London School of Economics and Political Science

Limit Order Markets and Bayesian Learning

Mingwei Lin

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

January 2026

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it). The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent. I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I confirm that Chapter 2 is a joint work with Professor Umut Çetin. Chapter 3 is a joint work with Professor Umut Çetin and Dr Giulia Livieri. Chapter 4 is a joint work with Dr Mauro Bernardi, Dr Giulia Livieri and Dr Luca Maestrini.

Abstract

This thesis explores two distinct research directions in mathematical finance and statistical learning. The first part develops microstructure models of trading in limit order markets with asymmetric information. The second part introduces a Bayesian deep learning framework designed to improve predictive performance and uncertainty quantification.

The first part investigates equilibrium formation, belief updating, and price impact in markets where liquidity suppliers face uncertainty about the presence and number of informed traders. We begin with a one-period model in which equilibrium arises from a fixed-point mapping that accommodates general distributions for asset value and insider count. We show that the resulting market impact function exhibits a power law under fat-tailed fundamentals and a logarithmic form under light-tailed ones, with exponents and coefficients characterised by numerically solvable fixed-point equations. Extending to a multiperiod setting, we study sequential trading with myopic insiders and heavy-tailed t-distributed noise. The dynamic equilibrium is characterised by a recursive system of fixed-point equations, and we prove that liquidity suppliers' beliefs about the asset value converge to the truth over time. The resulting price impact again follows a power-law decay, shaped by the noise tail index, insider competition, and the learning dynamics.

The second part proposes a flexible framework for Bayesian neural networks by introducing probabilistic activation functions and mixture-of-experts architectures. Leveraging a semiparametric variational inference scheme, the model improves prediction accuracy, uncertainty quantification, and interpretability. Applications to real-world regression and classification tasks demonstrate performance gains over standard deep learning models.

Acknowledgements

First of all, I am deeply grateful to my supervisors, Professor Umut Çetin and Dr Giulia Livieri, for their exceptional guidance, patience, and support throughout my PhD journey. Their expertise and encouragement have been instrumental. I have learned an enormous amount from their intellectual clarity, rigour, and generosity with feedback.

I would also like to thank my external supervisors and collaborators, Dr Mauro Bernardi, Dr Luca Maestrini, and Professor Kostas Kardaras, for their consistent encouragement and invaluable advice.

I feel fortunate to have been part of the Department of Statistics at the London School of Economics and Political Science. I thank the faculty and administrative staff for fostering an environment of academic excellence and collegiality. In particular, I appreciate the opportunities I've had to teach, learn, and discuss ideas with students and cohorts alike.

To my families. Heartfelt thanks to my sister Lin Chen and my brother-in-law Yubang He for accommodating me during my PhD years. Your whole family, with your two wonderful children William He and David Chen, brought me warmth, strength, and energy beyond words. Thank you to my father, Guoyou Chen, for your constant support across all these years, even from afar in China. A very special thank you to my mother, Yi Lin, you live on in my heart, and your eternal love and strength have been with me every step of the way.

To my friends, whose laughter, conversations, and presence reminded me that I was never alone. I am especially grateful to my cohorts and companions in this journey: Zetai Cen, Shakeel Gavioli-Akilagun, Ziqing Ho, Camilo Cárdenas Hurtado, Kaixin Liu, Xinhui Liu, Xinyi Liu, Tao Ma, Seyedpouya Mirrezaeiroudaki, Alexandros Pavlis, Pietro Maria Sparago, Di Su, Pingfan Su, Yutong Wang, Xuzhi Yang. Your support, humour, and solidarity made all the difference. I am also deeply thankful to my long-time friends, whether in London or across other cities and countries, and whether in academia or beyond. I feel incredibly lucky to have your unwavering support: Coco Chen, Yudong Chen, Zezhun Chen, Jun Cheng, Qinzi Deng, Hanzhang Fang, Justin Gwee, Xiaoshan Huang, Wenjin Lu, Rennie Sun, Han Wang, Xinya Wang, Yulin Wang, Richard Xia, Anita Xu, Jianing Yuan, Qingran Zhang, David Zheng. You mean more to me than you know. To those whose names I may not have mentioned, thank you as well. I am truly grateful to everyone who has surrounded me with warmth, encouragement, and care.

To Yadh Hafsi, your quiet confidence became the still force behind the steps I took. Thank you, deeply.

Contents

List of Figures

随遇而安。

To the people I love.

Chapter 1

Introduction

This thesis explores two distinct research directions in mathematical finance and statistical learning, presented in two separate parts.

The first part investigates the microstructure of limit order markets with asymmetric information, with particular emphasis on equilibrium characterisation, dynamic belief updating, and the price impact of large orders.

The second part develops a general framework for Bayesian neural networks, incorporating probabilistic activation functions, mixture-of-experts architectures, and a semiparametric variational inference scheme. The proposed approach aims to improve predictive performance and provide more informative uncertainty quantification in deep learning models.

Part I
Limit Order Market with Asymmetric Information

Before the rise of information-based models, early microstructure theories sought to explain the bid–ask spread and price formation through inventory costs, order-processing frictions, and risk management by dealers. Notable examples include Garman (1976), who models market-making as a stochastic process of order arrivals and price adjustments, and Ho and Stoll (1981), who develop a formal utility-maximising dealer model under inventory risk. See also Chapter 2 of O'Hara (1995) for a detailed review of these foundational contributions.

Alongside inventory and transaction-cost approaches, information-based models provide an alternative view that focus on asymmetric information and strategic trading. These models offer theoretical frameworks that explain key phenomena such as price formation, bid–ask spreads, and price dynamics, providing insights on how markets incorporate information over time.

They also allow for the analysis of strategic behaviour among both informed and uninformed traders, highlighting the role of asymmetric information in shaping trading outcomes.

One of the most influential contributions in this area is Kyle (1985), which studies an auction-style market where all orders are cleared at a single price. The model captures how private information affects market liquidity and depth. Informed traders strategically choose their order size to maximize profits while accounting for the impact of their trades on prices. The equilibrium pricing rule, summarised by the *Kyle lambda*, quantifies the price impact per unit of order flow and reflects how information is incorporated into prices through trading. Back (1992) develops a continuous-time analogue, preserving the core informational structure while introducing dynamic trading and gradual information revelation. These foundational works provide deep insight into the informational content of order flow and the resulting price adjustments.

In contrast to these auction-based models, many financial markets today operate via limit order books, where liquidity is provided through bid and ask limit prices. In such settings, the bid-ask spread naturally emerges as a mechanism for liquidity suppliers to protect themselves against *adverse selection* from informed trading. To see this, suppose the market contains the following agents: liquidity suppliers, who provide liquidity by posting quote prices to buy and sell; informed traders, who hold superior information about the trading asset; and uninformed traders, who trade for exogenous reasons. Informed traders buy when they know the asset's current price is too low and sell when they know it is too high. Moreover, they can choose how much to trade based on market conditions, and even whether to trade at all. Liquidity suppliers, however, do not have the flexibility to refrain from trading: they must always provide prices for both buying and selling. As a result, liquidity suppliers understand that when trading against informed traders, they are expected to incur losses. This motivates them to offset such losses by making gains from uninformed traders. These gains arise through the bid–ask spread.

Early important contributions include, for example, Glosten and Milgrom (1985), Easley and O'Hara (1987), and Glosten (1994). Although based on different frameworks, these models all illustrate how asymmetric information affects equilibrium limit prices, and thereby the bid–ask spread. In sequential trading settings, they also characterise how market makers dynamically update their beliefs about the fundamental value based on order flow, and how this belief updating in turn shapes the evolution of prices, essentially solving a Bayesian learning problem. Building on this literature, two main questions emerge and guide our investigation: (i) how to characterise the equilibrium, in terms of the price function and the trading strategies of the different market participants; and (ii) whether the prices updated by market makers converge to full-information values.

However, the existing models that attempt to answer these questions are subject to certain simplifying assumptions. Firstly, they typically restrict themselves to Gaussian settings, for example, assuming that the trading asset value and noise trade volumes are both normally distributed. While such assumptions offer analytical tractability and interpretability, they may not reflect the empirical realities of financial markets, which are often characterised by heavy tails and power-law behaviour; see, for example, Gabaix et al. (2003). Secondly, in sequential settings, orders are typically assumed to arrive in the market one by one, either from informed or uninformed traders, and usually with fixed trade sizes. Compared to batched order models, where orders from different types of traders are observed in aggregate, such sequential models, while more tractable, often lack the richer equilibrium structures and the more realistic decision environment faced by traders in actual markets.

A central feature of real-world limit order markets is *price impact*, the dependence of transaction prices on trade size. This arises because the marginal cost of trading increases with trade size, leading to a *discriminatory pricing* structure.

To illustrate this more concretely, when trading a small quantity, one can transact at the best ask (lowest selling price) or best bid (highest buying price). After such a trade, the best quotes may remain unchanged, and no price impact is incurred. However, if the trade size is large, the available quantity at the best price may be insufficient, and the remaining units must execute against less favourable, second-best prices. In this way, large trades move the market quotes and generate price impact. In the presence of informed traders, large trades become particularly relevant, as these traders aim to fully exploit their superior information to maximise profits. This makes the study of price impact especially relevant when investigating limit order markets with asymmetric information.

The equilibrium in Kyle (1985) is a linear equilibrium, meaning that the market price responds linearly to trading volume, implying a linear price impact function. However, this prediction is not supported by empirical findings, which consistently reflect *nonlinear*, concave price impact. However, the exact functional form remains debated, empirical studies suggest various nonlinear shapes, ranging from power-law to logarithmic impact, see (e.g., Barra, 1997; Potters and Bouchaud, 2003; Almgren et al., 2005; Moro et al., 2009; Bershova and and Rakhlin, 2013; Zarinelli et al., 2015; Tóth et al., 2016).

This thesis takes as its starting point the one-period limit order market model developed in Çetin and Waelbroeck (2024), which incorporates asymmetric information and allows the distribution of the traded asset to be general (subject to boundedness for theoretical justification). Importantly, their framework also characterises the market impact of large orders in equilibrium, contributing to the literature on power-law price impact.

In Chapter 2, we consider a similar one period equilibrium model involving noise traders, informed traders and competitive liquidity suppliers in a limit order market, where liquidity suppliers are uncertain about both the existence and the number of informed traders in the market. We discuss how informed traders, if they exist, seek to exploit their information advantage, and how liquidity suppliers adjust the limit prices in equilibrium based on their beliefs about these informed traders. When the informed trader does not exist, the order is solely submitted by one type of trader, namely noise traders. When informed traders are present, the order is a batched order that aggregates both informed and uninformed trades. This setup can be viewed as a hybrid between sequential models, in which only one type of trader acts per period, and batched models, where aggregate orders from different trader types are observed. Note that, in addition to allowing for uncertainty over the presence of informed traders, our model extends the analysis by accommodating multiple informed traders in equilibrium, capturing their aggregate strategic behaviour and its impact on the market. This introduction of uncertainty about informed traders nonlinearly affects the price impact asymptotics in the market. Although no closed-form expression characterises the exact nature of this nonlinearity, we provide fixed point equations for the exponent in the power-law or logarithmic price impact, which inherently capture the effect of insider uncertainty and can be numerically solved.

The chapter is organised as follows. We begin by introducing the market structure and equilibrium conditions. We then analyse two key cases: a benchmark model with a single uncertain insider, and a general framework where the number of insiders follows a known distribution. The subsequent sections examine asymptotic price impact through fixed point analysis and present numerical illustrations. We conclude with a discussion of implications and directions for future work.

In Chapter 3, we extend the one-period model to a discrete-time multiperiod limit order market model. Moreover, we assume a heavy-tailed location-scale t-distributed noise trading, which accommodates the possibility of large noise trades and enables the analysis of how such trades influence price formation in the market, and thereby the price impact associated with large orders. Furthermore, we show that the prices updated in the market converge to the true value of the trading asset. These results provide concrete answers to the two guiding questions raised earlier: characterise equilibrium trading and pricing under realistic market frictions, and demonstrate that market prices converge to the fundamental value through dynamic learning.

The chapter is organised as follows. We first introduce the multiperiod market structure and formalise the dynamic equilibrium. We then derive the equilibrium strategies and establish posterior consistency of liquidity suppliers' beliefs. The final sections explore asymptotic price impact and trading volume behaviour, followed by numerical results and concluding remarks.

Part II
Bayesian Neural Networks with Probabilistic Activations

Chapter 4 turns to a separate line of research, focusing on the development of Bayesian neural networks (BNNs) with improved predictive performance and more informative uncertainty quantification. We propose a general methodology for variational approximate Bayesian inference in BNNs, introducing *probabilistic activation functions* as a key architectural component.

While deep learning has achieved remarkable success across domains such as image analysis, speech recognition, and natural language processing, standard deep neural networks (DNNs) often exhibit overconfidence in predictions, limited interpretability, and poor uncertainty quantification—limitations that are particularly critical in high-stakes settings like finance or healthcare. BNNs address these issues by treating weights and activations as random variables, thereby enabling calibrated predictive uncertainty and a principled Bayesian learning framework (see, e.g., Ghahramani, 2015; Goan and Fookes, 2020; Jospin et al., 2022).

Our contribution builds on this idea by replacing deterministic activation functions (e.g., ReLU, logistic) with probabilistic counterparts. Inspired by the Pólya-Gamma data augmentation scheme of Polson et al. (2013), the Bayesian interpretation of deep learning in Polson and Sokolov (2017), and the stochastic layer design of Wang et al. (2022), we construct BNNs where nonlinearity is encoded via latent variables sampled from analytically tractable distributions. This yields a flexible probabilistic graphical model with enhanced regularisation and expressivity.

To enable efficient training and inference, we employ *semi-parametric mean-field variational Bayes*, which combines standard variational approximations for model parameters with structured parametric approximations for latent augmentation variables. This hybrid approach allows for scalable inference via optimisation, while retaining posterior flexibility.

We further integrate the framework with a *mixture of experts* (MoE) architecture (Jordan and Jacobs, 1994; Jacobs et al., 1991), where input-dependent gating networks allocate data points to specialised experts, each equipped with its own variational posterior. This improves both predictive performance and computational tractability, especially in settings with heterogeneous or multimodal data.

Our methodology is applicable to a wide range of learning problems. In this thesis, we demonstrate the proposed methodology's effectiveness in regression and classification tasks on real-world datasets. The results show gains in predictive accuracy, improved calibration of uncertainty, and enhanced robustness compared to standard approaches.

Chapter 4 is organised as follows. We begin by introducing the BNN architecture with probabilistic activations. We then describe the semi-parametric variational inference procedure and its integration with the MoE framework. The final sections present empirical evaluations and discuss implications and extensions.

Chapter 2

Equilibrium and Market Impact with Uncertain Insiders in One-period Limit Order Market

2.1 Introduction

One of the central concerns in market microstructure is how prices respond to order flow when some traders may hold private information. Kyle (1985) studies a market comprising a single risk-neutral informed trader, a risk neutral market maker, and a number of non-strategic uninformed traders. A single equilibrium price clears the market in each period, and a linear price impact arises, characterised by the *Kyle's lambda*. Back (1992) extends this model to continuous time and shows that the price impact remains linear, in line with the discrete-time model. In contrast to the auction-style models, where all trades are executed at a single market-clearing price, quote-driven and limit order book (LOB) models introduce price discrimination, as transaction prices vary across orders depending on the trade size, direction and LOB state. Glosten and Milgrom (1985) develops a sequential trade model in which a risk-neutral market maker posts the best bid and ask quotes for trades of fixed size. The resulting positive bid-ask spread reflects adverse selection which depends on the probability of informed trading. A large body of LOB models (e.g., Glosten, 1994; Parlour, 2015; Foucault, 1999; Rosu, 2009; Cont et al., 2014) further examines discriminatory pricing, order placement, adverse selection, and the dynamic evolution of book depth and non-linear price impact. Understanding price impact is crucial, as it links trading activity to price formation and reflects both informational asymmetry and market liquidity. Empirical data provide some evidence on the shape and scale of market impact. There are consistent beliefs and empirical studies that the price impact is a non-linear concave function of trade size. The price impact has been shown from square root to logarithmic (e.g., Barra, 1997; Potters and Bouchaud, 2003; Almgren et al., 2005; Moro et al., 2009; Bershova and and Rakhlin, 2013; Zarinelli et al., 2015; Tóth et al., 2016). While single-price auction models allow for relatively tractable analysis, LOB models are more complex, as equilibrium involves determining the entire pricing rule: a function mapping trade sizes to prices. This requires analysing the shape of the limit order book, making the study of price impact in such settings significantly more challenging. In this paper, we focus on a LOB model and contribute to the price impact literature using a rational expectation equilibrium approach.

In much of the theoretical literature on asymmetric information, the existence and number of informed traders are known to all market participants. These assumptions are embedded in classic frameworks that analyse how private information affects price formation, market efficiency, and equilibrium structure. For example, Grossman and Stiglitz (1980) studied how the number of informed traders affects the informativeness of prices, assuming that the proportion of informed traders is common knowledge among uninformed traders in a rational expectations equilibrium. Similarly, each investor in Hellwig (1980) is certain about the number of other investors. Arguably, this requires an unrealistic degree of knowledge about the economy.

In practice, market participants are often uncertain not only about asset fundamentals, but also about the presence and extent of private information in the market. Easley and O'Hara (1987) propose a sequential trade model in which market makers face uncertainty about whether trades are from informed or uninformed investors, and update their beliefs based on both the direction and discrete size of trades, which are assumed to take one of two possible quantities, leading to the formulation of the probability of informed trading (PIN) as a measure of adverse selection. Easley et al. (2014), by contrast, consider a different form of informational uncertainty within a Gaussian framework, where market participants are not uncertain about the existence of informed traders but about their strategic behaviour – a setting they refer to as opaque trading. Their framework illustrates how ambiguity over trading motives can amplify adverse selection and complicate price discovery. Banerjee and Green (2015) study a market in which uninformed traders are uncertain whether they are trading against rational informed traders or irrational noise traders, with only one type active in the market at any given time. Avery and Zemsky (1998) and Papadimitriou (2023) consider models in which the proportion of informed traders is uncertain and learned over time. Gao et al. (2013) introduces an uncertainty regarding the proportion of informed traders in a rational expectation equilibrium, which brings non-linear price impact in the market. Li (2013) and Back et al. (2013) extend

Kyle (1985) and Back (1992) in a continuous-time setting, and allow the uncertainty of the existence of a monopolistic informed trader.

Recent empirical work further demonstrates a growing interest in quantifying informed trading uncertainty, especially using data-driven methods. Collin-Dufresne and Fos (2015) provide empirical evidence that informed traders strategically time their participation, especially around major information events. Bogousslavsky et al. (2024) develop machine learning methods (e.g., gradient boosted trees) to estimate informed trading intensity from high-frequency data, improving the detection of informational asymmetries.

Our work extends the theoretical framework of Çetin and Waelbroeck (2024), which assumes that the number of informed traders is deterministic. We instead allow for uncertainty over both the existence and the number of informed traders in a one period limit order market model. This additional layer of uncertainty generates genuinely nonlinear effects in equilibrium pricing. In contrast to much of the existing literature, which focuses on single-price auction or quote-driven markets, our model is set in a limit order market that gives rise to a discriminatory pricing rule, whereby transaction prices depend on trade size and direction, capturing key features of modern electronic markets.

Moreover, while many prior models impose simple discrete or Gaussian distributional assumptions for analytical tractability, our framework accommodates general distributions for both the asset values and the number of informed traders. The latter serves to model the uncertainty over the presence and extent of informed trading. Our results show that the shape of market impact is jointly determined by the tail behaviour of asset returns and the distribution of informed trader quantities, which together influence how liquidity suppliers set limit prices in response to potential adverse selection.

Structure of the paper is as follows: We present the market structure and limit order book in the next section. In Section 2.3, we characterise the Nash equilibrium and establish its existence. Section 2.3.1 considers the case of an uncertain monopolistic informed trader, while Section 2.3.2 describes the general setting in which the number of informed traders is uncertain and governed by a known distribution. In Section 2.4, we discuss the market impact asymptotics for large trades in equilibrium. Section 2.5 presents numerical results that illustrate our theoretical findings. Section 2.6 contains the concluding remarks and directions for further study.

2.2 The Structure of the Limit Order Market

We are considering the one-period limit order market model introduced in Çetin and Waelbroeck (2024). We further allow the uncertainty of the asymmetric information in the market. In other words, from the liquidity suppliers' perspectives, there are random number of informed traders trading against their limit orders. All random variables discussed in this section are presumed to be defined on a complete probability space $(\Omega, \mathcal{F}, \mathcal{P})$, and E denotes the expectation operator corresponding to $\mathcal{P}$.

In this one-period limit order market, a single asset is traded through the limit order book. The trades are assumed to take place at $t = 0$ and $t = 1$. The risk-free interest rate is set to be $r = 0$. The fundamental value of the trading asset V will be revealed to the public at time $t = 1$. In the limit order market, four types of agents are interacting:

1. Competitive (infinitely many) liquidity suppliers. They do not know whether there are informed traders in the market or how many there are. Their only shared knowledge concerning the insiders is the distribution of the quantity of informed traders within the market. Based on their information set, they move first and place limit orders to the limit order book. The resulting limit order book is characterised by some not constant and non-decreasing function $h : \mathbb{R} \mapsto \mathbb{R}$. For $x > 0$, $h(x)$ represents the ask side of the book, while for $x < 0$ it represents the bid side. The quantity $h(0+) - h(0-)$ is the jump at zero and corresponds to the bid ask spread of the limit order book. Therefore, a market order of size x traded against the limit order book incurs the cost of

$$\int_0^x h(y)\, dy.$$

   Here, $h(y)$ can be viewed as the marginal price of the y-th share.

2. Noise traders. Their total trading demand is given by $Z \sim N(0, \sigma^2)$. They are non-strategic, i.e. Z is independent of the asset fundamental value V.

3. Risk neutral (symmetric) informed traders N. The number of informed traders N is a discrete random variable with probability function $p(n) := P(N = n)$, $n = 0, 1, \ldots$, $E[N] < \infty$. Moreover, the quantity of informed traders in the market is exogenous. That is, N, V and Z are mutually independent. They (if exist) know the terminal fundamental value of the asset V at time $t = 0$ and choose the (same) trade size x to maximise their expected profit conditional on their private information.

4. A trading desk. Behaving like a broker, it only gathers and executes orders for its clients. Specifically, the trading desk aggregates all orders and trades $Z + Nx$ with the liquidity suppliers.

Moreover, noise traders are assumed to arrive the desk simultaneously and prior to the informed orders (if exist), and the orders are charged proportional to their order size. That is, if U denotes the aggregate demand of the other informed traders, the cost of trading x units for an individual informed trader is

$$\frac{x}{U+x}\int_0^{U+x} h(Z+y)\,dy. \tag{2.2.1}$$

2.3 Insiders' Optimal Strategies and Limit Price in Equilibrium

In this section, our initial focus will be on the case that there is a monopolistic informed trader in the market, while the liquidity suppliers are unsure about it. Following this, we shall proceed to discuss the cases of more general random informed traders.

2.3.1 Uncertain Monopolistic Informed Trading

When there is a monopolistic informed trader who knows the terminal fundamental value of the asset V at time $t = 0$, she seeks to exploit her information advantage by determining the trade size x to maximise her expected profit. However, liquidity suppliers in the market are uncertain about the existence of such an informed trader. As a result, the monopolistic insider is poised to realise greater profitability within the limit order market. Meanwhile, liquidity suppliers shall adjust the bid-ask spread, and consequently, the money they gain from noise traders, according to their beliefs on the presence of the insider.

2.3.1.1 Optimal strategy for the monopolistic informed trader

The optimal trading size x for the monopolistic informed trader is

$$\arg\max_{x\in\mathbb{R}} E\left[Vx - \int_0^x h(Z+u)\,du \middle| V = v\right].$$

Since h is non-decreasing, first-order condition is sufficient for both existence and uniqueness of this optimisation problem. Define

$$F(x) := E\left[h(Z+x)\right] = \int_{-\infty}^{\infty} q(\sigma, z-x)h(z)\,dz,$$

which could be understood as the (expected) marginal cost of the x-th share. Then, given a limit order book h, the optimal strategy for the monopolistic insider is to trade x^* such that $F(x^*) = V$. Moreover, $F(x)$ is strictly increasing and one-to-one, since $h(\cdot)$ is not constant and non-decreasing.

2.3.1.2 The limit price and the equilibrium

The limit prices following Glosten (1994) are given by tail expectations. That is,

$$h(y) := \begin{cases} E[V|Y \geq y], \text{ if } y > 0, \\ E[V|Y < y], \text{ if } y < 0, \end{cases} \tag{2.3.1}$$

where Y is the total demand from insider (if exists) and noise traders. This limit prices guarantee the zero aggregate profit on average for them in the sense that

$$\begin{aligned} &E\Big[\int_0^Y (h(y)-V)\,dy\Big] \\ &= \int_0^{\infty} E[(h(y)-V)1_{[Y\geq y]}]\,dy + \int_{-\infty}^{0} \mathbb{E}[(V-h(y))1_{[Y\leq y]}]\,dy = 0. \end{aligned}$$

Moreover, the bid-ask spread is characterised by $h(0+) - h(0-) := \lim_{y\to 0^+} h(y) - \lim_{y\to 0^-} h(y)$.

These infinitely many liquidity suppliers are uncertain about whether there is an informed trader in the market. Their common beliefs are that the probability of having an insider in the market is p. In other words, $N \sim \text{Bernoulli}(p)$, $0 \leq p \leq 1$. Therefore, for $y > 0$, the marginal ask price is

$$\begin{aligned} h(y) &= \frac{E\left[V1_{[Y\geq y]}\right]}{P(Y\geq y)} = \frac{E\left[V1_{[Nx+Z\geq y]}\right]}{P(Nx+Z\geq y)} \\ &= \frac{(1-p)E\left[V1_{[Z\geq y]}\right] + pE\left[V1_{[x+Z\geq y]}\right]}{(1-p)P(Z\geq y)+pP(x+Z\geq y)}. \end{aligned}$$

In equilibrium, we have $x^* = F^{-1}(V)$. And denote the limit prices in equilibrium by $h^*(y)$, since $V \perp Z$,

$$\begin{aligned} h^*(y) &= \frac{(1-p)E[V]P(Z \geq y) + pE[V1_{[x^*+Z\geq y]}]}{(1-p)P(Z \geq y) + pP(x^* + Z \geq y)} \\ &= \frac{(1-p)E[V]P(Z \geq y) + pE[V1_{[V\geq F(y-Z)]}]}{(1-p)P(Z \geq y) + pP(V \geq F(y-Z))}. \end{aligned}$$

Example 2.3.1. *Let's consider a trading asset where $P(V = 1) = P(V = -1) = \frac{1}{2}$, in line with Example 3.4 in Çetin and Waelbroeck (2024), to see how bid-ask spread and expected profit are affected by the uncertainty surrounding the existence of a monopolistic informed trader.*

It is easily seen from the monotonicity of $F(x)$ that $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = -1$. That is, the insider would trade $x^* = +\infty$ when $V = 1$, and $x^* = -\infty$ when $V = -1$. Therefore, the ask price in limit order book in equilibrium, i.e. $y > 0$

$$\begin{aligned} h^*(y) &= \frac{p\left(P(V=1)\cdot 1 + P(V=-1)\cdot 0\right)}{(1-p)P(Z \geq y) + p\left(P(V=1)\cdot 1 + P(V=-1)\cdot 0\right)} \\ &= \frac{1}{\frac{2(1-p)}{p}P(Z \geq y) + 1}. \end{aligned}$$

Similarly, for $y < 0$, $h^*(y) = -\frac{1}{\frac{2(1-p)}{p}P(Z\geq y)+1}$. As a result, the limit price in equilibrium is written as $h^*(y) = \frac{1}{\frac{2(1-p)}{p}P(Z\geq y)+1}1_{[y>0]} - \frac{1}{\frac{2(1-p)}{p}P(Z\geq y)+1}1_{[y<0]}$. When $p = 0$, signifying no insider in the market, $h(y) = 0$ for all $y \in \mathbb{R}$, implying that there is no need for liquidity suppliers to create a bid-ask spread. On the other hand, when $p = 1$, which indicates that liquidity suppliers know there is a deterministic monopolistic insider in the market, the limit price $h(y) = 1_{[y>0]} - 1_{[y<0]}$ recovers the result for a deterministic monopolistic insider.

Note that the bid-ask spread in equilibrium becomes smaller, and is given by $h^*(0+) - h^*(0-) = 2p$. This indicates that as p increases, i.e. liquidity suppliers are more certain about the presence of the insider, there is a corresponding increase in the bid-ask spread. This increment reflects the liquidity suppliers' strategy to augment their earnings from noise traders as a means of compensating for the heightened risk of incurring greater losses from the monopolistic insider. In other words, the beliefs of liquidity suppliers would alter the bid-ask spread in equilibrium. This is in contrast to the deterministic monopolistic insider case, where the bid-ask spread remains independent of the number of insiders.

Additionally, the expected profit from the monopolistic insider remains finite even if she trades infinite amount. And as expected, it would be greater than the expected profit in the deterministic monopolistic insider case. For example, if $V = 1$, the expected profit in equilibrium of the insider is

$$\begin{aligned}&E\left[\int_0^{x^*} (V - h^*(Z+y))\,dy \middle| V = 1\right]\\ =&\sigma\sqrt{\frac{2}{\pi}} + E\left[\int_0^{\infty}\left(1 - \frac{1}{1+\frac{p}{2(1-p)P(Z\geq y)}}\right)(1_{[Z>-y]} - 1_{[Z<-y]})\,dy\right]\\ \geq&\sigma\sqrt{\frac{2}{\pi}}.\end{aligned}$$

2.3.2 Random Numbers of Informed Traders

Now, suppose that liquidity suppliers in the market are still uncertain about the existence of informed traders. Furthermore, there could be multiple informed traders in the market. Let's assume that all market participants, including liquidity suppliers, informed traders, and noise traders, share a common belief regarding informed traders in the market, characterised by an exogenous random variable $N = 0, 1, 2, \ldots$, $E[N] < \infty$. Consequently, liquidity suppliers provide limit prices based on this understanding of uncertainty. Meanwhile, informed traders will exploit their private information advantage while considering the presence of other insiders.

2.3.2.1 Optimal strategies for informed traders

For an individual informed trader, the number of insiders in the market follows a conditional distribution, conditioned on her own existence. It could be viewed as her understanding of the asymmetry information uncertainty. Specifically,

$$P(N = n | N \geq 1) = \begin{cases} 0, & n = 0, \\ \frac{p(n)}{1-p(0)}, & n = 1, 2, 3, \ldots. \end{cases}$$

Given the cost structure in (2.2.1), an individual informed trader (if he or she exists) hopes to maximise the expected profit, by choosing a trade size x conditioning on their private information on the asset value V and her understanding of uncertainty. That is,

$$\arg\max_{x\in\mathbb{R}} E\left[Vx - \frac{x}{U+x}\int_0^{U+x} h(Z+y)\,dy \middle| V = v, N \geq 1\right], \tag{2.3.2}$$

where U is the random variable, denoting the total trading amount from the other informed traders. The first order condition of the above maximisation problem gives us

$$V = E^v\left[\frac{x}{U+x}h(Z+U+x) + \frac{U}{(U+x)^2}\int_0^{U+x} h(Z+y)\,dy \middle| N \geq 1\right],$$

where E^v is the expectation operator for the informed trader with the private information $V = v$. Assume every insider has symmetric information and is risk neutral, i.e. $U = (N-1)x$, the first order condition associate with the above optimisation problem of an individual insider is given by

$$V = E^v\left[\frac{h(Z+Nx^*)}{N} + \frac{N-1}{N^2x^*}\int_0^{Nx^*} h(Z+u)\,du \middle| N \geq 1\right].$$

To find out the optimal strategy for each of the insider, we need to find out $V = F(x^*)$, where

$$\begin{aligned} F(x) :=& E^v\left[\frac{h(Z+Nx)}{N} + \frac{N-1}{N^2x}\int_0^{Nx} h(Z+u)\,du \middle| N \geq 1\right] \\ =& \sum_{n\geq 1}\frac{p(n)}{1-p(0)}\int_{-\infty}^{\infty}\left\{\frac{1}{n}q(\sigma, z-nx) + \frac{n-1}{n}\bar{q}(\sigma,n,x,z)\right\} h(z)\,dz, \end{aligned} \tag{2.3.3}$$

where $Z \sim N(0,\sigma^2)$, $q(\sigma,\cdot)$ is the probability density function of a mean-zero Gaussian random variable with variance σ^2, and $\bar{q}(\sigma,n,x,z) := 1_{[x\neq 0]}\frac{1}{x}\int_0^x q(\sigma, z-ny)\,dy + 1_{[x=0]}q(\sigma,z)$.

Follows from the monotonicity of h, F is strictly increasing, which allow us to obtain the optimal strategy by inverting F.

Lemma 2.3.1. *Given a non-constant and non-decreasing function h, F defined in (2.3.3) is strictly increasing.*

Proof. To see this, when $x \neq 0$, let's denote $g(x) := \frac{n-1}{n^2x}\int_0^{nx} h(z+u)\,du$. Then, $g'(x) = \frac{n-1}{n^2x^2}\left[nxh(z+nx) - \int_0^{nx} h(z+u)\,du\right] = \frac{n-1}{n^2x^2}\int_0^{nx}\left[h(z+nx) - h(z+u)\right]du > 0$, since h by definition is non-constant and non-decreasing. It will give us the desired strictly increasing property of F. And $F(0) = E^v[h(Z)]$, is interpreted by continuity. □

Therefore, given a h, the optimal strategy for an individual informed trader is $x^* = F^{-1}(V)$.

2.3.2.2 The limit price and the equilibrium

The limit price is in the same form as (2.3.1). But now, the total demand is $Y = Nx + Z$, equals to the total demand from insiders Nx and the total demand from noise traders Z.

Definition 2.3.2. *Suppose the support of the random variable V is (m, M), where $-\infty \leq m < M \leq \infty$. We define the following right-continuous functions:*

$$\begin{aligned}
&\Phi^+(y) := E[V1_{[V>y]}], \Pi^+(y) := P(V > y),\\
&\Phi^-(y) := E[V1_{[V\leq y]}], \Pi^-(y) := P(V \leq y) = 1 - \Pi^+(y).
\end{aligned}$$

Furthermore, define on the support of V:

$$\Psi^{\pm}(y) := \frac{\Phi^{\pm}(y)}{\Pi^{\pm}(y)},$$

so that $\Psi^+(y) = E[V|V > y]$ and $\Psi^-(y) := E[V|V \leq y]$.
Note that $\Phi^+(x-) = E[V1_{[V\geq x]}] = \Phi^+(x)$ for almost all x. Analogously for $\Pi^+(x-)$ and $\Psi^+(x-)$.

Let us consider the ask-side limit price, that is, when $y > 0$,

$$\begin{aligned}
h(y) = E[V|Y \geq y] &= \frac{E[V1_{[Y\geq y]}]}{P(Y \geq y)} = \frac{E[V1_{[Nx+Z\geq y]}]}{P(Nx + Z \geq y)}\\
&= \frac{p(0)E[V1_{[Z\geq y]}] + \sum_{n\geq 0} p(n)1_{[n\neq 0]}E[V1_{[nx+Z\geq y]}]}{p(0)P(Z \geq y) + \sum_{n\geq 0} p(n)1_{[n\neq 0]}P(nx + Z \geq y)}.
\end{aligned}$$

In equilibrium, $F(x^*) = V$, which gives us $x^* = F^{-1}(V)$. The limit price in equilibrium is therefore

$$
\begin{aligned}
h^*(y) &= \frac{p(0)E[V1_{[Z\geq y]}] + \sum_{n\geq 0} p(n)1_{[n\neq 0]}E[V1_{[V\geq F(\frac{y-Z}{n})]}]}{p(0)P(Z\geq y) + \sum_{n\geq 0} p(n)1_{[n\neq 0]}P(V\geq F(\frac{y-Z}{n}))} \\
&= \frac{E[V1_{[N=0]}1_{[Z\geq y]} + 1_{[N\neq 0]}\Phi^+(F(\frac{y-Z}{N}))]}{E[1_{[N=0]}1_{[Z\geq y]} + 1_{[N\neq 0]}\Pi^+(F(\frac{y-Z}{N}))]} \qquad (2.3.4) \\
&= \frac{E[V1_{[N=0]}1_{[B_1\leq 0]} + 1_{[N\neq 0]}\Phi^+(F(\frac{B_1}{N}))]}{E[1_{[N=0]}1_{[B_1\leq 0]} + 1_{[N\neq 0]}\Pi^+(F(\frac{B_1}{N}))]}, \qquad (2.3.5)
\end{aligned}
$$

where $B_t = y + \sigma\beta_t$, β_t is a standard Brownian motion. And B_1 is this Brownian motion at $t = 1$. Similarly for the bid side prices $h^*(y)$, when $y < 0$.
In order to obtain an equation of F in equilibrium, it is convenient to define the mappings associated with F.

Definition 2.3.3. *For any continuous function g, according to (2.3.4), define the mappings*

$$
\begin{aligned}
\phi_g^+(x) :=& \frac{E[V1_{[N=0]}1_{[Z\geq x]} + 1_{[N\neq 0]}\Phi^+(g(\frac{x-Z}{N}))]}{E[1_{[N=0]}1_{[Z\geq x]} + 1_{[N\neq 0]}\Pi^+(g(\frac{x-Z}{N}))]} \\
\phi_g^-(x) :=& \frac{E[V1_{[N=0]}1_{[Z\leq x]} + 1_{[N\neq 0]}\Phi^-(g(\frac{x-Z}{N}))]}{E[1_{[N=0]}1_{[Z\leq x]} + 1_{[N\neq 0]}\Pi^-(g(\frac{x-Z}{N}))]} \\
\phi_g(x) :=& \phi_g^+(x)1_{[x>0]} + \phi_g^-(x)1_{[x<0]}.
\end{aligned}
$$

In this way, the limit prices in equilibrium could be written as $h^*(x) = \phi_F(x)$, $x \in \mathbf{R}$. Now, combine with (2.3.3), the equation of F in equilibrium is obtained.

$$
F(x) = \sum_{n\geq 1} \frac{p(n)}{1-p(0)} \int_{-\infty}^{\infty} \left\{ \frac{1}{n} q(\sigma, z - nx) + \frac{n-1}{n} \bar{q}(\sigma, n, x, z) \right\} \phi_F(z)\, dz, \qquad (2.3.6)
$$

where

$$
\bar{q}(\sigma, n, x, z) := 1_{[x\neq 0]} \frac{1}{x} \int_0^x q(\sigma, z - ny)\, dy + 1_{[x=0]} q(\sigma, z). \qquad (2.3.7)
$$

The preceding derivations show that the existence of equilibrium in the limit order market is equivalent to find F so that (2.3.6) is satisfied. In other words, finding the equilibrium boils down to solve the fixed point mapping problem (2.3.6).

Theorem 2.3.4. *Equilibrium exists if and only if there exists a function $F : \mathbb{R} \mapsto \mathbb{R}$ that satisfies* (2.3.6)*. Given such a solution F, (x^*, ϕ_F) constitutes an equilibrium, where $x^* = F^{-1}(V)$ is the optimal strategy for an individual insider, and ϕ_F is the limit prices in equilibrium according to definition* 2.3.3 *and* (2.3.4)*.*

2.3.2.3 Existence of the equilibrium

We shall show the existence of an equilibrium when the value of the asset V is bounded between $[m, M]$, and when the number of informed traders N is almost surely finite. These two assumptions are imposed for technical reasons related to the existence of equilibrium. We later verify numerically that the equilibrium is unique even when the asset value and the number of insiders are unbounded.

First, we discuss the following important property of the marginal costs F.

Lemma 2.3.5. *Let F be a non-decreasing solution of* (2.3.6) *satisfying the integrable assumption:*

$$\int_{-\infty}^{\infty} \phi_F(z) q(\sigma, z)\, dz < \infty.$$

Then, $F(x)$ is strictly increasing with limits $\lim_{x\to\infty} F(x) = M$ and $\lim_{x\to-\infty} F(x) = m$.

The proof of lemma 2.3.5 is given in Appendix 2.A.

The marginal trading cost F increases strictly with trade size x. Furthermore, F converges to M, the upper bound of the asset's value, when an infinite buy order is placed. Conversely, F approaches the asset's lower bound m with an infinite selling order. This gives rise to the sufficient condition for equilibrium existence in Theorem 2.3.7. Before the formal statement of the equilibrium existence, the next lemma regarding to ϕ_g defined in Definition 2.3.3 is a key foundation.

Lemma 2.3.6. *Let $g : \mathbb{R} \mapsto \mathbb{R}$ be a continuous function. Suppose u^+(resp. u^-) be the unique solution of*

$$u_t + \frac{1}{2}\sigma^2 u_{xx} = 0,$$

$$u(1, x) = E_N\left[1_{[N=0]} 1_{[x\le 0]} + 1_{[N\neq 0]} \Pi^+ \left(g\left(\frac{x}{N}\right)\right)\right],$$

$$(resp.\, u(1, x) = E_N\left[1_{[N=0]} 1_{[x\ge 0]} + 1_{[N\neq 0]} \Pi^- \left(g\left(\frac{x}{N}\right)\right)\right]),$$

where N is random with probability mass function $p(n) := P(N = n)$, $n = 0, 1, \ldots$. and E_N in terminal condition is the expectation with respect to random variable N. Then, the following statements hold:

(1) There exists a solution B which is independent of N on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in(0,1]}, \mathbb{Q})$ to the following SDE:

$$dB_t = dW_t^{\mathbb{Q}} + \sigma^2 \frac{u_x(t, B_t)}{u(t, B_t)}\,dt, \qquad B_0 = x, \tag{2.3.8}$$

where u is either u^+ or u^-, and $W_t^{\mathbb{Q}}$ is a Brownian motion under $\mathbb{Q}$ with $W_0 = 0$.

(2) Suppose $(B, \mathbb{Q}^+)$ and $(B, \mathbb{Q}^-)$ are correspond to the solutions of (2.3.8) when $u = u^+$ and $u = u^-$, respectively, and $E^{\mathbb{Q}}$ stands for the expectation under $\mathbb{Q}$. Then, we have $\phi_g^+(x) = E^{\mathbb{Q}^+}\left[V1_{N=0}1_{B_1\le 0} + 1_{N\neq 0}\Psi^+\left(g\left(\frac{B_1}{N}\right)\right)\right]$ and $\phi_g^-(x) = E^{\mathbb{Q}^-}\left[V1_{N=0}1_{B_1\ge 0} + 1_{N\neq 0}\Psi^-\left(g\left(\frac{B_1}{N}\right)\right)\right]$.

(3) $\phi_g^+(0) > \phi_g^-(0)$.

(4) Suppose further that g is non-decreasing. Then, $\phi_g^\pm$ are non-decreasing, too. Consequently, ϕ_g is non-decreasing. Moreover,

$$\phi_g^+(x) \le E^{\mathbb{Q}^+}\left[V1_{[N=0]}1_{[\sigma W_1^{\mathbb{Q}^+}+x\le 0]} + 1_{[N\neq 0]}\Psi^+\left(g\left(\frac{\sigma W_1^{\mathbb{Q}^+}+x}{N}\right)\right)\right],$$
$$\phi_g^-(x) \ge E^{\mathbb{Q}^-}\left[V1_{[N=0]}1_{[\sigma W_1^{\mathbb{Q}^+}+x\ge 0]} + 1_{[N\neq 0]}\Psi^-\left(g\left(\frac{\sigma W_1^{\mathbb{Q}^-}+x}{N}\right)\right)\right].$$

The detailed proof is provided in Appendix 2.A.

With Lemma 2.3.5 and Lemma 2.3.6 established, we are now ready to prove the existence of an equilibrium. As discussed earlier, this requires demonstrating that the fixed-point equation (2.3.6) admits a solution. Assuming that the asset value lies in a bounded interval $[m, M]$ and that the number of informed traders N is almost surely finite, we apply Schauder's fixed-point theorem to conclude the existence of a function F satisfying the equilibrium condition.

Theorem 2.3.7. *Suppose $-\infty < m < M < \infty$ and that the number of informed traders N is almost surely finite. Then there exists an equilibrium.*

Proof. First, observe that

$$\begin{aligned}
\bar{q}(\sigma,n,x,z) &= 1_{[x\neq 0]}\frac{1}{x}\int_0^x q(\sigma,z-ny)\,dy + 1_{[x=0]}q(\sigma,z),\\
\frac{\partial\bar{q}(\sigma,n,x,z)}{\partial x} &= \frac{1}{x}q(\sigma,z-nx) - \frac{1}{x^2}\int_0^x q(\sigma,z-ny)\,dy\\
&= \frac{q(\sigma,z-nx)-\bar{q}(\sigma,n,x,z)}{x} = \frac{1}{x^2}\int_0^x \{q(\sigma,z-nx)-q(\sigma,z-ny)\}\,dy\\
&= \frac{1}{x^2}\int_0^x u q_u(\sigma,z-nu)\,du.
\end{aligned}$$

Therefore, according to (2.3.6),

$$\begin{aligned}
\left|\frac{dF(x)}{dx}\right| &\le \sum_{n\ge 1}\frac{p(n)}{1-p(0)}\left(\int_{-\infty}^{\infty}|\phi_F(z)|\left\{\frac{q_x(\sigma,z-nx)}{n}+\frac{n-1}{nx^2}\int_0^{|x|}u\,q_u(\sigma,z-nu)\,du\right\}dz\right)\\
&\le (|m|+M)\sum_{n\ge 1}\frac{p(n)}{1-p(0)}\left(\frac{1}{n}\int_{-\infty}^{\infty}|n(z-nx)|\frac{e^{-\frac{(z-nx)^2}{2\sigma^2}}}{\sigma^3\sqrt{2\pi}}\,dz\right.\\
&\qquad\left.+\frac{n-1}{nx^2}\int_0^{|x|}u\,du\int_{-\infty}^{\infty}|n(z-nu)|\frac{e^{-\frac{(z-nx)^2}{2\sigma^2}}}{\sigma^3\sqrt{2\pi}}\,dz\right)\\
&= (|m|+M)\left(\int_{-\infty}^{\infty}|z|\frac{e^{-\frac{z^2}{2\sigma^2}}}{\sigma^3\sqrt{2\pi}}\,dz\right)\sum_{n\ge 1}\frac{n+1}{2}\frac{p(n)}{1-p(0)}\\
&\le (|m|+M)\cdot\frac{1}{\sigma}\sqrt{\frac{1}{2\pi}}\left(E[N\mid N\ge 1]+1\right)<\infty.
\end{aligned}$$

In summary, F will possess a derivative that is bounded by

$$K_0 := (|m|+M)\cdot\frac{1}{\sigma}\sqrt{\frac{1}{2\pi}}\cdot\left(E[N\mid N\ge 1]+1\right).$$

We shall show the existence of a fixed point in the normed space $\mathcal{X} := L^2(\mathbb{R},\mu_0)$, i.e. the space of Borel measurable functions that are square integrable with respect to μ_0, where

$$\mu_0(dx) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx.$$

μ_0 is equivalent to the Lebesgue measure on $\mathbb{R}$. Next define

$$D_0 := \{g | g : \mathbb{R} \mapsto [m, M] \text{ is such that } |g(x) - g(y)| \leq K_0 |x - y|, \forall x, y \in \mathbb{R}\}$$

and let

$$D := \{g \in \mathcal{X} | g = g_0, \mu_0 - a.e. \text{ for some } g_0 \in D_0\}.$$

D is a convex subset of $\mathcal{X}$. Next, define the operator T on $\mathcal{X}$ by

$$Tg(x) := \sum_{n \geq 1} \left\{ \int_{-\infty}^{\infty} \left\{ \frac{1}{n} q(\sigma, z - nx) + \frac{n-1}{n} \bar{q}(\sigma, n, x, z) \right\} \phi_{\bar{g}(z)} \, dz \right\} \frac{p(n)}{1 - p(0)},$$

where $\bar{g} := (g \vee m) \wedge M$ and ϕ_g and $\bar{q}$ are defined as before. Furthermore, for a fixed n, for each x,

$$\left\{ \frac{1}{n} q(\sigma, z - nx) + \frac{n-1}{n} \bar{q}(\sigma, n, x, z) \right\} dz \tag{2.3.9}$$

is a probability measure on $\mathbb{R}$. Also recall that $\sum_{n \geq 1} \frac{p(n)}{1-p(0)} = 1$.

