ and satisfy

$$f_a^i(\delta, \vec{\delta}^{-i}) > 0, f_b^i(\delta, \vec{\delta}^{-i}) > 0, \quad \sum_{i=1}^N f_a^i(\delta, \vec{\delta}^{-i}) \leq 1, \quad \sum_{i=1}^N f_b^i(\delta, \vec{\delta}^{-i}) \leq 1$$

There exists a function $\Lambda(\delta) \in C^2(\mathbb{R})$, such that for $m \in \{a, b\}$, $0 < f_m^i(\delta, \vec{\delta}^{-i}) < \Lambda(\delta)$, and

$$\lim_{\delta \to \infty} \Lambda(\delta)\delta = 0, \Lambda'(\delta) < 0, \Lambda(\delta)\Lambda''(\delta) \leq 2(\Lambda'(\delta))^2$$

[3]The lower bound $-\delta_\infty$ is not only a practical concern but is also mathematically necessary for constraining the upper bound of intensity function as we will see in Assumption 2.1. This boundedness will be applied to validate Itô's formula on the objective function in (3.1.16). Without the lower bound δ_∞, it would be challenging to validate the application of Itô's formula.

Remark 3.1.2. $\Lambda(\delta)$ in Assumption 3.1.1 can be understood as the execution probability in the case of the single market maker as in [Guéant 2017]. The assumption corresponds to the intuition that competition lowers the execution rate for each market maker.

Assumption 3.1.3. The execution probabilities $\{f_m^i\}_{m\in\{a,b\}}$ satisfy: $\forall(\delta, \vec{\delta}^{-i}) \in \mathbb{R} \times \prod_{j\neq i} \mathbb{R}^{2\frac{Z_j}{\Delta}+1}, \forall j \neq i, \forall q_j \in \mathcal{Q}_j$,

$$\begin{aligned}
&\partial_1 f_m^i < 0, \partial_{j,q_j} f_m^i \geq 0, \qquad \frac{\partial_{11}^2 f_m^i \cdot f_m^i}{(\partial_1 f_m^i)^2} < 2 \\
&2(\partial_1 f_m^i)^2 - \partial_{11}^2 f_m^i \cdot f_m^i - \sum_{j\neq i} \sum_{q_j\in\mathcal{Q}_j} \left|\partial_1 f_m^i \cdot \partial_{j,q_j} f_m^i - f_m^i \cdot \partial_{j,q_j}\partial_1 f_m^i\right| > 0 \\
&\lim_{\delta\to+\infty} \frac{f_m^i(\delta, \vec{\delta}^{-i})}{\partial_1 f_m^i(\delta, \vec{\delta}^{-i})} < \infty, \qquad \forall \vec{\delta}^{-i} \in \prod_{j\neq i} \mathbb{R}^{2\frac{Z_j}{\Delta}+1}
\end{aligned} \tag{3.1.7}$$

Remark 3.1.4. Assumptions 3.1.1 and 3.1.3 are needed to prove the existence of the Nash equilibrium in Section 3.2. In Assumption 3.1.3 $\partial_1 f_m^i < 0, \partial_{j,q_j} f_m^i \geq 0$ corresponds to the monotonicity of the execution rate as functions of market maker and competitor quotes. Similarly to execution probabilities in [Guéant 2017] we need the function $\delta \to \delta \cdot f_m^i(\delta, \vec{\delta}^{-i})$ to reach a unique maximum on $[-\delta_\infty, \infty)$. Hence, we assume $\frac{\partial_{11}^2 f_m^i \cdot f_m^i}{(\partial_1 f_m^i)^2} < 2$, together with the assumption $\lim_{\delta\to+\infty} \frac{f_m^i}{\partial_1 f_m^i} < \infty$. The last two lines in (3.1.7) are the strongest conditions specifying the growth regularity of execution rate functions. These conditions are motivated by [Luo and Zheng 2021] but are generalized to the circumstance of N-player. This assumption is used in Proposition A.2.11 to prove the implicit function property of centered quotes $\boldsymbol{\delta}$ as a function of the value vector $\boldsymbol{p}$. Intuitively, the second line of (3.1.7) specifies the first variable δ, which is the market maker's own quote, dominates the growth of execution rate $f_m^i(\delta, \vec{\delta}^{-i})$ compared to competitors' quotes. The last line in (3.1.7) specifies that $\partial_1 f_m^i(\delta, \vec{\delta}^{-i})$ does not vanish too fast as a function of δ. An example is given in Proposition 3.1.5 that satisfies these assumptions.

Proposition 3.1.5. *The following function satisfies Assumptions 3.1.1 and 3.1.3.*

$$f_a^i(\delta, \vec{\delta}^{-i}) = \frac{1}{N} \frac{1}{1+e^{a\delta+b}} \frac{e^{\frac{1}{K_i}\sum_{j\neq i}\sum_{q_j\in\mathcal{Q}_j}(a\delta_{q_j}^j + b_{q_j}^j)}}{1+e^{a\delta+\frac{1}{K_i}\sum_{j\neq i}\sum_{q_j\in\mathcal{Q}_j}(a\delta_{q_j}^j+b_{q_j}^j)}} \tag{3.1.8}$$

where $a > 0, b \in \mathbb{R}, b_{q_j}^j \in \mathbb{R}$, *and* $K_i = \sum_{j\neq i, j\in\{1,\dots,N\}} (2\frac{Z_i}{\Delta}+1)$ *is the number of inventory levels of market maker* i*'s competitors.*

The proof is provided in Appendix A.1.

The specific form of the intensity function in (3.1.8) is chosen to both satisfy Assumption 3.1.1 and 3.1.3 from a mathematical perspective, and to align with economic intuition and micro-foundations. The first term $\frac{1}{1+e^{a\delta+b}}$ serves as an upper bound on the intensity, similar to the case of a single market maker. It takes a logistic form to model the probability that an RFQ is executed by the market maker, which reflects the decreasing likelihood as the market maker's own quote becomes less favorable. Similar logistic forms have been used in other single-agent market making models, such as in [Barzykin, Bergault, and Guéant 2021] and [Barzykin, Bergault, and Guéant 2024]. The second term can also be written as

$$\frac{1}{e^{-\frac{1}{K_i}\sum_{j\neq i}\sum_{q_j\in\mathcal{Q}_j}(a\delta^j_{q_j}+b^j_{q_j})}+e^{a\delta}} \tag{3.1.9}$$

which captures the coupling effects from the competition between the market maker i and her competitors. This term models the joint impact of market maker i's quote and the aggregate behavior of her competitors, where these effects act in opposite directions on the intensity. The average term

$$\frac{1}{K_i}\sum_{j\neq i}\sum_{q_j\in\mathcal{Q}_j}(a\delta^j_{q_j}+b^j_{q_j})$$

represents the aggregate influence of the competitors' quoting strategies on the execution probability for market maker i, which is an increasing function of the competitors' individual quotes. This term effectively smooths out extreme quotes by considering the average of competitors' quoting strategies, similar to a mean field effect, and also reflects a latency effect brought by competition. The coefficient a is kept the same for both the market maker i's quote and average term of the quoting strategies of her competitors, leading to the same absolute sensitivity value a of the market maker i and the aggregate effect of all competitors on the intensity function.

Market maker i's inventory q^i_t evolves according to

$$dq^i_t = \Delta(N^{b,i}(dt) - N^{a,i}(dt)) \tag{3.1.10}$$

and her cash holdings X^i_t evolve according to

$$\begin{aligned} dX^i_t &= \Delta(S_t+\varrho^{a,i}_t)N^{a,i}(dt) - \Delta(S_t-\varrho^{b,i}_t)N^{b,i}(dt) \\ &= \Delta\left(\varrho^{a,i}_t N^{a,i}(dt) + \varrho^{b,i}_t N^{b,i}(dt)\right) - S_t dq^i_t \end{aligned} \tag{3.1.11}$$

The objective of each market maker is to maximize the expected discounted profit. We consider the problem over an infinite horizon in order to study stationary feedback quoting strategies that are more readily accessible to reinforcement learning. Denoting $\boldsymbol{\delta}^i = (\vec{\delta}^{a,i}, \vec{\delta}^{b,i}), \boldsymbol{\delta}^{-i} = (\boldsymbol{\delta}^{a,-i}, \boldsymbol{\delta}^{b,-i})$, market maker i has objective function

$$\tilde{J}_i(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i}; q_0^i, x_0^i, s) = x_0^i + q_0^i s + \mathbb{E}\left[\int_0^\infty e^{-rt} d(X_t^i + q_t^i S_t) - \int_0^\infty e^{-rt}\psi_i(q_t^i)dt\right] \tag{3.1.12}$$

where $\psi_i : \mathbb{R} \to \mathbb{R}_+$ is the running cost for holding inventory.

Lemma 3.1.6. *The objective function in (3.1.12) can be written as:*

$$\tilde{J}_i(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i}; q_0^i, x_0^i, s) = x_0^i + q_0^i s + J_i(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i}; q_0^i) \tag{3.1.13}$$

where

$$\begin{aligned} J_i(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i}; q_0^i) =& \mathbb{E}^{q_0^i}\Bigg[\int_0^\infty e^{-rt}\big(\lambda^a \Delta \delta^{a,i}_{q^i_{t-}} f_a^i(\delta^{a,i}_{q^i_{t-}}, \boldsymbol{\delta}^{a,-i})\mathbb{I}(q^i_{t-} > -Z_i) \\ &+ \lambda^b \Delta \delta^{b,i}_{q^i_{t-}} f_b^i(\delta^{b,i}_{q^i_{t-}}, \boldsymbol{\delta}^{b,-i})\mathbb{I}(q^i_{t-} < Z_i)\big)dt - \int_0^\infty e^{-rt}\psi_i(q_t^i)dt\Bigg] \end{aligned} \tag{3.1.14}$$

Proof. Since q_t^i takes values from a finite set $\mathcal{Q}_i$, $\psi_i(q_t^i)$ is uniformly bounded for $t \geq 0$ and the expectation for the running cost term $\mathbb{E}[\int_0^\infty e^{-rt}\psi_i(q_t^i)dt] < \infty$ is well defined.

From Assumption 3.1.1 we have $f_a^i(\delta, \cdot) \leq \Lambda(\delta)$, $f_b^i(\delta, \cdot) \leq \Lambda(\delta), \forall \delta \in \mathbb{R}$, where $\Lambda(\delta)$ is a C^2 function. Given quoting strategies $(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i})$, since $\vec{\delta}^{a,i} = \left(\delta_q^{a,i}\right)_{q\in\mathcal{Q}_i}, \vec{\delta}^{b,i} = \left(\delta_q^{b,i}\right)_{q\in\mathcal{Q}_i}$ are vectors in space $\mathbb{R}^{2\frac{Z_i}{\Delta}+1}$, define $A := \sup_{q\in\mathcal{Q}_i}\left|\delta_q^{a,i}\right|, B := \sup_{q\in\mathcal{Q}_i}\left|\delta_q^{b,i}\right|$ and denote q_t^i market maker i's inventory process under quoting strategies $(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i})$. Recall that $\varrho_t^{a,i} = \delta^{a,i}_{q^i_{t-}}, \varrho_t^{b,i} = \delta^{b,i}_{q^i_{t-}}$, we also write $\boldsymbol{\varrho}_t^{a,-i}, \boldsymbol{\varrho}_t^{b,-i}$ to denote the quoting strategy process of other maker makers as functions of their corresponding inventory levels. We obtain

$$\begin{aligned} \int_0^\infty e^{-rt}\left(\lambda^a \varrho_t^{a,i} f_a^i(\varrho_t^{a,i}, \boldsymbol{\varrho}_t^{a,-i})\mathbb{I}(q^i_{t-} > -Z_i)\right)dt \leq A\Lambda(-\delta_\infty)\int_0^\infty e^{-rt}dt < \infty \\ \int_0^\infty e^{-rt}\left(\lambda^b \varrho_t^{b,i} f_b^i(\varrho_t^{b,i}, \boldsymbol{\varrho}_t^{b,-i})\mathbb{I}(q^i_{t-} < Z_i)\right)dt \leq B\Lambda(-\delta_\infty)\int_0^\infty e^{-rt}dt < \infty \end{aligned} \tag{3.1.15}$$

Therefore we have

$$\begin{aligned} \mathbb{E}\left[\int_0^\infty e^{-rt}\varrho_t^{a,i}N^{a,i}(dt)\right] &= \mathbb{E}\left[\int_0^\infty e^{-rt}\left(\lambda^a \varrho_t^{a,i} f_a^i(\varrho_t^{a,i}, \boldsymbol{\varrho}_t^{a,-i})\mathbb{I}(q^i_{t-} > -Z_i)\right)dt\right] \\ \mathbb{E}\left[\int_0^\infty e^{-rt}\varrho_t^{b,i}N^{b,i}(dt)\right] &= \mathbb{E}\left[\int_0^\infty e^{-rt}\left(\lambda^b \varrho_t^{b,i} f_b^i(\varrho_t^{b,i}, \boldsymbol{\varrho}_t^{b,-i})\mathbb{I}(q^i_{t-} < Z_i)\right)dt\right] \end{aligned}$$

We can apply Itô's lemma to $X_t^i + q_t^i S_t$ in objective function, to rewrite (3.1.12) as:

$$\begin{aligned}\tilde{J}_i(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i}; q_0^i, x_0^i, s) =& \mathbb{E}^{q_0^i}\Big[\int_0^\infty e^{-rt}\big(\lambda^a \Delta \varrho_t^{a,i} f_a^i(\varrho_t^{a,i}, \boldsymbol{\varrho}_t^{a,-i})\mathbb{I}(q_{t-}^i > -Z_i) \\ &+ \lambda^b \Delta \varrho_t^{b,i} f_b^i(\varrho_t^{b,i}, \boldsymbol{\varrho}_t^{b,-i})\mathbb{I}(q_{t-}^i < Z_i)\big)dt - \int_0^\infty e^{-rt}\psi_i(q_t^i)dt)\Big] \\ &+ x_0^i + q_0^i s \qquad (3.1.16) \\ =& x_0^i + q_0^i s + J_i(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i}; q_0^i) \qquad \square\end{aligned}$$

The expectation in (3.1.16) is taken with respect to the law of the process q^i whose evolution is given by (3.1.10), with the initial condition q_0^i. For simplicity of notation, we will use $\mathbb{E}$ instead of $\mathbb{E}^{q_0^i}$.

Remark 3.1.7. In the case where (3.1.10) admits a stationary solution (which will typically be the case in many realistic situations), this term does not depend on q_0^i, and we will denote this situation by $J_i(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i})$.

3.2 Competition among Market Makers: Nash Equilibrium

A situation of competition among market makers may be modelled through the concept of Nash equilibrium. In this section, we discuss the existence of Nash equilibrium and its characterization in terms of a Hamilton-Jacobi-Bellman system in the setting of our model.

We first show the existence of the Nash equilibrium under Assumptions 3.1.1 and 3.1.3. We then characterize Nash equilibrium quoting strategies in terms of an HJB equation (3.2.15) in Proposition 3.2.4.

There are N competing market makers whose quotes jointly affect the execution of the market order flow.

Definition 3.2.1 (Nash equilibrium)**.** A Nash equilibrium for system (3.1.14) is a tuple of quoting strategies

$$\vec{\boldsymbol{\delta}}^* = ((\boldsymbol{\delta}^1)^*, \dots, (\boldsymbol{\delta}^N)^*) \in \prod_{j=1}^N \Big((I_\delta)^{2\frac{Z_j}{\Delta}+1} \times (I_\delta)^{2\frac{Z_j}{\Delta}+1}\Big)$$

such that for any $\boldsymbol{q}_0 \in \mathbb{R}^N$ and any $i \in \{1, \dots, N\}$,

$$J_i((\boldsymbol{\delta}^i)^*, (\boldsymbol{\delta}^{-i})^*; q_0^i) \geq J_i(\boldsymbol{\delta}^i, (\boldsymbol{\delta}^{-i})^*; q_0^i), \quad \forall \boldsymbol{\delta}^i \in (I_\delta)^{2\frac{Z_i}{\Delta}+1} \times (I_\delta)^{2\frac{Z_i}{\Delta}+1} \qquad (3.2.1)$$

We say that $\vec{\boldsymbol{\delta}}^*$ is a stationary Nash equilibrium if under $\vec{\boldsymbol{\delta}}^*$ the inventories (3.1.10) admit a stationary solution (see Remark 3.1.7).

We denote

$$V_i(q_i) = J_i\big((\boldsymbol{\delta}^i)^*, (\boldsymbol{\delta}^{-i})^*; q_i\big) \tag{3.2.2}$$

the value function of player i under the Nash equilibrium quoting strategy, with initial condition $q_0^i = q_i$.

We first state a proposition on the uniform boundedness of the objective function $J_i(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i}; q_i)$. This result will be useful for proving the existence of Nash equilibrium.

Proposition 3.2.2. *Under Assumption 3.1.1 and 3.1.3, there exists a constant $J_{max} > 0$ such that for any strategy $(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i})$*

$$\forall i \in \{1, \dots, N\}, \forall q_i \in \mathcal{Q}_i, \qquad |J_i(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i}; q_i)| \leq J_{max} \tag{3.2.3}$$

Proof. From the definition of the objective function, we have

$$\begin{aligned} |J_i(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i}; q_i)| \leq \mathbb{E}\Bigg[& \left| \int_0^\infty e^{-rt} \left(\lambda^a \Delta \varrho_t^{a,i} f_a^i(\varrho_t^{a,i}, \boldsymbol{\varrho}_t^{a,-i}) \mathbb{I}(q_{t-}^i > -Z_i) \right) dt \right| \\ & + \left| \int_0^\infty e^{-rt} \left(\lambda^b \Delta \varrho_t^{b,i} f_b^i(\varrho_t^{b,i}, \boldsymbol{\varrho}_t^{b,-i}) \mathbb{I}(q_{t-}^i < Z_i) \right) dt \right| \\ & + \left| \int_0^\infty e^{-rt} \psi_i(q_i) dt \right| \Bigg] \end{aligned} \tag{3.2.4}$$

Since q_t^i takes values from a finite set $\mathcal{Q}_i = \{-Z_i, -Z_i + \Delta, \dots, Z_i - \Delta, Z_i\}$, $\psi_i(q_i)$ is uniformly bounded by $\Psi_i := \max_{q \in \mathcal{Q}_i} \psi_i(q_i)$. Hence,

$$\mathbb{E}\left[\left| \int_0^\infty e^{-rt} \psi_i(q_i) dt \right| \right] \leq \Psi_i \int_0^\infty e^{-rt} dt = \frac{\Psi_i}{r}$$

From Assumption 3.1.1, $\left|\varrho_t^{a,i} f_a^i(\varrho_t^{a,i}, \boldsymbol{\varrho}_t^{a,-i})\right| \leq \left|\varrho_t^{a,i} \Lambda(\varrho_t^{a,i})\right|$, $\lim_{\delta \to \infty} \Lambda(\delta)\delta = 0$, and $\varrho_t^{a,i} = \delta_{q_t^i}^{a,i}$ take values on $[-\delta_\infty, \infty)$, then there exists a constant $B > 0$ such that $\left|\varrho_t^{a,i} \Lambda(\varrho_t^{a,i})\right| \leq B$. In fact $\lim_{\delta \to \infty} \Lambda(\delta)\delta = 0$ implies that there exists a positive number $M > 0$, such that $\forall \delta > M$, we have $|\Lambda(\delta)\delta| < 1$. Also, $\Lambda(\delta)\delta$ is a continuous function on $[-\delta_\infty, M]$, and hence is bounded. We can then define $K := \max\Big(\sup_{\delta \in [-\delta_\infty, M]} \Lambda(\delta)\delta, 1 \Big)$. Hence $\left|\varrho_t^{a,i} f_a^i(\varrho_t^{a,i}, \boldsymbol{\varrho}_t^{a,-i})\right| \leq K$ uniformly.

Similarly, we have can prove for the bid side $\left|\varrho_t^{b,i} f_b^i(\varrho_t^{b,i}, \boldsymbol{\varrho}_t^{b,-i})\right| \leq K$. Therefore, from (3.2.4) we obtain the following.

$$|J_i(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i}; q_i)| \leq \frac{K(\lambda^a \Delta + \lambda^b \Delta) + \max_{i \in \{1, \dots, N\}} \Psi_i}{r} \tag{3.2.5}$$

The bound is uniform with respect to i, q_i, and $(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i})$. □

We denote the execution probabilities as $f_a^i\Big(\delta_{q_i}^{a,i}, (\vec{\delta}^{a,j})_{j\neq i}\Big)$ and $f_b^i\Big(\delta_{q_i}^{b,i}, (\vec{\delta}^{b,j})_{j\neq i}\Big)$ to emphasize that $\delta_{q_i}^{a,i}, \delta_{q_i}^{b,i}$ are the current ask and bid quotes by market maker i given her inventory level q_i. This notation will provide clarity in the optimality equations.

3.2.1 Dynamic Programming principle

The equilibrium value function (3.2.2) associated with the Nash equilibrium satisfies a Dynamic Programming Principle:

Lemma 3.2.3. *(Dynamic Programming Principle) Let $q_i \in \mathcal{Q}_i$. Given any finite stopping time θ the value function V_i defined by* (3.2.2)

$$V_i(q_i) = \sup_{\boldsymbol{\delta}^i \in \mathcal{A}_i} \mathbb{E}\Bigg[\int_0^\theta e^{-rt}\Bigg(\lambda^a \Delta \delta^{a,i}_{q^{i,q_i}_{t-}} \cdot f_a^i\big(\delta^{a,i}_{q^{i,q_i}_{t-}}, (\vec{\delta}^{a,j})^*_{j\neq i}\big)\mathbb{I}(q^{i,q_i}_{t-} > -Z_i)$$
$$+ \lambda^b \Delta \delta^{b,i}_{q^{i,q_i}_{t-}} \cdot f_b^i\big(\delta^{b,i}_{q^{i,q_i}_{t-}}, (\vec{\delta}^{b,j})^*_{j\neq i}\big)\mathbb{I}(q^{i,q_i}_{t-} < Z_i) - \psi_i(q_t^{i,q_i})\Bigg)dt + V_i(q_\theta^{i,q_i})\Bigg] \quad (3.2.6)$$

where q_t^{i,q_i} is the inventory process of market maker i under joint quoting strategy $(\boldsymbol{\delta}^i, (\boldsymbol{\delta}^{-i})^*)$, with $q_0^{i,q_i} = q_i$.

Proof. Given the quoting strategies $(\boldsymbol{\delta}^i, (\boldsymbol{\delta}^{-i})^*)$, and a finite stopping time $\theta < \infty$, from the Markovian property of q_t^{i,q_i}, we have

$$q_s^{i,q_i} = q_s^{i,q_\theta^{i,q_i}}, \forall s \geq \theta$$

From the properties of conditional expectation and change of variable, we have

$$\begin{aligned}
J_i(\boldsymbol{\delta}^i, (\boldsymbol{\delta}^{-i})^*, q_i) =& \mathbb{E}\left[(\int_0^\theta + \int_\theta^\infty)e^{-rt}d(X_t^i + q_t^i S_t) - (\int_0^\theta + \int_\theta^\infty)e^{-rt}\psi_i(q_t^i)dt\right] \\
=& \mathbb{E}\Bigg[\int_0^\theta \big(e^{-rt}d(X_t^i + q_t^{i,q_i}S_t) - e^{-rt}\psi_i(q_t^{i,q_i})dt\big) \\
&+ \mathbb{E}\left(\int_\theta^\infty e^{-rt}d(X_t^i + q_t^{i,q_i}S_t) - e^{-rt}\psi_i(q_t^{i,q_i})dt\Big| q_\theta^{i,q_i}\right)\Bigg] \\
=& \mathbb{E}\Bigg[\int_0^\theta \Big(e^{-rt}d(X_t^i + q_t^{i,q_i}S_t) - e^{-rt}\psi_i(q_t^{i,q_i})dt\Big) \\
&+ \mathbb{E}\Bigg(\int_0^\infty e^{-r(u+\theta)}d(X_{u+\theta}^i + q_{u+\theta}^{i,q_i}S_{u+\theta}) \\
&- e^{-r(u+\theta)}\psi_i(q_{u+\theta}^{i,q_i})du\Big| q_\theta^{i,q_i}\Bigg)\Bigg] \\
=& \mathbb{E}\Bigg[\int_0^\theta \Big(e^{-rt}d(X_t^i + q_t^{i,q_i}S_t) - e^{-rt}\psi_i(q_t^{i,q_i})dt\Big) \\
&+ e^{-r\theta}J_i\big(\boldsymbol{\delta}^i, (\boldsymbol{\delta}^{-i})^*; q_\theta^{i,q_i}\big)\Bigg] \qquad (3.2.7)
\end{aligned}$$

By Definition 3.2.1 $V_i(q_i) = J_i((\boldsymbol{\delta}^i)^*, (\boldsymbol{\delta}^{-i})^*, q_i) = \sup_{\boldsymbol{\delta}^i} J_i(\boldsymbol{\delta}^i, (\boldsymbol{\delta}^{-i})^*, q_i)$, we have

$$\begin{aligned} J_i(\boldsymbol{\delta}^i, (\boldsymbol{\delta}^{-i})^*, q_i) \leq & \mathbb{E}\Big[\int_0^\theta \Big(e^{-rt} d(X_t^i + q_t^{i,q_i} S_t) - e^{-rt} \psi_i(q_t^{i,q_i}) dt \Big) + e^{-r\theta} V_i(q_\theta^{i,q_i}) \Big] \\ \leq & \sup_{\boldsymbol{\delta}^i} \mathbb{E}\Big[\int_0^\theta \Big(e^{-rt} d(X_t^i + q_t^{i,q_i} S_t) - e^{-rt} \psi_i(q_t^{i,q_i}) dt \Big) \\ & + e^{-r\theta} V_i(q_\theta^{i,q_i})) \Big] \end{aligned} \tag{3.2.8}$$

Taking the supremum over $\boldsymbol{\delta}^i$ in the left-hand side of (3.2.8) we obtain:

$$V_i(q_i) \leq \sup_{\boldsymbol{\delta}^i} \mathbb{E}\left[\int_0^\theta \Big(e^{-rt} d(X_t^i + q_t^{i,q_i} S_t) - e^{-rt} \psi_i(q_t^{i,q_i}) dt \Big) + e^{-r\theta} V_i(q_\theta^{i,q_i})) \right] \tag{3.2.9}$$

On the other hand, for joint quoting strategy $(\boldsymbol{\delta}^i, (\boldsymbol{\delta}^{-i})^*)$ and given stopping time θ, from Definition 3.2.1 for any $\epsilon > 0, \omega \in \Omega$ there exists quoting strategy $\boldsymbol{\delta}^{i,\epsilon,\omega} = (\delta^{a,i,\epsilon,\omega}, \delta^{b,i,\epsilon,\omega}) \in (I_\delta)^{2\frac{Z_i}{\Delta}+1} \times (I_\delta)^{2\frac{Z_i}{\Delta}+1} \times (I_\delta)^{2\frac{Z_i}{\Delta}+1} \times (I_\delta)^{2\frac{Z_i}{\Delta}+1}$, such that $\boldsymbol{\delta}^{i,\epsilon,\omega}$ is an $\epsilon-$optimal control for value function $V_i(q_{\theta(\omega)}^{i,q_i}(\omega))$ starting at $q_{\theta(\omega)}^{i,q_i}(\omega)$:

$$V_i\big(q_{\theta(\omega)}^{i,q_i}(\omega)\big) - \epsilon \leq J_i\Big(\boldsymbol{\delta}^{i,\epsilon,\omega}, (\boldsymbol{\delta}^{-i})^*; q_{\theta(\omega)}^{i,q_i}(\omega)\Big)$$

Now define the following quoting strategy:

$$\hat{\boldsymbol{\delta}^i}_t(\omega) = \begin{cases} \delta^i_{q_{t-}^{i,q_i}(\omega)}\big(\omega\big) & t \in [0, \theta(\omega)] \\ \delta^{i,\epsilon,\omega}_{q_t^{i,q_i}(\omega)}\big(\omega\big) & t \in [\theta(\omega), \infty] \end{cases} \tag{3.2.10}$$

Using a measurable selection argument Bertsekas and Shreve 1978, the process $\hat{\boldsymbol{\delta}^i}$ is an admissible strategy. Then by the law of iterated conditional expectation we have

$$\begin{aligned} V(q_i) \geq & J_i(\hat{\boldsymbol{\delta}^i}, (\boldsymbol{\delta}^{-i})^*, q_i) = \mathbb{E}\Big[\int_0^\theta \Big(e^{-rt} d(X_t^i + q_t^{i,q_i} S_t) - e^{-rt} \psi_i(q_t^{i,q_i}) dt \Big) \\ & + e^{-r\theta} J_i(\delta^{i,\epsilon}, (\boldsymbol{\delta}^{-i})^*; q_\theta^{i,q_i}) \Big] \\ \geq & \mathbb{E}\left[\int_0^\theta \Big(e^{-rt} d(X_t^i + q_t^{i,q_i} S_t) - e^{-rt} \psi_i(q_t^{i,q_i}) dt \Big) + e^{-r\theta} V_i(q_\theta^{i,q_i}) - e^{-r\theta}\epsilon \right] \\ \geq & \mathbb{E}\left[\int_0^\theta \Big(e^{-rt} d(X_t^i + q_t^{i,q_i} S_t) - e^{-rt} \psi_i(q_t^{i,q_i}) dt \Big) + e^{-r\theta} V_i(q_\theta^{i,q_i}) \right] - \epsilon \end{aligned} \tag{3.2.11}$$

Since $\boldsymbol{\delta}^i$, θ and $\epsilon > 0$ are taken arbitrarily, we have the following.

$$V_i(q_i) \geq \sup_{\boldsymbol{\delta}^i} \mathbb{E}\left[\int_0^\theta \left(e^{-rt} d(X_t^i + q_t^{i,q_i} S_t) - e^{-rt}\psi_i(q_t^{i,q_i})dt\right) + e^{-r\theta} V_i(q_\theta^{i,q_i})\right] \tag{3.2.12}$$

Combining (3.2.9) and (3.2.12) we have

$$V_i(q_i) = \sup_{\boldsymbol{\delta}^i} \mathbb{E}\left[\int_0^\theta \left(e^{-rt} d(X_t^i + q_t^{i,q_i} S_t) - e^{-rt}\psi_i(q_t^{i,q_i})dt\right) + e^{-r\theta} V_i(q_\theta^{i,q_i})\right] \tag{3.2.13}$$

□

For arbitrarily given quoting strategies of N market makers $(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i})$, we define the objective function associated with this quoting strategy, denoted by $V_i^\delta(q_i)$, where

$$V_i^\delta(q_i) = J_i(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i}; q_i)$$

For consistency with reinforcement learning literature, we name V^δ the state-action value function. Note that if we denote V_i^δ by $V_i^{\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i}}$ to emphasize the dependence on the strategies of competitors $\boldsymbol{\delta}^{-i}$, then we have the relation $V_i(q_i) = \sup_{\boldsymbol{\delta}^i \in \mathcal{A}_i} V_i^{\boldsymbol{\delta}^i, (\boldsymbol{\delta}^{-i})^*}(q_i)$ where $\vec{\boldsymbol{\delta}}^*$ is a Nash equilibrium. The objective function V^δ associated with given quoting strategies satisfies following linear Bellman equation:[4] (Guéant and Manziuk 2019)

$$\begin{aligned}
&rV_i^\delta(q_i) + \psi_i(q_i) - \mathbb{I}(q_i > -Z_i)\lambda^a \Delta f_a^i\left(\delta_{q_i}^{a,i}, (\vec{\delta}^{a,j})_{j\neq i}\right)\cdot \\
&\left(\delta_{q_i}^{a,i} - \frac{V_i^\delta(q_i) - V_i^\delta(q_i - \Delta)}{\Delta}\right) - \mathbb{I}(q_i < Z_i)\lambda^b \Delta f_b^i\left(\delta_{q_i}^{b,i}, (\vec{\delta}^{b,j})_{j\neq i}\right)\cdot \\
&\left(\delta_{q_i}^{b,i} - \frac{V_i^\delta(q_i) - V_i^\delta(q_i + \Delta)}{\Delta}\right) = 0
\end{aligned} \tag{3.2.14}$$

The Dynamic Programming Principle thus leads to a system of Hamilton-Jacobi equations for the equilibrium value functions $V_i(q_i), i \in \{1, \ldots, N\}$:

$$\begin{aligned}
&rV_i(q_i) + \psi_i(q_i) - \mathbb{I}(q_i > -Z_i)\lambda^a \Delta \sup_{\delta \geq -\delta_\infty}\left[f_a^i\Big(\delta, ((\vec{\delta}^{a,j})^*)_{j\neq i}\Big)\cdot\right. \\
&\left.\left(\delta - \frac{V_i(q_i) - V_i(q_i - \Delta)}{\Delta}\right)\right] - \mathbb{I}(q_i < Z_i)\lambda^b \Delta \sup_{\delta \geq -\delta_\infty}\left[f_b^i\Big(\delta, ((\vec{\delta}^{b,j})^*)_{j\neq i}\Big)\cdot\right. \\
&\left.\left(\delta - \frac{V_i(q_i) - V_i(q_i + \Delta)}{\Delta}\right)\right] = 0
\end{aligned} \tag{3.2.15}$$

[4] Although notation $V_i^\delta(q_i)$ and $J_i(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i}; q_i)$ represent the same quantity, we will more frequently use $V_i^\delta(q_i)$ to emphasize the functional dependence of objective functions on q_i, which also benefits better formatting the linear Bellman equation and algorithm description subsequently.

where Nash equilibrium quoting strategy $\vec{\boldsymbol{\delta}}^*$ is such that the supremum in equation (3.2.15) is achieved simultaneously for $i \in \{1, \ldots, N\}$.

Proposition 3.2.4. *If there exists a Nash equilibrium quoting strategy*

$$\vec{\boldsymbol{\delta}}^* = ((\boldsymbol{\delta}^1)^*, \ldots, (\boldsymbol{\delta}^N)^*) \in \prod_{j=1}^{N} \left((I_\delta)^{2\frac{Z_j}{\Delta}+1} \times (I_\delta)^{2\frac{Z_j}{\Delta}+1} \right)$$

with corresponding equilibrium value functions $V_i(q_i), q_i \in \{-Z_i, \ldots, Z_i\}$, then the V_i satisfy the system of equations (3.2.15), where $\forall q_i \in \{-Z_i, \ldots, Z_i\}$

$$\begin{aligned}
(\delta_{q_i}^{a,i})^* &= \underset{\delta \geq -\delta_\infty}{\arg\max} \left[f_a^i\big(\delta, ((\vec{\delta}^{a,j})^*)_{j\neq i}\big) \left(\delta - \frac{V_i(q_i) - V_i(q_i - \Delta)}{\Delta} \right) \right] \\
(\delta_{q_i}^{b,i})^* &= \underset{\delta \geq -\delta_\infty}{\arg\max} \left[f_b^i\big(\delta, ((\vec{\delta}^{b,j})^*)_{j\neq i}\big) \left(\delta - \frac{V_i(q_i) - V_i(q_i - \Delta)}{\Delta} \right) \right]
\end{aligned} \tag{3.2.16}$$

Proof. The proof of Proposition 3.2.4 follows by a standard argument using the Dynamic Programming Principle in Lemma 3.2.3. □

The next proposition is a verification theorem stating that the solution to (3.2.15) is indeed a Nash equilibrium for the N market maker system.

Proposition 3.2.5. *Under Assumptions 3.1.1 and 3.1.3, if there exist quoting strategies $(\delta^{a,i})^*, (\delta^{b,i})^*, \forall i \in \{1, \ldots, N\}$ and functions $V_i(q_i), q_i \in \{-Z_i, \ldots, Z_i\}$ such that V_i satisfy system of equations (3.2.15), and $\forall q_i \in \{-Z_i, \ldots, Z_i\}$*

$$\begin{aligned}
(\delta_{q_i}^{a,i})^* &= \underset{\delta \geq -\delta_\infty}{\arg\max} \left[f_a^i\big(\delta, ((\vec{\delta}^{a,j})^*)_{j\neq i}\big) \left(\delta - \frac{V_i(q_i) - V_i(q_i - \Delta)}{\Delta} \right) \right] \\
(\delta_{q_i}^{b,i})^* &= \underset{\delta \geq -\delta_\infty}{\arg\max} \left[f_b^i\big(\delta, ((\vec{\delta}^{b,j})^*)_{j\neq i}\big) \left(\delta - \frac{V_i(q_i) - V_i(q_i - \Delta)}{\Delta} \right) \right]
\end{aligned} \tag{3.2.17}$$

then $(\delta^{a,i})^*, (\delta^{b,i})^*$ define a Nash equilibrium (Def. 3.2.1) and V_i are the corresponding equilibrium value functions.

Proof. To show that $(\boldsymbol{\delta}^i)^* = \left((\delta^{a,i})^*, (\delta^{b,i})^*\right), i \in \{1, \ldots, N\}$ are Nash equilibrium quoting strategies and $V_i(q_i)$ is the equilibrium value function, we need to verify $(\delta^{a,i})^*, (\delta^{b,i})^*$ and $V_i(q_i)$ satisfy Definition 3.2.1. Given any market maker i, we denote the joint quoting strategies of her competitors by $(\boldsymbol{\delta}^{-i})^*$. Assuming market maker i takes another quoting strategy $\boldsymbol{\delta}^i = (\delta^{a,i}, \delta^{b,i}) \in \mathbb{R}^{2\frac{Z_i}{\Delta}+1} \times \mathbb{R}^{2\frac{Z_i}{\Delta}+1}$ while her competitors still keep the joint quoting strategies $(\boldsymbol{\delta}^{-i})^*$, we need to show that

$$V_i(q_i) \geq J_i(\boldsymbol{\delta}^i, (\boldsymbol{\delta}^{-i})^*, q_i), \forall q_i \in \{-Z_i, \ldots, Z_i\}$$

Let $(X_t^{i,\delta})_{t\geq 0}$ denote the cash process and $(q_t^{i,\delta})_{t\geq 0}$ denote the inventory process of the market maker i with joint strategies $(\boldsymbol{\delta}^i, (\boldsymbol{\delta}^{-i})^*)$, where $q_0^{i,\delta} = q_i$. Since the process $q_t^{i,\delta}$ takes values from a finite set $\{-Z_i, -Z_i + \Delta, \ldots, Z_i - \Delta, Z_i\}$, the value function $V_i(q_t^{i,\delta})$ is uniformly bounded. Denote this uniform bound by $M (M > 0)$. We can then apply Ito's formula on the function $e^{-rt}V_i(q_t^{i,\delta})$.

For $T > 0$,

$$\begin{aligned} e^{-rT}V_i(q_T^{i,\delta}) &= V_i(q_i) - \int_0^T re^{-rt}V_i(q_t^{i,\delta})dt \\ &+ \int_0^T e^{-rt}\left[V_i(q_{t-}^{i,\delta} - \Delta) - V_i(q_{t-}^{i,\delta})\right] N^{a,i}(dt) \\ &+ \int_0^T e^{-rt}\left[V_i(q_{t-}^{i,\delta} + \Delta) - V_i(q_{t-}^{i,\delta})\right] N^{b,i}(dt) \end{aligned} \tag{3.2.18}$$

From Assumption 3.1.1, $f_a^i(\delta, \delta^{-i}) < \Lambda(\delta), f_b^i(\delta, \delta^{-i}) < \Lambda(\delta)$, and $\Lambda(\delta)$ is monotonically decreasing on $\mathbb{R}$. Moreover centered quotes are bounded from below by $-\delta_\infty$, we have

$$\begin{aligned} &\left|\mathbb{E}\int_0^T e^{-rt}\left[V_i(q_{t-}^{i,\delta} - \Delta) - V_i(q_{t-}^{i,\delta})\right] f_a^i\Big(\delta^{a,i}_{q_{t-}^{i,\delta}}, ((\vec{\delta}^{a,j})^*)_{j\neq i}\Big)\mathbb{I}(q_{t-}^{i,\delta} > -Z_i)dt\right| \\ \leq\ & 2M\cdot\mathbb{E}\left[\int_0^T e^{-rt}\Lambda\Big(\delta^{a,i}_{q_{t-}^{i,\delta}}\Big)dt\right] \leq 2M\cdot\mathbb{E}\left[\int_0^T e^{-rt}\Lambda(-\delta_\infty)dt\right] \\ \leq\ & 2M\Lambda(-\delta_\infty)\int_0^\infty e^{-rt}dt < \infty \end{aligned} \tag{3.2.19}$$

Similarly we also have

$$\left|\mathbb{E}\int_0^T e^{-rt}\left[V_i(q_{t-}^{i,\delta} + \Delta) - V_i(q_{t-}^{i,\delta})\right] f_b^i\Big(\delta^{b,i}_{q_{t-}^{i,\delta}}, ((\vec{\delta}^{b,j})^*)_{j\neq i}\Big)\mathbb{I}(q_{t-}^{i,\delta} < Z_i)dt\right| < \infty \tag{3.2.20}$$

Therefore we can take expectation on both sides of (3.2.18), and obtain

$$\begin{aligned} &\mathbb{E}\left[e^{-rT}V_i(q_T^{i,\delta})\right] = V_i(q_i) - \mathbb{E}\left[\int_0^T re^{-rt}V_i(q_t^{i,\delta})dt\right] \\ &+ \mathbb{E}\int_0^T e^{-rt}\left[V_i(q_{t-}^{i,\delta} - \Delta) - V_i(q_{t-}^{i,\delta})\right] f_a^i\Big(\delta^{a,i}_{q_{t-}^{i,\delta}}, ((\vec{\delta}^{a,j})^*)_{j\neq i}\Big)\mathbb{I}(q_{t-}^{i,\delta} > -Z_i)dt \\ &+ \mathbb{E}\int_0^T e^{-rt}\left[V_i(q_{t-}^{i,\delta} + \Delta) - V_i(q_{t-}^{i,\delta})\right] f_b^i\Big(\delta^{b,i}_{q_{t-}^{i,\delta}}, ((\vec{\delta}^{b,j})^*)_{j\neq i}\Big)\mathbb{I}(q_{t-}^{i,\delta} < Z_i)dt \end{aligned} \tag{3.2.21}$$

Since $V_i(q_i)$ satisfies HJB equation (3.2.15) for any $q_i \in \{-Z_i, \ldots, Z_i\}$ with $(\boldsymbol{\delta}^i)^*, (\boldsymbol{\delta}^{-i})^*$ being the Nash equilibrium quoting strategy, we can then replace

$(\boldsymbol{\delta}^i)^*$ by $\boldsymbol{\delta}^i$ and obtain inequality

$$
\begin{aligned}
&rV_i(q_t^{i,\delta}) + \psi_i(q_t^{i,\delta}) - \mathbb{I}(q_{t-}^{i,\delta} > -Z_i)\lambda^a \Delta \Big[f_a^i \Big(\delta^{a,i}_{q_{t-}^{i,\delta}}, ((\vec{\delta}^{a,j})^*)_{j\neq i} \Big) \cdot \\
&\left(\delta^{a,i}_{q_{t-}^{i,\delta}} - \frac{V_i(q_{t-}^{i,\delta}) - V_i(q_{t-}^{i,\delta} - \Delta)}{\Delta} \right) \Big] - \mathbb{I}(q_{t-}^{i,\delta} < Z_i)\lambda^b \Delta \Big[f_b^i \Big(\delta^{b,i}_{q_{t-}^{i,\delta}}, ((\vec{\delta}^{b,j})^*)_{j\neq i} \Big) \cdot \\
&\left(\delta^{b,i}_{q_{t-}^{i,\delta}} - \frac{V_i(q_{t-}^{i,\delta}) - V_i(q_{t-}^{i,\delta} + \Delta)}{\Delta} \right) \Big] \geq 0
\end{aligned} \tag{3.2.22}
$$

Combining (3.2.21) and (3.2.22), we obtain

$$
\begin{aligned}
&\mathbb{E}\left[e^{-rT} V_i(q_T^{i,\delta}) \right] \leq V_i(q_i) - \mathbb{E}\Big[\int_0^T e^{-rt} \Big(\mathbb{I}(q_{t-}^{i,\delta} > -Z_i)\lambda^a \Delta f_a^i \big(\delta^{a,i}_{q_{t-}^{i,\delta}}, ((\vec{\delta}^{a,j})^*)_{j\neq i} \big) \\
&+ \mathbb{I}(q_{t-}^{i,\delta} < Z_i)\lambda^b \Delta f_b^i \big(\delta^{b,i}_{q_{t-}^{i,\delta}}, ((\vec{\delta}^{b,j})^*)_{j\neq i} \big) \Big) dt \Big] + \mathbb{E}\left[\int_0^T e^{-rt} \psi_i(q_t^{i,\delta}) dt \right]
\end{aligned} \tag{3.2.23}
$$

From (3.2.19) and (3.2.20) and dominated convergence theorem, we can let $T \to \infty$ in both sides of (3.2.23). Since V_i is uniformly bounded, we have $\mathbb{E}\left[e^{-rT} V_i(q_T^{i,\delta}) \right] \xrightarrow{T\to\infty} 0$. We then obtain

$$
\begin{aligned}
V_i(q_i) \geq &\mathbb{E}\Big[\int_0^\infty e^{-rt} \Big(\mathbb{I}(q_{t-}^{i,\delta} > -Z_i)\lambda^a \Delta f_a^i \big(\delta^{a,i}_{q_{t-}^{i,\delta}}, ((\vec{\delta}^{a,j})^*)_{j\neq i} \big) \\
&+ \mathbb{I}(q_{t-}^{i,\delta} < Z_i)\lambda^b \Delta f_b^i \big(\delta^{b,i}_{q_{t-}^{i,\delta}}, ((\vec{\delta}^{b,j})^*)_{j\neq i} \big) \Big) dt \Big] - \mathbb{E}\left[\int_0^\infty e^{-rt} \psi_i(q_t^{i,\delta}) dt \right]
\end{aligned} \tag{3.2.24}
$$

The right hand side of (3.2.24) is exactly $J_i(\boldsymbol{\delta}^i, (\boldsymbol{\delta}^{-i})^*, q_i)$. Hence we have

$$
V_i(q_i) \geq J_i(\boldsymbol{\delta}^i, (\boldsymbol{\delta}^{-i})^*, q_i) \tag{3.2.25}
$$

$V_i(q_i)$ is the equilibrium value function in Definition 3.2.1.

Now, if market maker i takes the quoting strategy $(\boldsymbol{\delta}^i)^*$, while other market makers keep the strategies $(\boldsymbol{\delta}^{-i})^*$, the equality in (3.2.22) is achieved with $\boldsymbol{\delta}^i$ replaced by $(\boldsymbol{\delta}^i)^*$, as $(\boldsymbol{\delta}^i)^*$ is the maximum point in the equation (3.2.15). Subsequently, in (3.2.23) and (3.2.24) equality will be achieved with $\boldsymbol{\delta}^i$ replaced by $(\boldsymbol{\delta}^i)^*$. Therefore

$$
V_i(q_i) \geq J_i((\boldsymbol{\delta}^i)^*, (\boldsymbol{\delta}^{-i})^*, q_i) \tag{3.2.26}
$$

Therefore $\left\{ \left((\boldsymbol{\delta}^i)^*, (\boldsymbol{\delta}^{-i})^* \right), i \in \{1, \ldots, N\} \right\}$ is the Nash equilibrium quoting strategy in Definition 3.2.1. □

3.2.2 Existence of Nash Equilibrium

With Proposition 3.2.4 and Proposition 3.2.5 the existence Nash equilibrium can be established by seeking the solution to the system of equations (3.2.15). We state the following existence result of the Nash equilibrium. We briefly sketch the proof idea and complete the proof in Appendix A.2.

Theorem 3.2.6 (Existence of Nash equilibrium)**.** *Under Assumptions 3.1.1 and 3.1.3 a Nash equilibrium exists.*

The proof of Theorem 3.2.6 is motivated by [Luo and Zheng 2021]. We extend the proof given by [Luo and Zheng 2021] to the circumstance of multiple market makers. The first main result that we prove is Proposition A.2.9 that shows the fixed point property of the functions arg max defined in (A.2.3). Subsequently, we show by Proposition A.2.11 the uniqueness and continuity of the fixed point $\boldsymbol{\delta}_{\boldsymbol{p}} = \boldsymbol{\delta}(\boldsymbol{p})$ as a function of any given value function vector $\boldsymbol{p}$, using a global implicit function theorem A.2.10. Finally, we prove the system of non-linear equations (3.2.15) has a solution by Schauder's fixed-point theorem.

Remark 3.2.7. Both this thesis and the work of [Luo and Zheng 2021] address market making problems in the presence of competition. However, there are key differences in the model frameworks. [Luo and Zheng 2021] consider the problem from the perspective of a reference market maker. They incorporate the actions of competitors as exogenous variables in the intensity functions, where a single variable is used to represent the best ask and bid quotes of the competitors. In contrast, the thesis models the competition among N market makers endogenously, with the market makers' quoting strategies incorporated into the intensity functions. Conceptually, the model in [Luo and Zheng 2021] resembles a mean field game, where the mean field is replaced by the distribution of the most competitive market maker's inventory.

Regarding the proof of existence by Theorem 3.2.6, both our work and [Luo and Zheng 2021] use Schauder's fixed point theorem. While similar tools are employed, such as the fixed point property of the arg max functions of the Hamiltonian and the implicit function theorem, the application in our case involves a few extensions. [Luo and Zheng 2021] prove the existence of solutions to a system of ODEs using the Cauchy-Lipschitz theorem under finite horizon settings. Our proof deals with an infinite horizon setup and extends to N market makers, with the application of Schauder's theorem to address the existence of solutions to a system of non-linear equations, which are the main challenges and differences compared to [Luo and Zheng 2021].

3.3 Collusion and Pareto Optima

Collusion refers to the case where market makers coordinate their actions as a cartel in order to jointly maximize the sum of their objective functions, while sharing their inventory information ([Tirole 1988]). An explicit collusion strategy is thus equivalent to solving a 'central planner' problem whose objective is the sum of all market makers' profits.

In collusion, market makers make decisions based on both their own inventory and inventories of other market makers in the cartel. Consequently, the quoting strategy for the market maker i, depends on the entire set of inventories $\boldsymbol{q} \in \prod_{j=1}^{N} \mathcal{Q}_j$ of all the market makers. Note that this is an essential difference from the competitive setting studied in Sections 3.1 and 3.2, where the quoting strategies of the market maker i are based solely on its own inventory level q_i. Her quoting strategies $\vec{\delta}^{a,i}, \vec{\delta}^{b,i}$ are represented as vectors, with coordinates indexed by $\boldsymbol{q} \in \prod_{j=1}^{N} \mathcal{Q}_j$, instead of $q_i \in \mathcal{Q}_i$. Due to the difference in dimension of the quoting strategies, we need to adapt the execution probability functions f_a^i, f_b^i to be compatible with the expanded dimension, as we will see in (3.3.8)-(3.3.9).

Denoting the product space by $\mathcal{Q} = \prod_{j=1}^{N} \mathcal{Q}_j$, the ask and bid quotes have the form $\vec{\delta}^{a,i} = (\delta^{a,i}(\boldsymbol{q}))_{\boldsymbol{q}\in\mathcal{Q}}, \vec{\delta}^{b,i} = \left(\delta^{b,i}(\boldsymbol{q})\right)_{\boldsymbol{q}\in\mathcal{Q}}$, where $\vec{\delta}^{a,i}, \vec{\delta}^{b,i} \in (I_\delta)^{\prod_{j=1}^{N}(2\frac{Z_j}{\Delta}+1)}$. The quoting strategy of the market maker i can therefore be represented as $\boldsymbol{\delta}^i = (\vec{\delta}^{a,i}, \vec{\delta}^{b,i}) \in (I_\delta)^{\prod_{j=1}^{N} 2\frac{Z_i}{\Delta}+1} \times (I_\delta)^{\prod_{j=1}^{N} 2\frac{Z_i}{\Delta}+1}$. We denote the space of (joint) two-sided quoting strategies for all market makers by

$$\mathcal{S} = \prod_{j=1}^{N} \left((I_\delta)^{\prod_{j=1}^{N} 2\frac{Z_j}{\Delta}+1} \times (I_\delta)^{\prod_{j=1}^{N} 2\frac{Z_j}{\Delta}+1} \right)$$

Definition 3.3.1 (Collusion)**.** A set of quoting strategies

$$\vec{\boldsymbol{\delta}}^c = ((\boldsymbol{\delta}^1)^c, \ldots, (\boldsymbol{\delta}^N)^c) \in \mathcal{S}$$

represents *collusion* if it maximizes the sum of all market makers' objective functions for any $\boldsymbol{q} \in \mathbb{R}^N$:

$$\sum_{i=1}^{N} J_i((\boldsymbol{\delta}^i)^c, (\boldsymbol{\delta}^{-i})^c; q_i) \geq \sum_{i=1}^{N} J_i(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i}; q_i), \qquad \forall \vec{\boldsymbol{\delta}} \in \mathcal{S} \tag{3.3.1}$$

Remark 3.3.2. In addition to Definition 3.3.1, we acknowledge that it is also possible for the cartel in real-world scenarios to minimize based on a collective measure of risk, which we refer to as global risk. This concept arises when the objective changes from considering the aggregate individual inventory risks to managing the risks of the entire cartel in a collective manner. For example, when the penalty functions ψ_i are quadratic in inventory levels, the notion of global risk can be well represented by a quadratic form, such as $\frac{1}{2}\boldsymbol{q}^T\varsigma\boldsymbol{q}$, where ς is a matrix chosen by the cartel to integrate global risk in the inventories of the members of the cartel.

However, extending the concept of global risk beyond quadratic penalty functions is challenging in defining a proper unified risk measure. If the functions ψ_i take a more general form, it remains open how to combine these functions consistently that appropriately capture the global risk. We leave this topic to future research and apply the current Definition 3.3.1 of aggregate objective function throughout Chapter 3.

For given joint quoting strategies $\vec{\boldsymbol{\delta}} \in \mathcal{S}$, we denote by $\mathcal{J}$ the sum of objective functions of N market makers.

$$\mathcal{J}(\vec{\boldsymbol{\delta}}; \boldsymbol{q}) = \sum_{i=1}^{N} J_i(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i}; q_i) \tag{3.3.2}$$

The quantity

$$W(\boldsymbol{q}) = \mathcal{J}(\vec{\boldsymbol{\delta}}^c; \boldsymbol{q}) = \sup_{\vec{\boldsymbol{\delta}} \in \mathcal{S}} \mathcal{J}(\vec{\boldsymbol{\delta}}; \boldsymbol{q}) \tag{3.3.3}$$

represents the *cartel* value function. As we see from Definition 3.3.1, computing the cartel's strategy and the corresponding value functions amounts to solving a stochastic optimal control problem for a central agent with objective function 3.3.2 in policy space $\mathcal{S}$.

Motivated by [Cont, Guo, and Xu 2021], we show that collusion corresponds to a Pareto optimum:

Definition 3.3.3 (Pareto optimum)**.** $\vec{\boldsymbol{\delta}}^p = ((\boldsymbol{\delta}^1)^p, \ldots, (\boldsymbol{\delta}^N)^p) \in \mathcal{S}$ is a Pareto-optimal policy if and only if there does not exist $\vec{\boldsymbol{\delta}} \in \mathcal{S}$, such that for all $\boldsymbol{q} \in \prod_{j=1}^{N} \mathcal{Q}_j$,

$$\begin{aligned} &\forall i \in \{1, \ldots, N\}, J_i(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i}; q_i) \geq J_i((\boldsymbol{\delta}^i)^p, (\boldsymbol{\delta}^{-i})^p; q_i) \\ &\exists j \in \{1, \ldots, N\}, J_j(\boldsymbol{\delta}^j, \boldsymbol{\delta}^{-j}; q_j) > J_j((\boldsymbol{\delta}^j)^p, (\boldsymbol{\delta}^{-j})^p; q_i) \end{aligned} \tag{3.3.4}$$

The following result links the concept of collusion with Pareto optima of the N market maker system:

Proposition 3.3.4. *Any collusion strategy* $\vec{\boldsymbol{\delta}}^c = ((\boldsymbol{\delta}^1)^c, \ldots, (\boldsymbol{\delta}^N)^c)$ *in the sense of Definition 3.3.1 is a Pareto optimum as defined in Definition 3.3.3.*

Proof. By Definition 3.3.1, for any joint quoting strategy $\vec{\boldsymbol{\delta}} \in \mathcal{S}$,

$$W(\boldsymbol{q}) \geq \mathcal{J}(\vec{\boldsymbol{\delta}}; \boldsymbol{q}) = \sum_{i=1}^{N} J_i(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i}; q_i) \tag{3.3.5}$$

If there exists a quoting strategy $\vec{\boldsymbol{\delta}}'$ and $k \in \{1, \ldots, N\}$ such that

$$J_k((\boldsymbol{\delta}^i)', (\boldsymbol{\delta}^{-i})'; q_i) > J_k((\boldsymbol{\delta}^i)^c, (\boldsymbol{\delta}^{-i})^c; q_i) \tag{3.3.6}$$

meaning that for market maker k there is a strictly better joint quoting strategy, we need to prove that $\vec{\boldsymbol{\delta}}'$ cannot be an 'overall better' joint strategy for all market makers. Since

$$\sum_{j=1}^{N} J_j((\boldsymbol{\delta}^i)', (\boldsymbol{\delta}^{-i})'; q_i) \leq W(\boldsymbol{q}) = \sum_{j=1}^{N} J_j((\boldsymbol{\delta}^i)^c, (\boldsymbol{\delta}^{-i})^c; q_i) \tag{3.3.7}$$

There must exists another market maker $j \neq k$ such that her objective

$$J_j((\boldsymbol{\delta}^i)', (\boldsymbol{\delta}^{-i})'; q_i) < J_j((\boldsymbol{\delta}^i)^c, (\boldsymbol{\delta}^{-i})^c; q_i)$$

which can be verified by contradiction argument with (3.3.6-3.3.7). Therefore, $\vec{\boldsymbol{\delta}}'$ is not a 'globally better' strategy than $\vec{\boldsymbol{\delta}}^c$ for all market makers. By Definition 3.3.3, the explicit collusion strategy $\vec{\boldsymbol{\delta}}^c$ is a Pareto-optimal policy. □

Compared to (3.2.15) the dimension of the joint quoting strategy is changed since, under explicit collusion, each market maker's ask and bid quotes are functions of $\boldsymbol{q}$, instead of their own inventory level. We assume that under explicit collusion, the execution probability for the market maker i, denoted by $\tilde{f}_a^i$, $\tilde{f}_b^i$, is a function defined in $\mathbb{R} \times \prod_{j \neq i} \mathbb{R}^{\prod_{j=1}^{N}(2\frac{Z_j}{\Delta}+1)}$. For $\delta \in \mathbb{R}, \delta^{-i} \in \prod_{j \neq i} \mathbb{R}^{\prod_{j=1}^{N}(2\frac{Z_j}{\Delta}+1)}$,

$$\tilde{f}_a^i(\delta, \vec{\delta}^{-i}) : \mathbb{R} \times \prod_{k \neq i} \mathbb{R}^{\prod_{j=1}^{N}(2\frac{Z_j}{\Delta}+1)} \to \mathbb{R}^+, \tilde{f}_b^i(\delta, \vec{\delta}^{-i}) : \mathbb{R} \times \prod_{k \neq i} \mathbb{R}^{\prod_{j=1}^{N}(2\frac{Z_j}{\Delta}+1)} \to \mathbb{R}^+ \tag{3.3.8}$$

$\tilde{f}_a^i$, $\tilde{f}_b^i$ are interconnected through f_a^i, f_b^i in the following way. We take the example of ask quotes. For given joint inventory $\boldsymbol{q}$ and joint ask quoting strategies $(\delta_{\boldsymbol{q}}^{a,i}, ((\vec{\delta}_{\tilde{\boldsymbol{q}}}^{a,j})_{\tilde{\boldsymbol{q}} \in \mathcal{Q}})_{j \neq i})$,

$$\tilde{f}_a^i(\delta_{\boldsymbol{q}}^{a,i}, ((\vec{\delta}_{\tilde{\boldsymbol{q}}}^{a,j})_{\tilde{\boldsymbol{q}} \in \mathcal{Q}})_{j \neq i}) = f_a^i(\delta_{\boldsymbol{q}}^{a,i}, (\bar{\delta}^{a,j})_{j \neq i}) \tag{3.3.9}$$

where $\bar{\delta}^{a,j} \in \mathbb{R}^{2\frac{Z_j}{\Delta}+1}$ is a $(2\frac{Z_j}{\Delta}+1)$ dimensional vector indexed by $\mathcal{Q}_j$. $\bar{\delta}^{a,j}$ is defined by taking the average of market maker j's ask quotes at joint inventories where her inventory is q_j: $\forall q_j \in \mathcal{Q}_j$, $\bar{\delta}^{a,j}_{q_j} = \frac{1}{\prod_{k=1,k\neq j}^{N}(2\frac{Z_k}{\Delta}+1)} \sum_{\tilde{\boldsymbol{q}}\in\mathcal{Q},\tilde{\boldsymbol{q}}_j=q_j} \delta^{a,j}_{\tilde{\boldsymbol{q}}}$, where $\delta^{a,j}_{\tilde{\boldsymbol{q}}}$ is the coordinate of vector $(\vec{\delta}^{a,j}_{\tilde{\boldsymbol{q}}})_{\tilde{\boldsymbol{q}}\in\mathcal{Q}}$ at joint inventory index $\tilde{\boldsymbol{q}}$.

From the dynamic programming principle, the value function $W(\boldsymbol{q})$ satisfies the HJB equation: $\forall \boldsymbol{q} \in \prod_{j=1}^{N} \mathcal{Q}_j$,

$$
\begin{aligned}
&rW(\boldsymbol{q}) + \sum_{i=1}^{N}\psi_i(q_i) - \lambda^a\Delta \sup_{\vec{\boldsymbol{\delta}}\in\mathcal{S}} \sum_{i=1}^{N} \mathbb{I}(q_i > -Z_i)\Big[\tilde{f}^i_a\big(\delta^{a,i}_{\boldsymbol{q}}, ((\vec{\delta}^{a,j}_{\tilde{\boldsymbol{q}}})_{\tilde{\boldsymbol{q}}\in\mathcal{Q}})_{j\neq i}\big)\cdot \\
&\Big(\delta^{a,i}_{\boldsymbol{q}} - \frac{W(\boldsymbol{q}) - W(\boldsymbol{q}-\Delta\boldsymbol{e}_i)}{\Delta}\Big)\Big] - \lambda^b\Delta \sup_{\vec{\boldsymbol{\delta}}\in\mathcal{S}} \sum_{i=1}^{N} \mathbb{I}(q_i < Z_i)\Big[\tilde{f}^i_b\big(\delta^{b,i}_{\boldsymbol{q}}, ((\vec{\delta}^{b,j}_{\tilde{\boldsymbol{q}}})_{\tilde{\boldsymbol{q}}\in\mathcal{Q}})_{j\neq i}\big)\cdot \\
&\Big(\delta^{b,i}_{\boldsymbol{q}} - \frac{W(\boldsymbol{q}) - W(\boldsymbol{q}+\Delta\boldsymbol{e}_i)}{\Delta}\Big)\Big] = 0
\end{aligned} \tag{3.3.10}
$$

We further justify our choice of $\tilde{f}^i_a, \tilde{f}^i_b$ from (3.3.10). If we still formulate the central planner's problem with partially observed information, i.e. applying quoting strategies as functions of single q_i and keeping f^i_a, f^i_b in (3.3.10), (3.3.10) would be otherwise an over-determined system which derives, respectively, on ask and bid sides $\prod_{j=1}^{N}(2\frac{Z_i}{\Delta}+1)$ single variate optimization problem, but only $\sum_{j=1}^{N}(2\frac{Z_i}{\Delta}+1)$ variables on each side ($\delta^{a,i}_{q_j}$ and $\delta^{b,i}_{q_j}$). Hence, to formulate explicit collusion, we define the execution probability function $\tilde{f}^i$ that is close to $f^i_m, m \in \{a,b\}$, but compatible with the expanded dimension of the quoting strategies.

For given joint quoting strategies $\vec{\boldsymbol{\delta}}$, denote by $W^\delta(\boldsymbol{q})$ the value function associated with strategies $\vec{\boldsymbol{\delta}}$, then $W^\delta(\boldsymbol{q})$ satisfies the system of linear equations.

$$
\begin{aligned}
&rW^\delta(\boldsymbol{q}) + \sum_{i=1}^{N}\psi_i(q_i) - \lambda^a\Delta \sum_{i=1}^{N} \mathbb{I}(q_i > -Z_i)\Big[\tilde{f}^i_a\big(\delta^{a,i}_{\boldsymbol{q}}, ((\vec{\delta}^{a,j}_{\tilde{\boldsymbol{q}}})_{\tilde{\boldsymbol{q}}\in\mathcal{Q}})_{j\neq i}\big)\cdot \\
&\Big(\delta^{a,i}_{\boldsymbol{q}} - \frac{W^\delta(\boldsymbol{q}) - W^\delta(\boldsymbol{q}-\Delta\boldsymbol{e}_i)}{\Delta}\Big)\Big] - \lambda^b\Delta \sum_{i=1}^{N} \mathbb{I}(q_i < Z_i)\Big[\tilde{f}^i_b\big(\delta^{b,i}_{\boldsymbol{q}}, ((\vec{\delta}^{a,j}_{\tilde{\boldsymbol{q}}})_{\tilde{\boldsymbol{q}}\in\mathcal{Q}})_{j\neq i}\big)\cdot \\
&\Big(\delta^{b,i}_{\boldsymbol{q}} - \frac{W^\delta(\boldsymbol{q}) - W^\delta(\boldsymbol{q}+\Delta\boldsymbol{e}_i)}{\Delta}\Big)\Big] = 0
\end{aligned} \tag{3.3.11}
$$

We then obtain the following matrix representation of the system (3.3.11) by rearranging the terms in the equations:

$$
M^\delta \cdot W^\delta = A^\delta \tag{3.3.12}
$$

where W^δ represents the vector whose coordinates are the value function $W^\delta(\boldsymbol{q})$. M^δ is $\prod_{j=1}^{N}\left(2\frac{Z_j}{\Delta}+1\right)\times\prod_{j=1}^{N}\left(2\frac{Z_j}{\Delta}+1\right)$ matrix, and $W^\delta, A^\delta \in \mathbb{R}^{\prod_{j=1}^{N}\left(2\frac{Z_j}{\Delta}+1\right)}$. Data M^δ, A^δ are derived from (3.3.11). We now focus on constructing a numerical solution for the HJB equation (3.3.10, with the linear system (3.3.11) and (3.3.12) serving as important intermediate steps in the numerical scheme.

The existence and uniqueness of the solutions to (3.3.10) can be established with methods of stochastic control on graphs in [Guéant and Manziuk 2020]. The linear system of Bellman equations (3.3.11) allows to efficiently compute the value function W^δ for given joint quoting strategies. This naturally suggests an iterative approach to improving the quoting strategies step by step, eventually forming the basis for a policy iteration algorithm. The policy iteration method has been extensively studied across several decades, starting with the foundational work [Puterman and Brumelle 1979], which demonstrates the mathematical equivalence of policy iteration with a type of Newton's method in dynamic programming and analyzes its convergence for Markovian decision processes with a quadratic convergence rate. This is followed by [Puterman 1981], who extended these results to controlled diffusion processes. More recent advances have applied the policy iteration method to solving stochastic control problems and analyzed its convergence properties. [Santos and Rust 2004] studies the convergence properties for continuous state and control variables, establishing quadratic and superlinear convergence under specific conditions for discretized systems. [Jacka and Mijatović 2017] focus on the policy improvement algorithm for stochastic control problems in continuous time, establishing general conditions under which the algorithm is well-defined and convergent. [Kerimkulov, Šiška, and Szpruch 2020] investigate the global convergence properties of the policy iteration algorithm for controlled diffusion processes and prove the stability of the algorithm under perturbations. [Ito, Reisinger, and Zhang 2021] introduce a neural network-based policy iteration algorithm to solve high-dimensional stochastic games with HJBI boundary value problems, establishing H^2-superlinear convergence of the numerical scheme. Alternatively, there are approximation techniques available to solve the high-dimensional Hamilton-Jacobi equations similar to (3.3.10). In particular, [Bergault, Evangelista, et al. 2021] approximate the Hamiltonian functions in the equation by quadratic forms resulting in Riccati equations which can be solved in closed form. The closed-form solutions are then used as approximations for solutions to the original Hamilton-Jacobi

equations. However, compared to the multi-asset market making problems considered in [Bergault, Evangelista, et al. 2021], our settings do not yield a closed-form solution to (3.3.10) as a benchmark, making it challenging to assess the accuracy of such approximations. Therefore, we choose to solve (3.3.10) numerically using iterative algorithms.

Building on insights from the recent literature on policy iteration for stochastic control problems, we hereby propose a policy iteration scheme to numerically solve (3.3.10). Our policy iteration scheme sequentially improves the joint quoting strategies $\vec{\boldsymbol{\delta}}$ by breaking the problem down into **policy evaluation** and **policy improvement** in each iteration. The policy evaluation step is based on the linear Bellman equation (3.3.11), where the value function W^{δ} associated with a given joint quoting strategy $\vec{\boldsymbol{\delta}} \in \mathcal{S}$ is calculated by solving (3.3.12). Following the policy evaluation step, the policy improvement step solves the pointwise optimization of Hamiltonians in (3.3.10) for each inventory profile $\boldsymbol{q}$, based on the computed values of W^{δ} from the policy evaluation. However, direct optimization of Hamiltonians presents significant computational challenges, as it requires solving a high-dimensional multi-objective optimization problem for each inventory vector $\boldsymbol{q}$ simultaneously. To address this challenge, we first make the assumption that the market makers in the cartel adopt a uniform quoting strategy. This simplification is valid for homogeneous market makers, which is the problem we are interested in. Namely, for $\boldsymbol{q} \in \mathcal{Q}$, the approximated ask and bid quotes for homogeneous market makers are

$$\delta_{\boldsymbol{q}}^{a,i} = \delta_{\boldsymbol{q}}^{a} \qquad \delta_{\boldsymbol{q}}^{b,i} = \delta_{\boldsymbol{q}}^{b} \tag{3.3.13}$$

Moreover, in the policy improvement step, we fix the competitors' strategies in the intensity functions to be those from the previous iteration instead of optimizing all strategies simultaneously, as presented in (3.3.14). This effectively decouples high-dimensional optimization, allowing quoted spreads to be solved independently and efficiently in each inventory profile $\boldsymbol{q}$.

The details of the policy evaluation method are described in Algorithm 1.

Algorithm 1 Policy iteration of Pareto optimum for N market makers

Input: ϵ = error threshold, N =number of market makers, $\tilde{f}_a^i, \tilde{f}_b^i$, λ^a, λ^b: intensity of ask and bid order flow, Δ: unit order size, ψ_i: running cost for holding inventory to market maker i

Output: Approximated Pareto optimum quoting strategy

1: Initialize quoting strategies of N market makers, denoted by

$$\vec{\boldsymbol{\delta}}^{(0)} = \{(\vec{\delta}^{a,i,(0)}, \vec{\delta}^{b,i,(0)})\}$$

where $\vec{\delta}^{a,i,(0)} = \vec{\delta}^{a,(0)}, \vec{\delta}^{b,i,(0)} = \vec{\delta}^{b,(0)}, \forall i \in \{1, \dots, N\}$. Set $m \leftarrow 0$

2: **repeat**

3: **Policy evaluation**: Compute the values $\{W^{(m)}(\boldsymbol{q}), \boldsymbol{q} \in \prod_{j=1}^{N} \mathcal{Q}_j\}$ by solving the linear system (3.3.12) using joint quoting strategies $\vec{\boldsymbol{\delta}}^{(m)}$.

4: **Policy improvement**: Solve the pointwise optimization problems

$$\begin{aligned}
\left(\delta_{\boldsymbol{q}}^{a,(m+1)}\right)_{\boldsymbol{q}} =& \left(\arg\max_{\delta} \Big\{ \sum_{i=1}^{N} \mathbb{I}(q_i > -Z_i) \Big[\tilde{f}_a^i\big(\delta, ((\vec{\delta}_{\tilde{\boldsymbol{q}}}^{a,j,(m)})_{\tilde{\boldsymbol{q}} \in \mathcal{Q}})_{j \neq i}\big) \cdot \right. \\
& \left. \left(\delta - \frac{W^{(m)}(\boldsymbol{q}) - W^{(m)}(\boldsymbol{q} - \Delta \boldsymbol{e}_i)}{\Delta}\right) \Big] \Big\} \right)_{\boldsymbol{q}} \\
\left(\delta_{\boldsymbol{q}}^{b,(m+1)}\right)_{\boldsymbol{q}} =& \left(\arg\max_{\delta} \Big\{ \sum_{i=1}^{N} \mathbb{I}(q_i < Z_i) \Big[\tilde{f}_b^i\big(\delta, ((\vec{\delta}_{\tilde{\boldsymbol{q}}}^{b,j,(m)})_{\tilde{\boldsymbol{q}} \in \mathcal{Q}})_{j \neq i}\big) \cdot \right. \\
& \left. \left(\delta - \frac{W^{(m)}(\boldsymbol{q}) - W^{(m)}(\boldsymbol{q} + \Delta \boldsymbol{e}_i)}{\Delta}\right) \Big] \Big\} \right)_{\boldsymbol{q}} \qquad (3.3.14)
\end{aligned}$$

5: Increment m by 1: $m \leftarrow m + 1$.

6: **until** $m \geq 1$ and $\sum_{\boldsymbol{q}} |W^{(m+1)}(\boldsymbol{q}) - W^{(m)}(\boldsymbol{q})|^2 < \epsilon$

In general, the theoretical convergence of policy iteration can often be shown using techniques such as the contraction mapping theorem or by proving that the sequence of value functions generated by successive approximation converges monotonically to the optimal value function. For solving HJB equations that arise in optimal control problems, this usually relies on the regularity properties of the Hamiltonian that lead to a comparison principle ([Tran, Wang, and Zhang 2024]). For our specific problem, a rigorous convergence proof of the policy iteration scheme is challenging and remains an open question due to the discrete, high-dimensional state space and the Hamiltonian terms that include the directional difference of the value function, which does not necessarily hold a comparison principle. Instead, we focus on providing heuristic evidence of convergence through empirical results to solve (3.3.10) with Algorithm 1. We observe that the l^2-norm differences between the value functions and between the quoting strategies decrease significantly in the first few iterations and converge to levels close to 0, as shown in Figure 3.1. We also present the difference $\min_{\boldsymbol{q}} \big(W^{(m+1)} - W^{(m)}\big)(\boldsymbol{q})$ per iteration in Figure 3.1 that demonstrates that the value functions $\left(W^{(m)}\right)_m$ obtained from our policy iteration scheme form an increasing sequence. The computed Pareto

optimum quoting strategy is summarized in Figure 3.10 in comparison with the Nash equilibrium quoting strategy and the strategies given by the learning algorithms.

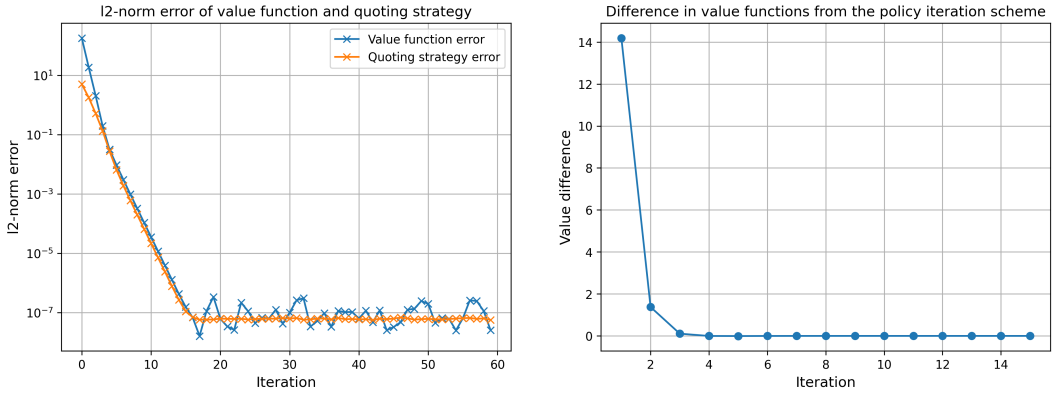


Figure 3.1: Left: The l^2-norm differences between the value functions and between the quoting strategies from the policy iteration scheme. Right: The difference $\min_{\boldsymbol{q}} \left(W^{(m+1)} - W^{(m)}\right)(\boldsymbol{q})$ per iteration from the policy iteration scheme.

As shown in Chapter 2, when market makers adjust their strategies by learning from market transactions, spreads can converge to levels consistent with collusion, even in the absence of coordination between market makers. We refer to this situation as *tacit collusion* [Tirole 1988]. As in [Calvano et al. 2020], our definition of tacit collusion is therefore operational, referring only to price outcomes.

3.4 Computation of Nash Equilibrium via Fictitious Play

Fictitious play is an iterative algorithm for computing Nash equilibrium. It was originally proposed in [Brown 1949] and [Brown 1951] to compute the value of a two-player zero-sum finite game, where two players each play iteratively the pure best response against the empirical distribution of their opponent's historical actions. The convergence of fictitious play for two-player zero-sum finite game is proved by [Robinson 1951]. [Monderer and Shapley 1996] proves convergence of fictitious play for potential games. However, there are no theoretical guarantees of convergence for general non-zero sum games, as is shown by a counterexample from [Shapley 1962]. Nevertheless, the idea of fictitious

play has been a standard tool of game theory for computing practically Nash equilibrium, and has been extended to broader applications such as mean field game learning problems (e.g. [Cardaliaguet and Hadikhanloo 2017], [Perrin et al. 2020]). In this section, we will be focusing on practical aspects, by applying fictitious play to compute numerically Nash equilibrium for multi-agent market making problem. Our numerical results in Figure 3.2-3.4 suggest fictitious play leads to convergent quotes. We shall leave theoretical convergence analysis of the fictitious play for our specific problem for future research.

Recently, fictitious play has been combined with deep learning methods to find (Markovian) Nash equilibria in dynamic games ([Han and Hu 2020]), with convergence analysis studied in [Han, Hu, and Long 2022]. [Hu 2021] has analyzed the convergence of a deep fictitious play algorithm to find open-loop Nash equilibria. Specifically, the deep fictitious play algorithm decouples the N-player game into N decision problems solved iteratively, in which each player solves the optimal strategy using deep neural networks, given that competitors' strategies remain fixed as in the previous iteration.

Inspired by [Han and Hu 2020], the fictitious play algorithm that we propose for the problem of N competing market makers is based on the systems of equations (3.2.14) and (3.2.15). For each equation in (3.2.15), the optimization problem at a given inventory level q is a single-variate optimization problem. Hence, at each iteration of fictitious play, one player needs to solve the single-variate optimization problem (3.4.5) for each of her inventory levels, while the competitors' strategies in the intensity functions are fixed as in the previous iteration. This single-variate optimization allows to avoid the curse of dimensionality due to the number of players and inventory levels of each player, because otherwise a policy evaluation method needs to optimize simultaneously the quotes of all market makers at all inventory levels at every iteration, leading to high-dimensional optimization problems. Note that the fictitious play algorithm is designed to numerically solve Nash equilibrium, instead of simulating real competition among market makers. In practice, market makers do not have information on their competitors' strategies, while in fictitious play each player optimizes for the next step based on competitors' current strategies. We shall tackle the simulation through Deep Reinforcement Learning in the next section.

Note that in the usual fictitious play algorithm, agents play the best response against the mixed strategy derived from their opponents' historical actions ([Brown 1949], [Brown 1951]). In our definition, best response is played

solely based on competitors' last stage actions. The algorithm based on last stage information is sometimes referred to as 'Iterated Best Response (IBR)' ([Lanctot et al. 2017]) or 'Best Response Dynamics' ([Roughgarden 2016]), in which an arbitrary agent is picked to update its best response to opponents' last-stage actions at each iteration. We shall adhere to our setting with the name 'fictitious play'. This will be explained in more detail in Remark 3.4.1.

Our fictitious play algorithm consists of iteratively executing 2 steps: **best response calculation** and **policy evaluation**. Policy evaluation computes the values $V_i^\delta(q_i)$ for joint quoting strategies based on the linear optimality equations (3.2.14). Given joint quoting strategies $(\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i})$, we reformulate (3.2.14) as a system of linear equations with values $V_i^\delta(q_i)$ being unknown variables. To simplify notations, we denote

$$f_a^i(\delta_{q_i}^{a,i}) = f_a^i(\delta_{q_i}^{a,i}, (\vec{\delta}^{a,j})_{j\neq i}), f_b^i(\delta_{q_i}^{b,i}) = f_b^i(\delta_{q_i}^{b,i}, (\vec{\delta}^{b,j})_{j\neq i})$$

The system of linear equations obtained is

$$\begin{cases} \left(r+\lambda^b f_b^i(\delta_{-Z_i}^{b,i})\right) V_i^\delta(-Z_i) - \lambda^b f_b^i(\delta_{-Z_i}^{b,i}) V_i^\delta(-Z_i+\Delta) = \lambda^b \Delta \delta_{-Z_i}^{b,i} f_b^i(\delta_{-Z_i}^{b,i}) - \psi_i(-Z_i) \\ -\lambda^a f_a^i(\delta_{Z_i}^{a,i}) V_i^\delta(Z_i-\Delta) + \left(r+\lambda^a f_a^i(\delta_{Z_i}^{a,i})\right) V_i^\delta(Z_i) = \lambda^a \Delta \delta_{Z_i}^{a,i} f_a^i(\delta_{-Z_i}^{b,i}) - \psi_i(Z_i) \\ -\lambda^a f_a^i(\delta_{q_i}^{a,i}) V_i^\delta(q_i-\Delta) + \left(r+\lambda^b f_b^i(\delta_{-Z_i}^{b,i}) + \lambda^a f_a^i(\delta_{Z_i}^{a,i})\right) V_i^\delta(q_i) - \lambda^b f_b^i(\delta_{q_i}^{b,i}) V_i^\delta(q_i+\Delta) \\ \qquad = \lambda^a \Delta \delta_{q_i}^{a,i} f_a^i(\delta_{q_i}^{b,i}) + \lambda^b \Delta \delta_{q_i}^{b,i} f_b^i(\delta_{q_i}^{b,i}) - \psi_i(q_i), \quad \text{if } q_i \in \mathcal{Q}_i \setminus \{-Z_i, Z_i\} \end{cases} \tag{3.4.1}$$

Let $\vec{V}_i^\delta := \left(V_i^\delta(-Z_i), V_i^\delta(-Z_i+\Delta), \ldots, V_i^\delta(Z_i-\Delta), V_i^\delta(Z_i)\right)$, then (3.4.1) can be formulated as the matrix representation.

$$M_i \cdot \vec{V}_i = A_i \tag{3.4.2}$$

where $M_i \in \mathbb{R}^{(2\frac{Z_i}{\Delta}+1)\times(2\frac{Z_i}{\Delta}+1)}$, $A_i \in \mathbb{R}^{2\frac{Z_i}{\Delta}+1}$.

$$M_i = \left[\begin{array}{c|c|c|c|c} r+\lambda^b f_b^i(\delta_{-Z_i}^{b,i}) & -\lambda^b f_b^i(\delta_{-Z_i}^{b,i}) & \cdots & 0 & 0 \\ -\lambda^a f_a^i(\delta_{-Z_i+\Delta}^{a,i}) & r+\lambda^b f_b^i(\delta_{-Z_i+\Delta}^{b,i}) + \lambda^a f_a^i(\delta_{-Z_i+\Delta}^{a,i}) & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & r+\lambda^b f_b^i(\delta_{Z_i-\Delta}^{b,i}) + \lambda^a f_a^i(\delta_{Z_i-\Delta}^{a,i}) & -\lambda^b f_b^i(\delta_{Z_i-\Delta}^{b,i}) \\ 0 & 0 & \cdots & -\lambda^a f_a^i(\delta_{Z_i}^{a,i}) & r+\lambda^a f_a^i(\delta_{Z_i}^{a,i}) \end{array}\right] \tag{3.4.3}$$

$$A_i = \begin{bmatrix} \lambda^b \Delta \delta_{-Z_i}^{b,i} f_b^i(\delta_{-Z_i}^{b,i}) - \psi_i(-Z_i) \\ \lambda^a \Delta \delta_{-Z_i+\Delta}^{a,i} f_a^i(\delta_{-Z_i+\Delta}^{b,i}) + \lambda^b \Delta \delta_{-Z_i+\Delta}^{b,i} f_b^i(\delta_{-Z_i+\Delta}^{b,i}) - \psi_i(-Z_i+\Delta) \\ \cdots \\ \lambda^a \Delta \delta_{Z_i-\Delta}^{a,i} f_a^i(\delta_{Z_i-\Delta}^{b,i}) + \lambda^b \Delta \delta_{Z_i-\Delta}^{b,i} f_b^i(\delta_{Z_i-\Delta}^{b,i}) - \psi_i(Z_i-\Delta) \\ \lambda^a \Delta \delta_{Z_i}^{a,i} f_a^i(\delta_{-Z_i}^{b,i}) - \psi_i(Z_i) \end{bmatrix} \tag{3.4.4}$$

Clearly, the matrix M_i is a diagonally dominant matrix, and hence is invertible. For given strategies, we can solve for value functions

$$V_i^{\delta} = M_i^{-1} \cdot A_i$$

Subsequently, in the best response calculation, each agent i solves her best response to $\boldsymbol{\delta}^{-i}$ using (3.4.5). [Han and Hu 2020] applies Deep BSDE method to solve this optimization problem. We directly apply the standard numerical optimization scheme since the objectives in (3.4.5) are single-variate, and hence not complicated with Assumption 3.1.3.

The details of fictitious play are presented in Algorithm 2.

Algorithm 2 Fictitious play for computation of Nash equilibrium for N market makers

Input: M =number of iterations, N =number of market makers, f_a^i, f_b^i, λ^a, λ^b: intensity of ask and bid order flow, Δ: unit order size, ψ_i: running cost for holding inventory to market maker i

Output: Approximated Nash equilibrium by fictitious play.

1: Initialize quoting strategies of N market makers, denoted by $\{(\vec{\delta}^{a,i,(0)}, \vec{\delta}^{b,i,(0)})\}$.

2: Compute initial values $\{V_i^{(0)}(q_i), q_i \in \mathcal{Q}_i, i \in \{1, \dots, N\}\}$ by solving linear system (3.4.1)

3: **for** $m \leftarrow 0$ to M-1 **do**

4: **Best response**: Solve optimal strategy for single market maker at every inventory level:

5: **for** $i \leftarrow 1$ to N **do**

6: **for** $q_i \in \mathcal{Q}_i$ **do**

7: Update

$$\begin{aligned}
\delta_{q_i}^{a,i,(m+1)} &= \arg\max_{\delta} f_a^i(\delta, (\vec{\delta}^{a,j,(m)})_{j\neq i})(\delta - \frac{V_i^{(m)}(q_i) - V_i^{(m)}(q_i - \Delta)}{\Delta}) \\
\delta_{q_i}^{b,i,(m+1)} &= \arg\max_{\delta} f_b^i(\delta, (\vec{\delta}^{b,j,(m)})_{j\neq i})(\delta - \frac{V_i^{(m)}(q_i) - V_i^{(m)}(q_i + \Delta)}{\Delta})
\end{aligned} \tag{3.4.5}$$

8: **end for**

9: **end for**

10: **Policy evaluation**: Compute the values $\{V_i^{(m+1)}(q_i), q_i \in \mathcal{Q}_i, i \in \{1, \dots, N\}\}$ by solving the linear system (3.4.1) using updated strategies

$$\{(\vec{\delta}^{a,i,(m+1)}, \vec{\delta}^{b,i,(m+1)})\}$$

11: **end for**

Remark 3.4.1. We are following the definition of fictitious play from [Han and Hu 2020], [Hu 2021] and [Han, Hu, and Long 2022], which is different from that originated from [Brown 1949] and [Brown 1951], in that the N agents update simultaneously their best responses against their competitors' pure strategies from the last stage. Alternately, one can choose to compute the best response against the average of opponents' past strategies in the spirit of classical fictitious play [Brown 1951]. However, as [Hu 2021] points out, convergence with the last stage information generally implies convergence with the average of past strategies, and switching to the latter tends to a better convergence rate for certain circumstances, but with additional computational cost ([Han and Hu 2020]). For our setting, since the problem we study already exhibits convergence with last stage information within a reasonable number of iterations, we adhere to this setting of best response using last stage information.

By Lemma A.2.1 there exist unique solutions $\delta_{q_i}^{a,i,(m+1)}, \delta_{q_i}^{b,i,(m+1)}$ to optimization problems in (3.4.5). Let $K_i = \sum_{j\neq i, j\in\{1,\dots,N\}} (2\frac{Z_i}{\Delta} + 1)$. K_i is the total number of possible inventory levels of the competitors of the market maker i'. We consider f_a^i and f_b^i of the form:

$$f_a^i(\delta, (\vec{\delta}^{a,j})_{j\neq i}) = \frac{1}{2} \cdot \frac{1}{1+\exp(\delta)} \cdot \frac{\exp(\frac{1}{K_i}\sum_{j\neq i}\sum_{q_j\in\mathcal{Q}_j} \delta_{q_j}^{a,j})}{1+\exp(\delta + \frac{1}{K_i}\sum_{j\neq i}\sum_{q_j\in\mathcal{Q}_j} \delta_{q_j}^{a,j})}$$
$$f_b^i(\delta, (\vec{\delta}^{b,j})_{j\neq i}) = \frac{1}{2} \cdot \frac{1}{1+\exp(\delta)} \cdot \frac{\exp(\frac{1}{K_i}\sum_{j\neq i}\sum_{q_j\in\mathcal{Q}_j} \delta_{q_j}^{b,j})}{1+\exp(\delta + \frac{1}{K_i}\sum_{j\neq i}\sum_{q_j\in\mathcal{Q}_j} \delta_{q_j}^{b,j})} \tag{3.4.6}$$

Since (3.4.6) is a special case of (3.1.8), we can verify that the execution probabilities (3.4.6) satisfy Assumptions 3.1.1 and 3.1.3 from Proposition 3.1.5. We also assume that each of the 2 market makers has the same running cost function $\psi_1(q) = \psi_2(q) = \frac{1}{2} \times 0.01q^2$. Market makers have the same inventory risk limit $Q_1 = Q_2 = 5$. Order size $\Delta = 1$. Interest rate $r = 0.01$ and order flow arrival intensities $\lambda^a = \lambda^b = 2$. For simplicity, we assume the units of all parameters are already scaled so that the unit of ask and bid quotes is basis point (bps). We apply fictitious play to the game with 2 market makers and compare the ask/bid quotes with a benchmark model where there is only 1 monopolistic market maker with the ask and bid intensity $\Lambda(\delta) = \frac{1}{1+e^{\delta}}$ where δ is the ask or bid quote of this monopolistic market maker. Figure 3.2 compares the equilibrium quotes obtained from the fictitious play algorithm when

there are 2 market makers, with the optimal centered quotes when there is one monopolistic market maker. The results show lower Nash equilibrium quotes compared to the monopolistic market maker's quotes, suggesting competition introduced through the execution probabilities (3.4.6) results in more competitive quotes than the monopolistic market maker case. Figure 3.2 also suggests skewing behavior by market makers caused by exposure to market risk of their nonzero inventories.

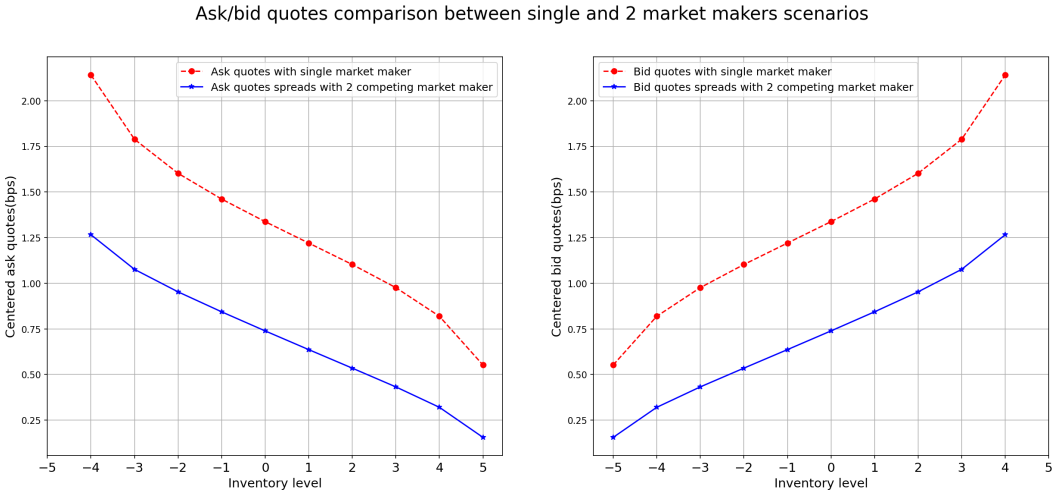


Figure 3.2: Comparison between 2-market-maker equilibrium ask and bid quotes and single-market-maker's quotes. The ask and bid quotes are centered at mid price.

In Figure 3.3, we plot the evolution of the l^2 norm error in the value functions and the quoting strategies by iteration to illustrate how the fictitious play algorithm leads to a stable state in our experiment setting. The equilibrium value function is also shown in Figure 3.3. The value function achieves its maximum at the 0 inventory level, which is in line with the market makers who expect a higher gain by keeping the inventory close to 0 since they pay less running cost for maintaining inventory and are exposed to a lower market risk on their inventories.

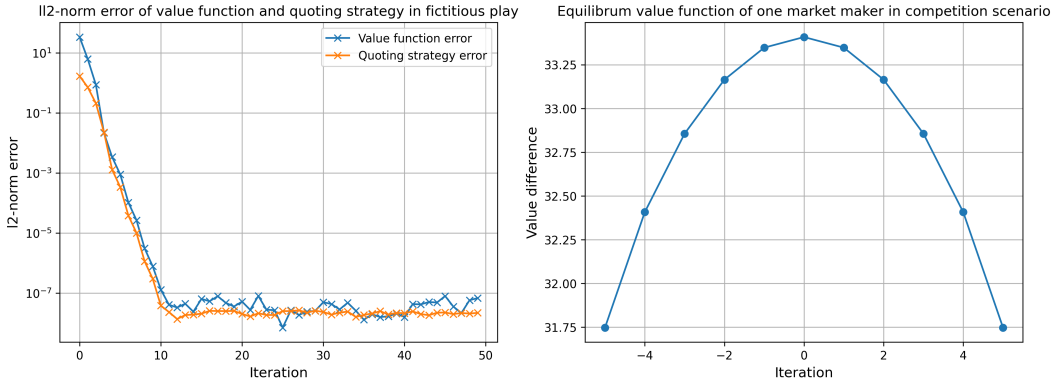


Figure 3.3: **Left**: evolution of the l^2 norm error in the value functions and the quoting strategies during fictitious play iterations; **Right**: equilibrium value function obtained from fictitious play algorithm.

To study the effects by the number of market makers in competition, Figure 3.4 compares the equilibrium quotes in scenarios of 2 and 5 market makers. Figure 3.4 suggests that more market makers tend to have more competitive quotes than the case of 2 market makers. The equilibrium quotes are lower at the same inventory level when the number of market makers increases to 5. However, the value of the quotes does not show an explicit decreasing trend in half of the inventory levels when there are 10 market makers. We think that this could be explained by the market makers' compensating effect from aggressively quoting negative quotes at the other half of inventory levels when there are many market makers. For example, the 10 market makers tend to keep their ask quotes high in negative inventories because they quote excessively negative values in positive inventories. Although the ask quotes are always above the bid quotes given by the learning algorithm, it should be noted that with 5 and 10 competing market makers when the inventory level is at the boundary values Q and $-Q$, the market makers even aggressively quote negative ask and bid quotes. This is due to the excessive running cost of holding larger inventory so that market makers would rather change their inventory state at the cost of losing profit from making the spread. The phenomenon of negative ask or bid quotes is even more remarkable when the number of market makers reaches 10, implying that when there are many market makers, they are more risk averse for holding nonzero inventory.

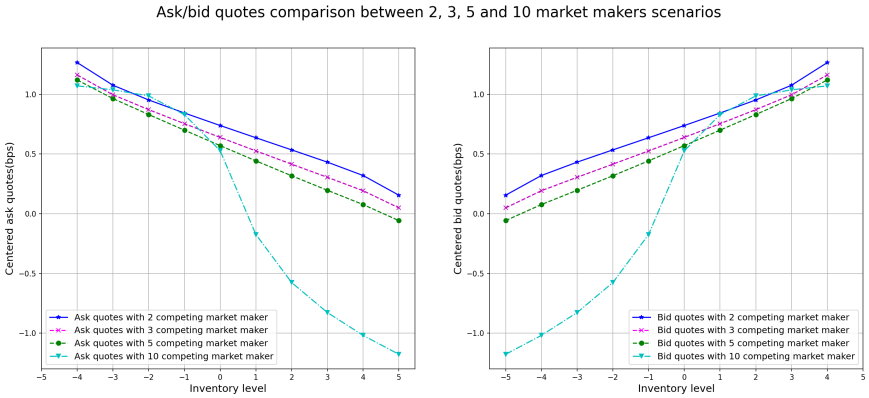


Figure 3.4: Comparison of equilibrium quotes with different number of market makers. The 0 value in vertical axis represents mid price.

3.5 Decentralized Learning and the Emergence of Tacit Collusion

The notions of Nash equilibrium and Pareto optimum described above refer to outcomes, but do not attempt to describe the process through which agents arrive at their quoting strategies. In practice, market makers rely on algorithms that update their quotes based on observed market data, and an interesting question is to understand whether agents' strategies converge to any stationary configuration in this learning process and how this limit can be characterized.

Market makers update their quotes dynamically and receive feedback in the form of profit from market transactions, based on which they adjust their quote. The automation of these market making algorithms naturally points to the use of Reinforcement Learning (RL) algorithms to model the resulting dynamics.

Also, in dealer markets, the market makers are not allowed to communicate and make decisions based on partial information related to their own inventory, quotes, and profits. These features require the RL algorithm to adapt to continuous state or action spaces. Hence we focus on *decentralized multi-agent policy gradient algorithms*, which account for a type of deep reinforcement learning algorithm that allows for continuous state and action spaces.

For each market maker, we introduce 2 types of neural networks: the critic and actor networks. The critic network is used to approximate the state-action value function $V_i^{\delta}(q_i)$ while the actor network approximates optimal quoting

strategies. The algorithm we apply is called decentralized Multi-Agent Deep Deterministic Policy Gradient (Decentralized MADDPG), inspired by [Lowe et al. 2017] and [Foerster et al. 2018]. [Lowe et al. 2017] and [Foerster et al. 2018] proposed a multi-agent actor-critic algorithm with decentralized actors and centralized critics, in which the actors are functions of each agent's local observation, and critics are functions of all agents' joint states and actions. In their algorithms, critics are trained in a centralized way, which implies certain communication during training steps needs to be allowed. We adapt the MADDPG algorithm to our market making model by constraining the critics to be decentralized as well.

An important feature of our Decentralized MADDPG algorithm is pre-training of critic and actor networks for initialization. Pre-training is important for the convergence of quoting strategies and has been considered by [Guéant and Manziuk 2019]. Recall that in Section 3.1 the upper bound $\Lambda(\delta)$ can be considered as the execution probability of a monopolistic market maker. We pre-train the critic networks for each market maker to the value function of a single monopolistic market maker with intensity $\Lambda(\delta)$, and pre-train the actor networks to the quoting strategies of the same monopolistic market maker, which is shown by the red curve in Figure 3.2. This pre-training step can be implemented by supervised learning on neural networks, since the value function and quoting strategy in the monopolistic case can be explicitly calculated. We think the motif for pre-training is consistent with practical scenarios in that even though market makers do not have information on the mechanics that influence their market shares, each market maker is supposed to have a prior estimate on the general form of the execution probability when there is no other competitor. In the meantime, the actor network is initialized by pre-training so that the ask and bid quotes are within a reasonable value range. For simplicity, we assume market makers know Λ.

3.5.1 Reformulation to Discrete-time Problem

Equations (3.2.14) and (3.2.15) are local versions of the Dynamic Programming Principle. To adapt to RL simulation, we first need to formulate the multi-agent market making problem into discrete-time Bellman equations. In the following derivations, we always assume that the joint quoting strategies are given and fixed $\vec{\boldsymbol{\delta}} \in \prod_{j=1}^{N} (I_\delta)^{2\frac{Z_j}{\Delta}+1}$. The corresponding state-action value functions are denoted by $V_i^\delta(q_i)$ where i refers to the index of market maker,

and q_i is the inventory level of market maker i at time 0. Let $\mathbb{E}_i[\cdot]$ denote the conditional expectation $\mathbb{E}[\cdot|q_0^i = q_i]$. From (3.1.12) we have

$$V_i^\delta(q_i) = \mathbb{E}_i\left[\int_0^\infty e^{-rt}\Big(\delta^{a,i}_{q^i_{t-}} N^{a,i}(dt) + \delta^{b,i}_{q^i_{t-}} N^{b,i}(dt)\Big) - \int_0^\infty e^{-rt}\psi_i(q_t^i)dt\right] \tag{3.5.1}$$

where $N^{a,i}(dt)$ and $N^{b,i}(dt)$ are order flow to market makers i whose intensities are defined in (3.1.4). Define 2 stopping times τ_a and τ_b denoting the arrival time of the first ask and bid RFQ after 0, that is

$$\begin{aligned}\tau_a &:= \inf\{t > 0, \int_0^t N^a(dt) > 0\}\\ \tau_b &:= \inf\{t > 0, \int_0^t N^b(dt) > 0\}\end{aligned} \tag{3.5.2}$$

Let $\tau := \tau_a \wedge \tau_b$ denote the first arrival time of an RFQ received by all market makers simultaneously. From the assumption on the independence of the ask and bid RFQs, τ_a and τ_b are independent random variables. We state a lemma on the probability distribution of τ_a and τ_b.

Lemma 3.5.1. *τ_a, τ_b are independent variables of exponential distribution with parameters λ_a and λ_b, respectively. Moreover,*

$$\begin{aligned}\mathbb{E}\left[\int_0^{\tau_a\wedge\tau_b} e^{-rt}dt\right] &= \frac{1}{r+\lambda_a+\lambda_b}\\ \mathbb{E}\left[e^{-r\tau_a}|\tau_a < \tau_b\right] = \mathbb{E}\left[e^{-r\tau_b}|\tau_b < \tau_a\right] &= \frac{\lambda_a+\lambda_b}{r+\lambda_a+\lambda_b}\end{aligned} \tag{3.5.3}$$

The proof of Lemma 3.5.1 follows easily from the fundamental calculation with exponentially distributed random variables.

By Dynamic Programming Principle, we obtain

$$\begin{aligned}V_i^\delta(q_i) =&\mathbb{E}_i\Big[-\int_0^\tau e^{-rt}\psi_i(q_i)dt + \int_0^\tau e^{-rt}\Big(\delta^{a,i}_{q^i_{t-}} N^{a,i}(dt) + \delta^{b,i}_{q^i_{t-}} N^{b,i}(dt)\Big)\\ &+ e^{-r\tau}V_i^\delta(q_\tau^i)\Big]\end{aligned} \tag{3.5.4}$$

We introduce random events R_a^i, R_b^i that indicate whether the market maker i wins the ask/bid RFQ.

$$\begin{aligned}R_a^i &:= \{\text{Market maker } i \text{ wins the ask RFQ}\}\\ R_b^i &:= \{\text{Market maker } i \text{ wins the bid RFQ}\}\end{aligned} \tag{3.5.5}$$

Upon arrival of an RFQ, the market maker i only profits if she wins the trade, that is when R_a^i or R_b^i takes place. Hence, we discuss different possible values for q_τ^i, with inventory risk limit taken into consideration.

If $-Z_i < q_i < Z_i$,

$$q_\tau^i(\omega) = \begin{cases} q_i - \Delta, & \text{if } \omega \in R_a^i \\ q_i + \Delta, & \text{if } \omega \in R_b^i \\ q_i, & \text{if } \omega \in (R_a^i)^c \cap (R_b^i)^c \end{cases} \tag{3.5.6}$$

If $q_i = -Z_i$,

$$q_\tau^i(\omega) = \begin{cases} -Z_i + \Delta, & \text{if } \omega \in R_b^i \\ -Z_i, & \text{if } \omega \in (R_b^i)^c \end{cases} \tag{3.5.7}$$

If $q_i = Z_i$,

$$q_\tau^i(\omega) = \begin{cases} Z_i - \Delta, & \text{if } \omega \in R_a^i \\ Z_i, & \text{if } \omega \in (R_a^i)^c \end{cases} \tag{3.5.8}$$

We have an important proposition relating $V_i^\delta(q_i)$ in terms of $\tau_a, \tau_b, R_a^i, R_b^i$. The proof of Proposition 3.5.2 is stated in Appendix A.3.