Step 1 (T maps D into itself): By Lemma 2.3.6 (4), $\phi_{\bar{g}(x)} \in [m, M]$. Combined with (2.3.9), Tg is continuous and takes values in $[m, M]$. Moreover, Tg is differentiable with a derivative bounded by K_0. Thus, $Tg \in D_0 \subset D$.

Step 2 (D is compact): Let $(g_m) \subset D$. Then there exists $(g_m^0) \subset D_0$ such that $\mu_0 - a.e.$ we have $g_m = g_m^0$ for each $m \geq 1$. Then by Arzela-Ascoli theorem, there exists a subsequence that converges uniformly on compacts to a continuous function g^0. Without loss of generality let us assume that g_m^0 converges to g^0. Note that $\left|g^0(x) - g^0(y)\right| \leq K_0 |x - y|$ for all $x, y \in \mathbb{R}$, i.e. $g^0 \in D_0$. Finally, as (g_m)s are uniformly bounded and μ_0 is a probability measure, the dominated convergence theorem yields $g_m \to g^0$ in $L^2(\mathbb{R}, \mu_0)$.

Step 3 ($T : D \mapsto D$ is continuous): Suppose $g_m \to g$ in D as $m \to \infty$. In view of the definition of D we may assume without loss of generality that (g_m) are continuous since changing g_m on a Lebesgue null set does not alter the value of Tg_m. By another application of Arzela-Ascoli theorem there exists a subsequence that converges pointwise to some continuous function, which we may identify with g due to the uniqueness of L^2-limits up to a null set. Thus, we may assume g is continuous, too.

Moreover, the same argument shows that every subsequence of g_m has a further subsequence that converges to g pointwise since continuous functions that agree on Lebesgue null sets should agree at every point. Thus, $g_m \to g$ pointwise as $n \to \infty$.
Next, since ϕ_{g_m} is uniformly bounded, the dominated convergence theorem yields

$$\lim_{m\to\infty} Tg_m(x) = \sum_{n\geq 1} \left(\int_{-\infty}^{\infty} dz \left\{ \frac{1}{n} q(\sigma, z - nx) + \frac{n-1}{n} \bar{q}(\sigma, n, x, z) \right\} \lim_{m\to\infty} \phi_{g_m}(z) \right) p(n),$$

Recall that in Lemma 2.3.6 (2), we have

$$\phi_{g_m}^{+}(z) = \frac{E^{\mathbb{P}}[V 1_{[N=0]} 1_{[B_1 \leq 0]} + 1_{[N\neq 0]} \Psi^{+}(g_m(\frac{B_1}{N}))\Pi^{+}(g_m(\frac{B_1}{N}))]}{E^{\mathbb{P}}[1_{[N=0]} 1_{[B_1 \leq 0]} + 1_{[N\neq 0]} \Pi^{+}(g_m(\frac{B_1}{N}))]}.$$

Since Φ^{+} and Π^{+} are continuous except on a Lebesgue null set, the dominated convergence theorem yields

$$\lim_{m\to\infty} E^{\mathbb{P}} \left[1_{[N\neq 0]} \Psi^{+} \left(g_m \left(\frac{B_1}{N} \right) \right) \Pi^{+} \left(g_m \left(\frac{B_1}{N} \right) \right) \right] = E^{\mathbb{P}} \left[1_{[N\neq 0]} \Psi^{+} \left(g \left(\frac{B_1}{N} \right) \right) \Pi^{+} \left(g \left(\frac{B_1}{N} \right) \right) \right],$$
$$\lim_{m\to\infty} E^{\mathbb{P}} \left[1_{[N\neq 0]} \Pi^{+} \left(g_m \left(\frac{B_1}{N} \right) \right) \right] = E^{\mathbb{P}} \left[1_{[N\neq 0]} \Pi^{+} \left(g \left(\frac{B_1}{N} \right) \right) \right].$$

The terms involving $N = 0$,

$$E^{\mathbb{P}} \left[V 1_{[N=0]} 1_{[B_1 \geq z]} \right] \quad \text{and} \quad E^{\mathbb{P}} \left[1_{[N=0]} 1_{[B_1 \geq z]} \right],$$

are constant with respect to m and remain unchanged in the limit.

Thus, we have shown $\lim_{m\to\infty} \phi_{g_m}(z) = \phi_g(z)$ for $z > 0$. Analogous argument yields the convergence for $z \leq 0$, which in turn establishes the pointwise convergence of Tg_m to Tg, i.e. for all $(g_m) \subset D$:

$$\lim_{m\to\infty} g_m(x) = g(x) \Rightarrow \lim_{m\to\infty} Tg_m(x) = Tg(x), \quad x \in \mathbb{R}.$$

Therefore, there exists a $g \in D$ such that $g = Tg$ by Schauder's fixed point theorem. Therefore, there exists an equilibrium. □

2.4 Price Impact Asymptotics

The shape of price impact as a function of trading volume plays important roles in practitioners trading strategies. In this section, we shall discuss the price impact asymptotics shapes based on our equilibrium model.

Definition 2.4.1. *A function $g : (0,\infty) \mapsto (0,\infty)$ is said to be regularly varying of index ρ at ∞ if*

$$\lim_{\lambda\to\infty} \frac{g(\lambda x)}{g(\lambda)} = x^{\rho}, \quad \forall x > 0.$$

Similarly, a function $g : (-\infty, 0) \mapsto (0,\infty)$ is said to be regularly varying of index ρ at $-\infty$ if $g(-x)$ is regularly varying of index ρ at ∞.
For $\rho = 0$, that is, if

$$\lim_{\lambda\to\infty} \frac{g(\lambda x)}{g(\lambda)} = 1, \quad \forall x > 0,$$

g is said to be slowly varying.

Remark 2.4.1. *Roughly speaking, a function of regular variation behaves like a power function asymptotically. On the other hand, for slowly varying functions, interesting examples include:* $\log x$, $(\log x)^{\alpha}$, *for $\alpha \in \mathbb{R}$,* $\log\log x$, *and* $\exp\{(\log x)^{\alpha}\}$, *for $\alpha \in (0,1)$.*

Assumption 2.4.2. Π^{+} *has a continuous derivative on (m, M). Moreover, $\Psi_x^{+}(M) := \lim_{x\to M} \frac{d}{dx}\Psi^{+}(x)$ and $\Psi_x^{-}(m) := \lim_{x\to m} \frac{d}{dx}\Psi^{-}(x)$ exists and equal to the left and right derivatives, respectively, at finite end points. Moreover, $\Psi_x^{+}(M)\Psi_x^{-}(m) \neq 0$.*

Note that $\Psi_x^{+}(M)$ $(resp.\,\Psi_x^{-}(m))$ describes the tail shape of the trading asset distribution V. Specifically, taking the right tail as an example: if $\Psi_x^{+}(M) < 1$, the trading asset is fat-tailed, whereas it has a light tail if $\Psi_x^{-}(M) = 1$. To see this, consider the case where V has a fat right tail with a power-law decay in probability as $V = v$ approaches the upper bound M. That is, $\Pi^{+}(x) = P(V > x) \propto (M - x)^{\alpha}$, $\alpha > 0$. In other words,

$$\frac{\Pi_x^{+}(x)}{\Pi^{+}(x)} \sim -\alpha(M - x)^{-1}, \quad \alpha > 0.$$

Beside, by standard integration by parts, we could see that $\Psi^{+}(x) - x = \frac{\int_x^M \Pi^{+}(y)\,dy}{\Pi^{+}(x)}$. This will further give us

$$\frac{M - \Psi^{+}(x)}{M - x} = \frac{\int_x^M \Pi^{+}(y)\,dy}{(M - x)\Pi^{+}(x)} + 1.$$

Now, apply L'Hopital rule as $x \to M$ to the equality above. It implies that $\Psi_x^+(M) = \frac{\alpha}{1+\alpha} < 1$, which is an indicator of a fat-tailed trading asset distribution.

In the following theorem, we shall show that if the distribution of the trading asset is fat-tailed, the price impact follows a power law, with the exponent satisfying a fix point equation.

Theorem 2.4.3. *Assume Assumption 2.4.2 holds, the number of insiders N is random with probability mass function $p(n)$, $n = 0, 1, 2, \ldots$, and the asset value is supported on a bounded interval $(m, M) \subset \mathbb{R}$. Let F be any solution to the fixed-point equation (2.3.6). If $0 < \Psi_x^+(M) < 1$ and $0 < \Psi_x^-(m) < 1$, then the following hold:*

(1) The function $M - F$ is regularly varying at ∞ with index ρ^+, where ρ^+ is a fixed point of the map $T : [-1, 0) \to [-1, \infty)$ given by

$$T(x) := \frac{\Psi_x^+(M)}{1 - p(0)} \frac{E\left[1_{[N\neq 0]} N^{-(k+1)x}\right]}{E\left[1_{[N\neq 0]} N^{-kx}\right]} E\left[1_{[N\neq 0]} (xN^{x-1} + N^x)\right] - 1, \tag{2.4.1}$$

where $k := \frac{\Psi_x^+(M)}{1-\Psi_x^+(M)}$. That is, $T(\rho^+) = \rho^+$ for some $\rho^+ \in [-1, 0)$.
Moreover, $\Pi^+(F)$ is regularly varying at ∞ with index $k\rho^+$.

(2) The function $F - m$ is regularly varying at $-\infty$ with index ρ^-, where ρ^- is a fixed point of the map $T : [-1, 0) \to [-1, \infty)$ defined by

$$T(x) := \frac{\Psi_x^-(m)}{1 - p(0)} \frac{E\left[1_{[N\neq 0]} N^{-(k+1)x}\right]}{E\left[1_{[N\neq 0]} N^{-kx}\right]} E\left[1_{[N\neq 0]} (xN^{x-1} + N^x)\right] - 1, \tag{2.4.2}$$

where $k := \frac{\Psi_x^-(m)}{1-\Psi_x^-(m)}$. That is, $T(\rho^-) = \rho^-$ for some $\rho^- \in [-1, 0)$.
Moreover, $\Pi^-(F)$ is regularly varying at $-\infty$ with index $k\rho^-$.

The full proof is provided in Appendix 2.B.

This result is consistent with the deterministic case studied in Çetin and Waelbroeck (2024). Specifically, when $N > 1$ is deterministic, the regular variation index of $M - F$ reduces to

$$\rho^+ = \frac{\Psi_x^+(M) - 1}{1 - \frac{\Psi_x^+(M)}{N}},$$

matching the asymptotic derived in their setting. Moreover, in the deterministic N case, one can show that the marginal price function $h(x)$ and the marginal cost function $F(x)$ exhibit similar asymptotic behaviour as trading volume becomes large. Specifically,

$$\lim_{x\to\infty}\frac{M-h(x)}{M-F(x)}=\Psi_x^+(M),$$

indicating that both $h(x)$ and $F(x)$ converge at the same rate as $x \to \infty$, with the decay governed by the operator Ψ_x^+.

We now analyse this limit in the case of random insider presence. In particular, we examine

$$\lim_{x\to\infty}\frac{M-h(x)}{M-F(x)}=\Psi_x^+(M).$$

Now, we look into $\lim_{x\to\infty}\frac{M-h(x)}{M-F(x)}$ in our random N case.

$$\begin{aligned}
\lim_{x\to\infty}\frac{M-h(x)}{M-F(x)} &= \lim_{x\to\infty}\frac{M-\phi_F^+(x)}{M-F(x)} \\
&= \lim_{x\to\infty}\frac{M-\frac{p(0)E[V]P(Z\geq x)+\sum_{n\geq 0}p(n)1_{[n\neq 0]}\int_{-\infty}^{\infty}dy\Phi^+(F(\frac{y}{n}))q(\sigma,x-y)}{p(0)P(Z\geq x)+\sum_{n\geq 0}p(n)1_{[n\neq 0]}\int_{-\infty}^{\infty}du\Pi^+(F(\frac{u}{n}))q(\sigma,x-u)}}{M-F(x)} \\
&= \lim_{x\to\infty}\sum_{n\geq 0}\frac{p(n)1_{[n\neq 0]}\int_{-\infty}^{\infty}dy\frac{\Pi^+(F(xy))}{\Pi^+(F(x))}q(\frac{\sigma}{nx},y-\frac{1}{n})}{\sum_{n\geq 0}p(n)1_{[n\neq 0]}\int_{-\infty}^{\infty}du\frac{\Pi^+(F(xu))}{\Pi^+(F(x))}q(\frac{\sigma}{nx},u-\frac{1}{n})}\frac{M-\Psi^+(F(xy))}{M-F(x)},
\end{aligned}$$

Using the regular variation of $\Pi^+(F(x))$ with index $k\rho^+$ as shown in Theorem 2.4.3, this expression simplifies to

$$\lim_{x\to\infty}\frac{M-h(x)}{M-F(x)}=\sum_{n\geq 1}\frac{p(n)\,n^{-(k+1)\rho^+}}{\sum_{n\geq 1}p(n)\,n^{-k\rho^+}}\Psi_x^+(M)=\Psi_x^+(M)\cdot\frac{E[1_{[N\neq 0]}N^{-(k+1)\rho^+}]}{E[1_{[N\neq 0]}N^{-k\rho^+}]},$$

which appears in the fixed-point condition for ρ^+ in equation (2.4.1). This result implies that even under uncertainty about insider presence, the marginal price function $h(x)$ and the marginal cost function $F(x)$ asymptotically behave similarly as trade size grows large. The difference is captured by a correction factor that depends on the distribution of N.

We are also interested in the asymptotic behaviour of the marginal cost $F(x)$ and the implementation shortfall $IS(x)$. When there is a random number of informed traders N, the implementation shortfall which describes the difference between a 'paper trading' benchmark and the actual trading costs are defined as below.

Definition 2.4.4. *The implementation shortfall associated with an individual insider trading x units is given by*

$$IS(x) := E\left[\frac{1}{Nx}\int_0^{Nx} h(Z+u)\,du \middle| N \geq 1\right].$$

Note that $IS(x)$ is the expected average cost of an individual insider trading x units, given random number of insiders are symmetric, i.e., they all make the same decisions in equilibrium. The expectation is with respect to the total number of inform traders N and the total demand from noise traders Z, where N and Z are independent. Additionally, by changing variable, equivalently, $IS(x) = \frac{1}{x}\int_0^x E[h(Z+Nu)\big|N \geq 1]\,du$. If we use this representation of $IS(x)$, we can derive the asymptotic relationship of $M - IS^*(x)$ and $M - F^*(x)$.

Corollary 2.4.5. *Assume Assumption 2.4.2, $N = 1, 2, 3, ...,$ is random with probability function $p(n)$ and $-\infty < m < M < \infty$, and let (h^*, x^*) be an equilibrium. Suppose that F^* is the fixed-point mapping in (2.3.6). Then,*

$$M - IS^*(x) \sim \frac{\Psi_x^+(M)}{1-p(0)} \frac{E[1_{[N\neq 0]}N^{-(k+1)\rho^+}]}{E[1_{[N\neq 0]}N^{-k\rho^+}]} \frac{E[1_{[N\neq 0]}N^{\rho^+}]}{\rho^+ + 1}(M - F^*(x)), \quad x \to \infty,$$
$$IS^*(x) - m \sim \frac{\Psi_x^-(m)}{1-p(0)} \frac{E[1_{[N\neq 0]}N^{-(k+1)\rho^-}]}{E[1_{[N\neq 0]}N^{-k\rho^-}]} \frac{E[1_{[N\neq 0]}N^{\rho^-}]}{\rho^- + 1}(F^*(x) - m), \quad x \to -\infty,$$

where ρ^+ and ρ^- are the fixed-point in Theorem 2.4.3.

This shows us that the asymptotic form of the marginal cost F will coincide with the implementation shortfall up to a scaling factor.

Furthermore, from Theorem 2.4.3, we know that, once we solve the fixed point equation (2.4.1) and get ρ^+, we have the asymptotic equilibrium distribution of the trading volume of a single informed trader. To be more specific, $\Pi^+(F)$ is regularly varying at ∞ of index $k\rho^+, k = \frac{\Psi_x^+(M)}{1-\Psi_x^+(M)}$. In equilibrium, for each informed trader, the probability that she trades more than $x > 0$ many shares are given by

$$P(x^* > x) = P(F^{-1}(V) > x) = P(V > F(x)) = \Pi^+(F(x)) = x^{k\rho^+}s(x),$$

where s is a slowly varying function.

In addition, the distribution of the aggregate order in equilibrium $Y^* = Nx^* + Z$ is also known. In short, $P(Y^* > y)$ is regularly varying at infinity with the same index.

Corollary 2.4.6. *Given $\Pi^{+}(F)$ is regularly varying at ∞ of index $k\rho^{+}$ as stated in Theorem 2.4.3, the aggregate order in equilibrium $Y^{*} = Nx^{*} + Z$ is regularly varying at ∞ of the same index $k\rho^{+}$, i.e.*

$$P(Y^{*} > y) = y^{k\rho^{+}} s(y), \quad y > 0,$$

where s is a slowly varying function.
Similarly, $\Pi^{-}(F)$ is regularly varying at $-\infty$ of index $k\rho^{-}$, then

$$P(Y^{*} < y) = y^{-k\rho^{-}} s(y), \quad y < 0,$$

where s is a slowly varying function.

The proof is provided in Appendix 2.B.

Recall that $\Psi_x^{+}(M) < 1$ serves as an indicator of fat-tailed behaviour in the asset value distribution. Up to this point, we have shown that when $\Psi_x^{+}(M) < 1$, the equilibrium exhibits corresponding power-law asymptotics in the marginal cost function $F(x)$, marginal price function $h(x)$, implementation shortfall (IS), insider trading volumes, and total market trading volume.

In contrast, when $\Psi_x^{+}(M) = 1$, corresponding to a light-tailed asset value distribution, the asymptotic behaviour of price impact is no longer polynomial. In this regime, the marginal cost function grows logarithmically in trade size. This is formalised in the following theorem.

Theorem 2.4.7. *Assume Assumption 2.4.2, N is random with probability mass function $p(n)$, $n = 0, 1, 2, \ldots$, and $-\infty < m < M < \infty$. Let F be any solution of (2.3.6). When $\Psi_x^{+}(M) = 1$ (resp. $\Psi_x^{-}(m) = 1$), the following statements are valid:*

(1) Suppose that there exists an integer $n \geq 1$ and a real constant $k \in (0, \infty)$ such that

$$\lim_{x \to M} \frac{\Psi^{+}(x) - x}{(M - x)^{n+1}} = \frac{1}{k} \quad (resp. \lim_{x \to m} \frac{\Psi^{-}(x) - x}{(x - m)^{n+1}} = \frac{1}{k}) \tag{2.4.3}$$

Then, $\Pi^{+}(F)$ is regularly varying at ∞ of index $-k\rho$, while $\Pi^{-}(F)$ is regularly varying at $-\infty$ of index $-k\rho$. Here, k is the constant in (2.4.3), and $\rho > 0$ is a fixed point that solves

$$\rho = \frac{1}{k} + \frac{\rho}{1 - p(0)} E\left[1_{[N \neq 0]} \left(logN + \frac{1}{N}\right)\right] - \rho \frac{E[1_{[N \neq 0]} N^{k\rho} logN]}{E[1_{[N \neq 0]} N^{k\rho}]}, \rho > 0$$

(2) More importantly, the following asymptotics holds:

$$M - F(x) \sim (n\rho)^{-\frac{1}{n}}(log x)^{-\frac{1}{n}}, \quad x \to \infty,$$
$$F(x) - m \sim (n\rho)^{-\frac{1}{n}}(log|x|)^{-\frac{1}{n}}, \quad x \to -\infty.$$

The full proof is provided in Appendix 2.B.

Note that when the distribution of informed traders degenerates to a deterministic value, the results above recover the logarithmic price impact for light-tailed asset distributions, as established in Çetin and Waelbroeck (2024).

2.5 Numerical Experiments

In this section, we numerically compute equilibrium outcomes in markets with uncertainty over informed trader participation. We consider two types of distributions for the asset value: a heavy-tailed Student-t distribution and a light-tailed Gaussian distribution. For each case, we explore how the number and uncertainty of informed traders affect the shape of the marginal cost function and price impact.

We examine two distributions for the number of informed traders N:

- In the *two-point distribution*, $N \in \{0, n\}$ for fixed $n \in \{2, 3, 10\}$, with probability mass function
$$P(N = n) = p, \quad P(N = 0) = 1 - p,$$
where $p \in \{0.1, 0.5, 0.9\}$. That is, either no insider is present, or exactly n symmetric insiders participate in the market.

- In the *geometric distribution*, the number of insiders $N \in \mathbb{N}$ follows
$$P(N = n) = (1 - p)^n p, \quad n = 0, 1, 2, \ldots,$$
where $p \in \{0.3, 0.5, 0.9\}$. This introduces a more diffuse form of uncertainty over insider presence and trading intensity, with higher values of p placing more weight on smaller insider counts.

While the existence result in Theorem 2.3.7 requires the boundedness of the asset value distribution, the numerical characterisation of equilibrium ultimately reduces to solving the fixed-point equation (2.3.6). In practice, we find that the mapping remains well-defined and numerically stable even when the asset value follows an unbounded distribution.

The numerical procedure is as follows. We first solve the fixed-point equation (2.3.6) via iteration to obtain the equilibrium marginal cost function $F(x)$. Given this solution, we then compute the corresponding equilibrium limit price function $h(x)$ using (2.3.4). Finally, the bid–ask spread is defined as the jump discontinuity at the origin, i.e., $h(0^+) - h(0^-)$.

In the subsections below, we illustrate the equilibrium marginal cost function $F(x)$ and the associated marginal limit price function $h(x)$ under each combination of asset value distribution and insider participation distribution.

2.5.1 Student-*t* Distributed Asset Value

2.5.1.1 Two-point Distributed Informed Traders

We assume that the fundamental value of the trading asset V follows a Student-t distribution, denoted $V \sim t_\nu(0, \sigma_V)$, where $\nu > 2$ is the degrees of freedom and $\sigma_V > 0$ is the scale parameter. The parameter ν controls the heaviness of the tails: smaller values correspond to fatter tails. The distribution has zero mean, and variance given by

$$\mathrm{Var}(V) = \frac{\nu\sigma_V^2}{\nu - 2}, \quad \text{for } \nu > 2.$$

The induced distributional operators $\Pi^\pm$, $\Phi^\pm$, and $\Psi^\pm$ are defined in Definition 2.3.2, and are computed numerically for the given t-distribution. In the following experiments, we fix the asset value parameters to $\nu = 3$ and $\sigma_V = 1$, and model the noise trades as Gaussian: $Z \sim \mathcal{N}(0, 1)$. The number of informed traders N is distributed according to the two-point distribution described above, with fixed values $n \in \{2, 3, 10\}$ and presence probabilities $p \in \{0.1, 0.5, 0.9\}$.

The resulting equilibrium marginal cost function $F(x)$ and marginal price function $h(x)$ are shown in Figure 2.1. For fixed p, as N increases, the price impact becomes smaller for large orders. This is because increased competition among informed traders leads each insider to trade more aggressively to capture a share of the informational advantage. The resulting increase in aggregate informed order flow improves price informativeness, which in turn reduces the marginal response of prices to additional volume. As a result, the marginal cost function $F(x)$ becomes flatter in the tails, reflecting lower price impact per unit of trade size. This mechanism is consistent with findings in the literature on strategic interaction among multiple informed agents as in Holden and Subrahmanyam (1992); Çetin and Waelbroeck (2024).

In Figure 2.2, we fit a power law to $F(x)$ over the region $x > 1.5$ to highlight the asymptotic behaviour of the marginal cost function, as discussed in Section 2.4. That is, in the regime of large trade sizes, the shape of $F(x)$ exhibits power-law shape. Recall from (2.4.1) that when N follows a two-point distribution as described above, the regularly varying index simplifies to

$$\rho^+ = \frac{\Psi_x^+(M) - 1}{1 - \frac{\Psi_x^+(M)}{N}},$$

which depends only on N and not on the presence probability p. This implies that for fixed N, the tail behaviour of $F(x)$ is asymptotically invariant to p. As seen in Figure 2.1, the curves of $F(x)$ converge to the same slope as $x \to \infty$, confirming the theoretical prediction.

Figure 2.3 illustrates how the bid-ask spread varies with the insider probability p and the insider count N. We observe that the spread increases with both p and N, reflecting greater adverse selection as the likelihood and intensity of informed trading rise.

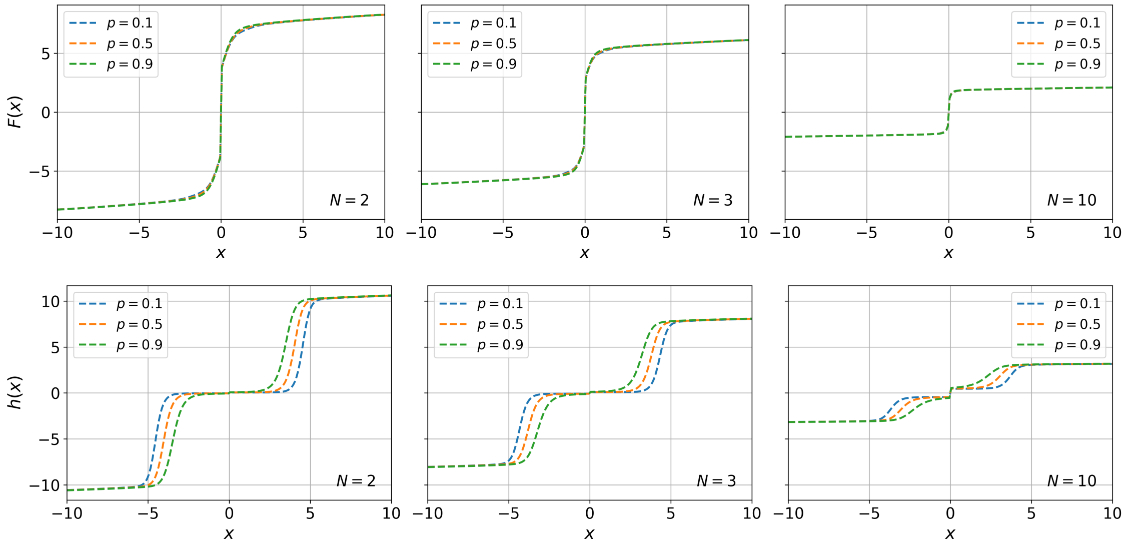


FIGURE 2.1: Equilibrium marginal cost function $F(x)$ (top row) and marginal limit price function $h(x)$ (bottom row) under Student-t distributed asset value with $\nu = 3$ degrees of freedom and scale parameter $\sigma_V = 1$. The noise trade distribution is standard Gaussian, $Z \sim \mathcal{N}(0,1)$. Each column corresponds to a fixed number of insiders $N \in \{2,3,10\}$, and within each panel we vary the probability of insider presence $p \in \{0.1, 0.5, 0.9\}$.

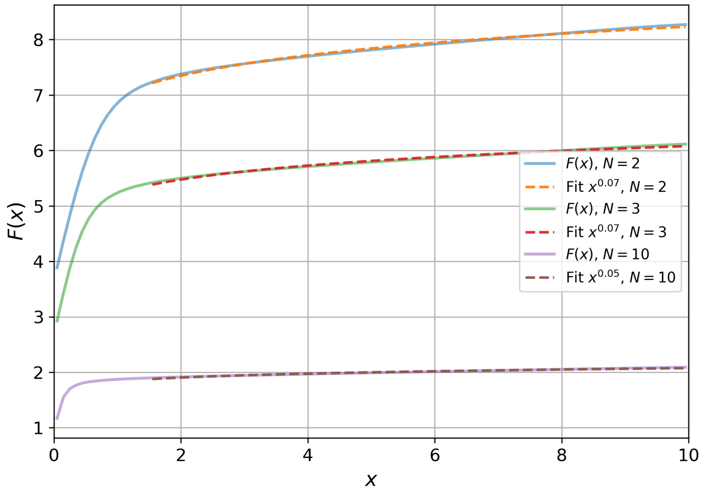


FIGURE 2.2: Power-law fits to the marginal cost function $F(x)$ for large trade sizes $x > 1.5$, under a Student-t distributed asset value with $\nu = 3$ and $\sigma_V = 1$, and standard Gaussian noise $Z \sim \mathcal{N}(0, 1)$. Each curve corresponds to a fixed number of insiders $N \in \{2, 3, 10\}$, with the insider presence probability fixed at $p = 0.9$. Dashed lines show power-law fits of the form $F(x) \sim x^{\rho^+}$.

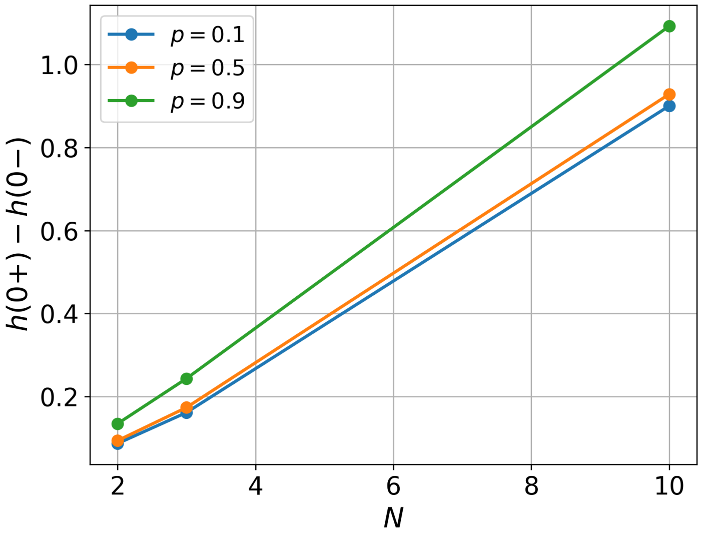


FIGURE 2.3: Bid-ask spread $h(0^+) - h(0^-)$ as a function of insider count $N \in \{2, 3, 10\}$, for various values of the insider presence probability $p \in \{0.1, 0.5, 0.9\}$. The asset value follows a Student-t distribution with $\nu = 3$, $\sigma_V = 1$, and noise trades are standard normal.

2.5.1.2 Geometric Distributed Informed Traders

We retain the same distributional assumptions for the trading asset and noise trades as in the previous subsection: the asset value $V \sim t_\nu(0, \sigma_V)$ with $\nu = 3$, $\sigma_V = 1$, and the noise trades are standard normal $Z \sim \mathcal{N}(0, 1)$.

In this setting, the number of informed traders $N \in \mathbb{N}$ follows a geometric distribution:

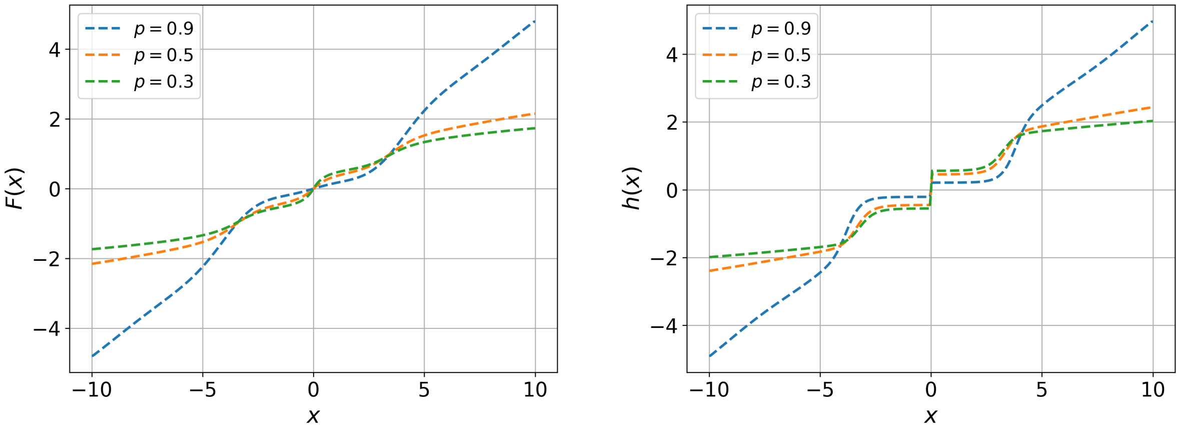
$$P(N = n) = (1 - p)^n p, \quad n = 0, 1, 2, \dots,$$

with geometric parameter $p \in \{0.3, 0.5, 0.9\}$. This specification introduces a more diffuse and persistent uncertainty over insider participation, encompassing both the possibility of no insider activity ($N = 0$) and varying degrees of competition when insiders are present. Smaller values of p place greater probability mass on larger insider counts, resulting in a heavier right tail in the insider distribution and more potential competition among informed traders.

As before, we numerically compute the equilibrium pricing functions $F(x)$ and $h(x)$, and compare their asymptotic behaviour across values of p. The effect of geometric uncertainty in insider participation exhibits notable non-linearities. As the parameter p decreases, the expected number of informed traders increases, thereby intensifying the degree of competition among insiders. This results in a flatter marginal cost function $F(x)$, particularly in the large-volume regime, consistent with the intuition that informed traders engage in more aggressive trading under competitive pressure. This can be observed in Figure 2.4. At the same time, the bid-ask spread becomes wider as shown in Figure 2.5, reflecting the heightened severity of adverse selection faced by liquidity suppliers in the presence of more informed trading. In Figure 2.6, we fit a power law to the marginal cost function $F(x)$ over the region $x > 5$, illustrating the asymptotic behaviour predicted by Theorem 2.4.3.

2.5.2 Gaussian Distributed Asset Value

We now assume that the asset value is Gaussian distributed, denoted by $V \sim \mathcal{N}(0, \sigma_V^2)$. The operators defined in Definition 2.3.2, which characterise this distribution, admit closed-form



(A) Equilibrium marginal cost function $F(x)$.

(B) Equilibrium marginal price function $h(x)$.

FIGURE 2.4: Marginal cost $F(x)$ and marginal price $h(x)$ functions under geometric uncertainty in the number of informed traders. The number of insiders follows a geometric distribution $N \sim \text{Geometric}(p)$ with $p \in \{0.3, 0.5, 0.9\}$, and asset values follow a Student-t distribution $V \sim t_3(0,1)$. Noise trades are Gaussian: $Z \sim \mathcal{N}(0,2)$.

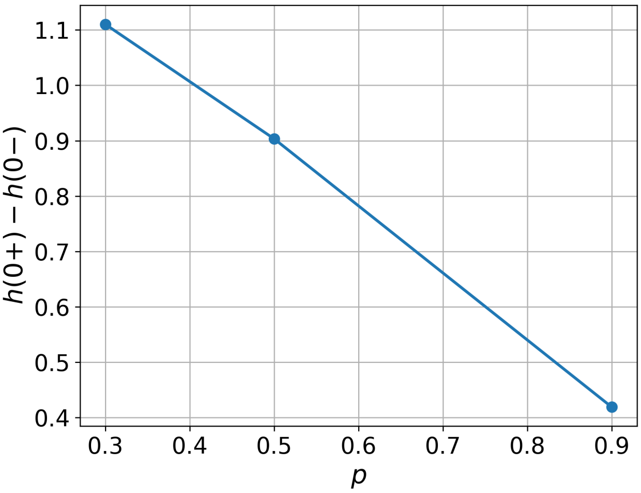


FIGURE 2.5: Bid–ask spread $h(0^+) - h(0^-)$ as a function of the geometric distribution parameter $p \in \{0.3, 0.5, 0.9\}$ for insider participation. The asset value follows a Student-t distribution with $\nu = 3$, $\sigma_V = 1$, and noise trades are standard normal $Z \sim \mathcal{N}(0,1)$.

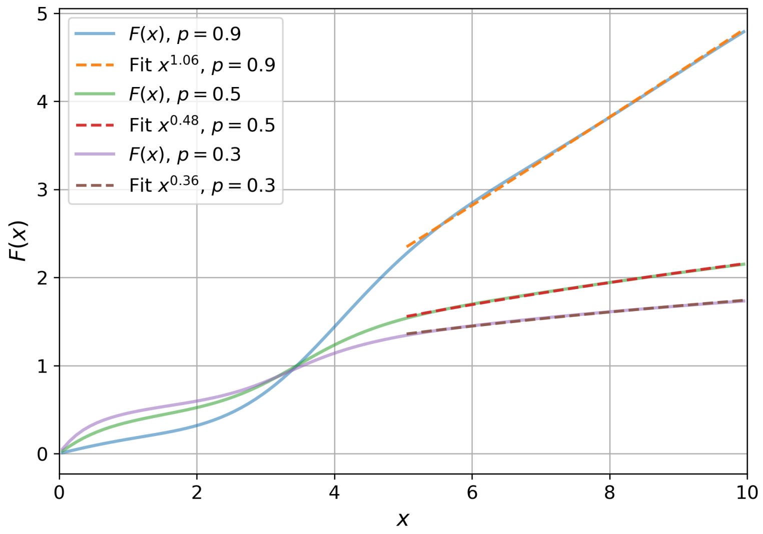


FIGURE 2.6: Power-law fits to the marginal cost function $F(x)$ in the region $x > 1.5$, under a Student-t distributed asset value with $\nu = 3$, $\sigma_V = 1$, and Gaussian noise trades $Z \sim \mathcal{N}(0,1)$. Each curve corresponds to a different value of the geometric insider distribution parameter $p \in \{0.3, 0.5, 0.9\}$.

expressions:

$$\Phi^+(y) = \sqrt{\frac{\sigma^2}{2\pi}} \exp\left(-\frac{y^2}{2\sigma^2}\right), \quad \Pi^+(y) = \frac{1}{2}\operatorname{erfc}\left(\frac{y}{\sqrt{2\sigma^2}}\right),$$
$$\Phi^-(y) = -\sqrt{\frac{\sigma^2}{2\pi}} \exp\left(-\frac{y^2}{2\sigma^2}\right), \quad \Pi^-(y) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{y}{\sqrt{2\sigma^2}}\right)\right].$$

In the following experiments, we fix the asset value variance at $\sigma_V^2 = 2$, and model aggregate noise trades as standard normal, i.e., $Z \sim \mathcal{N}(0,1)$. As before, we vary the insider presence probability parameter $p \in \{0.1, 0.5, 0.9\}$, and repeat the computations for each fixed insider count $N \in \{2, 3, 10\}$.

Figure 2.7 illustrates the equilibrium marginal cost function $F(x)$ and marginal price function $h(x)$ under varying values of p and N, alongside the deterministic benchmark. For a fixed probability p, increasing the number of insiders N leads to flatter profiles of both $F(x)$ and $h(x)$, reflecting reduced price impact due to intensified competition among informed traders. Comparing the uncertain insider setting to the deterministic case with the same N, we observe that uncertainty in insider presence reduces the price impact, resulting in a noticeably flatter limit order book.

Combined with the bid-ask spreads shown in Figure 2.8, we find that for fixed N, the spread increases with the insider probability p, reflecting heightened adverse selection as informed trading becomes more likely. Similarly, for fixed p, the spread increases with larger N, again indicating intensified adverse selection. Notably, in Figure 2.8b, we compare against the deterministic benchmark (dashed line): across all values of N, the spread is strictly smaller under the uncertain insider setting. This demonstrates that when insider presence is uncertain, liquidity suppliers demand less compensation for adverse selection risk, leading to a smaller bid ask spread. Intuitively, uncertainty over whether any informed trader is present dilutes the informational content of order flow, because any observed order imbalance may be generated by noise trading alone. This weakens the connection between trades and private information about the asset value, thereby reducing adverse selection risk.

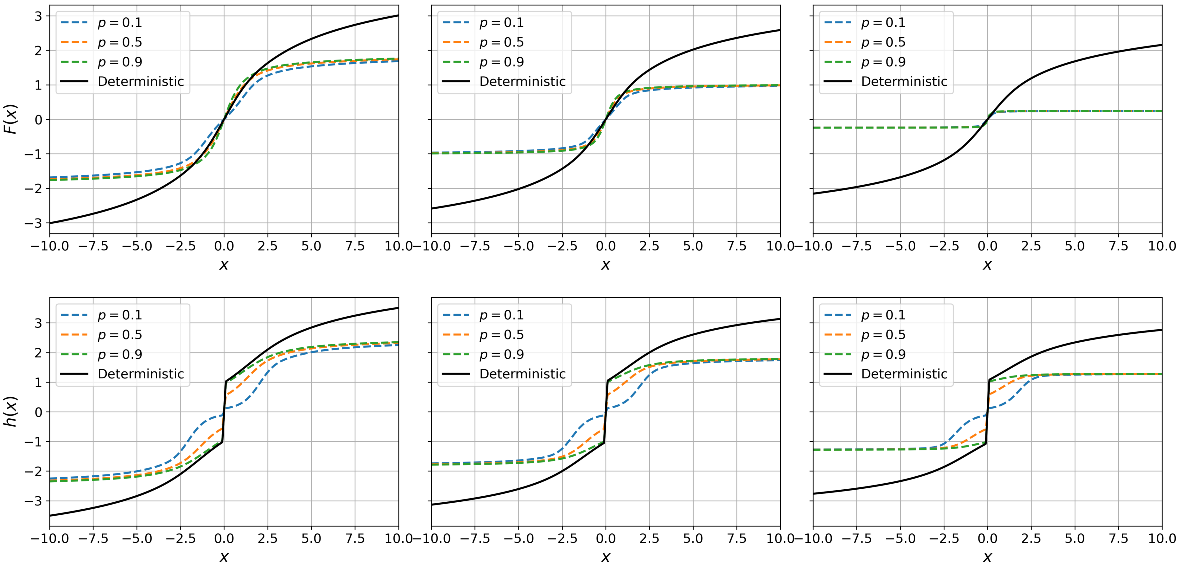
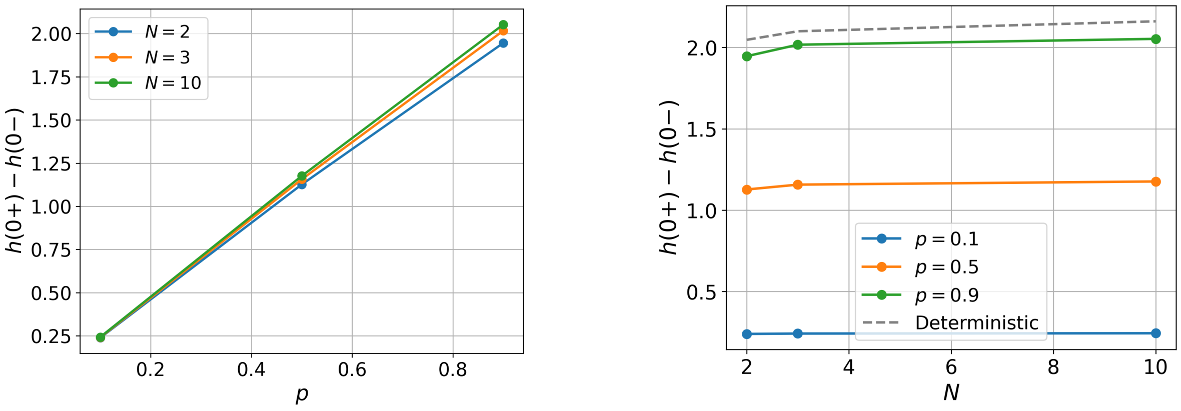


FIGURE 2.7: Comparison of marginal cost function $F(x)$ (top row) and marginal price function $h(x)$ (bottom row) under deterministic (solid black line) and random (dashed coloured lines) insider trading regimes. The asset value is Gaussian, $V \sim \mathcal{N}(0,2)$, and noise trades follow $Z \sim \mathcal{N}(0,1)$. Each column corresponds to a fixed number of potential insiders $N \in \{2,3,10\}$. In the random regime, the number of insiders follows a two-point distribution $\{0,N\}$ with insider presence probability $p \in \{0.1,0.5,0.9\}$.

2.6 Conclusion

We developed an equilibrium model with uncertain asymmetric information and provided a theoretical insights into the market impact of large trades, particularly relevant given the



(A) Spread vs insider probability p

(B) Spread vs number of insiders N

FIGURE 2.8: Bid-ask spread $h(0^+) - h(0^-)$ for Gaussian asset value $V \sim \mathcal{N}(0, 2)$ and noise trading $Z \sim \mathcal{N}(0, 1)$. (A) Fixes $N \in \{2, 3, 10\}$ and varies insider probability $p \in \{0.1, 0.5, 0.9\}$. (B) Fixes p and shows spread as a function of N, overlaying the deterministic spread as a dashed line.

scarcity of empirical data on very large trades. In equilibrium, competitive market makers determine limit prices based on their common beliefs about the asset return distribution and the distribution of the informed traders' quantities. We show that the asymptotic market impact is strongly non-linear in both: fat-tailed asset distributions lead to power-law market impact, while light-tailed asset distributions give rise to logarithmic impact, reflecting how informational uncertainty and tail risk jointly shape price formation.

Compared to the deterministic insider quantity in Çetin and Waelbroeck (2024), introducing uncertainty over both the existence and extent of informed traders yields a more realistic modelling framework, particularly in view of potential extensions to multiperiod settings. This uncertainty makes the power-law exponent more nonlinearly dependent on the distribution of informed trader quantities, as characterised through fixed-point equations.