Proposition 3.5.2. *The state-action value function $V_i^\delta(q_i)$ satisfies*

$$\begin{aligned} V_i^\delta(q_i) = &- \frac{\psi_i(q_i)}{r + \lambda_a + \lambda_b} + \mathbb{P}(\tau_a < \tau_b)\mathbb{E}_i\Big[\mathbb{I}(R_a^i)\left(e^{-r\tau_a}\delta_{q_i}^{a,i}\Delta + e^{-r\tau_a}V_i^\delta(q_i - \Delta)\right)\cdot \\ &\mathbb{I}(-Z_i < q_i \le Z_i) + \mathbb{I}((R_a^i)^c)e^{-r\tau_a}V_i^\delta(q_i)\Big|\tau_a < \tau_b\Big] \\ &+ \mathbb{P}(\tau_b < \tau_a)\mathbb{E}_i\Big[\mathbb{I}(R_b^i)\big(e^{-r\tau_b}\delta_{q_i}^{a,i}\Delta + e^{-r\tau_b}V_i^\delta(q_i + \Delta)\big)\mathbb{I}(-Z_i \le q_i < Z_i) \\ &+ \mathbb{I}((R_b^i)^c)e^{-r\tau_b}V_i^\delta(q_i)\Big|\tau_b < \tau_a\Big] \end{aligned} \tag{3.5.9}$$

The Bellman equation (3.5.9) is expressed in terms of conditional probabilities. The arrival of RFQ is incorporated into the simulation of the market environment, and hence it is exogenous to the market makers. The market makers modelled as learning agents receive the RFQ and improve their strategies through interactions with the simulated market environment. Assuming that ask and bid RFQs are independent, we have $\mathbb{P}(\tau_a < \tau_b) = \frac{\lambda_a}{\lambda_a+\lambda_b}$ and $\mathbb{P}(\tau_b < \tau_a) = \frac{\lambda_b}{\lambda_a+\lambda_b}$. We consider every iteration of the RL learning algorithm as an arrival of an RFQ with ask and bid requests at probabilities $\frac{\lambda_a}{\lambda_a+\lambda_b}$ and $\frac{\lambda_b}{\lambda_a+\lambda_b}$. Upon each arrival of the RFQ, the probability that the market maker i completes the transaction is proportional to the market share f_a^i or f_b^i depending on the sides of the RFQ.

Recall that the state space of the market maker i consists of all her possible inventory values $\mathcal{Q}_i = \{-Z_i, -Z_i + \Delta, \ldots, Z_i - \Delta, Z_i\}$. We consider discretized time steps $t = 0, 1, \ldots$, as the arrival of RFQs. Given the inventory state $q_t^i \in \mathcal{Q}_i$ at t, the market maker i sets up her ask and bid quotes $\left(\delta_{q_{t-}^i}^{a,i}, \delta_{q_{t-}^i}^{b,i}\right) \in I_\delta \times I_\delta$.

The quoting strategies $\left(\delta^{a,i}_{q^i_{t-}}, \delta^{b,i}_{q^i_{t-}}\right)$ are alternatively denoted as functions $\delta^i_a : q \in \mathcal{Q}_i \to \mathbb{R}, \delta^i_b : q \in \mathcal{Q}_i \to \mathbb{R}$. This function representation transmits quoting strategies to the neural network approximation, which will be introduced in Section 3.5.2.

The dealer market environment generates an ask or bid RFQ at each unit time step with probability $\frac{\lambda_a}{\lambda_a+\lambda_b}$ and $\frac{\lambda_b}{\lambda_a+\lambda_b}$.[5] The RFQ is sent to N market makers simultaneously. At time step t market maker i will set up centered ask quote $\delta^i_a = \pi^i_a(q^i_t|\theta^\pi_i)$ and centered bid quote $\delta^i_b = \pi^i_b(q^i_t|\theta^\pi_i)$. The market maker that executes the RFQ order is selected stochastically by the market environment. Recall the random events R^i_a, R^i_b defined in (3.5.5). $\{R^i_a, i \in \{1, \ldots, N\}\}$ form a collection of mutually exclusive pairwise events. This is the same case for $\{R^i_b, i \in \{1, \ldots, N\}\}$. We have the following conditional probability for events R^i_a, R^i_b.

$$\mathbb{P}(R^i_a|\tau_a < \tau_b) = \frac{f^i_a(\delta^i_a(q^i_t), \cdot)}{\sum_{j=1}^N f^j_a(\delta^j_a(q^j_t), \cdot)} \quad \text{(if RFQ is on ask side)} \tag{3.5.10}$$

$$\mathbb{P}(R^i_b|\tau_b < \tau_a) = \frac{f^i_b(\delta^i_b(q^i_t), \cdot)}{\sum_{j=1}^N f^j_b(\delta^j_b(q^j_t), \cdot)} \quad \text{(if RFQ is on bid side)}a \tag{3.5.11}$$

Again for simplicity in the notations, the dependence of f^i_a, f^i_b on competitors' quoting strategies is replaced with the symbol '$\cdot$' in (3.5.10). We apply this probability distribution based on Assumption 3.1.1 with $\sum_{j=1}^N f^i_a \leq 1, \sum_{j=1}^N f^i_b \leq 1$. We do not consider the circumstance where no market maker wins the RFQ, which has probability $1 - \sum_{j=1}^N f^i_a$ for an ask RFQ and $1 - \sum_{j=1}^N f^i_b$ for a bid RFQ. Hence, we normalize the execution probabilities in (3.5.10) so that they form probability distributions of random choice. Note that we hereby make a simplification that if market maker i wins the trade and her inventory is at the risk limit $\pm Z_i$, then i will not execute the trade that would drive her inventory out of the risk limit. For instance, when i has inventory $-Z_i$ she will not quote for an ask RFQ. In this case, the market environment will switch to the next time step that generates a new RFQ and selects a new market

[5] This is a simplification on arrival of order flows for tractability in Reinforcement Learning simulation. The arrival time τ_a, τ_b of RFQs follows exponential distributions. Instead of directly simulating τ_a, τ_b we consider RFQ arrivals in the sense of probabilistic expectation, assuming there is an RFQ arrival at each time step in RL simulation, which is more tractable.

maker stochastically for execution. This simplification is also reflected from conditional probabilities in equation (3.5.9).

Based on equation (3.5.9), we now define reward functions in the market environment. All market makers will have to pay the expected running cost.

$$\mathbb{E}_i\left[-\int_0^{\tau_a\wedge\tau_b} e^{-rt}dt\right]\psi_i(q_i) = -\frac{\psi_i(q_i)}{r+\lambda_a+\lambda_b}$$

where q_i is the inventory of the market maker indexed by $i \in \{1,\ldots,N\}$.

It is clear that at time t only the market maker who wins the RFQ receives the revenue from the spread. Hence, the market maker i's reward function consists of the inventory cost she has paid plus the time-discounted revenue from making the spread had she won the corresponding RFQ. Taking into account the risk limit and our simplification, we write formally our reward function $r_i(q_i,(\delta_a^i,\delta_b^i))$ for the market maker i.

$$\begin{aligned} r_i(q_i,(\delta_a^i,\delta_b^i)) = -\frac{\psi_i(q_i)}{r+\lambda_a+\lambda_b} + \frac{(\lambda_a+\lambda_b)\Delta}{r+\lambda_a+\lambda_b}\Big(&\mathbb{I}(R_a^i)\mathbb{I}(-Z_i < q_i \le Z_i)\cdot\delta_a^i \\ &+ \mathbb{I}(R_b^i)\mathbb{I}(-Z_i \le q_i < Z_i)\cdot\delta_b^i\Big) \end{aligned} \tag{3.5.12}$$

In terms of time steps in the simulation of RL, we include t as a subscript in the state q and the action δ, so that in the time step $t = 0, 1, \ldots$, the market maker i disposes of the inventory level q_t^i and sets the action ϱ_t^i. This notation should not cause ambiguity between subscripts used for inventory and quotes in Sections 3.1 to 3.2. Between time steps t and $t+1$ only the inventory of the market maker that wins the trade will possibly change depending on whether it reaches the inventory risk limit. The transition of market maker i's inventory q_t^i to q_{t+1}^i can be summarized depending on the side of the RFQ.

$$q_{t+1}^i = \begin{cases} q_t^i - \Delta\mathbb{I}(R_a^i)\mathbb{I}(q_t^i > -Z_i) & \text{for RFQ from ask side} \\ q_t^i + \Delta\mathbb{I}(R_b^i)\mathbb{I}(q_t^i < Z_i) & \text{for RFQ from bid side} \end{cases} \tag{3.5.13}$$

(3.5.10)-(3.5.13) define a Partially Observed Markov Decision Process (POMDP). Define the discount factor $\gamma = \frac{\lambda_a+\lambda_b}{r+\lambda_a+\lambda_b}$, with strong Markov property we have the following proposition that the objective function of market maker i can be written as a discrete-time format.

Proposition 3.5.3. *The state-action value function $V_i^{\delta}(q_i)$ in (3.5.1) can be written in the following format:*

$$V_i^{\delta}(q_i) = \mathbb{E}_i \left[\sum_{t=0}^{\infty} \gamma^t r_i \Big(q_t^i, \big(\delta_a^i(q_t^i), \delta_b^i(q_t^i) \big) \Big) \Big| q_0^i = q_i \right] \tag{3.5.14}$$

where $\gamma = \frac{\lambda_a + \lambda_b}{r + \lambda_a + \lambda_b}$ and reward functions $r_i(q_i, (\delta_a^i, \delta_b^i))$ are defined in (3.5.12).

The objective of the market maker i is to find the optimal quoting strategy $\delta_a^{i,*}, \delta_b^{i,*}$, which maximizes the expected future rewards discounted by a factor $\gamma = \frac{\lambda_a + \lambda_b}{r + \lambda_a + \lambda_b}$.

$$V_i(q_i) = \max_{\delta_a^i, \delta_b^i} \mathbb{E}_i \left[\sum_{t=0}^{\infty} \gamma^t r_i \Big(q_t^i, \big(\delta_a^i(q_t^i), \delta_b^i(q_t^i) \big) \Big) \Big| q_0^i = q_i \right] \tag{3.5.15}$$

3.5.2 The Multi-agent DDPG Algorithm for Market Makers

Now we elaborate on more details of our Decentralized Multi-agent DDPG algorithm. Our MADDPG algorithm is a type of actor-critic learning algorithm. The strategies of the market maker i, $\delta_a^i : q \in \mathcal{Q}_i \to \mathbb{R}$ and $\delta_b^i : q \in \mathcal{Q}_i \to \mathbb{R}$, are approximated by neural networks $\pi_a^i(q|\theta_i^{\pi}), \pi_b^i(q|\theta_i^{\pi})$. These are called actor networks. Similarly, the state-action value functions $V_i^{\delta}(q)$ of the market maker i are approximated by neural networks $Q_i(q, (\delta^a, \delta^b)|\theta_i^Q)$ where q is the inventory level, (δ^a, δ^b) are the ask and bid quotes, and θ_i^Q is the collection of network parameters. Neural networks $Q_i(q, (\delta^a, \delta^b)|\theta_i^Q)$ are named critic networks, which evaluate a given tuple of state-action combination $(q, (\delta^a, \delta^b))$. The network parameters θ_i^Q and θ_i^{π} are learned via interactions with the market environment, which sends RFQs to market makers, executes the transaction according to the execution probabilities of the market makers, and provides feedback rewards to each market maker. For the critic networks Q_i, Temporal Difference (TD) learning is applied to update the critic parameters θ_i^Q, while for the actor networks π_a^i, π_b^i, their parameters θ_i^{π} are updated by Stochastic Gradient Descent (SGD) step to minimize a given loss function. For simplicity in the notation, we omit parameters θ_i^Q and θ_i^{π} in neural networks Q_i and (π_a^i, π_b^i) when there is no ambiguity. Since quotes or actions (δ_a^i, δ_b^i) come from both the ask and the bid sides, to simplify the notation, we sometimes denote the action of the market maker i by $\delta^i = (\delta_a^i, \delta_b^i)$.

In addition to primal critic networks Q_i and actor networks δ_a^i, δ_b^i, target critic and actor networks are introduced coupling the critics and actors in

implementation. The target networks are indicated by $\tilde{Q}_i(q,(\delta^a,\delta^b)|\tilde{\theta}_i^Q)$ and $\tilde{\pi}_a^i(q|\tilde{\theta}_i^\pi), \tilde{\pi}_b^i(q|\tilde{\theta}_i^\pi)$. The introduction of a target network is a common practice in the implementation of deep reinforcement learning. It is intended for a more stable parameter update. While primal network parameters θ_i^Q and θ_i^π are updated at every training iteration, the target network parameters $\tilde{\theta}_i^Q$ and $\tilde{\theta}_i^\pi$ are updated slowly to make the training more stationary:

$$\begin{aligned} \tilde{\theta}_i^Q &\leftarrow \mu\theta_i^Q + (1-\mu)\tilde{\theta}_i^Q \\ \tilde{\theta}_i^\pi &\leftarrow \mu\theta_i^\pi + (1-\mu)\tilde{\theta}_i^\pi \end{aligned} \tag{3.5.16}$$

where μ represents the speed of updating target network parameters. A value μ close to 0 means a very slow transition of parameters obtained from the training steps to the target network parameters.

At each time step t, denote m_t the side of the RFQ, which takes values from $\{a,b\}$ with a for the ask RFQ and b for bid RFQ. Denote by $I_t^i \in \{0,1\}$ the indicator function whether the market maker i wins the RFQ. $I_t^i = 1$ suggests that the market maker i wins the RFQ at time step t. The interactions between the market maker i and the market environment generate a tuple of data $(q_t^i, \varrho_t^i, q_{t+1}^i, r_i(q_t^i,\varrho_t^i), I_t^i, m_t)$. In this data tuple q_t^i is market maker i's inventory level at t, ϱ_t^i refers to market maker i's centered ask and bid quotes given inventory q_t^i, q_{t+1}^i is the inventory level after taking action ϱ_t^i, and $r_i(q_t^i,\varrho_t^i)$ is the reward from the environment to the market maker i for state-action combination (q_t^i,ϱ_t^i). This data tuple is stored in an experience replay buffer of the corresponding agent.

Experience replay is a technique initially proposed by [Mnih, Kavukcuoglu, Silver, Graves, et al. 2013], widely used in reinforcement learning algorithms to improve the efficiency and stability of the learning process. It involves storing past experiences including tuples of state, action, reward, and next state into an experience replay buffer. Instead of learning only from the most recent experiences, the learning agent randomly samples batches of experiences from the replay buffer to update its parameters. The use of the replay buffer allows for more effective use of historical data, and breaks the temporal correlation between consecutive experiences in online reinforcement learning, which could lead to non-stationary distribution of learned policy. In fact, for parameter tuning, mini-batch data points are sampled from this experience replay buffer randomly, hence unbinding the sequential correlation of data samples from interactions with the environment in the hope that more robust quoting strategies can be learned when the agent is provided with more diverse training

samples. In our experiment, we assign the replay buffer to have a fixed length with first-in-first-out queue structure.

For parameter update, each iteration consists of a Temporal Difference (TD) learning phase and a policy improvement phase. In the TD learning phase, the parameters of the critic networks θ_i^Q are updated for every $i \in \{1, \ldots, N\}$, after which the policy improvement updates the parameters of the actor network θ_i^π. We first implement a mini-batched stochastic gradient descent on the following loss function for parameters θ_i^Q. This loss function is based on dynamic programming equations (A.3.1)-(A.3.3) with data sampled from experience replay.

$$\begin{aligned}\mathcal{L}_i^Q(\theta_i^Q) = \quad & \mathop{\mathbb{E}}_{q_i,\delta^i,q_i',I_i}\Bigg[\Bigg(r_i(q_i,\delta^i) + \gamma\Bigg(I_i\tilde{Q}_i\Big(q_i', \big(\tilde{\pi}_a^i(q_i'), \tilde{\pi}_b^i(q_i')\big)\,\Big|\,\tilde{\theta}_i^Q\Big) \\ & +(1-I_i)\tilde{Q}_i\Big(q_i, \big(\tilde{\pi}_a^i(q_i), \tilde{\pi}_b^i(q_i)\big)\,\Big|\,\tilde{\theta}_i^Q\Big)\Bigg) \\ & -Q_i\left(q_i, (\tilde{\pi}_a^i(q_i), \tilde{\pi}_b^i(q_i))|\theta_i^Q\right)\Bigg)^2\Bigg] \end{aligned} \tag{3.5.17}$$

where pairs $(q_i, \delta^i, q_i', I_i)$ are sampled from the experience replay buffer that stores the historical states, actions, and transitions of the market maker during the training steps. The reward $r_i(q_i, \delta^i)$ is calculated depending on ask or bid RFQ, based on (3.5.12). The actions $\tilde{\pi}_a^i(q_i'), \tilde{\pi}_b^i(q_i')$ are given by the target actor network of agent i. Then the stochastic gradient descent can be applied on (3.5.17) to calibrate the parameter θ_i^Q. Note that the parameters $\tilde{\theta}_i^Q$ of the target critic networks $\tilde{Q}_i$ are updated at this stage. The method applied to calibrate θ_i^Q is called Temporal Difference (TD) learning. More specifically, we apply a mini-batched stochastic gradient descent. Let K be the size of the minibatch. We sample K data points from the replay buffer of the experience, indicated by $\{(q_i^{(k)}, (\delta^i)^{(k)}, q_i^{\prime(k)}, I_i^{(k)}), k \in \{1, \ldots, K\}\}$, a stochastic gradient descent is implemented with a given learning rate α_Q.

$$\begin{aligned}\theta_i^Q \leftarrow \theta_i^Q + \alpha_Q \frac{1}{K}\sum_{k=1}^{K}\Bigg(\hat{\beta}_k - Q_i\left(q_i^{(k)}, (\tilde{\pi}_a^i(q_i^{(k)}), \tilde{\pi}_b^i(q_i^{(k)}))\right)\Bigg)\cdot \\ \nabla_{\theta_i^Q} Q_i\left(q_i^{(k)}, (\tilde{\pi}_a^i(q_i^{(k)}), \tilde{\pi}_b^i(q_i^{(k)}))|\theta_i^Q\right)\end{aligned} \tag{3.5.18}$$

where

$$\hat{\beta}_k = r_i(q_i^{(k)}, (\delta^i)^{(k)}) + \gamma \Bigg(I_i^{(k)} \tilde{Q}_i \Big(q_i'^{(k)}, \Big(\tilde{\pi}_a^i(q_i'^{(k)}), \tilde{\pi}_b^i(q_i'^{(k)}) \Big) \Big| \tilde{\theta}_i^Q \Big) + (1 - I_i^{(k)}) \tilde{Q}_i \Big(q_i^{(k)}, \Big(\tilde{\pi}_a^i(q_i^{(k)}), \tilde{\pi}_b^i(q_i^{(k)}) \Big) \Big| \tilde{\theta}_i^Q \Big) \Bigg) \tag{3.5.19}$$

When TD learning is performed, the policy improvement step updates parameters θ_i^π to improve the quoting strategy of each market maker. To calibrate the parameters of the actor network θ_i^π, the loss function we use is the value of the critic network.

$$\mathcal{L}_i(\theta_i^\pi) = -\mathop{\mathbb{E}}_{q_i} \Big[Q_i \Big(q_i, (\pi_a^i(q_i|\theta_i^\pi), \pi_b^i(q_i|\theta_i^\pi)) \Big| \theta_i^Q \Big) \Big] \tag{3.5.20}$$

we apply the policy gradient for actor networks:

$$\nabla_{\theta_i^\pi} \mathcal{L}_i(\theta_i^\pi) = -\mathop{\mathbb{E}}_{q_i} \Big[\nabla_{\delta_a^i} Q_i \Big(q_i, (\pi_a^i(q_i|\theta_i^\pi), \pi_b^i(q_i|\theta_i^\pi)) \Big| \theta_i^Q \Big) \nabla_{\theta_a^i} \pi_a^i(q_i|\theta_i^\pi) + \nabla_{\delta_b^i} Q_i \Big(q_i, (\pi_a^i(q_i|\theta_i^\pi), \delta_b^i(q_i|\theta_i^\pi)) \Big| \theta_i^Q \Big) \nabla_{\theta_b^i} \pi_b^i(q_i|\theta_i^\pi) \Big] \tag{3.5.21}$$

where q_i are sampled from the same mini-batch as in the TD learning phase. The stochastic gradient descent is then carried out in parameter θ_i^π with the learning rate α_δ.

$$\theta_i^\pi \leftarrow \theta_i^\pi + \alpha_\delta \frac{1}{K} \sum_{k=1}^{K} \Bigg(\nabla_{\delta_a^i} Q_i \Big(q_i^{(k)}, (\pi_a^i(q_i^{(k)}|\theta_i^\pi), \pi_b^i(q_i^{(k)}|\theta_i^\pi)) \Big| \theta_i^Q \Big) \nabla_{\theta_a^i} \pi_a^i(q_i^{(k)}|\theta_i^\pi) + \nabla_{\delta_b^i} Q_i \Big(q_i^{(k)}, (\pi_a^i(q_i^{(k)}|\theta_i^\pi), \pi_b^i(q_i^{(k)}|\theta_i^\pi)) \Big| \theta_i^Q \Big) \nabla_{\theta_b^i} \pi_b^i(q_i^{(k)}|\theta_i^\pi) \Bigg) \tag{3.5.22}$$

When designing the learning algorithm, we also take into consideration the trade-off between exploration and exploitation by introducing an exploration probability that decreases exponentially as a function of iteration steps in each episode. Exploration is an essential technique for reinforcement learning to avoid being stuck in the local minima. It allows RL agents to deviate from the action given by the learned policy according to certain exploration rules. In our context, exploration means that the agents switch to a randomly new action

contingently with an exploration probability, instead of picking the action given by actor networks. Exploration allows agents to 'experience' a more diversified situation to avoid being stuck on limited combinations of states and actions. More specifically, we add a random Gaussian noise to actions generated by actor networks when the agents are required to explore. At the beginning of each episode, each agent has a probability of $p_0 = 5\%$ to explore the actions. This exploration probability decreases exponentially $p_t = p_0 \cdot e^{-\eta t}$ where η is a constant, t is the index of the iteration step. In each episode, exploration is more often at the beginning with higher probability and decreases as iteration continues.

Equations (3.5.16)-(3.5.22) define the scheme of our decentralized Multi-agent Deep Deterministic Policy Gradient (Decentralized MADDPG) algorithm. Both critics Q_i and actors (π_a^i, π_b^i) are local functions of market maker i's own inventory q_i. This decentralized feature is an essential difference from the original DDPG algorithm in [Lowe et al. 2017; Foerster et al. 2018]. The interactions between market makers are realized through market shares f_a^i, f_b^i that are implicitly monitored by the market environment. As can be seen from the formulation of the algorithm, training of critics and actors network only requires local information by each market maker. Hence, our algorithm can provide a simulation for a scenario where all market makers apply automated learning algorithms for setting up ask and bid quotes. The simulation results will be useful for regulators in evaluating the effects of automated learning algorithms on market making. We hereby conclude the algorithm in Algorithm 3.

Algorithm 3 Decentralized Multi-agent Deep Deterministic Policy Gradient

Input: E = number of episodes, T = number of iteration steps in each episode, B = size of mini-batch. N = number of market makers, f_a^i, f_b^i: execution probabiltiies, λ^a, λ^b: intensity of ask and bid order flow, Δ: unit order size, ψ_i: running cost for holding inventory to market maker i, and hyperparameters for learning algorithm and optimization algorithm.

Output: The target actor networks $(\tilde{\pi}_a^i, \tilde{\pi}_b^i)$ of each market maker i.

1: Initialization of neural networks:
2: **for** i $\leftarrow$ 1 to N **do**
3: Pre-train critic network Q_i and actor networks π_a^i, π_b^i to value function and quoting strategy of a single monopolistic market maker with execution probability $\Lambda(\delta)$.
4: Let target networks equal to original networks. Namely $\tilde{Q}_i = Q_i$, $\pi_a^i =$

$\tilde{\pi}_a^i$, $\pi_b^i = \tilde{\pi}_b^i$.

end for

for Episode $\leftarrow 1$ to E **do**

  Initialize the inventory states of market makers, denoted by $(q_0^i)_{i\in\{1,\dots,N\}}$.

  for $t \leftarrow 0$ to $T-1$ **do**

    Market environment generates an ask or bid RFQ with probability $\frac{\lambda_a}{\lambda_a+\lambda_b}$ and $\frac{\lambda_b}{\lambda_a+\lambda_b}$.

    Compute action by target actor networks: $\delta_t^{a,i} = \tilde{\pi}_a^i(q_t^i), \delta_t^{b,i} = \tilde{\pi}_b^i(q_t^i)$. Exploration is considered with probability $p_0 \cdot e^{-\eta t}$.

    Market makers compete for the RFQ. I_t^i denote the indicator whether market maker i wins the RFQ.

    Next inventory level q_{t+1}^i is obtained for each market maker i.

    Data $(q_t^i, \varrho_t^i, q_{t+1}^i, I_t^i)$ is stored into market maker i's replay buffer.

    if Replay buffer contains more data points than mini-batch size B **then**

      for i $\leftarrow 1$ to N **do**

        Carry out mini-batch TD learning for critic parameters with (3.5.18)-(3.5.19).

        Carry out mini-batch Stochastic Gradient Descent for actor parameters with (3.5.22).

        Update target network parameters with (3.5.16).

      end for

    else

      Move on to next iteration $t+1$.

    end if

  end for

end for

3.5.3 Numerical Experiments

For the implementation, we consider the same numerical settings as in Section 3.4. More specifically, there are 2 market makers, with execution rates defined in (3.4.6). Other parameters are presented in Table 3.1

RFQ arrival rate $\lambda_a = \lambda_b$	Interest rate r	Order size Δ	Running cost $\psi_1(q) = \psi_2(q)$	Risk limit $Q_1 = Q_2$
2	0.01	1	$\psi_i(q) = -0.005q^2$	5

Table 3.1: RFQ and market makers' parameters for simulation

We use fully connected neural networks for both critic Q_i and actor (δ_a^i, δ_b^i). There are 3 layers in each neural network with 10 hidden units in each layer. We choose the Rectified Linear Unit (ReLU) function as the activation function for the hidden layers. The activation for the output layer of the actor network is a multiplier of tanh function. Specifically, if l is the value before activation in the output layer, then the final output of the actor network is $5 \cdot \tanh(l)$. For critic networks, linear activation is applied. We also introduce a layer normalization for each layer before passing the layer outputs to the activation function. Layer normalization ([Ba, Kiros, and Hinton 2016]) is a technique to obtain a more stationary distribution when data is passed between layers. It normalizes outputs from all units of the same hidden layer to avoid vanishing gradients or gradient explosions due to extreme outputs while maintaining the statistical properties of the values. We find layer normalization important for obtaining stationary quotes values given by actor networks. The detailed actor-critic structure of each market maker is shown in Figure 3.5.

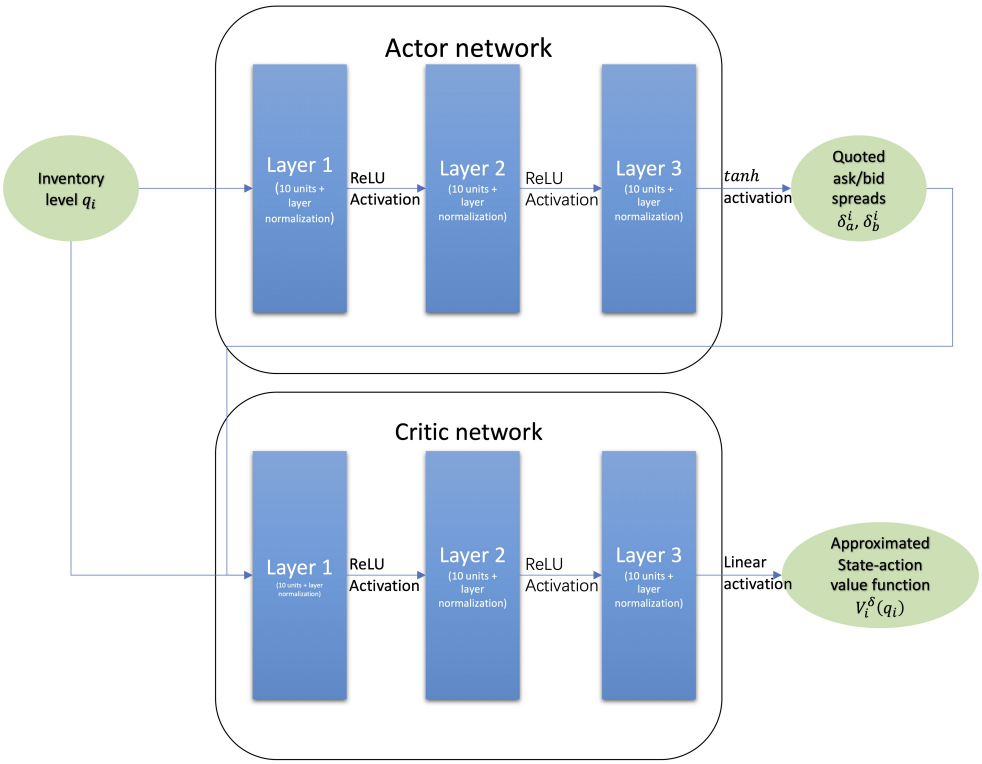


Figure 3.5: Actor-critic networks for market makers. Each hidden layer has 10 neuron units.

We use a standardized ADAM optimizer ([Kingma and Ba 2015]) for both critic and actor networks, with learning rate 0.001, momentum decay rates $(0.9, 0.99)$, and non-zero regularization 10^{-6}. In Reinforcement Learning, an 'episode' refers to a complete play of an agent interacting with the environment. Within an episode, the agent improves her strategy from these interactions through repeated iterations. An 'episode' starts from a randomized initial state and either terminates after a certain number of iterations or when the interactions reach a certain termination condition. In our market making context, an 'episode' refers to one round of repeated pricing game where market makers compete by posting centered quotes at each iteration, starting from a random initial inventory profile. Each iteration step updates the parameters of neural networks via stochastic gradient descent using the data from interactions with the environment. Since in dealer market the agents post centered quotes consecutively, there are no explicit termination conditions to reach for each episode. Hence, we set a fixed number of iterations within one episode.

The training step takes 500 episodes,[6] where in each episode there are 500 iterations. In other words, we set $E = 500, T = 500$ for Algorithm 3. The replay buffers have a fixed length of 10000, and the mini-batch size is 32.

After the training step, we then let the market makers play a repeated pricing game using trained quoting strategies, starting from all possible combinations of initial inventory levels. The market makers quote consecutively using the trained actor networks in the simulated market environment. This process is similar to the training step, however, without the stochastic gradient descent step. The difference between ask/bid quotes given by actor networks and equilibrium ask/bid quotes will be used as a measure for detecting collusion. Excessively higher quotes indicate a higher fee charged to clients.

To summarize, we call one round of simulation which consists of the following 2 steps:

- Training the critic and actor networks using Algorithm 3.
- Applying trained quoting strategies in automated pricing game and compare ask/bid quotes with equilibrium quotes.

We implement the training framework using PyTorch, based on OpenAI Gym ([Towers et al. 2024]), which is widely used for deep reinforcement learning tasks. Since OpenAI Gym only provides APIs for single-agent reinforcement learning, we developed the multi-agent DRL environment for market making from scratch. The training of the agents is conducted on an Intel Core i7 CPU. For the case of 2 market makers, each training step for a single round of 500 episodes takes approximately 40 minutes. To ensure statistical robustness, we repeat the training process for 100 independent rounds of simulations to study the average behavior of reinforcement learning algorithms applied in dealer market making. Consequently, it takes around 70 hours to obtain the complete training results for the scenario of 2 market makers.

Figure 3.6 shows the average cumulative reward per episode with 95% confidence interval taken over 100 independent simulations. The cumulative reward per episode refers to the sum of the rewards from the 500 iterations in each episode.

[6]We have also experimented training with 1000 episodes and find out that average cumulative reward curves exhibit relatively flat trend after first 500 episodes. The value of the learned quotes after the training episodes 1000 is at the same level as after the 500 episodes presented in Figure 3.10, which implies the numerical convergence of the learning algorithm. This comparison therefore validates our choice of using 500 episodes.

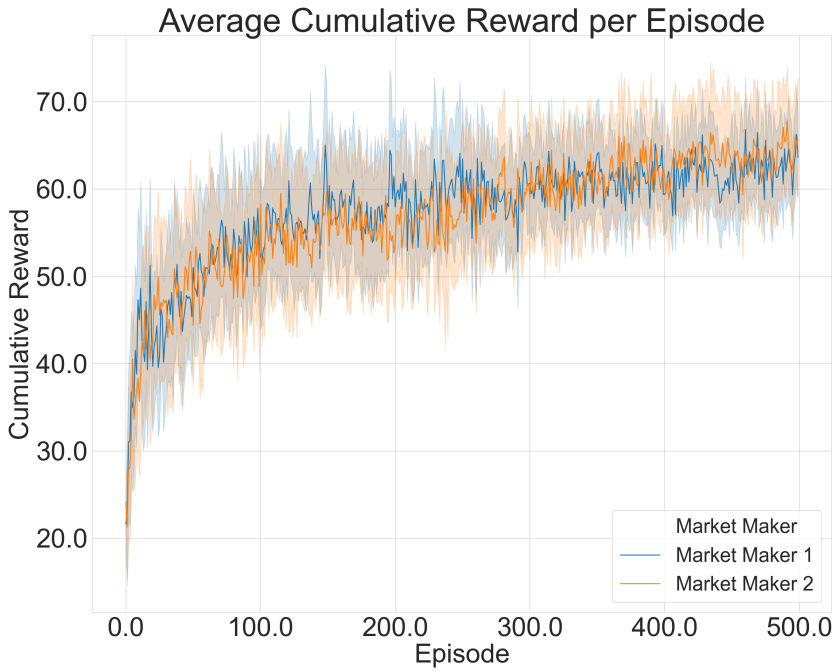


Figure 3.6: Cumulative reward per episode of market makers during training, averaged among 100 simulations with 95% confidence interval.

We can see that the average cumulative reward per episode increases during the training step, with a sharp increase during the first 10 iteration steps. This indicates that both market makers learn to set up more profitable ask/bid quotes as the training continues. There are oscillations in the rewards that suggest some exploration during training. The effectiveness of Decentralized MADDPG is hence demonstrated in that the learning algorithms have indeed learned patterns advantageous to each market maker. We also find that the average reward levels are similar between the 2 competing market makers. This is explained by their homogeneity with the same parameter values in Table 3.1, where neither market maker could gain a more favorable position than her competitor.

Figure 3.7 shows the distribution of the inventory states achieved by the 2 market makers during the training phase. The inventory distribution shows that different inventory values are visited sufficiently during the training step. The fact that inventory levels close to 0 are visited more frequently shows the effect of the running cost function $\psi_i(q_i)$ that penalizes holding nonzero inventory. Hence, Figure 3.7 provides evidence that exploration takes effect in our learning algorithm while the statistical property of inventory distribution

is still maintained.

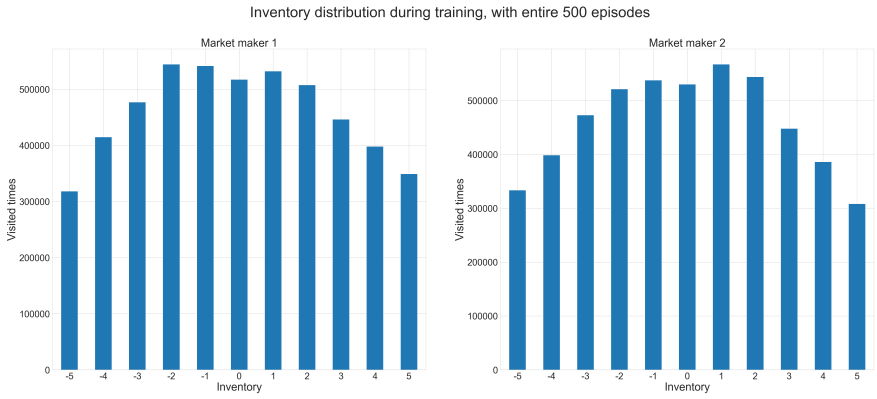


Figure 3.7: Distribution of inventory states for 2 market makers across 500 episodes

To understand how the distribution of inventory levels evolves during training, we separately plot the frequency of inventory aggregated from the first 250 episodes and the last 250 episodes in Figures 3.8 and 3.9. The bar plots show that as learning continues, both market makers have learned to keep their inventories more centered around neutral inventory, demonstrated by a more concentrated area around 0 inventory in the last 250 episodes compared to the first 250 episodes. This indicates that learning agents manage to avoid accumulating portfolios by interacting with the market directly, without knowing the competition mechanism. The learned quoting strategies by decentralized MADDPG are able to take into account by themselves the inventory constraints from running costs.

The concentration around 0 inventory presented by the learning algorithms is in line with the assumption of 0 drift on the dynamics of the asset price (3.1.1). [Vadori et al. 2024] has found that with a positive (negative) drift added in the price dynamics, the market makers' learning algorithm learns the increasing (decreasing) price trend and maintains a positive (negative) inventory to make profit. It is possible to extend our research with nonzero drift, in which we expect to observe a skewed inventory distribution from 0 given by learning algorithms, or more complicated price dynamics. We leave this topic for future research.

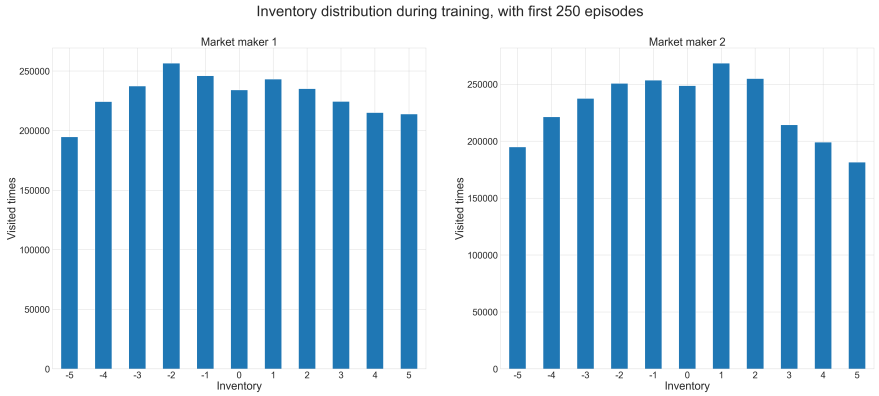


Figure 3.8: Distribution of visited inventory states of 2 market makers with first 250 episodes

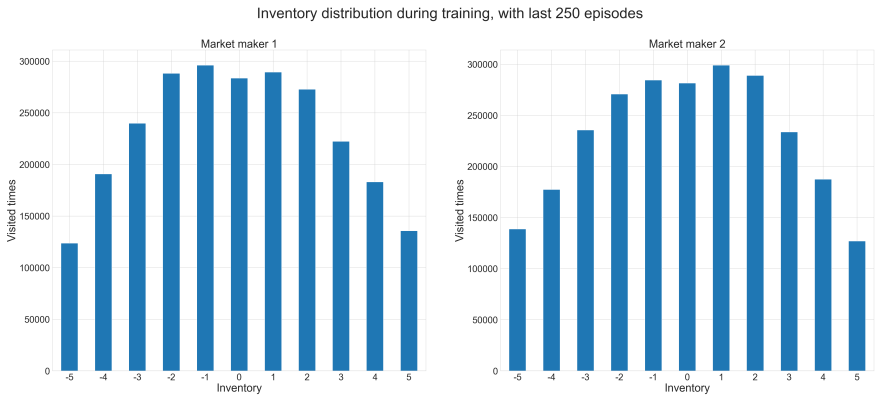


Figure 3.9: Distribution of visited inventory states of 2 market makers with last 250 episodes

We now study the ask and bid quotes by the actor networks after training. We obtain 100 independent quoting strategies from each market maker after 100 simulations. Figure 3.10 shows the average centered ask and bid quotes of the 100 quoting strategies with 95% confidence interval. The quotes are presented by inventory level. The equilibrium and collusive quotes are plotted as a benchmark. Note that the explicit collusion ask and bid quotes are approximated by linear functions of inventory, as described in Section 3.3.

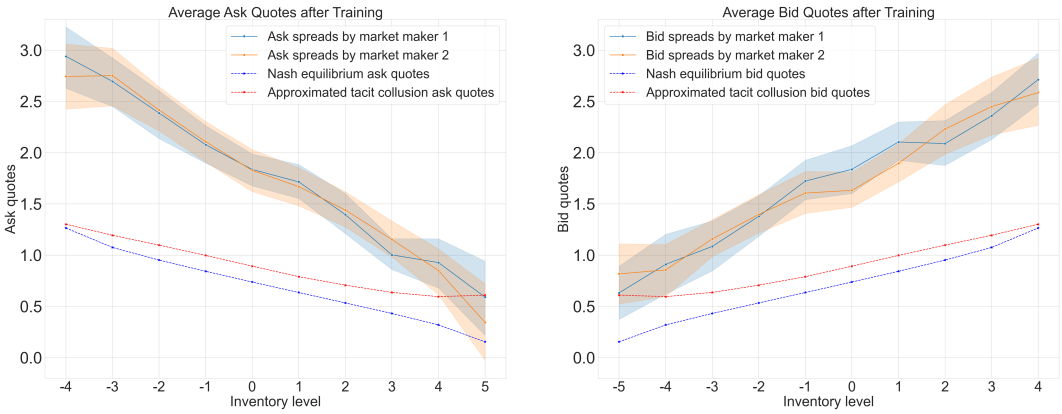


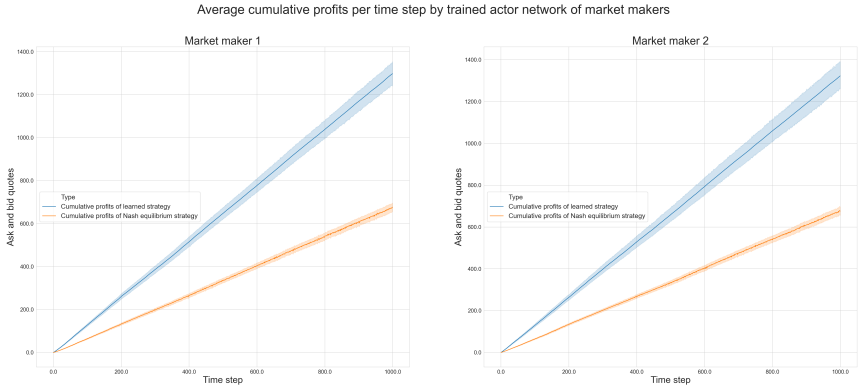
Figure 3.10: Average ask and bid quotes given by trained actor networks from 100 simulations, with 95% confidence interval. Nash equilibrium quotes are plotted in blue dashed lines, while approximated explicit collusion quotes in red dashed lines. The learned quotes are overall higher than competitive quotes in Nash equilibrium, and even higher than the collusion level across most inventory levels.

First of all, it is worth noting that the Decentralized MADDPG algorithm has learned a spread skewing pattern in terms of the inventory levels. The learned ask quote is a decreasing function of inventory, while the learned bid quote is increasing as inventory accumulates. Without knowing the competition mechanism, the learning algorithms are successful in learning the pricing pattern close to theoretical results. From Figure 3.10 we see that the algorithms learn to generate higher quotes than the Nash equilibrium at all inventory levels. Both market makers will quote higher than equilibrium regardless of inventory based on their learned quoting strategies. We have seen that each market maker's learning algorithm only takes information specific to the market maker without access to her competitors' information at all, meaning that the algorithms are not intended to collude by design. However, both algorithms still learn to quote systematically higher than the Nash equilibrium level.

For the next step, we simulate market making with competition using the trained actor networks in the same market environment. This step is also called a 'repeated game' in the game theory context. In this simulation, we let 2 market makers start from the inventory 0 at $t = 0$, after which they quote RFQs sent from the simulated market environment, using their trained actor networks. Since we have 100 independent scenarios, we are able to analyze the

average behavior in tacit collusion while avoiding bias from a certain specific simulation scenario. Figure 3.11 shows the average requests and bids and the cumulative profits of the 100 trained actor networks at each time step during the repeated game, compared to those of Nash equilibrium and collusion strategies. We see that the quotes achieved by Decentralized MADDPG algorithms are systematically wider than the equilibrium quotes, leading to higher cumulative profits earned by the learned strategies. This result is a dynamic version of Figure 3.10. The stationary ask and bid quotes in Figure 3.10 are applied in a repeated game and averaged on 100 possible inventory levels at each time step.

(a) Average ask and bid quotes of trained actor networks in repeated pricing game with 1000 time steps. 0 in vertical axis refers to mid price.



(b) Average cumulative profits by trained actor networks in repeated pricing game with 1000 time steps.

Figure 3.11: Average quotes and cumulative profits of trained strategies.

To further demonstrate the robustness of tacit collusion in terms of initial inventories at $t = 0$, we rerun the above repeated game with all combinations

of initial inventories (q_0^1, q_0^2) and calculate the average basis of 1000 time steps. The basis spreads are calculated by the difference between market makers' ask-bid spread and Nash equilibrium ask-bid spread at every time step. Note that the average is imposed both on 1000 time steps and 100 independent trained actor networks. Figure 3.12 shows a heat map of the average basis spreads of the market makers 2 with respect to the Nash equilibrium spread. In simulation, the risk limit $Q_1 = Q_2 = 5$ with unit order size $\Delta = 1$, therefore, each market maker has 11 possible inventory levels ranging from -5 to 5. With 2 market makers, there are 121 possible combinations of initial inventory levels. We see that the collusive behavior by trained actor networks is significant and robust in terms of initial states.

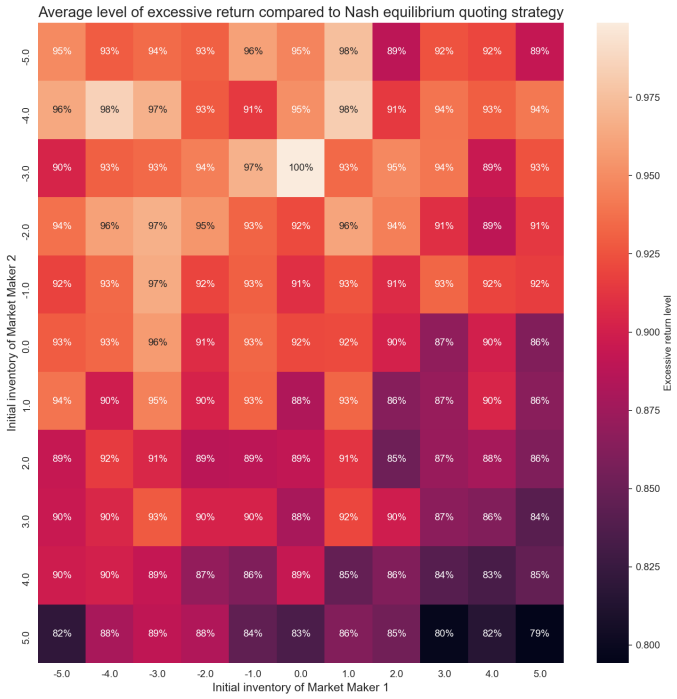
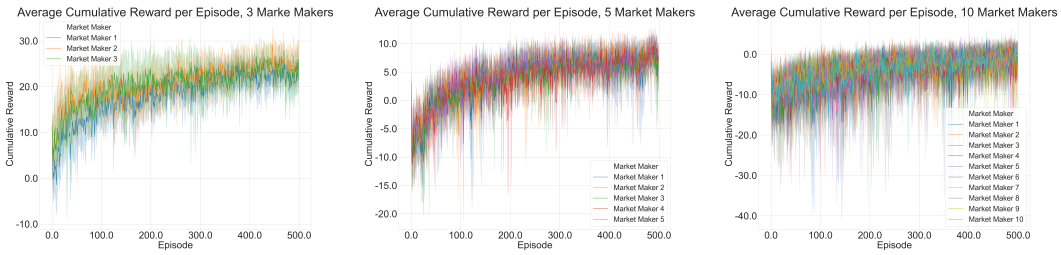


Figure 3.12: Average level of excess return in repeated pricing game with 1000 time steps, across all combinations of initial inventories.

To examine the impact of the number of competing market makers on learning results, we next conduct simulations under the same settings but with different numbers of market makers. The form of execution probabilities depends on (3.1.8) on the number of market makers N. We run 100 independent scenarios, each with 3, 5 and 10 market makers, and summarize the

results in Figures 3.13 and 3.14. Figure 3.13 shows the average cumulative rewards during the training episodes. We observe an increasing trend in the scenarios with the 3, 5 and 10 market makers, suggesting effective learning in all the 3 scenarios. However, with more market makers, the increasing trend in the cumulative rewards slows down, and the reward per market maker is significantly lower with more market makers. With 10 market makers, the cumulative rewards are negative most of the time. This result corresponds to intuition since more market makers induce more intense competition, hence the time for market makers to learn profitable strategies is longer.



(a) 3 Market Makers (b) 5 Market Makers (c) 10 Market Makers

Figure 3.13: Average cumulative reward per episode during training.

Figure 3.14 presents the learned average ask and bid quotes of 2, 3, 5 and 10 competing market markers compared to corresponding Nash equilibrium quotes. We see that compared to learned ask and bid quotes with 2 competing market makers, the learned ask and bid quoted prices with 3, 5 and 10 market makers are globally lower at all inventory levels. This suggests that the seemingly collusive phenomenon with 2 market makers is mitigated to some extent with more market makers. This mitigating trend is not apparent when the number of market makers increases above 3. We estimate that it is due to insufficient training when there are 5 or 10 market makers, because the rewards in Figure 3.13 still oscillate drastically below 0 with 500 training episodes. However, the learned quotes are still above Nash equilibrium levels for most of the inventory levels with 3, 5, and 10 market makers. It is worth noting that the shape of the learned ask and bid quotes resembles those of Nash equilibrium quotes, especially for the case of 10 market makers. This implies that the learning algorithms have replicated the behavior of Nash equilibrium strategy but produce higher quotes leading to 'tacit collusion'. Another important note is that when the number of market makers increases, the learned ask-bid spreads are more skewed to reduce inventory risk. With more market makers,

the learning algorithm tends to be more risk averse in that the ask-bid spreads are more skewed when inventory is none zero. For example, at the inventory level 5, the average ask quotes of 2 market makers is 0.47. This number for 10 competing market makers is -1.35. The negative ask quotes suggest that market makers are more eager to sell when they hold a large long position, even at the cost of losing money. One possible explanation for this skewing behavior is that with more market makers, the probability that the learning algorithms find profitable quotes is lower, hence they are more cautious to avoid holding large non-zero positions and more eager to reduce their exposure to inventory risk at such inventory levels.

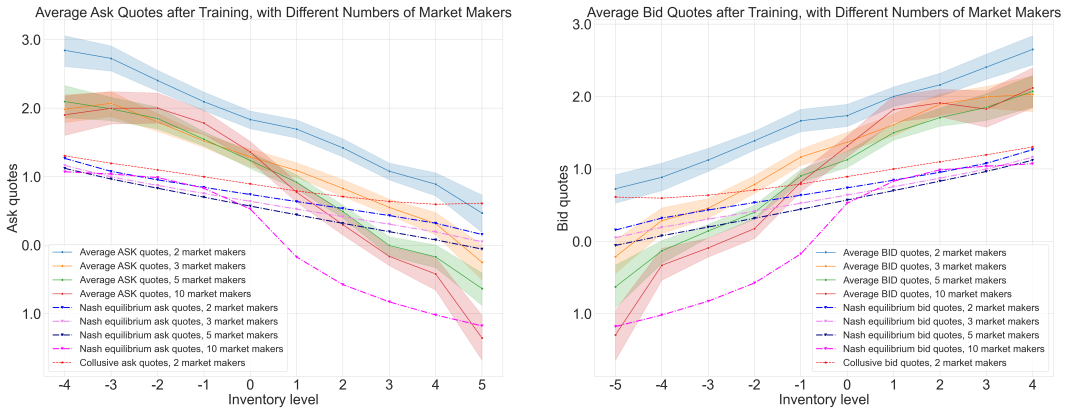


Figure 3.14: Influence of number of market makers on average ask and bid quotes.

In summary, this simulation using the decentralized MADDPG algorithm has seen a behavior similar to collusion between 2 competing market makers. The learning algorithm is effective in producing stylized features of ask and bid quotes, but with an overall higher level than equilibrium quotes. This tacit collusion is robust with different initial inventory levels. With more market makers, the learned algorithm tends to produce lower quotes than 2 market makers, and becomes more risk averse in that the ask-bid spreads are more skewed when market makers' inventory deviates away from 0.

3.6 A Review on Learning Dynamics during Multi-agent Training

Further investigation of the dynamics of inventory states during training steps reveals more information on the phenomenon of tacit collusion. Figure 3.15-3.18 present snapshots of training instances with different numbers of market makers. As indicated, each snapshot shows the dynamics of the inventory states during either the first or the last episodes of training of one specific experiment. It can be seen that the difference in the correlation pattern across the very first and very last episodes is phenomenal, regardless of the numbers of market makers. For example, in Figure 3.15, during episode 1 the inventory levels of the market maker 1 and the market maker 2 are coarsely uncorrelated, since the market maker 2 wins most of the order flow, resulting in the market maker 1' inventory being kept flat. However, after 500 episodes of training, inventories of the 2 market makers are closely correlated. Similar patterns are observed with 3, 5, 10 market makers, where inventories during the first few episodes are uncorrelated while during the final episodes they present a correlated pattern.

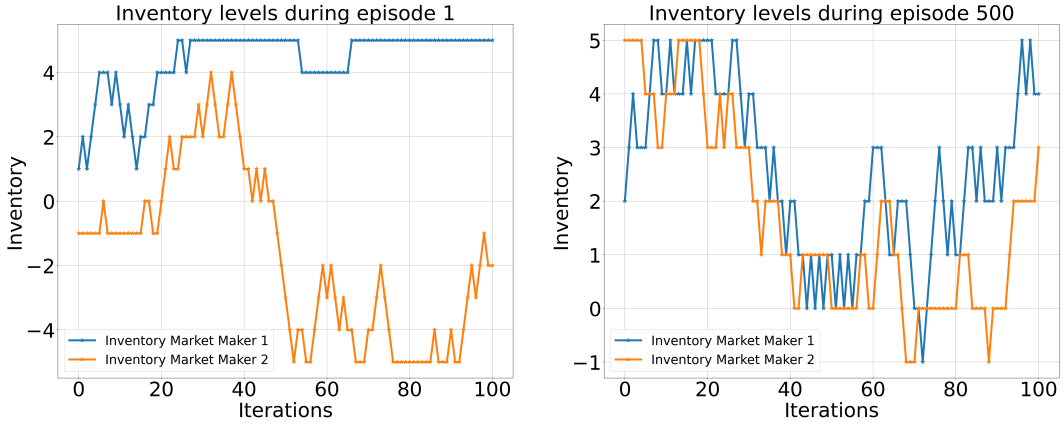


Figure 3.15: Inventory comparison between 2 learning market makers in first and last episodes.

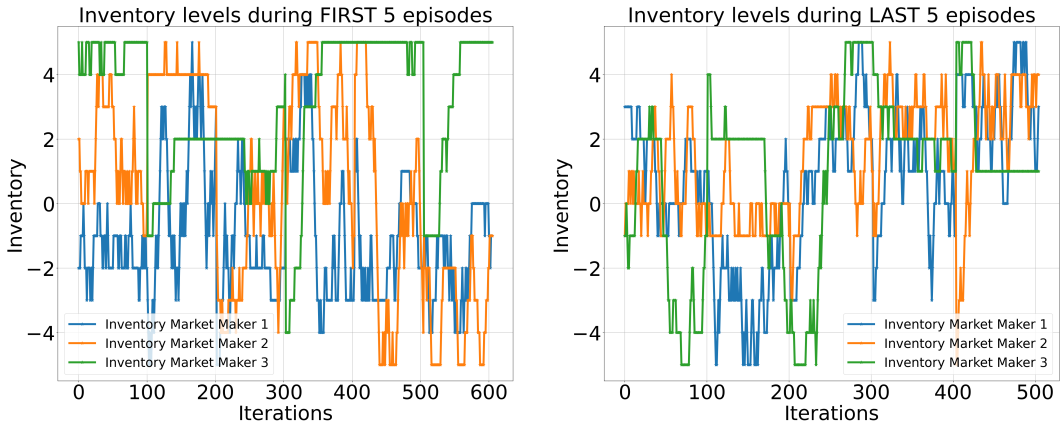


Figure 3.16: Inventory comparison between 3 learning market makers in first and last episodes.

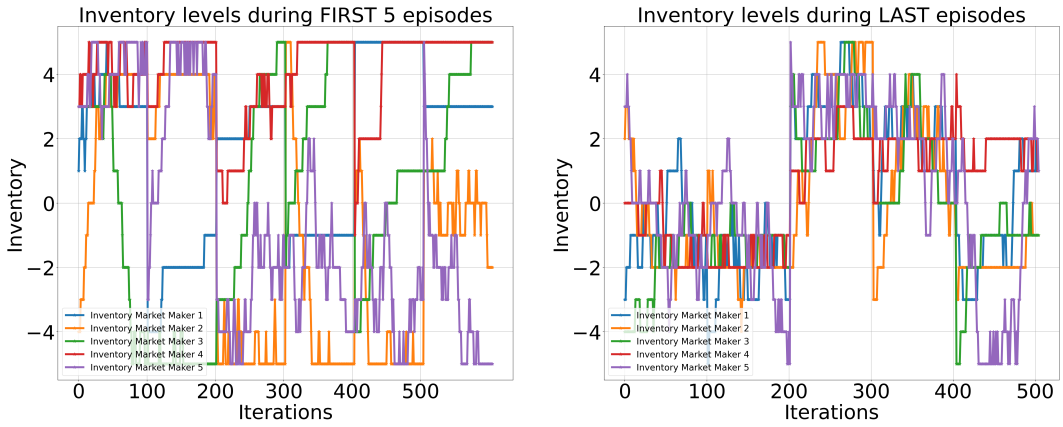


Figure 3.17: Inventory comparison between 5 learning market makers in first and last episodes.

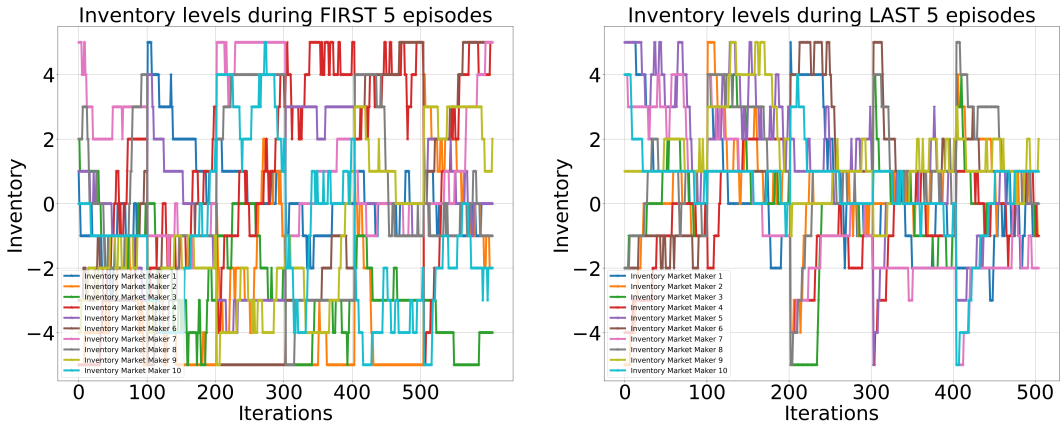


Figure 3.18: Inventory comparison between 10 learning market makers in first and last episodes.

Visual inspection of snapshots of training instances indeed shows the emergence of a positive correlation structure in market makers' inventories after the learning process. We further quantify this positive correlation pattern: When the number of market makers is N ($N \in \{2, 3, 5, 10\}$), within each episode, we calculate the correlation of inventory levels between all distinct combinations of market maker pairs after concatenating the inventories from the 100 independent training experiments. Hence, for each episode, there are $\binom{N}{2}$ correlations calculated for distinct market maker pairs. Then the average of these correlations across all market maker pairs is computed and used as a metric for quantification of the correlation pattern of each training episode. By doing this, we have incorporated several pairwise correlation coefficients into one single metric per episode, to represent the overall correlation when there are more than 2 market makers.

The result is summarized in Figure 3.19, which includes a more comprehensive overview of the correlation structure than the snapshots. A significantly increasing pattern in episode-wise correlation between market makers' inventories is shown across different numbers of market makers. The correlation oscillates around 0% in the very first episodes during training. After more episodes are trained, the average correlations increase to significantly positive levels. This trend is universal in 4 settings with different numbers of market makers. With the numbers of market makers increasing, the positive correlation reached in the final episodes tends to weaken but is still at a significantly positive level.

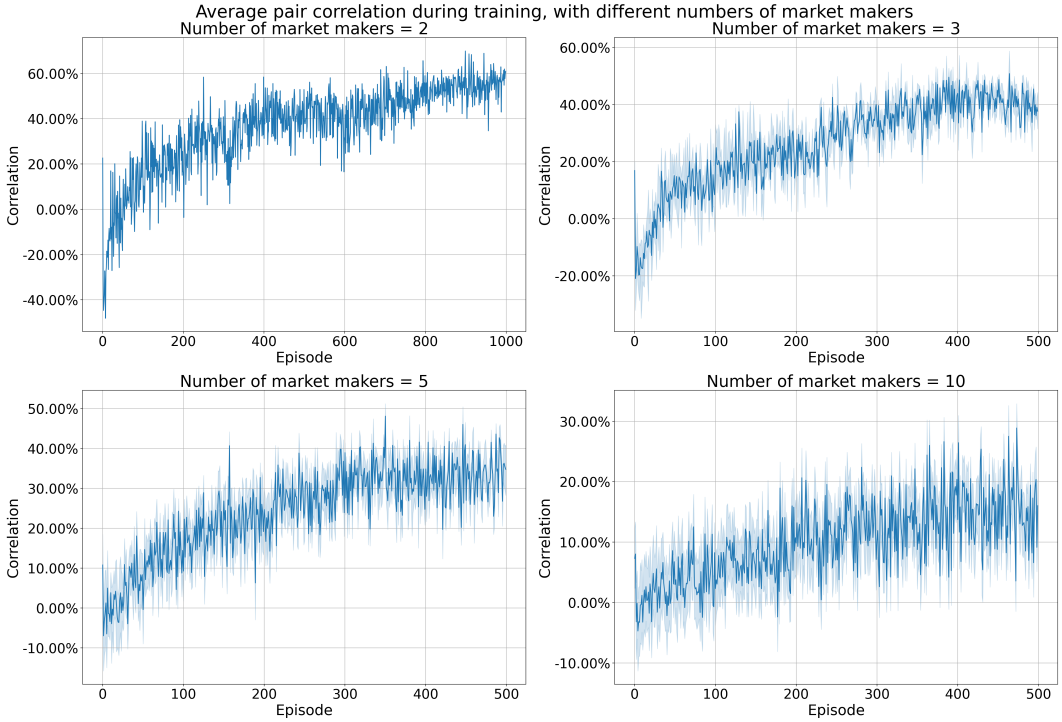


Figure 3.19: Average correlation of market maker pairs as function of training episodes with 95% confidence interval, for 2, 3, 5 and 10 market makers.

3.6.1 Granger Causality Test on the Inventory Processes

To further investigate whether the correlation observed between the market maker's inventories is due to a causal relationship, we perform a statistical analysis using Granger causality test ([Granger 1969]). Granger causality test is a statistical hypothesis test that aims to determine whether one time series can provide significant information to forecast another time series. The test includes fitting predictive models incorporating lagged values of both time series and assessing whether the inclusion of past values of one time series can improve the prediction accuracy for the other. If the improvement is statistically significant, it suggests the presence of a causal relationship in Granger causality sense.

Our analysis is carried out in the case of 2 market makers. Specifically, we test two null hypotheses:

(1) There is no Granger causality for the inventory process of Market Maker 1 on that of Market Maker 2.

(2) There is no Granger causality for the inventory process of Market Maker 2 on that of Market Maker 1.

For each hypothesis, we perform the tests across 500 episodes for each of 100 independent trials of experiments. The $p-$values obtained from these tests are averaged over the trials to provide a measure of statistical significance across different training episodes. If the $p-$value from the Granger causality test is less than 5%, we reject the null hypothesis of no causality. Otherwise, the null hypothesis is reserved. Our results show that the average $p-$values are well above the significance level of 5%, indicating that the null hypothesis cannot be rejected. This implies that there is no significant evidence of Granger causality between the inventory processes of the 2 market makers. We present the detailed results of the average $p-$values from Granger causality tests in Figure 3.20, which summarize the average $p-$values obtained for each null hypothesis across different episodes.

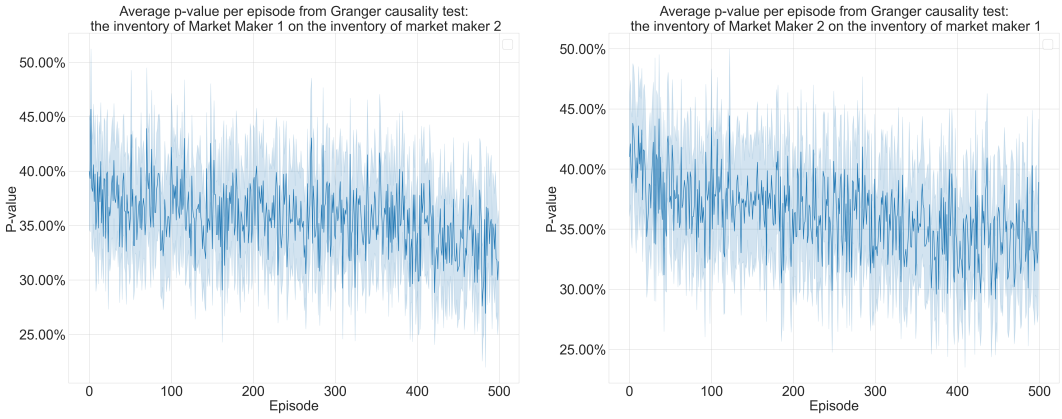


Figure 3.20: Average $p-$value per episode from Granger causality test on the inventory processes of 2 market maker.

The absence of Granger causality between the inventory processes suggests that the observed positive correlation between the inventories of the 2 market makers is not the result of one leading market maker that influences the inventory changes of the competitor.

3.6.2 Quote Reaction Analysis on Competitor's Quoted Values

We further conduct an analysis of how competitors' quotes evolve following an RFQ transaction by one of the market makers, in order to look into whether

the observed correlation between the market makers' inventories is due to an 'oscillating lead-lag' mechanism in the quotes of the competing market makers. This mechanism refers to a sequential pattern where one market maker's execution of an RFQ, which changes her inventory and thus skews her quotes, leads to her competitor more likely to win the subsequent RFQ in the same direction. If this mechanism exists, it can provide an explanation for the observed correlation in the inventory processes.

Our analysis is designed in 2 parts, focusing on the case of 2 market makers. First, we examine the quote changes of the competitor in the same direction (ask or bid) immediately following the execution of an RFQ by one of the market makers. The quote changes are measured for 1 to 4 iteration steps after an RFQ execution in the same episode. We perform the calculation across 500 episodes in all 100 independent experiment trials. The results are presented in the left panel of Figure 3.21, which shows that on average, there is a consistent reduction in the competitor's quotes after a same-side RFQ is executed by a market maker. This suggests that, during the training process, the learning algorithms tend to adjust the quotes more aggressively over the next few steps to increase the probability of winning the next RFQ if they do not win the RFQ at the current iteration step.

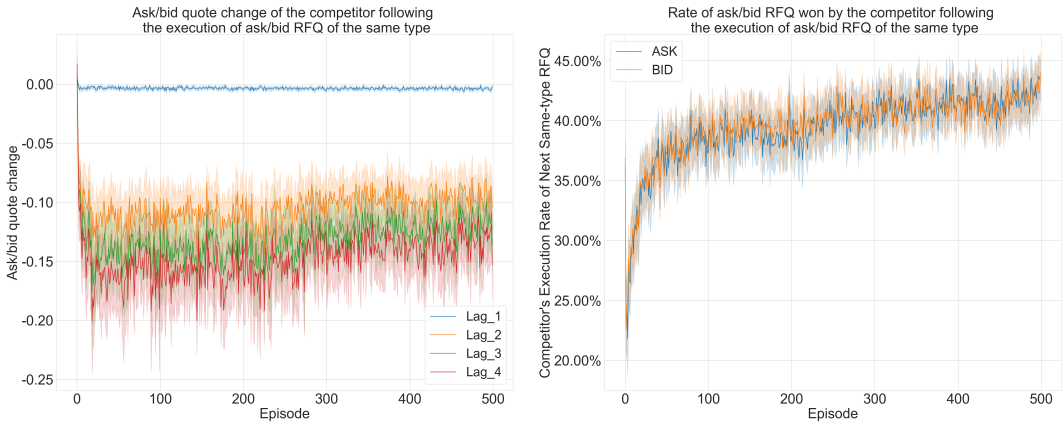


Figure 3.21: Average same-type quote changes and subsequent same-type quote execution probability of the competitor following an RFQ execution.

The second part of the analysis focuses on whether this quote adjustment pattern increases the probability of the competitor winning the next same-type RFQ. Within each episode, we calculate the empirical likelihood of the competitor winning the subsequent RFQ of the same type after an RFQ ex-

ecution by one of the market makers. The right panel of Figure 3.21 shows that, while there is an increasing trend during training in the probability that the competitor wins the next RFQ of the same type, this probability remains below 50%, suggesting that a subsequent same-type RFQ is not necessarily more likely to be executed by the competitor following an RFQ execution by one of the market makers. Therefore, there is no significant preference for the competitor to win the next same-type RFQ, and an 'oscillating lead-lag' mechanism is probably not statistically significant.

We present in Figure 3.22 a snapshot of correlated inventory processes from the episode 497 of one experiment trial, together with the associated ask or bid quotes of the market makers. The winner of the RFQs at each iteration is annotated. Visual inspection of this episode reveals that RFQ executions do not consistently lead to the competitor's execution of subsequent same-type RFQs.

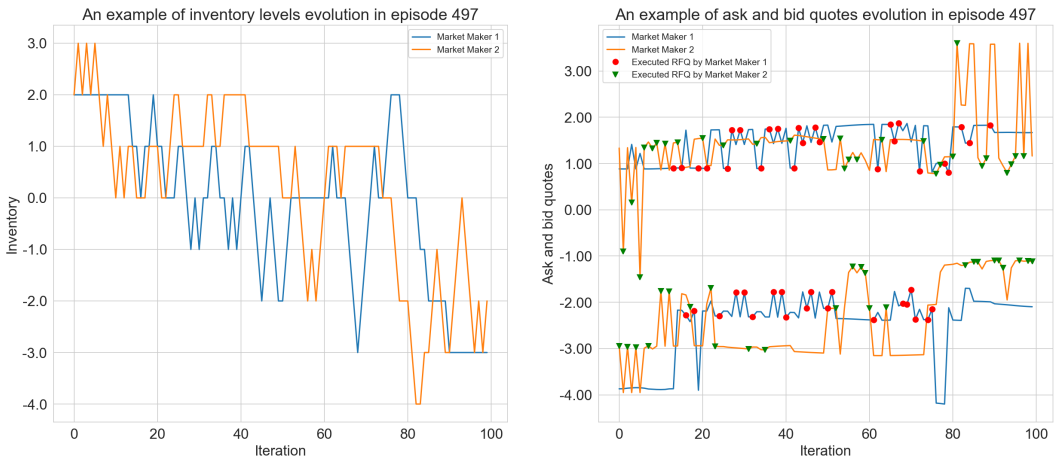


Figure 3.22: A snapshot of the evolution of the inventory and the ask/bid quotes during training, in episode 497 in an experiment trial. In the right panel, the curves with positive values are the ask quotes, and the curves with negative values are the bid quotes.

To conclude, the positive correlation pattern is nontrivial as the learning algorithm of each market maker does not have access to the inventory levels of her competitors. Note that since learning is decentralized and the market is assumed to be sufficiently liquid, we expect to observe uncorrelated inventory states, or, in other words, a distribution of inventory levels when the numbers of market makers are large, instead of correlated inventories. Nevertheless, these

decentralized learning algorithms are able to have learned quoting strategies that keep market makers' inventories correlated with each other. This suggests each market maker's algorithm implicitly learns about her competitors' inventory dynamics through interacting with the market environment, and learns to adjust their quotes to closely keep track of competitors' inventories, which could potentially contribute to market makers jointly quoting higher ask and bid prices than equilibrium. Although we leave the investigation of the exact mechanism driving the positively correlated inventory processes for future research, we believe that the analysis in this section sheds light on the impact of homogeneity on agents' learned quoting strategies. In our simulation, every market maker is applying the same category of learning algorithm, with a similar network structure and the same risk aversion parameters. The simulation is therefore based on a homogeneous setting. In Chapter 4, we shall look into one specific case of heterogeneity in a mean field modelling for a large population of market makers and show its impact on algorithm-learned strategy compared to homogeneous learning.

Chapter 4

Competition in Dealer Markets: Strategic Interactions, Learning and Heterogeneity

This chapter is based on [Assayag et al. 2024]. In this chapter, we study the interactions of market makers via mean field modelling. We consider the market makers to interact through their average bid and ask quotes in the intensity function and study the dynamics of mean field learning algorithms.

Mean Field Games (MFG) [Lasry and Lions 2007; Huang, Caines, and Malhamé 2007] provide a tractable approach to model a large population of market makers in an OTC market, and to study the resulting dynamics of deep reinforcement learning algorithms applied to make the market in such an MFG framework. This approach allows for a tractable analysis of the complex interactions and strategic decision making of a large number of market makers, which traditional N-player game-theoretic models struggle to handle efficiently.

The MFG framework, primarily characterized by a continuum of agents whose collective behavior impacts each individual, offers a novel perspective to understand the dynamics of decentralized decision making in financial markets. It enables us to model the average effect of all other market makers on a representative market maker, thus simplifying the analysis while capturing the essence of competitive and adaptive behaviors. This is particularly pertinent in the context of electronic OTC markets, where the number of participants can be large, and their interactions and quoting strategies are continually evolving.

Furthermore, the mean field deep reinforcement learning framework extends the scope studied in Chapter 3, where a decentralized multi-agent deep reinforcement learning algorithm is proposed for N-player stochastic differential game model of market makers. This learning process features the scenario

where a representative market maker applies a learning algorithm to quote ask and bid prices through interacting with the market environment and the mean field generated by collective market behavior, giving rise to the market dynamics that we aim to explore. These include equilibrium states, the potential for tacit collusion, and the impact of individual learning on overall market stability and efficiency.

By integrating the MFG framework with deep reinforcement learning algorithms, we undertake a comprehensive study to uncover the underlying mechanics of how market makers interact and learn in a decentralized manner. This thesis aims not only to contribute to the theoretical understanding of these complex systems but also to offer insights that could inform practical approaches in market regulation and the development of next-generation algorithmic trading strategies.

We envision the trend towards the popularized adoption of AI-driven algorithms by an increasing number of market participants, which can potentially introduce different market dynamics compared to rule-based algorithms. A recent report by [IMF 2024] has highlighted the potential changes and risks with the further adoption of AI in financial markets. Specifically, it has highlighted potential concerns such as algorithmic collusion and market manipulation risks, similarly raised in other literature cited in this thesis. Since RL-based algorithms are one of the most important AI techniques being applied in finance, and their learning through interactions with the environment can be naturally adapted to the application of market making, we believe that these RL-based algorithms are the most likely candidates for adoption by dealers to embrace AI in market making. Therefore, we believe that now is the time to proactively study the implications of the scenario in which a large number of market makers adopt RL-based quoting strategies. This includes extending the analysis of the models developed in Chapters 2 and 3 to a mean field game setting.

Outline Section 4.1 describes our model setting for a continuous-time dealer market with competition among market makers, formulated as a mean field game with control being the representative market maker's quoted ask and bid prices. We prove the existence of mean field Nash equilibrium and provide a condition for uniqueness. In Section 4.2 we design a mean field deep reinforcement learning algorithm applied by a representative market maker in a simulated market environment. In Section 4.3 we present simulation evidence that homogeneous mean field learning does not necessarily converge to mean

field Nash equilibrium, giving rise to 'tacit collusion'. We show that supra-competitive learning behavior is mitigated when heterogeneity is introduced where a learning agent interacts with a mean field equilibrium system.

4.1 Competition in Dealer Markets: a Mean Field Game Approach

The MFG framework relies on several key assumptions. One fundamental assumption is the existence of a large number of agents, where the individual agents have negligible impact on the overall system, so that the aggregate effect formed by the agents can be modelled by a mean field approximation. Another critical assumption is that the agents have rational expectations, meaning they respond optimally to the mean field, and have full knowledge on the evolution of the mean field based on other agents' strategies. The agents are typically assumed to be homogeneous and can be modelled by a representative agent to analyze equilibrium strategies. In terms of observability, mean field games assume that agents can observe the mean field, which reflects the population distribution, but do not have information about the states or strategies of individual agents. This aligns with our market making setting, where the market makers do not observe competitors' quoting strategies, but are influenced by the joint actions of competitors' quotes through the intensity function. In our model, this partial observability is represented by the dependence of the intensity functions on the average ask and bid quotes of the entire population, as discussed in this section.

We consider a single-asset market where the dynamics of the reference asset price follow Bachelier's model:

$$S_t = S_0 + \sigma W_t \tag{4.1.1}$$

We assume that there are infinitely many market makers competing in a dealer market, and the interactions occur through the mean bid and ask quotes in the intensity function. We hence study the behavior of a representative market maker instead of considering the complicated interactions in N-player games.

The bid and ask prices quoted by a representative market maker are indicated by S_t^a and S_t^b, in which the representative market maker quotes centered

spreads δ_t^a and δ_b on top of the reference price S_t.

$$\begin{aligned} S_t^b &= S_t - \delta_t^b \\ S_t^a &= S_t + \delta_t^a \end{aligned} \tag{4.1.2}$$

As discussed in Section 3.1, the ask price S_t^a is always above the bid price S_t^b. The centered quotes δ_t^a, δ_t^b are generally positive. However, one of the centered quotes δ_t^a or δ_t^b can take negative values, due to the aggressive quoting behavior of the market maker to attract order flow in the corresponding direction and to reduce her inventory risk. A lower bound $-\delta_\infty$ with $\delta_\infty > 0$ is imposed on the value of centered quotes δ_t^a, δ_t^b to limit negative quotes, namely $\delta_t^a \geq -\delta_\infty, \delta_t^b \geq -\delta_\infty$.

We let I_δ denote the interval $[-\delta_\infty, \infty)$, then introduce the admissible strategy space of the representative market maker starting from time t.

$$\begin{aligned} \mathcal{A}_t^T = \Big\{ \boldsymbol{\delta} = (\delta^a, \delta^b) \Big| &\text{for } k \in \{a, b\}, \delta^k \in C\big([t, T], (I_\delta)^{2H+1}\big); \\ &\forall s \in [t, T], \delta^k(s) = \big(\delta^k(s, q)\big)_{q \in \mathcal{Q}} \in (I_\delta)^{2H+1} \Big\} \end{aligned} \tag{4.1.3}$$

Particularly $\mathcal{A}_0^T$ denotes the space of admissible strategies that start at time 0.

We first introduce a few function spaces. We denote by $\mathcal{P}(\mathcal{Q})$ the space of probability distributions in $\mathcal{Q}$. Since $\mathcal{Q}$ is finite, we identify $\mathcal{P}(\mathcal{Q})$ with the $(2H+1)$-dimensional simplex $\mathcal{S}^{\mathcal{Q}}$:

$$\mathcal{S}^{\mathcal{Q}} = \{(p_q)_{q \in \mathcal{Q}} | p_q \geq 0, \sum_{q \in \mathcal{Q}} p_q = 1\}$$

$\mathcal{S}^{\mathcal{Q}}$ is a closed subset in $\mathcal{R}^{2H+1}$, equipped with the Euclidean norm $||.||_2$.

The market maker is characterized by their inventory level, representing their *state*, and their ask/bid quotes for the asset, representing their *action*. [Guo, Hu, et al. 2019] study generalized mean field games (GMFG) where the joint distribution of states and actions formulates the mean field. In this thesis, we are studying deterministic quoting strategies instead of randomized actions by market makers, and we shall see that the population's quoting strategies coincide with equilibrium quoting strategy when the system reaches mean field equilibrium. Therefore, we consider state distribution as the mean field formed by all market makers. More specifically, the mean field is characterized by a probability distribution flow of inventories. We describe the flow of the inventory distribution of the population using a (differentiable) map $\boldsymbol{M} : [0, T] \to (\mathcal{S}^{\mathcal{Q}}, ||\cdot||_\infty)$. We denote $\boldsymbol{M}(t) = m(t, \cdot)$ where

$m(t,\cdot) = (m(t,q))_{q\in\mathcal{Q}} \in \mathcal{S}^{\mathcal{Q}}$. Clearly $m(t,\cdot)$ defines a probability density function on $\mathcal{Q}$. We also define $\mathbb{B}([0,T]\times\mathcal{Q})$ as the space of bounded functions defined in domain $[0,T]\times\mathcal{Q}$, equipped with the norm $||\cdot||_\infty$.

We assume that the market maker population applies a common quoting strategy $\bar{\boldsymbol{\delta}} \in \mathcal{A}_0^T$ with $\bar{\boldsymbol{\delta}}_t = (\bar{\delta}^a(t,q), \bar{\delta}^b(t,q))$. The interactions between the representative market maker and the population are realized through the average of ask and bid quotes by the population.

$$\bar{\mu}_t^a = \int_Q \bar{\delta}^a(t,q)m(t,q)dq \qquad \bar{\mu}_t^b = \int_Q \bar{\delta}^b(t,q)m(t,q)dq \tag{4.1.4}$$

The inventory of the representative market maker is modelled by the point processes N_t^b and N_t^a representing the order flow from the bid and ask sides whose unit size is 1:

$$dq_t = dN_t^b - dN_t^a \tag{4.1.5}$$

The intensities ν_b, ν_a of point processes N_t^b, N_t^a satisfy

$$\nu_t^a = \lambda^a f_a(\delta_t^a, \bar{\mu}_t^a)\mathbb{I}(q_t > -Z) \quad \nu_t^b = \lambda^b f_b(\delta_t^b, \bar{\mu}_t^b)\mathbb{I}(q_t < Z) \tag{4.1.6}$$

where Z is the inventory limit imposed on the representative market maker, $\bar{\mu}^a, \bar{\mu}^b$ is the average ask and bid quotes from quoting strategies of the entire population. We assume that at each time t the population inventory distribution $m_t(dq)$ yields a probability density function denoted by $m(t,q)$. The inventory limit Z imposed results in the representative market maker's inventory taking values from the finite discrete set $\mathcal{Q} = \{-Z, -Z+1, \ldots, Z\}$ with $2Z+1$ elements.

The cash process X_t of the representative market maker consists of revenue from market making on both ask and bid sides. It equals the difference between the income from selling the asset and the payment for buying the asset.

$$dX_t = S_t^a dN_t^a - S_t^b dN_t^b = \delta_t^b dN_t^b + \delta_t^a dN_t^a - S_t dq_t \tag{4.1.7}$$

We consider Markovian controls where the quotes are functions of the market maker's inventory:

$$\delta_t^b = \delta^b(t, q_t), \delta_t^a = \delta^a(t, q_t) \tag{4.1.8}$$

The representative market maker faces an optimization problem with respect to her quoting strategy. For the objective function, we consider the profit and loss between a given time interval $[0,T]$ exponentially discounted

by the interest rate r, regularized by the running cost of holding nonzero inventory and terminal inventory. Namely, given a flow of population distribution $M : [0,T] \to \mathcal{S}^{\mathcal{Q}}$ and given a population quoting strategy $\bar{\boldsymbol{\delta}} \in \mathcal{A}_0^T$, at time 0 the representative market maker aims at maximizing:

$$J_0(\boldsymbol{\delta}; \bar{\boldsymbol{\delta}}, \boldsymbol{M}) = \mathbb{E}^{\boldsymbol{\delta}, \bar{\boldsymbol{\delta}}, \boldsymbol{M}} \Big[\int_0^T e^{-rt} d(X_t + q_t S_t) - e^{-rT}\phi(q_T) - \int_0^T e^{-rt}\psi(q_t)dt \Big] \tag{4.1.9}$$

where $\psi : \mathbb{R} \to \mathbb{R}_+$ is the running cost for holding inventory and $\phi : \mathbb{R} \to \mathbb{R}_+$ is the representative market maker's cost for liquidating non-zero inventory at terminal time T. The expectation $\mathbb{E}^{\boldsymbol{\delta}, \bar{\boldsymbol{\delta}}, \boldsymbol{M}}$ is taken for the inventory process q_t under the control of the representative market maker's quoting strategy $\boldsymbol{\delta}$, the population's quoting strategy $\bar{\boldsymbol{\delta}}$, and the population distribution flow M.

The value function at time t is defined as

$$\begin{aligned} V_t(\bar{\boldsymbol{\delta}}, \boldsymbol{M}) &= \sup_{\boldsymbol{\delta} \in \mathcal{A}_t^T} J_t(\boldsymbol{\delta}; \bar{\boldsymbol{\delta}}, \boldsymbol{M}) \\ &= \sup_{\boldsymbol{\delta} \in \mathcal{A}_t^T} \mathbb{E}^{\boldsymbol{\delta}, \bar{\boldsymbol{\delta}}, \boldsymbol{M}} \Big[\int_t^T e^{-r(u-t)} d(X_u + q_u S_u) - e^{-r(T-t)}\phi(q_T) \\ &\qquad - \int_t^T e^{-r(u-t)}\psi(q_u)du \Big| \mathcal{F}_t \Big] \end{aligned} \tag{4.1.10}$$

we denote the value function conditioned on the event $\{q_t = q\}$ by $V^{\bar{\boldsymbol{\delta}}, \boldsymbol{M}}(t,q)$, that is,

$$\begin{aligned} V^{\bar{\boldsymbol{\delta}}, \boldsymbol{M}}(t,q) = \sup_{\boldsymbol{\delta} \in \mathcal{A}_t^T} \mathbb{E}^{\boldsymbol{\delta}, \bar{\boldsymbol{\delta}}, \boldsymbol{M}} \Big[\int_t^T e^{-r(u-t)} d(X_u + q_u S_u) - e^{-r(T-t)}\phi(q_T) \\ - \int_t^T e^{-r(u-t)}\psi(q_u)du \Big| q_t = q \Big] \end{aligned} \tag{4.1.11}$$

Remark 4.1.1. N-player market making game can be formulated in an analogous format. In an N-player system, the inventory process q_t^i of the market maker indexed by i ($i = 1, \ldots, N$) is expressed by point processes $N_t^{a,i}$ and $N_t^{b,i}$: $dq_t^i = dN_t^{b,i} - dN_t^{a,i}$, where intensities of $N_t^{a,i}$ and $N_t^{b,i}$ are jointly affected by the market maker i and the average quotes of market makers:

$$\nu_t^{a,i} = \lambda^a f_a(\delta_t^{a,i}, \frac{1}{N}\sum_{j=1}^N \delta_t^{a,j})\mathbb{I}(q_t^i > -Z_i), \nu_t^{b,i} = \lambda^b f_b(\delta_t^{b,i}, \frac{1}{N}\sum_{j=1}^N \delta_t^{b,j})\mathbb{I}(q_t^i < Z_i)$$

Market maker i hence aims to maximize the objective functional:

$$J_0^i(\boldsymbol{\delta}^i; \boldsymbol{\delta}^{-i}) = \mathbb{E}^{\boldsymbol{\delta}^i, \boldsymbol{\delta}^{-i}} \Big[\int_0^T e^{-rt} d(X_t^i + q_t^i S_t) - e^{-rT}\phi(q_T^i) - \int_0^T e^{-rt}\psi(q_t^i)dt \Big]$$

The notations are analogous to the mean field game model with superscript i added. $\boldsymbol{\delta}^{-i}$ is the collection of all the competitors' quoting strategies. Propagation of chaos allows us to expect that when $N \to \infty$ the N-player Nash equilibrium converges to the mean field Nash equilibrium.

Some assumptions are made about the execution rate functions f_a and f_b:

Assumption 4.1.2. $f \in \{f_a, f_b\}$ are C^2 functions on $\mathbb{R}^2$, satisfying:

1. $\partial_\delta f(\delta, \mu) \leq 0, \quad \partial_\mu f(\delta, \mu) \geq 0$
2. $\forall \mu \in \mathbb{R}, \lim_{\delta \to \infty} f(\delta, \mu) = 0$; $\forall \delta \in \mathbb{R}, \lim_{\mu \to \infty} f(\delta, \mu) \leq 1$
3. There exists a function $\Lambda(\delta) \in C^2(\mathbb{R})$ such that $0 < f(\delta, \mu) \leq \Lambda(\delta)$, and
$$\lim_{\delta \to \infty} \Lambda(\delta)\delta = 0, \Lambda'(\delta) < 0, \Lambda(\delta)\Lambda''(\delta) \leq 2(\Lambda'(\delta))^2$$
4. There exists a constant $C > 0$ such that
$$2 - \frac{\partial^2_{\delta\delta} f(\delta, \mu) f(\delta, \mu)}{(\partial_\delta f(\delta, \mu))^2} \geq C \tag{4.1.12}$$
5. There exists a constant $K \geq 0$ such that
$$\left| \frac{\partial_\mu f(\delta, \mu)}{\partial_\delta f(\delta, \mu)} \right| \leq K \tag{4.1.13}$$
6. There exists a constant $c < 1$, such that $\forall \delta, \mu \in \mathbb{R}$,
$$\frac{|\partial_\delta f(\delta, \mu)\partial_\mu f(\delta, \mu) - f(\delta, \mu)\partial_\delta \partial_\mu f(\delta, \mu)|}{2(\partial_\delta f(\delta, \mu))^2 - \partial^2_{\delta\delta} f(\delta, \mu) f(\delta, \mu)} \leq c \tag{4.1.14}$$

Remark 4.1.3. Assumption 4.1.2 is made on the regularity of the execution rate functions f_a and f_b. Assumption 1 is intuitive that the execution rate decreases when the representative market maker increases her quotes and increases when the average population quotes increase. Assumption 2 is made on the asymptotic limit of the execution rates. Assumption 3 corresponds to the intuition that the execution probability in the mean field game model is dominated by $\Lambda(\delta)$ which can be understood as that of a single market maker case. This assumption is also made in Chapter 3 where a multi-agent model is proposed. Assumptions 4, 5 and 6 are regularity conditions for proving the existence of a mean field Nash equilibrium.

Moreover, we provide the economic rationale behind Assumptions 4, 5, and 6. The condition (4.1.12) makes sure that the function $\delta \to \delta f(\delta, \cdot)$ has

a unique maximum. The condition (4.1.13) represents a bounded sensitivity condition of the intensity function $f(\delta, \mu)$, where the impact of changes in the mean field μ on the intensity function f is limited and dominated by the impact of the representative market maker's quote δ. This means that the influence of the aggregate market behavior represented by $\partial_\mu f(\delta, \mu)$ is not disproportionally large compared to the influence of the agent's own quotes represented by $\partial_\delta f(\delta, \mu)$. Hence, it imposes a stability condition on the intensity function to prevent drastic changes in the intensity function in response to a small fluctuation in the mean quote μ. The condition (4.1.14) introduces another stability condition in terms of the higher-order sensitivity of f with respect to both δ and μ. The assumption is similar to one made by [Luo and Zheng 2021] to prove the general Issac's condition. The high-order or cross-term sensitivity of f involving both δ and μ should not be excessively large compared to that with respect to δ only. This ensures that the combined influence of δ and μ on f remains controlled, avoiding extreme sensitivity in the cross-terms.

We first show a uniform boundedness result of the objective function (4.1.9). The proof is similar to that in Chapter 3 where the third assumption of Assumption 4.1.2 plays the main role. We first need the following lemma to rewrite the objective function (4.1.9).

Proposition 4.1.4. *Under Assumption 4.1.2, the objective function (4.1.9) can be written as*

$$J_0(\boldsymbol{\delta}; \bar{\boldsymbol{\delta}}, \boldsymbol{M}) = \mathbb{E}^{\boldsymbol{\delta}, \bar{\boldsymbol{\delta}}, \boldsymbol{M}}\Big[\int_0^T e^{-rt}\Big(\lambda^a \delta_t^a f_a(\delta_t^a, \bar{\mu}_t^a)\mathbb{I}(q_t > -Z) + \lambda^b \delta_t^b f_b(\delta_t^b, \bar{\mu}_t^b)\mathbb{I}(q_t < Z)\Big)dt - e^{-rT}\phi(q_T) - \int_0^T e^{-rt}\psi(q_t)dt\Big] \tag{4.1.15}$$

There exists a constant $J_{max}(T) > 0$ such that for any quoting strategy δ and population distribution flow M

$$|J_0(\boldsymbol{\delta}; \bar{\boldsymbol{\delta}}, \boldsymbol{M})| \leq J_{max}(T). \tag{4.1.16}$$

Proof. First, observe that running cost $\psi(q_t)$ is uniformly bounded since the process q_t takes values from the finite set $\mathcal{Q}$. Then the expectation for the running cost term $\mathbb{E}[\int_0^T e^{-rt}\psi(q_t)dt] < \infty$ is uniformly bounded for any $T > 0$.

Assumption 4.1.2 implies that $|\delta_t^a f_a(\delta_t^a, \bar{\mu}_t^a)| \leq |\delta_t^a \Lambda(\delta_t^a)|$. Since $\lim_{\delta\to\infty} \Lambda(\delta)\delta = 0$, $\delta_t^a = \delta^a(t, q_t)$ takes values on $[-\delta_\infty, +\infty)$ then there exists a constant $\Delta > 0$, such that $\forall \delta > \Delta$, $|\Lambda(\delta)\delta| < 1$. From the continuity of function $\delta \to \Lambda(\delta)\delta$ in the closed interval $[-\delta_\infty, \Delta]$, this function is also bounded. We can then define $K_a := \max(\sup_{\delta\in[-\delta_\infty,\Delta]} \Lambda(\delta)\delta, 1)$ to obtain $|\delta_t^a f_a(\delta_t^a, \bar{\mu}_t^a)| \leq K_a$ uniformly.

Similarly, we obtain the uniform boundedness of $|\delta_t^b f_b(\delta_t^b, \bar{\mu}_t^b)|$ by a constant K_b.

Hence, we have the following estimation.

$$\begin{aligned}\int_0^T e^{-rt}\lambda^a \delta_t^a f_a(\delta_t^a, \bar{\mu}_t^a)\mathbb{I}(q_t > -Z)dt \leq \lambda_a K_a \int_0^T e^{-rt}dt < \infty \\ \int_0^T e^{-rt}\lambda^b \delta_t^b f_b(\delta_t^b, \bar{\mu}_t^b)\mathbb{I}(q_t < Z)dt \leq \lambda_b K_b \int_0^T e^{-rt}dt < \infty\end{aligned} \tag{4.1.17}$$

Therefore

$$\begin{aligned}\mathbb{E}\left[\int_0^T e^{-rt}\delta_t^a N^a(dt)\right] &= \mathbb{E}\left[\int_0^T e^{-rt}\lambda^a \delta_t^a f_a(\delta_t^a, \bar{\mu}_t^a)\mathbb{I}(q_t > -Z)dt\right] \\ \mathbb{E}\left[\int_0^T e^{-rt}\delta_t^b N^b(dt)\right] &= \mathbb{E}\left[\int_0^T e^{-rt}\lambda^b \delta_t^b f_b(\delta_t^b, \bar{\mu}_t^b)\mathbb{I}(q_t < Z)dt\right]\end{aligned} \tag{4.1.18}$$

By combining the dynamics of inventory and cash processes (4.1.5)-(4.1.7) and applying Itô's formula on process $(X_t + q_t S_t)$ one can write the objective function (4.1.9) as

$$\begin{aligned}J_0(\boldsymbol{\delta}; \bar{\boldsymbol{\delta}}, \boldsymbol{M}) &= \mathbb{E}^{\boldsymbol{\delta}, \bar{\boldsymbol{\delta}}, \boldsymbol{M}}\Big[\int_0^T e^{-rt}\Big(\delta_t^a N^a(dt) + \delta_t^b N^b(dt)\Big) - e^{-rT}\phi(q_T) - \int_0^T e^{-rt}\psi(q_t)dt\Big] \\ &= \mathbb{E}^{\boldsymbol{\delta}, \bar{\boldsymbol{\delta}}, \boldsymbol{M}}\Big[\int_0^T e^{-rt}\Big(\lambda^a \delta_t^a f_a(\delta_t^a, \bar{\mu}_t^a)\mathbb{I}(q_t > -Z) + \lambda^b \delta_t^b f_b(\delta_t^b, \bar{\mu}_t^b)\mathbb{I}(q_t < Z)\Big)dt \\ &\qquad - e^{-rT}\phi(q_T) - \int_0^T e^{-rt}\psi(q_t)dt\Big]\end{aligned} \tag{4.1.19}$$

Combining the estimation (4.1.17) with (4.1.19) we obtain

$$|J_0(\boldsymbol{\delta}; \bar{\boldsymbol{\delta}}, \boldsymbol{M})| \leq \frac{\lambda^a K_a + \lambda^b K_b + \max\limits_{q \in \mathcal{Q}} \psi(q)}{r} + e^{-rT}\max_{q \in \mathcal{Q}} \phi(q) \tag{4.1.20}$$

□

Remark 4.1.5. The upper bound in Proposition 4.1.4 is still valid for $J_t(\boldsymbol{\delta}; \bar{\boldsymbol{\delta}}, \boldsymbol{M})$. From the proof, we see that Proposition 4.1.4 applies to the case when T tends to ∞. Hence the objective function with infinite horizon is well-defined:

$$\begin{aligned}J^\infty(\boldsymbol{\delta}; \bar{\boldsymbol{\delta}}, \boldsymbol{M}) = \mathbb{E}^{\boldsymbol{\delta}; \bar{\boldsymbol{\delta}}, \boldsymbol{M}}\Big[\int_0^\infty e^{-rt}\Big(\lambda^a \delta_t^a f_a(\delta_t^a, \bar{\mu}_t^a)\mathbb{I}(q_t > -Z) + \lambda^b \delta_t^b f_b(\delta_t^b, \bar{\mu}_t^b)\mathbb{I}(q_t < Z)\Big)dt \\ - \int_0^\infty e^{-rt}\psi(q_t)dt\Big]\end{aligned} \tag{4.1.21}$$

where $\boldsymbol{\delta}$ take values from $\mathcal{A}_0^\infty$.

From Proposition 4.1.4 on the uniform boundedness of the value function $V^{\bar{\boldsymbol{\delta}}, \boldsymbol{M}}(t, q)$, for given t, $V^{\bar{\boldsymbol{\delta}}, \boldsymbol{M}}(t, \cdot)$ is an element of $\mathbb{R}^{2H+1}$ equipped with the

norm $||\cdot||_\infty$. Hence we can represent the value function as a map $\boldsymbol{V}^{\bar{\boldsymbol{\delta}},\boldsymbol{M}}$: $[0,T] \to (\mathbb{R}^{2H+1}, ||\cdot||_\infty)$ with $\boldsymbol{V}^{\bar{\boldsymbol{\delta}},\boldsymbol{M}}(t) = V^{\bar{\boldsymbol{\delta}},\boldsymbol{M}}(t,\cdot) \in \mathbb{R}^{2H+1}$. This notation provides the necessary framework for proving the existence of mean field Nash equilibrium, as is defined below.

Definition 4.1.6. (Mean field Nash equilibrium) We say a flow of population distribution $\boldsymbol{M}^*$ is a mean field Nash equilibrium (MFNE) for (4.1.2)-(4.1.9) if there exists a quoting strategy $\boldsymbol{\delta}^* \in \mathcal{A}_0^T$, such that

$$J_0(\boldsymbol{\delta}^*; \boldsymbol{\delta}^*, \boldsymbol{M}^*) = \max_{\boldsymbol{\delta} \in \mathcal{A}_0^T} J_0(\boldsymbol{\delta}; \boldsymbol{\delta}^*, \boldsymbol{M}^*) \tag{4.1.22}$$

and, for all $t \in [0,T]$, $\boldsymbol{M}^*(t) \in \mathcal{P}(\mathcal{Q})$ is the distribution of the representative market maker's inventory $q_t^{\boldsymbol{\delta}^*}$ controlled by the quoting strategy $\boldsymbol{\delta}^*$. We call $\boldsymbol{\delta}^*$ the mean field quoting strategy.

Remark 4.1.7. The optimality condition 4.1.22 means that, in equilibrium, market makers have no incentive to deviate from the equilibrium quoting strategy δ^* when the (rest of the) population is following this strategy.

Prior to prove the existence of mean field Nash equilibrium in our mean field game of market making model, we first define mappings Ξ^a and Ξ^b on $\mathbb{R}^2$ as the $argmax$ of following functions which will appear in (4.1.25):

$$\begin{aligned} \Xi^a(p,\mu) &= \arg\max_{\delta > -\delta_\infty} f_a(\delta,\mu)(\delta - p) \\ \Xi^b(p,\mu) &= \arg\max_{\delta > -\delta_\infty} f_b(\delta,\mu)(\delta - p) \end{aligned} \tag{4.1.23}$$

We shall see in Lemma B.1.1 that Ξ^a and Ξ^b are well-defined and continuously differentiable functions of (p,μ).

Define the Hamiltonian functions $\mathcal{H}^a, \mathcal{H}^b : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ as

$$\mathcal{H}^a(p,\mu) = \lambda_a \sup_{\delta > -\delta_\infty} f_a(\delta,\mu)(\delta - p) \qquad \mathcal{H}^b(p,\mu) = \lambda_b \sup_{\delta > -\delta_\infty} f_b(\delta,\mu)(\delta - p) \tag{4.1.24}$$

We can show that the mean field Nash equilibrium defined by Definition 4.1.6 is characterised by a system of coupled Hamilton-Jacobi equations and

Chapman-Kolmogorov equations:

$$\begin{cases} 0 = \partial_t V - rV - \psi(q) + \mathcal{H}^a\Big(V(t,q) - V(t,q-1), \mu^a(t)\Big)\mathbb{I}(q > -Z) \\ \qquad\qquad + \mathcal{H}^b\Big(V(t,q) - V(t,q+1), \mu^b(t)\Big)\mathbb{I}(q < Z) \\ V(T,q) = -\phi(q) \\ \mu^a(t) = \sum_{\xi \in \mathcal{Q}} \Xi^a\big(V(t,\xi) - V(t,\xi-1), \mu^a(t)\big) m(t,\xi)\mathbb{I}(\xi > -Z) \\ \mu^b(t) = \sum_{\xi \in \mathcal{Q}} \Xi^b\big(V(t,\xi) - V(t,\xi+1), \mu^b(t)\big) m(t,\xi)\mathbb{I}(\xi < Z) \\ 0 = \partial_t m(t,q) - \lambda_a f_a\Big(\Xi^a\big(V(t,q+1) - V(t,q), \mu^a(t)\big), \mu^a(t)\Big) m(t,q+1)\mathbb{I}(q < Z) \\ \qquad - \lambda_b f_b\Big(\Xi^b\big(V(t,q-1) - V(t,q), \mu^b(t)\big), \mu^b(t)\Big) m(t,q-1)\mathbb{I}(q > -Z) \\ \qquad + \lambda_a f_a\Big(\Xi^a\big(V(t,q) - V(t,q-1), \mu^a(t)\big), \mu^a(t)\Big) m(t,q)\mathbb{I}(q > -Z) \\ \qquad + \lambda_b f_b\Big(\Xi^b\big(V(t,q) - V(t,q+1), \mu^b(t)\big), \mu^b(t)\Big) m(t,q)\mathbb{I}(q < Z) \\ m(0,q) = m_0(q) \end{cases} \tag{4.1.25}$$

The following result, whose proof is provided in Appendix B.1, shows the existence of a mean field Nash equilibrium, characterized by the system of equations (4.1.25).

Theorem 4.1.8. *Under Assumption 4.1.2, there exists a solution (V^*, m^*) to (4.1.25), such that*

$$V^*, m^* \in \mathcal{B}([0,T] \times \mathcal{Q}), \qquad \text{and} \quad \forall q \in \mathcal{Q}, \quad V^*(\cdot,q), m^*(\cdot,q) \in C^1([0,T], \mathbb{R}). \tag{4.1.26}$$

Proposition 4.1.9. *(Verification) Let $(\tilde{V}^*, \tilde{m}^*)$ be a solution to (4.1.25) satisfying* (4.1.26) *and define the associated quoting strategy:*

$$\begin{aligned} \tilde{\delta}^{a,*}(t,q) &= \Xi^a\Big(\tilde{V}^*(t,q) - \tilde{V}^*(t,q-1), \tilde{\mu}^{a,*}(t)\Big) \\ \tilde{\delta}^{a,*}(t,q) &= \Xi^b\Big(\tilde{V}^*(t,q) - \tilde{V}^*(t,q+1), \tilde{\mu}^{b,*}(t)\Big) \end{aligned} \tag{4.1.27}$$

where Ξ^a, Ξ^b are defined in (4.1.23), and $\tilde{\mu}^{a,*}(t), \tilde{\mu}^{b,*}(t)$ are solutions to

$$\begin{aligned} \tilde{\mu}^{a,*}(t) &= \sum_{\xi \in \mathcal{Q}} \Xi^a\big(\tilde{V}^*(t,\xi) - \tilde{V}^*(t,\xi-1), \tilde{\mu}^{a,*}(t)\big) \tilde{m}^*(t,\xi)\mathbb{I}(\xi > -Z) \\ \tilde{\mu}^{b,*}(t) &= \sum_{\xi \in \mathcal{Q}} \Xi^b\big(\tilde{V}^*(t,\xi) - \tilde{V}^*(t,\xi+1), \tilde{\mu}^{b,*}(t)\big) \tilde{m}^*(t,\xi)\mathbb{I}(\xi < Z) \end{aligned} \tag{4.1.28}$$

Define the population distribution flow $\tilde{\boldsymbol{M}}^* : [0,T] \to \mathcal{P}(\mathcal{Q})$ associated with density $\tilde{m}^*(t,q)$ at time t, and write $\tilde{\boldsymbol{\delta}}^* = (\tilde{\delta}^{a,*}, \tilde{\delta}^{b,*})$. Then $\tilde{\boldsymbol{M}}^*$ is mean field Nash equilibrium defined in Definition 4.1.6, and $\tilde{\boldsymbol{\delta}}^*$ is the associated mean field quoting strategy.