In the static one-period setting, the equilibrium is characterised by a fixed-point mapping problem. Extending the model to a more realistic multiperiod model is more challenging, mainly because it requires keeping track of how information carries over time. However, under additional assumptions, such as myopic behaviour of traders, the framework could be extended to analyse how market makers dynamically update beliefs and set limit prices over time in response to evolving order flow.

Appendix

2.A Proofs for Equilibrium Existence

Proof of Lemma 2.3.5. Under the assumption, by monotone convergence theorem, we have

$$\lim_{x\to\infty}\int_{-\infty}^{\infty} q(\sigma,z)\phi_F(z+nx)\,dz = \int_{-\infty}^{\infty} q(\sigma,z)\lim_{x\to\infty}\phi_F(z+nx)\,dz.$$

Moreover, $\lim_{x\to\infty}\phi_F^+(z+nx) = \Psi^+(F(\infty)-)$, for all $n\in\mathbb{Z}_+$. From Definition 2.3.3 and (2.3.4), we have

$$\begin{aligned}\phi_F^+(z+nx) =& \frac{E[V\mathbb{1}_{[N=0]}\mathbb{1}_{[Z\geq z+nx]}]}{E[\mathbb{1}_{[N=0]}\mathbb{1}_{[Z\geq z+nx]}+\mathbb{1}_{[N\neq 0]}\Pi^+(F(\frac{z+nx-Z}{N}))]}\\ &+\frac{E[\mathbb{1}_{[N\neq 0]}\Phi^+(F(\frac{z+nx-Z}{N}))]}{E[\mathbb{1}_{[N=0]}\mathbb{1}_{[Z\geq z+nx]}+\mathbb{1}_{[N\neq 0]}\Pi^+(F(\frac{z+nx-Z}{N}))]}.\end{aligned}$$

The first term converges to 0 as $x\to\infty$. To see this,

$$\begin{aligned}&\lim_{x\to\infty}\frac{E[V\mathbb{1}_{[N=0]}\mathbb{1}_{[Z\geq z+nx]}]}{E[\mathbb{1}_{[N=0]}\mathbb{1}_{[Z\geq z+nx]}+\mathbb{1}_{[N\neq 0]}\Pi^+(F(\frac{z+nx-Z}{N}))]}\\
=&\lim_{x\to\infty}\frac{p(0)E[V]P(Z\geq z+nx)}{p(0)P(Z\geq z+nx)+\sum_{m\geq 1}mp(m)\int_{-\infty}^{\infty}q(\sigma,z+nx-mu)\Pi^+(F(u))\,du}\\
=&\lim_{x\to\infty}\frac{E[V]}{1+\sum_{m\geq 1}\frac{mp(m)}{p(0)}\frac{\int_{-\infty}^{\infty}q(\sigma,z+nx-mu)\Pi^+(F(u))\,du}{\int_{-\infty}^{0}q(\sigma,z+nx-u)\,du}}\\
\leq&\lim_{x\to\infty}\frac{E[V]}{1+\sum_{m\geq 1}\frac{mp(m)}{p(0)}\frac{\int_{a}^{\infty}\exp\{-\frac{1}{2\sigma^2}(z+nx-mu)^2\}\Pi^+(F(u))\,du}{\int_{-\infty}^{a}\exp\{-\frac{1}{2\sigma^2}(z+nx-u)^2\}\,du}} \quad \text{for some } a>0\\
\leq&\lim_{x\to\infty}\frac{E[V]}{1+\sum_{m\geq 1}\frac{mp(m)}{p(0)}\frac{\exp\{\frac{1}{\sigma^2}(z+nx)am\}\int_{a}^{\infty}\exp\{-\frac{1}{2\sigma^2}u^2m^2\}\Pi^+(F(u))\,du}{\exp\{\frac{1}{\sigma^2}(z+nx)a\}\int_{-\infty}^{a}\exp\{-\frac{1}{2\sigma^2}u^2\}\,du}}\\
\leq&\lim_{x\to\infty}\frac{E[V]}{1+\sum_{m\geq 2}\frac{mp(m)}{p(0)}\exp\{\frac{1}{\sigma^2}(z+nx)a(m-1)\}\frac{\int_{a}^{\infty}\exp\{-\frac{1}{2\sigma^2}u^2m^2\}\Pi^+(F(u))\,du}{\int_{-\infty}^{a}\exp\{-\frac{1}{2\sigma^2}u^2\}\,du}},\end{aligned}$$

converges to 0 as $x \to \infty$ due to the fact that $\exp\{\frac{1}{\sigma^2}(z+nx)a(m-1)\}$ in the denominator goes to infinity. Now, to see $\lim_{x\to\infty}\phi_F^+(z+nx)$ we could focus on the second term.

$$\begin{aligned}
&\frac{E[1_{[N\neq 0]}\Phi^+(F(\frac{z+nx-Z}{N}))]}{E[1_{[N=0]}1_{[Z\geq z+nx]}+1_{[N\neq 0]}\Pi^+(F(\frac{z+nx-Z}{N}))]} \\
&\leq \frac{E[1_{[N\neq 0]}\Phi^+(F(\frac{z+nx-Z}{N}))]}{E[1_{[N\neq 0]}\Pi^+(F(\frac{z+nx-Z}{N}))]} \\
&= \frac{\int_{-\infty}^{\infty}\sum_{m\geq 1} mp(m)q(\sigma,z+nx-mu)\Psi^+(F(u))\Pi^+(F(u))\,du}{\int_{-\infty}^{\infty}\sum_{m\geq 1} mp(m)q(\sigma,z+nx-mu)\Pi^+(F(u))\,du} \\
&= \frac{\int_{-\infty}^{\infty}\Psi^+(F(u))\sum_{m\geq 1} mp(m)q(\sigma,z+nx-mu)\Pi^+(F(u))\,du}{\int_{-\infty}^{\infty}\sum_{m\geq 1} mp(m)q(\sigma,z+nx-mu)\Pi^+(F(u))\,du}.
\end{aligned}$$

We shall show the probability measure

$$\frac{\sum_{m\geq 1} mp(m)q(\sigma,z+nx-mu)\Pi^+(F(u))}{\int_{-\infty}^{\infty}\sum_{m\geq 1} mp(m)q(\sigma,z+nx-mu)\Pi^+(F(u))\,du}\,du \tag{2.A.1}$$

converges to the point mass at ∞. For $0 < a < \infty$,

$$\begin{aligned}
&\frac{\int_{-\infty}^{a}\sum_{m\geq 1} mp(m)q(\sigma,z+nx-mu)\Pi^+(F(u))\,du}{\int_{-\infty}^{\infty}\sum_{m\geq 1} mp(m)q(\sigma,z+nx-mu)\Pi^+(F(u))\,du} \\
&= \frac{\sum_{m\geq 1} mp(m)\exp\{-\frac{(z+nx)^2}{2\sigma^2}\}\int_{-\infty}^{a}\Pi^+(F(u))\exp\{-\frac{(mu)^2}{2\sigma^2}+\frac{mu(z+nx)}{\sigma^2}\}\,du}{\sum_{m\geq 1} mp(m)\exp\{-\frac{(z+nx)^2}{2\sigma^2}\}\int_{-\infty}^{\infty}\Pi^+(F(u))\exp\{-\frac{(mu)^2}{2\sigma^2}+\frac{mu(z+nx)}{\sigma^2}\}\,du} \\
&\leq \frac{\sum_{m\geq 1} mp(m)\exp\{\frac{ma(z+nx)}{\sigma^2}\}\int_{-\infty}^{a}\Pi^+(F(u))\exp\{-\frac{(mu)^2}{2\sigma^2}\}\,du}{\sum_{m\geq 1} mp(m)\exp\{\frac{m2a(z+nx)}{\sigma^2}\}\int_{2a}^{\infty}\Pi^+(F(u))\exp\{-\frac{(mu)^2}{2\sigma^2}\}\,du} \\
&\leq \sum_{m\geq 1}\frac{mp(m)\exp\{\frac{ma(z+nx)}{\sigma^2}\}\int_{-\infty}^{a}\Pi^+(F(u))\exp\{-\frac{(mu)^2}{2\sigma^2}\}\,du}{\sum_{k=m} kp(k)\exp\{\frac{k2a(z+nx)}{\sigma^2}\}\int_{2a}^{\infty}\Pi^+(F(u))\exp\{-\frac{(ku)^2}{2\sigma^2}\}\,du} \\
&\leq \sum_{m\geq 1}\exp\left\{-\frac{a(z+nx)m}{\sigma^2}\right\}\frac{\int_{-\infty}^{a}\Pi^+(F(u))\exp\{-\frac{(mu)^2}{2\sigma^2}\}\,du}{\int_{2a}^{\infty}\Pi^+(F(u))\exp\{-\frac{(mu)^2}{2\sigma^2}\}\,du} \\
&\leq B\sum_{m\geq 1}\exp\{C(x)m\} \\
&= \frac{B}{\exp\{-C(x)\}-1},
\end{aligned}$$

where

$$B := \max_{1 \le m \le n_{\max}} \frac{\int_{-\infty}^{a} \Pi^+(F(u)) \exp\left\{-\frac{(mu)^2}{2\sigma^2}\right\} du}{\int_{2a}^{\infty} \Pi^+(F(u)) \exp\left\{-\frac{(mu)^2}{2\sigma^2}\right\} du},$$

and $C(x) := -\frac{a(z+nx)}{\sigma^2}$. Since the number of informed traders is almost surely bounded by $n_{\max} < \infty$, the supremum is taken over a finite set. Moreover, the denominator is strictly positive for each $1 \le m \le n_{\max}$. Hence $B < \infty$. Note that $C(x) \to -\infty$ as $x \to \infty$, which further shows the probability measure (2.A.1) is convergent to the point mass at ∞. As a result, $\lim_{x\to\infty} \phi_F^+(z + nx) = \Psi^+(F(\infty)-)$.

Now, from $F(x)$ represented in (2.3.6), since

$$\sum_{n\ge1} \frac{p(n)}{1-p(0)} \int_{-\infty}^{\infty} \left\{ \frac{1}{n} q(\sigma, z - nx) + \frac{n-1}{n} \bar{q}(\sigma, n, x, z) \right\} dz = 1,$$

we shall have $F(\infty) = \lim_{x\to\infty} \phi_F^+(z + nx) = \Psi^+(F(\infty)-)$. Observe that $\Psi^+(x-) := E[V|V \ge x] > x$ for any $x < M$, as $P(V > x) > 0$ whenever $x < M$. In other words, $\Psi^+(x-) = x$ only if $x = M$. Accordingly, $F(\infty) = M$. Similarly, $F(-\infty) = m$.

Since $m \neq M$, and that F solves (2.3.6), where ϕ_F is non-decreasing in view of Lemma 2.3.6(4), we see that F is strictly increasing. □

Proof of Lemma 2.3.6. The proof when $u = u^+$ is provided below, the correspondence for u^- could be proved analogously.

(1) Note that, for $0 \le t \le 1$,

$$\begin{aligned} u^+(t,x) &= E_N \left[1_{[N=0]} 1_{[B_1 \le 0]} + 1_{[N\neq0]} \Pi^+ \left(g\left(\frac{B_1}{N}\right)\right) |B_t = x \right] \\ &= \sum_{n\ge0} p(n) \int_{-\infty}^{\infty} \left\{ 1_{[n=0]} 1_{[y\le0]} + 1_{[n\neq0]} \Pi^+ \left(g\left(\frac{y}{n}\right)\right)\right\} \frac{1}{\sqrt{2\pi\sigma^2(1-t)}} \exp\left\{-\frac{(y-x)^2}{2\sigma^2(1-t)}\right\} dy \\ &= \sum_{n\ge0} p(n) \int_{-\infty}^{\infty} \left\{ 1_{[n=0]} 1_{[y+x\le0]} + 1_{[n\neq0]} \Pi^+ \left(g\left(\frac{y+x}{n}\right)\right)\right\} \frac{1}{\sqrt{2\pi\sigma^2(1-t)}} \exp\left\{-\frac{y^2}{2\sigma^2(1-t)}\right\} dy \end{aligned}$$

Now, define $\frac{d\mathbb{Q}}{d\mathbb{P}} = \frac{u(1,B_1)}{u(0,B_0)}$, $Z_t = u(t, B_t) = E\left[\frac{d\mathbb{Q}}{d\mathbb{P}}\middle|\mathcal{F}_t\right]$ is a bounded martingale with terminal condition $Z_1 = u(1, B_1) = E_N \left[1_{[N=0]} 1_{B_1\le0} + 1_{[N\neq0]} \Pi^+ \left(g\left(\frac{B_1}{N}\right)\right)\right]$. And $B_t = \sigma\beta_t^{\mathbb{P}} + x$, where $\beta_t^{\mathbb{P}}$ is a standard Brownian Motion in $\mathbb{P}$.

By Girsanov's Theorem,

$$dB_t = dW_t^{\mathbb{Q}} + \sigma^2 \frac{u_x(t,B_t)}{u(t,B_t)} dt, \quad B_0 = x, \tag{2.A.2}$$

which is the SDE in (2.3.8).

(2) According to Definition 2.3.3, assuming $B_1 = \sigma \beta_t^{\mathbb{P}} + x$,

$$\begin{aligned}
\phi_g^+(x) :=& \frac{E^{\mathbb{P}}[V 1_{[N=0]} 1_{[B_1 \le 0]} + 1_{[N \neq 0]} \Phi^+(g(\frac{B_1}{N}))]}{E^{\mathbb{P}}[1_{[N=0]} 1_{[B_1 \le 0]} + 1_{[N \neq 0]} \Pi^+(g(\frac{B_1}{N}))]} \\
=& \frac{E^{\mathbb{P}}[\left\{ V 1_{[N=0]} 1_{[B_1 \le 0]} + 1_{[N \neq 0]} \Psi^+(g(\frac{B_1}{N})) \right\} \left\{ 1_{[N=0]} 1_{[B_1 \le 0]} + 1_{[N \neq 0]} \Pi^+(g(\frac{B_1}{N})) \right\}]}{E^{\mathbb{P}}[1_{[N=0]} 1_{[B_1 \le 0]} + 1_{[N \neq 0]} \Pi^+(g(\frac{B_1}{N}))]} \\
=& E^{\mathbb{Q}} \left[V 1_{[N=0]} 1_{[B_1 \le 0]} + 1_{[N \neq 0]} \Psi^+(g(\frac{B_1}{N})) \right]
\end{aligned}$$

(3) From Definition 2.3.3, $\phi_g(\cdot)$ evaluated at $x = 0$ is

$$\begin{aligned}
\phi_g^+(0) =& \frac{p(0) E[V 1_{[Z \ge 0]}] + \sum_{n \ge 0} p(n) 1_{[n \neq 0]} \int_{-\infty}^{\infty} \Phi^+(g(\frac{z}{n})) q(\sigma, z)\, dz}{p(0) P(Z \ge 0) + \sum_{n \ge 0} p(n) 1_{[n \neq 0]} \int_{-\infty}^{\infty} \Pi^+(g(\frac{z}{n})) q(\sigma, z)\, dz} \\
=& \frac{\frac{1}{2} p(0) E[V] + \sum_{n \ge 0} p(n) 1_{[n \neq 0]} \int_{-\infty}^{\infty} \Phi^+(g(\frac{z}{n})) q(\sigma, z)\, dz}{\frac{1}{2} p(0) + \sum_{n \ge 0} p(n) 1_{[n \neq 0]} \int_{-\infty}^{\infty} \Pi^+(g(\frac{z}{n})) q(\sigma, z)\, dz}.
\end{aligned}$$

Similarly for $\phi_g^-(0)$. To show $\phi_g^+(0) > \phi_g^-(0)$, it's equivalent to show

$$\begin{aligned}
&\left(\tfrac{p(0)}{2} E[V] + \sum_{n \ge 1} p(n) \int_{-\infty}^{\infty} \Phi^+ \left(g\left(\tfrac{z}{n}\right)\right) q(\sigma, z)\, dz \right) \left(\tfrac{p(0)}{2} E[V] + \sum_{n \ge 1} p(n) \int_{-\infty}^{\infty} \Pi^- \left(g\left(\tfrac{z}{n}\right)\right) q(\sigma, z)\, dz \right) \\
- &\left(\tfrac{p(0)}{2} E[V] + \sum_{n \ge 1} p(n) \int_{-\infty}^{\infty} \Phi^- \left(g\left(\tfrac{z}{n}\right)\right) q(\sigma, z)\, dz \right) \left(\tfrac{p(0)}{2} E[V] + \sum_{n \ge 1} p(n) \int_{-\infty}^{\infty} \Pi^+ \left(g\left(\tfrac{z}{n}\right)\right) q(\sigma, z)\, dz \right) > 0.
\end{aligned}$$

By the relationship $\Phi^+ + \Phi^- = E[V]$, and $\Pi^+ + \Pi^- = 1$, it could be further equivalent to show

$$\sum_{n \ge 0} p(n) 1_{[n \neq 0]} \int_{-\infty}^{\infty} \left\{ \Psi^+(g(\frac{z}{n})) - E[V] \right\} \Pi^+(g(\frac{z}{n})) q(\sigma, z)\, dz > 0,$$

which is indeed the case, since $E[V] = \Psi^+(m)$, $g(\frac{z}{n}) \ge m$ for all $z \in \mathbb{R}$, $n \ge 0$ and $\Psi^+(\cdot)$

is non-decreasing and non-constant.
(4) Now, suppose g is non-decreasing, it is clear that $u^+_x(t,x) \leq 0$, since $1_{[y+x\leq 0]}$ and $\Pi^+\left(g\left(\frac{y+x}{n}\right)\right)$ are both non-increasing in y. Therefore, as $\frac{u^+_x(t,x)}{u(t,x)}$ is Lipschitz on $[0,t]$ for any $t < T$, the standard comparison results for SDEs applied to (2.A.2), and that Ψ^+ is increasing imply

$$\begin{aligned}&E^{\mathbb{Q}}\left[V1_{[N=0]}1_{[B_1\leq 0]}+1_{[N\neq 0]}\Psi^+(g(\frac{B_1}{N}))\Big|B_0=y\right]\\ \geq &E^{\mathbb{Q}}\left[V1_{[N=0]}1_{[B_1\leq 0]}+1_{[N\neq 0]}\Psi^+(g(\frac{B_1}{N}))\Big|B_0=x\right], \text{ if } y>x.\end{aligned}$$

This shows the desired monotonicity of ϕ^+_g.
Meanwhile, (2.A.2) is bounded from above by $\sigma W_t + x$ in the case of $u = u^+$, since $u^+_x \leq 0$. Combined with the monotonicity of $1_{[y+x\leq 0]}$ and $\Psi^+(g)$, we deduce that $\phi^+_g(x) \leq E^{\mathbb{Q}^+}\left[V1_{[N=0]}1_{[\sigma W_1^{\mathbb{Q}^+}+x\leq 0]}+1_{[N\neq 0]}\Psi^+\left(g\left(\frac{\sigma W_1^{\mathbb{Q}^+}+x}{N}\right)\right)\right]$. □

2.B Proofs for Price Impact Asymptotics

The following Karamata's Theorems, direct half and converse half, (see, for example, Bingham et al. (1987) Theorem 1.5.11 and Theorem 1.6.1) are essential in our price impact asymptotics derivations. These theorems describe the asymptotic behaviour of regularly varying functions when integrating against powers.

Theorem 2.B.1. *(Karamata's Theorem; direct half). Let f vary regularly with index ρ, and be locally bounded in $[X,\infty)$. Then*

(1) for any $\sigma \geq -(\rho+1)$,

$$\frac{x^{\sigma+1}f(x)}{\int_X^x t^\sigma f(t)dt} \to \sigma+\rho+1, \quad x\to\infty;$$

(2) for any $\sigma < -(\rho+1)$,

$$\frac{x^{\sigma+1}f(x)}{\int_x^\infty t^\sigma f(t)dt} \to -(\sigma+\rho+1), \quad x\to\infty.$$

Theorem 2.B.2. *(Karamata's Theorem; converse half). Let f be positive and locally integrable in $[X, \infty)$.*

(1) If for some $\sigma > -(\rho+1)$,

$$\frac{x^{\sigma+1} f(x)}{\int_X^x t^\sigma f(t) dt} \to \sigma + \rho + 1, \quad x \to \infty,$$

then f varies regularly with index ρ at ∞.

(2) If for some $\sigma < -(\rho+1)$,

$$\frac{x^{\sigma+1} f(x)}{\int_x^\infty t^\sigma f(t) dt} \to -(\sigma + \rho + 1), \quad x \to \infty,$$

then again f varies regularly with index ρ at ∞.

Proof of Theorem 2.4.3. We will only provide the proof of the first statement, while the second statement could be proved similarly. We shall first show the second claim about the regularly varying of $\Pi^+(F)$, by assuming ρ^+ as the regularly varying index of $M - F$. Then, we derive the fixed point equation that ρ^+ solves.

Suppose $0 < \Psi_x^+(M) < 1$, and $M - F$ is regularly varying at ∞ with index $\rho^+ \in [-1, 0)$. That is,

$$\gamma(y) := \lim_{\alpha \to \infty} \frac{M - F(\alpha y)}{M - F(\alpha)} = y^{\rho^+}.$$

To show that $\Pi^+(F(x))$ is regularly varying at ∞ of index $\frac{\Psi_x^+(M)}{1-\Psi_x^+(M)}\rho^+ =: k\rho^+$, we shall apply the converse part of Karamata's Theorem in Theorem 2.B.2.

Step 1: By direct manipulation, we have the identity $-\frac{\Pi_x^+(x)}{\Pi^+(x)} = \frac{\Psi_x^+(x)}{\Psi^+(x)-x}$. We want to use the Theorem 2.B.2 (2) to show the regularly varying property of $\Pi^+(F)$. First,

$$\lim_{x\to\infty} \frac{\Pi^+(F(x))}{\int_x^\infty \frac{\Pi^+(F(t))}{t} dt} = -\lim_{x\to\infty} \frac{x\Pi_x^+(F(x))F'(x)}{\Pi^+(F(x))} = \lim_{x\to\infty} \frac{x\Psi_x^+(F(x))F'(x)}{\Psi^+(F(x)) - F(x)}.$$

Step 2: $\Psi^+(F) - F$ is regularly varying at ∞ of index ρ^+.

$$\lim_{\alpha\to\infty} \frac{\Psi^+(F(\alpha x)) - F(\alpha x)}{\Psi^+(F(\alpha)) - F(\alpha)} = \lim_{\alpha\to\infty} \frac{\frac{\Psi^+(F(\alpha x))-F(\alpha x)}{M-F(\alpha x)} M - F(\alpha x)}{\frac{\Psi^+(F(\alpha))-F(\alpha)}{M-F(\alpha)} M - F(\alpha)} = \frac{M - F(\alpha x)}{M - F(\alpha)} = x^{\rho^+}.$$

This is because

$$\lim_{x\to\infty}\frac{\Psi^+(F(x))-F(x)}{M-F(x)}=\lim_{x\to\infty}\frac{(\Psi_x^+(F(x))-1)F'(x)}{-F'(x)}=1-\Psi_x^+(M).$$

Step 3: Now, in view of Theorem 2.B.1, let $\sigma=-1$

$$\lim_{x\to\infty}\frac{\Psi^+(F(x))-F(x)}{\int_x^\infty\frac{\Psi^+(F(t))-F(t)}{t}\,dt}=-\lim_{x\to\infty}\frac{x[\Psi_x^+(F(x))F'(x)-F'(x)]}{\Psi^+(F(x))-F(x)}=-\rho^+.$$

That is, $\lim_{x\to\infty}\frac{x\Psi_x^+(F(x))F'(x)}{\Psi^+(F(x))-F(x)}=\rho^++\lim_{x\to\infty}\frac{xF'(x)}{\Psi^+(F(x))-F(x)}$.
Step 4: Therefore,

$$\begin{aligned}\lim_{x\to\infty}\frac{\Pi^+(F(x))}{\int_x^\infty\frac{\Pi^+(F(t))}{t}\,dt}&=\lim_{x\to\infty}\frac{x\Psi_x^+(F(x))F'(x)}{\Psi^+(F(x))-F(x)}=\rho^++\lim_{x\to\infty}\frac{xF'(x)}{\Psi^+(F(x))-F(x)}\\&=\rho^++\lim_{x\to\infty}\frac{xF'(x)}{M-F(x)}\frac{M-F(x)}{\Psi^+(F(x))-F(x)}\\&=\rho^++\frac{1}{1-\Psi_x^+(M)}(-\rho^+)=-\frac{\Psi_x^+(M)}{1-\Psi_x^+(M)}\rho^+,\end{aligned}$$

since $\lim_{x\to\infty}\frac{xF'(x)}{M-F(x)}=-\rho^+$ from another application of Theorem 2.B.1 (2).
Step 5: Now, let $\sigma=-1$ in Theorem 2.B.2 (2), we shall have the desired regularly varying $\Pi^+(F)$.
To show the fix point equation for regularly varying index ρ^+, suppose F is of the form (2.3.6) in equilibrium. The asymptotics,

$$\begin{aligned}\lim_{\alpha\to\infty}\frac{M-F(\alpha x)}{M-F(\alpha)}&=\lim_{\alpha\to\infty}\sum_{n\geq1}\frac{p(n)}{1-p(0)}\int_{-\infty}^{\infty}dz\left\{\frac{1}{n}q(\sigma,z-n\alpha x)+\frac{n-1}{n}\bar{q}(\sigma,n,\alpha x,z)\right\}\frac{M-\phi_F(z)}{M-F(\alpha)}\\&=\lim_{\alpha\to\infty}\sum_{n\geq1}\frac{p(n)}{1-p(0)}\int_{-\infty}^{\infty}dz\left\{\frac{1}{n}q(\frac{\sigma}{\alpha},z-nx)+\frac{n-1}{n}\bar{q}(\frac{\sigma}{\alpha},n,x,z)\right\}\frac{M-\phi_F(\alpha z)}{M-F(\alpha)}\end{aligned}$$

Step 1: For $y>0$, by Mean Value Theorem, there exists some $y^*\geq F(\alpha y)$ such that

$$\frac{M-\Psi^+(F(\alpha y))}{M-F(\alpha)}=\frac{(M-F(\alpha y))\Psi_x^+(y^*)}{M-F(\alpha)}.$$

Therefore, as $\alpha \to \infty$,

$$\lim_{\alpha\to\infty} \frac{M - \Psi^+(F(\alpha y))}{M - F(\alpha)} = \lim_{\alpha\to\infty} \frac{(M - F(\alpha y))\Psi^+_x(y^*)}{M - F(\alpha)}$$
$$= \Psi^+_x(M) \lim_{\alpha\to\infty} \frac{M - F(\alpha y)}{M - F(\alpha)} = \Psi^+_x(M) y^{\rho^+}.$$

Step 2: We now look into $\lim_{\alpha\to\infty} \phi^+_F(\alpha z)$. By definition 2.3.3, and assuming $E[V] = 0$, we have

$$\lim_{\alpha\to\infty} \phi^+_F(\alpha z) = \lim_{\alpha\to\infty} \frac{p(0)E[V]P(Z \geq \alpha z) + \sum_{n\geq 0} p(n) 1_{[n\neq 0]} \int_{-\infty}^{\infty} dy \Pi^+\left(F\left(\frac{y}{n}\right)\right) q(\sigma, y - \alpha z) \Psi^+\left(F\left(\frac{y}{n}\right)\right)}{p(0)P(Z \geq \alpha z) + \sum_{n\geq 0} 1_{[n\neq 0]} \int_{-\infty}^{\infty} dy \Pi^+\left(F\left(\frac{y}{n}\right)\right) q(\sigma, y - \alpha z)}$$
$$= \lim_{\alpha\to\infty} \sum_{n\geq 0} \frac{p(n) 1_{[n\neq 0]} \int_{-\infty}^{\infty} dy \frac{\Pi^+(F(\alpha y))}{\Pi^+(F(\alpha))} q(\frac{\sigma}{\alpha n}, y - \frac{z}{n})}{\sum_{n\geq 0} p(n) 1_{[n\neq 0]} \int_{-\infty}^{\infty} dy \frac{\Pi^+(F(\alpha y))}{\Pi^+(F(\alpha))} q(\frac{\sigma}{\alpha n}, y - \frac{z}{n})} \Psi^+(F(\alpha y)).$$

Since for any $n \in \mathbb{N}_+$,

$$\frac{p(n) 1_{[n\neq 0]} \int_{-\infty}^{\infty} dy \frac{\Pi^+(F(\alpha y))}{\Pi^+(F(\alpha))} q(\frac{\sigma}{\alpha n}, y - \frac{z}{n})}{\sum_{n\geq 0} p(n) 1_{[n\neq 0]} \int_{-\infty}^{\infty} dy \frac{\Pi^+(F(\alpha y))}{\Pi^+(F(\alpha))} q(\frac{\sigma}{\alpha n}, y - \frac{z}{n})}$$

is a valid probability mass function, we have

$$\frac{M - \phi^+_F(\alpha z)}{M - F(\alpha)} = \sum_{n\geq 0} \frac{p(n) 1_{[n\neq 0]} \int_{-\infty}^{\infty} dy \frac{\Pi^+(F(\alpha y))}{\Pi^+(F(\alpha))} q(\frac{\sigma}{\alpha n}, y - \frac{z}{n})}{\sum_{n\geq 0} p(n) 1_{[n\neq 0]} \int_{-\infty}^{\infty} dy \frac{\Pi^+(F(\alpha y))}{\Pi^+(F(\alpha))} q(\frac{\sigma}{\alpha n}, y - \frac{z}{n})} \frac{M - \Psi^+(F(\alpha y))}{M - F(\alpha)}.$$

Now, send $\alpha \to \infty$. Since $q(\frac{\sigma}{\alpha n}, y - \frac{z}{n})\, dy$ for all $n \in \mathbb{N}_+$ converges to point mass at $\frac{z}{n}$, we shall have

$$\lim_{\alpha\to\infty} \frac{M - \phi^+_F(\alpha z)}{M - F(\alpha)} = \sum_{n\geq 0} \frac{p(n) 1_{[n\neq 0]} \left(\frac{z}{n}\right)^{k\rho^+}}{\sum_{n\geq 0} p(n) 1_{[n\neq 0]} \left(\frac{z}{n}\right)^{k\rho^+}} \Psi^+_x(M) \left(\frac{z}{n}\right)^{\rho^+}$$
$$= \Psi^+_x(M) z^{\rho^+} \frac{E[1_{[N\neq 0]} N^{-(k+1)\rho^+}]}{E[1_{[N\neq 0]} N^{-k\rho^+}]}. \tag{2.B.1}$$

Step 3: for $x > 0$, in view of the discussion in Step 1 and Step 2, the limit of $\frac{M-F(\alpha x)}{M-F(\alpha)}$ as $\alpha \to \infty$ can be written as

$$\begin{aligned}\lim_{\alpha\to\infty} \frac{M-F(\alpha x)}{M-F(\alpha)} &= \lim_{\alpha\to\infty} \sum_{n\geq 1} \frac{p(n)}{1-p(0)} \left\{ \frac{1}{n} \int_0^\infty dz q(\frac{\sigma}{\alpha}, z-nx) \frac{M-\phi_F^+(\alpha z)}{M-F(\alpha)} \right. \\ &\quad + \frac{n-1}{n} \int_0^\infty dz \bar{q}(\frac{\sigma}{\alpha}, n, x, z) \frac{M-\phi_F^+(\alpha z)}{M-F(\alpha)} \\ &\quad \left. + \int_{-\infty}^0 dz \left\{ \frac{1}{n} q(\frac{\sigma}{\alpha}, z-nx) + \frac{n-1}{n} \bar{q}(\frac{\sigma}{\alpha}, n, x, z) \right\} \frac{M-\phi_F^-(\alpha z)}{M-F(\alpha)} \right\} \\ x^{\rho^+} &= \Psi_x^+(M) \frac{E[1_{[N\neq 0]} N^{-(k+1)\rho^+}]}{E[1_{[N\neq 0]} N^{-k\rho^+}]} \sum_{n\geq 1} \frac{p(n)}{1-p(0)} \left\{ \frac{1}{n} (nx)^{\rho^+} + \frac{n-1}{nx} \int_0^x dy (ny)^{\rho^+} \right\}\end{aligned}$$

Direct manipulations yield

$$\rho^+ = \frac{\Psi_x^+(M)}{1-p(0)} \frac{E\left[1_{[N\neq 0]} N^{-(k+1)\rho^+}\right]}{E\left[1_{[N\neq 0]} N^{-k\rho^+}\right]} E\left[1_{[N\neq 0]} (\rho^+ N^{\rho^+-1} + N^{\rho^+})\right] - 1.$$

□

Proof of Corollary 2.4.5. We only provide the proof of asymptotic $M - IS^*(x)$ as $x \to \infty$, while the second statement can be similarly proved.

Start from the representation of $IS^*(x) = E[\frac{1}{x}\int_0^x h^*(Z+Nu)\,du | N \geq 1] = E[\frac{1}{x}\int_0^x \phi_{F^*}^+(Z+Nu)\,du | N \geq 1]$, we can compute that, when $x > 0$,

$$\begin{aligned}\frac{M-IS^*(x)}{M-F^*(x)} &= \frac{1}{x} \sum_{n\geq 1} \frac{p(n)}{1-p(0)} \int_0^x du \int_{-\infty}^\infty dz\, q(\sigma, z-nu) \frac{M-\phi_{F^*}(z)}{M-F^*(x)} \\ &= \sum_{n\geq 1} \frac{p(n)}{1-p(0)} \int_0^1 du \int_{-\infty}^\infty dz\, q(\sigma, z-nux) \frac{M-\phi_{F^*}(z)}{M-F^*(x)} \\ &= \sum_{n\geq 1} \frac{p(n)}{1-p(0)} \int_0^1 du \int_{-\infty}^\infty dz\, q(\frac{\sigma}{x}, z-nu) \frac{M-\phi_{F^*}(xz)}{M-F^*(x)}.\end{aligned}$$

As $x \to \infty$, observe that $q(\frac{\sigma}{x}, z - nu)dz$ converges to Dirac measure at nu, combined with (2.B.1),

$$\begin{aligned}
\lim_{x\to\infty} \frac{M - IS^*(x)}{M - F^*(x)} &= \lim_{x\to\infty} \sum_{n\geq 1} \frac{p(n)}{1-p(0)} \int_0^1 du \int_0^\infty dz\, q(\frac{\sigma}{x}, z - nu) \frac{M - \phi^+_{F^*}(xz)}{M - F^*(x)} \\
&= \Psi^+_x(M) \frac{E[1_{[N\neq 0]} N^{-(k+1)\rho^+}]}{E[1_{[N\neq 0]} N^{-k\rho^+}]} \sum_{n\geq 1} \frac{p(n)}{1-p(0)} \int_0^1 (nu)^{\rho^+} \\
&= \frac{\Psi^+_x(M)}{1-p(0)} \frac{E[1_{[N\neq 0]} N^{-(k+1)\rho^+}]}{E[1_{[N\neq 0]} N^{-k\rho^+}]} \frac{E[1_{[N\neq 0]} N^{\rho^+}]}{\rho^+ + 1}.
\end{aligned}$$

□

Proof of Corollary 2.4.6. For $y > 0$, we have

$$\begin{aligned}
P(Y^* > y) = P(Nx^* + Z > y) &= p(0)P(Z \geq y) + \sum_{n\geq 0} p(n) 1_{[n\neq 0]} \int_{-\infty}^\infty P\left(x^* > \frac{y-z}{n}\right) q(\sigma, z)\, dz \\
&= p(0)P(Z \geq y) + \sum_{n\geq 0} p(n) 1_{[n\neq 0]} \int_{-\infty}^\infty P\left(V > F\left(\frac{y-z}{n}\right)\right) q(\sigma, z)\, dz \\
&= p(0)P(Z \geq y) + \sum_{n\geq 0} p(n) n 1_{[n\neq 0]} \int_{-\infty}^\infty \Pi^+(F(z)) q(\sigma, y - nz)\, dz.
\end{aligned}$$

Now, we look into the regularly varying index of $P(Y^* > y)$.

$$\begin{aligned}
\frac{P(Y^* > \alpha y)}{P(Y^* > \alpha)} &= \frac{p(0)P(Z \geq \alpha y) + \sum_{n\geq 0} p(n) n 1_{[n\neq 0]} \int_{-\infty}^\infty dz \Pi^+(F(z)) q(\sigma, \alpha y - nz)}{p(0)P(Z \geq \alpha) + \sum_{n\geq 0} p(n) n 1_{[n\neq 0]} \int_{-\infty}^\infty dz \Pi^+(F(z)) q(\sigma, \alpha - nz)} \\
&= \frac{p(0)P(Z \geq \alpha y) + \sum_{n\geq 0} p(n) 1_{[n\neq 0]} \int_{-\infty}^\infty dz \Pi^+(F(\alpha z)) q(\frac{\sigma}{\alpha n}, \frac{y}{n} - z)}{p(0)P(Z \geq \alpha) + \sum_{n\geq 0} p(n) 1_{[n\neq 0]} \int_{-\infty}^\infty dz \Pi^+(F(\alpha z)) q(\frac{\sigma}{\alpha n}, \frac{1}{n} - z)} \\
&= \frac{p(0)P(Z \geq \alpha y) + \sum_{n\geq 0} p(n) 1_{[n\neq 0]} \int_{-\infty}^\infty dz \frac{\Pi^+(F(\alpha z))}{\Pi^+(F(\alpha))} q(\frac{\sigma}{\alpha n}, \frac{y}{n} - z)}{p(0)P(Z \geq \alpha) + \sum_{n\geq 0} p(n) 1_{[n\neq 0]} \int_{-\infty}^\infty dz \frac{\Pi^+(F(\alpha z))}{\Pi^+(F(\alpha))} q(\frac{\sigma}{\alpha n}, \frac{1}{n} - z)} \\
&\xrightarrow{\alpha\to\infty} \frac{\sum_{n\geq 0} p(n) 1_{[n\neq 0]} \left(\frac{y}{n}\right)^{k\rho^+}}{\sum_{n\geq 0} p(n) 1_{[n\neq 0]} \left(\frac{1}{n}\right)^{k\rho^+}} = y^{k\rho^+}.
\end{aligned}$$

Therefore, $P(Y^* > y)$ can be written as $y^{k\rho^+ s(x)}$, where s is a slowly varying function. The claim when $y < 0$ can be shown similarly. □

Proof of Theorem 2.4.7. Again, we shall only show the asymptotics for $M-F$. The one for $F-m$ can be shown in a similar way.
Step 1: Suppose $\Psi_x^+(M) = 1$, $M-F$ is slowly varying, i.e. $\lim_{\alpha\to\infty} \frac{M-F(\alpha x)}{M-F(\alpha)} = x^{\rho^+} = 1$, where $\rho^+ = 0$. Let $g(x) := \exp\{(M-F(x))^{-n}\}$, we can see that

$$\begin{aligned}\frac{g(\alpha x)}{g(\alpha)} &= \exp\left\{\frac{(M-F(\alpha))^n-(M-F(\alpha x))^n}{(M-F(\alpha))^n(M-F(\alpha x))^n}\right\}\\ &= \exp\left\{\frac{(F(\alpha x)-F(\alpha))\sum_{i=0}^{n-1}(M-F(\alpha))^{n-1-i}(M-F(\alpha x))^i}{(M-F(\alpha))^n(M-F(\alpha x))^n}\right\}\\ &= \exp\left\{\frac{F(\alpha x)-F(\alpha)}{(M-F(\alpha))^{n+1}}\frac{M-F(\alpha)}{M-F(\alpha x)}\sum_{i=1}^{n}\left(\frac{M-F(\alpha x)}{M-F(\alpha)}\right)^i\right\}.\end{aligned}$$

Denote $f(\alpha,x) := \frac{F(\alpha x)-F(\alpha)}{(M-F(\alpha))^{n+1}}$ and $\gamma(x) := \lim_{\alpha\to\infty} f(\alpha,x) = \rho\log x$ for some $\rho > 0$. Then, we have

$$\lim_{\alpha\to\infty}\frac{g(\alpha x)}{g(\alpha)} = \exp\{n\gamma(x)\} = x^{n\rho},$$

i.e. $g(x)$ is regularly varying at ∞ of index $n\rho$, which means that $g(x)$ can be written as $g(x) = x^{n\rho}s(x)$, $x>0$, where $s(x)$ is some slowly varying function at ∞. By Theorem 2.B.1 (2),

$$\begin{aligned}&\lim_{x\to\infty}\frac{g(x)}{\int_{-\infty}^{x}\frac{g(t)}{t}\,dt} = \lim_{x\to\infty}\frac{xg'(x)}{g(x)} = n\rho,\\ &\frac{xg'(x)}{g(x)} = \frac{xnF'(x)}{(M-F(x))^{n+1}},\\ &\lim_{x\to\infty}\frac{xF'(x)}{(M-F(x))^{n+1}} = \frac{1}{n}n\rho = \rho. \qquad (2.B.2)\end{aligned}$$

Step 2: Suppose that there exist an integer $n>0$ and a real constant $k\in(0,\infty)$ such that

$$\lim_{x\to M}\frac{\Psi^+(x)-x}{(M-x)^{n+1}} = \frac{1}{k}, \qquad (2.B.3)$$

then we shall have $\Psi^+(F) - F$ is slowly varying. To see this,

$$\lim_{\alpha\to\infty} \frac{\Psi^+(F(\alpha x)) - F(\alpha x)}{\Psi^+(F(\alpha)) - F(\alpha)} = \lim_{\alpha\to\infty} \frac{\frac{\Psi^+(F(\alpha x)) - F(\alpha x)}{(M - F(\alpha x))^{n+1}}}{\frac{\Psi^+(F(\alpha)) - F(\alpha)}{(M - F(\alpha))^{n+1}}} \left(\frac{M - F(\alpha x)}{M - F(\alpha)}\right)^{n+1} = (x^{\rho^+})^{n+1} = 1.$$

By Karamata's Theorem 2.B.1 (2),

$$\lim_{x\to\infty} \frac{\Psi^+(F(x)) - F(x)}{\int_x^\infty \frac{\Psi^+(F(t)) - F(t)}{t}\,dt} = -\lim_{x\to\infty} \frac{x(\Psi_x^+(F(x))F'(x) - F'(x))}{\Psi^+(F(x)) - F(x)} = -\rho^+ = 0.$$

In view of Karamata's Theorem 2.B.2 (2), since

$$\begin{aligned}\lim_{x\to\infty} \frac{\Pi^+(F(x))}{\int_x^\infty \frac{\Pi^+(F(t))}{t}\,dt} &= -\lim_{x\to\infty} \frac{x\Pi_x^+(F(x))F'(x)}{\Pi^+(F(x))} = \lim_{x\to\infty} \frac{x\Psi_x^+(F(x))F'(x)}{\Psi^+(F(x)) - F(x)} \\ &= \lim_{x\to\infty} \frac{xF'(x)}{\Psi^+(F(x)) - F(x)} = k \lim_{x\to\infty} \frac{xF'(x)}{(M - F(x))^{n+1}} = k\rho,\end{aligned}$$

$\Pi^+(F)$ is regularly varying at ∞ of index $-k\rho$, where $k > 0$ as a limit in (2.B.3) and $\rho > 0$. This finishes the first claim.