The proof is provided in Appendix B.2.

As a consequence of Theorem 4.1.8 and Proposition 4.1.9 we obtain:

Corollary 4.1.10 (Existence of Nash Equilibrium)**.** *Under Assumption 4.1.2, there exists a mean field Nash equilibrium for system (4.1.2)-(4.1.9).*

The uniqueness of equilibrium is far from obvious in mean field games, and requires assumptions on the monotonicity of the associated Hamiltonian with respect to the probability measure under separation of variables ([Lasry and Lions 2006a; Lasry and Lions 2006b; Lasry and Lions 2007]). For mean field games on a finite state space similar monotonicity and separability conditions have been used by [Gomes, Mohr, and Souza 2013; Guéant 2011] to derive uniqueness. However, in our model, the functions $\mathcal{H}^a, \mathcal{H}^b$ defined in (4.1.24) do not necessarily possess a separability property in the variables p and μ. [Guéant 2015] proposes an algebraic condition for uniqueness in a more general framework without separability property. In spirit of [Guéant 2015] we provide, in Appendix B.3, an algebraic condition for uniqueness of Nash equilibrium.

4.2 Learning Dynamics: Mean Field Deep Reinforcement Learning

The mean field Nash equilibrium characterized by system of equations (4.1.25) results from the interactions of a representative market maker with the mean field generated by the distribution of all market makers' states. This equilibrium relies implicitly on the rational expectation assumption, where each agent is assumed to have complete knowledge of the population dynamics. Agents respond optimally to the overall population distribution, and anticipates that other agents will also behave rationally. They consistently expect to know the evolution of the population distribution, leading to a fixed point distribution to which they respond with an optimal policy, and which coincides with the population distribution resulting from this optimal policy used by all the agents ([Guéant, Lasry, and Lions 2011]).

However, the notion of equilibrium does not necessarily characterize the procedure of reaching this equilibrium state. In practice, market makers employ automated algorithms that dynamically adjust their quotes based on observed market data without knowing the underlying market model. Algorithm automation specifically leads us to study reinforcement learning algorithms that enable agents to learn their decision strategy through interactions with

the environment by trial and error. In this case, the assumption of rational expectation is relaxed, and we are interested in the potential market dynamics resulted from the learning algorithms with adaptive expectations, which can be reflected by our algorithm. Compared to multi-agent reinforcement learning algorithms, mean field game learning has the advantage of efficiency resolving computational intractability when number of agents is large while maintaining realistic nature of agents' learning behavior. We hereby study the effects brought by the mean field reinforcement learning algorithm applied to market making problems.

Algorithms for learning mean field games usually involves alternately updating the population distribution and representative agent's strategy. [Guo, Hu, et al. 2019] study the learning problem associated with the generalized mean field game (GMFG) and prove the convergence of the mean field Q-learning algorithm. We refer to [Laurière et al. 2022] for a comprehensive survey on learning problems in mean field games. Note that in this literature, learning algorithms are primarily employed as numerical method to compute equilibrium. Our study distinctly focuses on exploring the dynamics resulting from the application of these algorithms by market makers without prior knowledge of market dynamics, shedding light on the real-world behavioral patterns and strategic interactions of market makers influenced by learning algorithms.

4.2.1 Computing Nash Equilibria

The numerical methods for solving mean field games have been studied extensively along with its theoretical development over the past decade. The foundational works by [Achdou and Capuzzo-Dolcetta 2010] and [Achdou, Camilli, and Capuzzo-Dolcetta 2012] develop finite different schemes to solve the coupled HJB and FPK equations of MFG systems in both infinite and finite time horizon. The authors analyze existence and uniqueness of solution to the numerical scheme, and provide bounds on the solution. Examples of convergence results have also been shown. Further advancements include the work of [Achdou and Kobeissi 2021], who introduce finite difference approximations for mean field games of controls, in which the agents interact through distribution of both their states and strategies. [Achdou and Laurière 2020] provide a comprehensive overview of different numerical methods with an emphasis on the iterative schemes and applications to various MFG problems. [Laurière 2021] further surveys numerical approaches for both MFGs and mean field type

control with discussions on both the numerical schemes and the algorithms to solve these schemes. The authors also include more recent developments such as the monotonic operator method ([Almulla, Ferreira, and Gomes 2017]) and neural network approximations ([Carmona and Laurière 2021]).

In this section, we propose a Picard fixed-point iteration scheme to numerically solve the system of MFG equations (4.1.25), as described by [Laurière 2021]. We consider intensity functions f_a, f_b of the form:

$$\begin{aligned} f_a(\delta, \mu) &= \frac{1}{C_a e^{\delta} + C_{am} e^{k_{am}\delta - k_{a\mu}\mu}} \\ f_b(\delta, \mu) &= \frac{1}{C_b e^{\delta} + C_{bm} e^{k_{bm}\delta - k_{b\mu}\mu}} \end{aligned} \tag{4.2.1}$$

Meanwhile, we define a quadratic inventory cost function $\psi(q) = \frac{1}{2} \times 0.01q^2$ and a quadratic terminal cost $\phi(q) = -0.01q^2$. The interest rate is set $r = 0.01$. An inventory limit of $H = 10$ is imposed on the representative market maker. The order flow arrival rates are symmetric on the ask and bid side $\lambda_a = \lambda_b = 5$. For the intensity function, we specify the parameters as follows: $k_a = k_b = 2, k_{am} = k_{bm} = K_{a\mu} = K_{b\mu} = 3, C_a = C_b = 1, C_{am} = C_{bm} = 1$. We can verify that (4.2.1) satisfies Assumption 4.1.2, hence a mean field market making problem with (4.2.1) yields a Nash equilibrium. Figure 4.1 shows the intensity as functions of the representative market maker's centered quotes. These intensities are decreasing in terms of centered quotes, but increasing as functions of average quoted prices by competitors. The intensities are bounded by $\Lambda(\delta) = e^{-\delta}$ which can be regarded as the intensity when there is a single market maker, as well as the upper bound in Assumption 4.1.2.

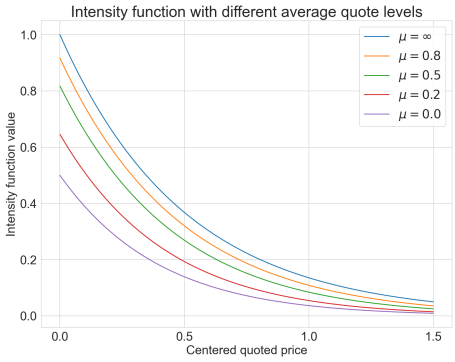


Figure 4.1: Intensity functions with different average centered quotes, which are increasing functions of μ.

We compute mean field Nash equilibrium ask and bid quotes by numeri-

cally solving the system of equations ((4.1.25)) using a Euler finite-difference scheme, coupled with an iterative method for computing the fixed point. Define the time grid $0 = t_0 < t_1 < \ldots < t_N = T$ with $\Delta t_n = t_{n+1} - t_n$, and compute the approximate value function $\mathcal{V} = \big((\tilde{V}_n(q))_{q\in\mathcal{Q}}\big)_{n\in\{0,\ldots,N\}}$ and the approximate flow of the density function $\mathcal{M} = \big((\tilde{m}_n(q))_{q\in\mathcal{Q}}\big)_{n\in\{0,\ldots,N\}}$ by fixed point iteration, where $\tilde{V}_n(q) = \tilde{V}(t_n, q), \tilde{m}_n(q) = \tilde{m}(t_n, q)$. Within each fixed point iteration $\mathcal{V}$ is computed by a backward iteration starting with the terminal condition $\tilde{V}_N(q) = -\phi(q)$ and $\mathcal{M}$ is updated by a forward iteration starting from $\tilde{m}_0(q) = m_0(q)$. At k^{th} fixed point iterations, the values $\mathcal{V}^{(k)}$ are updated according to the backward Euler scheme (4.2.2) using the density flow $\mathcal{M}^{(k-1)}$ from the $(k-1)^{th}$ fixed point iteration.

$$\begin{cases} \tilde{V}_n^k(q) = \tilde{V}_{n+1}^k(q) - \Delta t_n \Big(r\tilde{V}_{n+1}^k(q) + \psi(q) - \mathcal{H}^a\Big(\tilde{V}_{n+1}^k(q) - \tilde{V}_{n+1}^k(q-1), \mu_{n+1}^a\Big)\mathbb{I}(q > -Z) \\ \qquad\qquad -\mathcal{H}^b\Big(\tilde{V}_{n+1}^k(q) - \tilde{V}_{n+1}^k(q+1), \mu_{n+1}^b\Big)\mathbb{I}(q < Z)\Big) \\ \mu_{n+1}^a = \sum\limits_{\xi\in\mathcal{Q}} \Xi^a\big((\tilde{V}_{n+1}^k(\xi) - \tilde{V}_{n+1}^k(\xi-1), \mu_{n+1}^a\big)\tilde{m}_{n+1}^{k-1}(\xi)\mathbb{I}(\xi > -Z) \\ \mu_{n+1}^b = \sum\limits_{\xi\in\mathcal{Q}} \Xi^b\big(\tilde{V}_{n+1}^k(\xi) - \tilde{V}_{n+1}^k(\xi+1), \mu_{n+1}^b\big)\tilde{m}_{n+1}^{k-1}(\xi)\mathbb{I}(\xi < Z) \\ \tilde{V}_N^k(q) = -\phi(q) \end{cases} \tag{4.2.2}$$

After calculating $\mathcal{V}^{(k)}$, the density function flow $\mathcal{M}^{(k)}$ is updated via the forward Euler scheme (4.2.3) using the calculated value functions $\mathcal{V}^{(k)}$ within k^{th} fixed point iteration.

$$\begin{cases} \tilde{m}_{n+1}^k(q) = \tilde{m}_n^k(q) + \Delta t_n \Big(\lambda_a f_a\Big(\Xi^a\big(\tilde{V}_n^k(q+1) - \tilde{V}_n^k(q), \mu_n^a\big), \mu_n^a\Big)\tilde{m}_n^k(q+1)\mathbb{I}(q < Z) + \\ \qquad\qquad \lambda_b f_b\Big(\Xi^b\big(\tilde{V}_n^k(q-1) - \tilde{V}_n^k(q), \mu_n^b\big), \mu_n^b\Big)\tilde{m}_n^k(q-1)\mathbb{I}(q > -Z) - \\ \qquad\qquad \lambda_a f_a\Big(\Xi^a\big(\tilde{V}_n^k(q) - \tilde{V}_n^k(q-1), \mu_n^a\big), \mu_n^a\Big)\tilde{m}_n^k(q)\mathbb{I}(q > -Z) - \\ \qquad\qquad \lambda_b f_b\Big(\Xi^b\big(\tilde{V}_n^k(q) - \tilde{V}_n^k(q+1), \mu_n^b\big), \mu_n^b\Big)\tilde{m}_n^k(q)\mathbb{I}(q < Z) \\ \mu_n^a = \sum\limits_{\xi\in\mathcal{Q}} \Xi^a\big((\tilde{V}_n^k(\xi) - \tilde{V}_n^k(\xi-1), \mu_n^a\big)\tilde{m}_n^k(\xi)\mathbb{I}(\xi > -Z) \\ \mu_n^b = \sum\limits_{\xi\in\mathcal{Q}} \Xi^b\big(\tilde{V}_n^k(\xi) - \tilde{V}_n^k(\xi+1), \mu_n^b\big)\tilde{m}_n^k(\xi)\mathbb{I}(\xi < Z) \\ m_0(q) = m_0(q) \end{cases} \tag{4.2.3}$$

The numerical scheme is summarized below (Algorithm 4). We study the numerical convergence based on L^2 norm of difference in values and densities between two consecutive fixed point iteration $||\mathcal{V}^{(k)} - \mathcal{V}^{(k-1)}||_2$ and $||\mathcal{M}^{(k)} - \mathcal{M}^{(k)}||_2$.

Algorithm 4 Finite difference scheme for computing mean field Nash equilibrium

Input: K = number of fixed point iterations, N =number of discrete time steps on interval $[0, T]$, f_a, f_b, λ^a, λ^b: intensity of ask and bid order flow, ψ: inventory cost function, ϕ: terminal cost function

Output: Approximated value function, population distribution and mean field Nash equilibrium

1: Initialize values $\mathcal{V}^{(0)}$ and densities $\mathcal{M}^{(0)}$.
2: **for** $k \leftarrow 1$ to K **do**
3: **Value iteration**: Compute $\mathcal{V}^{(k)} = \left((\tilde{V}_n^k(q))_{q\in\mathcal{Q}}\right)_{n\in\{0,\dots,N\}}$ with backward Euler scheme (4.2.2) where the density function flow is fixed as $\mathcal{M}^{(k-1)}$.
4: **Density iteration**: Compute $\mathcal{M}^{(k)} = \left((\tilde{m}_n^k(q))_{q\in\mathcal{Q}}\right)_{n\in\{0,\dots,N\}}$ with forward Euler scheme (4.2.3) where the value function is fixed as $\mathcal{V}^{(k)}$.
5: **end for**
6: Compute approximated Nash equilibrium quoting strategy using $\mathcal{V}^{(K)}$ and $\mathcal{M}^{(K)}$ from (4.1.27) and (4.1.28).

Remark 4.2.1. Algorithm 4 is relatively straightforward compared to some of the more advanced methods available in the literature. However, our chosen approach has several advantages that align well in the scope of our framework. The numerical scheme we propose is essentially a Picard fixed-point iteration method as discussed in [Laurière 2021]. This iterative approach is appropriate for our setting where the MFG equations (4.1.25) involves a finite state space, resulting in a system of coupled forward and backward ODEs without spatial derivative terms. Since only the time component needs to be discretized, the explicit Euler schemes (4.2.2) and (4.2.3) are a practical and effective choice. Moreover, our approach is computationally efficient and straightforward to implement. We aim at providing a benchmark solution to the MFG system that can be used for comparison with the learning algorithm simulation. This numerical scheme is shown to be sufficient with numerical convergence in the first few iterations presented in Figure 4.2.

In the numerical simulation, we set $T = 3000$ with a constant time step $\Delta t_n = \Delta t = 0.1$. We are interested in stationary behavior of representative agent in mean field game when $T \to \infty$.[1] The stationary setting is usually required for

[1]The general convergence of finite time horizon mean field game to stationary mean field game when $T \to \infty$ needs a rigorous proof. In our particular mean field game market making problem we do not approach this but only look at numerical evidence for convergence.

designing learning algorithm that learns strategies independent of time, which will be studied in Section 4.2.2. Therefore we set a large value $T = 3000$ for an approximation of the stationary mean field game when $T \to \infty$. $T = 3000$ is a reasonable time horizon as the numerical MFG value function $\tilde{V}_0(q)$ does not change when T goes beyond 3000, indicating good approximation for the stationary case. Figure 4.2 shows the L_2 norm of difference in value and density functions between fixed point iteration, demonstrating the numerical convergence of the algorithm. After 5 iterations, the error is below 10^{-8}, from which we can conclude that a fixed point is found through iteration corresponding to the mean field Nash equilibrium.

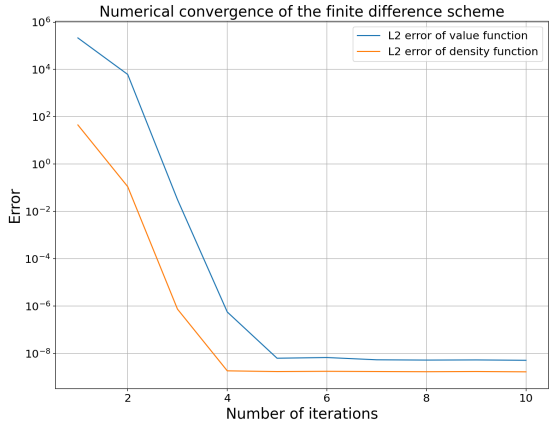


Figure 4.2: Numerical convergence of the finite difference scheme.

The MFG value function at $t = 0$ and the population density function at $t = T$ are shown in Figure 4.3. As discussed with a large terminal $T = 3000$, the numerical MFG value function $\tilde{V}(0, q)$ approximates the stationary MFG value function when $T \to \infty$. The inventory distribution $\tilde{m}(T, q)$ approximates the population distribution of stationary MFG. Both functions achieve maximum in inventory $q = 0$, indicating the market maker's aversion to inventory risk.

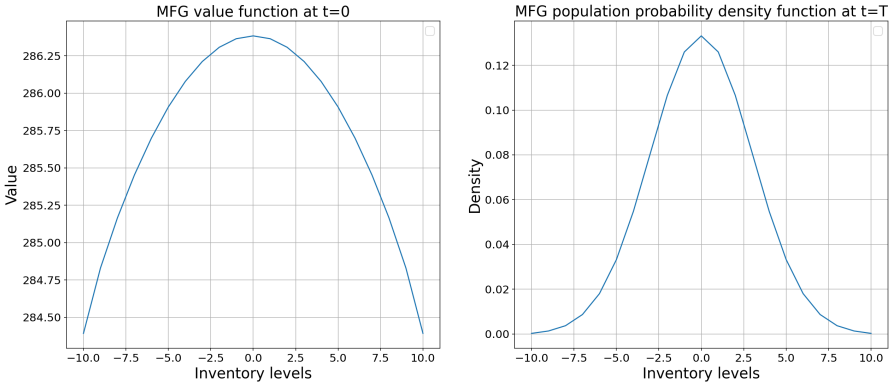


Figure 4.3: Mean field game value function ($t = 0$) and population density function ($t = T$) from finite difference scheme.

We apply (4.1.27) and (4.1.28) to the numerical MFG value functions $\tilde{V}(0, q)$ to compute the ask and bid quotes at $t = 0$ by a representative market maker. Using a finite difference scheme similar to (4.2.2) but replacing the MFG intensity functions with the upper bound function $\Lambda(\delta)$ from Assumption 4.1.2, we can calculate the monopolistic ask and bid quotes at $t = 0$. Figure 4.4 shows the monopolistic value function with the intensity function $\Lambda(\delta)$ and the comparison between the MFG and the monopolistic quoting strategies, where the monopolistic quotes are above the MFG quotes. We consider MFG and monopolistic quotes as 2 benchmark cases in the learning scenario in the next subsection.

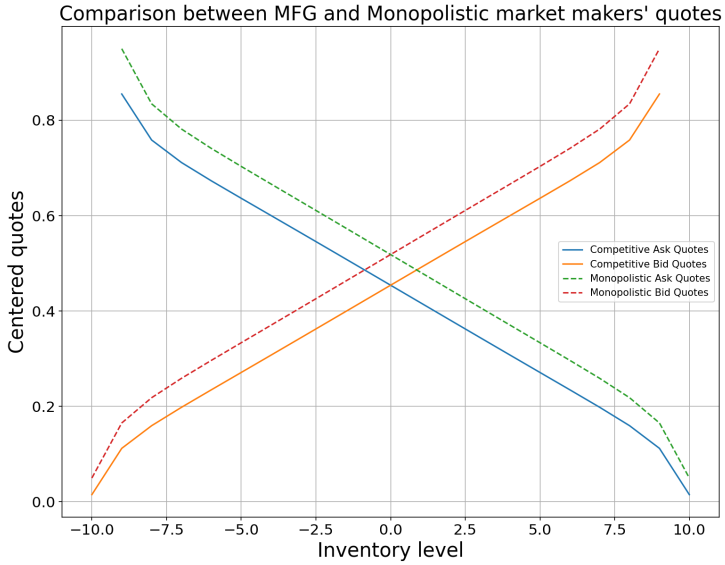


Figure 4.4: Mean field game ask and bid quotes at $t = 0$ compared to monopolistic ask and bid quotes at $t = 0$.

4.2.2 Decentralized Deep Reinforcement Learning

Classical reinforcement learning setup is usually based on the Bellman equation derived from the Markov decision process ([Sutton and Barto 2018]). In Chapter 3, we reformulate the stochastic differential game to multi-agent Markov decision process (MDP) and design a multi-agent deep reinforcement learning algorithm corresponding to this system. Analogously to Chapter 3, we convert the mean field game system in Section 4.1 into a Markov decision process with population distribution incorporated as state variables, and propose a mean field deep reinforcement learning algorithm for simulation of market makers.

In practical scenarios, reinforcement learning algorithm is often based on Bellman equation for infinite horizon Markov decision process, in order to derive a stationary state-action value function hence a stationary strategy that does not vary with time. The adaption of infinite horizon favors our simulation of continuous market making without a terminal time, which particularly corresponds to the case of OTC markets such as FX and bond markets. In this section, we adopt the model setting with infinite horizon and leave the theoretical convergence of finite horizon to infinite horizon mean field game for future research. Another simplification adapted in our simulation is that we equate the population quoting strategy with the representative market maker's quoting strategy $\bar{\boldsymbol{\delta}} = \boldsymbol{\delta}$ in (4.1.21). This simplification is based on the def-

inition 4.1.6 that the population quoting strategy coincides with the agent quoting strategy in equilibrium. Instead of searching in quoting strategy space of the population we equate it with agent quoting strategy in simulation hence reducing the complexity of our mean field game market making problem. For simplicity, we use $\mathbb{E}^{\boldsymbol{\delta};\boldsymbol{M}}$ to denote the expectation $\mathbb{E}^{\boldsymbol{\delta};\boldsymbol{\delta},\boldsymbol{M}}$ used in (4.1.9). Therefore, given a population distribution flow $M : [0,\infty] \to \mathcal{S}^{\mathcal{Q}}$ and given a quoting strategy $\bar{\boldsymbol{\delta}} \in \mathcal{A}_0^\infty$, the value function is defined as

$$V^{\boldsymbol{\delta};\boldsymbol{M}}(q) = \mathbb{E}^{\boldsymbol{\delta};\boldsymbol{M}}\Big[\int_0^\infty e^{-rt}\Big(\lambda^a\delta_t^a f_a(\delta_t^a,\bar{\mu}_t^a)\mathbb{I}(q_t > -Z) + \lambda^b\delta_t^b f_b(\delta_t^b,\bar{\mu}_t^b)\mathbb{I}(q_t < Z)\Big)dt - \int_0^\infty e^{-rt}\psi(q_t)dt\Big] \tag{4.2.4}$$

Under a given population distribution flow M, the objective of a representative market maker is to learn a Markovian quoting strategy solely dependent on inventory level and independent of time by interacting with the market environment. We restrict quoting strategies to Markovian: $\delta_k(q), k \in \{a,b\} : \mathcal{Q} \to \mathbb{R}$, or equivalently, a Markovian quoting strategy is regarded as a vector in space $(I_\delta)^{2H+1}$.

We formulate a Markov decision process to simulate the practical scenario with a representative market maker learning a strategy in mean field game. Given the population inventory distribution flow M, at time t for the representative market maker with inventory level q_t and quoting strategies $\delta_t^a, \delta_t^b \in (I_\delta)^{2H+1}$, the probability of winning the next ask and bid RFQ is $f_a(\delta_t^a,\mu_t^a)$ and $f_b(\delta_t^b,\mu_t^b)$, where $\mu_t^a = \delta_t^a \cdot M_t, \mu_t^b = \delta_t^b \cdot M_t$ are the mean quoted ask and bid.

Similar to Chapter 3, we hereby introduce indicators I_t^a, I_t^b:

$$\begin{aligned} I_t^a &= \mathbb{I}(\text{Representative market maker wins the ask RFQ at time } t) \\ I_t^b &= \mathbb{I}(\text{Representative market maker wins the bid RFQ at time } t) \end{aligned} \tag{4.2.5}$$

We have $\mathbb{P}(I_t^a = 1) = f_a(\delta_t^a,\mu_t^a)$ and $\mathbb{P}(I_t^b = 1) = f_b(\delta_t^b,\mu_t^b)$.

Define stopping time $\tau := \tau_a \wedge \tau_b$ where τ_a and τ_b are the first arrival time of ask and bid RFQs after time 0.

$$\tau_a := \inf\{t > 0, \int_0^t N^a(dt) > 0\}, \quad \tau_b := \inf\{t > 0, \int_0^t N^b(dt) > 0\} \tag{4.2.6}$$

The state transition of the representative market maker's inventory is hence

$$q_\tau = q_{\tau-} - I_\tau^a\mathbb{I}(\tau_a < \tau_b)\mathbb{I}(q_{\tau-} > -Z) + I_\tau^b\mathbb{I}(\tau_b < \tau_a)\mathbb{I}(q_{\tau-} < Z) \tag{4.2.7}$$

From Lemma 3.5.1 in Chapter 3, stopping times τ_a and τ_b satisfy

$$\mathbb{E}\left[\int_0^{\tau_a\wedge\tau_b} e^{-rt}dt\right] = \frac{1}{r+\lambda_a+\lambda_b}$$
$$\mathbb{E}\left[e^{-r\tau_a}|\tau_a<\tau_b\right] = \mathbb{E}\left[e^{-r\tau_b}|\tau_b<\tau_a\right] = \frac{\lambda_a+\lambda_b}{r+\lambda_a+\lambda_b}$$

Following the methodology used in Chapter 3, we can show that the value function $V^{\boldsymbol{\delta};\boldsymbol{M}}(q)$ satisfies the Bellman equation:

$$\begin{aligned}
V^{\boldsymbol{\delta};\boldsymbol{M}}(q) = &-\frac{\psi(q)}{r+\lambda_a+\lambda_b}\\
&+\mathbb{P}(\tau_a<\tau_b)\mathbb{E}_q^{\boldsymbol{\delta};\boldsymbol{M}}\Big[I^a_{\tau_a}\left(e^{-r\tau_a}\delta^a_q + e^{-r\tau_a}V^{\boldsymbol{\delta};\boldsymbol{M}}(q-1)\right)\mathbb{I}(-Z<q\le Z)\\
&+(1-I^a_{\tau_a})e^{-r\tau_a}V^{\boldsymbol{\delta};\boldsymbol{M}}(q)\Big|\tau_a<\tau_b\Big]\\
&+\mathbb{P}(\tau_b<\tau_a)\mathbb{E}_q^{\boldsymbol{\delta};\boldsymbol{M}}\Big[I^b_{\tau_b}\left(e^{-r\tau_b}\delta^b_q + e^{-r\tau_b}V^{\boldsymbol{\delta};\boldsymbol{M}}(q+1)\right)\mathbb{I}(-Z\le q<Z)\\
&+(1-I^b_{\tau_b})e^{-r\tau_b}V^{\boldsymbol{\delta};\boldsymbol{M}}(q)\Big|\tau_b<\tau_a\Big]
\end{aligned} \tag{4.2.8}$$

It is important to note that the mean field, introduced as a probability distribution flow $M : [0,\infty) \to M_t \in \mathcal{S}^{\mathcal{Q}}$ which is dependent on time, does not affect the Markov property that underlies the derivation of (4.2.8). At the arrival time of an RFQ τ, the intensities $f_a(\delta^a_q, \mu^a)$ and $f_a(\delta^a_q, \mu^a)$ are determined by the population inventory distribution M_τ at time τ. The probabilities of the ask and bid RFQs are independent, defined as $\mathbb{P}(\tau_a < \tau_b) = \frac{\lambda_a}{\lambda_a+\lambda_b}$ and $\mathbb{P}(\tau_b < \tau_a) = \frac{\lambda_b}{\lambda_a+\lambda_b}$. The representative market maker, as a learning agent, responds to RFQs to refine her quoting strategy. The stationary quoting strategy $\delta^a, \delta^b \in (I_\delta)^{2H+1}$ is alternatively expressed as mappings $\delta_a : q \in \mathcal{Q} \to I_\delta$ and $\delta_b : q \in \mathcal{Q} \to I_\delta$, which facilitates functional representation of quoting strategies using neural networks in the next section.

From (4.2.8), we can define the reward function of the representative market maker at the arrival time of the RFQ τ. Note that the reward function $r(q,(\delta_a,\delta_b))$ is implicitly dependent on the population distribution M as well, since the probability of winning the RFQ is a function of average quotes.

$$\begin{aligned}
r(q,(\delta_a,\delta_b)) = &-\frac{\psi(q)}{r+\lambda_a+\lambda_b} + \frac{\lambda_a+\lambda_b}{r+\lambda_a+\lambda_b}\Big(I^a_{\tau_a}\mathbb{I}(\tau_a<\tau_b)\mathbb{I}(-Z<q\le Z)\cdot\delta_a\\
&+I^b_{\tau_b}\mathbb{I}(\tau_b<\tau_a)\mathbb{I}(-Z\le q<Z)\cdot\delta_b\Big)
\end{aligned} \tag{4.2.9}$$

A final gap between our formulation of the Bellman equation (4.1.3) and the implementation of the reinforcement learning algorithm is the discrepancy in time steps. In our framework, interval length between RFQs is stochastic, while the reinforcement learning algorithm usually requires fixed time step to run simulations. To resolve this gap, we still consider fixed time steps in the simulation that incorporate the expectation of RFQ arrival times in the discount factor. An RFQ always arrives at each time step t. From t to $t+1$, the inventory of the representative market maker changes with probability equal to the intensity right before $(t+1)$.

We further formulate (4.2.8) into a discrete-time Markov decision process in discrete time using the reward function (4.2.9) and the discount factor $\gamma = \frac{\lambda_a+\lambda_b}{r+\lambda_a+\lambda_b}$.

Let q_t be the inventory process of the representative market maker under quoting strategy $\boldsymbol{\delta}$ and population inventory distribution flow $\boldsymbol{M}$, then the state-action value function $V^{\boldsymbol{\delta};\boldsymbol{M}}(q)$ in (4.2.8) is written in the following format:

$$V^{\boldsymbol{\delta};\boldsymbol{M}}(q) = \mathbb{E}_q^{\boldsymbol{\delta};\boldsymbol{M}}\left[\sum_{t=0}^{\infty}\gamma^t r\left(q_t,(\delta^a_{q_t},\delta^b_{q_t})\right)\middle| q_0 = q\right] \tag{4.2.10}$$

where $\gamma = \frac{\lambda_a+\lambda_b}{r+\lambda_a+\lambda_b}$ and reward functions $r(q,(\delta_a,\delta_b))$ is defined in (4.2.9).

When the flow of population inventory distribution $\boldsymbol{M}$ is fixed, the representative market maker seeks to find optimal quoting strategy $\boldsymbol{\delta}^* = (\delta^*_a, \delta^*_b)$.

$$V^{\boldsymbol{M}}(q) = \max_{\boldsymbol{\delta}} \mathbb{E}_q^{\boldsymbol{\delta};\boldsymbol{M}}\left[\sum_{t=0}^{\infty}\gamma^t r\left(q_t,(\delta^a_{q_t},\delta^b_{q_t})\right)\middle| q_0 = q\right] \tag{4.2.11}$$

After finding $\boldsymbol{\delta}^*$, the flow of population inventory distribution $\boldsymbol{M'}$ is generated from the current optimal quoting strategy $\boldsymbol{\delta}^*$, under which the representative market maker learns a new optimal quoting strategy under $\boldsymbol{M'}$, until convergence to a fixed point. This fixed point iteration follows ideas on learning in mean field games [Guo, Hu, et al. 2019; Guo, Hu, et al. 2023; Laurière et al. 2022].

We apply an actor-critic learning algorithm to simulate the representative market maker learning a quoting strategy. In Chapter 3, we propose a decentralized multi-agent deep deterministic policy gradient (DDPG) to simulate the learning process of multiple market makers. The quoting strategy of the representative market maker (actor) is a pair of maps $\delta_a : q \in \mathcal{Q} \to \mathbb{R}$ and $\delta_b : q \in \mathcal{Q} \to \mathbb{R}$ parametrized as neural networks $\pi_a(q|\theta^\pi), \pi_b(q|\theta^\pi)$. A critic neural network $Q(q,(\delta^a,\delta^b),M|\theta^Q)$ is trained to approximate the value func-

tion $V^{\boldsymbol{\delta};\boldsymbol{M}}(q)$. The inputs for the networks are interpreted as follows: q is inventory level, (δ^a, δ^b) are ask and bid quotes, and $M \in \mathcal{S}^{\mathcal{Q}}$ refers to the population inventory distribution. Mean field game setting is incorporated to the algorithm through the population distribution M. The critic $Q\Big(q, (\delta^a, \delta^b), M|\theta^Q\Big)$ evaluates a combination of state-action pair $(q, (\delta^a, \delta^b))$ under a given population distribution M, and actor neural networks $\pi_a(q|\theta^\pi), \pi_b(q|\theta^\pi)$ generates the ask and bid quotes at a given state of the inventory level q. Parameters θ^Q and θ^π are trained through the interactions of the critic and the actor with the market environment.

Similar to the algorithm used in Chapter 3, we incorporate usage of the target critic and actor networks denoted by

$$\tilde{Q}\Big(q, (\delta^a, \delta^b), M|\tilde{\theta}^Q\Big)$$

and

$$\tilde{\pi}_a(q|\tilde{\theta}^\pi), \tilde{\pi}_b(q|\tilde{\theta}^\pi)$$

whose parameters are updated from θ^Q and θ^π the parameters of primal networks, but in a more stationary manner.

$$\begin{aligned} \tilde{\theta}^Q &\leftarrow \mu\theta^Q + (1-\mu)\tilde{\theta}^Q \\ \tilde{\theta}^\pi &\leftarrow \mu\theta^\pi + (1-\mu)\tilde{\theta}^\pi \end{aligned} \tag{4.2.12}$$

The training data is collected from the interactions of the representative market maker with the market environment. At each iteration from time t to $t+1$, a tuple $(q_t, \delta_t, q_{t+1}, r(q_t, \delta_t), I_t, M_t, \mu_t, d_t)$ is collected and saved in the replay buffer. q_t and q_{t+1} are the inventory levels at time t and $t+1$. δ_t is market maker's ask and quoting strategies given by the target actor network $\tilde{\pi}$. $r(q_t, \delta_t)$ is reward from market environment to the market maker for state-action pair (q_t, δ_t). I_t is the indicator whether a representative market maker wins the RFQ. M_t is the population distribution at time t. μ_t are the average ask and bid quotes computed using δ_t and M_t. d_t is the side of the RFQ, either an ASK or BID. After storing the data tuple into replay buffer, a stochastic gradient descent (SGD) step is conducted using mini-batch samples from the replay buffer.

The stochastic gradient descent is exercised on loss functions of critic and actor network parameters θ^Q and θ^π, respectively. The loss functions for critic

and actor networks are defined as:

$$\mathcal{L}^{Q}_{MFG}(\theta^Q) = \mathop{\mathbb{E}}_{q,\delta,q',I,\mu,M}\Bigg[\Bigg(r(q,\delta) + \gamma\Bigg(I \cdot \tilde{Q}\Big(q', (\tilde{\pi}_a(q'), \tilde{\pi}_b(q'), M)\,\Big|\tilde{\theta}^Q\Big) + (1-I)\tilde{Q}\Big(q, (\tilde{\pi}_a(q), \tilde{\pi}_b(q), M)\,\Big|\tilde{\theta}^Q\Big)\Bigg) - Q\left(q, (\tilde{\pi}_a(q), \tilde{\pi}_b(q)), M|\theta^Q\right)\Bigg)^2\Bigg] \tag{4.2.13}$$

$$\mathcal{L}(\theta^\pi) = -\mathop{\mathbb{E}}_{q,M}\Bigg[Q\left(q, (\pi_a(q|\theta^\pi), \pi_b(q|\theta^\pi))\big|\theta^Q\right)\Bigg] \tag{4.2.14}$$

The derivation of the SGD step is similar to (3.5.18) and (3.5.22) in Chapter 3, which is omitted in this section.

Prior to the reinforcement learning process, we fit the critic and actor networks by monopolistic value function and quoting strategy. An exploration rate of $p_0 e^{-\eta t}$ is also applied in the iteration step t. For exploration we add Gaussian noise to quotes from actor network. The SGD step for loss functions (4.2.13) and (4.2.14) defines the learning phase of mean field Deep Deterministic Policy Gradient under given population distribution M. The mean field learning approach incorporates the update of the population distribution using the trained target actor network $\tilde{\pi}(q|\theta^\pi)$ after each training episode. Concretely, at the population distribution update phase of episode e, denote the current population distribution as M_e. The representative market maker interacts with the market environment using $\tilde{\pi}(q|\theta^\pi)$ in given time steps to collect the inventory states of the representative market maker. The population inventory distribution is subsequently updated using a weighted sum of M_e and empirical distribution of the representative market maker's inventory. This reflects the main difference of the learning framework with the rational expectation assumption. We are not assuming competitors' quoting strategies or derive the exact population distribution based on the current quoting strategies. Instead, we directly take the representative market maker's empirical distribution to update next episode's population distribution, while in the theoretical framework, these 2 coincides under mean field Nash equilibrium.

Therefore, the input M to the critic network $Q(q, (\delta^a, \delta^b), M|\theta^Q)$ reflects actually the representative market maker's estimation on the population distribution rather than the exact distribution. For tractability, we stick to this simplification so that the learning problem always focuses on the critic and

actor networks of the representative market maker. We summarize the simulation approach in Algorithm 5.

Algorithm 5 Mean Field Deep Deterministic Policy Gradient

Input: Initial population state distribution M_0, E = number of episodes, T = number of iteration steps in each episode, B = size of mini-batch. N = number of market makers, f_a, f_b, λ^a, λ^b: intensity of ask and bid order flow, ψ: inventory cost function, ϕ: terminal cost function

Output: The target actor networks $(\tilde{\pi}_a, \tilde{\pi}_b)$ of the representative market maker.

Initializing the critic and actor neural networks:
Pre-train critic network Q and actor networks π_a, π_b to value function and quoting strategy of a single monopolistic market maker with execution probability $\Lambda(\delta)$.
Set target networks equal to the original networks: $\tilde{Q} = Q$, $\pi_a = \tilde{\pi}_a$, $\pi_b = \tilde{\pi}_b$.
for Episode $\leftarrow 1$ to E **do**
    Initialize the inventory q_0.
    for $t \leftarrow 0$ to $T-1$ **do**
        Ask and bid RFQ generated by market environment with probability $\frac{\lambda_a}{\lambda_a+\lambda_b}$ and $\frac{\lambda_b}{\lambda_a+\lambda_b}$.
        Obtain quoting strategy using target actor networks: $\delta_t^a = \tilde{\pi}_a(q_t), \delta_t^b = \tilde{\pi}_b(q_t)$. Compute μ_t^a, μ_t^b using the state distribution M_t and the quoting strategy given by the target actor network $\tilde{\pi}_a, \tilde{\pi}_b$. With probability $p_0 \cdot e^{-\eta t}$ a Gaussian noise is added to explore the quoting strategy.
        With probability $f_a(\delta_t^a, \mu_t^a), f_b(\delta_t^b, \mu_t^b)$, the representative market maker wins the ask and bid RFQ, I_t denotes the indicator whether the representative market maker wins the RFQ.
        Set next state q_{t+1}.
        Data tuple $(q_t, \delta_t, q_{t+1}, I_t, \mu_a^t, \mu_b^t, M_t)$ is stored into market maker i's replay buffer.
        Carry out mini-batch TD learning for critic and Stochastic Gradient Descent for actor network.
        Update target network parameters.
    end for
    Update state distribution M_{t+1} by sampling using the target networks $(\tilde{\pi}_a, \tilde{\pi}_b)$.
end for

4.3 Numerical Experiments: Heterogeneity, Learning and Non-competitive outcomes

The mean field DRL algorithm defined in Section 4.2 applies to a homogeneous population in which one *representative* agent, whose strategy is parameterized as a multilayer neural network, learns by updating their parameters iteratively. However, using a slight modification of Algorithm 5, one can also investigate the impact of heterogeneity. For example, one can study the situation where a new dealer is introduced into a market in which the other dealers are distributed according to the mean field Nash equilibrium. To model this situation, we fix the population distribution M_t in the equilibrium distribution computed using Algorithm 5. Subsequently, the learning agent is trained on the samples generated according to the equilibrium distribution M_t. Another method is to simulate the learning behavior of the new agent via reinforcement learning while interacting with an environment composed of $N-1$ (other) agents following the mean field equilibrium quoting strategies.

We use (4.2.1) as the underlying intensity function for simulation, with $k_a = k_b = 2, k_{am} = k_{bm} = 3, C_a = C_b = 1, C_{am} = C_{bm} = 1$. The ask and bid orders are point processes with intensity $\lambda_a = \lambda_b = 5$. The interest rate to discount future profits and losses is set as $r = 0.01$. Inventory limit $H = 10$. Both the critic and the network take the form of a fully connected neural network. We apply a one-hot encoding at the inventory level q since it takes values in a discrete set $\mathcal{Q}$. Note that the critic network $Q(q, (\delta^a, \delta^b), M|\theta^Q)$ takes the estimated population distribution M as input, while the actor networks $\pi_a(q|\theta^\pi), \pi_b(q|\theta^\pi)$ only have the inventory level q as input.

In mean field game learning, exploitability $\epsilon(\boldsymbol{\delta})$ is a measure used to quantify how much a given strategy $\boldsymbol{\delta}$ deviates from the equilibrium strategy. ([Perrin et al. 2020]) It can be formally defined as

$$\epsilon(\boldsymbol{\delta}) = J_0(\boldsymbol{\delta}^*; \boldsymbol{\delta}^*, \boldsymbol{M}^*) - J_0(\boldsymbol{\delta}; \boldsymbol{\delta}^*, \boldsymbol{M}^*) \tag{4.3.1}$$

Tracking the exploitability $\epsilon(\boldsymbol{\delta})$ during the training steps can usually reveal the convergence property of the learning algorithms. However, computing the exploitability requires calculating the value function $J_0(\boldsymbol{\delta}; \boldsymbol{\delta}^*, \boldsymbol{M}^*)$ at each iteration, which becomes computationally intractable given the large number of iterations. Since we are primarily concerned about the market dynamics from learning reflected principally by quoted prices, especially the possibility of 'tacit collusion', we can directly track the difference between learned quoting

strategy and equilibrium strategy based on a simplified distance metric.

We denote $d_k(\delta^k, \delta^{k,*}), k \in \{a, b\}$ the distance between the learned quoting strategy $\boldsymbol{\delta} = (\delta^a, \delta^b)$ and the mean field Nash equilibrium $\boldsymbol{\delta}^* = (\delta^{a,*}, \delta^{b,*})$:

$$\begin{aligned} d_a(\delta^a, \delta^{a,*}) &= \frac{1}{|\mathcal{Q}| - 1} \sum_{q \in \mathcal{Q} \setminus \{-H\}} \big(\delta^a(q) - \delta^{a,*}(q)\big) \\ d_b(\delta^b, \delta^{b,*}) &= \frac{1}{|\mathcal{Q}| - 1} \sum_{q \in \mathcal{Q} \setminus \{H\}} \big(\delta^b(q) - \delta^{b,*}(q)\big) \end{aligned} \tag{4.3.2}$$

$d_k(\delta^k, \delta^{k,*})$ measures the aggregate surplus at all inventory levels from learned quoting strategy compared to the equilibrium level. This metric can be roughly regarded as quantification of 'tacit collusion' level of a market maker: a positive $d_k(\delta^k, \delta^{k,*})$ means the learned quoting strategy δ^k quotes higher prices at most inventory levels, more likely to generate excessive return. We track this distance during training steps to study whether the learning algorithm leads to 'tacit collusion' phenomenon.

4.3.1 Learning in a Homogeneous Population of Dealers

Direct application of Algorithm 5 leads to the homogeneous learning scenario where the rational expectation assumption is not predominantly introduced. We train the critic and actor networks with $E = 1000$ episodes and $T = 1000$ iterations per episode in each experiment and run 100 independent experiments. The reward curves and critic/actor losses from 100 independent simulations are presented in Figure 4.5. The average cumulative reward per episode presents a steep increasing trend in the first 100 episode, then the increasing trend slows down as we train more episodes. The representative market maker in fact learns to adjust the ask and bid quotes to a more profitable direction, numerically demonstrating the convergence of the learning algorithm. The average loss of actor and critic networks per episode is shown in the right graph of Figure 4.5, further demonstrating the convergence of the learning algorithm. Both losses show a downward trend. Specifically, the critic loss oscillates in the first 200 episodes, which is likely due to exploration, then stays at a close level to 0.

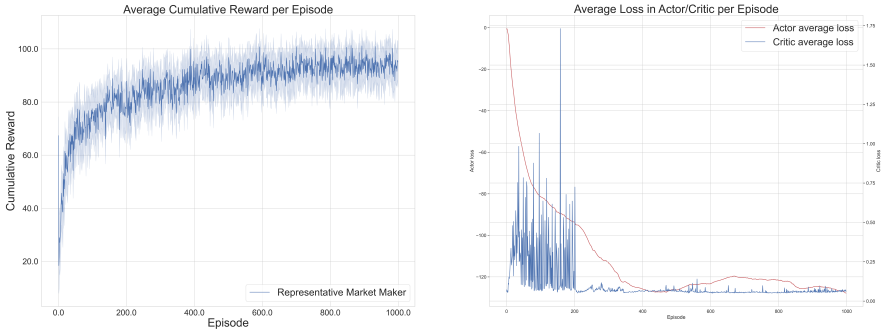


Figure 4.5: Mean field DRL learning results in homogeneous learning scenario. Left: cumulative reward per episode during training, with 95% confidence interval from 100 independent simulations. Right: averaged losses in actor and critic networks showing convergence of algorithm.

Figure 4.5 shows that the learning algorithm converges. But does the learned strategy converge to the mean field Nash equilibrium strategy? After 100 independent experiments, each with 1000 episodes, we plot the output from the actor networks as the learned quoting strategy of the representative market maker. We plot the ask and bid quotes given by actor networks at each inventory level in Figure 4.6, compared to 2 benchmark cases of mean field Nash equilibrium and monopolistic quotes. At each inventory level a 95% confidence level accompanies the average ask and bid quotes. This result reveals several interesting aspects about the learning dynamics from mean field game of market making.

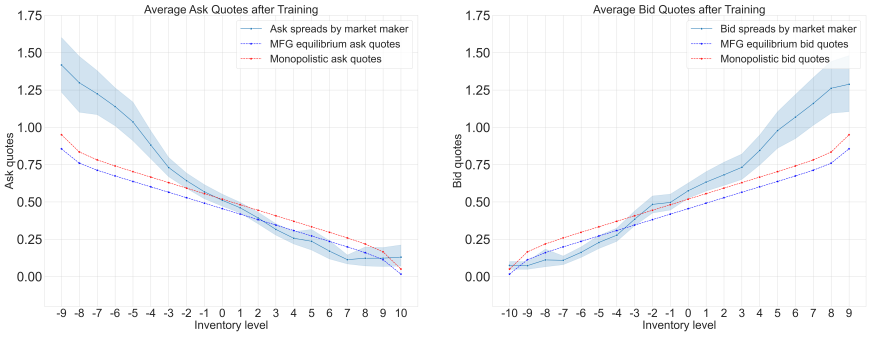


Figure 4.6: Average ask and bid quoting strategies learned by mean field DDPG algorithm from 100 simulations, with 95% confidence interval. The benchmarks are Nash equilibrium (in blue) and monopolistic quoting strategy (in red).

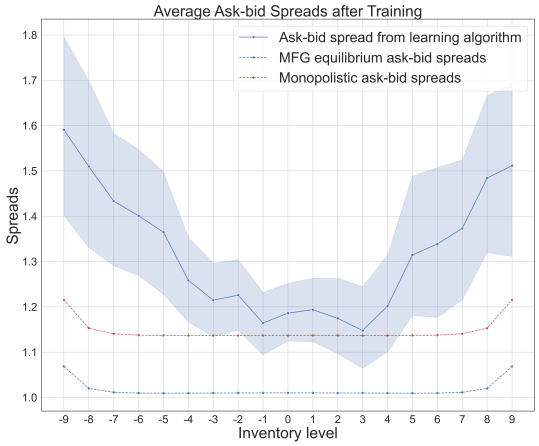


Figure 4.7: Average ask and bid spreads learned by mean field DDPG algorithm, compared to benchmarks.

Figures 4.6 and 4.7 reveal insights into homogeneous learning dynamics. The quotes from 100 independent experiments exhibit a robust trend as functions of inventory levels. The ask quotes are raised when inventory turns to negative position, while bid quotes are increasing when inventory accumulates to the positive side. The algorithm learns to skew the quoted prices as function of inventory. This corresponds to the theoretical result and practical observation that ask and bid quotes are skewed according to inventory changes, since market makers are subject to inventory risk due to market price movement of

their accumulated position. This spread-skewing behavior is not intrinsically pre-set in the algorithm but learned by the agent through interactions with the market environment.

Another important outcome of learning is the emergence of supra-competitive quoting behavior in presence of inventory risk. The learned quoting strategies tend to be more conservative than the mean field equilibrium quotes. When inventory level decreases from zero to negative inventory limit, the ask quotes are increasingly above the corresponding benchmarks including the equilibrium and monopolistic quotes. A similar trend is observed on the bid quotes with higher bid quotes above the benchmarks when inventory increases from zero to a positive inventory limit. In contrast, the ask (bid) quotes stay close or below the mean field Nash equilibrium levels when the inventory level is positive (resp. negative). This together results in learned quoting spreads above equilibrium levels, as shown in Figure 4.7. The result suggests that the supra-competitive quoted spreads applied by the population exhibit a phenomenon of 'tacit collusion' as we observe the prices above competitive level. Overall, the learned quoting strategy is more inventory risk averse in that it tends to quote higher when inventory approaches risk limit to compensate more for its exposure to inventory risk.

The dynamics of the quoting strategy during training is represented by the distance to equilibrium metric (4.3.2). Figure 4.8 shows that at the evolution of the average distance between learned quoting strategy and mean field Nash equilibrium strategy as function of episodes, which stabilizes around 20% after 1000 training episodes. The average levels of the learned quoting strategy are robustly above 20% more than those of the equilibrium quoting strategies. This shows that homogeneous learning leads to agents maintaining their quoting strategy at supra-competitive levels.

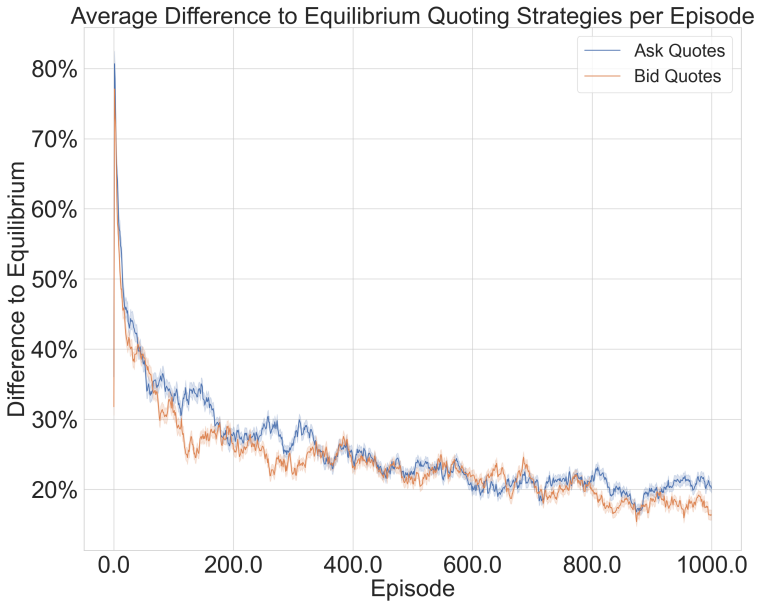


Figure 4.8: Distance of learned quoting strategy to equilibrium quoting strategy in homogeneous learning scenario.

Figure 4.9 compares the distribution of inventory sizes at (Nash) equilibrium with the one obtained under learning dynamics. Both distributions are centered and symmetric around zero, indicating a preference for a balanced inventory due to the positive inventory cost. This shows the effectiveness of reinforcement learning in learning to balance inventory. However, there is a visible discrepancy between the two distributions: the inventory distribution under learning dynamics is more heavy-tailed than under Nash equilibrium. There is a connection between this heavy-tailed inventory distribution and learned supra-competitive quoting strategies. We find that both empirical distribution generated by the trained agent and that during training have a heavy-tailed feature. The agent encounters large inventory levels more frequently than under equilibrium, hence learns to widen the quotes to compensate for higher inventory cost at such levels. Consequently, the actor networks produce higher quotes above equilibrium levels. Therefore, a homogeneous population of dealers learns to jointly increase bid-ask spreads, resulting in a phenomenon of 'algorithmic collusion' or tacit collusion. We regard the difference between the learned bid-ask spread and the MFG equilibrium spread as a metric to measure the extent of tacit collusion.

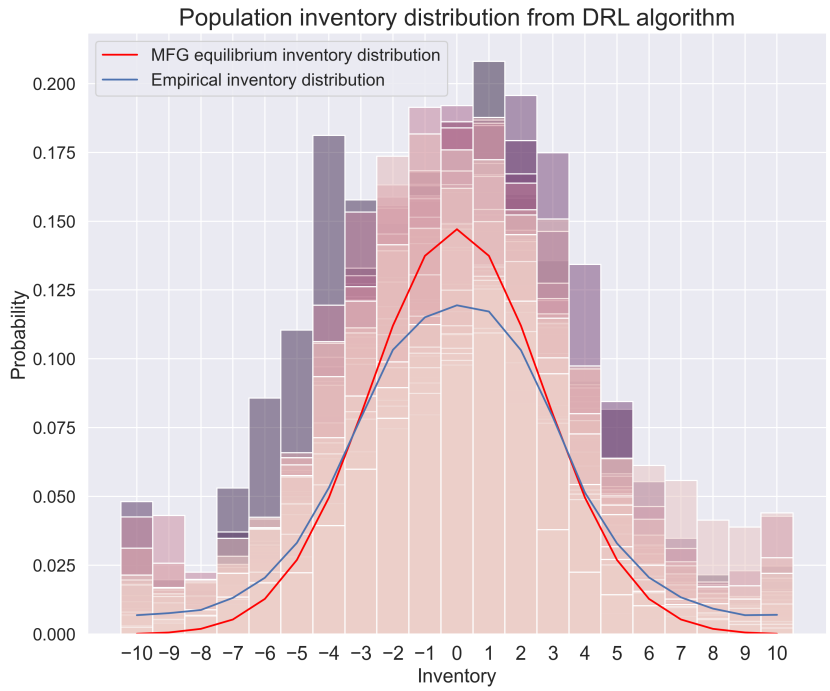


Figure 4.9: Homogeneous learning: comparison between MFG equilibrium population distribution and empirical distribution from DRL algorithm. The learned quoting strategy generates an empirical distribution with heavier tail in connection with higher quotes at extreme inventory levels.

4.3.2 The Impact of Heterogeneity

We now examine the impact on equilibrium by modifying Algorithm 5. In Section 4.3.1 the critic network uses a population distribution estimated empirically by the learning agent, which indicates the absence of the rational expectation assumption. Here, we study the learning dynamics in a mean field that is already in equilibrium (Definition 4.1.6), introducing a heterogeneous learning agent interacting with the homogeneous dealers following equilibrium quoting strategies. Heterogeneity is introduced through the following two independent model settings.

(A) In Algorithm 5, we fix the population distribution M_t^* as the mean field Nash equilibrium. Meanwhile, at each stochastic gradient descent step, mini-batches are sampled according to the MFNE distribution from the experience replay buffer.

(B) Apply an $N-$player learning algorithm similar to Algorithm 3, with one learning agent competing with $N-1$ market makers who use the quoting strategy δ_t^* associated with the MFNE.

We can see that in both approaches, the learning agent is placed within an MFNE equilibrium, hence rational expectation assumption is implicitly imposed in the environment.

Approach (A) We start with the approach (A), which we call 'adjusted sampling' approach. 100 independent experiments are conducted. After training, we compare bid-ask spreads between the learned quoting strategy and the benchmark quoting strategies. The results are summarized in Figure 4.10. With 'adjusted sampling' approach according to the MFNE distribution, the representative market maker has learned to keep their bid-ask spread closer to the mean field Nash equilibrium level, with significant difference from simulation results in Section 4.3.1.

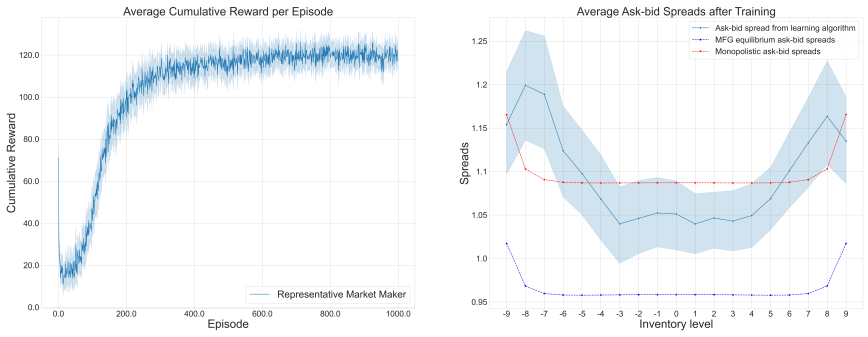


Figure 4.10: Learning outcomes: 'adjusted sampling' simulation (Approach (A)). Left: cumulative reward per episode. Right: comparison of bid-ask spread between learned quoting strategy and benchmark quoting strategies

This 'adjusted sampling' simulation is equivalent to training a representative market maker in an environment where the population inventory distribution is already at mean field equilibrium. Note that in Section 4.3.1, the mini-batches used for stochastic gradient descent are sampled uniformly from the replay buffer, which can lead the learning agent to encounter extreme inventory levels more frequently than under equilibrium distribution, thus resulting in homogeneous learning agent being more significantly risk averse. The 'adjusted sampling' in the approach (A) is able to mitigate market maker's risk

averse behavior and lead to bid-ask spread close to equilibrium spread. The comparison between Section 4.3.1 and the approach (A) suggests that when there is a large population of market makers applying homogeneous learning strategies, the learning population simultaneously raises the quoted spreads, leading to a phenomenon similar to 'tacit collusion. But in an MFNE equilibrium system, a deviating market maker applying the learning algorithm does not necessarily destabilize the system and retain at a closer level to equilibrium spreads.

Approach (B) Compared to one representative agent's learning simulation, we move on to the approach (B). We run a simplified version of N-player simulation that consists of one 'learning market maker' and numerous 'background market makers', based on Remark 4.1.1. The 'learning market maker' refers to the representative market maker applying learning algorithm, while the 'background market makers' fix mean field equilibrium strategy as their quoting strategy. In our experiment, we set $N = 100$, with 1 market maker applying the mean field DDPG Algorithm 5 against 99 market makers fixed at the mean field Nash equilibrium quoting strategy. All other numerical configurations are invariant with homogeneous learning simulation. We run 1000 episodes with 1000 iteration per episode. Figure 4.11 shows the learning curve of the representative market maker, together with the quoted bid-ask spread from the learned strategy. It can be seen that the learned spreads stay closer to the mean field Nash equilibrium spreads compared to Figure 4.7. In this case, the supra-competitive quoting pattern found in Section 4.3.1 has been mitigated when heterogeneity is introduced.

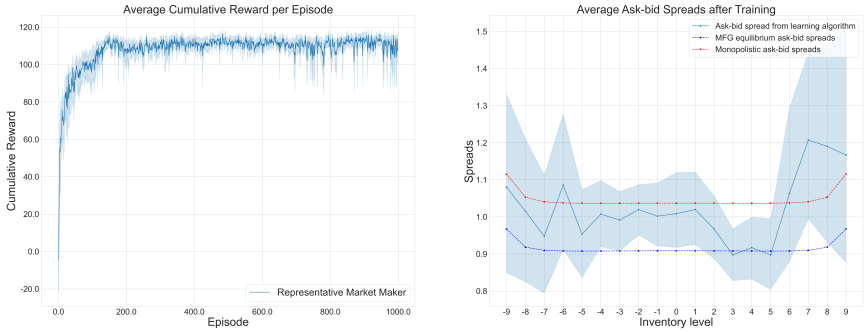


Figure 4.11: Learning results of heterogeneous N-player simulation (Approach (B)) Left: cumulative reward per episode. Right: comparison of bid-ask spread between learned quoting strategy and benchmark quoting strategies

Figure 4.12 demonstrates a similar pattern of a representative market maker's learned quoting strategy as homogeneous learning. However, compared to Figure 4.6, the learned ask and bid quotes are closer to equilibrium benchmark levels. We quantify this difference in Figure 4.13 by pointing out that the distance between the learned quoting strategy and equilibrium is systematically lower compared to homogeneous learning. Hence we see that heterogeneity of agent faced with mean field Nash equilibrium underlies as an essential part in this MFG learning model.

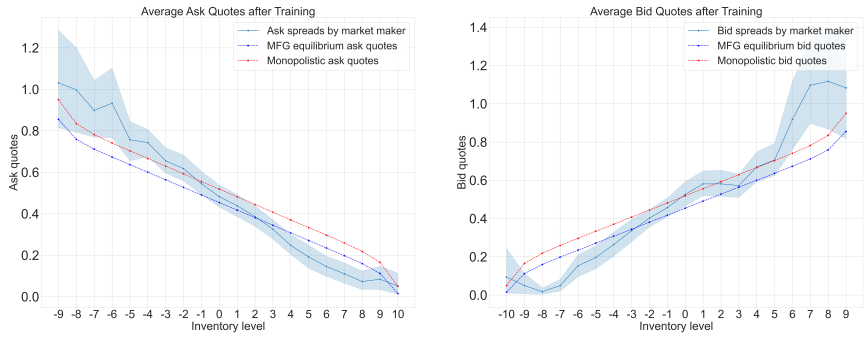


Figure 4.12: Learned ask and bid quotes of simple heterogeneous N-player simulation. The learning market maker exhibits a risk-neutral pattern when training in a market already in equilibrium.

The distance to equilibrium metrics as a function of episodes is plotted in Figure 4.13. During learning steps, the distance decreases to levels below 10%

rather quickly after the first 200 episodes and towards level 0. This means that the extent of the supra-competitive quoting pattern discovered in Section 4.3.1 has been mitigated when heterogeneity is considered in the learning environment. Note that in Section 4.3.1, homogeneity plays the role when the representative market maker learns simultaneously as the other competitors.

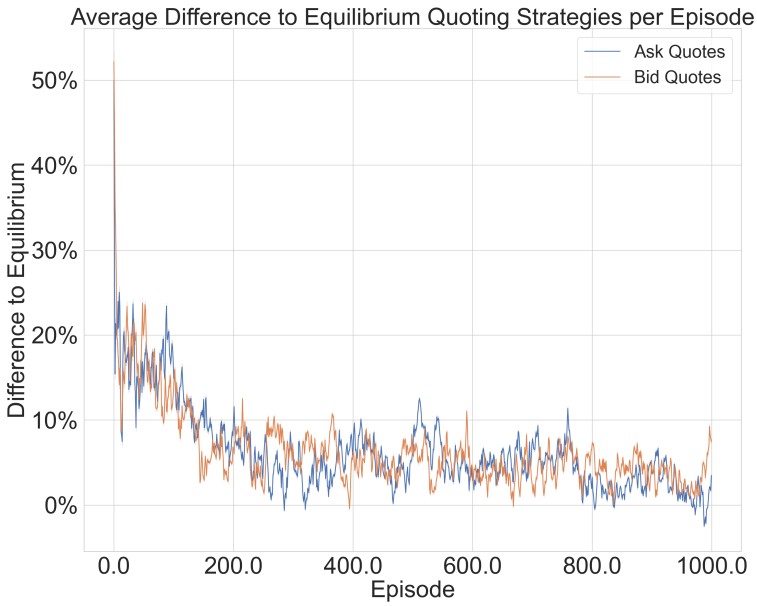


Figure 4.13: Distance of learned quoting strategy to equilibrium quoting strategy in N-player heterogeneous learning scenario.

The empirical population distribution under N-player MFG heterogeneous learning is shown in Figure 4.14. Compared to Figure 4.9, the population distribution in heterogeneous learning is more centered on the inventory 0 with a thinner tail. Overall, the representative market maker encounters less extreme inventory levels, which distinctively explains a learned quoting strategy closer to equilibrium benchmark.

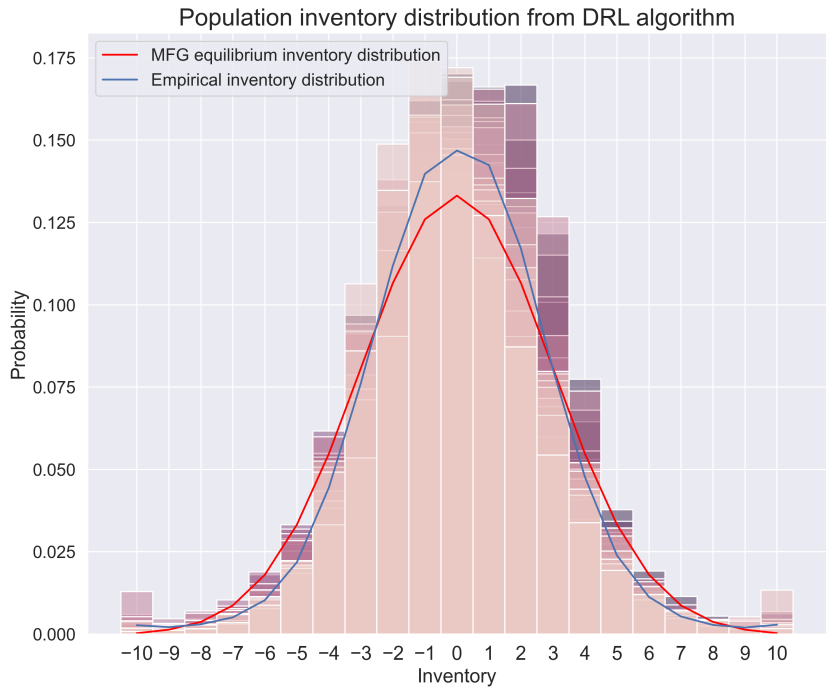


Figure 4.14: Heterogeneous learning: comparison between MFG equilibrium population distribution and empirical distribution from DRL algorithm.

4.3.3 Summary

In summary, our detailed numerical experiments indicate several interesting phenomena that may arise from learning and competition among dealers:

1. In a homogeneous population of dealers who learn by interacting with order flow, as modelled by the mean field reinforcement learning applied to a representative market maker, we observe the emergence of 'tacit collusion', exemplified by supra-competitive bid and ask quote levels.

2. Introducing heterogeneity mitigates the 'tacit collusion' phenomenon observed in the homogeneous case: introducing a new learning agent interacting with other dealers at (mean field) equilibrium leads to lower spread levels.

4.4 Conclusion

We have provided a theoretical framework as well as a simulation-based tool for exploring the dynamics and interactions of autonomous market making

algorithms in electronic OTC markets, and shown how the incorporation of learning dynamics may lead to outcomes different from a competitive Nash equilibrium.

Our methodology is to model strategic interactions among market makers via a finite-state mean field game. The competitive market case is then represented by a mean field Nash equilibrium. We prove the existence and provide a sufficient condition for the uniqueness of such an equilibrium, characterized as a solution of a system of backward Hamilton-Jacobi equations coupled with a forward Chapman-Kolmogorov equation.

We have compared this competitive equilibrium with the outcome of learning dynamics using a *mean field deep reinforcement learning* algorithm. This comparison points to the possibility of supra-competitive quoting strategies emerging as a byproduct of algorithmic interactions and learning dynamics, leading to a situation of *tacit collusion*. The emergence of tacit collusion is associated with heavy-tailed inventory distributions. We also observe that the introduction of heterogeneity can mitigate the emergence of tacit collusion.

This study advances our understanding of algorithm-driven market dynamics, but also points to new research questions on the impact of automated market making strategies and machine learning on market dynamics.

Bibliography

[1] Ibrahim Abada and Xavier Lambin. "Artificial Intelligence: Can Seemingly Collusive Outcomes Be Avoided?" *Management Science* 69 (9 Sept. 2023), 5042–5065.

[2] Yves Achdou, Fabio Camilli, and Italo Capuzzo-Dolcetta. "Mean Field Games: Numerical Methods for the Planning Problem". *SIAM Journal on Control and Optimization* 50 (1 Jan. 2012), 77–109.

[3] Yves Achdou and Italo Capuzzo-Dolcetta. "Mean Field Games: Numerical Methods". *SIAM Journal on Numerical Analysis* 48 (3 Jan. 2010), 1136–1162.

[4] Yves Achdou and Ziad Kobeissi. "Mean field games of controls: Finite difference approximations". *Mathematics In Engineering* 3 (3 2021), 1–32.

[5] Yves Achdou and Mathieu Laurière. "Mean Field Games and Applications: Numerical Aspects". *Mean Field Games. Lecture Notes in Mathematics* (Mar. 2020).

[6] Noha Almulla, Rita Ferreira, and Diogo Gomes. "Two Numerical Approaches to Stationary Mean-Field Games". *Dynamic Games and Applications* 7 (4 Dec. 2017), 657–682.

[7] Leo Ardon, Nelson Vadori, Thomas Spooner, Mengda Xu, Jared Vann, and Sumitra Ganesh. "Towards a Fully RL-based Market Simulator". *ICAIF 2021 - 2nd ACM International Conference on AI in Finance*. Vol. 1. Association for Computing Machinery, 2021.

[8] John Asker, Chaim Fershtman, and Ariel Pakes. "Artificial Intelligence, Algorithm Design, and Pricing". *AEA Papers and Proceedings* 112 (May 2022), 452–56.

[9] Stephanie Assad, Emilio Calvano, Giacomo Calzolari, Robert Clark, Vincenzo Denicolò, Daniel Ershov, Justin Johnson, Sergio Pastorello, Andrew Rhodes, Lei Xu, and Matthijs Wildenbeest. "Autonomous algorithmic collusion: Economic research and policy implications". *Oxford Review of Economic Policy* 37.3 (2021), 459–478.

[10] Hanna Assayag, Alexander Barzykin, Rama Cont, and Wei Xiong. "Competition and learning in dealer markets". *SSRN Electronic Journal* (2024), 1–42.

[11] Marco Avellaneda and Sasha Stoikov. "High-frequency trading in a limit order book". *Quantitative Finance* 8.3 (2008), 217–224.

[12] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. "Layer Normalization". *arXiv* (2016).

[13] Bastien Baldacci, Philippe Bergault, and Dylan Possamaï. "A Mean-Field Game of Market-Making against Strategic Traders". *SIAM Journal on Financial Mathematics* 14.4 (2023), 1080–1112.

[14] Alexander Barzykin, Philippe Bergault, and Olivier Guéant. "Algorithmic market making in dealer markets with hedging and market impact". *Mathematical Finance* 33.1 (Dec. 2023), 41–79.

[15] Alexander Barzykin, Philippe Bergault, and Olivier Guéant. "Algorithmic Market Making in Spot Precious Metals" (Apr. 2024).

[16] Alexander Barzykin, Philippe Bergault, and Olivier Guéant. "Market making by an FX dealer: tiers, pricing ladders and hedging rates for optimal risk control" (Dec. 2021).

[17] Alain Bensoussan, Jens Frehse, and Phillip Yam. *Mean Field Games and Mean Field Type Control Theory*. New York, NY: Springer New York, 2013.

[18] Philippe Bergault, David Evangelista, Olivier Guéant, and Douglas Vieira. "Closed-form Approximations in Multi-asset Market Making". *Applied Mathematical Finance* 28.2 (2021), 101–142.

[19] Philippe Bergault and Olivier Guéant. "Size matters for OTC market makers: General results and dimensionality reduction techniques". *Mathematical Finance* 31.1 (2021), 279–322.

[20] Claude Berge. *Topological Spaces*. Oliver & Boyd, 1963.

[21] M. Bernasconi-de-Luca, E. Vittori, F. Trovò, and M. Restelli. "Dealer markets: A reinforcement learning mean field game approach". *North American Journal of Economics and Finance* 68.August (2023), 101974.

[22] Dimitri P. Bertsekas and Steven E. Shreve. *Stochastic optimal control : the discrete time case*. New York: Academic Press, 1978.

[23] P. Bremaud. *Point Processes and Queues: Martingale Dynamics*. Advances in Physical Geochemistry. Springer, 1981.

[24] Ariela Briani and Pierre Cardaliaguet. "Stable solutions in potential mean field game systems". *Nonlinear Differential Equations and Applications* 25.1 (2018), 1–22.

[25] George W. Brown. "Iterative Solution of Games by Fictitious Play". *Activity Analysis of Production and Allocation*. Ed. by T. C. Koopmans. New York: Wiley, 1951.

[26] George W. Brown. "Some Notes on Computation of Games Solutions". *RAND Corporation* (1949), RM-125–PR.

[27] Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. "Artificial Intelligence, Algorithmic Pricing, and Collusion". *American Economic Review* 110.10 (Oct. 2020), 3267–97.

[28] Pierre Cardaliaguet, Francois Delarue, Jean Michel Lasry, and Pierre Louis Lions. "The master equation and the convergence problem in mean field games". *Annals of Mathematics Studies* 2019-Janua.201 (2019), 1–222.

[29] Pierre Cardaliaguet and Saeed Hadikhanloo. "Learning in mean field games: The fictitious play". *ESAIM - Control, Optimisation and Calculus of Variations* 23.2 (2017), 569–591.

[30] Pierre Cardaliaguet and Charles Albert Lehalle. "Mean field game of controls and an application to trade crowding". *Mathematics and Financial Economics* 12.3 (2018), 335–363.

[31] René Carmona and François Delarue. "Probabilistic analysis of mean-field games". *SIAM Journal on Control and Optimization* 51.4 (2013), 2705–2734.

[32] René Carmona, François Delarue, and Daniel Lacker. "Mean Field Games with Common Noise". *Annals of Probability* 48.5 (2020), 2644–2646.

[33] René Carmona and Mathieu Laurière. "Convergence Analysis of Machine Learning Algorithms for the Numerical Solution of Mean Field Control and Games I: The Ergodic Case". *SIAM Journal on Numerical Analysis* 59 (3 Jan. 2021), 1455–1485.

[34] Álvaro Cartea, Patrick Chang, Mateusz Mroczka, and Roel Oomen. "AI-driven liquidity provision in OTC financial markets". *Quantitative Finance* (Oct. 2022), 1–34.

[35] Álvaro Cartea, Patrick Chang, and José Penalva. "Algorithmic Collusion in Electronic Markets: The Impact of Tick Size". *SSRN Electronic Journal* (2022).

[36] Álvaro Cartea, Sebastian Jaimungal, and Jason Ricci. "Buy Low, Sell High: A High Frequency Trading Perspective". *SIAM Journal on Financial Mathematics* 5.1 (Jan. 2014), 415–444.

[37] Philippe Casgrain and Sebastian Jaimungal. "Mean-field games with differing beliefs for algorithmic trading". *Mathematical Finance* 30.3 (2020), 995–1034.

[38] William G. Christie and Paul H. Schultz. "Why do NASDAQ Market Makers Avoid Odd-Eighth Quotes". *The Journal of Finance* 49.5 (1994), 1813–1840.

[39] Jean-Edouard Colliard, Thierry Foucault, and Stefano Lovo. "Algorithmic Pricing and Liquidity in Securities Markets". *SSRN Electronic Journal* 7 (2022).

[40] Competition & Markets Authority. *Algorithms: How they can reduce competition and harm consumers*. Tech. rep. 2021.

[41] Rama Cont, Xin Guo, and Renyuan Xu. "Interbank lending with benchmark rates: Pareto optima for a class of singular control games". *Mathematical Finance* 31 (4 2021), 1–32.

[42] Rama Cont and Wei Xiong. "Dynamics of market making algorithms in dealer markets: Learning and tacit collusion". *Mathematical Finance* 34.2 (Apr. 2024), 467–521.

[43] Josu Doncel, Nicolas Gast, and Bruno Gaujal. "Discrete mean field games: Existence of equilibria and convergence". *Journal of Dynamics and Games* 6.3 (2019), 221–239.

[44] Winston Wei Dou, Itay Goldstein, and Yan Ji. "AI-Powered Trading, Algorithmic Collusion, and Price Efficiency". *SSRN Electronic Journal* (2023).

[45] Prajit K. Dutta and Ananth Madhavan. "Competition and Collusion in Dealer Markets". *The Journal of Finance* 52.1 (1997), 245.

[46] European Securities and Markets Authority. *Markets in Financial Instruments Directive II (MiFID II)*. 2018.

[47] Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. "Global convergence of policy gradient methods for the linear quadratic regulator". *International Conference on Machine Learning*. ICML. 2018, 1467–1476.

[48] Jean-David Fermanian, Olivier Guéant, and Jiang Pu. "The behavior of dealers and clients on the European corporate bond market: the case of Multi-Dealer-to-Client platforms". *Market microstructure and liquidity* 2.03n04 (2016), 1750004.

[49] Financial Industry Regulatory Authority. *Rule 5310. Best Execution and Interpositioning*. 2014.

[50] Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. "Counterfactual multi-agent policy gradients". *32nd AAAI Conference on Artificial Intelligence, AAAI 2018* (2018), 2974–2982.

[51] Marek Galewski and Marius Rădulescu. "On a global implicit function theorem for locally Lipschitz maps via non-smooth critical point theory". *Quaestiones Mathematicae* 41.4 (2018), 515–528.

[52] Lawrence R. Glosten and Paul R. Milgrom. "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders". *Journal of Financial Economics* 14.1 (1985), 71–100.

[53] Diogo A. Gomes, Joana Mohr, and Rafael Rigão Souza. "Continuous time finite state mean field games". *Applied Mathematics and Optimization* 68 (1 2013), 99–143.

[54] C.W.J. Granger. "Investigating Causal Relations by Econometric Models Published by : The Econometric Society Stable URL : https://www.jstor.org/stable/1912 to Econometrica". *Econometrica* 37.3 (1969), 424–438.

[55] Olivier Guéant. "Existence and Uniqueness Result for Mean Field Games with Congestion Effect on Graphs". *Applied Mathematics and Optimization* 72 (2 2015), 291–303.

[56] Olivier Guéant. "From infinity to one: The reduction of some mean field games to a global control problem". *arXiv* (2011).

[57] Olivier Guéant. "Optimal market making". *Applied Mathematical Finance* 24.2 (2017), 112–154.

[58] Olivier Guéant, Jean-Michel Lasry, and Pierre-Louis Lions. "Mean Field Games and Applications". *Paris-Princeton Lectures on Mathematical Finance 2010. Lecture Notes in Mathematics* 203 (2011), 205–266.

[59] 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 (2013), 477–507.

[60] Olivier Guéant and Iuliia Manziuk. "Deep Reinforcement Learning for Market Making in Corporate Bonds: Beating the Curse of Dimensionality". *Applied Mathematical Finance* 26.5 (2019), 387–452.

[61] Olivier Guéant and Iuliia Manziuk. "Optimal control on graphs: Existence, uniqueness, and long-term behavior". *ESAIM - Control, Optimisation and Calculus of Variations* 26 (2020), 1–14.

[62] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. "A General Framework for Learning Mean-Field Games". *Mathematics of Operations Research* 48.2 (2023), 656–686.

[63] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. "Learning mean-field games". *Advances in Neural Information Processing Systems* 32 (2019), 1–18.

[64] Xin Guo and Renyuan Xu. "Stochastic games for fuel follower problem: N versus mean field game". *SIAM Journal on Control and Optimization* 57.1 (2019), 659–692.

[65] Ben Hambly, Renyuan Xu, and Huining Yang. "Recent advances in reinforcement learning in finance". *Mathematical Finance* 2020-July (February Apr. 2023), 4751–4756.

[66] Bingyan Han. "Understanding algorithmic collusion with experience replay". *arXiv* (2021).

[67] Jiequn Han and Ruimeng Hu. "Deep Fictitious Play for Finding Markovian Nash Equilibrium in Multi-Agent Games". *Proceedings of Machine Learning Research* 107 (2020), 221–245.

[68] Jiequn Han, Ruimeng Hu, and Jihao Long. "Convergence of deep fictitious play for stochastic differential games". *Frontiers of Mathematical Finance* 1 (2 2022), 287.

[69] Matthias Hettich. "Algorithmic Collusion: Insights from Deep Learning". *SSRN Electronic Journal* (2021), 1–19.

[70] Thomas S. Y. Ho and Hans R. Stoll. "Optimal dealer pricing under transactions and return uncertainty". *Journal of Financial Economics* 8.1 (1980), 47–73.

[71] Thomas S. Y. Ho and Hans R. Stoll. "The Dynamics of Dealer Markets Under Competition". *The Journal of Finance* 38.4 (Sept. 1983), 1053.

[72] Junling Hu and Michael P Wellman. "Multiagent reinforcement learning: Theoretical framework and an algorithm". *Proceedings of the fifteenth international conference on machine learning* 242 (1998), 250.

[73] Ruimeng Hu. "Deep fictitious play for stochastic differential games". *Communications in Mathematical Sciences* 19 (2 2021), 325–353.

[74] Minyi Huang, Peter E. Caines, and Roland P. Malhamé. "Large-population cost-coupled LQG problems with nonuniform agents: Individual-mass behavior and decentralized ϵ-nash equilibria". *IEEE Transactions on Automatic Control* 52.9 (2007), 1560–1571.

[75] Xuancheng Huang, Sebastian Jaimungal, and Mojtaba Nourian. "Mean-Field Game Strategies for Optimal Execution". *Applied Mathematical Finance* 26.2 (2019), 153–185.

[76] IMF. *Global Financial Stability Report: Steadying the Course: Uncertainty, Artificial Intelligence, and Financial Stability*. Tech. rep. International Monetary Fund, 2024.

[77] Kazufumi Ito, Christoph Reisinger, and Yufei Zhang. "A Neural Network-Based Policy Iteration Algorithm with Global H2 -Superlinear Convergence for Stochastic Games on Domains". *Foundations of Computational Mathematics* 21.2 (2021), 331–374.

[78] Marc Ivaldi, Bruno Jullien, Patrick Rey, Paul Seabright, and Jean Tirole. "The Economics of Tacit Collusion". *IDEI Working Papers, Institut d'Économie Industrielle (IDEI), Toulouse* 186.March (2003).

[79] Saul D. Jacka and Aleksandar Mijatović. “On the policy improvement algorithm in continuous time”. *Stochastics* 89.1 (2017), 348–359.

[80] Ioannis Karatzas. “A Class of Singular Stochastic Control Problems”. *Advances in Applied Probability* 15.2 (1983), 225–254.

[81] Bekzhan Kerimkulov, David Šiška, and Lukasz Szpruch. “Exponential convergence and stability of Howard’s policy improvement algorithm for controlled diffusions”. *SIAM Journal on Control and Optimization* 58.3 (2020), 1314–1340.

[82] Diederik P. Kingma and Jimmy Lei Ba. “Adam: A method for stochastic optimization”. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2015), 1–15.

[83] Albert S. Kyle. “Continuous Auctions and Insider Trading”. *Econometrica* 53.6 (Nov. 1985), 1315.

[84] Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. “A unified game-theoretic approach to multiagent reinforcement learning”. *Advances in Neural Information Processing Systems*. Vol. 2017-Decem. 2017, 4191–4204.

[85] Jean Michel Lasry and Pierre Louis Lions. “Jeux à champ moyen. II - Horizon fini et contrôle optimal”. *Comptes Rendus Mathematique* 343.10 (2006), 679–684.

[86] Jean Michel Lasry and Pierre Louis Lions. “Mean field games”. *Japanese Journal of Mathematics* 2 (1 2007), 229–260.

[87] Jean Michel Lasry and Pierre Louis Lions. “Mean field games. I - The stationary case”. *Comptes Rendus Mathematique* 343.9 (2006), 619–625.

[88] Mathieu Laurière. “Numerical methods for mean field games and mean field type control”. *arXiv* (2021), 221–282.

[89] Mathieu Laurière, Sarah Perrin, Matthieu Geist, and Olivier Pietquin. “Learning Mean Field Games: A Survey” (2022), 1–50.

[90] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. “Continuous control with deep reinforcement learning.” *ICLR*. Ed. by Yoshua Bengio and Yann LeCun. 2016.

[91] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. "Multi-agent actor-critic for mixed cooperative-competitive environments". *Advances in Neural Information Processing Systems* 2017-Decem (2017), 6380–6391.

[92] Jialiang Luo and Harry Zheng. "Dynamic Equilibrium of Market Making with Price Competition". *Dynamic Games and Applications* 11.3 (2021), 556–579.

[93] J.L. Menaldi and M. I. Taksar. "Optimal Correction of a Multidimensional Stochastic System". *Automatica* 25.2 (1989), 223–232.

[94] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. "Playing Atari with Deep Reinforcement Learning". *NeurIPS Deep Learning Workshop* (2013).

[95] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. "Human-level control through deep reinforcement learning". *Nature* 518.7540 (Feb. 2015), 529–533.

[96] Dov Monderer and Lloyd S Shapley. "Fictitious Play Property for Games with Identical Interests". *Journal of Economic Theory* 68 (1 Jan. 1996), 258–265.

[97] Eyal Neuman and Moritz Voß. "Trading with the crowd". *Mathematical Finance* 33.3 (July 2023), 548–617.

[98] Sarah Perrin, Julien Perolat, Mathieu Laurière, Matthieu Geist, Romuald Elie, and Olivier Pietquin. "Fictitious play for mean field games: Continuous time analysis and applications". *Advances in Neural Information Processing Systems*. Vol. 2020-Decem. 2020, 13199–13213.

[99] Huyn Pham. *Continuous-time Stochastic Control and Optimization with Financial Applications*. 1st. Springer Publishing Company, Incorporated, 2009.

[100] Martin L. Puterman. "On the convergence of policy iteration for controlled diffusions". *Journal of Optimization Theory and Applications* 33.1 (1981), 137–144.

[101] Martin L. Puterman and Shelby L. Brumelle. “On the Convergence of Policy Iteration in Stationary Dynamic Programming”. *Mathematics of Operations Research* 4.1 (1979), 60–69.

[102] Lynn Riggs, Esen Onur, David Reiffen, and Haoxiang Zhu. “Swap trading after Dodd-Frank: Evidence from index CDS”. *Journal of Financial Economics* 137 (3 Sept. 2020), 857–886.

[103] Julia Robinson. “An Iterative Method of Solving a Game”. *Annals of Mathematics* 54 (2 1951), 296–301.

[104] Tim Roughgarden. “Best-Response Dynamics”. *Twenty Lectures on Algorithmic Game Theory*. Cambridge University Press, Aug. 2016, 216–229.

[105] Manuel S. Santos and John Rust. “Convergence properties of policy iteration”. *SIAM Journal on Control and Optimization* 42.6 (2004), 2094–2115.

[106] Securities and Exchange Commission. *Regulation Best Execution*. 2023.

[107] Lloyd S. Shapley. *On the Nonconvergence of Fictitious Play*. RAND Corporation, 1962.

[108] Andrzej Skrzypacz and Hugo Hopenhayn. “Tacit collusion in repeated auctions”. *Journal of Economic Theory* 114.1 (2004), 153–169.

[109] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book, 2018.

[110] Jean Tirole. *The Theory of Industrial Organization*. Cambridge, MA, USA: MIT Press, 1988.

[111] Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. “Gymnasium: A Standard Interface for Reinforcement Learning Environments”. *arXiv* (2024), 1–10.

[112] Hung Vinh Tran, Zhenhua Wang, and Yuming Paul Zhang. “Policy Iteration for Exploratory Hamilton-Jacobi-Bellman Equations”. *arXiv* (2024), 1–21.

[113] 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". *Mathematical Finance* 34.2 (2024), 262–347.

[114] Ludo Waltman and Uzay Kaymak. "Q-learning agents in a Cournot oligopoly model". *Journal of Economic Dynamics and Control* 32.10 (2008), 3275–3293.

[115] Chaojun Wang. "The limits of multi-dealer platforms". *Journal of Financial Economics* 149 (3 Sept. 2023), 434–450.

[116] Christopher J C H Watkins. "Learning from delayed rewards". PhD thesis. King's College, University of Cambridge, 1989.

[117] Wei Xiong and Rama Cont. "Interactions of Market Making Algorithms : a Study on Perceived Collusion". *ICAIF '21: Proceedings of the Second ACM International Conference on AI in Finance*. Association for Computing Machinery, 2021, Article No.: 32, Pages 1–9.

Appendix A

Appendix of Chapter 3

A.1 Proof of Proposition 3.1.5

Proof. For simplicity of notations, we denote the average term in (3.1.8) by

$$\chi = \frac{1}{K_i} \sum_{j \neq i} \sum_{q_j \in \mathcal{Q}_j} (a\delta^j_{q_j} + b^j_{q_j})$$

It is straightforward to verify that the intensity function

$$f^i_a(\delta, \vec{\delta}^{-i}) = \frac{1}{N} \frac{1}{1 + e^{a\delta + b}} \frac{e^{\chi}}{1 + e^{a\delta + \chi}} \tag{A.1.1}$$

satisfies Assumption 3.1.1, since $f^i_a(\delta, \vec{\delta}^{-i}) > 0$, and

$$\sum_{i=1}^{N} f^i_a(\delta, \vec{\delta}^{-i}) \leq N \times \frac{1}{N} = 1$$

Moreover, f^i_a is dominated by the upper bound function

$$\Lambda(\delta) = \frac{1}{1 + e^{a\delta + b}}$$

Therefore, Assumption 3.1.1 is satisfied by (3.1.8).

Then, we calculate the derivatives of f^i_a, and obtain:

$$\begin{aligned} \partial_1 f^i_a &= -\frac{a e^{a\delta + \chi}\big(e^b + e^{\chi} + 2e^{a\delta + b + \chi}\big)}{N\big(1 + e^{a\delta + b}\big)^2\big(1 + e^{a\delta + \chi}\big)^2} < 0 \\ \partial_{j, q_j} f^i_a &= \frac{a e^{\chi}}{N K_i\big(1 + e^{a\delta + b}\big)\big(1 + e^{a\delta + \chi}\big)} > 0 \end{aligned} \tag{A.1.2}$$

Furthermore, we compute the term

$$2(\partial_1 f^i_a)^2 - \partial^2_{11} f^i_a \cdot f^i_a = \frac{a^2 e^{a\delta + 2\chi}\big(e^b + e^{\chi} + 4e^{a\delta + b + \chi}\big)}{N^2\big(1 + e^{a\delta + b}\big)^3\big(1 + e^{a\delta + \chi}\big)^3} > 0 \tag{A.1.3}$$

Therefore, the conditions in the first line of Assumption 3.1.3 are satisfied. Next we compute the term

$$\partial_1 f_a^i \cdot \partial_{j,q_j} f_a^i - f_a^i \cdot \partial_{j,q_j}\partial_1 f_a^i = \frac{a^2 e^{a\delta+3\chi}}{N^2 K_i \big(1+e^{a\delta+b}\big)^2 \big(1+e^{a\delta+\chi}\big)^4} > 0 \tag{A.1.4}$$

Hence, we obtain the following:

$$\begin{aligned}
&2(\partial_1 f_a^i)^2 - \partial_{11}^2 f_a^i \cdot f_a^i - \sum_{j\neq i}\sum_{q_j\in\mathcal{Q}_j} \left|\partial_1 f_a^i \cdot \partial_{j,q_j} f_a^i - f_a^i \cdot \partial_{j,q_j}\partial_1 f_a^i\right| \\
&= 2(\partial_1 f_a^i)^2 - \partial_{11}^2 f_a^i \cdot f_a^i - \sum_{j\neq i}\sum_{q_j\in\mathcal{Q}_j} \left(\partial_1 f_a^i \cdot \partial_{j,q_j} f_a^i - f_a^i \cdot \partial_{j,q_j}\partial_1 f_a^i\right) \\
&= \frac{1}{N^2}\left[\frac{a^2 e^{a\delta+2\chi}\big(e^b+e^\chi+4e^{a\delta+b+\chi}\big)}{\big(1+e^{a\delta+b}\big)^3\big(1+e^{a\delta+\chi}\big)^3} - \frac{a^2 e^{a\delta+3\chi}}{\big(1+e^{a\delta+b}\big)^2\big(1+e^{a\delta+\chi}\big)^4}\right] \\
&= \frac{a^2 e^{a\delta+2\chi}\Big(e^b+4e^{a\delta+b+\chi}+e^{a\delta+2\chi}+4e^{2a\delta+b+2\chi}\Big)}{N^2\big(1+e^{a\delta+b}\big)^3\big(1+e^{a\delta+\chi}\big)^4} > 0
\end{aligned} \tag{A.1.5}$$

Therefore, the second line of Assumption 3.1.3 is satisfied.

Finally we compute:

$$\frac{f_a^i(\delta,\vec{\delta}^{-i})}{\partial_1 f_a^i(\delta,\vec{\delta}^{-i})} = -\frac{e^{-a\delta}\big(1+e^{a\delta+b}\big)\big(1+e^{a\delta+\chi}\big)}{a\big(e^b+e^\chi+2e^{a\delta+b+\chi}\big)} \tag{A.1.6}$$

By comparing the orders in the exponential terms in (A.1.6), we have that for any given $\vec{\delta}^{-i}$,

$$\lim_{\delta\to+\infty} \frac{f_a^i(\delta,\vec{\delta}^{-i})}{\partial_1 f_a^i(\delta,\vec{\delta}^{-i})} = -\frac{1}{2a} < \infty \tag{A.1.7}$$

And the last line of Assumption 3.1.3 is satisfied. Therefore, the intensity function (3.1.8) satisfies Assumption 3.1.1 and Assumption 3.1.3. □

A.2 Proof of Theorem 3.2.6

From Proposition 3.2.4 and the verification theorem 3.2.5, to prove Theorem 3.2.6 it suffices to prove the system of the HJB equation (3.2.15) admits a solution.

The structure of the proof follows the existence of the Nash equilibrium presented in [Luo and Zheng 2021]. We employ similar mathematical tools, including hemicontinuity properties, Berge's Maximum Theorem, the implicit function theorem, and Schauder's fixed point theorem. As noted in Remark

3.2.7, our proof adapts these tools specifically for the multi-agent setting considered in this thesis.

We define $H^i_{q_i}(\delta) := (\delta - p^i_{q_i}) f^i_a\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big)$, and give a lemma on the properties of the maximum point of $H^i_{q_i}$.

Lemma A.2.1. *Suppose that Assumptions 3.1.1 and 3.1.3 are satisfied by intensity functions f^i_a, then for given $p^i_{q_i} \in \mathbb{R}$ and $(\vec{\delta}^{a,j})_{j\neq i} \in \prod_{j\neq i} \mathbb{R}^{2^{\frac{Z_j}{\Delta}+1}}$, function $H^i_{q_i}(\delta) := (\delta - p^i_{q_i}) f^i_a\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big)$ has a unique maximum point δ^* on $\mathbb{R}$. The maximum point δ^* satisfies*

$$\frac{\partial H^i_{q_i}}{\partial \delta}(\delta^*) = 0 \tag{A.2.1}$$

Proof. Using notations in (3.1.6), we have

$$\begin{aligned}(H^i_{q_i})'(\delta) &= (\delta - p^i_{q_i})\partial_1 f^i_a\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big) + f^i_a\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big) \\ &= \partial_1 f^i_a\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big)\left[\delta - p^i_{q_i} + \frac{f^i_a\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big)}{\partial_1 f^i_a\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big)}\right] \end{aligned} \tag{A.2.2}$$

Define

$$h(\delta) = \delta - p^i_{q_i} + \frac{f^i_a\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big)}{\partial_1 f^i_a\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big)}$$

Then

$$h'(\delta) = 2 - \frac{f^i_a\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big)\partial^2_{ii} f^i_a\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big))^2}{\Big(\partial_1 f^i_a\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big)\Big)^2} > 0$$

from Assumption 3.1.3. Hence $h(\delta)$ is an increasing function in δ. Again, from Assumption 3.1.3, $\partial_1 f^i_a < 0$, $(H^i_{q_i})'(\delta)$ is a decreasing function in δ. When $\delta < p^i_{q_i}$, we have $h(\delta) < 0$. Hence, when $\delta < p^i_{q_i}$, we have

$$(H^i_{q_i})'(\delta) = \partial_1 f^i_a\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big)h(\delta) > 0$$

From Assumption 3.1.3, we have

$$\lim_{\delta\to+\infty} \frac{f^i_a(\delta, \vec{\delta}^{-i})}{\partial_1 f^i_a(\delta, \vec{\delta}^{-i})} < \infty$$

Then, $\lim_{\delta\to+\infty} h(\delta) = +\infty$. We obtain that, when δ is large enough,

$$(H^i_{q_i})'(\delta) < 0$$

By the mean value theorem applied on the continuous function $(H^i_{q_i})'(\delta)$, there exists $\delta^* > p^i_{q_i}$ such that $(H^i_{q_i})'(\delta^*) = 0$. When $\delta < \delta^*$, $(H^i_{q_i})'(\delta) > 0$. When $\delta > \delta^*$, $(H^i_{q_i})'(\delta) < 0$. Hence δ^* is the unique maximum point of $H^i_{q_i}$, and $(H^i_{q_i})'(\delta) = 0$. □

We consider a family of mappings

$$\mathcal{T}^a_{\boldsymbol{p}} : \prod_{j=1}^{N} (I_\delta)^{2^{\frac{Z_j}{\Delta}+1}} \to \prod_{j=1}^{N} (I_\delta)^{2^{\frac{Z_j}{\Delta}+1}}$$

indexed by a vector $\boldsymbol{p} = (\vec{p}^1, \dots, \vec{p}^N) \in \prod_{j=1}^{N} \mathbb{R}^{2^{\frac{Z_j}{\Delta}+1}}$, where each $\vec{p}^i = (p^i_q)_{q\in\mathcal{Q}_i} \in \mathbb{R}^{2^{\frac{Z_i}{\Delta}+1}}$ is indexed by $q \in \mathcal{Q}_i = \{-Z_i, -Z_i+\Delta, \dots, Z_i-\Delta, Z_i\}$. $\mathcal{T}^a_{\boldsymbol{p}}$ is such that $\forall \boldsymbol{\delta} = (\vec{\delta}^{a,1}, \dots, \vec{\delta}^{a,N}) \in \prod_{j=1}^{N} (I_\delta)^{2^{\frac{Z_j}{\Delta}+1}}$ where $\vec{\delta}^{a,i} = (\delta^{a,i}_q)_{q\in\mathcal{Q}_i} \in (I_\delta)^{2^{\frac{Z_i}{\Delta}+1}}$ is a vector in $\mathbb{R}^{2^{\frac{Z_i}{\Delta}+1}}$ indexed by $\mathcal{Q}_i$, we have

$$\mathcal{T}^a_{\boldsymbol{p}}(\boldsymbol{\delta}) = \left(\left(\underset{\delta^{a,i}_{q_i} \geq -\delta_\infty}{\arg\max} \left[(\delta^{a,i}_{q_i} - p^i_{q_i}) f^i_a\big(\delta^{a,i}_{q_i}, (\vec{\delta}^{a,j})_{j\neq i}\big) \right] \right)_{q_i\in\mathcal{Q}_i} \right)_{i\in\{1,\dots,N\}} \tag{A.2.3}$$

From Lemma A.2.1, given $\boldsymbol{p}$, $\underset{\delta\in\mathbb{R}}{\arg\max}\, H^i_{q_i}(\delta)$ is unique. Hence, the mapping $\mathcal{T}^a_{\boldsymbol{p}}$ is well defined.

Remark A.2.2. Note that in (A.2.3) arg max is taken in the interval $[-\delta_\infty, \infty)$, but $\mathcal{T}^a_{\boldsymbol{p}}$ is still well defined. Since if the maximum point δ^* of function $H^i_{q_i}(\delta)$ satisfies $\delta^* \leq -\delta_\infty$, then

$$\underset{\delta\geq -\delta_\infty}{\arg\max} (\delta - p^i_{q_i}) f^i_a\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big) = -\delta_\infty$$

We also symmetrically define the mappings $\mathcal{T}^b_{\boldsymbol{p}}$ for the bid quoting strategy side. From now on we shall focus on $\mathcal{T}^a_{\boldsymbol{p}}$, and the results follow immediately for $\mathcal{T}^b_{\boldsymbol{p}}$.

For any given $\boldsymbol{p} \in \prod_{j=1}^{N} \mathbb{R}^{2^{\frac{Z_j}{\Delta}+1}}$, we study the existence of a fixed point for $\mathcal{T}^a_{\boldsymbol{p}}$. If $\mathcal{T}^a_{\boldsymbol{p}}$ has a fixed point, we can then take $\boldsymbol{p} = \left((\frac{V_i(q_i) - V_i(q_i-\Delta)}{\Delta})_{q_i\in\mathcal{Q}_i} \right)_{i\in\{1,\dots,N\}}$ and transfer the system of HJB equations (3.2.15) to a system of non-linear equations where the unknown variables are $\{V_i(q_i), q_i \in \mathcal{Q}_i, i \in \{1, \dots, N\}\}$.

Proposition A.2.3. *For any given* $\boldsymbol{p} \in \prod_{j=1}^{N} \mathbb{R}^{2^{\frac{Z_j}{\Delta}+1}}$, *there exists a nonempty compact set* $K_{\boldsymbol{p}} \subseteq \prod_{j=1}^{N} (I_\delta)^{2^{\frac{Z_j}{\Delta}+1}}$ *such that* $\mathcal{T}^a_{\boldsymbol{p}}(K_{\boldsymbol{p}}) \subseteq K_{\boldsymbol{p}}$.

Proof. We prove that for any $\boldsymbol{\delta}$, $\mathcal{T}^a_{\boldsymbol{p}}(\boldsymbol{\delta})$ is uniformly bounded.

Define

$$p_m = \min_{i\in\{1,\dots,N\}, q_i\in\mathcal{Q}_i} p^i_{q_i}, p_M = \max_{i\in\{1,\dots,N\}, q_i\in\mathcal{Q}_i} p^i_{q_i}$$

For given i and q_i denote the coordinate $q_i \in \mathcal{Q}_i$ of the i^{th} vector in $\mathcal{T}^a_{\boldsymbol{p}}(\boldsymbol{\delta})$ by $g^i_{q_i}$, i.e.

$$g^i_{q_i} := \mathcal{T}^{a,i,q_i}_{\boldsymbol{p}}(\boldsymbol{\delta}) := \arg\max_{\delta\in\mathbb{R}}(\delta - p^i_{q_i}) f^i_a\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big) \tag{A.2.4}$$

From Assumption 3.1.1, $f^i_a\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big) > 0$, so when $\delta > p^i_{q_i}$ we have

$$(\delta - p^i_{q_i}) f^i_a\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big) > 0$$

Therefore, the maximum in (A.2.4) must be attained in the interval $(p^i_{q_i}, \infty)$. We obtain the lower bound p_m

$$g^i_{q_i} > p^i_{q_i} \geq p_m, \forall i \in \{1, \ldots, N\}, q_i \in \mathcal{Q}_i \tag{A.2.5}$$

Define $\delta_m = \max(p_m, -\delta_\infty)$, then we have

$$\arg\max_{\delta\geq -\delta_\infty}(\delta - p^i_{q_i}) f^i_a\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big) \geq \delta_m \tag{A.2.6}$$

On the other hand, for any $\boldsymbol{\delta}$ such that $\delta^{a,j}_{q_j} \geq \delta_m$ where $q_j \in \mathcal{Q}_j, j \in \{1, \ldots, N\}$, we seek an upper bound for $g^i_{q_i}$ where i and q_i are arbitrary. Replace the coordinate $\delta^{a,i}_{q_i}$ inside $\boldsymbol{\delta}$ by $\hat{\delta} \equiv p_M + 1$ and form a new vector $\hat{\boldsymbol{\delta}} \in \prod_{j=1}^{N} \mathbb{R}^{2\frac{Z_j}{\Delta}+1}$. We then consider the value $G^i_{q_i}$ defined by

$$G^i_{q_i} = (\hat{\delta} - p^i_{q_i}) f^i_a\big(\hat{\delta}, (\vec{\delta}^{a,j})_{j\neq i}\big) \tag{A.2.7}$$

From Assumption 3.1.3, f^i_a is increasing function in $\delta^j_{q_j}, \forall q_j \in \mathcal{Q}_j, \forall j \neq i$, and $\delta^j_{q_j} \geq p_m, \forall q_j \in \mathcal{Q}_j, \forall j \neq i$, we have

$$G^i_{q_i} = (\hat{\delta} - p^i_{q_i}) f^i_a\big(\hat{\delta}, (\vec{\delta}^{a,j})_{j\neq i}\big) \geq (\hat{\delta} - p^i_{q_i}) f^i_a\big(\hat{\delta}, (p_m)_{q_j\in\mathcal{Q}_j, j\neq i}\big) \geq f^i_a\big(\hat{\delta}, (p_m)_{q_j\in\mathcal{Q}_j, j\neq i}\big) \tag{A.2.8}$$

From Assumption 3.1.1, the upper bound function Λ of f^i_a satisfies

$$\lim_{\delta\to\infty} \Lambda(\delta)\delta = 0$$

We can also derive $\lim_{\delta\to\infty} \Lambda(\delta) = 0$. There exists $\delta_M > \max(p_M + 1, -\delta_\infty)$ such that

$$f^i_a\big(p_M + 1, (p_m)_{q_j\in\mathcal{Q}_j, j\neq i}\big) > \max_{q_i\in\mathcal{Q}_i} \Lambda(\delta_M)(\delta_M - p^i_{q_i}) > \max_{q_i\in\mathcal{Q}_i} f^i_a\big(\delta_M, (\vec{\delta}^{a,j})_{j\neq i}\big)(\delta_M - p^i_{q_i}) \tag{A.2.9}$$

Combining (A.2.8) and (A.2.9) we obtain

$$(\hat{\delta} - p^i_{q_i}) f^i_a\big(\hat{\delta}, (\vec{\delta}^{a,j})_{j\neq i}\big) > \max_{q_i\in\mathcal{Q}_i} f^i_a\big(\delta_M, (\vec{\delta}^{a,j})_{j\neq i}\big)(\delta_M - p^i_{q_i}) \tag{A.2.10}$$

Hence the maximum point δ^* of function $H^i_{q_i}(\delta) := (\delta - p^i_{q_i}) f^i_a\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big)$ can only be achieved on interval $[p_m, \delta_M]$, since from Lemma A.2.1 $H^i_{q_i}$ is increasing function on $[p_m, \delta^*]$ and decreasing function on $[\delta^*, \infty)$. Therefore we have

$$\delta^* \leq \delta_M$$

Since i and q_i are arbitrary we finally obtain

$$g^i_{q_i} \leq \delta_M, \forall q_i \in \mathcal{Q}_i, i \in \{1, \ldots, N\} \tag{A.2.11}$$

Therefore the compact set $K_{\boldsymbol{p}} = \prod_{j=1}^{N} [\delta_m, \delta_M]^{(2\frac{Z_j}{\Delta}+1)} \subseteq \prod_{j=1}^{N} (I_\delta)^{2\frac{Z_j}{\Delta}+1}$ satisfies

$$\mathcal{T}^a_{\boldsymbol{p}}(K_{\boldsymbol{p}}) \subseteq K_{\boldsymbol{p}}$$

□

Next, we prove that $\mathcal{T}^a_{\boldsymbol{p}}$ is continuous on $K_{\boldsymbol{p}}$. To proceed we first introduce the notions of upper and lower hemicontinuity for set-valued functions (or, by another name, correspondence). We denote a correspondence that maps from A to subsets of B by $\Gamma : A \Rightarrow B$ such that $\forall x \in A$, $\Gamma(x) \subseteq B$.