Step 2: To show the second claim regarding to the fixed point equation that ρ satisfies, we shall look at the limit of $f(\alpha, x)$ defined above as $\alpha \to \infty$. Firstly,

$$\begin{aligned}\lim_{\alpha\to\infty} \frac{\Psi^+(F(\alpha y)) - F(\alpha)}{(M - F(\alpha))^{n+1}} &= \lim_{\alpha\to\infty} \frac{\Psi^+(F(\alpha y)) - F(\alpha y)}{(M - F(\alpha))^{n+1}} + \lim_{\alpha\to\infty} \frac{F(\alpha y) - F(\alpha)}{(M - F(\alpha))^{n+1}} \\ &= \frac{1}{k} + \lim_{\alpha\to\infty} f(\alpha, x).\end{aligned}$$

Now, we shall see

$$\begin{aligned}\lim_{\alpha\to\infty} f(\alpha, x) &= \lim_{\alpha\to\infty} \frac{F(\alpha x) - F(\alpha)}{(M - F(\alpha))^{n+1}} \\ &= \lim_{\alpha\to\infty} \sum_{n\geq 1} \frac{p(n)}{1 - p(0)} \left\{ \int_0^\infty dz \left[\frac{1}{n} q\left(\frac{\sigma}{\alpha}, z - nx\right) + \frac{n-1}{n} \bar{q}\left(\frac{\sigma}{\alpha}, n, x, z\right)\right] \frac{\phi_F^-(\alpha x) - F(\alpha)}{(M - F(\alpha))^{n+1}} \right. \\ &\qquad \left. + \int_0^\infty dz \left[\frac{1}{n} q\left(\frac{\sigma}{\alpha}, z - nx\right) + \frac{n-1}{n} \bar{q}\left(\frac{\sigma}{\alpha}, n, x, z\right)\right] \frac{\phi_F^+(\alpha x) - F(\alpha)}{(M - F(\alpha))^{n+1}} \right\} \\ &= \lim_{\alpha\to\infty} \sum_{n\geq 1} \frac{p(n)}{1 - p(0)} \left\{ \frac{1}{n} \int_0^\infty dz\, q\left(\frac{\sigma}{\alpha}, z - nx\right) \frac{\phi_F^+(\alpha x) - F(\alpha)}{(M - F(\alpha))^{n+1}} \right. \\ &\qquad \left. + \frac{n-1}{n} \int_0^\infty dz\, \bar{q}\left(\frac{\sigma}{\alpha}, n, x, z\right) \frac{\phi_F^+(\alpha x) - F(\alpha)}{(M - F(\alpha))^{n+1}} \right\}.\end{aligned}$$

Note that

$$\lim_{\alpha\to\infty}\frac{\phi_F^+(\alpha x)-F(\alpha)}{(M-F(\alpha))^{n+1}}$$
$$=\sum_{n\geq 0}\frac{p(n)1_{[n\neq 0]}\int_{-\infty}^{\infty}dy\frac{\Pi^+(F(\alpha y))}{\Pi^+(F(\alpha))}q(\frac{\sigma}{\alpha n},y-\frac{z}{n})}{\sum_{n\geq 0}p(n)1_{[n\neq 0]}\int_{-\infty}^{\infty}du\frac{\Pi^+(F(\alpha u))}{\Pi^+(F(\alpha))}q(\frac{\sigma}{\alpha u},u-\frac{z}{n})}\left(\frac{1}{k}+f(\alpha,y)\right).$$

From Step 1, we have seen that $\frac{\Pi^+(F(\alpha y))}{\Pi^+(F(\alpha))}\to y^{-k\rho}$ as $\alpha\to\infty$. Combine with the fact that $q(\frac{\sigma}{\alpha n},y-\frac{z}{n})dy$ converges to the point mass at $\frac{z}{n}$ for any $z\in(-\infty,\infty)$ and $n\in\mathbb{N}_+$,

$$\lim_{\alpha\to\infty}\sum_{n\geq 0}\frac{p(n)1_{[n\neq 0]}\int_{-\infty}^{\infty}dy\frac{\Pi^+(F(\alpha y))}{\Pi^+(F(\alpha))}q(\frac{\sigma}{\alpha n},y-\frac{z}{n})}{\sum_{n\geq 0}p(n)1_{[n\neq 0]}\int_{-\infty}^{\infty}du\frac{\Pi^+(F(\alpha u))}{\Pi^+(F(\alpha))}q(\frac{\sigma}{\alpha u},u-\frac{z}{n})}f(\alpha,y)$$
$$=\sum_{n\geq 0}\frac{p(n)1_{[n\neq 0]}\left(\frac{z}{n}\right)^{-k\rho}}{\sum_{n\geq 0}p(n)1_{[n\neq 0]}\left(\frac{z}{n}\right)^{-k\rho}}\gamma(\frac{z}{n})=\sum_{n\geq 0}\frac{p(n)1_{[n\neq 0]}n^{k\rho}}{\sum_{n\geq 0}p(n)1_{[n\neq 0]}n^{k\rho}}\gamma(\frac{z}{n}),$$

where $\gamma(x):=\lim_{\alpha\to\infty}f(\alpha,x)=\rho\log x$ for some $\rho>0$. Now, define the probability function

$$p^1(n;k,\rho):=\frac{p(n)1_{[n\neq 0]}n^{k\rho}}{\sum_{n\geq 0}p(n)1_{[n\neq 0]}n^{k\rho}},$$

we can rewrite the limit

$$\lim_{\alpha\to\infty}\frac{\phi_F^+(\alpha x)-F(\alpha)}{(M-F(\alpha))^{n+1}}=\frac{1}{k}+E_{p^1}\left[\gamma\left(\frac{z}{N}\right)\right],\ \text{ where } N\sim p^1 \text{ defined as above.}$$

Therefore, according our assumption $\gamma(x)=\rho\log x$, we shall derive the desired fixed point equation for $\rho>0$.

$$\lim_{\alpha\to\infty}f(\alpha,x)=\gamma(x)=\frac{1}{k}+\sum_{n\geq 1}\frac{p(n)}{1-p(0)}1_{[n\neq 0]}\left\{\frac{1}{n}E_{p^1}\left[\gamma\left(\frac{nx}{N}\right)\right]+\frac{n-1}{nx}\int_0^x dy\,E_{p^1}\left[\gamma\left(\frac{ny}{N}\right)\right]\right\},$$
$$x\rho\log x=\frac{x}{k}+\sum_{n\geq 1}\frac{p(n)}{1-p(0)}1_{[n\neq 0]}\left\{\frac{x}{n}E_{p^1}\left[\rho\log\frac{nx}{N}\right]+\frac{n-1}{n}\int_0^x dy\,E_{p^1}\left[\rho\log\frac{ny}{N}\right]\right\},$$
$$\rho=\frac{1}{k}+\frac{\rho}{1-p(0)}E\left[1_{[N\neq 0]}\left(\log N+\frac{1}{N}\right)\right]-\rho\frac{E\left[1_{[N\neq 0]}N^{k\rho}\log N\right]}{E\left[1_{[N\neq 0]}N^{k\rho}\right]},\quad \rho>0.$$

Step 3: Finally, we can look into the asymptotics of $M-F$. From Step 1, we have the definition

$\lim_{\alpha\to\infty}\frac{g(\alpha x)}{g(\alpha)} = \exp(n\gamma(x)) = x^{n\rho}$, which implies that $g(x) = x^{n\rho}s(x)$ where $s(x)$ is some slowly varying function (as $x \to \infty$). In other words, $\log g(x) \equiv (M - F(x))^{-n} = n\rho \log x + \log s(x)$. Since $\lim_{x\to\infty}\frac{\log s(x)}{\log x} = 0$, we have

$$M - F(x) \sim (n\rho)^{-\frac{1}{n}}(\log x)^{-\frac{1}{n}}, \quad x \to \infty,$$

where ρ solves the fixed point equation in Step 2. □

Chapter 3

Equilibrium and Market Impact in Multi-period Limit Order Market

3.1 Introduction

There is a large literature on the dynamics of the limit order book (LOB). One important strand focuses on traders' decisions to place limit versus market orders. For example, Rosu (2009) develops a dynamic LOB model in which patient and impatient traders arrive according to independent Poisson processes and submit limit and market orders, respectively. The LOB evolves as a Markov process, and equilibrium is characterised by a recursive system for the expected utility of limit order traders. Related contributions include Goettler et al. (2009); Cohen et al. (1981); Foucault (1999); Parlour (2015), who also study the choice between market and limit orders in dynamic environments under different modelling assumptions.

In contrast, our multiperiod limit order market model does not consider the choice between limit and market orders. Instead, we assume a market structure: competitive liquidity suppliers post limit orders, while both noise and informed traders submit market orders. Our focus is on how informed trading interacts with the LOB over time, how the equilibrium is characterised, and how the book evolves through Bayesian updates of the liquidity suppliers' beliefs.

Assuming the presence of asymmetric information, our market structure in each trading period closely follows that of Çetin and Waelbroeck (2024). A single risky asset is traded in a limit order market, where competitive liquidity suppliers submit limit orders that are executed against batched market orders from noise and informed traders. After each trade, the limit order book is cleared, and liquidity suppliers revise their beliefs and resubmit their quotes. We therefore abstract away from order cancellations or modifications, since each trading period begins with a fresh book.

This structure combines elements from several canonical models. The interaction between informed traders, noise traders, and liquidity suppliers resembles Kyle (1985), although pricing in our setting arises from a limit order book rather than a single clearing price. The limit prices submitted by competitive (infinitely many) liquidity suppliers follow the equilibrium condition of Glosten (1994), based on conditional tail expectations that evolve dynamically over time. Belief updating by liquidity suppliers follows a Bayesian mechanism akin to Glosten and Milgrom (1985), where trade outcomes gradually reveal information about the asset's fundamental value. In contrast to the sequential framework of Glosten and Milgrom (1985), where a single trader (either informed or uninformed) arrives in each period and submits a unit-sized buy or sell order, our model allows each trading period to feature a *batched* order flow, in which both noise and informed traders submit their orders concurrently. However, within each batch, we assume that noise trades are given *execution priority* in the limit order book: their orders are consumed first, followed by those of informed traders. Liquidity suppliers cannot observe the decomposition of total order flow into noise and informed components, but informed traders are aware of this execution rule and understand that their orders are queued behind the noise trades. Moreover, both types of traders can submit *multi-unit orders*. The noise trade size is modelled as an exogenous random variable, independent of other sources of randomness in the market, while informed traders determine their trade sizes by maximising expected profits based on the observed market states and their private information. This structure more accurately reflects the aggregated nature of trading activity in modern electronic markets, where multiple orders are submitted in parallel and execution is not restricted to unit sizes.

Our model of a multiperiod limit order market with asymmetric information makes the following contributions:

- We model noise trades using a *location-scale t-distribution* with fixed degrees of freedom $\nu > 2$, capturing heavy-tailed noise order flow and allowing large-volume trades that contribute to price impact and pricing uncertainty.

- Informed traders are assumed to be *myopic*, leading to aggressive trading that enhances price discovery and enables tractable equilibrium characterisation.

- We construct the equilibrium via a sequence of fixed-point equations in Theorem 3.3.3 and Theorem 3.3.7, which facilitate the characterisation of optimal insider strategies and the shape of the LOB in equilibrium.

- We prove posterior consistency of liquidity suppliers' beliefs in Theorem 3.4.11, where belief updating is based on the full history of observed order flow. This setting departs from standard i.i.d. frameworks and captures the dependence structure induced by sequential trading data.
- We derive power-law asymptotics for market impact in Theorem 3.5.3, showing how the shape of the LOB evolves over time in response to the presence of informed traders and fat-tailed noise trading.

Noise trading distribution. At each trading date, noise traders submit aggregate market orders that are exogenous and unrelated to the asset's fundamental value. To better capture the empirical behaviour of large and irregular order flows, we model the noise trading volume using a location-scale t-distribution, in contrast to the Gaussian specification used in Çetin and Waelbroeck (2024). This generalisation accommodates large noise trades and yields a richer equilibrium structure. Empirical studies such as Gabaix et al. (2003) show that fluctuations in asset prices, trading volume, and the number of trades exhibit heavy-tailed power-law behaviour, challenging the adequacy of Gaussian models in capturing extreme market movements. In related theoretical work, such as Appendix D.2 of Saddier and Marsili (2024), the noise trading volume is assumed to follow an even distribution with tail decay $p(z) \sim C/|z|^{1+\nu}$, where ν is the tail index. This is consistent with observed tail exponents in financial markets, typically around $\nu \approx 5/2$ (e.g., Gopikrishnan et al., 2000; Farmer and Lillo, 2004; Lillo et al., 2005; Bouchaud et al., 2018).

Motivated by these findings, we assume that at each time t, the aggregate noise trading volume follows a *location-scale t-distribution*, $Z_t \sim t_\nu(0, \sigma)$, with fixed degrees of freedom $\nu > 2$ and scale parameter $\sigma > 0$. This specification allows for heavy-tailed behaviour, with the thickness of the tails governed by the parameter ν. Specifically, the distribution satisfies a power-law decay: $P(|Z_t| > z) \sim z^{-\nu}$ as $z \to \infty$. Smaller values of ν correspond to fatter tails and a higher likelihood of extreme noise trades, while as $\nu \to \infty$, the distribution becomes increasingly light-tailed and converges to a Gaussian distribution. For notational convenience, we absorb any dependence on the time partition (e.g., $\sqrt{\Delta t}$) into the scale parameter σ; this is standard in models with evenly spaced trading intervals and does not affect the probabilistic structure or asymptotic behaviour. Our primary focus is to understand the equilibrium price impact of large trades in the presence of such fat-tailed noise.

Myopic informed traders. We extend the model of Çetin and Waelbroeck (2024) by introducing a multiperiod setting. At each trading date, there are N_t *myopic* insiders (see, e.g., Brown and Jennings, 1989; Grundy and McNichols, 1989), each of whom knows the true fundamental value of the asset. These insiders seek to maximise expected profits within the current period and exit the market immediately afterward. The myopic assumption improves analytical tractability, since insiders do not split orders across periods, and captures the empirical tendency of informed traders to trade aggressively to exploit their informational advantage, thereby pushing prices toward the true value.

Equilibrium characterisation. Equilibrium is characterised through a recursive sequence of fixed-point equations, as given in (3.3.8). To establish the existence of an equilibrium, we prove the existence of a fixed point for the associated mapping using Schauder's fixed point theorem in Theorem 3.3.3. A key technical step in the proof is to show the monotonicity of the equilibrium limit price function, for which we leverage the Normal–Inverse Gamma representation of the location-scale t-distribution. Specifically, we construct a Brownian motion stopped at an independent inverse-gamma distributed time and apply a suitable Doob's h-transform to obtain the desired monotonicity result in Lemma 3.3.5.

Posterior consistency. Since liquidity suppliers must infer the fundamental value from aggregated order flow, which includes both noise and informed trades, they update their beliefs using Bayes' rule. A natural question is whether these beliefs become accurate over time. That is, whether the posterior concentrates around the true asset value as more trading volume is observed. This property, known as *posterior consistency*, refers to the convergence of the posterior distribution to the true parameter as the number of observations increases. It captures the idea that, under suitable conditions, the Bayesian updating mechanism learns the true value asymptotically.

The earliest rigorous result is due to Doob (1949), who proved almost sure convergence of the posterior under a prior with full support, assuming the model is well-specified. However, consistency at every parameter in a set of prior probability one is not sufficient. The importance of selecting an appropriate prior was subsequently emphasized by Freedman (1963, 1965); Diaconis and Freedman (1986), who constructed striking counterexamples where the posterior is inconsistent. The breakthrough result of Schwartz (1965) established posterior consistency under two key conditions: (i) the true distribution lies in the Kullback–Leibler support of the prior, and (ii) there exists a uniformly consistent sequence of tests for distinguishing the true model from alternatives. The latter is guaranteed by a strong separation condition, often expressed via the affinity between probability distributions, which is closely related to the Hellinger distance (see Choi and Ramamoorthi, 2008). Barron et al. (1999) further investigated the connection between posterior consistency and the existence of uniformly consistent tests, providing necessary and sufficient conditions. These fundamental developments are comprehensively reviewed in (Ghosh and Ramamoorthi, 2003, Chapter 4). Building on the foundational work of Schwartz (1965), later research extended posterior consistency beyond compact Euclidean settings. Notably, Barron et al. (1999), Ghosal et al. (1999), and Walker (2004) developed the theory under more general topologies, including the L^1 and weak topologies. A comprehensive overview of these developments is provided in (Ghosal and van der Vaart, 2017, Chapter 6) and Choi and Ramamoorthi (2008).

The theory has also been extended to settings with independent but non-identically distributed observations. For instance, Amewou-Atisso et al. (2003) established strong posterior consistency in semiparametric regression models, while Choi and Schervish (2007) considered general nonparametric regression. Choudhuri et al. (2004) focused on spectral density estimation and proved convergence in probability under milder conditions. These works show that posterior consistency can be achieved in a broad class of independent, non-identically distributed models. Ghosal and van der Vaart (2007) developed a general posterior contraction theory beyond the i.i.d. setting, with applications to non-identically distributed data, Markov processes, stationary Gaussian time series, and white noise models. While these cover important dependent sequences, the dependence structures considered are relatively structured and limited in generality. In contrast, Shalizi (2009) established posterior consistency under model misspecification for a much broader class of dependent data, including settings that go beyond Markovian or stationary assumptions. His framework accommodates general time series with minimal assumptions on the dependence structure.

In our multiperiod limit order market, trading volumes at each period depend on the full history of past order flows. Building on the framework of Shalizi (2009), but without allowing for model misspecification, we establish posterior consistency of liquidity suppliers' beliefs about the asset's distribution. This is achieved by interpreting the asset value as the latent parameter governing the law of total observed trading volume. Despite the strong dependence across time induced by strategic interaction, we show in Theorem 3.4.11 that under mild conditions on the prior, the posterior beliefs concentrate strongly around the true value as trading volume data accumulates.

Power-law price impact. Finally, since myopic informed traders tend to submit large orders, we analyse the asymptotic price impact as trade size tends to infinity. We show that, under mild conditions on the asset value distribution, the asymptotic price impact follows a power law. The tail exponent depends on the fat-tailedness of noise trading, the tail behaviour of the asset distribution, the degree of competition among insiders (i.e., the number of them), and the accumulation of historical price impact, as described in Theorem 3.5.3. The result implies that when noise trades are more likely to be large, the LOB becomes steeper, leading to greater price impact. This is likely because liquidity suppliers find it more difficult to disentangle informed trades from the batched order flow. By contrast, increased competition among informed traders enhances price discovery, resulting in a flatter LOB. This is consistent with the theoretical insights of Holden and Subrahmanyam (1992). Moreover, the price impact declines over time as the LOB flattens, a consequence of posterior consistency, whereby liquidity suppliers gradually learn the true value of the asset from the trading volume history.

The rest of the paper is organised as follows. Section 3.2 introduces the multiperiod trading model and market structure. Section 3.3 formulates the equilibrium conditions, characterises insider strategies and liquidity supplier beliefs, and establishes equilibrium existence. In Section 3.4, we prove posterior consistency of liquidity suppliers' beliefs about the fundamental asset value. Section 3.5 analyses the asymptotic market impact in equilibrium and derives related results on trading volume behaviour. Section 3.6 presents numerical illustrations of the model's implications. Finally, Section 3.7 presents some interesting observations arising from the model.

3.2 The Model and Market Structure

We consider a discrete-time financial market that operates over T trading periods, where T may be finite or infinite. Trading periods are indexed by $t = 0, 1, \ldots, T$. All random variables introduced in this section are defined on a complete probability space $(\Omega, \mathcal{F}, \mathcal{P})$, and E denotes the expectation operator associated with P.

In each period t, there are four types of agents in the market: *competitive (infinitely many) liquidity suppliers*, *a trading desk*, *noise traders*, and $N_t \geq 1$ risk-neutral *informed traders* (insiders). The asset being traded has a non-constant, continuously distributed fundamental value $V \in [m, M] \subset \mathbb{R}$, which is revealed at the terminal time. We assume the boundedness of V as a sufficient condition for equilibrium existence.

In the following, we describe how agents interact in the market at each trading period $t \in \{0, 1, \ldots, T\}$:

1. *Liquidity suppliers* move first and post limit orders, which collectively form the limit order book (LOB). The LOB is characterised by a non-decreasing pricing function

$$h_t(y) := h(t, Y_0^{t-1}, y) : \mathbb{N} \times \mathbb{R}^{t-1} \times \mathbb{R} \to \mathbb{R},$$

where $Y_0^{t-1} = (Y_1, \ldots, Y_{t-1})$ denotes the history of aggregate order flows up to period $t-1$, and $y \in \mathbb{R}$ is the current order size under consideration. The total cost of executing a market order of size x at time t is then given by

$$\int_0^x h_t(y)\, dy,$$

where $h_t(y)$ represents the marginal price of the y-th share. Liquidity suppliers are risk-neutral and competitive, setting prices to break even based on their current posterior beliefs about V.

2. *Noise traders* submit their aggregate market orders prior to the informed traders. Their total order flow at time t is denoted by a random variable $Z_t \sim t_\nu(0, \sigma)$, where $\nu > 2$ is the degrees of freedom and $\sigma > 0$ is a fixed scale parameter. Noise traders are assumed to be non-strategic: their orders are independent of the asset's fundamental value V and are not based on private information.

3. There are $N_t \geq 1$ *informed traders* at each time t, each of whom observes the true asset value V. These traders are risk-neutral and *myopic*: they maximise expected profits in the current period, taking the pricing function $h_t(\cdot)$ as given. They exit the market after trading once and do not consider future periods. Their actions are thus independent across time. The number of insiders N_t is deterministic and common knowledge.

   Let $x_{i,t}$ denote the trade of insider i at time t, and define the total insider demand:

$$X_t := \sum_{i=1}^{N_t} x_{i,t}.$$

   We assume symmetry among insiders, i.e., $x_{i,t} = x_t$ for all i, yielding

$$X_t = N_t x_t.$$

4. The *trading desk* acts as a broker and does not hold inventory. It aggregates orders from noise traders and insiders and executes them against the LOB. Any market order of size y is executed at total cost

$$\int_0^y h_t(Y + u)\, du, \tag{3.2.1}$$

where Y is the cumulative order flow already submitted in period t. While the value of Y is not observable by traders, the pricing rule (3.2.1) is publicly known and applies to all orders.

Noise traders and informed traders both act as clients of the trading desk, submitting market orders within each trading period. We assume that noise trades are submitted slightly earlier than informed trades and therefore receive execution priority. The total order flow at time t is given by

$$Y_t = Z_t + X_t,$$

where Z_t is the aggregate noise trade and X_t is the aggregate informed trade. For instance, if a single insider places a trade of size x, and the noise order in the same period is Z_t, the total execution cost incurred by the insider is

$$\int_0^x h_t(Z_t + u)\, du,$$

which reflects the price impact induced by queueing behind the earlier noise order.

3.3 The Equilibrium

In this section, we characterise the market equilibrium in our multiperiod trading setting. The equilibrium in each trading period consists of two interacting components:

1. The limit prices in the LOB set by competitive liquidity suppliers, based on their posterior beliefs about the asset's fundamental value.

2. The optimal trading strategies of myopic informed traders, who know the true fundamental value and choose their trade size to maximise expected profits within the current period.

Since informed traders are myopic and do not act across periods, we analyse equilibrium in a period-by-period fashion. In each period, an equilibrium is reached when the LOB pricing function reflects rational belief-based pricing, and the informed traders submit optimal order sizes given the pricing rule and their private information. The dynamic aspect of the model arises through the sequential belief updates of the liquidity suppliers: after each period, they observe the aggregate order flow and update their posterior accordingly. This sequence of belief-driven static equilibria across periods constitutes the overall multi-period equilibrium of the market.

3.3.1 Limit Prices in Equilibrium

As in Glosten (1994); Çetin and Waelbroeck (2024), we model the limit prices in the LOB using *tail expectations*. That is, at each time t, the LOB is characterised by the marginal pricing function:

$$h_t(y) := h(t, Y_0^{t-1}, y) = \begin{cases} E[V \mid Y_0^{t-1}, Y_t \geq y], & \text{if } y > 0, \\ E[V \mid Y_0^{t-1}, Y_t \leq y], & \text{if } y < 0, \end{cases} \tag{3.3.1}$$

where $Y_t = X_t + Z_t$ denotes the total order flow at time t, composed of the aggregate demand from N_t informed traders X_t, and the noise traders' demand Z_t. The first case in (3.3.1) defines the *limit ask price* for the dy-th share when $y > 0$, corresponding to a market buy order of size y. The second case defines the *limit bid price* for a market sell order of size $|y|$. The equilibrium limit prices (3.3.1) ensure that the expected aggregate profit of liquidity suppliers is zero, consistent with the assumption of perfect competition among infinitely many suppliers. If any supplier were to earn positive expected profit, others would undercut by offering more competitive pricing until the profit margin vanishes.

To see this, consider the expected profit from executing a total order Y_t, conditional on past order flow Y_0^{t-1}:

$$\begin{aligned} E\left[\int_0^{Y_t} (h_t(y) - V)\, dy \,\middle|\, Y_0^{t-1}\right] &= \int_0^{\infty} E\left[(h_t(y) - V)\,\mathbf{1}_{\{Y_t \geq y\}} \,\middle|\, Y_0^{t-1}\right] dy \\ &\quad + \int_{-\infty}^{0} E\left[(V - h_t(y))\,\mathbf{1}_{\{Y_t \leq y\}} \,\middle|\, Y_0^{t-1}\right] dy \\ &= 0, \end{aligned}$$

which follows from Fubini's theorem and the definition of $h_t(y)$ as the conditional expectation of V given $Y_t \geq y$ or $Y_t \leq y$.

The best ask and best bid prices correspond to the right and left limits at the origin, defined as

$$h_t(0+) := \lim_{y \downarrow 0} h_t(y), \qquad h_t(0-) := \lim_{y \uparrow 0} h_t(y),$$

and the bid-ask spread is given by the discontinuity $h_t(0+) - h_t(0-)$ in the marginal pricing function around zero.

3.3.2 Informed traders optimal strategies

Given the limit order book h_t at time t, a myopic insider chooses a trade size x_t to maximize her expected profit, conditional on the true fundamental value $V = v_0$ and the observed history of aggregate order flow Y_0^{t-1}. Her optimization problem is:

$$\max_{x_t \geq 0} E^{v_0}\left[Vx_t - \frac{x_t}{U_t + x_t} \int_0^{U_t + x_t} h_t(y + Z_t)\, dy \,\middle|\, Y_0^{t-1} \right], \tag{3.3.2}$$

where $E^{v_0}[\cdot]$ denotes expectation conditional on $V = v_0$, U_t is the aggregate demand from the other $N_t - 1$ informed traders at time t, and Z_t is the noise trade, which is assumed to arrive prior to the insider's trade.

Since the insider is myopic and only cares about immediate-period profit, she has an incentive to trade aggressively, aiming to push the market price closer to the true value v_0. In the limit, her optimal trade size would equate the marginal execution price with the asset's true fundamental value, exploiting her informational advantage while disregarding any future consequences.

Because the LOB is characterised by a non-decreasing pricing function h_t, the insider's objective function is concave in trade size x_t. Hence, the first-order condition is sufficient to characterise the unique optimal strategy. In the special case of a monopolistic insider, i.e., $N_t = 1$, the optimal trade size $x_t^\star$ satisfies:

$$x_t^\star = F_t^{-1}(v_0),$$

where the function F_t is defined as

$$\begin{aligned} F_t(x) := F(t, Y_0^{t-1}, x) &= \int_{-\infty}^{+\infty} h(t, Y_0^{t-1}, z + x)\, q_\nu(\sigma, z)\, dz \\ &= \int_{-\infty}^{+\infty} h_t(z + x)\, q_\nu(\sigma, z)\, dz, \end{aligned}$$

and $q_\nu(\sigma, \cdot)$ denotes the density of a location-scale t-distribution with location 0, scale σ, and degrees of freedom ν. The function $F_t(x)$ represents the expected marginal price faced by the insider, integrated over the distribution of noise trades. Since h_t is non-decreasing, F_t is strictly increasing, ensuring invertibility and uniqueness of the optimal trade size.

Suppose there are $N_t > 1$ *symmetric, myopic* informed traders at time t, each observing the true fundamental value $V = v_0$. An individual insider, taking the behaviour of the others as given, chooses a trade size x_t to maximise her expected profit:

$$E^{v_0}\left[Vx_t - \frac{x_t}{U_t + x_t}\int_0^{U_t+x_t} h_t(Z_t + y)\, dy \,\middle|\, Y_0^{t-1}\right],$$

where U_t denotes the aggregate demand from the other $N_t - 1$ insiders, and Z_t is the noise trade.

Taking the first-order condition with respect to x_t, we obtain:

$$v_0 = E^{v_0}\left[\frac{x_t}{U_t + x_t} h_t(Z_t + U_t + x_t) + \frac{U_t}{(U_t + x_t)^2}\int_0^{U_t+x_t} h_t(Z_t + y)\, dy \,\middle|\, Y_0^{t-1}\right].$$

Assuming symmetric behaviour among the myopic insiders, i.e., $x_{i,t} = x_t^\star$ for all i, we define:

$$X_t^\star := \sum_{i=1}^{N_t} x_{i,t} = N_t x_t^\star, \qquad U_t = (N_t - 1)x_t^\star.$$

Under this assumption, the first-order condition simplifies to:

$$v_0 = E^{v_0}\left[\frac{1}{N_t} h_t(Z_t + X_t^\star) + \frac{N_t - 1}{N_t X_t^\star}\int_0^{X_t^\star} h_t(Z_t + u)\, du \,\middle|\, Y_0^{t-1} = y_0^{t-1}\right].$$

We then define the function $F_t : \mathbb{R} \to \mathbb{R}$ by:

$$F_t(x) := F(t, Y_0^{t-1}, x) := E^{v_0}\left[\frac{1}{N_t} h_t(Z_t + x) + \frac{N_t - 1}{N_t x}\int_0^{x} h_t(Z_t + u)\, du \,\middle|\, Y_0^{t-1}\right], \tag{3.3.3}$$

which represents the expected marginal price perceived by an insider submitting a trade of size x, given the observed order flow history y_0^{t-1} and the true value $V = v_0$.

Given that the marginal pricing function $h_t(\cdot)$ is non-decreasing, the function $F_t(\cdot)$ in (3.3.3) is strictly increasing. Consequently, the equilibrium aggregate and individual insider demands are given by

$$X_t^\star = F_t^{-1}(v_0), \qquad x_t^\star = \frac{1}{N_t} F_t^{-1}(v_0). \tag{3.3.4}$$

3.3.3 Liquidity suppliers' belief updates

Liquidity suppliers update their beliefs about the fundamental value V using Bayes' rule, conditioned on the observed history of aggregate order flows. Following Glosten and Milgrom (1985), prior to trading at time t, after observing the sequence $Y_0^{t-1} = (Y_0, \ldots, Y_{t-1})$, their conditional belief about V satisfies:

$$P(V \in dv \mid Y_0^{t-1}) \propto P(Y_{t-1} \in dy_{t-1} \mid Y_0^{t-2}, V = v) \cdot P(Y_0^{t-2} \in dY_0^{t-2}, V \in dv).$$

From the equilibrium strategy of the insiders described in Section 3.3.2, the aggregate insider demand at time $t-1$ is deterministic given the history and fundamental value:

$$X_{t-1}^{\star} = F_{t-1}^{-1}(v),$$

where v is the realisation of the true fundamental value. The total order flow then satisfies

$$Y_{t-1} = X_{t-1}^{\star} + Z_{t-1},$$

with $Z_{t-1} \sim t_\nu(0, \sigma)$ representing noise trader demand, independent of V. It follows that the conditional distribution of Y_{t-1} given V and Y_0^{t-2} is:

$$(Y_{t-1} \mid Y_0^{t-2}, V) \sim t_\nu(F_{t-1}^{-1}(V), \sigma).$$

This implies the recursive form for the posterior distribution of V given Y_0^{t-1}:

$$\begin{aligned} P(V \in dv \mid Y_0^{t-1}) \propto & \left(1 + \frac{1}{\nu}\left(\frac{y_{t-1} - F_{t-1}^{-1}(v)}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}} \\ & \times P(Y_0^{t-2} \in dY_0^{t-2}, V \in dv), \end{aligned}$$

and by recursion back to time 0:

$$P(V \in dv \mid Y_0^{t-1}) \propto P(V \in dv) \prod_{i=0}^{t-1} \left(1 + \frac{1}{\nu}\left(\frac{y_i - F_i^{-1}(v)}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}}. \tag{3.3.5}$$

This expression captures how, in equilibrium, liquidity suppliers revise their beliefs about the fundamental value using the full history of observed order flows, and how informed traders respond optimally given their private information.

Definition 3.3.1. *Let $E_t[\cdot] := E[\cdot \mid Y_0^{t-1} = y_0^{t-1}]$ and $P_t(\cdot) := P(\cdot \mid Y_0^{t-1} = y_0^{t-1})$ denote the conditional expectation and probability operators given the observed order flow history up to time $t-1$.*

The posterior distribution (3.3.5) feeds directly into the liquidity suppliers' pricing through the equilibrium condition (3.3.1). At each time t, the updated belief about V, encoded in P_t, determines the conditional expectations $E_t[\cdot]$ used to calculate the marginal prices in the limit order book.

3.3.4 Equilibrium characterisation

We now formally define a dynamic equilibrium as a sequence of pricing and trading strategies that are mutually optimal and consistent with Bayesian belief updates.

Definition 3.3.2 (Equilibrium)**.** *The sequence $(h_t^\star, x_t^\star)_{t\in\mathbb{N}}$ constitutes an equilibrium if the limit pricing functions $h_t^\star$ are non-decreasing and non-constant for all $t \in \mathbb{N}$, and the insider trade sizes $x_t^\star \in \mathbb{R}$, such that:*

(i) Each insider's trade size $x_t^\star$ maximises expected profit given the pricing function $h_t^\star$, as in (3.3.3) *and* (3.3.4)*;*

(ii) The pricing function $h_t^\star$ satisfies the zero-profit condition for competitive liquidity suppliers, as given in (3.3.1)*, with total order flow $Y_t^\star = N_t x_t^\star + Z_t$.*

To characterise the equilibrium pricing rule $h_t^\star(\cdot)$, we define the following conditional expectation and probability functions with respect to the posterior belief over the fundamental value V, based on the observed trading volume history up to time $t-1$:

$$\Phi_t^+(y) := E_t\left[V\mathbb{1}_{\{V\geq y\}}\right], \quad \Pi_t^+(y) := P_t\left(V \geq y\right),$$
$$\Phi_t^-(y) := E_t\left[V\mathbb{1}_{\{V< y\}}\right], \quad \Pi_t^-(y) := P_t\left(V < y\right),$$

and define the corresponding conditional tail expectations on the support of V:

$$\Psi_t^\pm(y) := \frac{\Phi_t^\pm(y)}{\Pi_t^\pm(y)},$$

whenever the denominator is strictly positive. That is,

$$\Psi_t^+(y) = E_t[V \mid V \geq y], \qquad \Psi_t^-(y) = E_t[V \mid V < y].$$

These functions determine the marginal prices in the limit order book via the equilibrium condition (3.3.1). To see this, consider the limit ask price at time t for an incoming market buy order of size $y > 0$. In equilibrium, we have:

$$\begin{aligned} h_t^\star(y) = E_t[V \mid Y_t^\star \geq y] &= \frac{E_t\left[V 1_{\{X_t^\star + Z_t \geq y\}}\right]}{P_t\left(X_t^\star + Z_t \geq y\right)} \\ &= \frac{E_t\left[V 1_{\{V \geq F_t(y - Z_t)\}}\right]}{P_t\left(V \geq F_t(y - Z_t)\right)} = \frac{\int_{-\infty}^{\infty} \Phi_t^+\left(F_t(y-z)\right) q_\nu(\sigma, z)\, dz}{\int_{-\infty}^{\infty} \Pi_t^+\left(F_t(y-z)\right) q_\nu(\sigma, z)\, dz}, \end{aligned} \tag{3.3.6}$$

where $q_\nu(\sigma, z)$ denotes the density of a location-scale t-distribution with location parameter 0, scale σ, and degrees of freedom ν.

Similarly, for $y < 0$, the equilibrium limit bid price is given by:

$$h_t^\star(y) = \frac{\int_{-\infty}^{\infty} \Phi_t^-\left(F_t(y-z)\right) q_\nu(\sigma, z)\, dz}{\int_{-\infty}^{\infty} \Pi_t^-\left(F_t(y-z)\right) q_\nu(\sigma, z)\, dz}. \tag{3.3.7}$$

We shall soon see that the equilibrium is characterised by a fixed-point mapping equation involving F_t for each t. To facilitate this, we define the following mappings associated with any continuous function $g_t : \mathbb{R} \to \mathbb{R}$:

$$\begin{aligned} \phi_{g_t}^+(x) &:= \frac{\int_{-\infty}^{\infty} \Phi_t^+\left(g_t(z)\right) q_\nu(\sigma, x - z)\, dz}{\int_{-\infty}^{\infty} \Pi_t^+\left(g_t(z)\right) q_\nu(\sigma, x - z)\, dz}, \\ \phi_{g_t}^-(x) &:= \frac{\int_{-\infty}^{\infty} \Phi_t^-\left(g_t(z)\right) q_\nu(\sigma, x - z)\, dz}{\int_{-\infty}^{\infty} \Pi_t^-\left(g_t(z)\right) q_\nu(\sigma, x - z)\, dz}. \end{aligned}$$

An important modeling advantage arises here: the normalising constant of the posterior distribution P_t, which is embedded in both $\Phi_t^\pm$ and $\Pi_t^\pm$, cancels out in the above expressions. As a result, the liquidity suppliers' belief update rule (3.3.5) is sufficient for characterising the equilibrium, and no further normalisation is required.

We also define the piecewise mapping $\phi_{g_t} : \mathbb{R} \to \mathbb{R}$ by:

$$\phi_{g_t}(x) := \phi_{g_t}^+(x) \cdot 1_{\{x \geq 0\}} + \phi_{g_t}^-(x) \cdot 1_{\{x < 0\}}.$$

Combining the definition of ϕ_{g_t} with the insider's optimality condition (3.3.3) and the liquidity suppliers' equilibrium pricing conditions (3.3.6)–(3.3.7), we obtain the following fixed-point equation for F_t:

$$F_t(x) = \int_{-\infty}^{+\infty} \left\{ \frac{1}{N_t} q_v(\sigma, x - z) + \frac{N_t - 1}{N_t} \bar{q}_v(\sigma, x, z) \right\} \phi_{F_t}(z)\, dz, \tag{3.3.8}$$

where $\bar{q}_v$ is defined by:

$$\bar{q}_v(\sigma, x, z) := 1_{\{x \neq 0\}} \cdot \frac{1}{x} \int_0^x q_v(\sigma, y - z)\, dy + 1_{\{x=0\}} \cdot q_v(\sigma, z).$$

The function F_t must satisfy the fixed-point equation (3.3.8), which is a necessary condition for the existence of an equilibrium.

Theorem 3.3.3. *Equilibrium exists if and only if there exists a function $F_t : \mathbb{R} \to \mathbb{R}$ for each $t \in \mathbb{N}$ that satisfies (3.3.8). Given such a solution $\{F_t\}$, the pair $(h_t^\star, x_t^\star)_{t \in \mathbb{N}}$ constitutes an equilibrium, where $x_t^\star = \frac{1}{N_t} F_t^{-1}(V)$ and $h_t^\star$ is defined via (3.3.1).*

3.3.5 Existence of Equilibrium

To establish the existence of an equilibrium, we first specify the support of the fundamental value random variable V. Let $\operatorname{supp}(V) = [m, M]$, where we assume that V is non-degenerate, so that $-\infty < m < M < \infty$.

In addition, we impose the following integrability condition, which ensures that we can interchange limits and integration in the analysis that follows.

Assumption 3.3.4. *For each t, we assume that $\phi_{F_t}(z)$ satisfies the integrability condition*

$$\int_{-\infty}^{\infty} \phi_{F_t}(z)\, q_v(\sigma, z)\, dz < \infty$$

This assumption is automatically satisfied when V is bounded, since in that case $\phi_{F_t}(z)$ is uniformly bounded as well.

In view of Theorem 3.3.3, the existence of an equilibrium reduces to showing that the fixed-point integral equation (3.3.8) admits a solution. Before presenting the full proof, we introduce the following Lemma 3.3.5, which establishes key properties of the fixed-point mapping and lays the groundwork for applying Schauder's fixed-point theorem to prove the existence of a market equilibrium as defined in Definition 3.3.2.

A similar monotonicity result was established in (Çetin and Waelbroeck, 2024), under the assumption of Gaussian-distributed noise trades. There, the monotonicity of the mapping

$\phi_g(y)$ in y, for any non-decreasing function g, is proved by constructing a bounded martingale driven by Brownian motion and applying Doob's h-transform.

In our setting, however, the noise trades follow a fat-tailed location-scale Student-t distribution, for which the Gaussian-based arguments do not apply directly. To address this, we exploit the Normal–Inverse Gamma representation of the location-scale t-distribution; see Appendix 3.A.2. Specifically, a location-scale t_ν random variable can be expressed as a Gaussian random variable with variance scaled by an independent inverse-gamma random variable. This enables us to reinterpret the noise trade as a Brownian motion stopped at a random (inverse-gamma distributed) time. Using this representation, we adapt the original martingale and h-transform construction to our setting, thereby recovering the desired monotonicity and regularity properties of ϕ_g under location-scale t distributed heavy-tailed noise.

To simplify notation, we state the following result for generic functions $\Pi^{\pm}$, $\Psi^{\pm}$, and g without an explicit time index. However, in our model these quantities depend on the trading period t, so the lemma should be understood as applying to each trading date separately.

Lemma 3.3.5. *Let $g : \mathbb{R} \mapsto [m, M]$ be a continuous function, and consider a Brownian motion B and an independent random variable $T \sim$ Inv-Gamma$(\frac{\nu}{2}, \frac{\sigma^2\nu}{2})$ on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Define the functions*

$$u^+(t,z) = \mathbb{E}\big[\Pi^+(g(B_T))1_{T>t}\big|B_t = z\big],$$
$$u^-(t,z) = \mathbb{E}\big[\Pi^-(g(B_T))1_{T>t}\big|B_t = z\big].$$

Let $(\mathcal{F}_t)_{t\geq 0}$ minimal filtration satisfying the usual conditions such that T is an $(\mathcal{F}_t)_{t\geq 0}$-stopping time and B is an $(\mathcal{F}_t)_{t\geq 0}$-Brownian motion. Then, the following hold:
(1) Define the probability measures $(\mathbb{Q}^+, \mathbb{Q}^-)$ on $\mathcal{F}$ by

$$\frac{d\mathbb{Q}^{\pm}}{d\mathbb{P}} = \frac{\Pi^{\pm}(g(B_T))}{\mathbb{E}[\Pi^{\pm}(g(B_T))]}.$$

Then, on $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{Q})$, B follows

$$dB_t = dW_t^{\mathbb{Q}^{\pm}} + \frac{u_z(t, B_t)}{u(t, B_t)}dt, \quad t < T, \quad B_0 = x, \tag{3.3.9}$$

where $W^{\mathbb{Q}^{\pm}}$ is a standard Brownian motion under $\mathbb{Q}^{\pm}$ with $W_0 = 0$.
(2)Suppose $B_0 = x$, $\mathbb{P}$-a.s.. Then, $\phi_g^+(x) = E^{\mathbb{Q}^+}[\Psi^+(g(B_T)))]$ and $\phi_g^-(x) =$

$E^{\mathbb{Q}^-}[\Psi^-(g(B_T))]$.

(3) $\phi_g^+(0) > \phi_g^-(0)$.

(4) $\phi_g^\pm(x)$ *is non-decreasing, if* g *is non-decreasing. Consequently,* ϕ_g *is non-decreasing.*

Proof. We present the proof for u^+. The argument for u^- follows analogously.

(1) Note that $M_t := \mathbb{E}[\Pi^+(g(B_T))|\mathcal{F}_t]$ is a uniformly integrable $(\mathbb{P}, \mathcal{F}_t)$-martingale. Moreover,

$$\begin{aligned} 1_{T>t}M_t &= 1_{T>t}\mathbb{E}[\Pi^+(g(B_T))|\mathcal{F}_t] \\ &= 1_{T>t}\frac{\mathbb{E}[\Pi^+(g(B_T))1_{T>t}|\mathcal{F}_t^B]}{\mathbb{P}(T>t)} \\ &= 1_{T>t}\frac{u^+(t,B_t)}{\mathbb{P}(T>t)}, \end{aligned}$$

where $\mathcal{F}^B$ is the natural filtration of B. In above, the second line follows from the independence of T and B under $\mathbb{P}$, and the third line is due to the Markov property of B and, once again, its independence from T under $\mathbb{P}$. The claim now follows from a direct application of Girsanov's theorem.

(2) Recall that $\phi_g^+(x) = \frac{E[\Phi^+(g(Y))]}{E[\Pi^+(g(Y))]}$, where $Y \sim t_\nu(x, \sigma)$. From the normal-inverse gamma representation of the location-scale Student-t distribution (see 3.A.2), suppose

$$\begin{aligned} Y \mid T &\sim \mathcal{N}(x, T), \\ T &\sim \text{Inv-Gamma}\left(\frac{\nu}{2}, \frac{\sigma^2\nu}{2}\right), \end{aligned}$$

then the marginal distribution of Y is the location-scale Student-t distribution:

$$Y \sim t_\nu(x, \sigma).$$

This implies that $Y := B_T$, where $B_0 = x$, can be interpreted as a Brownian motion starting from x and observed at an independent random time $T \sim \text{Inv-Gamma}(\nu/2, \sigma^2\nu/2)$.

Therefore,

$$\begin{aligned}\phi_g^+(x) &= \frac{E[\Phi^+(g(Y))]}{E[\Pi^+(g(Y))]} = \frac{E[\Phi^+(g(B_T))]}{E[\Pi^+(g(B_T))]}\\ &= \frac{E[\Psi^+(g(B_T))\Pi^+(g(B_T))]}{E[\Pi^+(g(B_T))]}\\ &= E^{\mathbb{Q}^+}[\Psi^+(g(B_T))].\end{aligned}$$

Similarly, $\phi_g^-(x) = E^{\mathbb{Q}^-}[\Psi^-(g(B_T))]$.
(3) Note that $\phi_g^\pm(x)$ can be written as $\phi_g^\pm(x) = \frac{E[\Phi^\pm(g(x+R))]}{E[\Pi^\pm(g(x+R))]}$, where $R \sim t_\nu(0,1)$. To show $\phi_g^+(0) > \phi_g^-(0)$, it is equivalent to verify

$$E[\Phi^+(g(R))]E[\Pi^-(g(R))] - E[\Phi^-(g(R))]E[\Pi^+(g(R))] > 0.$$

Recall that $\Pi^+ = 1 - \Pi^-$ and $\Phi^+ + \Phi^- = E[V]$, the above can be written as

$$E[\Phi^+(g(R))] - E[V]E[\Pi^+(g(R))] > 0,$$

which is further equivalent to

$$\int_{-\infty}^{\infty} \left(\Psi^+(g(r)) - E[V]\right) \Pi^+(g(r))q_\nu(\sigma,r)\,dr > 0.$$

This is true as for any x we have $\Psi^+(x) \geq \Psi^+(m) = E[V]$ and Ψ^+ is not constant.

(4) Now, we want to show that $\phi_g^+(x)$ is non-decreasing given a non-decreasing g. First, note that $\phi_g^+(x) = E^{\mathbb{Q}^+}[\Psi^+(g(B_T))]$, where B_t solves (3.3.9) for $t < T$. Since g is non-decreasing and Π^+ is non-increasing, we have $u_z^+(t,z) \leq 0$. Therefore, since the drift $\frac{u_z^+(t,z)}{u^+(t,z)}$ is locally Lipschitz in z, we can construct strong solutions to (3.3.9) starting from different initial values $x \in \mathbb{R}$ on a common probability space using the same Brownian motion. This allows the use of pathwise comparison, if $y \geq x$,

$$B_t^y \geq B_t^x \quad \text{for all } t < T, \quad \mathbb{Q}^+\text{-a.s.}$$

As Ψ^+ is non-decreasing:

$$\phi_g^+(y) = \mathbb{E}^{\mathbb{Q}^+}\left[\Psi^+(g(B_T^{(y)}))\right] \geq \mathbb{E}^{\mathbb{Q}^+}\left[\Psi^+(g(B_T^{(x)}))\right] = \phi_g^+(x), \quad y \geq x.$$

Similar arguments apply to $\phi_g^-(x)$ in conjunction with $\phi_g^+(0) > \phi_g^-(0)$, we obtain the desired monotonicity of ϕ_g. □

Based on this lemma, combined with the Assumption 3.3.4, we shall have one of the key properties of the marginal cost function F_t in each trading period. That is, it is strictly increasing, with well-defined limits at both extremes.

Lemma 3.3.6. *Assume Assumption 3.3.4, in each trading date $t \in \mathbb{N}$, let F_t be a continuous non-decreasing solution of* (3.3.8)*. If $-\infty < m < M < \infty$ is the support of the trading asset,* $\lim_{x\to\infty} F_t(x) = M$ *and* $\lim_{x\to-\infty} F_t(x) = m$*. Consequently, F is strictly increasing.*

Proof. Given that F_t is the solution of (3.3.8), and $V \in [m, M]$, it follows that $F_t(x) \leq M$, for all $x \in \mathbb{R}$. We now prove $\lim_{x\to\infty} F_t(x) = M$ by contradiction. Assume $\lim_{x\to\infty} F_t(x) := L < M$.