Definition A.2.4. (Upper hemicontinuity) A correspondence $\Gamma : A \Rightarrow B$ is upper hemicontinuous at $a \in A$, if for any open neighborhood V of $\Gamma(a)$ (i.e. $\Gamma(a) \subseteq V$), there exists a neighborhood U of a, such that for any $x \in U$, $\Gamma(x) \subseteq V$.

Definition A.2.5. (Lower hemicontinuity) A correspondence $\Gamma : A \Rightarrow B$ is lower hemicontinuous at $a \in A$, if for any open set V such that $V \cap \Gamma(a) \neq \emptyset$, there exists a neighborhood U of a, such that for any $x \in U$, $\Gamma(x) \cap V \neq \emptyset$.

We will need below Berge's Maximum Theorem ([Berge 1963]) for the continuity of arg max function.

Lemma A.2.6. *(Berge's Maximum Theorem) A function $f : X \times \Theta \to \mathbb{R}$ is continuous on $X \times \Theta$. A correspondence $D : \Theta \Rightarrow X$ is compact-valued, i.e. $\forall \theta \in \Theta, D(\theta)$ is a compact subset of X. Define the maximum function $f^*(\theta) = \sup\{f(x, \theta), x \in D(\theta)\}$ and $D^* : \Theta \Rightarrow X$ by $D^*(\theta) = \arg\sup\{f(x, \theta), x \in D(\theta)\} = \{x \in D(\theta) : f(x, \theta) = f^*(\theta)\}$. If D is both upper and lower hemicontinuous at θ, then $f^*(\theta)$ is continuous, and $D^*(\theta)$ is upper hemicontinuous with nonempty and compact values.*

For a single-valued mapping, we have following lemma connecting upper hemicontinuity and the continuity of function.

Lemma A.2.7. *Let X, Y be 2 topological spaces and $\Gamma : X \Rightarrow Y$ be single-valued mapping. If Γ is upper hemicontinuous, then Γ is also a continuous function of $\Gamma : X \to Y$.*

Proof. The proof is straightforward. For $x \in X$, let $V \subseteq Y$ be an open set containing $f(x)$, i.e., $\{f(x)\} \subseteq V$. Since Γ is upper hemicontinuous and V is a neighborhood with $\{f(x)\} \subseteq V$, from Definition A.2.4 there exists a neighborhood U of x, such that for any $u \in U$, $\{\Gamma(u)\} \subseteq V$. Since U is a neighborhood of x, there exists an open set O that satisfies $x \in O, O \subseteq U$. Moreover $\forall u \in O$, $\Gamma(u) \in V$. Hence Γ is continuous in $x \in X$, $\forall x \in X$. Therefore, $\Gamma : X \to Y$ is a continuous function. □

Proposition A.2.8. *For any given $\boldsymbol{p} \in \prod_{j=1}^{N} \mathbb{R}^{2\frac{Z_j}{\Delta}+1}$, let $K_{\boldsymbol{p}} \subseteq \prod_{j=1}^{N} (I_\delta)^{2\frac{Z_j}{\Delta}+1}$ be the compact set defined in Proposition A.2.3. Then $\mathcal{T}_{\boldsymbol{p}}^{a} : K_{\boldsymbol{p}} \to K_{\boldsymbol{p}}$ is continuous.*

Proof. It suffices to verify the continuity of $\mathcal{T}_{\boldsymbol{p}}^{a,i,q_i}$ for given index i, q_i, where

$$\mathcal{T}_{\boldsymbol{p}}^{a,i,q_i}(\boldsymbol{\delta}) := \arg\max_{\delta \geq -\delta_\infty} (\delta - p_{q_i}^i) f_a^i\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big)$$

in other words to prove $\mathcal{T}_{\boldsymbol{p}}^{a,i,q_i}$ is continuous in terms of $(\vec{\delta}^{a,j})_{j\neq i}$. Write function $H_{q_i}^i(\delta) := (\delta - p_{q_i}^i) f_a^i\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big)$. From Lemma A.2.1, arg max of $H_{q_i}^i(\delta)$ exists and is unique for any $i \in \{1, \ldots, N\}, q_i \in \mathcal{Q}_i$. Hence $\mathcal{T}_{\boldsymbol{p}}^{a}$ is well defined as a single-valued mapping on $K_{\boldsymbol{p}}$.

$H_{q_i}^i$ is continuous in terms of $(\vec{\delta}^{a,j})_{j\neq i}$, and $\arg\max_{\delta} H_{q_i}^i(\delta)$ is taken in a compact set, denoted by $K_{\boldsymbol{p}}^{i,q_i} \subseteq \mathbb{R}$. Hence, the conditions in Lemma A.2.6 are satisfied. In fact, we take $f = H_{q_i}^i$, $X = \mathbb{R}$, $\Theta = \prod_{j\neq i} \mathbb{R}^{2\frac{Z_i}{\Delta}+1}$, $x = \delta \in X$, $\theta = (\vec{\delta}^{a,j})_{j\neq i} \in \Theta$. And $D(\theta) \equiv K'$ where K' is the projection of $K_{\boldsymbol{p}}$ on the subspace $\prod_{j\neq i} \mathbb{R}^{2\frac{Z_i}{\Delta}+1}$. Then $D(\theta) = K'$ is also a compact set. D as a constant mapping is both upper and lower hemicontinuous. Therefore, from Lemma A.2.6 $\mathcal{T}_{\boldsymbol{p}}^{a,i,q_i}$ is continuous as a function of $(\vec{\delta}^{a,j})_{j\neq i}$. Combining all coordinates $q_i \in \mathcal{Q}_i$ and $i \in \{1, \ldots, N\}$, we obtain $\mathcal{T}_{\boldsymbol{p}}^{a,i,q_i}$ is a continuous mapping on $K_{\boldsymbol{p}}$. □

Proposition A.2.9. *For any given $\boldsymbol{p} \in \prod_{j=1}^{N} \mathbb{R}^{2\frac{Z_j}{\Delta}+1}$, the mapping $\mathcal{T}_{\boldsymbol{p}}^{a} : K_{\boldsymbol{p}} \to K_{\boldsymbol{p}}$ has a fixed point. That is, there exists $\boldsymbol{\delta}_{\boldsymbol{p}} \in K_{\boldsymbol{p}}$ such that $\mathcal{T}_{\boldsymbol{p}}^{a}(\boldsymbol{\delta}_{\boldsymbol{p}}) = \boldsymbol{\delta}_{\boldsymbol{p}}$.*

Proof. By Proposition A.2.3, $K_{\boldsymbol{p}}$ is a compact set in $\prod_{j=1}^{N} \mathbb{R}^{2\frac{Z_j}{\Delta}+1}$, and therefore is closed. By construction in the proof of Proposition A.2.3, $K_{\boldsymbol{p}}$ is also a convex

set. By Proposition A.2.8, $\mathcal{T}^a_{\boldsymbol{p}}$ is continuous on $K_{\boldsymbol{p}}$. Then, by Schauder's fixed point theorem, $\mathcal{T}^a_{\boldsymbol{p}}$ has a fixed point in $K_{\boldsymbol{p}}$.

□

To proceed, we also need the uniqueness of the fixed point $\boldsymbol{\delta_p}$ and the continuity of the map $\boldsymbol{p} \mapsto \boldsymbol{\delta_p}$. To derive these results we use the following global implicit function theorem ([Galewski and Rădulescu 2018]):

Lemma A.2.10. *Let $F \in C^1(\mathbb{R}^n \times \mathbb{R}^m, \mathbb{R}^n)$ be a C^1 mapping which satisfies*

- *$\forall y \in \mathbb{R}^m$ the function $\phi_y(x)$ defined by $\phi_y(x) = \frac{1}{2}||F(x,y)||^2$ is coercive, i.e.*
$$\lim_{||x||\to\infty} \phi_y(x) = +\infty$$
- *The Jacobian matrix $\partial_x F(x,y)$ is non-singular for any $(x,y) \in \mathbb{R}^n \times \mathbb{R}^m$.*

Then there exists a unique function $f : \mathbb{R}^m \to \mathbb{R}^n$ such that $f \in C^1(\mathbb{R}^m, \mathbb{R}^n)$ and

$$\{(x,y) \in \mathbb{R}^n \times \mathbb{R}^m, F(x,y) = 0\} = \{(x,y) \in \mathbb{R}^n \times \mathbb{R}^m, x = f(y)\}.$$

Proposition A.2.11. *For any $\boldsymbol{p} \in \prod_{j=1}^N \mathbb{R}^{2^{\frac{Z_j}{\Delta}+1}}$, the fixed point $\boldsymbol{\delta_p}$ from Proposition A.2.9 is unique and the mapping*

$$\boldsymbol{\delta_p} = \boldsymbol{\delta}(\boldsymbol{p}) : \prod_{j=1}^N \mathbb{R}^{2^{\frac{Z_j}{\Delta}+1}} \to \prod_{j=1}^N (I_\delta)^{2^{\frac{Z_j}{\Delta}+1}}$$

is continuous in $\boldsymbol{p}$.

Proof. Given i and $q_i \in \mathcal{Q}_i$, from Lemma A.2.1, the maximal point of $H^i_{q_i}(\delta) = (\delta - p^i_{q_i}) f^i_a\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big)$ satisfies first order condition:

$$(\delta - p^i_{q_i})\partial_1 f^i_a\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big) + f^i_a\big(\delta, (\vec{\delta}^{a,j})_{j\neq i}\big) = 0 \tag{A.2.12}$$

Define a mapping $\mathcal{L}^i_{q_i}(\boldsymbol{\delta}, \boldsymbol{p}) : \prod_{j=1}^N \mathbb{R}^{2^{\frac{Z_j}{\Delta}+1}} \times \prod_{j=1}^N \mathbb{R}^{2^{\frac{Z_j}{\Delta}+1}} \to \mathbb{R}$.

$$\mathcal{L}^i_{q_i}(\boldsymbol{\delta}, \boldsymbol{p}) = -\frac{f^i_a\big(\delta^{a,i}_{q_i}, (\vec{\delta}^{a,j})_{j\neq i}\big)}{\partial_1 f^i_a\big(\delta^{a,i}_{q_i}, (\vec{\delta}^{a,j})_{j\neq i}\big)} - \delta^{a,i}_{q_i} + p^i_{q_i} \tag{A.2.13}$$

Then define the mapping

$$\mathcal{L}(\boldsymbol{\delta}, \boldsymbol{p}) = \big((\mathcal{L}^i_{q_i})_{q_i \in \mathcal{Q}_i}\big)_{i\in\{1,\dots,N\}} : \prod_{j=1}^N \mathbb{R}^{2^{\frac{Z_j}{\Delta}+1}} \times \prod_{j=1}^N \mathbb{R}^{2^{\frac{Z_j}{\Delta}+1}} \to \prod_{j=1}^N \mathbb{R}^{2^{\frac{Z_j}{\Delta}+1}}$$

We then compute the gradient of $\mathcal{L}$.

$$\frac{\partial \mathcal{L}^i_{q_i}}{\partial \delta^{a,i}_{q_i}} = -2 + \frac{\partial^2_{ii} f^i_a \cdot f^i_a}{(\partial_1 f^i_a)^2} \tag{A.2.14}$$

$$\frac{\partial \mathcal{L}^i_{q_i}}{\partial \delta^{a,i}_{q_j}} = 0, \forall q_j \neq q_i \tag{A.2.15}$$

$$\frac{\partial \mathcal{L}^i_{q_i}}{\partial \delta^{a,j}_{q_j}} = -\frac{(\partial_1 f^i_a)(\partial^j_{q_j} f^i_a) - f^i_a(\partial^j_{q_j}\partial_1 f^i_a)}{(\partial_1 f^i_a)^2}, \forall j \neq i \tag{A.2.16}$$

Then we have

$$\begin{aligned}
&\left|\frac{\partial \mathcal{L}^i_{q_i}}{\partial \delta^{a,i}_{q_i}}\right| - \sum_{q_j \neq q_i} \left|\frac{\partial \mathcal{L}^i_{q_i}}{\partial \delta^{a,i}_{q_j}}\right| - \sum_{j\neq i}\sum_{q_j \in \mathcal{Q}_j} \left|\frac{\partial \mathcal{L}^i_{q_i}}{\partial \delta^{a,j}_{q_j}}\right| \\
&= \frac{2(\partial_1 f^i_a)^2 - \partial^2_{ii} f^i_a \cdot f^i_a - \sum_{j\neq i}\sum_{q_j\in\mathcal{Q}_j}\left|(\partial_1 f^i_a)(\partial^j_{q_j} f^i_a) - f^i_a(\partial^j_{q_j}\partial_1 f^i_a)\right|}{(\partial_1 f^i_a)^2}
\end{aligned} \tag{A.2.17}$$

From Assumption 3.1.3, we have $\left|\frac{\partial \mathcal{L}^i_{q_i}}{\partial \delta^{a,i}_{q_i}}\right| - \sum_{q_j \neq q_i} \left|\frac{\partial \mathcal{L}^i_{q_i}}{\partial \delta^{a,i}_{q_j}}\right| - \sum_{j\neq i}\sum_{q_j \in \mathcal{Q}_j} \left|\frac{\partial \mathcal{L}^i_{q_i}}{\partial \delta^{a,j}_{q_j}}\right| > 0$ Hence the Jacobian matrix $\nabla_{\boldsymbol{\delta}} \mathcal{L}(\boldsymbol{\delta}, \boldsymbol{p})$ is diagonally dominant, and hence is nonsingular. Therefore, $\nabla_{\boldsymbol{\delta}} \mathcal{L}(\boldsymbol{\delta}, \boldsymbol{p})$ is bijective for any $(\boldsymbol{\delta}, \boldsymbol{p}) \in \prod_{j=1}^{N} \mathbb{R}^{2\frac{Z_j}{\Delta}+1} \times \prod_{j=1}^{N} \mathbb{R}^{2\frac{Z_j}{\Delta}+1}$.

Now it remains to prove $\boldsymbol{\mathcal{L}}(\boldsymbol{\delta}, \boldsymbol{p})$ is coercive. Given a sequence $\{\boldsymbol{\delta}^{(n)}\} \subseteq \prod_{j=1}^{N} \mathbb{R}^{2\frac{Z_j}{\Delta}+1}$ such that $||\boldsymbol{\delta}^{(n)}|| \to \infty$, there must exist a subsequence $\{(\delta^{a,i_n}_{q_{i_n}})^{(k_n)}\}$ such that $\lim_{n\to\infty} |(\delta^{a,i_n}_{q_{i_n}})^{(k_n)}| = \infty$. Otherwise, there exists a constant $M > 0$, such that for any $K > 0$ there exists $k > K$ and $|(\delta^i_{q_i})^{(k)}| < M$ for any i, q_i. Hence

$$||\boldsymbol{\delta}^{(k)}|| < M\sqrt{\prod_{i=1}^{N}(2\frac{Z_i}{\Delta}+1)}$$

This contradicts $||\boldsymbol{\delta}^{(n)}|| \to \infty$.

When $(\delta^{a,i_n}_{q_{i_n}})^{(k_n)} \to -\infty$, since

$$-\frac{f^i_a\big(\delta^{a,i}_{q_i}, (\vec{\delta}^{a,j})_{j\neq i}\big)}{\partial_1 f^i_a\big(\delta^{a,i}_{q_i}, (\vec{\delta}^{a,j})_{j\neq i}\big)} > 0$$

we have

$$\begin{aligned}
\boldsymbol{\mathcal{L}}^{i_n}_{q_{i_n}}(\boldsymbol{\delta}^{(k_n)}, \boldsymbol{p}) &= -\frac{f^i_a\Big((\delta^{a,i_n}_{q_{i_n}})^{(k_n)}, \big((\vec{\delta}^{a,j})^{(k_n)}\big)_{j\neq i}\Big)}{\partial_1 f^i_a\Big((\delta^{a,i_n}_{q_{i_n}})^{(k_n)}, \big((\vec{\delta}^{a,j})^{(k_n)}\big)_{j\neq i}\Big)} - (\delta^{a,i_n}_{q_{i_n}})^{(k_n)} + p^{i_n}_{q_{i_n}} \\
&> -(\delta^{a,i_n}_{q_{i_n}})^{(k_n)} + p^{i_n}_{q_{i_n}} \to \infty
\end{aligned} \tag{A.2.18}$$

When $(\delta^{a,i_n}_{q_{i_n}})^{(k_n)} \to +\infty$, from Assumption 3.1.3 let $Q = \lim_{\delta\to+\infty} \frac{f^i_a(\delta,\cdot)}{\partial_1 f^i_a(\delta,\cdot)} < \infty$ there exists $R > 0$ such that for any $n > R$ we have

$$\begin{aligned}\mathcal{L}^{i_n}_{q_{i_n}}(\boldsymbol{\delta}^{(k_n)}, \boldsymbol{p}) &= -\frac{f^i_a\Big((\delta^{a,i_n}_{q_{i_n}})^{(k_n)}, \big((\vec{\delta}^{a,j})^{(k_n)}\big)_{j\neq i}\Big)}{\partial_1 f^i_a\Big((\delta^{a,i_n}_{q_{i_n}})^{(k_n)}, \big((\vec{\delta}^{a,j})^{(k_n)}\big)_{j\neq i}\Big)} - (\delta^{a,i_n}_{q_{i_n}})^{(k_n)} + p^{i_n}_{q_{i_n}} \\ &\leq -Q + 1 - (\delta^{a,i_n}_{q_{i_n}})^{(k_n)} + p^{i_n}_{q_{i_n}} \to -\infty \end{aligned} \tag{A.2.19}$$

When there are two subsequences of $(\delta^{a,i_n}_{q_{i_n}})^{(k_n)} \to +\infty$ converging respectively to $+\infty$ and $-\infty$ from (A.2.18) and (A.2.19) we still have

$$\lim_{||\boldsymbol{\delta}^{k_n}||\to\infty} ||\mathcal{L}(\boldsymbol{\delta}^{k_n}, \boldsymbol{p})|| = +\infty$$

Therefore, $\mathcal{L}(\boldsymbol{\delta}, \boldsymbol{p})$ satisfies the conditions in Lemma A.2.10, so there exists a unique mapping $\boldsymbol{\delta} = \boldsymbol{\delta}(\boldsymbol{p})$ which is C^1 in $\boldsymbol{p}$. Define $\boldsymbol{\delta}_{\boldsymbol{p}} = \max(\boldsymbol{\delta}(\boldsymbol{p}), -\delta_\infty)$, then $\boldsymbol{\delta}_{\boldsymbol{p}}$ is continuous in $\boldsymbol{p}$. □

We can now prove Theorem 3.2.6.

Proof. (**Theorem 3.2.6**) Denote $\vec{V} = \Big(\big(V_i(q_i)\big)_{q_i\in\mathcal{Q}_i}\Big)_{i\in\{1,\dots,N\}}$ as an unknown vector that we want to solve in space $\prod_{j=1}^{N} \mathbb{R}^{2\frac{Z_j}{\Delta}+1}$. Then we take a specific $\hat{\boldsymbol{p}} \in \prod_{j=1}^{N} \mathbb{R}^{2\frac{Z_j}{\Delta}+1}$ such that

$$\hat{p}^i_{q_i} = \frac{V_i(q_i) - V_i(q_i - \Delta)}{\Delta}$$

From Proposition A.2.9 we can express the fixed point of $\mathcal{T}^a_{\hat{\boldsymbol{p}}}$ as a function of $\hat{\boldsymbol{p}}$, that is, a function of $\vec{V}$. We denote this fixed point by $\boldsymbol{\delta}(\vec{V}) = ((\delta^{a,i}_{q_i}(\vec{V}))_{q_i\in\mathcal{Q}_i})_{i\in\{1,\dots,N\}}$. Note that by Proposition A.2.11 $\boldsymbol{\delta}(\vec{V})$ is unique given $\vec{V}$, and is continuous in $\vec{V}$.

Equation (3.2.15) can be written as

$$\begin{aligned} &rV_i(q_i) + \psi_i(q_i) - \mathbb{I}(q_i > -Z_i)\lambda^a\Delta\Bigg[f^i_a\Big(\delta^{a,i}_{q_i}(\vec{V}), (\vec{\delta}^{a,j}(\vec{V}))_{j\neq i}\Big)\cdot \\ &\Big(\delta^{a,i}_{q_i}(\vec{V}) - \frac{V_i(q_i) - V_i(q_i-\Delta)}{\Delta}\Big)\Bigg] - \mathbb{I}(q_i < Z_i)\lambda^b\Delta\Bigg[f^i_b\Big(\delta^{b,i}_{q_i}(\vec{V}), (\vec{\delta}^{b,j}(\vec{V}))_{j\neq i}\Big)\cdot \\ &\Big(\delta^{b,i}_{q_i}(\vec{V}) - \frac{V_i(q_i) - V_i(q_i+\Delta)}{\Delta}\Big)\Bigg] = 0 \end{aligned} \tag{A.2.20}$$

To prove that there exists a solution to equation (3.2.15), it suffices to prove there exists a vector $\vec{V} = \Big(\big(V_i(q_i)\big)_{q_i\in\mathcal{Q}_i}\Big)_{i\in\{1,\dots,N\}} \in \prod_{j=1}^{N}\mathbb{R}^{2^{\frac{Z_j}{\Delta}+1}}$ satisfying the system of nonlinear equations (A.2.20).

We define $\mathbb{R}$-valued mappings $\mathcal{H}^{a,i}_{q_i}$ and $\mathcal{H}^{b,i}_{q_i}$ defined on $\prod_{j=1}^{N}\mathbb{R}^{2^{\frac{Z_j}{\Delta}+1}}$, such that for $\vec{V} \in \prod_{j=1}^{N}\mathbb{R}^{2^{\frac{Z_j}{\Delta}+1}}$,

$$\begin{aligned}
\mathcal{H}^{a,i}_{q_i}(\vec{V}) = &\quad \mathbb{I}(q_i > -Z_i) f^i_a\Big(\delta^{a,i}_{q_i}(\vec{V}), \big(\vec{\delta}^{a,j}(\vec{V})\big)_{j\neq i}\Big)\Big(\delta^{a,i}_{q_i}(\vec{V}) - \frac{V_i(q_i) - V_i(q_i-\Delta)}{\Delta}\Big)\\
\mathcal{H}^{b,i}_{q_i}(\vec{V}) = &\quad \mathbb{I}(q_i < Z_i) f^i_b\Big(\delta^{b,i}_{q_i}(\vec{V}), \big(\vec{\delta}^{b,j}(\vec{V})\big)_{j\neq i}\Big)\Big(\delta^{b,i}_{q_i}(\vec{V}) - \frac{V_i(q_i) - V_i(q_i+\Delta)}{\Delta}\Big)
\end{aligned} \tag{A.2.21}$$

Then $\mathcal{H}^{a,i}_{q_i}$ and $\mathcal{H}^{b,i}_{q_i}$ form mappings $\mathcal{H}^a, \mathcal{H}^b : \prod_{j=1}^{N}\mathbb{R}^{2^{\frac{Z_j}{\Delta}+1}} \to \prod_{j=1}^{N}\mathbb{R}^{2^{\frac{Z_j}{\Delta}+1}}$ when we group together $q_i \in \mathcal{Q}_i$ and $i \in \{1,\dots,N\}$, defined by

$$\mathcal{H}^a(\vec{V}) = \Big(\big(\mathcal{H}^{a,i}_{q_i}(\vec{V})\big)_{q_i\in\mathcal{Q}_i}\Big)_{i\in\{1,\dots,N\}}, \mathcal{H}^b(\vec{V}) = \Big(\big(\mathcal{H}^{b,i}_{q_i}(\vec{V})\big)_{q_i\in\mathcal{Q}_i}\Big)_{i\in\{1,\dots,N\}} \tag{A.2.22}$$

From equation (A.2.20) and notations in (A.2.22) we define mapping $\Phi : \prod_{j=1}^{N}\mathbb{R}^{2^{\frac{Z_j}{\Delta}+1}} \to \prod_{j=1}^{N}\mathbb{R}^{2^{\frac{Z_j}{\Delta}+1}}$, where $\Phi(\vec{V}) = \Big(\big(\Phi^i_{q_i}(\vec{V})\big)_{q_i\in\mathcal{Q}_i}\Big)_{i\in\{1,\dots,N\}}$

$$\Phi^i_{q_i}(\vec{V}) = (1-r)V_i(q_i) - \psi_i(q_i) + \lambda^a\Delta\cdot\mathcal{H}^{a,i}_{q_i}(\vec{V}) + \lambda^b\Delta\cdot\mathcal{H}^{b,i}_{q_i}(\vec{V}) \tag{A.2.23}$$

From Proposition 3.2.2, any equilibrium value function $V_i(q_i)$ is uniformly bounded by a constant M. Hence the vector $\vec{V}$ can be restricted on a convex and compact set $K \subseteq \prod_{j=1}^{N}\mathbb{R}^{2^{\frac{Z_j}{\Delta}+1}}$. $\psi_i(q_i)$ is also uniformly bounded $\forall q_i \in \mathcal{Q}_i, \forall i \in \{1,\dots,N\}$ by $\max_{i\in\{1,\dots,N\}}\Psi_i$ defined in the proof of Proposition 3.2.2. We now prove that $\mathcal{H}^{a,i}_{q_i}(\vec{V})$ and $\mathcal{H}^{b,i}_{q_i}(\vec{V})$ are also uniformly bounded. From (A.2.22), we obtain

$$|\mathcal{H}^{a,i}_{q_i}(\vec{V})| \le \Big|\delta^{a,i}_{q_i}(\vec{V})\Lambda\big(\delta^{a,i}_{q_i}(\vec{V})\big)\Big| + \Big|\Lambda\big(\delta^{a,i}_{q_i}(\vec{V})\big)\Big|\,\frac{2L}{\Delta} \tag{A.2.24}$$

where L is the uniform bound of $V_i(q_i)$ defined in Proposition 3.2.2. From Proposition A.2.3 $\delta^{a,i}_{q_i}(\vec{V}))$ takes value from a compact set $K^i_{q_i}$, the functions $\Lambda(\delta)$ and $\delta\Lambda(\delta)$ are continuous in δ, hence they are bounded on the compact set $K^i_{q_i}$. Therefore there exists a constant $C^i_{q_i}$ such that

$$|\mathcal{H}^{a,i}_{q_i}(\vec{V})| \le C^i_{q_i}$$

We then take $C := \max_{i,q_i} C^i_{q_i}$, then C is the uniform upper bound for $|\mathcal{H}^{a,i}_{q_i}(\vec{V})|$ regardless of q_i and i. We can similarly prove $|\mathcal{H}^{b,i}_{q_i}(\vec{V})|$ is also uniformly bounded by a positive constant.

Therefore $\forall q_i \in \mathcal{Q}_i, \forall i \in \{1,\dots,N\}$, the mapping $\Phi^i_{q_i}(\vec{V})$ is uniformly bounded by a closed interval $I^i_{q_i}$ regardless of $\vec{V}$. Define the set $A = \prod_{i=1}^{N}(\prod_{q_i\in\mathcal{Q}_i} I^i_{q_i})$, then $A \subseteq \prod_{j=1}^{N} \mathbb{R}^{2\frac{Z_j}{\Delta}+1}$ and A is a compact and convex set. Meanwhile, $\Phi(A) \subseteq A$.

Finally, by Proposition A.2.9 $\delta^{a,i}_{q_i}(\vec{V})$ and $\delta^{b,i}_{q_i}(\vec{V})$ are continuous functions of $\vec{V}$, then $\mathcal{H}^{a,i}_{q_i}(\vec{V})$ and $\mathcal{H}^{b,i}_{q_i}(\vec{V})$ are also continuous functions of $\vec{V}$. Therefore, $\Phi(\vec{V}) : \prod_{j=1}^{N} \mathbb{R}^{2\frac{Z_j}{\Delta}+1} \to \prod_{j=1}^{N} \mathbb{R}^{2\frac{Z_j}{\Delta}+1}$ is also a continuous mapping in terms of $\vec{V}$. Applying Schauder's fixed point theorem, Φ has a fixed point $\vec{V}^* \in A \subseteq \prod_{j=1}^{N} \mathbb{R}^{2\frac{Z_j}{\Delta}+1}$.

$$\Phi(\vec{V}^*) = \vec{V}^* \tag{A.2.25}$$

$\vec{V}^*$ satisfies the system of linear equations (A.2.20). Hence $\vec{V}^*$ satisfies the system of HJB equations (3.2.15). By the verification theorem Proposition 3.2.5, $\vec{V}^*$ is the equilibrium value functions of N market makers, while $\boldsymbol{\delta}(\vec{V}^*)$ is the joint quoting strategy under Nash equilibrium. □

A.3 Proof of Proposition 3.5.2

From Lemma 3.5.1, the running cost of the market maker is

$$\mathbb{E}_i\left[-\int_0^{\tau_a\wedge\tau_b} e^{-rt}\psi_i(q_i)dt\right] = -\frac{\psi_i(q_i)}{r+\lambda_a+\lambda_b}$$

Hence from (3.5.4) we obtain for $-Z_i < q_i < Z_i$:

$$\begin{aligned}
V^\delta_i(q_i) =& -\frac{\psi_i(q_i)}{r+\lambda_a+\lambda_b} + \mathbb{E}_i\Big[\mathbb{I}(R^i_a)\mathbb{I}(\tau_a<\tau_b)\left(e^{-r\tau_a}\delta^{a,i}_{q_i}\Delta + e^{-r\tau_a}V^\delta_i(q_i-\Delta)\right)\\
&+\mathbb{I}(R^i_b)\mathbb{I}(\tau_b<\tau_a)\big(e^{-r\tau_b}\delta^{b,i}_{q_i}\Delta + e^{-r\tau_b}V^\delta_i(q_i+\Delta)\big) + \mathbb{I}\big((R^i_a)^c\cap(R^i_b)^c\big)e^{-r\tau}V^\delta_i(q_i)\Big]\\
=& -\frac{\psi_i(q_i)}{r+\lambda_a+\lambda_b}\\
&+\mathbb{P}(\tau_a<\tau_b)\mathbb{E}_i\Big[\mathbb{I}(R^i_a)\big(e^{-r\tau_a}\delta^{a,i}_{q_i}\Delta + e^{-r\tau_a}V^\delta_i(q_i-\Delta)\big) + \mathbb{I}\big((R^i_a)^c\big)e^{-r\tau_a}V^\delta_i(q_i)\Big|\tau_a<\tau_b\Big]\\
&+\mathbb{P}(\tau_b<\tau_a)\mathbb{E}_i\Big[\mathbb{I}(R^i_b)\big(e^{-r\tau_b}\delta^{a,i}_{q_i}\Delta + e^{-r\tau_b}V^\delta_i(q_i+\Delta)\big) + \mathbb{I}\big((R^i_b)^c\big)e^{-r\tau_b}V^\delta_i(q_i)\Big|\tau_b<\tau_a\Big]
\end{aligned} \tag{A.3.1}$$

For $q_i = -Z_i$

$$
\begin{aligned}
V_i^{\delta}(-Z_i) =& -\frac{\psi_i(-Z_i)}{r+\lambda_a+\lambda_b} + \mathbb{E}_i\Big[\mathbb{I}(\tau_a<\tau_b)\big(e^{-r\tau_a}V_i^{\delta}(-Z_i)\big) \\
&+ \mathbb{I}(\tau_b<\tau_a)\Big(\mathbb{I}(R_b^i)\big(e^{-r\tau_b}\delta^{b,i}_{-Z_i}\Delta + e^{-r\tau_b}V_i^{\delta}(-Z_i+\Delta)\big) + \mathbb{I}\big((R_b^i)^c\big)e^{-r\tau_b}V_i^{\delta}(-Z_i)\Big)\Big] \\
=& -\frac{\psi_i(-Z_i)}{r+\lambda_a+\lambda_b} + \mathbb{P}(\tau_a<\tau_b)\mathbb{E}_i\Big[e^{-r\tau_a}V_i^{\delta}(-Z_i)\Big|\tau_a<\tau_b\Big] \\
&+ \mathbb{P}(\tau_b<\tau_a)\mathbb{E}_i\Big[\mathbb{I}(R_b^i)\left(e^{-r\tau_b}\delta^{b,i}_{-Z_i}\Delta + e^{-r\tau_b}V_i^{\delta}(-Z_i+\Delta)\right) \\
&+ \mathbb{I}\big((R_b^i)^c\big)e^{-r\tau_b}V_i^{\delta}(-Z_i)\Big|\tau_b<\tau_a\Big]
\end{aligned}
\tag{A.3.2}
$$

For $q_i = Z_i$

$$
\begin{aligned}
V_i^{\delta}(Z_i) =& -\frac{\psi_i(Z_i)}{r+\lambda_a+\lambda_b} + \mathbb{E}_i\Big[\mathbb{I}(\tau_b<\tau_a)\big(e^{-r\tau_b}V_i^{\delta}(Z_i)\big) \\
&+ \mathbb{I}(\tau_a<\tau_b)\Big(\mathbb{I}(R_a^i)\big(e^{-r\tau_a}\delta^{a,i}_{Z_i}\Delta + e^{-r\tau_a}V_i^{\delta}(Z_i-\Delta)\big) + \mathbb{I}\big((R_a^i)^c\big)e^{-r\tau_a}V_i^{\delta}(Z_i)\Big)\Big] \\
=& -\frac{\psi_i(Z_i)}{r+\lambda_a+\lambda_b} + \mathbb{P}(\tau_b<\tau_a)\mathbb{E}_i\Big[e^{-r\tau_b}V_i^{\delta}(Z_i)\Big|\tau_b<\tau_a\Big] \\
&+ \mathbb{P}(\tau_a<\tau_b)\mathbb{E}_i\Big[\mathbb{I}(R_a^i)\left(e^{-r\tau_a}\delta^{a,i}_{Z_i}\Delta + e^{-r\tau_a}V_i^{\delta}(Z_i-\Delta)\right) \\
&+ \mathbb{I}\big((R_a^i)^c\big)e^{-r\tau_a}V_i^{\delta}(Z_i)\Big|\tau_a<\tau_b\Big]
\end{aligned}
\tag{A.3.3}
$$

By combining equations (A.3.1)-(A.3.3) with indicator functions $\mathbb{I}(-Z_i < q_i \leq Z_i)$ and $\mathbb{I}(-Z_i \leq q_i < Z_i)$ we obtain (3.5.9).

Appendix B

Appendix of Chapter 4

B.1 Proof of Theorem 4.1.8

The proof of the existence of a mean field Nash equilibrium is based on demonstrating the Lipschitz continuity of the operators derived from the forward and backward equations. We then apply Schauder's fixed point theorem on the function space of population distribution flows. We first provide original proofs for Lemma B.1.1 and Lemma B.1.2. The proof of Theorem 4.1.8 follows the general framework established in the mean field game literature, including the existence results from [Gomes, Mohr, and Souza 2013] and [Guéant 2015].

We first study the property of Ξ^a defined in (4.1.23). We have the following lemma.

Lemma B.1.1. *Under Assumption 4.1.2, for any $p, \mu \in \mathbb{R}$, $\Xi^a(p,\mu)$ and $\Xi^b(p,\mu)$ in (4.1.23) are well defined. $\Xi^a(p,\mu)$ and $\Xi^b(p,\mu)$ are continuous functions of (p,μ). For any given $p \in \mathbb{R}$ there exist constants m_p, M_p such that $m_p \leq \Xi^a(p,\mu) \leq M_p$ and $m_p \leq \Xi^b(p,\mu) \leq M_p$.*

Proof. It suffices to prove the case for Ξ^a. We first define $\zeta_{p,\mu}(\delta) = f_a(\delta,\mu)(\delta - p)$. To show that Ξ^a is well defined, we need to prove that $\zeta_{p,\mu}(\delta)$ achieves a unique maximum point on I_δ. The first order derivative of $\zeta_{p,\mu}$ yields

$$\begin{aligned}\zeta'_{p,\mu}(\delta) &= (\delta - p)\partial_\delta f_a(\delta,\mu) + f_a(\delta,\mu) \\ &= \partial_\delta f_a(\delta,\mu)\left[\delta - p + \frac{f_a(\delta,\mu)}{\partial_\delta f_a(\delta,\mu)}\right] \end{aligned} \tag{B.1.1}$$

Using the same argument as from Lemma A.1 of [Cont and Xiong 2024] we can show that there exists a unique point $\delta^{max}(p,\mu) > p$ in $\mathbb{R}$ such that $\zeta'_{p,\mu}(\delta) > 0, \forall \delta < \delta^{max}(p,\mu)$, $\zeta'_{p,\mu}(\delta) < 0, \forall \delta > \delta^{max}(p,\mu)$, and $\zeta'_{p,\mu}(\delta^{max}(p,\mu)) = 0$.

That is, $\delta^{max}(p,\mu)$ is the unique maximum point of $\zeta_{p,\mu}(\delta)$ in $\mathbb{R}$. $\delta^{max}(p,\mu)$ is characterized by

$$\delta^{max}(p,\mu) - p + \frac{f_a(\delta^{max}(p,\mu),\mu)}{\partial_\delta f_a(\delta^{max}(p,\mu),\mu)} = 0 \tag{B.1.2}$$

From Assumption 4.1.2 and by the Implicit Function Theorem, $\delta^{max}(p,\mu)$ is continuously differentiable in (p,μ) and we have

$$\begin{cases} \frac{\partial \delta^{max}(p,\mu)}{\partial p} = \frac{\left(\partial_\delta f_a\left(\delta^{max}(p,\mu),\mu\right)\right)^2}{2\left(\partial_\delta f_a\left(\delta^{max}(p,\mu),\mu\right)\right)^2 - f_a\left(\delta^{max}(p,\mu),\mu\right)\cdot\partial^2_{\delta\delta} f_a\left(\delta^{max}(p,\mu),\mu\right)} \\ \frac{\partial \delta^{max}(p,\mu)}{\partial \mu} = \frac{-\partial_\delta f_a\left(\delta^{max}(p,\mu),\mu\right)\partial_\mu f_a\left(\delta^{max}(p,\mu),\mu\right) + f_a\left(\delta^{max}(p,\mu),\mu\right)\cdot\partial_\delta\partial_\mu f_a\left(\delta^{max}(p,\mu),\mu\right)}{2\left(\partial_\delta f_a\left(\delta^{max}(p,\mu),\mu\right)\right)^2 - f_a\left(\delta^{max}(p,\mu),\mu\right)\cdot\partial^2_{\delta\delta} f_a\left(\delta^{max}(p,\mu),\mu\right)} \end{cases} \tag{B.1.3}$$

Therefore, the value of $\Xi^a(p,\mu)$ is well defined achieving a unique value, with

$$\Xi^a(p,\mu) = \max\big(-\delta_\infty, \delta^{max}(p,\mu)\big) \tag{B.1.4}$$

The continuity of $\Xi^a(p,\mu)$ follows immediately from (B.1.3) and (B.1.4). Define the closed set $\Upsilon = \{(p,\mu)|\delta^{max}(p,\mu) \le -\delta_\infty\}$, then $\Xi^a(p,\mu) \equiv -\delta_\infty$ in Υ and $\Xi^a(p,\mu) = \delta^{max}(p,\mu)$ on $\mathbb{R}^2 \setminus \Upsilon$.

From Assumption 4.1.2, $f_a(\delta,\mu) > 0$, so whenever $\delta > p$ we have $f_a(\delta,\mu)(\delta - p) > 0$. Therefore, $\arg\max_\delta f_a(\delta,\mu)(\delta - p)$ is achieved when $\delta > p$, hence $\Xi^a(p,\mu) \ge \min(-\delta_\infty, p)$.

On the other hand, take $\hat{\delta} = p+1$. Since $f_a(\delta,\mu) \le \Lambda(\delta)$, and $\lim_{\delta\to\infty} \Lambda(\delta)\delta = 0$, we also have $\lim_{\delta\to\infty} \Lambda(\delta) = 0$. So we obtain that there exists a large enough $\delta_M > \max(-\delta_\infty, p+1)$ such that

$$f_a(\hat{\delta},\mu)(\hat{\delta} - p) = f_a(\hat{\delta},\mu) \ge \Lambda(\delta_M)(\delta_M - p) \ge f_a(\delta_M,\mu)(\delta_M - p) \tag{B.1.5}$$

where δ_M is dependent on p. Therefore, the maximum of $f_a(\delta,\mu)(\delta - p)$ is obtained between $(p, \delta_M]$. Assigning $m_p = \max(-\delta_\infty, p)$ and $M_p = \delta_M$, we conclude that

$$m_p \le \Xi^a(p,\mu) \le M_p$$

□

Lemma B.1.2. *Define the mappings* $G^a, G^b : \mathbb{R} \times \mathbb{R}^{2H+1} \times \mathcal{S}^{\mathcal{Q}} \to \mathbb{R}$*:*

$$\begin{aligned} G^a(\mu,u,\theta) &= \sum_{\xi\in\mathcal{Q}} \Xi^a\big(u(\xi) - u(\xi-1),\mu\big)\theta(\xi)\mathbb{I}(\xi > -Z) \\ G^b(\mu,u,\theta) &= \sum_{\xi\in\mathcal{Q}} \Xi^a\big(u(\xi) - u(\xi+1),\mu\big)\theta(\xi)\mathbb{I}(\xi < Z) \end{aligned} \tag{B.1.6}$$

Under Assumption 4.1.2, for any pair $(u,\theta)\in\mathbb{R}^{2H+1}\times\mathcal{S}^{\mathcal{Q}}$ there exist a unique fixed point $\mu^a(u,\theta)\in\mathbb{R}$ of G^a and a unique fixed point $\mu^b(u,\theta)\in\mathbb{R}$ of G^b. Furthermore, $\mu^a(u,\theta)$ and $\mu^b(u,\theta)$ are locally Lipschitz continuous functions of (u,θ).

Proof. It suffices to prove the existence and uniqueness of a fixed point of mapping G^a. From Lemma B.1.1 we obtain the regularity property of $\Xi^a(p,\mu)$ with respect to (p,μ). We take the closed set Υ defined in the proof of Lemma B.1.1. Whenever $(p,\mu)\in\mathbb{R}^2\setminus\Upsilon$ we can write the derivatives of Ξ^a in the implicit form of Ξ^a:

$$\begin{cases}\frac{\partial\Xi^a(p,\mu)}{\partial p}=\frac{\Big(\partial_\delta f_a\Big(\Xi^a(p,\mu),\mu\Big)\Big)^2}{2\Big(\partial_\delta f_a\Big(\Xi^a(p,\mu),\mu\Big)\Big)^2-f_a\Big(\Xi^a(p,\mu),\mu\Big)\cdot\partial^2_{\delta\delta}f_a\Big(\Xi^a(p,\mu),\mu\Big)}\\ \frac{\partial\Xi^a(p,\mu)}{\partial\mu}=\frac{-\partial_\delta f_a\Big(\Xi^a(p,\mu),\mu\Big)\partial_\mu f_a\Big(\Xi^a(p,\mu),\mu\Big)+f_a\Big(\Xi^a(p,\mu),\mu\Big)\cdot\partial_\delta\partial_\mu f_a\Big(\Xi^a(p,\mu),\mu\Big)}{2\Big(\partial_\delta f_a\Big(\Xi^a(p,\mu),\mu\Big)\Big)^2-f_a\Big(\Xi^a(p,\mu),\mu\Big)\cdot\partial^2_{\delta\delta}f_a\Big(\Xi^a(p,\mu),\mu\Big)}\end{cases}\tag{B.1.7}$$

From Point 6 in Assumption 4.1.2, we have

$$\Big|\frac{\partial\Xi^a(p,\mu)}{\partial\mu}\Big|=\Big|\frac{\partial_\delta f\cdot\partial_\mu f-f\cdot\partial_\delta\partial_\mu f}{2(\partial_\delta f)^2-\partial^2_{\delta\delta}f\cdot f}\Big|(\Xi^a(p,\mu),\mu)\le c<1\tag{B.1.8}$$

When $(p,\mu)\in\Upsilon$, $\Xi^a(p,\mu)$ is constant, therefore the partial derivatives $\partial_p\Xi^a,\partial_\mu\Xi^a$ are 0 at the interior of Υ. For any point (p,μ) on the boundary $\partial\Upsilon$ we have $\limsup_{x\to\mu}|\frac{\Xi^a(p,x)-\Xi^a(p,\mu)}{x-\mu}|\le c<1$.

Hence whenever the derivatives in (B.1.9) are valid we have:

$$\begin{aligned}\Big|\frac{\partial G^a(\mu,u,\theta)}{\partial\mu}\Big|&=\Big|\sum_{\xi\in\mathcal{Q}}\frac{\partial\Xi^a\big(u(\xi)-u(\xi-1),\mu\big)}{\partial\mu}\theta(\xi)\mathbb{I}(\xi>-Z)\Big|\\&\le c\sum_{\xi\in\mathcal{Q}}\theta(\xi)\mathbb{I}(q>-Z)\le c<1\end{aligned}\tag{B.1.9}$$

For any μ where the derivatives in (B.1.9) are not well-defined we still have

$$\Big|\limsup_{x\to\mu}\frac{G^a(x,u,\theta)-G^a(\mu,u,\theta)}{x-\mu}\Big|\le c\sum_{\xi\in\mathcal{Q}}\theta(\xi)\mathbb{I}(q>-Z)\le c<1\tag{B.1.10}$$

Therefore, for given (u,θ), $G^a(\mu,u,\theta)$ is a contraction mapping on $\mathbb{R}$. By Banach's Fixed Point Theorem, there exists a unique fixed point $\mu^a(u,\theta)$ of $G^a(\mu,u,\theta)$.

For simplicity and without loss of generality, we now assume $\Xi^a(p,\mu)$ is C^1 for all (p,μ). Note then that G^a is C^1 in terms of u and θ, and $\theta\in\mathcal{S}^{\mathcal{Q}}$ where

$\mathcal{S}^{\mathcal{Q}}$ is closed and bounded. To show the local Lipschitz continuity of $\mu^a(u,\theta)$, we apply Implicit Function Theorem on following equation:

$$G^a(\mu,u,\theta)-\mu=0 \tag{B.1.11}$$

(B.1.11) is well defined from the existence and uniqueness of a fixed point $\mu^a(u,\theta)$. From (B.1.9) we see that $\frac{\partial G^a(\mu,u,\theta)}{\partial \mu}-1\neq 0$. By the Implicit Function Theorem, there exists an open neighborhood $U=U_1\times U_2$ of (u,θ) where $U_1\subset\mathbb{R}, U_2\subset\mathcal{S}^{\mathcal{Q}}$ such that $\mu^a(u,\theta)$ is continuously differentiable on $U_1\times U_2$.

From (B.1.9) we obtain a bound for $\frac{\partial G^a(\mu,u,\theta)}{\partial \mu}-1$:

$$-c-1\leq \frac{\partial G^a(\mu,u,\theta)}{\partial \mu}-1\leq c-1<0 \tag{B.1.12}$$

From Lemma B.1.1 $\Xi^a(p,\mu)$ is a continuous function of (p,μ). We take closed intervals $I_1\times I_2\subset U_1\times U_2$ such that $(u,\theta)\in I_1\times I_2$, then the gradient $\nabla_\theta G^a(\mu,u,\theta)|_{\mu=\mu^a(u,\theta)}=\Big(\big(\Xi^a\big(u(\xi)-u(\xi-1),\mu^a(u,\theta)\big)\mathbb{I}(\xi>-Z)\big)\Big)_{\xi\in\mathcal{Q}}$ is bounded uniformly on the closed and bounded interval $I_1\times I_2$ because of the continuity of $\Xi^a(u(\xi)-u(\xi-1),\mu^a(u,\theta))$ in (u,θ) for all $\xi\in\mathcal{Q}$.

We then calculate $\nabla_u G^a(\mu,u,\theta)|_{\mu=\mu^a(u,\theta)}$:

$$\begin{aligned}\nabla_u G^a(\mu,u,\theta)|_{\mu=\mu^a(u,\theta)}=&\Big(\frac{\partial\Xi^a}{\partial p}\big(u(\xi)-u(\xi-1),\mu^a(u,\theta)\big)\theta(\xi)\mathbb{I}(\xi>-Z)\\&-\frac{\partial\Xi^a}{\partial p}\big(u(\xi+1)-u(\xi),\mu^a(u,\theta)\big)\theta(\xi+1)\mathbb{I}(\xi+1>-Z)\Big)_{\xi\in\mathcal{Q}}\end{aligned} \tag{B.1.13}$$

From (B.1.7) and Assumption 4.1.2, we obtain $0<\frac{\partial\Xi^a(p,\mu)}{\partial p}\leq\frac{1}{C}$ where C is the constant at Point 4 of Assumption 4.1.2. Hence, the coordinates of $\nabla_u G^a(\mu,u,\theta)|_{\mu=\mu^a(u,\theta)}$ in (B.1.13) are uniformly bounded.

Therefore by Implicit Function Theorem within the neighborhood U we can write the derivatives $(\nabla_u\mu^a,\nabla_\theta\mu^a)$:

$$(\nabla_u\mu^a,\nabla_\theta\mu^a)=-(\frac{\partial G^a(\mu,u,\theta)}{\partial\mu}-1)(\nabla_u G^a,\nabla_\theta G^a) \tag{B.1.14}$$

$(\nabla_u\mu^a,\nabla_\theta\mu^a)$ is uniformly bounded on $I_1\times I_2$. Therefore, μ^a is locally Lipschitz continuous in (u,θ). □

Proof. (**Proof of Theorem 4.1.8**) For a given continuous flow of probability distribution $\boldsymbol{M}:[0,T]\to\mathcal{S}^{\mathcal{Q}}$, we seek the existence and uniqueness of solution $\boldsymbol{V}:[0,T]\to(\mathbb{R}^{2H+1},||\cdot||_\infty)$ to the following backward differential equation:

$$\begin{cases}\frac{d\boldsymbol{V}}{dt}(t)=\Phi(\boldsymbol{V}(t),\boldsymbol{M}(t))\\ \boldsymbol{V}(T)(\cdot)=-\phi(\cdot)\end{cases} \tag{B.1.15}$$

where the mapping $\Phi : (u,\theta) \in \mathbb{R}^{2H+1} \times \mathcal{S}^{\mathcal{Q}} \to \mathbb{R}^{2H+1}$ is defined from the HJB equation in (4.1.25):

$$\begin{aligned}\Phi(u,\theta)(q) = ru(q) + \psi(q) - \mathcal{H}^a(u(q) - u(q-1), \mu^a(u,\theta))\mathbb{I}(q > -Z) \\ - \mathcal{H}^b(u(q) - u(q+1), \mu^b(u,\theta))\mathbb{I}(q < Z) \quad \text{(B.1.16)}\end{aligned}$$

with $\mu^a, \mu^b : (u,\theta) \in \mathbb{R}^{2H+1} \times \mathcal{S}^{\mathcal{Q}} \to \mathbb{R}$ defined by

$$\begin{aligned}\mu^a(u,\theta) &= \sum_{\xi \in \mathcal{Q}} \Xi^a\big(u(\xi) - u(\xi-1), \mu^a(u,\theta)\big)\theta(\xi)\mathbb{I}(\xi > -Z) \\ \mu^b(u,\theta) &= \sum_{\xi \in \mathcal{Q}} \Xi^b\big(u(\xi) - u(\xi+1), \mu^b(u,\theta)\big)\theta(\xi)\mathbb{I}(\xi < Z) \quad \text{(B.1.17)}\end{aligned}$$

From Lemma B.1.2, (B.1.17) admits a unique solution $\big(\mu^a(u,\theta), \mu^b(u,\theta)\big)$ for each $(u,\theta) \in \mathbb{R}^{2H+1} \times \mathcal{S}^{\mathcal{Q}}$ and $\big(\mu^a(u,\theta), \mu^b(u,\theta)\big)$ are locally Lipschitz on $\mathbb{R}^{2H+1} \times \mathcal{S}^{\mathcal{Q}}$.

Using the definitions of Ξ^a we can calculate the derivatives of function $\mathcal{H}^a(p,\mu)$.

$$\begin{aligned}\frac{\partial \mathcal{H}^a}{\partial p}(p,\mu) &= -\lambda_a f_a(\Xi^a(p,\mu),\mu) \\ \frac{\partial \mathcal{H}^a}{\partial \mu}(p,\mu) &= -\lambda_a f_a(\Xi^a(p,\mu),\mu)\frac{\partial_\mu f_a(\Xi^a(p,\mu),\mu)}{\partial_\delta f_a(\Xi^a(p,\mu),\mu)} \quad \text{(B.1.18)}\end{aligned}$$

From Assumption 4.1.2 $\left|\frac{\partial_\mu f_a(\delta,\mu)}{\partial_\delta f_a(\delta,\mu)}\right| \le K$ and $f_a(\delta,\mu)$ is uniformly bounded when $\delta > -\delta_\infty$, we see that both $\left|\frac{\partial \mathcal{H}^a}{\partial p}\right|$ and $\left|\frac{\partial \mathcal{H}^a}{\partial \mu}\right|$ are uniformly bounded. Let us take $(u,\theta) \in \mathbb{R}^{2H+1} \times \mathcal{S}^{\mathcal{Q}}$. From Lemma B.1.2 $\mu^a(u,\theta)$ is locally Lipschitz. There exists a constant $K_{u,\theta}$ such that μ^a is Lipschitz continuous within a neighborhood O of (u,θ). We then take a u' such that $(u',\theta) \in O$, and compute

$$\begin{aligned}\Big(\Phi(u,\theta) - \Phi(u',\theta)\Big)(q) = r(u-u')(q) - \mathcal{H}^a(u(q) - u(q-1), \mu^a(u,\theta))\mathbb{I}(q > -Z) \\ - \mathcal{H}^b(u(q) - u(q+1), \mu^b(u,\theta))\mathbb{I}(q < Z) \\ + \mathcal{H}^a(u'(q) - u'(q-1), \mu^a(u',\theta))\mathbb{I}(q > -Z) \\ + \mathcal{H}^b(u'(q) - u'(q+1), \mu^b(u',\theta))\mathbb{I}(q < Z) \quad \text{(B.1.19)}\end{aligned}$$

We shall now study the term

$$\begin{aligned}&\mathcal{H}^a(u(q) - u(q-1), \mu^a(u,\theta)) - \mathcal{H}^a(u'(q) - u'(q-1), \mu^a(u',\theta)) \\ =&\Big(\mathcal{H}^a(u(q) - u(q-1), \mu^a(u,\theta)) - \mathcal{H}^a(u'(q) - u'(q-1), \mu^a(u,\theta))\Big) \\ &+ \Big(\mathcal{H}^a(u'(q) - u'(q-1), \mu^a(u,\theta)) - \mathcal{H}^a(u'(q) - u'(q-1), \mu^a(u',\theta))\Big) \quad \text{(B.1.20)}\end{aligned}$$

Since $\left|\frac{\partial \mathcal{H}^a}{\partial p}\right|$ is uniformly bounded, let L_p be its upper bound, application of Mean Value Theorem yields

$$\begin{aligned}&\Big|\mathcal{H}^a(u(q)-u(q-1),\mu^a(u,\theta))-\mathcal{H}^a(u'(q)-u'(q-1),\mu^a(u,\theta))\Big|\\&\qquad\leq L_p|u(q)-u(q-1)-u'(q)+u'(q-1)|\leq 2L_p||u-u'||_\infty\end{aligned}\tag{B.1.21}$$

Since $\left|\frac{\partial \mathcal{H}^a}{\partial \mu}\right|$ is uniformly bounded, let L_μ denote its upper bound, then we see that

$$\begin{aligned}&\Big|\mathcal{H}^a(u'(q)-u'(q-1),\mu^a(u,\theta))-\mathcal{H}^a(u'(q)-u'(q-1),\mu^a(u',\theta))\Big|\\&\qquad\leq L_\mu\Big|\mu^a(u,\theta)-\mu^a(u',\theta)\Big|\leq L_\mu K_{u,\theta}||u-u'||_\infty\end{aligned}\tag{B.1.22}$$

Combining (B.1.19)-(B.1.22) we obtain that $\mathcal{H}^a(u(q)-u(q-1),\mu^a(u,\theta))$ is locally Lipschitz in u. Same conclusion can be drawn for $\mathcal{H}^b(u(q)-u(q+1),\mu^a(u,\theta))$. Therefore when θ is given, the mapping $\Phi(u,\theta)$ is locally Lipschitz continuous in terms of u. Moreover we can verify that $\Phi(u,\theta)$ is continuous in θ. Since the mapping $\boldsymbol{M}:[0,T]\to\mathcal{S}^{\mathcal{Q}}$ is continuous, from Cauchy-Lipschitz Theorem, there exists a unique solution $\boldsymbol{V}^{\boldsymbol{M}}$ of (B.1.15), and $\boldsymbol{V}^{\boldsymbol{M}}$ depends continuously on the parameter $\boldsymbol{M}$.

As a second step for a given continuous mapping $\boldsymbol{V}:[0,T]\to(\mathbb{R}^{2H+1},||\cdot||_\infty)$, we prove the existence and uniqueness of solution $\boldsymbol{M}:[0,T]\to\mathcal{S}^{\mathcal{Q}}$ to the forward differential equation:

$$\begin{cases}\frac{d\boldsymbol{M}}{dt}(t)=\Psi(\boldsymbol{M}(t),\boldsymbol{V}(t))\\ \boldsymbol{M}(0)(\cdot)=m_0(\cdot)\end{cases}\tag{B.1.23}$$

where the mapping $\Psi:(\theta,u)\in\mathcal{S}^{\mathcal{Q}}\times\mathbb{R}^{2H+1}\to\mathcal{S}^{\mathcal{Q}}$ is defined from the forward Fokker-Planck equation in (4.1.25):

$$\begin{aligned}\Psi(\theta,u)(q)=&\lambda_a f_a\Big(\Xi^a\big(u(q+1)-u(q),\mu^a(u,\theta)\big),\mu^a(u,\theta)\Big)\theta(q+1)\mathbb{I}(q<Z)\\&+\lambda_b f_b\Big(\Xi^b\big(u(q-1)-u(q),\mu^b(u,\theta)\big),\mu^b(u,\theta)\Big)\theta(q-1)\mathbb{I}(q>-Z)\\&-\lambda_a f_a\Big(\Xi^a\big(u(q)-u(q-1),\mu^a(u,\theta)\big),\mu^a(u,\theta)\Big)\theta(q)\mathbb{I}(q>-Z)\\&-\lambda_b f_b\Big(\Xi^b\big(u(q)-u(q+1),\mu^b(u,\theta)\big),\mu^b(u,\theta)\Big)\theta(q)\mathbb{I}(q<Z)\end{aligned}\tag{B.1.24}$$

with $\mu^a, \mu^b : (u, \theta) \in \mathbb{R}^{2H+1} \times \mathcal{S}^{\mathcal{Q}} \to \mathbb{R}$ defined in (B.1.17). It is straightforward to verify that $\Psi(\theta, u)$ is Lipschitz continuous in θ since f_a, f_b are uniformly bounded by a constant B_f when $\delta \geq -\delta_\infty$. And $\Psi(\theta, u)$ is also continuous in u, from the continuity of functions $f_a, f_b, \Xi^a, \Xi^b, \mu^a, \mu^b$. Therefore, by the Cauchy-Lipschitz Theorem, there exists a unique solution $\boldsymbol{M}^{\boldsymbol{V}}$ of (B.1.23). And $\boldsymbol{M}^{\boldsymbol{V}}$ depends continuously on the parameter$\boldsymbol{V}$.

Finally we consider a subset $\mathcal{C}$ of $C([0,T] \to \mathcal{S}^{\mathcal{Q}})$ defined by

$$\mathcal{C} = \left\{ f \in C([0,T] \to \mathcal{S}^{\mathcal{Q}}) \middle| \sup_{s \neq t} \frac{||f(t) - f(s)||_\infty}{|t-s|} \leq 2(\lambda_a + \lambda_b) B_f \right\} \quad \text{(B.1.25)}$$

Then we can verify that $\mathcal{C}$ is a closed and convex set. And by definition $\mathcal{C}$ is equicontinuous. Since $\mathcal{S}^{\mathcal{Q}}$ is a bounded subset of $\mathbb{R}^{2H+1}$, $\mathcal{C}$ is uniformly bounded in $||\cdot||_\infty$ norm, therefore by Arzela-Ascoli Theorem $\mathcal{C}$ is relatively compact in $C([0,T] \to \mathcal{S}^{\mathcal{Q}})$.

We define the mapping $\mathcal{T} : C([0,T] \to \mathcal{S}^{\mathcal{Q}}) \to C([0,T] \to \mathcal{S}^{\mathcal{Q}})$: for any $\boldsymbol{M} \in C([0,T] \to \mathcal{S}^{\mathcal{Q}})$, we can obtain a unique solution $\boldsymbol{V}^{\boldsymbol{M}}$ to differential equation (B.1.15) with $\boldsymbol{M}$ being its parameter. Then considering $\boldsymbol{V}^{\boldsymbol{M}}$ as the parameter of the differential equation (B.1.23) there exists a unique solution $\tilde{\boldsymbol{M}}$ to (B.1.23). We therefore define

$$\tilde{\boldsymbol{M}} = \mathcal{T}(\boldsymbol{M})$$

Mapping $\mathcal{T}$ is well defined and continuous under uniform norms. For any $\boldsymbol{M} \in \mathcal{C}$, since $\tilde{\boldsymbol{M}} = \mathcal{T}(\boldsymbol{M})$ solves equation (B.1.23), from the upper boundedness of operator Ψ by $2(\lambda_a + \lambda_b)B_f$ we have

$$\sup_{s \neq t} \frac{||\tilde{\boldsymbol{M}}(t) - \tilde{\boldsymbol{M}}(s)||_\infty}{t-s} \leq 2(\lambda_a + \lambda_b) B_f$$

Therefore $\tilde{\boldsymbol{M}} \in \mathcal{C}$. We therefore obtain a continuous mapping $\mathcal{T} : \mathcal{C} \to \mathcal{C}$

By Schauder's Fixed Point Theorem, there exists a fixed point $\boldsymbol{M}^* \in \mathcal{C}$ of $\mathcal{T}$. Let $\boldsymbol{V}^*$ denote the corresponding solution to equation (B.1.15) and define $V^*(t,q) = \boldsymbol{V}(t)(q), m^*(t,q) = \boldsymbol{M}^*(t)(q)$, then (V^*, m^*) is a classical solution to system (4.1.25). □

B.2 Proof of Proposition 4.1.9

The proof of Proposition 4.1.9 follows a standard argument from the optimal control literature, as detailed in [Pham 2009]. A verification theorem specific to the mean field game of major-minor market making problem can be found in [Bergault and Guéant 2021].

Proof. We first show that for any given population distribution flow $\tilde{\boldsymbol{M}} : [0,T] \to \mathcal{S}^{\mathcal{Q}}$ associated with probability density function $\tilde{m}(t,q)$ at time t, if a function $V(t,q)$ satisfy the following Hamilton-Jacobi equation B.2.1

$$\begin{cases} 0 = \partial_t V - rV - \psi(q) + \mathcal{H}^a\Big(V(t,q) - V(t,q-1), \tilde{\mu}^a(t)\Big)\mathbb{I}(q > -Z) \\ \qquad\quad + \mathcal{H}^b\Big(V(t,q) - V(t,q+1), \tilde{\mu}^b(t)\Big)\mathbb{I}(q < Z) \\ V(T,q) = -\phi(q) \\ \tilde{\mu}^a(t) = \sum\limits_{\xi \in \mathcal{Q}} \Xi^a\big(V(t,\xi) - V(t,\xi-1), \tilde{\mu}^a(t)\big)\tilde{m}(t,\xi)\mathbb{I}(\xi > -Z) \\ \tilde{\mu}^b(t) = \sum\limits_{\xi \in \mathcal{Q}} \Xi^b\big(V(t,\xi) - V(t,\xi+1), \tilde{\mu}^b(t)\big)\tilde{m}(t,\xi)\mathbb{I}(\xi < Z) \end{cases} \tag{B.2.1}$$

Then $V(t,q)$ is the value function $V_t(\tilde{\boldsymbol{\delta}}, \tilde{\boldsymbol{M}})$ of representative market maker's optimization problem defined in (4.1.10), with the population's quoting strategy $\tilde{\boldsymbol{\delta}}$ defined by functions Ξ^a, Ξ^b:

$$\begin{aligned} \tilde{\delta}^a(t,q) &= \Xi^a\Big(V(t,q) - V(t,q-1), \tilde{\mu}^a(t)\Big) \\ \tilde{\delta}^a(t,q) &= \Xi^b\Big(V(t,q) - V(t,q+1), \tilde{\mu}^b(t)\Big) \end{aligned} \tag{B.2.2}$$

$\tilde{\mu}^a(t), \tilde{\mu}^b(t)$ are defined by solving the last 2 equations in $(B.2.1)$, which yield unique solution from Lemma B.1.2.

To show above, assume that the representative market maker takes an arbitrary quoting strategy $\boldsymbol{\delta} \in \mathcal{A}_t^T$, and that her controlled inventory process is $(q_s^{\boldsymbol{\delta}})_{s\in[t,T]}$ with $q_t^{\boldsymbol{\delta}} = q$ at time t.

Appying Itô's formula to function $e^{-rt}V(t,q)$ on the interval $[t,T]$ we obtain

$$
\begin{aligned}
e^{-rT}V(T,q_T^{\boldsymbol{\delta}}) &= e^{-rt}V(t,q) + \int_t^T e^{-rs}\Big(\partial_s V(s,q_s^{\boldsymbol{\delta}}) - rV(s,q_s^{\boldsymbol{\delta}})\Big)ds \\
&\quad + \int_t^T e^{-rs}\left[V(s,q_s^{\boldsymbol{\delta}}-1) - V(s,q_s^{\boldsymbol{\delta}})\right]N^a(ds) \\
&\quad + \int_t^T e^{-rs}\left[V(s,q_s^{\boldsymbol{\delta}}+1) - V(s,q_s^{\boldsymbol{\delta}})\right]N^b(ds) \\
&= e^{-rt}V(t,q) + \int_t^T e^{-rs}\Big(\partial_s V(s,q_s^{\boldsymbol{\delta}}) - rV(s,q_s^{\boldsymbol{\delta}})\Big)dt \\
&\quad + \lambda_a\int_t^T e^{-rs}\left[V(s,q_s^{\boldsymbol{\delta}}-1) - V(s,q_s^{\boldsymbol{\delta}})\right]f_a\big(\delta^a(s,q_s^{\boldsymbol{\delta}}),\tilde{\mu}^a(s)\big)\mathbb{I}(q_s^{\boldsymbol{\delta}} > -Z)ds \\
&\quad + \lambda_b\int_t^T e^{-rs}\left[V(s,q_s^{\boldsymbol{\delta}}+1) - V(s,q_s^{\boldsymbol{\delta}})\right]f_b\big(\delta^b(s,q_s^{\boldsymbol{\delta}}),\tilde{\mu}^b(s)\big)\mathbb{I}(q_s^{\boldsymbol{\delta}} < Z)ds \\
&\quad + \int_t^T e^{-rs}\left[V(s,q_s^{\boldsymbol{\delta}}-1) - V(s,q_s^{\boldsymbol{\delta}})\right]\tilde{N}^a(ds) \\
&\quad + \int_t^T e^{-rs}\left[V(s,q_s^{\boldsymbol{\delta}}+1) - V(s,q_s^{\boldsymbol{\delta}})\right]\tilde{N}^b(ds) \qquad \text{(B.2.3)}
\end{aligned}
$$

where $\tilde{N}^a, \tilde{N}^b$ are the compensated processes of N^a, N^b.

From Assumption 4.1.2, $f_a(\delta,\mu) < \Lambda(\delta)$, $f_b(\delta,\mu) < \Lambda(\delta)$, and $\Lambda(\delta)$ is monotonically decreasing on $\mathbb{R}$. Since centered quotes are bounded from below by $-\delta_\infty$ and the function $V(t,q)$ is bounded on its domain $[0,T]\times\mathcal{Q}$, let $V_{max} = \sup_{(t,q)\in[0,T]\times\mathcal{Q}} |V(t,q)|$, we can deduce

$$
\begin{aligned}
&\left|\mathbb{E}\int_t^T e^{-rs}\left[V(s,q_s^{\boldsymbol{\delta}}-1) - V(s,q_s^{\boldsymbol{\delta}})\right]f_a\big(\delta^a(s,q_s^{\boldsymbol{\delta}}),\tilde{\mu}^a(s)\big)\mathbb{I}(q_s^{\boldsymbol{\delta}} > -Z)ds\right| \\
&\leq 2V_{max}\cdot\mathbb{E}\left[\int_0^T e^{-rs}\Lambda(\delta^a(s,q_s^{\boldsymbol{\delta}}))ds\right] \leq 2V_{max}\cdot\mathbb{E}\left[\int_0^T e^{-rs}\Lambda(-\delta_\infty)ds\right] \\
&\leq 2V_{max}\Lambda(-\delta_\infty)\int_0^\infty e^{-rt}dt < \infty \qquad \text{(B.2.4)}
\end{aligned}
$$

Similarly

$$
\left|\mathbb{E}\int_t^T e^{-rs}\left[V(s,q_s^{\boldsymbol{\delta}}+1) - V(s,q_s^{\boldsymbol{\delta}})\right]f_b\big(\delta^b(s,q_s^{\boldsymbol{\delta}}),\tilde{\mu}^b(s)\big)\mathbb{I}(q_s^{\boldsymbol{\delta}} < Z)ds\right| < \infty \qquad \text{(B.2.5)}
$$

From the boundedness of the function $V(t,q)$ and the finiteness result in Proposition 4.1.4, $\tilde{N}^a, \tilde{N}^b$ are martingales, hence

$$
\begin{aligned}
&\mathbb{E}\left[\int_t^T e^{-rs}\left[V(s,q_s^{\boldsymbol{\delta}}-1)-V(s,q_s^{\boldsymbol{\delta}})\right]\tilde{N}^a(ds)\right]\\
&=\mathbb{E}\left[\int_t^T e^{-rs}\left[V(s,q_s^{\boldsymbol{\delta}}+1)-V(s,q_s^{\boldsymbol{\delta}})\right]\tilde{N}^b(ds)\right]\\
&=0
\end{aligned}
\tag{B.2.6}
$$

Therefore we can take expectation on both sides of (B.2.3), and obtain

$$
\begin{aligned}
&\mathbb{E}\left[e^{-rT}V(T,q_T^{\boldsymbol{\delta}})\right]=e^{-rt}V(t,q)+\mathbb{E}\left[\int_t^T e^{-rs}\Big(\partial_s V(s,q_s^{\boldsymbol{\delta}})-rV(s,q_s^{\boldsymbol{\delta}})-\psi(q_s^{\boldsymbol{\delta}})\Big)ds\right]\\
&+\lambda_a\mathbb{E}\int_t^T e^{-rs}\left[V(s,q_s^{\boldsymbol{\delta}}-1)-V(s,q_s^{\boldsymbol{\delta}})\right]f_a\big(\delta^a(s,q_s^{\boldsymbol{\delta}}),\tilde{\mu}^a(s)\big)\mathbb{I}(q_s^{\boldsymbol{\delta}}>-Z)ds\\
&+\lambda_b\mathbb{E}\int_t^T e^{-rs}\left[V(s,q_s^{\boldsymbol{\delta}}+1)-V(s,q_s^{\boldsymbol{\delta}})\right]f_b\big(\delta^b(s,q_s^{\boldsymbol{\delta}}),\tilde{\mu}^b(s)\big)\mathbb{I}(q_s^{\boldsymbol{\delta}}<Z)ds\\
&+\mathbb{E}\left[\int_t^T e^{-rs}\psi(q_s^{\boldsymbol{\delta}})ds\right]
\end{aligned}
\tag{B.2.7}
$$

Since $V(t,q)$ solves the HJB equation (B.2.1) and the quoting strategy $\boldsymbol{\delta}$ is arbitrary, we obtain the following.

$$
\begin{aligned}
&\partial_t V-rV-\psi(q)\\
&+\lambda_a\left[\delta^a(s,q_s^{\boldsymbol{\delta}})+V(s,q_s^{\boldsymbol{\delta}}-1)-V(s,q_s^{\boldsymbol{\delta}})\right]f_a\big(\delta^a(s,q_s^{\boldsymbol{\delta}}),\tilde{\mu}^a(s)\big)\mathbb{I}(q_s^{\boldsymbol{\delta}}>-Z)\\
&+\lambda_b\left[\delta^b(s,q_s^{\boldsymbol{\delta}})+V(s,q_s^{\boldsymbol{\delta}}+1)-V(s,q_s^{\boldsymbol{\delta}})\right]f_b\big(\delta^b(s,q_s^{\boldsymbol{\delta}}),\tilde{\mu}^b(s)\big)\mathbb{I}(q_s^{\boldsymbol{\delta}}<Z)\leq 0
\end{aligned}
\tag{B.2.8}
$$

Combining (B.2.7) and (B.2.8), we obtain

$$
\begin{aligned}
\mathbb{E}\left[e^{-rT}V(T,q_T^{\boldsymbol{\delta}})\right]\leq&e^{-rt}V(t,q)+\mathbb{E}\left[\int_t^T e^{-rs}\psi(q_s^{\boldsymbol{\delta}})ds\right]\\
&-\mathbb{E}\Big[\int_t^T e^{-rs}\Big(\delta^a(s,q_s^{\boldsymbol{\delta}})f_a(\delta^a(s,q_s^{\boldsymbol{\delta}}),\tilde{\mu}^a(s))\mathbb{I}(q_s^{\boldsymbol{\delta}}>-Z)ds\Big]\\
&-\mathbb{E}\Big[\int_t^T e^{-rs}\Big(\delta^b(s,q_s^{\boldsymbol{\delta}})f_b(\delta^b(s,q_s^{\boldsymbol{\delta}}),\tilde{\mu}^b(s))\mathbb{I}(q_s^{\boldsymbol{\delta}}<Z)ds\Big]
\end{aligned}
\tag{B.2.9}
$$

Combining (B.2.9) with the terminal condition $V(T,q)=\phi(q)$, We thereby

obtain

$$\begin{aligned} V(t,q) \geq & - e^{-r(T-t)}\mathbb{E}\Big[\phi(q_T^{\boldsymbol{\delta}})\Big] \\ & + \mathbb{E}\Big[\int_t^T e^{-r(s-t)}\Big(\delta^a(s,q_s^{\boldsymbol{\delta}})f_a(\delta^a(s,q_s^{\boldsymbol{\delta}}),\tilde{\mu}^a(s))\mathbb{I}(q_s^{\boldsymbol{\delta}} > -Z) \\ & + \delta^b(s,q_s^{\boldsymbol{\delta}})f_b(\delta^b(s,q_s^{\boldsymbol{\delta}}),\tilde{\mu}^b(s))\mathbb{I}(q_s^{\boldsymbol{\delta}} < Z) - \psi(q_s^{\boldsymbol{\delta}})\Big)ds\Big] \end{aligned} \tag{B.2.10}$$

The right-hand side of (B.2.10) is exactly $J_t(\boldsymbol{\delta};\tilde{\boldsymbol{\delta}},\tilde{\boldsymbol{M}})$. Hence we have

$$V(t,q) \geq J_t(\boldsymbol{\delta};\tilde{\boldsymbol{\delta}},\tilde{\boldsymbol{M}}) \tag{B.2.11}$$

The equality is achieved when $\boldsymbol{\delta} = \tilde{\boldsymbol{\delta}}$. By definition $V(t,q) = V_t(\tilde{\boldsymbol{\delta}},\tilde{\boldsymbol{M}})$

Now given $(\tilde{V}^*, \tilde{m}^*)$ classical solution to system (4.1.25), $\tilde{V}^*(t,q)$ is the solution to the HJB equation (B.2.1) with the population distribution $\tilde{\boldsymbol{M}}^*(t)$ in t and the population quoting strategy $\tilde{\boldsymbol{\delta}}^*$ defined by (4.1.27)-(4.1.28). From previous arguments we have

$$\tilde{V}^*(t,q) = V_t(\tilde{\boldsymbol{\delta}}^*,\tilde{\boldsymbol{M}}^*) = \sup_{\boldsymbol{\delta}\in\mathcal{A}_t^T} J_t(\boldsymbol{\delta};\tilde{\boldsymbol{\delta}}^*,\tilde{\boldsymbol{M}}^*) \tag{B.2.12}$$

The equality is achieved when the representative market maker quotes $\boldsymbol{\delta} = \tilde{\boldsymbol{\delta}}^*$.

Furthermore, we see that last 2 equations in system (4.1.25) are the Kolmogorov forward equation of the controlled process $q_t^{\tilde{\boldsymbol{\delta}}^*}$, hence the solution $\tilde{m}^*(t,q)$ defines the distribution of $q_t^{\tilde{\boldsymbol{\delta}}^*}$. Therefore, by Definition 4.1.6 the $\tilde{\boldsymbol{M}}^*$ is a mean field Nash equilibrium, and $\tilde{\boldsymbol{\delta}}^*$ is the associated mean field quoting strategy. □

B.3 A Uniqueness Condition for Nash Equilibrium

In the following proposition, we index the space $\mathbb{R}^{2H+1}$ by the ordered finite set $\{-H, -H+1, \ldots, H\}$. A similar index convention is used for the matrices $\mathbb{R}^{(2H+1)\times(2H+1)}$. Define, for $(u,m) \in \mathbb{R}^{2H+1} \times \mathcal{S}^{\mathcal{Q}} \to \mathbb{R}^{(6H+2)\times(6H+2)}$

$$M^a(u,m) = \begin{bmatrix} A^a(u,m) & \frac{1}{2}C^a(u,m)^T & \frac{1}{2}B^a(u,m) \\ \frac{1}{2}C^a(u,m) & D^a(u,m) & \frac{1}{2}E^a(u,m) \\ \frac{1}{2}B^a(u,m)^T & \frac{1}{2}E^a(u,m)^T & 0 \end{bmatrix} \tag{B.3.1}$$

$$M^b(u,m) = \begin{bmatrix} A^b(u,m) & \frac{1}{2}C^b(u,m)^T & \frac{1}{2}B^b(u,m) \\ \frac{1}{2}C^b(u,m) & D^b(u,m) & \frac{1}{2}E^b(u,m) \\ \frac{1}{2}B^b(u,m)^T & \frac{1}{2}E^b(u,m)^T & 0 \end{bmatrix} \tag{B.3.2}$$

where

$A^a : \mathbb{R}^{2H+1} \times \mathcal{S}^{\mathcal{Q}} \to \mathbb{R}^{(2H+1)\times(2H+1)}$:

$$A^a_{q,\xi}(u,m) = \begin{cases} 0, & q = -Z, \xi \in \mathcal{Q} \\ -\frac{\partial \mathcal{H}^a}{\partial \mu}\big(u_q - u_{q-1}, \mu^a(u,m)\big)\frac{\partial \mu^a}{\partial m_\xi}(u,m), & q \in \mathcal{Q}\backslash\{-Z\}, \xi \in \mathcal{Q} \end{cases}$$

$B^a : \mathbb{R}^{2H+1} \times \mathcal{S}^{\mathcal{Q}} \to \mathbb{R}^{(2H+1)\times(2H+1)}$:

$$B^a_{q,\xi}(u,m) = \begin{cases} 0, & q = -Z, \xi \in \mathcal{Q} \\ -\frac{\partial \mathcal{H}^a}{\partial \mu}\big(u_q - u_{q-1}, \mu^a(u,m)\big)\frac{\partial \mu^a}{\partial u_\xi}(u,m), & q \in \mathcal{Q}\backslash\{-Z\}, \xi \in \mathcal{Q} \end{cases}$$

$C^a : \mathbb{R}^{2H+1} \times \mathcal{S}^{\mathcal{Q}} \to \mathbb{R}^{2H\times(2H+1)}$:

$$C^a_{q,\xi}(u,m) = m_q \frac{\partial^2 \mathcal{H}^a}{\partial \mu \partial p}\big(u_q - u_{q-1}, \mu^a(u,m)\big)\frac{\partial \mu^a}{\partial m_\xi}(u,m), \quad q \in \mathcal{Q}\backslash\{-Z\}, \xi \in \mathcal{Q}$$

$D^a : \mathbb{R}^{2H+1} \times \mathcal{S}^{\mathcal{Q}} \to \mathbb{R}^{2H\times 2H}$:

$$D^a_{q,\xi}(u,m) = m_q \frac{\partial^2 \mathcal{H}^a}{\partial p^2}\big(u_q - u_{q-1}, \mu^a(u,m)\big)\mathbb{I}(q=\xi), \quad q \in \mathcal{Q}\backslash\{-Z\}, \xi \in \mathcal{Q}\backslash\{-Z\}$$

$E^a : \mathbb{R}^{2H+1} \times \mathcal{S}^{\mathcal{Q}} \to \mathbb{R}^{2H\times(2H+1)}$:

$$E^a_{q,\xi}(u,m) = m_q \frac{\partial^2 \mathcal{H}^a}{\partial \mu \partial p}\big(u_q - u_{q-1}, \mu^a(u,m)\big)\frac{\partial \mu^b}{\partial \mu_\xi}(u,m), \quad q \in \mathcal{Q}\backslash\{-Z\}, \xi \in \mathcal{Q}$$

$A^b : \mathbb{R}^{2H+1} \times \mathcal{S}^{\mathcal{Q}} \to \mathbb{R}^{(2H+1)\times(2H+1)}$:

$$A^b_{q,\xi}(u,m) = \begin{cases} -\frac{\partial \mathcal{H}^b}{\partial \mu}\big(u_q - u_{q+1}, \mu^b(u,m)\big)\frac{\partial \mu^b}{\partial m_\xi}(u,m), & q \in \mathcal{Q}\backslash\{Z\}, \xi \in \mathcal{Q} \\ 0, & q = Z, \xi \in \mathcal{Q} \end{cases}$$

$B^b : \mathbb{R}^{2H+1} \times \mathcal{S}^{\mathcal{Q}} \to \mathbb{R}^{(2H+1)\times(2H+1)}$:

$$B^b_{q,\xi}(u,m) = \begin{cases} -\frac{\partial \mathcal{H}^b}{\partial \mu}\big(u_q - u_{q+1}, \mu^b(u,m)\big)\frac{\partial \mu^b}{\partial u_\xi}(u,m), & q \in \mathcal{Q}\backslash\{Z\}, \xi \in \mathcal{Q} \\ 0, & q = Z, \xi \in \mathcal{Q} \end{cases}$$

$C^b : \mathbb{R}^{2H+1} \times \mathcal{S}^{\mathcal{Q}} \to \mathbb{R}^{2H\times(2H+1)}$:

$$C^b_{q,\xi}(u,m) = m_q \frac{\partial^2 \mathcal{H}^b}{\partial \mu \partial p}\big(u_q - u_{q+1}, \mu^b(u,m)\big)\frac{\partial \mu^b}{\partial m_\xi}(u,m), \quad q \in \mathcal{Q}\backslash\{Z\}, \xi \in \mathcal{Q}$$

$D^b : \mathbb{R}^{2H+1} \times \mathcal{S}^{\mathcal{Q}} \to \mathbb{R}^{2H\times 2H}$:

$$D^b_{q,\xi}(u,m) = m_q \frac{\partial^2 \mathcal{H}^b}{\partial p^2}\big(u_q - u_{q+1}, \mu^b(u,m)\big)\mathbb{I}(q=\xi), \quad q \in \mathcal{Q}\backslash\{Z\}, \xi \in \mathcal{Q}\backslash\{Z\}$$

$E^b : \mathbb{R}^{2H+1} \times \mathcal{S}^{\mathcal{Q}} \to \mathbb{R}^{2H\times(2H+1)}$:

$$E^b_{q,\xi}(u,m) = m_q \frac{\partial^2 \mathcal{H}^b}{\partial \mu \partial p}\big(u_q - u_{q+1}, \mu^b(u,m)\big)\frac{\partial \mu^b}{\partial \mu_\xi}(u,m), \quad q \in \mathcal{Q}\backslash\{Z\}, \xi \in \mathcal{Q}$$

where $\mu^a(u,m), \mu^b(u,m)$ are the unique fixed points defined in (B.1.2).

Proposition B.3.1 (Uniqueness of Nash equilibrium)**.** *Assume that*

$$M^a(u,m), M^b(u,m)$$

are either positive definite for all $(u,m) \in \mathbb{R}^{2H+1} \times \mathcal{S}^{\mathcal{Q}}$*, or* $M^a(u,m), M^b(u,m)$ *are negative definite for any* $(u,m) \in \mathbb{R}^{2H+1} \times \mathcal{S}^{\mathcal{Q}}$*.*

Then if (V,m) *and* $(\tilde{V},\tilde{m})$ *are solutions to (4.1.25), then it must hold that* $m = \tilde{m}$ *and* $V = \tilde{V}$*.*

Proof. We compute the derivative $\frac{d}{dt}\Big(e^{-rt}(V-\tilde{V})(m-\tilde{m})\Big)$

$$\begin{aligned}
&\frac{d}{dt}\Big(e^{-rt}\Big(V(t,q)-\tilde{V}(t,q)\Big)\Big(m(t,q)-\tilde{m}(t,q)\Big)\Big)\\
=&e^{-rt}\Big(\frac{\partial V}{\partial t}-\frac{\partial \tilde{V}}{\partial t}\Big)(m-\tilde{m})+e^{-rt}(V-\tilde{V})\Big(\frac{\partial m}{\partial t}-\frac{\partial \tilde{m}}{\partial t}\Big)\\
&-re^{-rt}(V-\tilde{V})\Big(m(t,q)-\tilde{m}(t,q)\Big)\\
=&e^{-rt}\Big[-\Big(\mathcal{H}^a\Big(V(t,q)-V(t,q-1),\mu^a\big(V(t,\cdot),m\big)\Big)\mathbb{I}(q>-Z)\\
&-\mathcal{H}^a\Big(\tilde{V}(t,q)-\tilde{V}(t,q-1),\mu^a\big(\tilde{V}(t,\cdot),\tilde{m}\big)\Big)\mathbb{I}(q>-Z)\Big)\\
&-\Big(\mathcal{H}^b\Big(V(t,q)-V(t,q+1),\mu^b\big(V(t,\cdot),m\big)\Big)\mathbb{I}(q<Z)\\
&-\mathcal{H}^b\Big(\tilde{V}(t,q)-\tilde{V}(t,q+1),\mu^b\big(\tilde{V}(t,\cdot),\tilde{m}\big)\Big)\mathbb{I}(q<Z)\Big)\Big]\Big(m(t,q)-\tilde{m}(t,q)\Big)\\
&+e^{-rt}\Big(V(t,q)-\tilde{V}(t,q)\Big)\Big(\Psi\big(m(t,\cdot),V(t,\cdot)\big)-\Psi\big(\tilde{m}(t,\cdot),\tilde{V}(t,\cdot)\big)\Big) \quad \text{(B.3.3)}
\end{aligned}$$

where Ψ is defined in (B.1.24).

We define functions $H^a, H^b : \mathcal{Q} \times \mathbb{R}^{2H+1} \times \mathcal{S}^{\mathcal{Q}} \to \mathbb{R}$:

$$\begin{aligned}
H^a(q,u,m) &= \mathcal{H}^a(u_q - u_{q-1}, \mu^a(u,m))\mathbb{I}(q>-Z)\\
H^b(q,u,m) &= \mathcal{H}^b(u_q - u_{q+1}, \mu^b(u,m))\mathbb{I}(q<Z) \quad \text{(B.3.4)}
\end{aligned}$$

From Lemma B.1.2 without loss of generality we assume $\mu^a(u,m), \mu^b(u,m)$ are C^1 functions in (u,m). We denote their partial derivatives with respect to $u, m \in \mathbb{R}^{2H+1}$ by

$$\nabla_u \mu^k = (\frac{\partial \mu^k}{\partial u_\xi})_{\xi\in\mathcal{Q}}, \nabla_m \mu^k = (\frac{\partial \mu^k}{\partial m_\xi})_{\xi\in\mathcal{Q}}$$

where vectors u, m are indexed by finite set $\mathcal{Q}$.

We also define functions Δ^a, Δ^b using the derivatives $\frac{\partial \mathcal{H}^a}{\partial p}(p,\mu)$, $\frac{\partial \mathcal{H}^b}{\partial p}(p,\mu)$:

$$\begin{aligned}
\Delta^a(q,u,m) &= \frac{\partial \mathcal{H}^a}{\partial p}\Big(u_q - u_{q-1}, \mu^a(u,m)\Big)\mathbb{I}(q>-Z)\\
\Delta^a(q,u,m) &= \frac{\partial \mathcal{H}^b}{\partial p}\Big(u_q - u_{q+1}, \mu^b(u,m)\Big)\mathbb{I}(q<Z) \quad \text{(B.3.5)}
\end{aligned}$$

For simplicity of notation, we apply functions $H^a, H^b, \Delta^a, \Delta^b$ introduced in (B.3.4)-(B.3.5), and we write $V_t := V(t,\cdot) \in \mathbb{R}^{2H+1}, m_t := m(t,\cdot) \in \mathcal{S}^{\mathcal{Q}}$. From (B.1.18) we also have the relation

$$\begin{aligned}\Delta^a(q,u,m) &= -\lambda_a f_a\Big(\Xi^a(u_q - u_{q-1}, \mu^a(u,m)), \mu^a(u,m)\Big)\\ \Delta^b(q,u,m) &= -\lambda_b f_b\Big(\Xi^a(u_q - u_{q+1}, \mu^b(u,m)), \mu^b(u,m)\Big)\end{aligned} \tag{B.3.6}$$

Without ambiguity, we further introduce the following simplified notation: for $k \in \{a,b\}$,

$$\begin{aligned}H_t^k(q) = H^k(q, V_t, m_t), \quad \tilde{H}_t^k(q) = H^k(q, \tilde{V}_t, \tilde{m}_t)\\ \Delta_t^k(q) = \Delta^k(q, V_t, m_t), \quad \tilde{\Delta}_t^k(q) = \Delta^k(q, \tilde{V}_t, \tilde{m}_t)\end{aligned} \tag{B.3.7}$$

Then the right-hand-side of (B.3.3) becomes

$$\begin{aligned}&e^{-rt}\Big[\Big(-\big(H_t^a(q) - \tilde{H}_t^a(q)\big) - \big(H_t^b(q) - \tilde{H}_t^b(q)\big)\Big)\big(m(t,q) - \tilde{m}(t,q)\big)\\ &+ \big(V(t,q) - \tilde{V}(t,q)\big)\Big(-\Delta_t^a(q+1)m(t,q+1) + \tilde{\Delta}_t^a(q+1)\tilde{m}(t,q+1)\\ &- \Delta^b(q-1)m(t,q-1) + \tilde{\Delta}_t^b(q-1)\tilde{m}(t,q-1)\\ &+ \Delta_t^a(q)m(t,q) - \tilde{\Delta}_t^a(q)\tilde{m}(t,q) + \Delta_t^b(q)m(t,q) - \tilde{\Delta}_t^b(q)\tilde{m}(t,q)\Big)\Big]\end{aligned} \tag{B.3.8}$$

On the other hand, using the initial and terminal conditions of (4.1.25) and integrating the left-hand-side of (B.3.3) on $[0,T] \times \mathcal{Q}$, we have

$$\begin{aligned}&\sum_{q\in\mathcal{Q}} \int_0^T \frac{d}{dt}\Big(e^{-rt}\big(V(t,q) - \tilde{V}(t,q)\big)\big(m(t,q) - \tilde{m}(t,q)\big)\Big)dt\\ =&\sum_{q\in\mathcal{Q}}\Big[e^{-rT}\Big(V(T,q) - \tilde{V}(T,q)\Big)\Big(m(T,q) - \tilde{m}(T,q)\Big)\\ &- \Big(V(0,q) - \tilde{V}(0,q)\Big)\Big(m(0,q) - \tilde{m}(0,q)\Big)\Big]\\ =&0\end{aligned} \tag{B.3.9}$$

Therefore after integrating (B.3.8) on $[0,T] \times \mathcal{Q}$ and rearranging the terms

we have

$$
\begin{aligned}
&\int_0^T \sum_{q\in\mathcal{Q}} e^{-rt}\Big[\Big(-\big(H_t^a(q)-\tilde{H}_t^a(q)\big)-\big(H_t^b(q)-\tilde{H}_t^b(q)\big)\Big)\big(m(t,q)-\tilde{m}(t,q)\big)\\
&+\big(V(t,q)-\tilde{V}(t,q)\big)\Big(-\Delta_t^a(q+1)m(t,q+1)+\tilde{\Delta}_t^a(q+1)\tilde{m}(t,q+1)\\
&-\Delta^b(q-1)m(t,q-1)+\tilde{\Delta}_t^b(q-1)\tilde{m}(t,q-1)\\
&+\Delta_t^a(q)m(t,q)-\tilde{\Delta}_t^a(q)\tilde{m}(t,q)+\Delta_t^b(q)m(t,q)-\tilde{\Delta}_t^b(q)\tilde{m}(t,q)\Big)\Big]dt\\
=&\int_0^T \sum_{q\in\mathcal{Q}} e^{-rt}\Big(-\big(H_t^a(q)-\tilde{H}_t^a(q)\big)-\big(H_t^b(q)-\tilde{H}_t^b(q)\big)\Big)\big(m(t,q)-\tilde{m}(t,q)\big)dt\\
&+\int_0^T \sum_{q\in\mathcal{Q}} e^{-rt}\Big(\Delta_t^a(q)m(t,q)-\tilde{\Delta}_t^a(q)\tilde{m}(t,q)\Big)\Big(\big(V(t,q)-V(t,q-1)\big)\\
&-\big(\tilde{V}(t,q)-\tilde{V}(t,q-1)\big)\Big)dt\\
&+\int_0^T \sum_{q\in\mathcal{Q}} e^{-rt}\Big(\Delta_t^b(q)m(t,q)-\tilde{\Delta}_t^b(q)\tilde{m}(t,q)\Big)\Big(\big(V(t,q)-V(t,q+1)\big)\\
&-\big(\tilde{V}(t,q)-\tilde{V}(t,q+1)\big)\Big)dt
\end{aligned}
\tag{B.3.10}
$$

We denote $V^\theta=\tilde{V}+\theta(V-\tilde{V}), m^\theta=\tilde{m}+\theta(m-\tilde{m})$. Applying fundamental theorem of calculus on the terms related to ask side in right-hand-side of (B.3.10) we obtain:

$$
\begin{aligned}
&-\Big(H_t^a(q)-\tilde{H}_t^a(q)\Big)=\\
&\mathbb{I}(q>-Z)\cdot\int_0^1\Big[-\frac{\partial\mathcal{H}^a}{\partial p}\Big(V^\theta(t,q)-V^\theta(t,q-1),\mu^a(V^\theta,m^\theta)\Big)\cdot\\
&\Big(V(t,q)-\tilde{V}(t,q)-V(t,q-1)+\tilde{V}(t,q-1)\Big)\\
&-\frac{\partial\mathcal{H}^a}{\partial\mu}\Big(V^\theta(t,q)-V^\theta(t,q-1),\mu^a(V^\theta,m^\theta)\Big)\Big(\nabla_u\mu^a(V^\theta,m^\theta)\cdot(V(t,\cdot)-\tilde{V}(t,\cdot))\Big)\\
&-\frac{\partial\mathcal{H}^a}{\partial\mu}\Big(V^\theta(t,q)-V^\theta(t,q-1),\mu^a(V^\theta,m^\theta)\Big)\Big(\nabla_m\mu^a(V^\theta,m^\theta)\cdot(m(t,\cdot)-\tilde{m}(t,\cdot))\Big)\Big]d\theta
\end{aligned}
\tag{B.3.11}
$$

$$
\begin{aligned}
&\Delta_t^a(q)m(t,q) - \tilde{\Delta}_t^a(q)\tilde{m}(t,q) = \\
&\int_0^1 \Big\{ \Big(m(t,q) - \tilde{m}(t,q)\Big) \frac{\partial \mathcal{H}^a}{\partial p}\Big(V^\theta(t,q) - V^\theta(t,q-1), \mu^a(V^\theta, m^\theta)\Big) \\
&+ m^\theta(t,q) \frac{\partial^2 \mathcal{H}^a}{\partial p^2}\Big(V^\theta(t,q) - V^\theta(t,q-1), \mu^a(V^\theta, m^\theta)\Big) \cdot \\
&\Big(V(t,q) - \tilde{V}(t,q) - V(t,q-1) + \tilde{V}(t,q-1)\Big) \\
&+ m^\theta(t,q) \frac{\partial^2 \mathcal{H}^a}{\partial \mu \partial p}\Big(V^\theta(t,q) - V^\theta(t,q-1), \mu^a(V^\theta, m^\theta)\Big)\Big(\nabla_u \mu^a(V^\theta, m^\theta) \cdot \big(V(t,\cdot) - \tilde{V}(t,\cdot)\big) \\
&+ \nabla_m \mu^a(V^\theta, m^\theta) \cdot \big(m(t,\cdot) - \tilde{m}(t,\cdot)\big)\Big)\Big\} d\theta \cdot \mathbb{I}(q > -Z)
\end{aligned} \tag{B.3.12}
$$

Similar calculation are conducted for $-\big(H_t^b(q) - \tilde{H}_t^b(q)\big)$ and $\Delta_t^b(q)m(t,q) - \tilde{\Delta}_t^b(q)\tilde{m}(t,q)$. Taking (B.3.11)-(B.3.12) into (B.3.10) we obtain that the right-hand-side of (B.3.10) yields

$$
\begin{aligned}
&\int_0^T \sum_{q\in\mathcal{Q}} e^{-rt} \int_0^1 \Big\{ \Big(m(t,q) - \tilde{m}(t,q)\Big) \Big[- \frac{\partial \mathcal{H}^a}{\partial \mu}\Big(V^\theta(t,q) - V^\theta(t,q-1), \mu^a(V^\theta, m^\theta)\Big) \cdot \\
&\Big(\nabla_u \mu^a(V^\theta, m^\theta) \cdot \big(V(t,\cdot) - \tilde{V}(t,\cdot)\big)\Big) \\
&- \frac{\partial \mathcal{H}^a}{\partial \mu}\Big(V^\theta(t,q) - V^\theta(t,q-1), \mu^a(V^\theta, m^\theta)\Big)\Big(\nabla_m \mu^a(V^\theta, m^\theta) \cdot \big(m(t,\cdot) - \tilde{m}(t,\cdot)\big)\Big)\Big] \\
&+ \Big(V(t,q) - \tilde{V}(t,q) - V(t,q-1) + \tilde{V}(t,q-1)\Big) \cdot \\
&\Big[\Big(m(t,q) - \tilde{m}(t,q)\Big) \frac{\partial \mathcal{H}^a}{\partial p}\Big(V^\theta(t,q) - V^\theta(t,q-1), \mu^a(V^\theta, m^\theta)\Big) \\
&+ m^\theta(t,q) \frac{\partial^2 \mathcal{H}^a}{\partial p^2}\Big(V^\theta(t,q) - V^\theta(t,q-1), \mu^a(V^\theta, m^\theta)\Big) \cdot \\
&\Big(V(t,q) - \tilde{V}(t,q) - V(t,q-1) + \tilde{V}(t,q-1)\Big) \\
&+ m^\theta(t,q) \frac{\partial^2 \mathcal{H}^a}{\partial \mu \partial p}\Big(V^\theta(t,q) - V^\theta(t,q-1), \mu^a(V^\theta, m^\theta)\Big)\Big(\nabla_u \mu^a(V^\theta, m^\theta) \cdot (V(t,\cdot) - \tilde{V}(t,\cdot)) \\
&+ \nabla_m \mu^a(V^\theta, m^\theta) \cdot \big(m(t,\cdot) - \tilde{m}(t,\cdot)\big)\Big)\Big]\Big\} d\theta \cdot \mathbb{I}(q > -Z) + \\
&\int_0^T \sum_{q\in\mathcal{Q}} e^{-rt} \int_0^1 \Big\{ \Big(m(t,q) - \tilde{m}(t,q)\Big) \Big[- \frac{\partial \mathcal{H}^b}{\partial \mu}\Big(V^\theta(t,q) - V^\theta(t,q+1), \mu^b(V^\theta, m^\theta)\Big) \cdot \\
&\Big(\nabla_u \mu^b(V^\theta, m^\theta) \cdot (V(t,\cdot) - \tilde{V}(t,\cdot))\Big) \\
&- \frac{\partial \mathcal{H}^b}{\partial \mu}\Big(V^\theta(t,q) - V^\theta(t,q+1), \mu^b(V^\theta, m^\theta)\Big)\Big(\nabla_m \mu^b(V^\theta, m^\theta) \cdot (m(t,\cdot) - \tilde{m}(t,\cdot))\Big)\Big] \\
&+ \Big(V(t,q) - \tilde{V}(t,q) - V(t,q+1) + \tilde{V}(t,q+1)\Big) \cdot \\
&\Big[\Big(m(t,q) - \tilde{m}(t,q)\Big) \frac{\partial \mathcal{H}^b}{\partial p}\Big(V^\theta(t,q) - V^\theta(t,q+1), \mu^b(V^\theta, m^\theta)\Big) \\
&+ m^\theta(t,q) \frac{\partial^2 \mathcal{H}^b}{\partial p^2}\Big(V^\theta(t,q) - V^\theta(t,q+1), \mu^b(V^\theta, m^\theta)\Big) \cdot \\
&\Big(V(t,q) - \tilde{V}(t,q) - V(t,q+1) + \tilde{V}(t,q+1)\Big) \\
&+ m^\theta(t,q) \frac{\partial^2 \mathcal{H}^b}{\partial \mu \partial p}\Big(V^\theta(t,q) - V^\theta(t,q+1), \mu^b(V^\theta, m^\theta)\Big)\Big(\nabla_u \mu^b(V^\theta, m^\theta) \cdot (V(t,\cdot) - \tilde{V}(t,\cdot)) \\
&+ \nabla_m \mu^b(V^\theta, m^\theta) \cdot (m(t,\cdot) - \tilde{m}(t,\cdot))\Big)\Big]\Big\} d\theta \cdot \mathbb{I}(q < Z) \\
=& \int_0^T \sum_{q\in\mathcal{Q}} e^{-rt} \Big[\int_0^1 P_a(t)^T \cdot M^a\Big(V^\theta(t,\cdot), m^\theta(t,\cdot)\Big) \cdot P_a(t) d\theta \\
&+ \int_0^1 P_b(t)^T \cdot M^b\Big(V^\theta(t,\cdot), m^\theta(t,\cdot)\Big) \cdot P_b(t) d\theta \Big]
\end{aligned}
\tag{B.3.13}
$$

where $P_a(t), P_b(t)$ are vectors with values in $\mathbb{R}^{6H+2}$ defined below

$$\begin{aligned}
P_a(t) =& \Big[\Big(m(t,\xi)-\tilde{m}(t,\xi)\Big)_{\xi\in\mathcal{Q}}, \Big(V(t,\xi)-\tilde{V}(t,\xi)-V(t,\xi-1)+\tilde{V}(t,\xi-1)\Big)_{\xi\in\mathcal{Q}\setminus\{-Z\}}, \\
& \Big(V(t,\xi)-\tilde{V}(t,\xi)\Big)_{\xi\in\mathcal{Q}}\Big]^T \\
P_b(t) =& \Big[\Big(m(t,\xi)-\tilde{m}(t,\xi)\Big)_{\xi\in\mathcal{Q}}, \Big(V(t,\xi)-\tilde{V}(t,\xi)-V(t,\xi+1)+\tilde{V}(t,\xi+1)\Big)_{\xi\in\mathcal{Q}\setminus\{Z\}}, \\
& \Big(V(t,\xi)-\tilde{V}(t,\xi)\Big)_{\xi\in\mathcal{Q}}\Big]^T
\end{aligned} \tag{B.3.14}$$

with each term in the bracket being a row vector indexed by sorted elements from $\mathcal{Q}$.

Combining (B.3.9) and (B.3.13)

$$\begin{aligned}
&\int_0^T \sum_{q\in\mathcal{Q}} e^{-rt}\Big[\int_0^1 P_a(t)^T \cdot M^a(V^\theta(t,\cdot), m^\theta(t,\cdot)) \cdot P_a(t) d\theta \\
&+ \int_0^1 P_b(t)^T \cdot M^b(V^\theta(t,\cdot), m^\theta(t,\cdot)) \cdot P_b(t) d\theta\Big] = 0
\end{aligned} \tag{B.3.15}$$

From the assumptions on M^a, M^b that they are either simultaneously strictly positive or strictly negative, we obtain that $P_a(t) = 0, P_b(t) = 0$. Therefore, we have $m = \tilde{m}, V = \tilde{V}$. □