Firstly, Monotone convergence theorem in conjunction with integrable assumption implies

$$\lim_{x\to\infty} \int_{-\infty}^{\infty} q_\nu(\sigma, z)\phi_{F_t}(x+z)\,dz = \int_{-\infty}^{\infty} q_\nu(\sigma, z) \lim_{x\to\infty} \phi_{F_t}(x+z)\,dz.$$

Moreover, $\lim_{x\to\infty} \phi_{F_t}(x+z) = \Psi^+(F_t(\infty)-)$. To see this, first note that, for $x+z>0$,

$$\phi_{F_t}^+(z+x) = \frac{\int_{-\infty}^{\infty} \Psi_t^+(F_t(u))\Pi_t^+(F_t(u)) \left(1 + \frac{(x+z-u)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} du}{\int_{-\infty}^{\infty} \Pi_t^+(F_t(u)) \left(1 + \frac{(x+z-u)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} du}. \tag{3.3.10}$$

Next, the measure

$$\frac{\Pi_t^+(F_t(u)) \left(1 + \frac{(x+z-u)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} du}{\int_{-\infty}^{\infty} \Pi_t^+(F_t(u)) \left(1 + \frac{(x+z-u)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} du}$$

converges to the point mass at ∞. That is, we need to show that, for any $0 < a < \infty$, we have

$$\lim_{x\to\infty} \frac{\int_{-\infty}^{a} \Pi_t^+(F_t(u)) \left(1 + \frac{(x+z-u)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} du}{\int_{-\infty}^{\infty} \Pi_t^+(F_t(u)) \left(1 + \frac{(x+z-u)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} du} = 0$$

Firstly, we observe that the upper bound of the numerator goes to 0 when $x \to \infty$:

$$\begin{aligned}
&\int_{-\infty}^{a} \Pi_t^+(F_t(u)) \left(1 + \frac{(x+z-u)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} du \\
\leq &\int_{-\infty}^{a} \left(1 + \frac{(x+z-u)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} du \\
\leq &\int_{-\infty}^{a} \left(\frac{(x+z-u)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} du \\
= &(\nu\sigma^2)^{\frac{\nu+1}{2}} \int_{-\infty}^{a} (x+z-u)^{-\nu-1}\, du \\
= &\nu^{\frac{\nu-1}{2}} \sigma^{\nu+1} (x+z-a)^{-\nu},
\end{aligned}$$

where $\nu > 2$.

Then, we need to show the lower bound of the denominator is bounded below away from 0:

$$\begin{aligned}
&\int_{-\infty}^{\infty} \Pi_t^+(F_t(u)) \left(1 + \frac{(x+z-u)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} du \\
= &\int_{-\infty}^{\infty} \Pi^+(F_t(x+z+k)) \left(1 + \frac{k^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} dk \\
\geq &\int_{-\delta}^{\delta} \Pi_t^+(F_t(x+z+k)) \left(1 + \frac{k^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} dk \\
\geq &\left(1 + \frac{\delta^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} \int_{-\delta}^{\delta} \Pi_t^+(F_t(x+z+k))\, dk > 0.
\end{aligned}$$

This is because we assume $F_t(x) \to L < M$ and $\Pi_t^+(F_t(u)) = P_t(V > F_t(u))$, and since $P_t(V > v) > 0$ for all $v < M$, the integrand is bounded below away from 0 on $[-\delta, \delta]$ for large x. Putting the numerator and the denominator together, for any $0 < a < \infty$ and $\delta > 0$,

$$\begin{aligned}
&\frac{\int_{-\infty}^{a} \Pi_t^+(F_t(u)) \left(1 + \frac{(x+z-u)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} du}{\int_{-\infty}^{\infty} \Pi_t^+(F_t(u)) \left(1 + \frac{(x+z-u)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} du} \\
\leq &\frac{\nu^{\frac{\nu-1}{2}} \sigma^{\nu+1} (x+z-a)^{-\nu}}{\left(1 + \frac{\delta^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} \int_{-\delta}^{\delta} \Pi_t^+(F_t(x+z+k))\, dk},
\end{aligned}$$

which goes to 0 as $x \to \infty$. Returning to (3.3.10), we conclude

$$\lim_{x\to\infty} \phi_{F_t}(x+z) = \Psi_t^+(F_t(\infty)-).$$

Plugging this into the representation of F_t in (3.3.8), we deduce

$$F_t(\infty) = \Psi_t^+(F_t(\infty)-). \tag{3.3.11}$$

However, the map Ψ_t^+ satisfies $\Psi_t^+(x-) > x$ for all $x < M$, so the identity in (3.3.11) contradicts the assumption $\lim_{x\to\infty} F(x) < M$. Therefore, we must have

$$\lim_{x\to\infty} F_t(x) = M.$$

Similarly, $\lim_{x\to-\infty} F_t(x) = m$. Since $m \neq M$, and that F_t solves (3.3.8), where $\phi_{F_t}(x)$ is non-decreasing in view of Lemma 3.3.5 (4), we see that F_t is strictly increasing. □

With this in place, we can establish the existence of an equilibrium as defined in Definition 3.3.2. Note that when V has bounded support, as assumed in the following theorem, Assumption 3.3.4 is automatically satisfied.

Theorem 3.3.7. *Suppose the fundamental value V has support contained in a bounded interval $[m, M] \subset \mathbb{R}$, with $-\infty < m < M < \infty$. Then, at each trading date $t \in \mathbb{N}$, there exists an equilibrium.*

Proof of Theorem 3.3.7. The proof proceeds via an application of Schauder's fixed point theorem. Since $V \in [m, M]$ by assumption, the conditional expectations $\Psi_t^\pm(y)$ also take values in $[m, M]$. Thus, the mapping ϕ_{g_t} defined in terms of $\Psi_t^\pm$ inherits boundedness in this interval in view of Lemma 3.3.5.

We now consider the fixed-point equation (3.3.8) at each trading date t. If a solution $F_t \in \mathcal{C}$, then this yields an equilibrium according to Definition 3.3.2.

First, observe that

$$\frac{\partial}{\partial x} q_\nu(\sigma, z-x) = -q_\nu(\sigma, z-x)\frac{(\nu+1)(x-z)}{\nu\sigma^2 + (x-z)^2}, \tag{3.3.12}$$

$$\begin{aligned}\frac{\partial \bar{q}_\nu(\sigma, x, z)}{\partial x} &= \frac{q_\nu(\sigma, x-z) - \bar{q}_\nu(\sigma, x, z)}{x} = \frac{1}{x^2}\int_0^x \{q_\nu(\sigma, x-z) - q_\nu(\sigma, y-z)\}\, dy \\ &= \frac{1}{x^2}\int_0^x u\, \frac{\partial}{\partial u} q_\nu(\sigma, u-z)\, du.\end{aligned} \tag{3.3.13}$$

Define

$$I := \int_{-\infty}^{\infty} \left| \frac{\partial}{\partial u} q_\nu(\sigma, u) \right| du. \tag{3.3.14}$$

As shown in Lemma 3.A.1, I is strictly positive and finite. Therefore, we have the following bound:

$$\begin{aligned}
\left| \frac{d}{dx} F_t(x) \right| &\le \int_{-\infty}^{\infty} |\phi_{F_t}(z)| \left| \frac{1}{N_t} \frac{\partial}{\partial x} q_\nu(\sigma, x - z) + \frac{N-1}{N} \frac{\partial}{\partial x} \bar{q}_\nu(\sigma, x, z) \right| dz \\
&\le (|m| + |M|) \left\{ \frac{1}{N_t} \int_{-\infty}^{\infty} \left| \frac{\partial}{\partial x} q_\nu(\sigma, z - x) \right| dz + \frac{N_t - 1}{N_t x^2} \int_0^x |u| \, du \int_{-\infty}^{\infty} \left| \frac{\partial}{\partial u} q_\nu(\sigma, u - z) \right| dz \right\} \\
&= (|m| + |M|) \left\{ \frac{1}{N_t} I + \frac{N_t - 1}{N_t x^2} \int_0^x |u| \cdot I \, du \right\} \\
&= (|m| + |M|) \left\{ \frac{1}{N_t} I + \frac{N_t - 1}{2N_t} I \right\} \\
&= (|m| + |M|) \cdot \frac{N_t + 1}{2N_t} \cdot I.
\end{aligned}$$

Define the constant

$$K_0 := (|m| + |M|) \cdot \frac{N_t + 1}{2N_t} \cdot I < \infty.$$

Then F is differentiable with derivative uniformly bounded by K_0, i.e.,

$$\left| \frac{d}{dx} F_t(x) \right| \le K_0 \quad \text{for all } x \in \mathbb{R}.$$

We shall show the existence of a fixed point in the normed space $\chi := L^2(\mathbb{R}, \mu_0)$, i.e., the space of Borel measurable functions that are square integrable with respect to μ_0, where

$$\mu_0(dx) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma} \left(1 + \frac{x^2}{\nu\sigma^2} \right)^{-\frac{\nu+1}{2}} dx.$$

Define

$$D_0 = \{ g_t \mid g_t : \mathbb{R} \mapsto [m, M] \text{ is such that } |g_t(x) - g_t(y)| \le K_0 |x - y|, \forall x, y \in \mathbb{R} \},$$

and

$$D = \{ g_t \in \chi \mid g_t = g_0, \ \mu_0\text{-a.e. for some } g_0 \in D_0 \}.$$

We can see that D is a convex subset of χ.

Now, define the operator T on χ via

$$Tg_t(x) := \int_{-\infty}^{\infty} \left\{ \frac{1}{N_t} q_\nu(\sigma, x-z) + \frac{N_t - 1}{N_t} \bar{q}_\nu(\sigma, x, z) \right\} \phi_{\bar{g}_t}(z)\, dz,$$

where $\bar{g}_t := (g_t \vee m) \wedge M$, and define:

$$\begin{aligned} \phi^{\pm}_{g_t}(x) &= \frac{\int_{-\infty}^{\infty} \Phi^{\pm}_t(g_t(y)) q_\nu(\sigma, y-x)\, dy}{\int_{-\infty}^{\infty} \Pi^{\pm}_t(g_t(y)) q_\nu(\sigma, y-x)\, dy}, \\ \phi_{g_t}(x) &= 1_{x>0}\phi^{+}_{g_t}(x) + 1_{x<0}\phi^{+}_{g_t}(x), \\ \bar{q}_\nu(\sigma, x, z) &= 1_{\{x \neq 0\}} \frac{1}{x} \int_0^x q_\nu(\sigma, y-z)\, dy + 1_{\{x=0\}} q_\nu(\sigma, z). \end{aligned}$$

Step 1 (*T maps D into itself*): From Lemma 3.3.5, we have $\phi^{+}_{g_t}(x) = \mathbb{E}^{\mathbb{Q}^+}[\Psi^{+}_t(g_t(B_T))]$, where B solves the SDE (3.3.9) and $T \sim \text{InvGamma}(\nu/2, \nu\sigma^2/2)$ is independent of B. It implies $\phi^{+}_{g_t}(x) \in [m, M]$.

Moreover,

$$\left\{ \frac{1}{N_t} q_\nu(\sigma, x-z) + \frac{N_t - 1}{N_t} \bar{q}_\nu(\sigma, x, z) \right\} dz$$

is a probability measure. Thus, Tg_t is continuous, takes values in $[m, M]$, and has derivative bounded by K_0, so $Tg_t \in D_0 \subset D$.

Step 2 (D *is compact*): Let $(g_{t,n}) \subset D$. Then there exists a sequence $(g^0_{t,n}) \subset D_0$ such that μ_0-a.e. we have $g_{t,n} = g^0_{t,n}$.

By Arzelà–Ascoli, a subsequence g^0_{t,n_k} converges uniformly on compacts to some $g^0_t \in D_0$. Then $g_{t,n} \to g_t := g^0_t$ in $L^2(\mu_0)$, and D is compact.

Step 3 (*$T : D_t \to D_t$ is continuous*): Suppose $g_{t,n} \to g_t$ in D_t as $n \to \infty$. By the definition of D_t, we may assume without loss of generality that each $g_{t,n} \in D_{t,0}$ is continuous, since changing on a μ_0-null set does not affect $Tg_{t,n}$. By the Arzelà–Ascoli theorem, the sequence $(g_{t,n}) \subset D_{t,0}$ admits a subsequence converging uniformly on compacts to a continuous function $g_t \in D_{t,0}$, since the functions are equicontinuous (with Lipschitz constant K_0) and uniformly bounded in $[m, M]$. Without loss of generality, assume $g_{t,n} \to g_t$ uniformly on compacts. Because convergence in $L^2(\mu_0)$ implies convergence in measure (under a finite measure), and since continuous functions that agree almost everywhere must agree everywhere, we conclude that $g_{t,n} \to g_t$ pointwise.

We now show pointwise convergence of $Tg_{t,n}(x) \to Tg_t(x)$. Recall:

$$Tg_t(x) := \int_{-\infty}^{\infty} \left\{ \frac{1}{N_t} q_\nu(\sigma, x - z) + \frac{N_t - 1}{N_t} \bar{q}_\nu(\sigma, x, z) \right\} \phi_{\bar{g}_t}(z)\, dz.$$

From Lemma 3.3.5, for $z > 0$, we have:

$$\phi_{g_{t,n}}(z) = \frac{E[\Psi_t^+(g_{t,n}(B_T)) \cdot \Pi_t^+(g_{t,n}(B_T))]}{E[\Pi_t^+(g_{t,n}(B_T))]},$$

where B_T is a Brownian motion stopped at an independent inverse-gamma random time $T \sim \text{Inv-Gamma}(\frac{\nu}{2}, \frac{\nu\sigma^2}{2})$.

Since $g_{t,n} \to g_t$ pointwise and Ψ_t^+, Π_t^+ are continuous (on a set of full probability), and the outputs of $g_{t,n}$ remain uniformly bounded in $[m, M]$, we may apply the dominated convergence theorem:

$$\begin{aligned} E[\Psi_t^+(g_{t,n}(B_T)) \cdot \Pi_t^+(g_{t,n}(B_T))] &\to E[\Psi_t^+(g_t(B_T)) \cdot \Pi_t^+(g_t(B_T))], \\ E[\Pi_t^+(g_{t,n}(B_T))] &\to E[\Pi_t^+(g_t(B_T))], \end{aligned}$$

so that $\phi_{g_{t,n}}(z) \to \phi_{g_t}(z)$ pointwise for each $z \in \mathbb{R}$.

Since the kernel

$$\left\{ \frac{1}{N_t} q_\nu(\sigma, x - z) + \frac{N_t - 1}{N_t} \bar{q}_\nu(\sigma, x, z) \right\}$$

is bounded and integrates to one in z, and $\phi_{\bar{g}_{t,n}}(z) \to \phi_{\bar{g}_t}(z)$ pointwise with uniform boundedness by $[m, M]$, the dominated convergence theorem again yields

$$\lim_{n\to\infty} Tg_{t,n}(x) = Tg_t(x), \quad \text{for all } x \in \mathbb{R}.$$

Hence, T is pointwise continuous on $D_t \subset \chi$, and since the image remains in a convex, compact subset of a Banach space, Schauder's fixed point theorem applies. Therefore, there exists a fixed point $g_t = Tg_t \in D_t$, completing the proof. □

For the results in Section 3.4 and 3.5, we need to prove that $F_t'(x)$ is strictly positive, and it will turn out that Lemma 3.3.5 is a key result. First, we write down $F_t(x)$ in the following way

$$\begin{aligned} F_t(x) &= \frac{1}{N_t} \int_{-\infty}^{+\infty} \mathrm{q}_\nu(\sigma, x - z) \phi_{F_t}(z)\, dz + \frac{N_t - 1}{N_t} \int_{-\infty}^{+\infty} \bar{q}_\nu(\sigma, x, z) \phi_{F_t}(z)\, dz \\ &=: \frac{1}{N_t} G_t(x) + \frac{N_t - 1}{N_t} H_t(x). \end{aligned} \tag{3.3.15}$$

Lemma 3.3.5-(3) implies that $\Delta_{F_t} := \phi^{+}_{F_t}(0) - \phi^{-}_{F_t}(0) > 0$. Stieltjes' derivatives is therefore a positive measure that assigns a strictly positive measure Δ to the point zero. We now compute

$$\begin{aligned} G_t^{'}(x) &= \int_{-\infty}^{+\infty} \partial_x q_v(\sigma, x-z)\phi_{F_t}(z)\, dz = \int_{-\infty}^{+\infty} q_v(\sigma, x-z)\, \phi_{F_t}(dz) \\ &\geq \Delta_{F_t}\, q_v(\sigma, x) > 0, \quad \forall x \in \mathbb{R}. \end{aligned}$$

Regarding $H_t(x)$, for $x \neq 0$ we have

$$H_t(x) = \frac{1}{x}\int_0^x G_t(u)\, du.$$

So

$$H_t'(x) = \frac{1}{x}G_t(x) - \frac{1}{x^2}\int_0^x G_t(u)\, du.$$

From the previous step, $G_t(x)$ is strictly increasing. If $x > 0$, then $\frac{1}{x}\int_0^x G_t(u)\, du$ is strictly less than $G_t(x)$, which implies that $H_t'(x) > 0$. Instead, if $x < 0$, the previous average over $[x, 0]$ is strictly greater than $G_t(x)$, and since $\frac{1}{x} < 0$, we have $H_t'(x) > 0$. Finally, for $x = 0$, we have

$$H_t'(0) = \frac{1}{2}G_t'(0) = \frac{1}{2}\int_{-\infty}^{\infty} q_v(\sigma, -z)\phi_{F_t}(dz) > 0.$$

Putting together the previous observations, we obtain $F_t'(x) > 0$, $\forall x \in \mathbb{R}$.

3.4 Liquidity suppliers' belief posterior consistency

In this section, we establish the posterior consistency of the liquidity suppliers' belief about the fundamental value of the traded asset. Specifically, we show that as the trading volume is observed over time, their posterior distribution concentrates around the true asset value known to the informed traders.

We view the asset distribution V as the parameter governing the law of total trading volume over time, and study posterior concentration with respect to this parameter.

Our consistency result is closely aligned with the general framework of Shalizi (2009), as the data-generating process, namely, the sequence of trading volumes, is time dependent and therefore falls outside the classical i.i.d. setting. However, unlike the misspecified models considered in Shalizi (2009), our framework assumes correct specification: the aggregate trading volume sequence is parameterised by the true fundamental value v_0, which is known to the informed traders but latent to the liquidity suppliers. We show that, starting from a class of suitable priors satisfying mild regularity conditions, the posterior belief of liquidity suppliers concentrates strongly around v_0 as trading volume observations accumulate. This yields a strong posterior consistency result under dependent data.

To establish posterior consistency in our model, we first introduce general preliminaries in Section 3.4.1. The notation and definitions are formulated independently of the trading context and provide the technical foundation for our result. In Section 3.4.2, we specialise this framework to our setting and verify the required conditions for posterior consistency of the liquidity suppliers' beliefs.

3.4.1 Posterior consistency preliminaries

Let Θ denote the parameter space, assumed to be a compact and measurable subset of $\mathbb{R}$. Let $(\mathcal{X}, \mathcal{A})$ be a measurable space representing the observation space at each time point. For each $t \in \mathbb{N}$, define the finite product space $(\mathcal{X}^t, \mathcal{A}^t)$ as the space of length-t sequences $(x_1, \ldots, x_t)$, equipped with the product σ-algebra $\mathcal{A}^t := \bigotimes_{i=1}^t \mathcal{A}$. Define the infinite product space $(\mathcal{X}^\infty, \mathcal{A}^\infty)$ as the space of all infinite sequences $(x_1, x_2, \ldots)$, equipped with the product σ-algebra $\mathcal{A}^\infty := \bigotimes_{i=1}^\infty \mathcal{A}$. For each t, let $\mathcal{F}_t := \sigma(X_1, \ldots, X_t)$ denote the natural filtration generated by the first t observations.

Let $X_1^t := (X_1, X_2, \ldots, X_t)$ denote the finite sequence of observations up to time t, and let $X_1^\infty := (X_1, X_2, \ldots)$ denote the entire infinite sequence. We assume that X_1^∞ is generated sequentially under a fixed but unknown true parameter $\theta_0 \in \Theta$, with joint distribution $P_{\theta_0}^\infty$ defined on $(\mathcal{X}^\infty, \mathcal{A}^\infty)$. Conditional on θ, the joint density admits the following factorisation:

$$f_\theta(X_1^t) = f_\theta(x_1, x_2, \ldots, x_t) := \prod_{i=1}^t f_{i,\theta}(x_i \mid x_1^{i-1}),$$

where $f_{i,\theta}(\cdot \mid x_1^{i-1})$ denotes the conditional density of X_i given the history $x_1^{i-1} := (x_1, \ldots, x_{i-1})$, with the convention that $x_1^0 := \varnothing$. The goal is to perform Bayesian inference on θ based on the observations X_1^t. In a Bayesian framework, a prior distribution Π_0 is posited on θ, and the posterior distribution given $X_1^t = (X_1, \ldots, X_t)$ is denoted by $\Pi(\cdot \mid X_1^t)$. For any measurable subset A of Θ, the posterior distribution $\Pi(A|X_1^t)$ is

$$\Pi(A|X_1^t) = \frac{\int_A f_\theta(X_1^t)\Pi_0(d\theta)}{\int_\Theta f_\theta(X_1^t)\Pi_0(d\theta)} = \frac{\int_A \frac{f_\theta(X_1^t)}{f_{\theta_0}(X_1^t)}\Pi_0(d\theta)}{\int_\Theta \frac{f_\theta(X_1^t)}{f_{\theta_0}(X_1^t)}\Pi_0(d\theta)} = \frac{J_A(X_1^t)}{J(X_1^t)},$$

where

$$J_A(X_1^t) = \int_A \frac{f_\theta(X_1^t)}{f_{\theta_0}(X_1^t)} \Pi_0(d\theta),$$
$$J(X_1^t) = \int_\Theta \frac{f_\theta(X_1^t)}{f_{\theta_0}(X_1^t)} \Pi_0(d\theta).$$

Definition 3.4.1 (KL divergence rate, KL neighbourhood, and KL support)**.** *Let $\theta_0 \in \Theta$ denote the true parameter value. The* Kullback–Leibler (KL) divergence rate *between any $\theta \in \Theta$ and θ_0 is defined as*

$$K(\theta, \theta_0) := \lim_{t\to\infty} \frac{1}{t} \mathbb{E}_{\theta_0} \left[\log \frac{f_{\theta_0}(X_1^t)}{f_\theta(X_1^t)} \right],$$

where $\mathbb{E}_{\theta_0}[\cdot]$ is the expectation under the marginal distribution $P_{\theta_0}^t$ of the sequence X_1^t. We assume the limit exists and is finite for all $\theta \in \Theta$, and that the map $\theta \mapsto K(\theta, \theta_0)$ is Borel measurable.

For any $\epsilon > 0$, the KL neighbourhood *of θ_0 is defined as*

$$K_\epsilon(\theta_0) := \{\theta \in \Theta : K(\theta, \theta_0) < \epsilon\}.$$

The true parameter θ_0 is said to lie in the KL support *of a prior Π_0 if, for all $\epsilon > 0$,*

$$\Pi_0(K_\epsilon(\theta_0)) > 0.$$

Remark 3.4.1 (Uniform separation of KL divergence)**.** *Assuming the map $\theta \mapsto K(\theta, \theta_0)$ is continuous and that $K(\theta, \theta_0) = 0$ if and only if $\theta = \theta_0$, then for any $\eta > 0$, since $\Theta \setminus B_\eta(\theta_0)$ is compact and K is continuous and strictly positive on this set, there exists $\delta > 0$ such that*

$$\inf_{\theta \in \Theta \setminus B_\eta(\theta_0)} K(\theta, \theta_0) \geq \delta > 0.$$

This uniform lower bound ensures that the KL divergence rate is bounded away from zero outside any neighbourhood of the true parameter θ_0.

Assumption 3.4.2 (KL support of the true parameter)**.** *The true parameter $\theta_0 \in \Theta$ lies in the Kullback–Leibler (KL) support of the prior Π_0. That is, for every $\epsilon > 0$,*

$$\Pi_0(K_\epsilon(\theta_0)) > 0.$$

This assumption ensures that the prior does not exclude any KL neighbourhood of the true parameter. It guarantees that the prior gives positive probability to sets where the average log-likelihood under θ is close to that under θ_0, which is essential for the posterior to concentrate around the truth as more data arrive.

Our target is to show the posterior distribution on $A \subset \Theta$ is converging to 0 as $t \to \infty$, where $K(\theta, \theta_0) > 0$ for all $\theta \in A$. Intuitively, it requires the numerator J_A goes to 0 fast, while the denominator does not go to 0 faster than the numerator.

Assumption 3.4.3. *For each $\theta \in \Theta$, the empirical log-likelihood ratio converges to the KL-divergence rate:*

$$\lim_{t\to\infty} \frac{1}{t} \log \frac{f_\theta(X_1^t)}{f_{\theta_0}(X_1^t)} = -K(\theta, \theta_0) \quad P_{\theta_0}^\infty \; a.s. \tag{3.4.1}$$

Assumption 3.4.3 is a mild ergodic-type condition requiring a law of large numbers condition on the log likelihood ratio process. It requires that, under the true parameter θ_0, the time averaged log likelihood ratio converges almost surely to the Kullback–Leibler divergence rate. Such an assumption is central to general posterior consistency results.

Similarly to Shalizi (2009), we assume the existence of a sequence of parameter sets on which the likelihood ratio decays sufficiently fast, and the prior Π_0 concentrates most of its mass on these sets. In contrast, the sets where the likelihood ratio performs poorly are assigned vanishingly small prior mass. This division allows us to partition any measurable set $A \subset \Theta$ into these two regimes, which in turn facilitates the proof of posterior consistency. The sequence of parameter sets is structured as follows.

Assumption 3.4.4. *There exists a sequence of sets $G_t \to \Theta$ such that*

1. $\Pi_0(G_t) \geq 1 - \alpha \exp\{-\beta t\}$, *for some $\alpha, \beta > 0$.*

2. *The convergence in (3.4.1) is uniform in θ over G_t.*

The following lemma controls the denominator of the posterior to not decay too fast.

Lemma 3.4.5. *Under Assumptions 3.4.2 and 3.4.3, for every $\epsilon > 0$,*

$$P_{\theta_0}^\infty \left(\{ J(X_1^t) \leq \exp\{-t\epsilon\} \; i.o. \} \right) = 0.$$

Proof. Define the set of parameters for which the log-likelihood ratio converges to its KL divergence rate:

$$Q := \left\{ (\theta, X_1^\infty) : \lim_{t\to\infty} \frac{1}{t} \log \frac{f_\theta(X_1^t)}{f_{\theta_0}(X_1^t)} = -K(\theta, \theta_0) \right\},$$
$$Q_{X_1^\infty} := \{\theta \in \Theta : (\theta, X_1^\infty) \in Q\}.$$

By Assumption 3.4.3, for $P_{\theta_0}^\infty$-almost every X_1^∞, we have $\Pi_0(Q_{X_1^\infty}) = 1$.

Fix $\epsilon > 0$. Consider the posterior denominator:

$$J(X_1^\infty) = \int_\Theta \frac{f_\theta(X_1^\infty)}{f_{\theta_0}(X_1^\infty)} \Pi_0(d\theta).$$

Then,

$$\begin{aligned} \exp\{t\epsilon\} J(X_1^\infty) &= \int_\Theta \exp\left\{ t\epsilon + \log \frac{f_\theta(X_1^t)}{f_{\theta_0}(X_1^t)} \right\} \Pi_0(d\theta) \\ &\geq \int_{K_{\epsilon/2}(\theta_0) \cap Q_{X_1^\infty}} \exp\left\{ t\left(\epsilon + \frac{1}{t} \log \frac{f_\theta(X_1^t)}{f_{\theta_0}(X_1^t)} \right) \right\} \Pi_0(d\theta). \end{aligned}$$

For $\theta \in Q_{X_1^\infty}$, the integrand satisfies

$$\epsilon + \frac{1}{t} \log \frac{f_\theta(X_1^t)}{f_{\theta_0}(X_1^t)} \longrightarrow \epsilon - K(\theta, \theta_0).$$

For $\theta \in K_{\epsilon/2}(\theta_0)$, we have $K(\theta, \theta_0) < \epsilon/2$, so the limiting value exceeds $\epsilon/2$. Since the integrand is positive and convergence is pointwise over the fixed measurable set $K_{\epsilon/2}(\theta_0) \cap Q_{X_1^\infty}$, we may apply Fatou's Lemma:

$$\begin{aligned} \liminf_{t\to\infty} \exp\{t\epsilon\} J(X_1^\infty) &\geq \int_{K_{\epsilon/2}(\theta_0) \cap Q_{X_1^\infty}} \liminf_{t\to\infty} \exp\left\{ t\left(\epsilon + \frac{1}{t} \log \frac{f_\theta(X_1^t)}{f_{\theta_0}(X_1^t)} \right) \right\} \Pi_0(d\theta) \\ &= \int_{K_{\epsilon/2}(\theta_0) \cap Q_{X_1^\infty}} \exp\left\{ t(\epsilon - K(\theta, \theta_0)) \right\} \Pi_0(d\theta) \\ &\geq \exp\left\{ \tfrac{1}{2}\epsilon t \right\} \Pi_0\left(K_{\epsilon/2}(\theta_0) \cap Q_{X_1^\infty} \right). \end{aligned}$$

By Assumption 3.4.2, $\Pi_0(K_{\epsilon/2}(\theta_0)) > 0$, and since $\Pi_0(Q_{X_1^\infty}) = 1$, their intersection has positive prior mass:

$$\Pi_0\left(K_{\epsilon/2}(\theta_0) \cap Q_{X_1^\infty}\right) > 0 \quad \text{for } P_{\theta_0}^\infty\text{-a.e. } X_1^\infty.$$

Therefore, $\exp\{t\epsilon\}J(X_1^\infty) \to \infty$, i.e.,

$$J(X_1^t) > \exp\{-t\epsilon\} \quad \text{eventually, almost surely.}$$

Equivalently, the event $\{J(X_1^t) \leq \exp\{-t\epsilon\} \text{ i.o.}\}$ has zero probability. □

Now, the following two lemmas facilitate us to show that the numerator of the posterior is decaying quickly enough.

Lemma 3.4.6. *With Assumption 3.4.3 and 3.4.4,*

$$J_{G_t^c}(X_1^t) \leq \alpha \exp\left\{-\frac{t\beta}{2}\right\}, \tag{3.4.2}$$

for all but finitely many t.

Proof. First, note that by Fubini's theorem,

$$\begin{aligned}
E_{\theta_0}[J_{G_t^c}(X_1^t)] &= \int_{\mathcal{X}^t}\int_{G_t^c} \frac{f_\theta(X_1^t)}{f_{\theta_0}(X_1^t)}\Pi_0(d\theta)P_{\theta_0}(dX_1^t) \\
&= \int_{G_t^c}\int_{\mathcal{X}^t} dP_\theta(dX_1^t)\Pi_0(d\theta) \\
&= \Pi_0(G_t^c).
\end{aligned}$$

Therefore, for some $\alpha, \beta > 0$ such that $\Pi_0(G_t) \geq 1 - \alpha\exp\{-\beta t\}$ as in Assumption 3.4.4, and for all t,

$$\begin{aligned}
P_{\theta_0}\left(\left\{X_1^t : J_{G_t^c}(X_1^t) > \exp\left\{-\frac{t\beta}{2}\right\}\right\}\right) &\leq \exp\left\{\frac{t\beta}{2}\right\} E[J_{G_t^c}(X_1^t)] \\
&= \exp\left\{\frac{t\beta}{2}\right\}\Pi_0(G_t^c) \\
&\leq \alpha\exp\left\{-\frac{t\beta}{2}\right\},
\end{aligned}$$

where the last inequality is from the Assumption 3.4.4, $\Pi_0(G_t^c) = 1 - \Pi_0(G_t) \leq \alpha \exp\{-t\beta\}$. By Borel-Cantelli lemma, we have

$$J_{G_t^c}(X_1^t) \leq \alpha \exp\left\{-\frac{t\beta}{2}\right\},$$

for all but finitely many t.

□

Lemma 3.4.7. *Make Assumption 3.4.4. For any measurable set $A \subset \Theta$ with $\Pi_0(A) < \infty$, and any $\epsilon \in (0, \delta/2)$ (where $\delta > 0$ is the uniform lower bound on the KL divergence rate over A, as in Remark 3.4.1), there exists $T > 0$ such that for all $t \geq T$,*

$$\Pi(A \cap G_t \mid X_1^t) \leq \Pi_0(A) \exp\{-t(\delta - 2\epsilon)\}, \quad P_{\theta_0}^{\infty}\text{-}a.s.$$

Proof. As noted in Remark 3.4.1, the uniform lower bound on the KL divergence rate over sets such as $A = \Theta \setminus B_\eta(\theta_0)$ guarantees the existence of $\delta > 0$ such that $\inf_{\theta \in A} K(\theta, \theta_0) > \delta$.

By Assumption 3.4.4, the convergence in (3.4.1) is uniform in θ over G_t. That is, for any $\theta \in G_t$,

$$\lim_{t\to\infty} \frac{1}{t} \log \frac{f_\theta(X_1^t)}{f_{\theta_0}(X_1^t)} = -K(\theta, \theta_0), \quad P_{\theta_0}^{\infty}\text{-a.s.}$$

Hence, for any $\epsilon \in (0, \delta/2)$, there exists $T > 0$ such that for all $t \geq T$ and all $\theta \in G_t$,

$$\left| \frac{1}{t} \log \frac{f_\theta(X_1^t)}{f_{\theta_0}(X_1^t)} + K(\theta, \theta_0) \right| < \epsilon.$$

Since $K(\theta, \theta_0) \geq \delta$ for $\theta \in A$, it follows that

$$\frac{f_\theta(X_1^t)}{f_{\theta_0}(X_1^t)} \leq \exp\{-t(K(\theta, \theta_0) - \epsilon)\} \leq \exp\{-t(\delta - \epsilon)\}.$$

Therefore, for large t, the numerator of the posterior becomes

$$\begin{aligned} J_{A\cap G_t}(X_1^t) &= \int_{A\cap G_t} \frac{f_\theta(X_1^t)}{f_{\theta_0}(X_1^t)} \Pi_0(d\theta) \leq \int_{A\cap G_t} \exp\{-t(\delta - \epsilon)\} \Pi_0(d\theta) \\ &= \Pi_0(A \cap G_t) \exp\{-t(\delta - \epsilon)\} \leq \Pi_0(A) \exp\{-t(\delta - \epsilon)\}. \end{aligned}$$

Meanwhile, by Lemma 3.4.5, for sufficiently large t, the denominator satisfies

$$J(X_1^t) \geq \exp\{-t\epsilon\}, \quad P_{\theta_0}^{\infty}\text{-a.s.}$$

Putting numerator and denominator together, we obtain

$$\Pi(A \cap G_t \mid X_1^t) \leq \frac{\Pi_0(A)\exp\{-t(\delta - \epsilon)\}}{\exp\{-t\epsilon\}} = \Pi_0(A)\exp\left\{-t(\delta - 2\epsilon)\right\},$$

where $\delta - 2\epsilon > 0$. This completes the proof. □

Theorem 3.4.8 (Posterior Consistency). *Let $\Theta \subset \mathbb{R}$ be compact, and suppose Assumptions 3.4.4 and 3.4.3 hold. Then, for any measurable set $A \subset \Theta$ satisfying $\Pi_0(A) > 0$, $\Pi_0(A) < \infty$, and $\inf_{\theta \in A} K(\theta, \theta_0) > \delta$ for some $\delta > 0$, we have*

$$\Pi(A \mid X_1^t) \to 0 \quad \textit{as } t \to \infty, \quad P_{\theta_0}^{\infty}\textit{-almost surely.}$$

Proof. Fix any $\epsilon \in (0, \delta/2)$. Decompose the set A into two disjoint subsets:

$$\Pi(A \mid X_1^t) = \Pi(A \cap G_t \mid X_1^t) + \Pi(A \cap G_t^c \mid X_1^t) \leq \Pi(A \cap G_t \mid X_1^t) + \Pi(G_t^c \mid X_1^t).$$

We bound both terms separately:

1. Upper bound for $\Pi(G_t^c \mid X_1^t)$:

By Lemmas 3.4.5 and 3.4.6, we have

$$\Pi(G_t^c \mid X_1^t) \leq \frac{J_{G_t^c}(X_1^t)}{J(X_1^t)} \leq \frac{\alpha \exp\{-\frac{t}{2}\beta\}}{\exp\{-t\epsilon\}} = \exp\left\{t(\epsilon - \tfrac{1}{2}\beta)\right\}.$$

Since $\epsilon < \frac{1}{2}\beta$, this tends to 0 as $t \to \infty$, $P_{\theta_0}^{\infty}$-almost surely.

2. Upper bound for $\Pi(A \cap G_t \mid X_1^t)$:

By Lemma 3.4.7, and using the uniform separation $\inf_{\theta \in A} K(\theta, \theta_0) > \delta$ (guaranteed by Remark 3.4.1), we have

$$\Pi(A \cap G_t \mid X_1^t) \leq \Pi_0(A)\exp\left\{-t(\delta - 2\epsilon)\right\},$$

which vanishes as $t \to \infty$, $P_{\theta_0}^{\infty}$-a.s.

Combining the two bounds:

$$\Pi(A \mid X_1^t) \leq \Pi_0(A) \exp\{-t(\delta - 2\epsilon)\} + \exp\{t(\epsilon - \tfrac{1}{2}\beta)\} \to 0 \quad \text{a.s.}$$

This completes the proof. □

Example 3.4.1 (Posterior consistency in the Euclidean topology)**.** *Theorem 3.4.8 implies posterior consistency in the Euclidean topology on* $\Theta \subset \mathbb{R}$*. For any open neighborhood* $B_\eta(\theta_0) \subset \Theta$*, the complement* $A := \Theta \setminus B_\eta(\theta_0)$ *satisfies* $K(\theta, \theta_0) \geq \delta > 0$ *uniformly by Remark 3.4.1. Applying the theorem yields*

$$\Pi(\Theta \setminus B_\eta(\theta_0) \mid X_1^t) \to 0 \quad \textit{almost surely.}$$

Hence, the posterior concentrates in every Euclidean neighborhood of θ_0*, establishing consistency in this topology.*

3.4.2 Liquidity suppliers' posterior beliefs

At each trading period t, liquidity suppliers observe the aggregate order flow composed of informed and noise trades, $Y_t = X_t + Z_t$. In equilibrium, informed traders submit orders $X_t^* = F_t^{-1}(v_0)$, where, for any $\epsilon > 0$, $v_0 \in (m + \epsilon, M - \epsilon)$ is the true asset value. Let $Y_0^t := (Y_0, Y_1, \ldots, Y_t)$ denote the history of observed aggregate trading volumes up to time t. We assume Y_0 is fixed deterministically (e.g., $Y_0 = 0$) to initialize the filtration, which is defined as $\mathcal{F}_t := \sigma(Y_0, Y_1, \ldots, Y_t)$.

Conditional on asset value v, the joint distribution of trading volumes Y_0^t is given by

$$f_v(Y_0^t) = \prod_{i=0}^{t} f_{i,v}(Y_i \mid Y_0^{i-1}),$$

where each conditional density $f_{i,v}(y_i \mid y_0^{i-1})$ follows a location-scale t-distribution with ν degrees of freedom:

$$f_{i,v}(y_i \mid y_0^{i-1}) \sim t_\nu \left(y_i \,\middle|\, F_i^{-1}(v), \sigma \right),$$

and depends on the past order flow y_0^{i-1} and the asset value v. Here, F_i^{-1} denotes the inverse of the equilibrium marginal cost function F_i defined in (3.3.8).

Liquidity suppliers begin with a prior distribution Π_0 over V, and update their beliefs via Bayes' rule. After observing the trade flow history Y_0^t, their posterior belief is

$$\Pi(dv \mid Y_0^t) = \frac{f_v(Y_0^t)\,\Pi_0(dv)}{\int_m^M f_v(Y_0^t)\,\Pi_0(dv)}. \tag{3.4.3}$$

We will show in Theorem 3.4.11 that the posterior distribution $\Pi(\cdot|Y_0^t)$ eventually concentrates around the true parameter v_0, i.e., that posterior consistency holds.

To ease the discussion, we assume $N_t = N$ for all $t \in \mathbb{N}$. Since N_t is deterministic and plays no role across different trading periods, this simplification does not affect the validity of our result.

Firstly, we will show that Assumption 3.4.3 is satisfied. To see this, we require that for each $v \in (m+\epsilon, M-\epsilon)$, the log-likelihood ratio is integrable and has finite second moment.

1. Lipschitz continuity of F_t^{-1} in $(m+\epsilon, M-\epsilon)$, for any $\epsilon > 0$.

Since F_t is strictly increasing and continuously differentiable, the inverse function F_t^{-1} is differentiable with

$$\left|\frac{d}{dx}F_t^{-1}(x)\right| = \frac{1}{|F_t'\left(F_t^{-1}(x)\right)|}.$$

Since $F_t'(x) > 0$ and $|F_t'(x)| \le K_0$ shown in Theorem 3.3.7, F_t^{-1} is Lipschitz on every $(m+\epsilon, M-\epsilon)$. We denote the Lipschitz constant of F_t^{-1} as L_ϵ.

2. Finite second moment of the log-likelihood ratio when $\nu > 2$.

We verify that

$$\sup_{i,v} E_{v_0}\left[\left(\log\frac{f_{i,v}(Y_i \mid Y_0^{i-1})}{f_{i,v_0}(Y_i \mid Y_0^{i-1})}\right)^2 \middle| \mathcal{F}_{i-1}\right] < \infty. \tag{3.4.4}$$

To see this, recall that

$$f_{i,v}(y_i \mid y_0^{i-1}) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma}\left(1+\frac{1}{\nu}\left(\frac{y_i - F_i^{-1}(v)}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}}.$$

Hence, by an application of the mean value theorem for some $\xi_i \in \left(1+\frac{(Y_i-F_i^{-1}(v))^2}{\nu\sigma^2}, 1+\frac{(Y_i-F_i^{-1}(v_0))^2}{\nu\sigma^2}\right)$ $(0 < \frac{1}{\xi_i^2} \le 1)$, the Lipschitz continuity of F_i^{-1} in any $(m+\epsilon, M-\epsilon)$, the fact that in equilibrium $Y_i = Z_i + F_i^{-1}(v_0)$, and $Z_i \sim t_\nu(0,\sigma)$ with

Z_i independent of the fundamental value, we obtain

$$
\begin{aligned}
&E_{v_0}\left[\left(\log\frac{f_{i,v}(Y_i \mid Y_1^{i-1})}{f_{i,v_0}(Y_i \mid Y_1^{i-1})}\right)^2 \middle| \mathcal{F}_{i-1}\right] \\
&= \frac{(\nu+1)^2}{4} E_{v_0}\left[\left\{\log\left(1+\frac{(Y_i - F_i^{-1}(v))^2}{\nu\sigma^2}\right) - \log\left(1+\frac{(Y_i - F_i^{-1}(v_0))^2}{\nu\sigma^2}\right)\right\}^2 \middle| \mathcal{F}_{i-1}\right] \\
&= \frac{(\nu+1)^2}{4\nu^2\sigma^4} E_{v_0}\left[\frac{1}{\xi_i^2}\left(4Z_i^2(F_i^{-1}(v) - F_i^{-1}(v_0))^2 + (F_i^{-1}(v) - F_i^{-1}(v_0))^4\right) \middle| \mathcal{F}_{i-1}\right] \\
&\leq C_1 E[Z_i^2] + C_2 < \infty,
\end{aligned}
$$

where $C_1 = \frac{\nu+1}{2\nu\sigma^2}2L_\epsilon$, $C_2 = \frac{\nu+1}{2\nu\sigma^2}2L_\epsilon^2$, and $E[Z_i^2]$ is finite and independent of i. This implies Equation 3.4.4, as well as an analogous bound for the corresponding first absolute moment.

Lemma 3.4.9. *Suppose $\nu > 2$, let $v \in (m+\epsilon, M-\epsilon)$ for any $\epsilon > 0$ such that $K(v, v_0) > 0$, where*

$$K(v, v_0) := \lim_{t\to\infty} \frac{1}{t} E_{v_0}\left[\log\frac{f_{v_0}(Y_0^t)}{f_v(Y_0^t)}\right]$$

exists and is finite. Then,

$$\lim_{t\to\infty} \frac{1}{t} \log\frac{f_v(Y_0^t)}{f_{v_0}(Y_0^t)} = -K(v, v_0), \quad \mathbb{P}_{v_0}\text{-a.s.}$$

Moreover, the convergence is uniform on $(m+\epsilon, M-\epsilon)$

$$\lim_{t\to\infty} \sup_{v\in(m+\epsilon, M-\epsilon)} \left|\frac{1}{t}\log\frac{f_v(Y_0^t)}{f_{v_0}(Y_0^t)} + K(v, v_0)\right| = 0, \quad \mathbb{P}_{v_0}\text{-a.s.}$$

Proof. Let $\ell_i(v) := \log\frac{f_{i,v}(Y_i|Y_0^{i-1})}{f_{i,v_0}(Y_i|Y_0^{i-1})}$ and $\mathcal{F}_i := \sigma(Y_0, Y_1, \ldots, Y_i)$. Define the martingale difference sequence

$$D_i(v) := \ell_i(v) - E_{v_0}[\ell_i(v) \mid \mathcal{F}_{i-1}].$$

Under the assumption $\nu > 2$, the log-likelihood ratio has uniformly bounded conditional second moments (see (3.4.4)), so by the martingale strong law of large numbers (see, e.g., Williams, 1991),

$$\frac{1}{t}\sum_{i=1}^{t} D_i(v) \to 0 \quad \mathbb{P}_{v_0}\text{-a.s.}$$

Write

$$\frac{1}{t}\sum_{i=1}^{t} E_{v_0}[\ell_i(v) \mid \mathcal{F}_{i-1}] = \frac{1}{t}\sum_{i=1}^{t} E_{v_0}[\ell_i(v)] + \frac{1}{t}\sum_{i=1}^{t} W_i,$$

where $W_i := E_{v_0}[\ell_i(v) \mid \mathcal{F}_{i-1}] - E_{v_0}[\ell_i(v)]$. The sequence $\{W_i\}$ is again a martingale difference with bounded second moments, so

$$\frac{1}{t}\sum_{i=1}^{t} W_i \to 0 \quad P_{v_0}\text{-a.s.}$$

By assumption, $\frac{1}{t}\sum_{i=1}^{t} E_{v_0}[\ell_i(v)] \to -K(v, v_0)$. Combining everything,

$$\frac{1}{t}\log\frac{f_v(Y_0^t)}{f_{v_0}(Y_0^t)} = \frac{1}{t}\sum_{i=1}^{t}\ell_i(v) = \underbrace{\frac{1}{t}\sum_{i=1}^{t} D_i(v)}_{\to 0} + \underbrace{\frac{1}{t}\sum_{i=1}^{t} E_{v_0}[\ell_i(v) \mid \mathcal{F}_{i-1}]}_{\to -K(v,v_0)} \to -K(v, v_0) \quad \mathbb{P}_{v_0}\text{-a.s.}$$

It remains to establish uniform convergence on $(m+\epsilon, M-\epsilon)$. Define

$$g_t(v) := \frac{1}{t}\log\frac{f_v(Y_0^t)}{f_{v_0}(Y_0^t)} = \frac{1}{t}\sum_{i=1}^{t}\ell_i(v).$$

We show that the family $\{g_t(\cdot)\}_{t\geq 1}$ is uniformly Lipschitz on $(m+\epsilon, M-\epsilon)$. To simplify notation, set

$$m_i(v) := F_i^{-1}(v),$$

so that $f_{i,v}(Y_i \mid Y_0^{i-1})$ is a t_ν density with location $m_i(v)$ and scale σ. Then

$$\partial_v \log f_{i,v}(Y_i \mid Y_0^{i-1}) = \partial_m \log f_{i,v}(Y_i \mid Y_0^{i-1}) \cdot \frac{d}{dv} m_i(v).$$

Moreover,

$$\partial_m \log f_{i,v}(Y_i \mid Y_0^{i-1}) = \frac{(\nu+1)(Y_i - m_i(v))}{\nu\sigma^2 + (Y_i - m_i(v))^2},$$

and hence

$$\sup_{y,m}\left|\partial_m \log f_{i,v}(y \mid Y_0^{i-1})\right| = \frac{\nu+1}{2\sigma\sqrt{\nu}} =: C_0 < \infty.$$

Since $F_i^{-1}(\cdot)$ is Lipschitz on $(m+\epsilon, M-\epsilon)$ with constant L_ϵ, we obtain

$$\sup_{i\geq 1}\ \sup_{v\in(m+\epsilon,M-\epsilon)}\left|\partial_v \log f_{i,v}(Y_i \mid Y_0^{i-1})\right| \leq C_0 L_\epsilon.$$

Therefore, for all $v_1, v_2 \in (m+\epsilon, M-\epsilon)$,

$$|g_t(v_1) - g_t(v_2)| \le \frac{1}{t}\sum_{i=1}^{t} |\ell_i(v_1) - \ell_i(v_2)| \le C_0 L_\epsilon |v_1 - v_2|,$$

so $\{g_t\}$ is uniformly Lipschitz on $(m+\epsilon, M-\epsilon)$.

Finally, since $g_t(v) \to -K(v, v_0)$ $\mathbb{P}_{v_0}$-a.s. for each fixed $v \in (m+\epsilon, M-\epsilon)$, and $\{g_t\}$ is uniformly Lipschitz on $(m+\epsilon, M-\epsilon)$, the convergence is uniform on $(m+\epsilon, M-\epsilon)$, which yields

$$\lim_{t\to\infty} \sup_{v\in(m+\epsilon, M-\epsilon)} |g_t(v) + K(v, v_0)| = 0, \quad \mathbb{P}_{v_0}\text{-a.s.}$$

This concludes the proof. □

In the following lemma, we will show that under some mild conditions, we can construct a sequence of set G_t such that Assumption 3.4.4 is satisfied.

Lemma 3.4.10. *Define the sequence of sets*

$$G_t^\epsilon := \left\{ v \in (m+\epsilon, M-\epsilon) : K(v, v_0) \ge (M-m-2\epsilon)^2 \frac{\alpha^2}{4} \exp\{-2\beta t\},\ \alpha, \beta > 0 \right\}. \tag{3.4.5}$$

If the prior Π_0 satisfies the following conditions:

1. *The prior assigns positive mass to every KL neighbourhood of v_0 inside $(m+\epsilon, M-\epsilon)$, that is, $\Pi_0(K_\delta(v_0) \cap (m+\epsilon, M-\epsilon)) > 0$ for any $\delta > 0$, where $K_\delta(v_0)$ is a δ-KL neighbourhood.*

2. $\Pi_0((G_t^\epsilon)^c) \le \alpha \exp\{-\beta t\}$,

then Assumption 3.4.4 is satisfied on $(m+\epsilon, M-\epsilon)$.

Proof. To check that the defined G_t^ϵ satisfies Assumption 3.4.4:
Firstly, as $t \to \infty$, $\alpha \exp\{-\beta t\} \to 0$, and hence G_t^ϵ expands and eventually covers the full parameter space $(m+\epsilon, M-\epsilon)$, that is, $G_t^\epsilon \to (m+\epsilon, M-\epsilon)$.

Secondly, by the condition on the prior,

$$\Pi_0(G_t^\epsilon) \ge 1 - \alpha \exp\{-\beta t\}.$$

Finally, the uniform convergence (3.4.1) on G_t^ϵ follows directly from Lemma 3.4.9, since the convergence is uniform for all $v \in (m+\epsilon, M-\epsilon)$. □

Example 3.4.2. *If the prior Π_0 has a continuous density π_0 supported on the compact interval $(m+\epsilon, M-\epsilon)$, then the conditions in Lemma 3.4.10 are satisfied.*

Proof. Since π_0 is continuous on the compact interval $[m+\epsilon, M-\epsilon]$, the Extreme Value Theorem implies that π_0 is bounded. That is, there exists a constant $0 < C < \infty$ such that

$$\sup_{v\in[m+\epsilon, M-\epsilon]} \pi_0(v) = C(\epsilon)'.$$

Then, for the sequence of sets $(G_t^\epsilon)^c \subseteq (m+\epsilon, M-\epsilon)$, we have

$$\begin{aligned}\Pi_0\big((G_t^\epsilon)^c\big) = \int_{(G_t^\epsilon)^c} \pi_0(v)\,\mu(dv) &\leq \sup_{v\in[m+\epsilon, M-\epsilon]} \pi_0(v)\cdot\mu\big((G_t^\epsilon)^c\big)\\ &\leq C(\epsilon)'\cdot\mu\big((G_t^\epsilon)^c\big).\end{aligned}$$

By the definition of G_t^ϵ, one can show that there exists a constant $C(\epsilon)$ such that

$$(G_t^\epsilon)^c \subseteq \Big\{v \in (m+\epsilon, M-\epsilon) : |v-v_0| \leq C(\epsilon)^{-1/2}(M-m-2\epsilon)\tfrac{\alpha}{2}e^{-\beta t}\Big\},$$

and hence

$$\mu\big((G_t^\epsilon)^c\big) \leq C(\epsilon)^{-1/2}(M-m-2\epsilon)\,\alpha e^{-\beta t}.$$

Therefore,

$$\Pi_0\big((G_t^\epsilon)^c\big) \leq C(\epsilon)'\cdot C(\epsilon)^{-1/2}(M-m-2\epsilon)\,\alpha e^{-\beta t}.$$

Letting $\alpha' := C(\epsilon)'C(\epsilon)^{-1/2}(M-m-2\epsilon)\alpha$, we conclude that

$$\Pi_0\big((G_t^\epsilon)^c\big) \leq \alpha' e^{-\beta t},$$

as required. □

Now, we are ready to state the posterior consistency result for the liquidity suppliers' beliefs. Fix any $\epsilon > 0$. For any set $A \subseteq (m+\epsilon, M-\epsilon)$ that is separated from the true fundamental value v_0 in the sense of KL divergence, the posterior mass on A vanishes almost surely.

Theorem 3.4.11 (Liquidity suppliers' posterior consistency)**.** *Suppose the fundamental value of the trading asset is $V \in [m, M]$, where $-\infty < m < M < \infty$, and the true value is $v_0 \in (m, M)$. For any $\varepsilon > 0$, define the truncated parameter space*

$$\Theta_\varepsilon := (m+\varepsilon,\, M-\varepsilon),$$

and suppose that $v_0 \in \Theta_\varepsilon$.

Liquidity suppliers observe the sequence of aggregate order flows from informed and noise traders, denoted by $Y_0^t = (Y_0, Y_1, \ldots, Y_t)$*, which is generated sequentially under* P_{v_0} *with density*

$$f_{v_0}(Y_1^t) = \prod_{i=1}^{t} f_{i,v_0}(Y_i \mid Y_0^{i-1}),$$

where each conditional density satisfies

$$f_{i,v_0}(y_i \mid y_0^{i-1}) \sim t_\nu\big(y_i \mid F_i^{-1}(v_0), \sigma\big),$$

and F_i^{-1} *is defined as the solution to the fixed point equation* (3.3.8) *based on the past trading data* y_0^{i-1}.

Assume the prior Π_0 *is supported on* Θ_ε *and satisfies the conditions of Lemma 3.4.10 on* Θ_ε*, and that liquidity suppliers update their beliefs according to the Bayesian recursion* (3.4.3)*. Then for any measurable set* $A \subseteq \Theta_\varepsilon$ *such that:*

- $\Pi_0(A) > 0$*, and*
- $K(v, v_0) > 0$ *for all* $v \in A$,

the posterior contracts onto v_0*; that is,*

$$\Pi(A \mid Y_0^t) \to 0, \quad \text{as } t \to \infty, \quad \text{almost surely under } P_{v_0}^\infty.$$

In other words, for every $\varepsilon > 0$*, the posterior becomes consistent at the true fundamental value* v_0 *on the compact subset* Θ_ε*, and asymptotically places no mass on any region in* Θ_ε *that is KL-separated from it.*

Proof. In view of Lemma 3.4.9 and Lemma 3.4.10, the Assumption 3.4.3 and Assumption 3.4.4 are both satisfied. Then, by an application of Theorem 3.4.8, the liquidity suppliers' posterior consistency follows. □

3.5 Asymptotic price impact

In markets with asymmetric information, informed traders who possess private knowledge of the asset's true value tend to submit large orders to exploit their informational advantage.

In our setting, where informed traders are myopic and act only once, this effect is even more pronounced, as they trade aggressively to maximise their profits in each period.

This motivates the study of asymptotic market impact, which characterises how the cost of trading evolves as the order size becomes large. In particular, we analyse the tail behaviour of the marginal cost function $F_t(x)$ at each trading date t, using tools from the theory of regular variation.

Definition 3.5.1. *A function $g : (0, \infty) \mapsto (0, \infty)$ is said to be regularly varying of index ρ at ∞ if*

$$\lim_{\lambda \to \infty} \frac{g(\lambda x)}{g(\lambda)} = x^{\rho}, \quad \forall x > 0.$$

Similarly, a function $g : (-\infty, 0) \mapsto (0, \infty)$ is said to be regularly varying of index ρ at $-\infty$ if $g(-x)$ is regularly varying of index ρ at ∞.
For $\rho = 0$, that is, if

$$\lim_{\lambda \to \infty} \frac{g(\lambda x)}{g(\lambda)} = 1, \quad \forall x > 0,$$

g is said to be slowly varying.

Remark 3.5.1. *Roughly speaking, a function of regular variation behaves like a power function asymptotically. On the other hand, for slowly varying functions, interesting examples include: $\log x$, $(\log x)^{\alpha}$, for $\alpha \in \mathbb{R}$, $\log \log x$, and $\exp\{(\log x)^{\alpha}\}$, for $\alpha \in (0, 1)$.*

Denote $\partial_x \Psi_t^+(M) = \lim_{x \to M} \frac{M - \Psi_t^+(x)}{M - x}$, $\partial_x \Psi_t^-(m) = \lim_{x \to m} \frac{\Psi_t^-(x) - m}{x - m}$. The following lemma is important to show that $\partial_x \Psi_t^{\pm}$ is an indication of fat tails.

Lemma 3.5.2. *When the support of V is supp$(V) = [m, M]$,we have*

$$\partial_x \Psi_t^+(M) = \lim_{x \to M} \frac{1}{-1 + (M - x)\frac{\partial_x \Pi_t^+(x)}{\Pi_t^+(x)}} + 1, \tag{3.5.1}$$

$$\partial_x \Psi_t^-(m) = \lim_{x \to m} -\frac{1}{1 + (x - m)\frac{\partial_x \Pi_t^-(x)}{\Pi_t^-(x)}} + 1. \tag{3.5.2}$$

Proof. First, note that $\Pi_t^+(x) = \frac{\int_x^M \Pi_t^+(y)\, dy}{\Psi_t^+(x) - x}$. To show this equality, it is equivalent to show $\Phi_t^+(x) - x\Pi_t^+(x) = \int_x^M \Pi_t^+(y)\, dy$, which is the result of an application of integration by parts. Then, by direct manipulation, we can obtain

$$\frac{M - \Psi_t^+(x)}{M - x} = -\frac{\int_x^M \Pi_t^+(y)\, dy}{(M-x)\Pi_t^+(x)} + 1.$$

Sending x to M and applying L'Hôpital's rule, we get the desired result. Similarly for $\partial_x \Psi_t^-(m)$. □

In the next result, we shall show that, under mild condition on trading asset distribution, the price impact obeys a power law.

Theorem 3.5.3. *For all trading periods $t \in \mathbb{N}$, suppose the derivatives $\partial_x \Psi_t^+(M)$ and $\partial_x \Psi_t^-(m)$ exist, and that $N_t > 1$, with the asset distribution supported on $\operatorname{supp}(V) = [m, M] \subset \mathbb{R}$. For each trading period t, the following statements hold:*

1. *Assume there exists a constant $L \in \mathbb{R}$ such that*

$$\lim_{x \to M^-} (M - x) \frac{P'(V \in dx)}{P(V \in dx)} = L.$$

Then, $M - F_t(x)$ is regularly varying at $+\infty$ with index

$$\rho_t^+ = \frac{\partial_x \Psi_t^+(M) - 1}{1 - \frac{\partial_x \Psi_t^+(M)}{N_t}},$$

where

$$\partial_x \Psi_t^+(M) = \frac{1 - L - (\nu + 1)\sum_{i=0}^{t-1} \frac{1}{\rho_i^+}}{2 - L - (\nu + 1)\sum_{i=0}^{t-1} \frac{1}{\rho_i^+}} \in (0, 1).$$

2. *Assume there exists a constant $L' \in \mathbb{R}$ such that*

$$\lim_{x \to m^+} (x - m) \frac{P'(V \in dx)}{P(V \in dx)} = L'.$$

Then, $F_t(x) - m$ is regularly varying at $-\infty$ with index

$$\rho_t^- = \frac{\partial_x \Psi_t^-(m) - 1}{1 - \frac{\partial_x \Psi_t^-(m)}{N_t}},$$

where

$$\partial_x \Psi_t^-(m) = \frac{1 + L' - (\nu+1)\sum_{i=0}^{t-1} \frac{1}{\rho_i^-}}{2 + L' - (\nu+1)\sum_{i=0}^{t-1} \frac{1}{\rho_i^-}} \in (0,1).$$

Furthermore, as $t \to \infty$, the marginal response flattens:

$$\lim_{t\to\infty} \partial_x \Psi_t^+(M) = 1, \quad \lim_{t\to\infty} \partial_x \Psi_t^-(m) = 1, \quad \lim_{t\to\infty} |\rho_t^+| = 0, \quad \lim_{t\to\infty} |\rho_t^-| = 0.$$

Proof. Recall that $\Pi_t^+(x) = P(V \geq x \mid Y_0^{t-1} = y_0^{t-1}) = \int_x^M P(V \in dv \mid Y_0^{t-1} = y_0^{t-1})\, dv$, where $P(V \in dv \mid Y_0^{t-1} = y_0^{t-1})$ is as in (3.3.5). Denote $p\left(y \mid y_0^{t-1}\right) := P(V \in dy \mid Y_0^{t-1} = y_0^{t-1})$, by L'Hôpital's rule, we can write

$$\begin{aligned}
\lim_{x\to M} (M-x)\frac{\partial_x \Pi_t^+(x)}{\Pi_t^+(x)} &= \lim_{x\to M} -\frac{(M-x)p\left(x \mid y_0^{t-1}\right)}{\int_x^M p\left(v \mid y_0^{t-1}\right)\, dv} \\
&= -1 + \lim_{x\to M} (M-x)\frac{p'\left(x \mid y_0^{t-1}\right)}{p\left(x \mid y_0^{t-1}\right)} \\
&= -1 + \lim_{x\to M} (M-x)\left(\log p\left(x \mid y_0^{t-1}\right)\right)'.
\end{aligned}$$

From liquidity suppliers' Bayesian updating formula in (3.3.5), we have

$$\log p\left(x \mid y_0^{t-1}\right) = \text{const} + \log p(V \in dx) - \frac{\nu+1}{2}\sum_{i=0}^{t-1} \log\left(1 + \frac{(F_i^{-1}(x) - y_i)^2}{\nu\sigma^2}\right),$$

$$\begin{aligned}
(M-x)\left(\log p\left(x \mid y_0^{t-1}\right)\right)' &= (M-x)\frac{p'(V \in dx)}{p(V \in dx)} \\
&\quad - (\nu+1)\sum_{i=0}^{t-1} \frac{M-x}{F_i'(F_i^{-1}(x))} \cdot \frac{F_i^{-1}(x) - y_i}{\nu\sigma^2 + (F_i^{-1}(x) - y_i)^2}.
\end{aligned}$$

Since each $F_i(x)$ is monotonically increasing in x, and $F_i(x) \to M$ as $x \to \infty$, sending x to M, we have

$$\begin{aligned}
&\lim_{x\to M}(M-x)\frac{\partial_x \Pi_t^+(x)}{\Pi_t^+(x)} \\
&= -1 + \lim_{x\to M}(M-x)\frac{p'(V\in dx)}{p(V\in dx)} - (\nu+1)\sum_{i=0}^{t-1}\lim_{x\to\infty}\frac{1}{(x-y_i)+\frac{\nu\sigma^2}{x-y_i}}\frac{M-F_i(x)}{F_i'(x)} \\
&= -1 + \lim_{x\to M}(M-x)\frac{p'(V\in dx)}{p(V\in dx)} + (\nu+1)\sum_{i=0}^{t-1}\lim_{x\to\infty}\frac{1}{(x-y_i)+\frac{\nu\sigma^2}{x-y_i}}\frac{M-F_i(x)}{(M-F_i(x))'} \\
&= -1 + L + (\nu+1)\sum_{i=0}^{t-1}\lim_{x\to\infty}\frac{1}{\rho_i^+}\frac{1}{1-\frac{y_i}{x}+\frac{\nu\sigma^2}{x(x-y_i)}} \\
&= -1 + L + (\nu+1)\sum_{i=0}^{t-1}\frac{1}{\rho_i^+}.
\end{aligned}$$

We shall see later that, at all trading dates $i = 0, \ldots, t-1$ before t, $M - F_i(x)$ is regularly varying at ∞ with the corresponding index ρ_i^+. Therefore, $\frac{M-F_i(x)}{(M-F_i(x))'} \sim \frac{x}{\rho_i^+}$ as $x \to \infty$, where $\rho_i^+ \in (-1, 0)$, which gives us the desired next-to-last equality.

Now, according to (3.5.1),

$$\partial_x \Psi_t^+(M) = \frac{1 - L - (\nu+1)\sum_{i=0}^{t-1}\frac{1}{\rho_i^+}}{2 - L - (\nu+1)\sum_{i=0}^{t-1}\frac{1}{\rho_i^+}} \in (0,1). \tag{3.5.3}$$

It then follows that $\lim_{t\to\infty}\partial_x \Psi_t^+(M) = 1$.

To see the asymptotic price impact at each trading date t,

$$\begin{aligned}
\lim_{\alpha\to\infty}\frac{M-F_t(\alpha x)}{M-F_t(\alpha)} &= \lim_{\alpha\to\infty}\int_{-\infty}^{+\infty}\left\{\frac{1}{N_t}q_\nu(\sigma,\alpha x - z) + \frac{N_t-1}{N_t}\bar{q}_\nu(\sigma,\alpha x, z)\right\}\frac{M-\phi_{F_t}(z)}{M-F_t(\alpha)}dz \\
&= \lim_{\alpha\to\infty}\int_{-\infty}^{+\infty}\left\{\frac{1}{N_t}q_\nu\left(\frac{\sigma}{\alpha}, x-z\right) + \frac{N_t-1}{N_t}\bar{q}_\nu\left(\frac{\sigma}{\alpha}, x, z\right)\right\}\frac{M-\phi_{F_t}(\alpha z)}{M-F_t(\alpha)}dz
\end{aligned}$$

Note that when $z > 0$,

$$\begin{aligned}
\phi_{F_t}(\alpha z) &= \frac{\int_{-\infty}^{\infty} dy\, \Phi_t^+ (F_t(y))\, q_v(\sigma, \alpha z - y)}{\int_{-\infty}^{\infty} dy\, \Pi_t^+ (F_t(y))\, q_v(\sigma, \alpha z - y)} \\
&= \frac{\int_{-\infty}^{\infty} dy\, \Pi_t^+ (F_t(\alpha y))\, q_v(\frac{\sigma}{\alpha}, z - y)}{\int_{-\infty}^{\infty} dy\, \Pi_t^+ (F_t(\alpha y))\, q_v(\frac{\sigma}{\alpha}, z - y)} \Psi_t^+ (F_t(\alpha y)) \\
&=: \int_{-\infty}^{\infty} \Lambda_t(\alpha, z, dy) \Psi_t^+ (F_t(\alpha y)) \\
&\longrightarrow \Psi_t^+ (F_t(\alpha z))
\end{aligned}$$

as the probability measure

$$\Lambda_t(\alpha, z, dy) := \frac{\Pi_t^+ (F_t(\alpha y))\, q_v(\frac{\sigma}{\alpha}, z - y)}{\int_{-\infty}^{\infty} \Pi_t^+ (F_t(\alpha u))\, q_v(\frac{\sigma}{\alpha}, z - u)\, du} dy \tag{3.5.4}$$

converges to the point mass at z as $\alpha \to \infty$. By the Mean Value Theorem, for each $\alpha > 0$, there exists $z^* \in [F_t(\alpha z), M]$ such that

$$\frac{M - \Psi_t^+(F_t(\alpha z))}{M - F_t(\alpha)} = \partial_x \Psi_t^+(z^*) \cdot \frac{M - F_t(\alpha z)}{M - F_t(\alpha)}.$$

Assume the limit

$$\lim_{\alpha \to \infty} \frac{M - F_t(\alpha z)}{M - F_t(\alpha)} =: \gamma(t, z)$$

exists. Since $F_t(\alpha z) \to M$ as $\alpha \to \infty$, we have $z^* \to M$, and since $\partial_x \Psi_t^+$ is continuous near M by assumption, we conclude:

$$\lim_{\alpha \to \infty} \frac{M - \Psi_t^+(F_t(\alpha z))}{M - F_t(\alpha)} = \partial_x \Psi_t^+(M) \cdot \gamma(t, z).$$

Then, we need to solve

$$\gamma(t, x) = \frac{\partial_x \Psi_t^+ (M)}{N_t} \gamma(t, x) + \frac{N_t - 1}{N_t x} \partial_x \Psi_t^+ (M) \int_0^x \gamma(t, y)\, dy, \tag{3.5.5}$$

with initial condition $\gamma(t, 1) = 1$. Assume the form $\gamma(t, x) = x^{\rho_t^+}$, then

$$\rho_t^+ = \frac{\partial_x \Psi_t^+(M) - 1}{1 - \frac{\partial_x \Psi_t^+(M)}{N_t}}.$$

Similar arguments for $\partial_x \Psi_t^- (m)$ and $F_t(x) - m$. □

Remark 3.5.2. *When the asset distribution follows a power law near the upper bound, i.e., $P(V > x) \propto (M - x)^\alpha$ for some $\alpha > 0$, the condition*

$$\lim_{x \to M^-} (M - x) \left(\log p(V \in dx) \right)' = L$$

is satisfied with $L = -\alpha$.

Similarly, for lighter-tailed distributions such as

$$p(V \in dx) \sim \exp\{-(M - x)^\beta\}, \quad \textit{for } \beta \geq 0,$$

the same condition holds with $L = 0$. For instance, a truncated Gaussian distribution on $[m, M]$ falls under this category.

According to equation (3.5.3), the exponent ρ_t^+ can be written as

$$\rho_t^+ = -\frac{1}{1 + \frac{N_t - 1}{N_t} \alpha_t},$$

where

$$\alpha_t := 1 - L - (\nu + 1) \sum_{i=0}^{t-1} \frac{1}{\rho_i^+}, \quad \text{with } \alpha_0 = 1 - L.$$

This direct form highlights the dependence of the market impact exponent ρ_t^+ on the tail decay of the prior (via L), the noise distribution (via ν), and the entire trading history up to period t.

Moreover, note that α_t admits the recursive representation

$$\alpha_t = \alpha_{t-1} - \frac{\nu + 1}{\rho_{t-1}^+},$$

from which it is clear that $\alpha_t > \alpha_{t-1}$ since $\rho_{t-1}^+ \in (-1, 0)$. Hence, α_t increases strictly with t, and the absolute value of the impact exponent decreases:

$$\lim_{t \to \infty} |\rho_t^+| = \frac{1}{1 + \frac{N_t - 1}{N_t} \alpha_t} = 0 \quad .$$

This reflects the economic intuition that the marginal informativeness of each trade declines as more trades are observed, leading to a progressively flatter LOB over time. Furthermore, for fixed t, a larger number of insiders N_t reduces the price impact due to increased competition among them. Conversely, a smaller value of ν, corresponding to heavier tails in the noise distribution, amplifies the price impact by making it harder for liquidity suppliers to separate informed trades from noise.

Compared with the Gaussian noise assumption in Çetin and Waelbroeck (2024), which corresponds to the limiting case $\nu \to \infty$, our model with fat-tailed location-scale t-distributed noise captures a steeper LOB. In the Gaussian case, noise traders do not submit large trades, making it easier for liquidity suppliers to infer the asset value from observed prices. By contrast, under heavy-tailed noise, large trades by noise traders obscure the signal from insiders, allowing insiders to strategically hide their trades and profit more from the trading.

Furthermore, in equilibrium, h_t and F_t behave similarly for large values. In other words, $M - h_t$ and $M - F_t$ are regularly varying with the same index. To see this, note that

$$\begin{aligned}\phi^+_{F_t}(x) &= \frac{\int_{-\infty}^{\infty} \Phi^+_t(F_t(y))q_\nu(\sigma, x-y)\,dy}{\int_{-\infty}^{\infty} \Pi^+_t(F_t(y))q_\nu(\sigma, x-y)\,dy}\\ &= \lim_{x\to\infty} \frac{\int_{-\infty}^{\infty} \Pi^+_t(F_t(xy))\Psi^+_t(F_t(xy))q_\nu(\frac{\sigma}{x}, 1-y)\,dy}{\int_{-\infty}^{\infty} \Pi^+_t(F_t(xy))q_\nu(\frac{\sigma}{x}, 1-y)\,dy}\\ &= \int_{-\infty}^{\infty} \Lambda_t(x, 1, dy)\Psi^+_t(F_t(xy)),\end{aligned}$$

where Λ_t is the probability measure defined in (3.5.4). And it converges to a Dirac measure at 1 as $x \to \infty$.

$$\begin{aligned}&\lim_{x\to\infty} \frac{M - h_t(x)}{M - F_t(x)} = \lim_{x\to\infty} \frac{M - \phi^+_{F_t}(x)}{M - F_t(x)}\\ =& \lim_{x\to\infty} \int_{-\infty}^{\infty} \Lambda_t(x, 1, dy)\frac{M - \Psi^+_t(F_t(xy))}{M - F_t(x)}\\ =& \lim_{x\to\infty} \int_{-\infty}^{\infty} \Lambda_t(x, 1, dy)\partial_x \Psi^+_t(y^*)\frac{M - F_t(xy)}{M - F_t(x)}\\ =&\partial_x \Psi^+_t(M),\end{aligned}$$

where the second-to-last equality is from Mean Value Theorem for some $y^* \in [F_t(xy), M]$, and the last equality is from the fact that the probability measure $\Lambda_t(x, 1, dy)$ converges to the point mass at 1 as $x \to \infty$, and that $\lim_{x\to\infty} F(xy) = M$.

Given the power-law like regularly varying asymptotics of the marginal cost function F_t, the following corollary further tells us the distribution of total trading volume in equilibrium.

Corollary 3.5.4. *Assume* $\partial_x \Psi_t^+ (M)$ *and* $\partial_x \Psi_t^- (m)$ *exists,* $N_t > 1$ *and* $-\infty < m < M < \infty$*, then*

1. *If* $M - F_t(x)$ *is regularly varying of index* ρ_t^+ *at* ∞*, then* $\Pi_t^+ (F_t(x))$ *is regularly varying of index* $\frac{\partial_x \Psi_t^+(M)}{1-\partial_x \Psi_t^+(M)} \rho_t^+$ *at* ∞*.*

2. *If* $F_t(x) - m$ *is regularly varying of index* ρ_t^- *at* $-\infty$*, then* $\Pi_t^- (F_t(x))$ *is regularly varying of index* $\frac{\partial_x \Psi_t^-(m)}{1-\partial_x \Psi_t^-(m)} \rho_t^-$ *at* $-\infty$*.*

Proof. To show that $\Pi_t^+(F_t(x))$ is regularly varying at ∞ of index $\frac{\partial_x \Psi_t^+(M)}{1-\partial_x \Psi_t^+(M)} \rho_t^+$, we shall apply the converse part of Karamata's Theorem in Theorem 2.B.2. That is, we want to show

$$\lim_{x\to\infty} \frac{\Pi_t^+(F_t(x))}{\int_x^\infty \frac{\Pi_t^+(F_t(u))}{u}\,du} = -\frac{\partial_x \Psi_t^+(M)}{1-\partial_x \Psi_t^+(M)} \rho_t^+.$$

Step 1, by standard integration by parts, we can have $\Pi_t^+(x) = \frac{\int_x^M \Pi_t^+(y)\,dy}{\Psi_t^+(x)-x}$, which further gives us $-\frac{\partial_x \Pi_t^+(x)}{\Pi_t^+(x)} = \frac{\partial_x \Psi_t^+(x)}{\Psi_t^+(x)-x}$. Therefore,

$$\lim_{x\to\infty} \frac{\Pi_t^+(F_t(x))}{\int_x^\infty \frac{\Pi_t^+(F_t(u))}{u}\,du} = \lim_{x\to\infty} \frac{x\partial_x \Pi_t^+(F_t(x))F_t'(x)}{-\Pi_t^+(x)} = \lim_{x\to\infty} \frac{x\partial_x \Psi_t^+(F_t(x))F_t'(x)}{\Psi_t^+(F_t(x)) - F_t(x)}.$$

Step 2, $\Psi_t^+(F_t) - F_t$ is regularly varying of index ρ_t^+ at ∞. To see this,

$$\lim_{\alpha\to\infty} \frac{\Psi_t^+(F_t(\alpha x)) - F_t(\alpha x)}{\Psi_t^+(F_t(\alpha)) - F_t(\alpha)} = \lim_{\alpha\to\infty} \frac{\frac{\Psi_t^+(F_t(\alpha x))-F_t(\alpha x)}{M-F_t(\alpha x)}}{\frac{\Psi_t^+(F_t(\alpha))-F_t(\alpha)}{M-F_t(\alpha)}} \frac{M-F_t(\alpha x)}{M-F_t(\alpha)} = \lim_{\alpha\to\infty} \frac{M-F_t(\alpha x)}{M-F_t(\alpha)} = x^{\rho_t^+},$$

where the next-to-last equality is from

$$\lim_{x\to\infty} \frac{\Psi_t^+(F_t(x)) - F_t(x)}{M - F_t(x)} = \lim_{x\to\infty} \frac{(\partial_x \Psi_t^+(F_t(x)) - 1)F_t'(x)}{-F_t'(x)} = 1 - \partial_x \Psi_t^+(M).$$

Step 3, by the direct part of Karamata's Theorem 2.B.1, let $\sigma = -1$, we have

$$\lim_{x\to\infty} \frac{\Psi_t^+(F_t(x)) - F_t(x)}{\int_x^\infty \frac{\Psi_t^+(F_t(u))-F_t(u)}{u}\,du} = -\rho_t^+.$$

In other words,

$$\begin{aligned}\lim_{x\to\infty}\frac{x\partial_x\Psi_t^+(F_t(x))F_t'(x)}{\Psi_t^+(F_t(x))-F_t(x)} =&\rho_t^+ + \lim_{x\to\infty}\frac{xF_t'(x)}{\Psi_t^+(F_t(x))-F_t(x)}\\ =&\rho_t^+ + \lim_{x\to\infty}\frac{xF_t'(x)}{M-F_t(x)}\frac{M-F_t(x)}{\Psi_t^+(F_t(x))-F_t(x)}.\end{aligned}$$

Note that, $\lim_{x\to\infty}\frac{M-F_t(x)}{\Psi_t^+(F_t(x))-F_t(x)} = \frac{1}{1-\partial_x\Psi_t^+(M)}$. And another application of Karamata's Theorem give us $\lim_{x\to\infty}\frac{M-F_t(x)}{\int_x^\infty\frac{M-F_t(u)}{u}du} = \lim_{x\to\infty}\frac{xF_t'(x)}{M-F_t(x)} = -\rho_t^+$.
Step 4, combining step 1 and step 3, we obatin

$$\lim_{x\to\infty}\frac{\Pi_t^+(F_t(x))}{\int_x^\infty\frac{\Pi_t^+(F_t(u))}{u}du} = -\frac{\partial_x\Psi_t^+(M)}{1-\partial_x\Psi_t^+(M)}\rho_t^+.$$

By converse half of Karamata's Theorem, we have $\Pi_t^+(F_t)$ is regularly varying with index $\frac{\partial_x\Psi_t^+(M)}{1-\partial_x\Psi_t^+(M)}\rho_t^+$ at ∞. Similar derivation can be done for $\Pi_t^-(F_t)$. □

This result characterises the tail behaviour of the equilibrium aggregate trading volume from insiders:

$$P(X_t^* > x) = P_t\left(F_t^{-1}(V) > x\right) = P_t\left(V > F_t(x)\right) = \Pi_t^+\left(F_t(x)\right),$$

which is regularly varying at $+\infty$ with index

$$\frac{\partial_x\Psi_t^+(M)}{1-\partial_x\Psi_t^+(M)}\cdot\rho_t^+ = -\frac{\partial_x\Psi_t^+(M)}{1-\frac{\partial_x\Psi_t^+(M)}{N_t}}.$$

Moreover, the total order flow $Y_t^* = X_t^* + Z_t$, which includes both insider and noise trader orders, is also regularly varying at $+\infty$ with the same index.

To see this, note that

$$
\begin{aligned}
P_t(Y_t^* > y) = P_t(X_t^* + Z_t > y) &= \int_{-\infty}^{\infty} P_t(X_t^* > y - z)\, q_\nu(\sigma, z)\, dz \\
&= \int_{-\infty}^{\infty} P_t\left(V > F_t(y - z)\right)\, q_\nu(\sigma, z)\, dz \\
&= \int_{-\infty}^{\infty} \Pi_t^+(F_t(y - z))\, q_\nu(\sigma, z)\, dz \\
&= \int_{-\infty}^{\infty} \Pi_t^+(F_t(z))\, q_\nu(\sigma, y - z)\, dz.
\end{aligned}
$$

To establish regular variation and identify the same tail index, we compute the limit:

$$
\begin{aligned}
\lim_{\alpha\to\infty} \frac{P(Y_t^* > \alpha y)}{P(Y_t^* > \alpha)} &= \lim_{\alpha\to\infty} \frac{\int_{-\infty}^{\infty} \Pi_t^+(F_t(z))\, q_\nu(\sigma, \alpha y - z)\, dz}{\int_{-\infty}^{\infty} \Pi_t^+(F_t(z))\, q_\nu(\sigma, \alpha - z)\, dz} \\
&= \lim_{\alpha\to\infty} \frac{\int_{-\infty}^{\infty} \Pi_t^+(F_t(\alpha z))\, q_\nu\left(\frac{\sigma}{\alpha}, y - z\right)\, dz}{\int_{-\infty}^{\infty} \Pi_t^+(F_t(\alpha z))\, q_\nu\left(\frac{\sigma}{\alpha}, 1 - z\right)\, dz} \\
&= \lim_{\alpha\to\infty} \frac{\Pi_t^+(F_t(\alpha y))}{\Pi_t^+(F_t(\alpha))},
\end{aligned}
$$

where we used the scaling property of the t-density in the second step.

Hence, Y_t^* inherits the regular variation from the insider demand X_t^* with the same tail exponent.

3.6 Numerical experiments

In this section, we conduct numerical simulations to investigate the equilibrium and asymptotic price impact in our multiperiod LOB model under different settings on the prior distribution of the asset value. Specifically, we consider two settings: a Pareto prior and a Gaussian prior. Although the theoretical model as in Theorem 3.3.3 and Theorem 3.5.3 assumes a bounded asset distribution with $\operatorname{supp}(V) = [m, M]$, we find that the fixed point equation (3.3.8) converges numerically even when the asset distribution is unbounded. This enables us to simulate and study the equilibrium marginal cost function F_t, the marginal price function h_t, belief updates of liquidity suppliers, bid-ask spreads, and asymptotic market impact.

In these two asset distribution settings, we illustrate the marginal cost function F_t and the limit prices h_t, which together characterise the equilibrium price impact and the shape of the LOB. We examine how these objects evolve over trading periods in markets with varying numbers of informed traders. In addition, we investigate the posterior consistency of liquidity suppliers' beliefs about the asset value across time.

3.6.1 Pareto asset distribution

Suppose the fundamental value of the trading asset follows a Pareto distribution with shape parameter $\alpha > 0$, denoted by $V \sim \text{Pareto}(\alpha)$. Then, for $x \geq 1$, the density, distribution function, and expectation are given by

$$f_V(x) = \frac{\alpha}{x^{\alpha+1}}, \quad F_V(x) = 1 - \frac{1}{x^{\alpha}}, \quad E[V] = \frac{\alpha}{\alpha - 1}, \quad \text{for } \alpha > 1.$$

The associated tail and conditional expectation operators are

$$\begin{aligned} \Phi^+(y) &= \frac{\alpha}{\alpha-1} y^{-\alpha+1}, \quad \Pi^+(y) = y^{-\alpha}, \quad \Psi^+(y) = \frac{\alpha}{\alpha-1} y, \\ \Phi^-(y) &= \frac{\alpha}{1-\alpha} y^{-\alpha+1} + \frac{\alpha}{\alpha-1}, \quad \Pi^-(y) = 1 - y^{-\alpha}. \end{aligned}$$

In our simulation setup, we set $\alpha = 3$, so the prior expectation is $E[V] = 1.5$, with variance $\text{Var}(V) = 0.75$. This corresponds to a common belief that the asset follows a heavy-tailed Pareto distribution: its density decays as $f_V(x) \sim x^{-4}$, while the survival function exhibits power-law tail behaviour $P(V > x) \sim x^{-3}$. We assume the true fundamental value is $v_0 = 3$, which is known to the informed traders but unknown to noise traders and liquidity suppliers.

Noise trades are modelled by a location-scale Student's t-distribution with $\nu = 3$ degrees of freedom and scale parameter $\sigma = 0.2$, i.e.,

$$Z \sim t_3(0, \sigma).$$

The variance of this distribution is $\text{Var}(Z) = \frac{\sigma^2 \nu}{\nu - 2} = 0.12$. The location parameter is set to zero so that noise traders, on average, submit zero-mean orders. This reflects their non-strategic and non-informational role in the market.

In Figure 3.1, we present simulation results over six trading periods, $t = 0, \ldots, 5$. In each experiment, the number of informed traders per period is fixed as $N_t = N \in \{2, 3, 5, 10\}$, and we investigate the effect of varying N across runs. Informed traders are assumed to be myopic: a fresh set of N informed traders enters the market at each period t, trades once based on their private knowledge of the true fundamental value, and exits thereafter. These traders act independently across time and do not coordinate intertemporally.

Within each period, trading proceeds as follows. Noise trades and informed trades are aggregated and submitted as a batch to the market. These orders are executed against a limit order book formed by competitive liquidity suppliers, who quote prices based on their current posterior beliefs about the asset's value. After observing the net order flow, liquidity suppliers update their beliefs, and a new limit order book is constructed for the next trading period based on the updated posterior. Numerically, at each trading period t, we solve the fixed-point equation (3.3.8) to obtain the equilibrium marginal cost function F_t, from which the marginal pricing function h_t is computed via (3.3.6) and (3.3.7). Liquidity suppliers then update their beliefs according to (3.3.5).

Although simulations are conducted over six trading periods, Figure 3.1 displays results from periods $t = 1, \ldots, 5$, in order to highlight the evolution of the belief-updating mechanism beyond the initial prior. In Figure 3.2, we fit power-law curves to each F_t. The numerical results are consistent with Theorem 3.5.3, exhibiting both the expected regular variation in the tail shape and the decay of the regularly varying index $\rho_t^+ \to 0$ as $t \to \infty$.

Figure 3.3 illustrates the evolution of the bid–ask spread across trading periods $t = 1, \ldots, 5$, for varying numbers of informed traders $N \in \{2, 3, 5, 10\}$. The spread is measured as $h_t(0^+) - h_t(0^-)$, capturing the discontinuity in the marginal pricing function at the origin. This spread arises due to adverse selection: liquidity suppliers widen the spread to protect themselves from potential losses when trading against insiders. When more informed traders are present, competition among them intensifies, leading to more aggressive trading behaviour. This accelerates price discovery, as trading more effectively reveals information about the asset's true value. As a result, simulations with larger N exhibit consistently tighter bid–ask spreads across all trading periods.

Moreover, as trading progresses, liquidity suppliers update their beliefs based on the observed aggregate order flow. The posterior distributions gradually concentrate around the true fundamental value, reducing informational asymmetry and improving price efficiency. Figure 3.4 illustrates this learning dynamic. The solid curves represent posterior densities from a representative simulation, while dashed and dotted lines show results from four additional runs with different random seeds. Across all simulations, beliefs converge towards the true value $v_0 = 3$, reflecting posterior consistency over time. In these simulations, we fix the number of informed traders at $N_t = 10$ in every period, and adopt the same parameter configuration used in previous figures. Noise trades follow a location-scale Student-t

distribution with degrees of freedom $\nu = 3$ and scale parameter $\sigma = 0.2$. The prior distribution of the asset value V is Pareto with shape parameter $\alpha = 3$, and the true fundamental value is set at $v_0 = 3$. At each trading period $t \in \{0, \ldots, 5\}$, we solve the fixed-point equation (3.3.8) to obtain the equilibrium pricing function F_t, and then compute the posterior via (3.3.5).

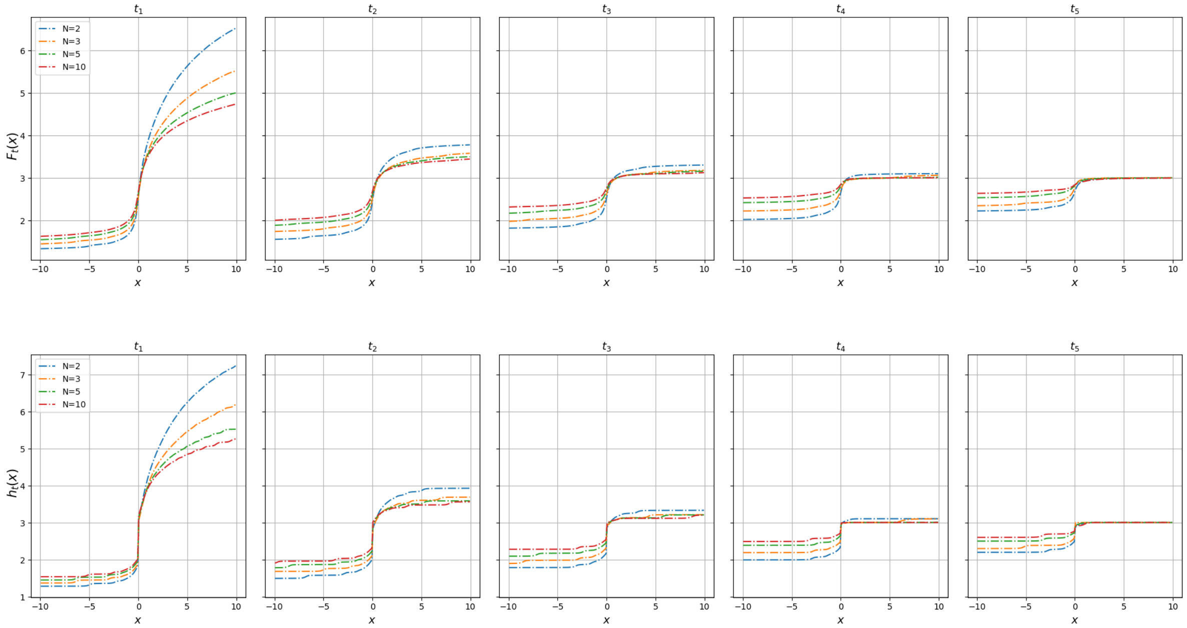


FIGURE 3.1: Marginal cost function $F_t(x)$ (top) and marginal price function $h_t(x)$ (bottom) in equilibrium over trading periods $t = 1, \ldots, 5$, with varying numbers of informed traders $N_t \in \{2, 3, 5, 10\}$. In each setting, N_t is held constant across all periods but represents a new, independent cohort of myopic informed traders who participate in a single trading round and then exit the market. The true fundamental value is $v_0 = 3$, drawn from a Pareto distribution $V \sim \text{Pareto}(3)$, and noise trades follow a location-scale Student's t-distribution $Z_t \sim t_3(0, 0.2)$.

3.6.2 Gaussian Asset Distribution

We consider a setting in which the distribution of the asset value V is commonly agreed upon by all market participants to be Gaussian: $V \sim \mathcal{N}(0, \sigma^2)$, with variance $\sigma^2 = 4$. This prior represents the market consensus at time $t = 0$, and in particular, the belief held by liquidity suppliers. The true value of the asset is fixed at $v_0 = 1$, known only to the informed traders. Noise trades are independently drawn from a heavy-tailed distribution $t_3(0, 1)$, capturing random order flow that is unrelated to fundamental value.

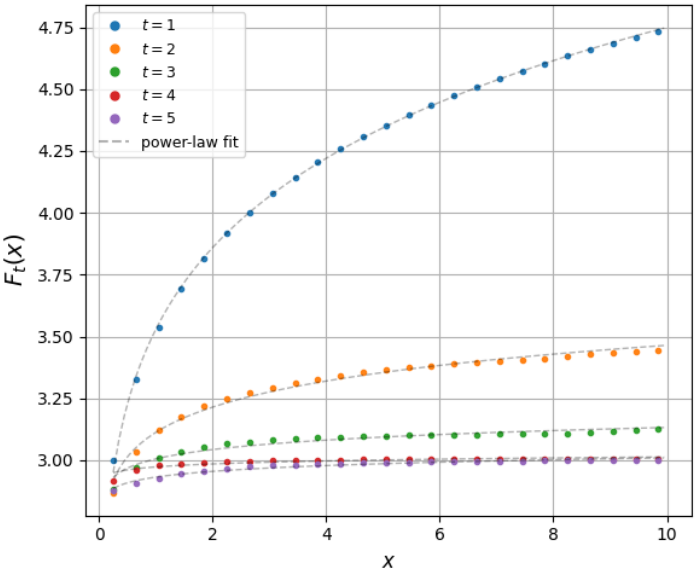


FIGURE 3.2: Power-law fits to the pricing function $F_t(x)$ across trading periods $t = 1, \ldots, 5$, with the number of insiders fixed at $N_t = 10$. Dots represent the numerically computed $F_t(x)$ obtained from solving the fixed-point equation (3.3.8), while the black dashed lines indicate fitted power-law curves of the form $F_t(x) \sim x^{\rho_t^+}$. The fitted exponents ρ_t^+ capture the regular variation behaviour of F_t in the right tail. The true fundamental value is $v_0 = 3$, drawn from a Pareto distribution $V \sim \text{Pareto}(3)$, and noise trades follow a location-scale t-distribution $Z_t \sim t_3(0, 0.2)$.

Given this Gaussian prior, the tail probabilities and tail expectations used by liquidity suppliers at $t = 0$ are:

$$\Phi^+(y) = \sqrt{\frac{\sigma^2}{2\pi}} \exp\left(-\frac{y^2}{2\sigma^2}\right), \quad \Pi^+(y) = \frac{1}{2} \operatorname{erfc}\left(\frac{y}{\sqrt{2\sigma^2}}\right),$$
$$\Phi^-(y) = -\sqrt{\frac{\sigma^2}{2\pi}} \exp\left(-\frac{y^2}{2\sigma^2}\right), \quad \Pi^-(y) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{y}{\sqrt{2\sigma^2}}\right)\right].$$

The beliefs of liquidity suppliers are then updated dynamically using the Bayesian updating formula (3.3.5) after each observed trade.

Throughout the trading horizon $t = 0, 1, \ldots, 4$, the number of informed traders is fixed at $N_t = 10$, and an equilibrium is computed at each period. As trading progresses and new prices are observed, the posterior beliefs of liquidity suppliers become increasingly concentrated around the true value v_0, reflecting information aggregation through prices.

The marginal cost function $F_t(x)$ and the marginal pricing function $h_t(x)$ both evolve

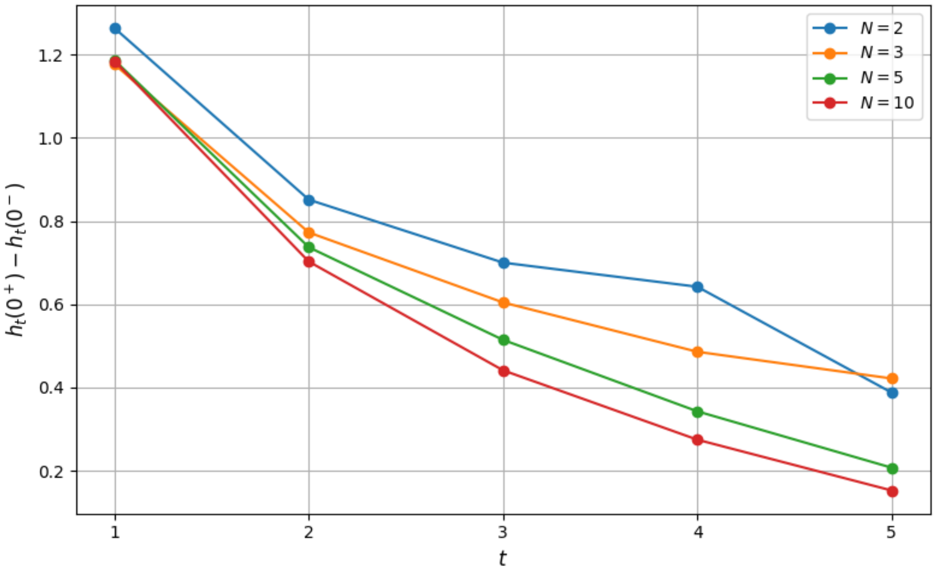


FIGURE 3.3: Bid-ask spread over trading periods $t = 1, \ldots, 5$ for different numbers of informed traders $N \in \{2, 3, 5, 10\}$. The true fundamental value is $v_0 = 3$, drawn from a Pareto distribution $V \sim \text{Pareto}(3)$, and noise trades follow a location-scale t-distribution $Z_t \sim t_3(0, 0.2)$. Each line corresponds to a fixed N, representing a separate simulation where the same number of myopic informed traders enters the market in each period. The spread narrows both over time within each simulation and across simulations as N increases, indicating that a larger number of informed traders leads to tighter markets.

over time, with the latter flattening around the true value v_0, indicating increased price informativeness. This is visible in Figure 3.5. The narrowing bid-ask spread over time, shown in Figure 3.7, further quantifies the improvement in market efficiency. Figure 3.6 illustrates the full dynamics of belief updating across trading periods. Starting from the initial Gaussian prior centred at zero, liquidity suppliers gradually filter out the noise component of trades and shift their posterior beliefs toward the true value v_0. This learning process is a direct consequence of inferring the asset's fundamental value from observed aggregate trading volumes, which contain informed and noise trades.

3.7 Discussion

In the previous setting, we assumed that informed traders know the distribution of noise trades, given by $t_\nu(0, \sigma)$. We now consider an alternative specification in which insiders do

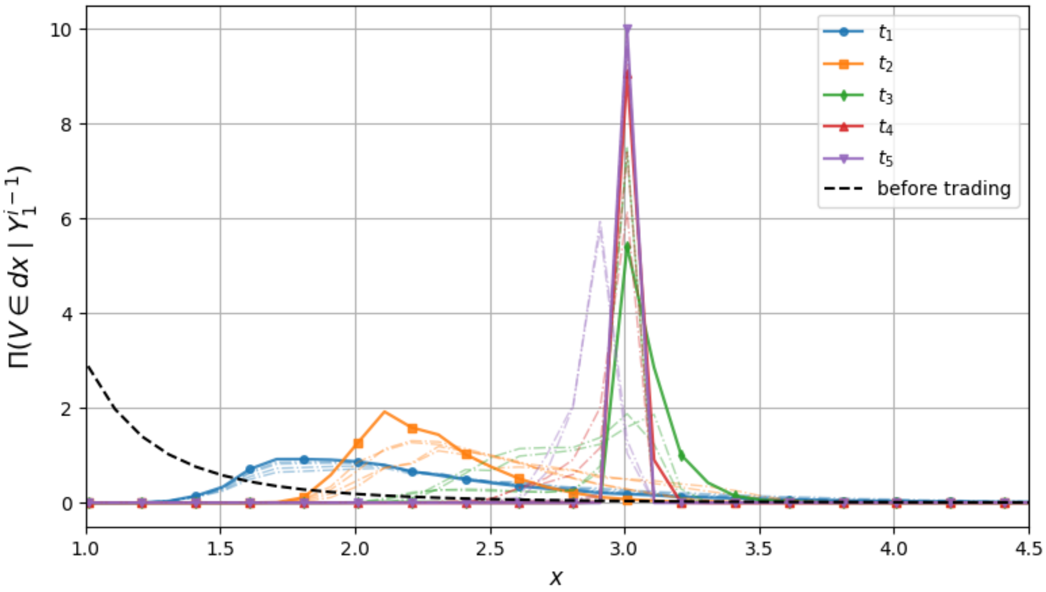


FIGURE 3.4: Evolution of the liquidity suppliers' posterior beliefs about the asset value V across trading periods $t = 1, \ldots, 5$. Solid curves show the posterior densities from one representative simulation; dashed and dotted lines show four additional runs with different random seeds. The initial prior (black dashed line) is Pareto with shape $\alpha = 3$, and the true value is $v_0 = 3$. Noise trades follow a Student-t distribution $t_3(0, 0.2)$, and the number of informed traders is fixed at $N_t = 10$. As trading progresses, beliefs concentrate around the true value, indicating posterior consistency.

not know the exact distribution of noise. Instead, they begin with a prior belief about the noise variance and update it sequentially via Bayesian learning.

Suppose the aggregate noise trade in period t is denoted $Z_t \sim t_{2\alpha}(0, \sqrt{\alpha/\beta})$, where (α, β) are *unknown* to the informed traders. These traders place an inverse-gamma prior on the variance σ^2 and update their beliefs as new order flow observations arrive. Let $\mathcal{F}_t^I := \sigma(Y_0^{t-1}, V)$ denote the insiders' information at time t, consisting of the trading history up to time $t-1$ and the true asset value $V = v_0$.

If the variance parameters (α, β) were known, then the conditional law of Y_t admits a Gaussian mixture representation with inverse-gamma mixing density:

$$P(Y_t \in dy \mid \mathcal{F}_t^I) \propto \int_0^\infty P(Y_t \in dy \mid \sigma^2, F_t^{-1}(v_0)) \cdot p(\sigma^2 \mid \alpha, \beta)\, d\sigma^2,$$

where $Y_t \mid \sigma^2, F_t^{-1}(v_0) \sim \mathcal{N}(F_t^{-1}(v_0), \sigma^2)$ and $\sigma^2 \sim \text{Inv-Gamma}(\alpha, \beta)$. By the standard representation of the location-scale t-distribution as a normal-inverse-gamma mixture (see

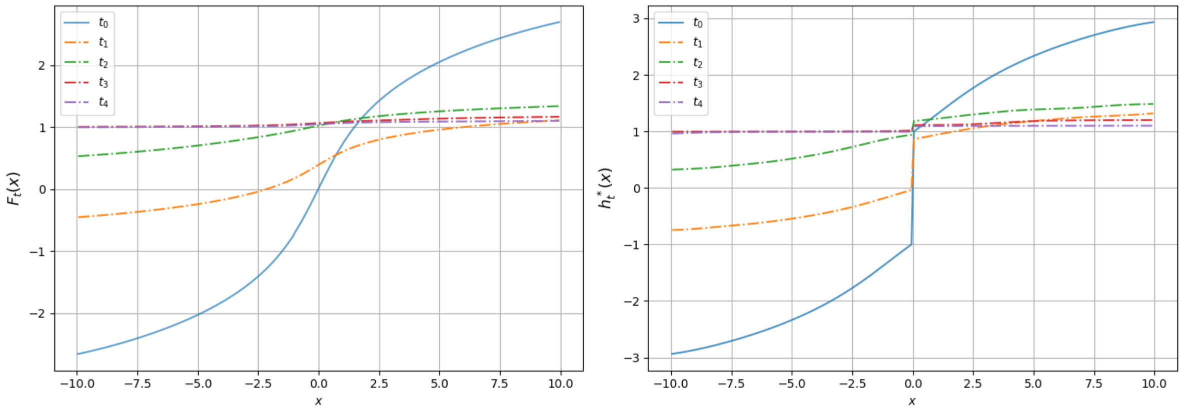


FIGURE 3.5: Marginal cost function $F_t(x)$ (left) and marginal price function $h_t(x)$ (right) over trading periods $t = 0, \dots, 4$, under a Gaussian prior $\mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 4$, and true value fixed at $v_0 = 1$. The functions evolve as liquidity suppliers update their beliefs from trading data. As t increases, the marginal cost flattens near the true value, and the marginal price function becomes steeper, leading to a tighter bid-ask spread.

Appendix 3.A.2), it follows that

$$Y_t \mid \mathcal{F}_t^I \sim t_{2\alpha}(F_t^{-1}(v_0), \sqrt{\alpha/\beta}).$$

In the learning setting, however, insiders treat σ^2 as unknown and estimate it recursively from observed trading volumes. The inverse-gamma prior is conjugate to the Gaussian likelihood, and right after observing y_{t-1}, the posterior distribution for σ^2 becomes:

$$\begin{aligned}\pi_t(\sigma^2 \mid Y_0^{t-1}, V) &\propto \underbrace{p(y_{t-1} \mid \sigma^2, F_{t-1}^{-1}(v_0))}_{\text{likelihood}} \cdot \underbrace{p(\sigma^2 \mid \alpha_{t-1}, \beta_{t-1})}_{\text{prior}} \\ &\propto (\sigma^2)^{-\alpha_{t-1}-\frac{1}{2}-1} \exp\left\{-\frac{1}{\sigma^2}\left[\beta_{t-1} + \frac{1}{2}(y_{t-1} - F_{t-1}^{-1}(v_0))^2\right]\right\} \\ &\propto \text{Inv-Gamma}(\alpha_t, \beta_t),\end{aligned}$$

where the updated parameters are:

$$\alpha_t = \alpha_{t-1} + \frac{1}{2}, \quad \beta_t = \beta_{t-1} + \frac{1}{2}(y_{t-1} - F_{t-1}^{-1}(v_0))^2.$$

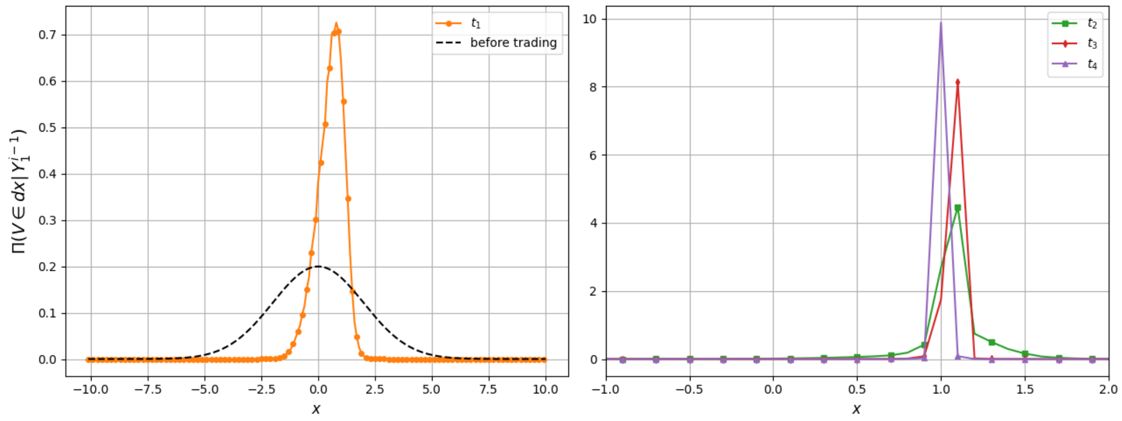


FIGURE 3.6: Evolution of the liquidity suppliers' posterior beliefs about the asset value V across trading periods $t = 1, \ldots, 4$. The initial belief (black dashed line) follows a Gaussian prior distribution, $\mathcal{N}(0, \sigma^2)$, with variance $\sigma^2 = 4$, and the true value is fixed at $v_0 = 1$. As more trading data are observed, the posteriors concentrate around the true value, illustrating posterior consistency from the perspective of the liquidity suppliers.

This recursive structure yields a natural estimate of the prevailing noise variance at time t, given by the posterior mean:

$$E[\sigma^2 \mid \mathcal{F}_t^I] = \frac{\beta_t}{\alpha_t - 1}, \quad \text{for } \alpha_t > 1.$$

This formulation enables insiders to gradually infer the volatility of noise trades from observed order flow, rather than assuming full knowledge of its distribution.

These observations highlight the flexibility of our framework in accommodating informational asymmetries and uncertainty about market noise. While the benchmark setting assumes complete knowledge of the noise distribution, the analysis above shows that informed traders can begin with a prior and sequentially learn the intensity of noise trades. This extension preserves equilibrium tractability while allowing for richer belief dynamics within a Bayesian learning environment.

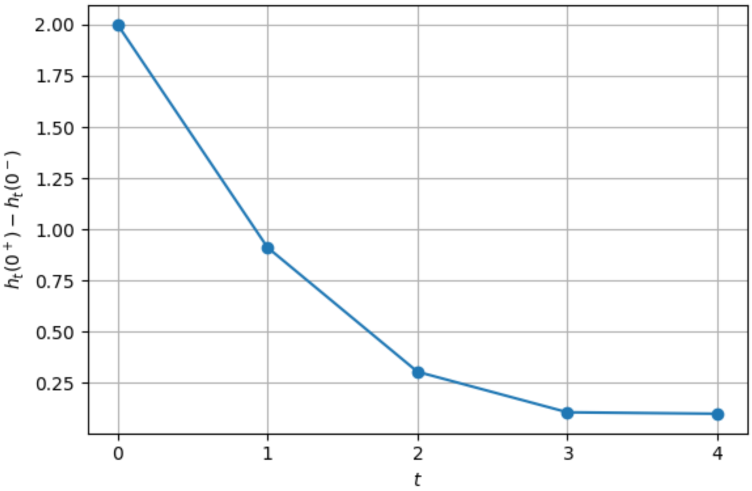


FIGURE 3.7: Evolution of the bid-ask spread across trading periods $t = 0, \ldots, 4$, with the number of informed traders fixed at $N_t = 10$ for all t. The spread narrows over time as liquidity suppliers learn the true asset value, resulting in more efficient and tighter markets. The noise trader's signal is distributed as $t_3(0, 0.2)$, and the prior on V is Gaussian with variance $\sigma^2 = 4$.

Appendix

3.A Probability distributions

3.A.1 Location-scale Student's t distribution

A continuous random variable X has a location-scale *Student's t* distribution with location μ, scale σ, and degree of freedom ν, written $x \sim \mathsf{T}_\nu(\mu, \sigma)$, if the density function of X is

$$p_\nu(\sigma, \mu, x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}\sigma}\left\{1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right\}^{-\frac{(\nu+1)}{2}}, \quad x \in \mathbb{R},$$

where $\Gamma(\cdot)$ denotes the gamma function. We have $E[x] = \mu$, if $\nu > 1$, and $Var[x] = \sigma^2 \frac{\nu}{\nu-2}$, if $\nu > 2$.

3.A.2 Normal-Inverse Gamma representation

The location-scale t distribution can be written as a mixture of Gaussians, where the variance follows an Inverse-Gamma distribution. That is, if we have a hierarchical representation

$$\begin{aligned} Y \mid \mu, \sigma^2 &\sim \mathcal{N}(\mu, \sigma^2) \\ \sigma^2 &\sim \text{Inv-Gamma}(\alpha, \beta) \end{aligned}$$

Then, by marginalising out σ^2, we shall have

$$Y \mid \ \mu \sim t_{2\alpha}(\mu, \sqrt{\frac{\beta}{\alpha}}).$$

To see this, the key integral identity is $\lambda^{-\nu+1}\Gamma(\nu-1) = \int_0^\infty x^{-\nu}e^{-\frac{\lambda}{x}}\,dx$ for all $\nu > 1$.

$$\begin{aligned}
p(y|\mu) &= \int_0^\infty p(y|\mu,\sigma^2)p(\sigma^2)\,d\sigma^2 \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)\frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}\exp\left(-\frac{\beta}{\sigma^2}\right)d\sigma^2 \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)\sqrt{2\pi}}\int_0^\infty (\sigma^2)^{-\alpha-\frac{3}{2}}\exp\left(-\frac{1}{\sigma^2}\left(\beta+\frac{(y-\mu)^2}{2}\right)\right)d\sigma^2 \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)\sqrt{2\pi}}\left(\beta+\frac{(y-\mu)^2}{2}\right)^{-(\alpha+\frac{3}{2})+1}\Gamma((\alpha+\frac{3}{2})-1) \\
&= \frac{\Gamma(\frac{2\alpha+1}{2})}{\Gamma(\frac{2\alpha}{2})\sqrt{\pi 2\alpha}\sqrt{\frac{\beta}{\alpha}}}\left(1+\frac{(y-\mu)^2}{2\alpha\frac{\beta}{\alpha}}\right)^{-\frac{2\alpha+1}{2}}.
\end{aligned}$$

It provides conjugacy property of posterior update of σ^2:

$$\begin{aligned}
p(\sigma^2 \mid y) &\propto p(y \mid \sigma^2)p(\sigma^2) \\
&\propto \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)\frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}\exp\left(-\frac{\beta}{\sigma^2}\right) \\
&\propto (\sigma^2)^{-\alpha-\frac{1}{2}-1}\exp\left(-\frac{1}{\sigma^2}(\beta+\frac{1}{2}(y-\mu)^2)\right) \\
&\propto \text{Inv-Gamma}(\alpha',\beta'),
\end{aligned}$$

where

$$\alpha' = \alpha+\frac{1}{2}, \quad \beta' = \beta+\frac{1}{2}(y-\mu)^2.$$

Note that, for inverse gamma distribution, we have $E[\cdot] = \frac{\beta}{\alpha-1}$, for $\alpha > 1$, $Var(\cdot) = \frac{\beta}{(\alpha-1)^2(\alpha-2)}$, for $\alpha > 2$.

3.A.3 Derivative property

Lemma 3.A.1. *Let $\nu > 0$ and $\sigma > 0$. Then the integral*

$$I := \int_{-\infty}^\infty \left|\frac{\partial}{\partial u}q_\nu(\sigma,u)\right| du = \int_{-\infty}^\infty \frac{(\nu+1)|u|}{\nu\sigma^2+u^2}q_\nu(\sigma,u)\,du$$

is finite and strictly positive.

Proof. We shall split the integral into two parts:

$$I = \int_{|u|\leq 1} \frac{(\nu+1)|u|}{\nu\sigma^2+u^2}\, q_\nu(\sigma,u)\, du + \int_{|u|>1} \frac{(\nu+1)|u|}{\nu\sigma^2+u^2}\, q_\nu(\sigma,u)\, du.$$

Finiteness:

On the region $|u| > 1$, observe:

$$\frac{(\nu+1)|u|}{\nu\sigma^2+u^2} \leq (\nu+1)\cdot\frac{1}{|u|},$$

and the density satisfies, for large $|u|$,

$$q_\nu(\sigma,u) = C_\nu\left(1+\frac{u^2}{\nu\sigma^2}\right)^{-(\nu+1)/2} \leq \frac{C'_\nu}{|u|^{\nu+1}} \quad \text{for some constant } C'_\nu > 0.$$

Thus,

$$\frac{(\nu+1)|u|}{\nu\sigma^2+u^2}\cdot q_\nu(\sigma,u) \leq \frac{C}{|u|^{\nu+2}} \quad \text{for } |u| > 1,$$

and the integrand is integrable on $|u| > 1$ since $\nu+2 > 1$.

On the region $|u| \leq 1$,

$$\frac{(\nu+1)|u|}{\nu\sigma^2+u^2}\cdot q_\nu(\sigma,u)$$

is continuous and bounded, hence integrable. Therefore, $I < \infty$.

Positivity: The integrand is strictly positive for all $u \neq 0$, and in particular on any interval $[\delta,1]\cup[-1,-\delta]$ with $\delta > 0$, which has positive measure. Hence,

$$\int_{|u|\leq 1} \frac{(\nu+1)|u|}{\nu\sigma^2+u^2}\, q_\nu(\sigma,u)\, du > 0,$$

and thus $I > 0$. □

Chapter 4

Probabilistic Activations and Variational Inference via Mixture of Experts in Bayesian Neural Networks

4.1 Introduction

Over the past decade, deep learning has significantly outperformed other statistical learning methods in various domains, including image analysis (e.g., Szegedy et al., 2015), speech recognition (e.g., Graves et al., 2013), and natural language processing (e.g., Hinton et al., 2012). However, deep learning techniques such as deep neural networks (DNNs; LeCun et al., 2015; Goodfellow et al., 2016) face several challenges, including overconfidence in predictions, and lack of interpretability and explainability with many models considered “black boxes" (Nguyen et al., 2015; Lipton, 2018). Furthermore, effectiveness and reliability of DNNs is often limited by the lack of reliable uncertainty estimates (Guo et al., 2017; Ashukha et al., 2020). Bayesian neural networks (BNNs) represent an attempt to overcome this issue: they combine the advantages of DNNs and Bayesian inference (Ghahramani, 2015) by means of probability distributions that allow for the expression of uncertainty, and provide greater insight into the accuracy of output and predictions (Lampinen and Vehtari, 2001; Titterington, 2004; Goan and Fookes, 2020; Jospin et al., 2022).

DNNs are a parametrized composition of simple functions that can effectively model complex relationships when arranged in multiple layers, and have become a powerful tool for function approximation. Each layer operates through a linear transformation followed by

[1]Università degli Studi di Padova
[2]London School of Economics and Political Science
[3]The Australian National University

an element-wise non-linear transformation, known as the activation function. A common example of an activation function is the rectified linear unit (ReLU), defined as $\max(x, 0)$. A DNN, specifically a feed forward neural network (FFNN), is characterized by its depth L (i.e. the number or hidden layers) and a vector $\boldsymbol{p} = (p_0, \ldots, p_{L+1})^\top$ that lists the width of each layer. We follow the indexing convention of Wang et al. (2022) and use p_0 to represent the dimension of the output layer, $p_1, \ldots, p_L$ to denote the dimensions of the L hidden layers and p_{L+1} for the dimension of the input variable. The DNN architecture includes the weight matrices $\boldsymbol{W}_\ell \in \mathbb{R}^{p_\ell \times p_{\ell+1}}$, $\ell = 0, \ldots, L$, that connect the $(\ell+1)^{th}$ layer to the ℓ^{th} layer, and the vectors $\boldsymbol{b}_\ell$, $\ell = 0, \ldots, L$, representing the bias terms associated with the ℓ^{th} layer.

The core of training DNNs lies in gradient-based optimization, executed primarily through a technique known as backpropagation (Goodfellow et al., 2016). For larger datasets, this process is frequently replaced by stochastic gradient descent (SGD), which approximates gradients on randomly selected subsets called batches. However, SGD can lead to challenges such as local minima traps and overfitting, largely due to the non-convex nature of the optimization problem. A compelling solution to these issues is to introduce stochastic elements into the activation functions. Shridhar et al. (2019) developed an innovative probabilistic activation function named *ProbAct* which is decomposed into a mean and variance, allowing for the output value to be sampled from the formed distribution and making the activation mechanism stochastic. Additionally, Smith et al. (2021) presented the *bow tie network*, a deep generative model that enhances rectified linear FFNNs by integrating stochastic activations. This model cleverly employs the Pólya-gamma data augmentation scheme proposed by Polson and Scott (2013) to create a conditionally conjugate model, further demonstrating the potential of stochastic activations in DNNs (see also Sheinkman and Wade, 2024).

Data augmentation strategies play a crucial role in improving model performance, as demonstrated in Polson et al. (2013) and Wang et al. (2022). Notably, the latter study utilizes hierarchical representations of models with regularizations to find maximum a posteriori (MAP) estimates (Polson and Scott, 2011). By exploiting the Bayesian framework, Wang et al. (2022) transform the output layer of a DNN into a stochastic layer whose parameters $\boldsymbol{W}_0$ and $\boldsymbol{b}_0$ are updated through data augmentation of an appropriate conditional distribution that facilitates MAP estimation. The parameters $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_L$ and $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L$ of the lower layers are refined using SGD. In this work, we advance the proposal of Wang et al. (2022) by substituting the deterministic layers and/or activation functions in the deep architecture with stochastic layers and/or probabilistic equivalents, exploiting data augmentation to enhance model robustness. The BNN framework we propose is particularly versatile in a sense that it can be easily adapted to various types of responses (e.g., binary outcomes, categorical data, count data, and continuous variables) by adding a linear top layer connected to the output.

One of the primary challenges in Bayesian machine learning is performing inference, particularly because the posterior distribution of complex models is a high-dimensional and highly non-convex probability distribution function (e.g., Izmailov et al., 2021). Computing posterior distributions 'exactly' using Markov Chain Monte Carlo (MCMC) methods (e.g., Hastings, 1970; Casella and George, 1992; Chib and Greenberg, 1995) can be time-consuming and often impractical. A popular alternative is variational Bayes (e.g., David M. Blei and McAuliffe, 2017; Zhang et al., 2019), which provides approximate posterior distributions that can be typically obtained by means of computationally efficient optimization algorithms (e.g., Ormerod and Wand, 2010). Mean field variational Bayes (MFVB) assumes that the joint posterior density function is factorized as a product of density functions (Wainwright et al., 2008) and is arguably the simplest and most commonly used variational Bayes approach. Some of the densities of this density-product approximation can be assumed to be part of pre-specified parametric families and this can yield significant practical advantages, especially when one or more MFVB approximations to the marginal posterior densities arising from the model are intractable (e.g., Barber and Bishop, 1997; Tan and Nott, 2013). Here, we refer to this relaxation of MFVB as *semi-parametric mean-field variational Bayes*, adopting the nomenclature of Rohde and Wand (2016). Specifically, in our BNNs with probabilistic activation functions we approximate the marginal posterior densities of weight matrices and bias terms using regular MFVB. The marginal posterior approximating densities of augmented variables used to hierarchically represent probabilistic activation functions are instead assumed to be from pre-specified parametric families.

A key aspect of our work is the integration of mixture of experts (MoE; Jordan and Jacobs, 1994; Bishop and Svenskn, 2003; Jacobs et al., 1991; Waterhouse et al., 1995) into our framework. Traditional statistics often operates under the assumption of a universal relationship among variables across an entire population, which can lead to the dismissal of observations that do not conform to the model as mere outliers. In certain instances, this viewpoint may oversimplify the complexity of reality, where multiple subpopulations likely exhibit unique behaviours that deserve investigation. MoE models are known for their computational efficiency, especially in large-scale applications, and can boost predictive accuracy by capturing complex input-output relationships. This is accomplished through a gating network that adaptively partitions the input space and assigns specialized experts to specific regions, allowing for more accurate and tailored inferences. Importantly, the choice of the experts is not tied to a particular selection mechanism and can be performed, for instance, by means of decision tree classifiers, K-means clustering, Gaussian mixture models and neural networks. In the context of our work, the significance of MoE extends beyond these general advantages. They play a crucial role in facilitating the prediction process, especially when employing variational approximations on augmented variables. For a new test input, the learned gating network selects an expert based on the input's structure; the assigned expert then samples the posteriors of the augmented variables from the optimal parametric variational densities it has learned, ensuring accurate and efficient inference.

The following section, Section 4.2, describes the proposed BNNs. In Section 4.3 we describe our approach for learning the proposed BNNs, as well as semi-parametric MFVB, MoE and the procedures for inference and prediction. Sections 4.4 and 4.5.2 illustrate the application of our methodology to regression and classification, respectively. Section 4.6 demonstrates the positive performance of our approach in some numerical studies, and Section 4.7 provides conclusions and a discussion.

4.2 Bayesian neural networks

We start by briefly reviewing FFNNs, which are a particular type of DNNs where information flows in one direction from the input to the output layers, and their use for function approximation. We then introduce BNNs with stochastic (Gaussian) layers and probabilistic activation functions in Subsections 4.2.1.1 and 4.2.1.2.

Let L be a positive integer and $\boldsymbol{p} = (p_0, \dots, p_{L+1})^\top$ a vector of positive integers, where p_0 and p_{L+1} are respectively the dimensions of the output and input layers. A $(L+1)$-layer FFNN with layer width $\boldsymbol{p}$ belongs to the following composite function class:

$$\mathfrak{F}_{\boldsymbol{\theta}} := \{f_{\boldsymbol{\theta}}(\cdot) : f_{\boldsymbol{\theta}}(\cdot) = (f_{\boldsymbol{W}_0,\boldsymbol{b}_0} \circ f_{\boldsymbol{W}_1,\boldsymbol{b}_1} \circ \cdots \circ f_{\boldsymbol{W}_L,\boldsymbol{b}_L})(\cdot)\}, \tag{4.2.1}$$

where $\circ$ denotes function composition, and $\boldsymbol{\theta} = \{(\boldsymbol{W}_0, \boldsymbol{b}_0), \dots, (\boldsymbol{W}_L, \boldsymbol{b}_L)\}$ are trainable parameters. In detail, $\boldsymbol{W}_\ell \in \mathbb{R}^{p_\ell \times p_{\ell+1}}$ and $\boldsymbol{b}_\ell \in \mathbb{R}^{p_\ell}$, $\ell = 0, \dots, L$, are respectively weight matrices connecting the $(\ell+1)^{th}$ layer to the ℓ^{th} layer, and vectors containing bias terms associated with the ℓ^{th} layer. For any vector $\boldsymbol{x} \in \mathbb{R}^{p_{\ell+1}}$, the functions $f_{\boldsymbol{W}_\ell,\boldsymbol{b}_\ell}(\cdot)$ are defined as $f_{\boldsymbol{W}_\ell,\boldsymbol{b}_\ell}(\boldsymbol{v}) = \sigma_\ell(\boldsymbol{W}_\ell \boldsymbol{v} + \boldsymbol{b}_\ell)$, where σ_ℓ are simple non-linear transformations that operate component-wise, that is, $\sigma_\ell((v_1, \dots, v_{p_\ell})^\top) = (\sigma_\ell(v_1), \dots, \sigma_\ell(v_{p_\ell}))^\top$.

Popular choices of activation functions are the ReLU (Nair and Hinton, 2010) and the logistic functions. Let $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{p_{L+1}} \times \mathbb{R}^{p_0}$ — or $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{p_{L+1}} \times \{0, 1\}^{p_0}$ in case of a classification task with two possible labels treated later on — be a pair of random vectors satisfying $\mathsf{E}(\boldsymbol{y}^\top \boldsymbol{y}) < \infty$. Given a sample of size n of $(\boldsymbol{x}, \boldsymbol{y})$, that is, given a training data set $\mathcal{D}_n = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \dots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$, where $(\boldsymbol{x}_1, \boldsymbol{y}_1), \dots, (\boldsymbol{x}_n, \boldsymbol{y}_n)$ are i.i.d., the aim is to estimate $\boldsymbol{\theta}$ as

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\{\boldsymbol{y}_i, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\}, \tag{4.2.2}$$

where $\mathcal{L}(\cdot, \cdot)$ is a fixed loss function which depends on the learning task.

A regularization term is commonly included to control the predictive bias-variance trade-off (e.g., Moradi et al., 2020). In this context, the estimator $\widehat{\boldsymbol{\theta}}$ is such that

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\{\boldsymbol{y}_i, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\} + \lambda h(\boldsymbol{\theta}), \tag{4.2.3}$$

for some regularization coefficient $\lambda \in \mathbb{R}^+$ and regularization penalty function $h(\cdot)$. This optimization problem for training the deep learner $f_{\boldsymbol{\theta}}(\cdot)$ can be interpreted as finding maximum a posteriori (MAP) estimates (e.g., Polson and Sokolov, 2017; Moradi et al., 2020; Wang et al., 2022). To see this, note that the MAP estimator is

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}_{\text{MAP}} &= \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \log \mathfrak{p}(\boldsymbol{\theta} \mid \mathcal{D}_n) \\
&= \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \log \mathfrak{p}(\mathcal{D}_n \mid \boldsymbol{\theta}) + \log \mathfrak{p}(\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{n} \log \mathfrak{p}(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) + \log \mathfrak{p}(\boldsymbol{\theta})
\end{aligned} \tag{4.2.4}$$

where $\log \mathfrak{p}(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$ is the log-likelihood, and $\mathfrak{p}(\boldsymbol{\theta})$ is the prior. If $\mathcal{L}(\boldsymbol{y}_i, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) = -\log \mathfrak{p}(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$ and $\mathfrak{p}(\boldsymbol{\theta}) \propto \exp\{-\lambda h(\boldsymbol{\theta})\}$, the MAP estimator in (4.2.4) corresponds to the estimator in (4.2.3), and if one assumes an uninformative prior $\mathfrak{p}(\boldsymbol{\theta}) \propto 1$, the MAP estimator is equivalent to the estimator in (4.2.2). Hence, the duality between optimization and Bayesian inference shows that the likelihood function choice is directly linked to the form of loss functions, while the prior specification is connected to the regularization. For example, a Gaussian likelihood yields the mean square error (MSE) loss function which is commonly used for regression tasks, and a Bernoulli likelihood yields a cross-entropy loss which is widely used for classification tasks. Regarding priors and regularizations, examples include the equivalence between Gaussian priors and L^2-norm regularization, or between Laplace

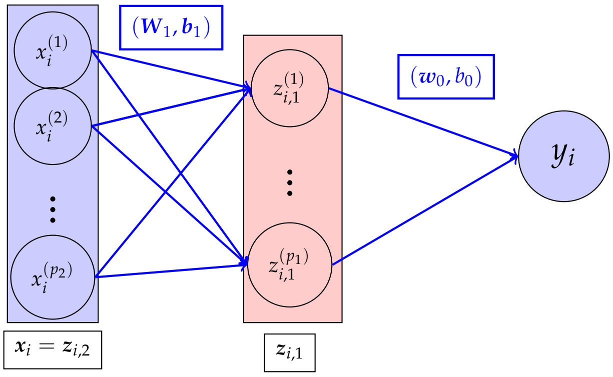

$x_i^{(1)}$
$x_i^{(2)}$
$x_i^{(p_2)}$
$(\boldsymbol{W}_1, \boldsymbol{b}_1)$
$z_{i,1}^{(1)}$
$z_{i,1}^{(p_1)}$
$(\boldsymbol{w}_0, b_0)$
y_i
$\boldsymbol{x}_i = \boldsymbol{z}_{i,2}$
$\boldsymbol{z}_{i,1}$

FIGURE 4.1: Example of a simple FFNN with one hidden layer ($L = 1$). The input vector $\boldsymbol{x}_i = \left(x_i^{(1)}, \ldots, x_i^{(p_2)}\right)^\top$ is passed to the augmented variable vector $\boldsymbol{z}_{i,1} = \left(z_{i,1}^{(1)}, \ldots, z_{i,1}^{(p_1)}\right)^\top$, which is then used to predict the output y_i.

priors and L^1-norm regularization.

Example 4.2.1. *Let us consider a FFNN defined with one hidden layer, that is, $L = 1$, and layer dimensions specified by the tuple $\boldsymbol{p} = (p_0, p_1, p_2)$, where $p_0 = 1$. The network parameters are given by $\boldsymbol{\theta} = \left\{(\boldsymbol{w}_0, b_0), (\boldsymbol{W}_1, \boldsymbol{b}_1)\right\}$, where $\boldsymbol{w}_0 \in \mathbb{R}^{p_0 \times p_1}$ and $\boldsymbol{W}_1 \in \mathbb{R}^{p_1 \times p_2}$, then, for any $\boldsymbol{x}_i \in \mathbb{R}^{p_2}$ are the weight matrices, and $\boldsymbol{b}_0$, $\boldsymbol{b}_1$ are the corresponding bias vectors. Then, for any input $\boldsymbol{x}_i \in \mathbb{R}^{p_2}$, the network output is given by*

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \sigma_0\big(\boldsymbol{w}_0(\sigma_1(\boldsymbol{W}_1\boldsymbol{x}_i + \boldsymbol{b}_1) + b_0\big), \tag{4.2.5}$$

for $i = 1, \ldots, n$, where σ_1 and σ_0 denote the inner and outer activation functions, respectively. Moreover, we have

$$\begin{aligned} y_i &= \sigma_0(\boldsymbol{w}_0 \boldsymbol{z}_{i,1} + b_0) \\ \boldsymbol{z}_{i,1} &= \sigma_1(\boldsymbol{W}_1 \boldsymbol{z}_{i,2} + \boldsymbol{b}_1) \\ \boldsymbol{z}_{i,2} &= \boldsymbol{x}_i \end{aligned}$$

See Figure 4.1 for an illustration.

Throughout this paper, we focus on the case $p_0 = 1$, that is, $y_i \in \mathbb{R}$ or $\{0, 1\}$ for the sake of conciseness and clarity; however, our methodology can easily be extended to the multivariate case.

Let $\boldsymbol{z}_{i,\ell} \in \mathbb{R}^{p_\ell}$ be the output of the ℓ^{th} layer corresponding to the i^{th} element of the training data set, with $z_{i,0} = y_i \in \mathbb{R}$ and $\boldsymbol{z}_{i,L+1} = \boldsymbol{x}_i \in \mathbb{R}^p$. As in Example 4.2.1, we can rewrite each element in $\mathfrak{F}_{\boldsymbol{\theta}}$ in terms of $\boldsymbol{z}_{i,\ell}$, $\ell = 0, \dots, L+1$, in the following way:

$$\begin{aligned} y_i &= \sigma_0(\boldsymbol{w}_0 \boldsymbol{z}_{i,1} + b_0), \\ \boldsymbol{z}_{i,\ell} &= \sigma_\ell(\boldsymbol{W}_\ell \boldsymbol{z}_{i,\ell+1} + \boldsymbol{b}_\ell), \quad \ell = 1, \dots, L, \\ \boldsymbol{z}_{i,L+1} &= \boldsymbol{x}_i. \end{aligned} \tag{4.2.6}$$

The FFNN illustrated in (4.2.6) represents a *deterministic* approach to modeling. In this work, we harness the power of BNNs, a class of neural networks that are trained through the principles of Bayesian inference (MacKay, 1992), and propose a *stochastic* modeling approach. Unlike conventional FFNNs, BNNs introduce stochastic activations or weights, enabling exploration over a distribution of models characterized by $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and governed by a prior density $\mathfrak{p}(\boldsymbol{\theta})$. This structure can be viewed as a Bayesian ensemble, where predictive performance improves through integration over model uncertainty. We adopt the following BNN framework:

$$\begin{aligned} y_i &\sim \mathfrak{p}\left(y_i \mid \boldsymbol{z}_{i,1}\right), \\ \boldsymbol{z}_{i,\ell} &\sim \mathfrak{p}_\ell\left(\boldsymbol{z}_{i,\ell} \mid \boldsymbol{W}_\ell \boldsymbol{z}_{i,\ell+1} + \boldsymbol{b}_\ell\right), \quad \ell = L, \dots, 1, \\ \boldsymbol{z}_{i,L+1} &= \boldsymbol{x}_i, \\ \boldsymbol{W}_\ell &\sim \mathfrak{p}(\boldsymbol{W}_\ell), \quad \boldsymbol{b}_\ell \sim \mathfrak{p}(\boldsymbol{b}_\ell), \quad \ell = 1, \dots, L. \end{aligned}$$

Here, $\boldsymbol{W}_\ell \sim \mathfrak{p}(\boldsymbol{W}_\ell)$ and $\boldsymbol{b}_\ell \sim \mathfrak{p}(\boldsymbol{b}_\ell)$ are assigned prior distributions, while $\mathfrak{p}_\ell(\cdot)$ denotes the probabilistic activation at layer ℓ. The specific forms of these components will be discussed in Section 4.2.1.

In addition, BNNs allow for uncertainty quantification, which is crucial for robust decision making. In the BNN framework, we treat all model parameters $\boldsymbol{\theta}$, including weights, biases, and parameters of the probabilistic activations $\mathfrak{p}_\ell(\cdot)$, as random variables with prior density $\mathfrak{p}(\boldsymbol{\theta})$ over $\boldsymbol{\Theta}$. Given a dataset $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, the posterior distribution is

$$\mathfrak{p}(\boldsymbol{\theta} \mid \mathcal{D}_n) \propto \mathfrak{p}(\mathcal{D}_n \mid \boldsymbol{\theta})\, \mathfrak{p}(\boldsymbol{\theta}),$$

and the posterior predictive distribution for a new input $\boldsymbol{x}_i^{\text{new}}$ is

$$\mathfrak{p}(y_i^{\text{new}} \mid \boldsymbol{x}_i^{\text{new}}, \mathcal{D}_n) = \int_{\boldsymbol{\Theta}} \mathfrak{p}(y_i^{\text{new}} \mid \boldsymbol{x}_i^{\text{new}}, \boldsymbol{\theta})\, \mathfrak{p}(\boldsymbol{\theta} \mid \mathcal{D}_n)\, \mathrm{d}\boldsymbol{\theta}.$$

4.2.1 Hidden layers

In the following, we illustrate how the identity map and probabilistic activation functions can be integrated into the hidden layer architecture, emphasizing their distinct roles and influence on model behavior.

4.2.1.1 Gaussian layer

Let $\boldsymbol{W}_\ell^{(j\cdot)} \in \mathbb{R}^{p_{\ell+1}}$ be a column vector corresponding to the jth row of $\boldsymbol{W}_\ell$ and $b_\ell^{(j)}$ the jth element of $\boldsymbol{b}_\ell$, for $\ell = 0, \dots, L$, and consider the following structure:

$$\begin{aligned}
&\boldsymbol{W}_\ell^{(j\cdot)} \sim \mathsf{Normal}\Big(\boldsymbol{\mu}_{\boldsymbol{W}_\ell^{(j\cdot)}}, \boldsymbol{\Sigma}_{\boldsymbol{W}_\ell^{(j\cdot)}}\Big), \quad \boldsymbol{b}_\ell \sim \mathsf{Normal}(\boldsymbol{\mu}_{\boldsymbol{b}_\ell}, \boldsymbol{\Sigma}_{\boldsymbol{b}_\ell}), \\
&a_{i,\ell}^{(j)} = \big(\boldsymbol{W}_\ell^{(j\cdot)}\big)^\top \boldsymbol{z}_{i,\ell+1} + b_\ell^{(j)}, \quad \boldsymbol{z}_{i,\ell} = \boldsymbol{a}_{i,\ell}, \quad j = 1, \dots, p_\ell,
\end{aligned} \tag{4.2.7}$$

where $a_{i,\ell}^{(j)}$ is the j^{th} component of the p_ℓ-dimensional vector $\boldsymbol{a}_{i,\ell} = (a_{i,\ell}^{(1)}, \dots, a_{i,\ell}^{(p_\ell)})^\top$. In this situation, the activation function is the identity map between $\boldsymbol{z}_{i,\ell}$ and $\boldsymbol{a}_{i,\ell}$ and yields a linear Gaussian layer. In fact, under these prior distributional assumptions on $\boldsymbol{W}_\ell$ and $\boldsymbol{b}_\ell$, the distribution of the j^{th} element of $\boldsymbol{z}_{i,\ell}$ conditionally on $\boldsymbol{z}_{i,\ell+1}$ is

$$z_{i,\ell}^{(j)} \mid \boldsymbol{z}_{i,\ell+1} \sim \mathsf{Normal}\Big(\big(\boldsymbol{\mu}_{\boldsymbol{W}_\ell^{(j\cdot)}}\big)^\top \boldsymbol{z}_{i,\ell+1} + \mu_{\boldsymbol{b}_\ell}^{(j)},\, \boldsymbol{z}_{i,\ell+1}^\top \boldsymbol{\Sigma}_{\boldsymbol{W}_\ell^{(j\cdot)}} \boldsymbol{z}_{i,\ell+1} + \Sigma_{\boldsymbol{b}_\ell}^{(jj)}\Big), \tag{4.2.8}$$

where $\mu_{\boldsymbol{b}_\ell}^{(j)}$ is the j^{th} component of $\boldsymbol{\mu}_\ell$, and $\Sigma_{\boldsymbol{b}_\ell}^{jj}$ refers to the (j,j)-th entry of the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{b}_\ell}$. A Gaussian layer can be viewed as a Bayesian linear model. When $p_\ell < p_{\ell+1}$, it performs a probabilistic dimension reduction, mapping the higher-dimensional layer $z_{i,\ell+1}$ to a lower-dimensional stochastic projection $z_{i,\ell}$. When $p_\ell > p_{\ell+1}$, under the standard Gaussian prior in (4.2.7), the conditional distribution in (4.2.8) takes the likelihood form of probabilistic PCA (Tipping and Bishop, 1999), with $\boldsymbol{M}_{\boldsymbol{W}_\ell} := (\boldsymbol{\mu}_{\boldsymbol{W}_{\ell,1\cdot}}, \dots, \boldsymbol{\mu}_{\boldsymbol{W}_{\ell,p_\ell\cdot}})$ spanning the principal subspace in the probabilistic PCA sense.

4.2.1.2 Probabilistic activation functions

The conditional distribution $\boldsymbol{z}_{i,\ell} \mid \boldsymbol{a}_{i,\ell} \sim \mathfrak{p}(\sigma_\ell(\boldsymbol{a}_{i,\ell}))$, where $\sigma_\ell(\cdot)$ is a non-linear transformation applied element-wise to the random vector $\boldsymbol{a}_{i,\ell}$, is generally intractable. To overcome this intractability issue, we propose the incorporation of activation functions by means of data augmentation strategies (e.g., Wang et al., 2022) and name these activation functions *probabilistic activation functions*. Throughout this paper, we focus on the ReLU activation function for its widespread use (e.g., LeCun et al., 2015), empirical success (e.g., Krizhevsky et al., 2012), and theoretical support (e.g., Schmidt-Hieber, 2020). However, the concepts discussed here can be similarly applied to other probabilistic activation functions (see, e.g., Wang et al., 2022, Table 1).

For every $i = 1, \dots, n$ and $\ell = 1, \dots, L$, let $\boldsymbol{z}_{i,\ell} = (z_{i,\ell}^{(1)}, \dots, z_{i,\ell}^{(p_\ell)})^\top$ be auxiliary (or augmented) variables in the BNN. As in (4.2.7), let $a_{i,\ell}^{(j)} = \left(\boldsymbol{W}_\ell^{(j\cdot)}\right)^\top \boldsymbol{z}_{i,\ell+1} + b_\ell^{(j)}$, but now assume

$$\mathfrak{p}(a_{i,\ell}^{(j)}) \propto \exp\left\{-2\max(a_{i,\ell}^{(j)}, 0)\right\},$$

where $\sigma_\ell(a_{i,\ell}^{(j)}) = \max(a_{i,\ell}^{(j)}, 0)$ is the ReLU function. From expression (4.B.2) in the supplementary material, and assuming that $a_{i,\ell}^{(j)} \mid z_{i,\ell}^{(j)} \sim \mathsf{Normal}(-z_{i,\ell}^{(j)}, z_{i,\ell}^{(j)})$ and $z_{i,\ell}^{(j)} \sim \mathsf{GIG}(1,0,0)$, with GIG denoting a Generalized Inverse Gaussian distribution, the right-hand side of the expression above can be rewritten as

$$\begin{aligned}\exp\left\{-2\max(a_{i,\ell}^{(j)}, 0)\right\} &= \int_0^\infty \mathfrak{p}(a_{i,\ell}^{(j)} \mid z_{i,\ell}^{(j)})\mathfrak{p}(z_{i,\ell}^{(j)})\, \mathrm{d}z_{i,\ell}^{(j)} \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi z_{i,\ell}^{(j)}}} \exp\left\{-\frac{1}{2z_{i,\ell}^{(j)}}\left(a_{i,\ell}^{(j)} + z_{i,\ell}^{(j)}\right)^2\right\} \mathrm{d}z_{i,\ell}^{(j)}.\end{aligned}$$

Here, a continuous random variable $x > 0$ has a Generalized Inverse Gaussian distribution with parameters $p \in \mathbb{R}, a > 0$ and $b > 0$, written $x \overset{\text{d.}}{\sim} \mathsf{GIG}(p, a, b)$, if the density function of x is $\mathfrak{p}(x \mid p, a, b) \propto x^{p-1}\exp\{-(b/x + ax)/2\}$. See supplementary material 4.A.1 for more details and results involving the Generalized Inverse Gaussian distribution.

An important advantage of introducing an augmented variable lies in the ability to model complex marginal distributions. Even when both the prior $\mathfrak{p}(z_{i,\ell}^{(j)})$ and the conditional distribution $\mathfrak{p}(a_{i,\ell}^{(j)} \mid z_{i,\ell}^{(j)})$ are relatively simple, the induced marginal $\mathfrak{p}(a_{i,\ell}^{(j)})$ can be highly flexible. Crucially, the full conditional distribution remains analytically tractable and available in closed form, enabling efficient inference. From this formulation, it follows that the full conditional distribution of the augmented variable is given by

$$z_{i,\ell}^{(j)} \mid a_{i,\ell}^{(j)} \sim \mathsf{GIG}\left(\frac{1}{2}, 1, (a_{i,\ell}^{(j)})^2\right). \tag{4.2.9}$$

To summarize, in a FFNN with probabilistic ReLU activation functions, each neuron first computes a linear combination $a_{i,\ell}^{(j)}$, analogous to deterministic neural networks. The neuron's output, $z_{i,\ell}^{(j)}$, is then modeled as an augmented random variable whose conditional distribution is given by (4.2.9). This probabilistic formulation facilitates tractable inference via data augmentation methods (Polson et al., 2013; Polson and Scott, 2013; Wang et al., 2022).

4.2.1.3 Top layers

The proposed Bayesian neural network (BNN) framework can naturally accommodate a broad class of response types by incorporating a generalized linear model (GLM) output layer. In this case, the network is reformulated as:

$$
\begin{gathered}
g\{\mathsf{E}(y_i \mid z_{i,1})\} = \boldsymbol{W}_0 z_{i,1} + b_0, \quad i = 1, \dots, n, \\
\mathfrak{p}(z_{i,\ell} \mid \boldsymbol{a}_{i,\ell}) \propto \mathfrak{p}(z_{i,\ell}, \boldsymbol{a}_{i,\ell}), \quad \boldsymbol{a}_{i,\ell} = \boldsymbol{W}_\ell z_{i,\ell+1} + \boldsymbol{b}_{i,\ell}, \quad \ell = 1, \dots, L, \\
z_{i,L+1} = \boldsymbol{x}_i,
\end{gathered}
$$

where $g(\cdot)$ is a monotonic, invertible link function selected according to the type of response variable y_i. This formulation aligns with classical GLMs (see, e.g. Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989), where the conditional mean of the response is linked to a linear predictor via g. For example, when the response variable y_i is binary, taking values in $\{0, 1\}$, the appropriate choice is either the logit link function, $g(\mu) = \log\{\mu/(1-\mu)\}$, which corresponds to logistic regression, or the probit link $g(\mu) = \Phi^{-1}(\mu)$, where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. Both models are widely used for binary classification and differ primarily in their underlying assumptions: logistic regression assumes a logistic distribution for the latent variable, whereas probit regression assumes a standard normal distribution. For count data, where $y_i \in \mathbb{N}$, a commonly used link function is the logarithmic link, $g(\mu) = \log(\mu)$, yielding the Poisson regression model. This choice ensures that the mean response μ remains strictly positive and aligns with the natural exponential family form of the Poisson distribution. Conversely, for continuous real-valued responses $y_i \in \mathbb{R}$, the identity link function $g(\mu) = \mu$ is adopted, resulting in a Gaussian regression framework. This model assumes that the conditional distribution of the response is normally distributed with mean μ and constant variance.

Such model specifications are firmly rooted in the theory of generalized linear models (GLMs), originally introduced by Nelder and Wedderburn (1972), which unify a broad class of regression models through the use of link functions and probability distributions from the exponential family. By incorporating these link functions into Bayesian neural networks, the resulting models retain interpretability while gaining the flexibility to accommodate a diverse range of response types, thereby enabling efficient and principled inference. Moreover, for each of the aforementioned response types, valid data augmentation strategies have been developed to facilitate posterior computation. For binary responses with a probit link, Albert and Chib (1993) proposed a latent Gaussian formulation. In the context of logit models, Polson et al. (2013) introduced an efficient augmentation approach based on Pólya-Gamma variables, while Durante (2019) and Onorati and Liseo (2025) proposed a related scheme relying on the Skew-Normal distribution for probit and logit links, respectively. These methodologies have been recently extended by Anceschi et al. (2023) to accommodate more complex settings, including tobit and multinomial probit models. For count data, Bradley et al. (2018) developed a tractable data augmentation technique for Poisson regression. More broadly, Castiglione and Bernardi (2025) proposed general extensions of generalized linear and mixed models to settings involving intractable likelihoods, further broadening the scope of models amenable to efficient Bayesian inference.

This flexible modeling framework extends the expressive power of GLMs (see McCullagh and Nelder, 1989) to deep learning architectures, enabling robust modeling across diverse data types including binary, count, and continuous outcomes. Moreover, incorporating probabilistic activation functions through data augmentation not only enriches the modeling capacity but also facilitates efficient posterior inference and scalable training algorithms. This is further explored in the next section.

4.3 Training, inference and prediction

Here we present details of the proposed methodology for training our BNNs with probabilistic activation functions and making inferences about the parameters $\boldsymbol{\theta}$. We take advantage of two powerful techniques, semiparametric MFVB and MoE, which are detailed below. The first technique is introduced to resolve the intractability issues arising from the application of generic MFVB, and the latter is used to enhance the predictive performance of BNNs. Section 4.3.2 provides details on the computation of the posterior predictive distribution that is then used in supervised learning tasks.

4.3.1 Semiparametric mean-field variational Bayes

Semiparametric MFVB (Rohde and Wand, 2016; Opper and Archambeau, 2009) is an approximate inference method from the class of variational approximations. As in Section 4.2, let $\boldsymbol{\theta}$ represent all model parameters and $\mathcal{D}_n$ be the training data set. The logarithm of the marginal likelihood satisfies

$$\begin{aligned}\log \mathfrak{p}(\mathcal{D}_n) &= \log \mathfrak{p}(\mathcal{D}_n) \int \mathfrak{q}(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} = \int \mathfrak{q}(\boldsymbol{\theta}) \log \mathfrak{p}(\mathcal{D}_n)\, \mathrm{d}\boldsymbol{\theta} \\ &= \int \mathfrak{q}(\boldsymbol{\theta}) \log \left\{ \frac{\mathfrak{p}(\mathcal{D}_n, \boldsymbol{\theta})/\mathfrak{q}(\boldsymbol{\theta})}{\mathfrak{p}(\boldsymbol{\theta} \mid \mathcal{D}_n)/\mathfrak{q}(\boldsymbol{\theta})} \right\} \mathrm{d}\boldsymbol{\theta} \\ &= \mathsf{ELBO}_{\mathfrak{q}}(\mathcal{D}_n; \boldsymbol{\theta}) + \mathsf{KL}\{\mathfrak{q}(\boldsymbol{\theta}) \| \mathfrak{p}(\boldsymbol{\theta} \mid \mathcal{D}_n)\},\end{aligned} \tag{4.3.1}$$

where

$$\mathsf{ELBO}_{\mathfrak{q}}(\mathcal{D}_n; \boldsymbol{\theta}) \equiv \int \mathfrak{q}(\boldsymbol{\theta}) \log \left\{ \frac{\mathfrak{p}(\mathcal{D}_n, \boldsymbol{\theta})}{\mathfrak{q}(\boldsymbol{\theta})} \right\} \mathrm{d}\boldsymbol{\theta}$$

is the so-called *evidence lower bound*, which depends on an arbitrary density function $\mathfrak{q}(\boldsymbol{\theta})$, and

$$\mathsf{KL}\{\mathfrak{q}(\boldsymbol{\theta}) \| \mathfrak{p}(\boldsymbol{\theta} \mid \mathcal{D}_n)\} \equiv \int \mathfrak{q}(\boldsymbol{\theta}) \log \left\{ \frac{\mathfrak{q}(\boldsymbol{\theta})}{\mathfrak{p}(\boldsymbol{\theta} \mid \mathcal{D}_n)} \right\} \mathrm{d}\boldsymbol{\theta}$$

is the Kullback-Leibler divergence between $\mathfrak{q}(\boldsymbol{\theta})$ and $\mathfrak{p}(\boldsymbol{\theta} \mid \mathcal{D}_n)$, which is strictly non-negative.

Let $\boldsymbol{z}$ denote all the model augmented variables that are present in the model and consider steps analogous to those in (4.3.1) but for the pair $(\boldsymbol{\theta}, \boldsymbol{z})$ and $\mathfrak{q}(\boldsymbol{\theta})$ replaced by $\mathfrak{q}(\boldsymbol{\theta}, \boldsymbol{z})$. Suppose that $\mathfrak{q}(\boldsymbol{\theta}, \boldsymbol{z})$ is selected from a family of densities $\mathcal{Q}$. In variational Bayes, the optimal approximating density is the minimizer of the Kullback-Leibler divergence between the variational approximating density itself and the true posterior density, that is

$$\mathfrak{q}^{\star}(\boldsymbol{\theta}, \boldsymbol{z}) = \arg\min_{\mathfrak{q} \in \mathcal{Q}} \mathsf{KL}\{\mathfrak{q}(\boldsymbol{\theta}, \boldsymbol{z}) \| \mathfrak{p}(\boldsymbol{\theta}, \boldsymbol{z} \mid \mathcal{D}_n)\}, \tag{4.3.2}$$

From (4.3.1) it is clear that minimizing the Kullback-Leibler divergence (4.3.2) is equivalent to maximizing the corresponding evidence lower bound

$$\mathsf{ELBO}_{\mathfrak{q}}(\mathcal{D}_n; \boldsymbol{\theta}, \boldsymbol{z}) = \int \mathfrak{q}(\boldsymbol{\theta}, \boldsymbol{z}) \log \left\{ \frac{\mathfrak{p}(\mathcal{D}_n, \boldsymbol{\theta}, \boldsymbol{z})}{\mathfrak{q}(\boldsymbol{\theta}, \boldsymbol{z})} \right\} \mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{z}.$$

Imposing a restriction on $\mathcal{Q}$ can facilitate the optimization task. In the case of MFVB, $\mathcal{Q}$ is factorized as a product of approximating densities such as

$$\mathcal{Q} = \{\mathfrak{q}(\boldsymbol{\theta}, \boldsymbol{z}) : \mathfrak{q}(\boldsymbol{\theta}, \boldsymbol{z}) = \mathfrak{q}(\boldsymbol{\theta}_1) \cdots \mathfrak{q}(\boldsymbol{\theta}_M)\mathfrak{q}(\boldsymbol{z})\} \text{ for some partition } \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\} \text{ of } \boldsymbol{\theta}.$$

Under this restriction, the optimal variational approximation can be obtained via a coordinate ascent variational inference algorithm (see, e.g., Bishop, 2006; Ormerod and Wand, 2010), and

$$\begin{aligned} \mathfrak{q}^\star(\boldsymbol{\theta}_j) &\propto \exp\left[\mathsf{E}_{-\boldsymbol{\theta}_j}\{\log \mathfrak{p}(\boldsymbol{\theta}_j \mid \text{rest})\}\right], \quad j = 1, \ldots, M, \\ \mathfrak{q}^\star(\boldsymbol{z}) &\propto \exp\left[\mathsf{E}_{\boldsymbol{\theta}}\{\log \mathfrak{p}(\boldsymbol{z} \mid \text{rest})\}\right], \end{aligned} \tag{4.3.3}$$

where $\mathsf{E}_{-\boldsymbol{\theta}_j}(\cdot)$ and $\mathsf{E}_{\boldsymbol{\theta}}(\cdot)$ respectively symbolize expectations with respect to $\prod_{k \neq j} \mathfrak{q}(\boldsymbol{\theta}_k)\mathfrak{q}(\boldsymbol{z})$ and $\prod_{j=1}^{M} \mathfrak{q}(\boldsymbol{\theta}_j)$. For models that exhibit specific conjugacy properties, the optimal variational densities $\mathfrak{q}^\star(\boldsymbol{\theta}_j)$ and $\mathfrak{q}^\star(\boldsymbol{z})$ can be available in closed form. When these approximating densities do not belong to any conventional family of distributions, other solutions can be applied to avoid intractability and numerical integration. For such situations, here we consider semiparametric MFVB, which is a relaxation of ordinary mean field variational Bayes based on assuming that some density functions in the product density restriction are from pre-specified parametric families of densities that can be conveniently chosen to achieve tractability. We adopt the nomenclature *semiparametric MFVB* introduced by Rohde and Wand (2016) acknowledging that terms such as, for instance, fixed-form variational Bayes (Honkela et al., 2010) and non-conjugate variational message passing (Knowles and Minka, 2011) have been used in the message passing literature as alternatives.

In semiparametric MFVB, the restriction on $\mathcal{Q}$ becomes

$$\mathcal{Q} = \{\mathfrak{q}(\boldsymbol{\theta}, \boldsymbol{z}) : \mathfrak{q}(\boldsymbol{\theta}, \boldsymbol{z}) = \mathfrak{q}(\boldsymbol{\theta}_1) \cdots \mathfrak{q}(\boldsymbol{\theta}_M)\tilde{\mathfrak{q}}(\boldsymbol{z}; \boldsymbol{\xi}),\ \boldsymbol{\xi} \in \Xi\}, \tag{4.3.4}$$

where $\tilde{\mathfrak{q}}(\boldsymbol{z}; \boldsymbol{\xi})$ is a pre-specified parametric family of density functions in $\boldsymbol{z}$ depending on some parameters $\boldsymbol{\xi} \in \Xi$. Under restriction (4.3.4), the optimal variational approximations are

$$\mathfrak{q}^\star(\boldsymbol{\theta}_j) \propto \exp\left[\mathsf{E}_{-\boldsymbol{\theta}_j}\{\log \mathfrak{p}(\boldsymbol{\theta}_j \mid \text{rest})\}\right], \quad j = 1, \ldots, M, \tag{4.3.5}$$

and

$$\tilde{\mathfrak{q}}^\star(\boldsymbol{z}; \boldsymbol{\xi}) = \arg \max_{\mathfrak{q} \in \{\tilde{\mathfrak{q}}(\boldsymbol{z}; \boldsymbol{\xi}) : \boldsymbol{\xi} \in \Xi\}} \mathsf{ELBO}_{\mathfrak{q}}(\mathcal{D}_n; \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{\xi}), \tag{4.3.6}$$

where $\mathsf{E}_{-\boldsymbol{\theta}_j}(\cdot)$ denotes expectation with respect to $\prod_{h \neq j} \mathfrak{q}(\boldsymbol{\theta}_h)\tilde{\mathfrak{q}}(\boldsymbol{z}; \boldsymbol{\xi})$, and $\mathsf{ELBO}_{\mathfrak{q}}(\mathcal{D}_n; \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{\xi})$ is an evidence lower bound similar to those presented above but depending on $\tilde{\mathfrak{q}}(\boldsymbol{z}; \boldsymbol{\xi})$. A generic procedure for obtaining the optimal variational densities is listed as in Algorithm 2 in Rohde and Wand (2016).

4.3.2 Mixture of experts

Mixture of expert (MoE) models are celebrated for their high computational efficiency, making them a prime choice for large-scale applications (Jacobs et al., 1991; Bishop and Svenskn, 2003; Jordan and Jacobs, 1994; Waterhouse et al., 1995). Their ability to enhance predictive accuracy is also particularly noteworthy, as they are able to model intricate input-output relationships by adaptively partitioning the input space. In our framework, MoE models crucially support both the variational inference process, especially the application of variational approximations to augmented variables, and the prediction procedure (see Section 4.3.4). This dual functionality underscores their crucial role in our proposed methodology.

Consider the predictive posterior distribution arising in a supervised learning task involving BNNs. Given a new input $\boldsymbol{x}_i^{\text{new}}$ and observed data $\mathcal{D}_n = (\boldsymbol{x}_i, y_i)_{i=1}^n$, the predictive distribution of the corresponding response y_i^{new} is given by

$$\mathfrak{p}(y_i^{\text{new}} \mid \boldsymbol{x}_i^{\text{new}}, \mathcal{D}_n) = \int \mathfrak{p}(y_i^{\text{new}} \mid \boldsymbol{z})\mathfrak{p}(\boldsymbol{z} \mid \boldsymbol{x}_i^{\text{new}}, \boldsymbol{\theta}, \mathcal{D}_n)\mathfrak{p}(\boldsymbol{\theta} \mid \mathcal{D}_n)\mathfrak{p}(\boldsymbol{\theta})\mathfrak{p}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\boldsymbol{\theta}. \tag{4.3.7}$$

When the methodology described in Subsection 4.3.1 is applied, the conditional densities of $\boldsymbol{\theta}$ and $\boldsymbol{z}$ on the right-hand side of (4.3.7) can be replaced by the corresponding optimal variational Bayes densities of $\boldsymbol{\theta}_j$ and $\boldsymbol{z}$ given by expressions (4.3.5) and (4.3.6), respectively. However, when dealing with a new test data point with input $\boldsymbol{x}_i^{\text{new}}$, the issue is finding a unique (parametric) variational approximation to the marginal posterior distribution of $\boldsymbol{z}$, that is $\mathfrak{p}(\boldsymbol{z} \mid \boldsymbol{x}_i^{\text{new}}, \boldsymbol{\theta}, \mathcal{D}_n)$. In fact, traditional variational inference methods are typically "local" in the sense that they focus on optimizing variational parameters that have a one-to-one correspondence to each data point. Therefore, when presented with a new input $\boldsymbol{x}_i^{\text{new}}$ for prediction, it is unclear which variational parameters should be used to obtain a variational approximation to $\mathfrak{p}(\boldsymbol{z} \mid \boldsymbol{x}_i^{\text{new}}, \boldsymbol{\theta}, \mathcal{D}_n)$ in (4.3.7). Variational auto-encoders (VAEs; e.g., Kingma, 2013; Rezende et al., 2014; Kingma et al., 2019), for instance, offer a "global" solution to this problem based on a single encoder neural network for performing posterior inference on all training data; this approach is known as amortized variational inference (e.g., Gershman and Goodman, 2014). A drawback of VAEs is that they may be slow to fit, particularly when navigating large and high-dimensional datasets (Cremer et al., 2018). Furthermore, their instability may be a concern, as each parameter is intricately linked to all the data and slight modifications to the data can substantially impact the model's overall fit (Razavi et al., 2019).

Rather than adopting a solution depending on a single global model or a large number of narrowly focused local models, we propose using MoEs (Jacobs et al., 1991), which leverage multiple local models (referred to as *experts*) to capture the relationship between inputs and outputs within subpopulations that exhibit similar functional behaviors. This approach emphasizes modeling local response patterns rather than merely clustering input covariates based on similarity. Here we employ the hard gating mechanism of MoEs introduced by Jacobs et al. (1991), wherein the gating network deterministically/stochastically selects a single expert model for each input, rather than distributing responsibility across multiple experts. Each resulting localized expert is approximated locally by its own variational density. The gating network itself can take various forms, including but not limited to decision tree classifiers, k-means clustering, Gaussian mixture models, or neural networks, offering flexibility in how the expert selection mechanism is modeled. When the gating network is implemented using a supervised learning approach, its parameters can be updated iteratively. In contrast, if the gating network is based on unsupervised methods, the gating network held fixed throughout the procedure.

In the following, we provide two examples of gating networks used to assign inputs to experts.

Example 4.3.1 (K-means clustering as a gating network)**.** *Let $z_i \in \mathbb{R}^p$ denote the input to the gating network. Given K centroids $\{\mu_k\}_{k=1}^K \subset \mathbb{R}^p$, obtained by applying standard K-means clustering to $\{z_i\}_{i=1}^n$, the gating function $T : \mathbb{R}^p \to \{1, \ldots, K\}$ is defined by*

$$T(z_i) = \arg \min_{k \in \{1,\ldots,K\}} \|z_i - \mu_k\|_2^2.$$

Each centroid defines a region in $\mathbb{R}^p$ associated with a unique expert, and $T(z_i)$ selects the expert corresponding to the closest centroid. The centroids are held fixed during training.

Unlike the purely unsupervised approach in the previous example, the following strategy incorporates label information by clustering in the joint input–output space. This allows the gating structure to better reflect the underlying input–output mapping, often leading to more meaningful expert specialization.

Example 4.3.2 (Decision tree classifier trained via K-means initialization)**.** *Let $z_i \in \mathbb{R}^p$ denote the input to the gating network. To better capture the input–output relationship, K-means clustering is first applied to the joint space $\{(z_i, y_i)\}_{i=1}^n$ to obtain soft expert assignment probabilities $\pi_{ik} \in [0,1]$, where $\sum_{k=1}^K \pi_{ik} = 1$ for each i.*

A decision tree classifier $\mathcal{G} : \mathbb{R}^p \to [0,1]^K$ is then trained to predict the assignment vector $\boldsymbol{\pi}_i := (\pi_{i1}, \ldots, \pi_{iK})$ from z_i. Let $\widehat{\pi}_{ik} := \mathcal{G}_k(z_i)$ denote the predicted probability for expert k, satisfying $\sum_{k=1}^K \widehat{\pi}_{ik} = 1$.

The hard gating function $T : \mathbb{R}^p \to \{1,\dots,K\}$ is then defined by

$$T(\boldsymbol{z}_i) = \arg \max_{k \in \{1,\dots,K\}} \widehat{\pi}_{ik}.$$

The classifier $\mathcal{G}$ is updated iteratively during training; see Algorithm 1 as an example.

The next subsection presents the proposed procedure for incorporating MoEs within BNNs.

4.3.3 Training via mixture of experts

Let $\boldsymbol{T}$ be the hard gating network in a MoEs. Consider a BNN as depicted in Figure 4.2, where we have a number n_{global} of global hidden layers and a number n_{expert} of hidden layers in each expert, so that $n_{\text{global}} + n_{\text{expert}} = L$. To ensure identifiability under the linear Gaussian specification below, we fix $n_{\text{global}} = 1$ throughout, so that the global layer corresponds to $\ell = L$. While other choices of global probabilistic layers, such as those arising from conjugate exponential family structures, are possible, we restrict attention to the Gaussian case for clarity and tractability.

The overall procedure shown in Figure 4.2 can be heuristically understood as follows: the input first passes through the global layer(s), which serve as a form of preprocessing or dimension reduction. The resulting representation is then routed through the gating network, which selects a single expert. The selected expert processes the input through its own n_{expert} hidden layers to produce the final output. The corresponding inference procedure is summarized below.

(i) **Variational inference for the BNN's global hidden layer.** In the global hidden layer $\ell = L$, we assume

$$\boldsymbol{z}_{i,\ell} \mid \boldsymbol{a}_{i,\ell}, \sigma_\ell^2 \sim \mathsf{Normal}(\boldsymbol{a}_{i,\ell}, \sigma_\ell^2 \boldsymbol{I}_{p_\ell}), \quad \boldsymbol{a}_{i,\ell} = \boldsymbol{W}_\ell \boldsymbol{z}_{i,\ell+1} + \boldsymbol{b}_\ell,$$

$$\mathfrak{p}(\boldsymbol{W}_\ell) \propto 1, \quad \mathfrak{p}(\boldsymbol{b}_\ell) \propto 1, \quad \mathfrak{p}(\sigma_\ell^2) \propto (\sigma_\ell^2)^{-1/2},$$

where $\boldsymbol{W}_\ell \in \mathbb{R}^{p_\ell \times p_{\ell+1}}$, $\boldsymbol{b}_\ell \in \mathbb{R}^{p_\ell}$, $\boldsymbol{z}_{i,\ell} \in \mathbb{R}^{p_\ell}$, $\boldsymbol{z}_{i,L+1} = \boldsymbol{x}_i \in \mathbb{R}^p$, and $\boldsymbol{I}_{p_\ell}$ is a p_ℓ-dimensional identity matrix. Denote $\boldsymbol{Z}_\ell = (\boldsymbol{z}_{1,\ell}, \dots, \boldsymbol{z}_{n,\ell})^\top \in \mathbb{R}^{n \times p_\ell}$. At each iteration of our algorithm, the *global parameters* in each layer $\boldsymbol{\theta}_{\text{global}} = \{\boldsymbol{Z}_\ell, \boldsymbol{W}_\ell, \boldsymbol{b}_\ell, \sigma_\ell^2\}$ are updated via MFVB. Let $\boldsymbol{W}_{\ell,j\cdot} \in \mathbb{R}^{1 \times p_{\ell+1}}$ as the j-th row of the matrix $\boldsymbol{W}_\ell$ and $\dot{\boldsymbol{W}}_{\ell,j\cdot} = (\boldsymbol{b}_{\ell,j}, \boldsymbol{W}_{\ell,j\cdot})^\top \in \mathbb{R}^{p_{\ell+1}+1}$, $\dot{\boldsymbol{W}}_\ell = (\boldsymbol{b}_\ell, \boldsymbol{W}_\ell)^\top \in \mathbb{R}^{(p_{\ell+1}+1) \times p_\ell}$, and $\boldsymbol{Z}_{\ell,\cdot j} \in \mathbb{R}^n$ is the j-th column of the matrix $\boldsymbol{Z}_\ell$, $\dot{\boldsymbol{Z}}_\ell = (\boldsymbol{1}_n, \boldsymbol{Z}_\ell) \in \mathbb{R}^{n \times (p_\ell+1)}$, then from application of (4.3.3) it follows that

$$\begin{aligned}
\mathfrak{q}^{\star}(z_{i,\ell}) &\propto \exp\left[\mathsf{E}_{-z_{i,\ell}}\left\{\log \mathfrak{p}(z_{i,\ell} \mid \text{rest})\right\}\right], \quad i = 1, \dots, n, \\
\mathfrak{q}^{\star}(\dot{\boldsymbol{W}}_{\ell,j\cdot}) &\propto \exp\left[\mathsf{E}_{-\dot{\boldsymbol{W}}_{\ell,j\cdot}}\left\{\log \mathfrak{p}(\dot{\boldsymbol{W}}_{\ell,j\cdot} \mid \text{rest})\right\}\right], \quad j = 1, \dots, p_\ell, \\
\mathfrak{q}^{\star}(\sigma_\ell^2) &\propto \exp\left[\mathsf{E}_{-\sigma_\ell^2}\left\{\log \mathfrak{p}(\sigma_\ell^2 \mid \text{rest})\right\}\right].
\end{aligned}$$

By direct manipulation, the optimal mean field approximations of augmented variables $z_{i,\ell}$, $i = 1, \dots, n$ are Gaussian:

$$\begin{aligned}
\mathfrak{q}^{\star}(z_{i,\ell}) &\sim \mathsf{Normal}(\boldsymbol{\mu}_{\mathfrak{q}(z_{i,\ell})}, \boldsymbol{\Sigma}_{\mathfrak{q}(z_{i,\ell})}), \\
\boldsymbol{\mu}_{\mathfrak{q}(z_{i,\ell})} &= \boldsymbol{\Sigma}_{\mathfrak{q}(z_{i,\ell})}\Big(\mathsf{E}_{\mathfrak{q}(\sigma_\ell^2)}[\sigma_\ell^{-2}]\boldsymbol{I}_{p_\ell}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{W}_\ell)}\boldsymbol{\mu}_{\mathfrak{q}(z_{i,\ell+1})} + \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{b}_\ell)}\right) \\
&\qquad + \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{W}_{\ell-1})}^{\top}\mathsf{E}_{\mathfrak{q}(\sigma_{\ell-1}^2)}[\sigma_{\ell-1}^{-2}]\boldsymbol{I}_{p_{\ell-1}}(\boldsymbol{\mu}_{\mathfrak{q}(z_{i,\ell-1})} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{b}_{\ell-1})})\Big), \\
\boldsymbol{\Sigma}_{\mathfrak{q}(z_{i,\ell})} &= \left(\mathsf{E}_{\mathfrak{q}(\sigma_\ell^2)}[\sigma_\ell^{-2}]\boldsymbol{I}_{p_\ell} + \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{W}_{\ell-1})}^{\top}\mathsf{E}_{\mathfrak{q}(\sigma_{\ell-1}^2)}[\sigma_{\ell-1}^{-2}]\boldsymbol{I}_{p_{\ell-1}}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{W}_{\ell-1})}\right)^{-1}.
\end{aligned}$$

Let $\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{Z}_\ell)} = \left(\boldsymbol{\mu}_{\mathfrak{q}(z_{1,\ell})}, \dots, \boldsymbol{\mu}_{\mathfrak{q}(z_{n,\ell})}\right)^{\top} \in \mathbb{R}^{n \times p_\ell}$ and $\boldsymbol{M}_{\mathfrak{q}(\dot{\boldsymbol{Z}}_\ell)} = \left(\boldsymbol{1}_n, \boldsymbol{M}_{\mathfrak{q}(\boldsymbol{Z}_\ell)}\right) \in \mathbb{R}^{n \times (p_\ell + 1)}$, the optimal mean field approximations of weights and biases $\dot{\boldsymbol{W}}_{\ell,j\cdot}$, $j = 1, \dots, p_\ell$, are Gaussian as well:

$$\begin{aligned}
\mathfrak{q}^{\star}(\dot{\boldsymbol{W}}_{\ell,j\cdot}) &\sim \mathsf{Normal}(\boldsymbol{\mu}_{\mathfrak{q}(\dot{\boldsymbol{W}}_{\ell,j\cdot})}, \boldsymbol{\Sigma}_{\mathfrak{q}(\dot{\boldsymbol{W}}_{\ell,j\cdot})}), \\
\boldsymbol{\mu}_{\mathfrak{q}(\dot{\boldsymbol{W}}_{\ell,j\cdot})} &= \boldsymbol{\Sigma}_{\mathfrak{q}(\dot{\boldsymbol{W}}_{\ell,j\cdot})}\boldsymbol{M}_{\mathfrak{q}(\dot{\boldsymbol{Z}}_{\ell+1})}^{\top}\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{Z}_{\ell,\cdot j})}, \\
\boldsymbol{\Sigma}_{\mathfrak{q}(\dot{\boldsymbol{W}}_{\ell,j\cdot})} &= \mathsf{E}_{\mathfrak{q}(\sigma_\ell^2)}[\sigma_\ell^2]\left(\boldsymbol{M}_{\mathfrak{q}(\dot{\boldsymbol{Z}}_{\ell+1})}^{\top}\boldsymbol{M}_{\mathfrak{q}(\dot{\boldsymbol{Z}}_{\ell+1})}\right)^{-1}.
\end{aligned}$$

Similarly, let $\boldsymbol{M}_{\mathfrak{q}(\dot{\boldsymbol{W}}_\ell)} = \left(\boldsymbol{\mu}_{\mathfrak{q}(\dot{\boldsymbol{W}}_{\ell,1\cdot})}, \dots, \boldsymbol{\mu}_{\mathfrak{q}(\dot{\boldsymbol{W}}_{\ell,p_\ell\cdot})}\right) \in \mathbb{R}^{(p_{\ell+1}+1) \times p_\ell}$. Then, the optimal mean field approximation of variance σ_ℓ^2 is Inverse-Gamma distribution:

$$\begin{aligned}
\mathfrak{q}^{\star}(\sigma_\ell^2) &\sim \mathsf{Inv\text{-}Gamma}(\alpha_{\mathfrak{q}(\sigma_\ell^2)}, \beta_{\mathfrak{q}(\sigma_\ell^2)}), \\
\alpha_{\mathfrak{q}(\sigma_\ell^2)} &= \frac{np_\ell - 1}{2}, \\
\beta_{\mathfrak{q}(\sigma_\ell^2)} &= \frac{1}{2}\mathsf{Trace}\left(\left(\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{Z}_\ell)} - \boldsymbol{M}_{\mathfrak{q}(\dot{\boldsymbol{Z}}_{\ell+1})}\boldsymbol{M}_{\mathfrak{q}(\dot{\boldsymbol{W}}_\ell)}\right)^{\top}\left(\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{Z}_\ell)} - \boldsymbol{M}_{\mathfrak{q}(\dot{\boldsymbol{Z}}_{\ell+1})}\boldsymbol{M}_{\mathfrak{q}(\dot{\boldsymbol{W}}_\ell)}\right)\right).
\end{aligned}$$

(ii) **Expert choice routing.** Each input $\boldsymbol{x}_i$ for $i = 1, \dots, n$ is propagated forward through

n_{global} global hidden layers, which we recall is fixed to $n_{\text{global}} = 1$ in our setup. This yields the representation $\boldsymbol{z}_{i,L}$ before expert assignment. A linear transformation $\boldsymbol{W}_{L-1}\boldsymbol{z}_{i,L} + \boldsymbol{b}_{L-1} \in \mathbb{R}^{p_{L-1}}$ is then applied and serves as input to the gating network.

Each input is assigned to a single expert via the gating function $\boldsymbol{T} : \mathbb{R}^{p_{L-1}} \rightarrow \{1, \dots, K\}$, which partitions the space $\mathbb{R}^{p_{L-1}}$ into K disjoint regions via axis-aligned splits in a decision tree gating structure.

As a result, the training dataset $\mathcal{D}_n$ is partitioned into K groups, denoted by C_k for $k = 1, \dots, K$. The subset assigned to expert k is denoted $\mathcal{D}_{C_k}$. For the ℓ^{th} expert hidden layer, define $\mathcal{D}^{\ell}_{C_k} = \{(\boldsymbol{z}_{i,\ell+1}, \boldsymbol{z}_{i,\ell-1}) : i \in C_k\}$. The parameters of the gating network are collectively denoted by $\boldsymbol{\eta}$.

(iii) **Variational inference for the BNN's expert hidden layers.** For each expert $k = 1, \dots, K$, its parameters $\boldsymbol{\theta}^{[k]}_{\text{hidden}} = \left\{\boldsymbol{Z}^{[k]}_{\ell}, \dot{\boldsymbol{W}}^{[k]}_{\ell}, (\sigma^2_{\ell})^{[k]}\right\}_{\ell=1}^{\ell=n_{\text{expert}}}$ are updated independently with categorized data $\mathcal{D}^{\ell}_{C_k}$. If a hidden layer is a Gaussian layer as in Section 4.2.1.1, then all the parameters in the hidden layer $\boldsymbol{\theta}^{[k]}_{\text{hidden},\ell}$ will be updated via MFVB. On the other hand, if a probabilistic activation function as in Section 4.2.1.2 is applied in the hidden layer of the expert k, the augmented variables $\boldsymbol{z}^{[k]}_{i,\ell}$ are updated via semiparametric MFVB, the weight and bias $\dot{\boldsymbol{W}}^{[k]}_{\ell}$ is updated via the mean-field variational approach, and $(\sigma^2_{\ell})^{[k]}$ is not required. Let $\boldsymbol{\xi}^{[k]}_{\ell}$ be parameters that characterise the parametric distribution of the k^{th}- expert's parameters in the ℓ^{th} expert hidden layer. In the ℓ^{th} hidden layer, the objective ELBO for the k^{th}-expert is defined as:

$$
\begin{aligned}
\mathsf{ELBO}_{\mathsf{q}}(\mathcal{D}^{\ell}_{C_k}; \dot{\boldsymbol{W}}^{[k]}_{\ell}, \boldsymbol{Z}^{[k]}_{\ell}, \boldsymbol{\xi}^{[k]}_{\ell}) &= \sum_{i \in C_k} \mathsf{ELBO}_{\mathsf{q}}\left((\boldsymbol{z}^{[k]}_{i,\ell+1}, \boldsymbol{z}^{[k]}_{i,\ell-1}); \dot{\boldsymbol{W}}^{[k]}_{\ell}, \boldsymbol{z}^{[k]}_{i,\ell}, \boldsymbol{\xi}^{[k]}_{\ell}\right) \\
&= \sum_{i \in C_k} \mathsf{E}_{\mathsf{q}}\left\{\frac{\mathfrak{p}(\boldsymbol{z}^{[k]}_{i,\ell+1}, \boldsymbol{z}^{[k]}_{i,\ell-1}, \boldsymbol{z}^{[k]}_{i,\ell}, \dot{\boldsymbol{W}}^{[k]}_{\ell})}{\mathsf{q}(\boldsymbol{z}^{[k]}_{i,\ell}; \boldsymbol{\xi}^{[k]}_{\ell})\mathsf{q}(\dot{\boldsymbol{W}}^{[k]}_{\ell})}\right\} \\
&= \sum_{i \in C_k} \mathsf{E}_{\mathsf{q}}[\log\{\mathfrak{p}(\boldsymbol{z}^{[k]}_{i,\ell+1}, \boldsymbol{z}^{[k]}_{i,\ell-1}, \boldsymbol{z}^{[k]}_{i,\ell}, \dot{\boldsymbol{W}}^{[k]}_{\ell})\}] \\
&\quad + \sum_{i \in C_k} \mathsf{Entropy}\{\mathsf{q}(\boldsymbol{z}^{[k]}_{i,\ell}; \boldsymbol{\xi}^{[k]}_{\ell})\} + \mathsf{Entropy}\{\mathsf{q}(\dot{\boldsymbol{W}}^{[k]}_{\ell})\},
\end{aligned}
\tag{4.3.8}
$$

where the $\mathsf{E}_{\mathsf{q}}(\cdot)$ symbolizes the expectation with respect to q; see Rohde and Wand (2016) for a justification of the above decomposition. The optimised parameters $(\boldsymbol{\xi}^{[k]}_{\ell})^{\star}$

of the k^{th} expert is the optimiser of the following maximization problem

$$(\boldsymbol{\xi}_\ell^{[k]})^\star = \arg\max_{\boldsymbol{\xi}_\ell^{[k]} \in \Xi} \mathsf{ELBO}_{\mathfrak{q}}(\mathcal{D}_{C_k}^\ell; \dot{\boldsymbol{W}}_\ell^{[k]}, \boldsymbol{Z}_\ell^{[k]}, \boldsymbol{\xi}_\ell^{[k]}), \tag{4.3.9}$$

which gives the optimal variational density $\tilde{q}^\star(\boldsymbol{z}_\ell^{[k]}; \boldsymbol{\xi}_\ell^{[k]})$. On the other hand, the variational approximations on $\dot{\boldsymbol{W}}_\ell^{[k]}$ are obtained through mean field variational Bayes. That is,

$$\mathfrak{q}^\star(\dot{\boldsymbol{W}}_\ell^{[k]}) \propto \exp\left[\mathsf{E}_{-\dot{\boldsymbol{W}}_\ell^{[k]}}\{\log \mathfrak{p}(\dot{\boldsymbol{W}}_\ell^{[k]} \mid \text{rest})\}\right],$$

where the expectation $\mathsf{E}_{-\dot{\boldsymbol{W}}_\ell^{[k]}}(\cdot)$ is equivalent to the expectation $\mathsf{E}_{\tilde{q}(\boldsymbol{z}_\ell^{[k]}; \boldsymbol{\xi}_\ell^{[k]})}(\cdot)$.

(iv) **Update the BNN's top layer.** As mentioned in Section 4.2.1.2, the top layer can be viewed as a generalized linear regression, which can be adapted according to the types of responses. The updates of the top layer's parameters $\boldsymbol{\phi}^{[k]}$ depend on the specific response types. For example, in the regression task, one can find MFVB approximations on $\boldsymbol{\phi}^{[k]}$, as in Section 4.4. For binary classification, see Section 4.5.2, another data augmentation application enables the MFVB approximations on $\boldsymbol{\phi}^{[k]}$.

(v) **Update the gating network $\boldsymbol{T}$ (if needed).** If the gating network is based on unsupervised methods, an expectation maximization (EM) algorithm can be applied to update the gating network. In the latter, for the i^{th} observation, $i = 1, \dots, n$, consider

$$\boldsymbol{\Gamma}_i \overset{\text{d.}}{\sim} \mathsf{Multinomial}(1, \boldsymbol{\pi}_i),$$

where $\boldsymbol{\Gamma}_i \sim \mathsf{Multinomial}(1, \boldsymbol{\pi}_i)$ is a multinomial random variable representing the selection of an expert for the i^{th} observation. Notice that, $\gamma_{ik} = \{0, 1\}$ for all $k = 1, \dots, K$. and $\sum_{k=1}^K \gamma_{ik} = 1$, meaning that only one of the expert is selected, i.e., hard gating network. Here, $\boldsymbol{\pi}_i = (\pi_{i1}, \dots, \pi_{iK})$ denotes the probability vector over the experts for that particular leaf node with $\pi_{ik} \geq 0$, for all $k = 1, \dots, K$, and $\sum_{k=1}^K \pi_{ik=1}$. The log-likelihood function for the gating network's parameter $\boldsymbol{\eta}$ is given by

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\eta}) &= \sum_{i=1}^n \log\{f(z_{i,n_{\text{expert}}}, \dots, z_{i,1}, y_i; \boldsymbol{\eta})\} \\
&= \sum_{i=1}^n \log\left\{\sum_{k=1}^K \pi_{ik} f_k(z_{i,n_{\text{expert}}}, \dots, z_{i,1}, y_i; \boldsymbol{\eta})\right\},
\end{aligned}$$

Algorithm 1: *Algorithm for the inference of the BNN with probabilistic activation function by using a decision tree classifier as hard gating network for the MoE model.*

Data Input: Training dataset $\mathcal{D}_n$, number of experts K.

Initialize:

- Experts assignments by K-means clustering using both input and output data.
- The the decision tree classifier $\boldsymbol{T}$.
- The probabilities of belonging to each expert k, i.e., π_{ik} for each $i = 1, \ldots, n$ and $k = 1, \ldots, K$.

Cycle until convergence:

- Update $\boldsymbol{\theta}_{\text{global}}$, the parameters in global hidden layers.
- Assign the i^{th} observation to the k^{th} category via the decision tree classifier $\boldsymbol{T}$ with probability π_{ik} for each $i = 1, \ldots, n$ and $k = 1, \ldots, K$.
- Update the corresponding selected experts hidden layers parameters $\boldsymbol{\theta}^{[k]}_{\text{expert}}$ via either MFVB, or semiparametric MFVB, for all $i = 1, \ldots, n$ and $k = 1, \ldots, K$.
- Update the decision tree classifier gating network $\boldsymbol{T}$.

where $f_k(\boldsymbol{x}_i, y_i; \boldsymbol{\eta}_k)$ is the likelihood function of the k^{th} expert. By direct manipulation (e.g., McLachlan and Peel, 2000, Section 2.8), the posterior probability that $(\boldsymbol{x}_i, y_i)$ belongs to the k-th expert, also called *responsibilities*, is

$$\tau_{ik} = \frac{\pi_{ik} f_k(z_{i,n_{\text{expert}}}, \ldots, z_{i,1}, y_i; \boldsymbol{\eta}_k)}{\sum_{k=1}^{K} \pi_{ik} f_k(z_{i,n_{\text{expert}}}, \ldots, z_{i,1}, y_i; \boldsymbol{\eta}_k)}. \tag{4.3.10}$$

Finally, one assigns the i^{th} observation to the k^{th} expert with probability τ_{ik} for each $i = 1, \ldots, n$ and refit the gating network $\boldsymbol{T}$.

The above procedure continues until the proportional changes of the ELBOs are negligible and convergence of the parameters learning is achieved. The inference procedure in BNN with a decision tree classifier as the gating network is summarized in Algorithm 1.

4.3.4 Prediction via mixture of experts

After training the proposed BNN with a probabilistic activation function, the posterior predictive distribution (4.3.7) can be used to predict an output $y^{\text{new}} \in \mathbb{R}$ corresponding to

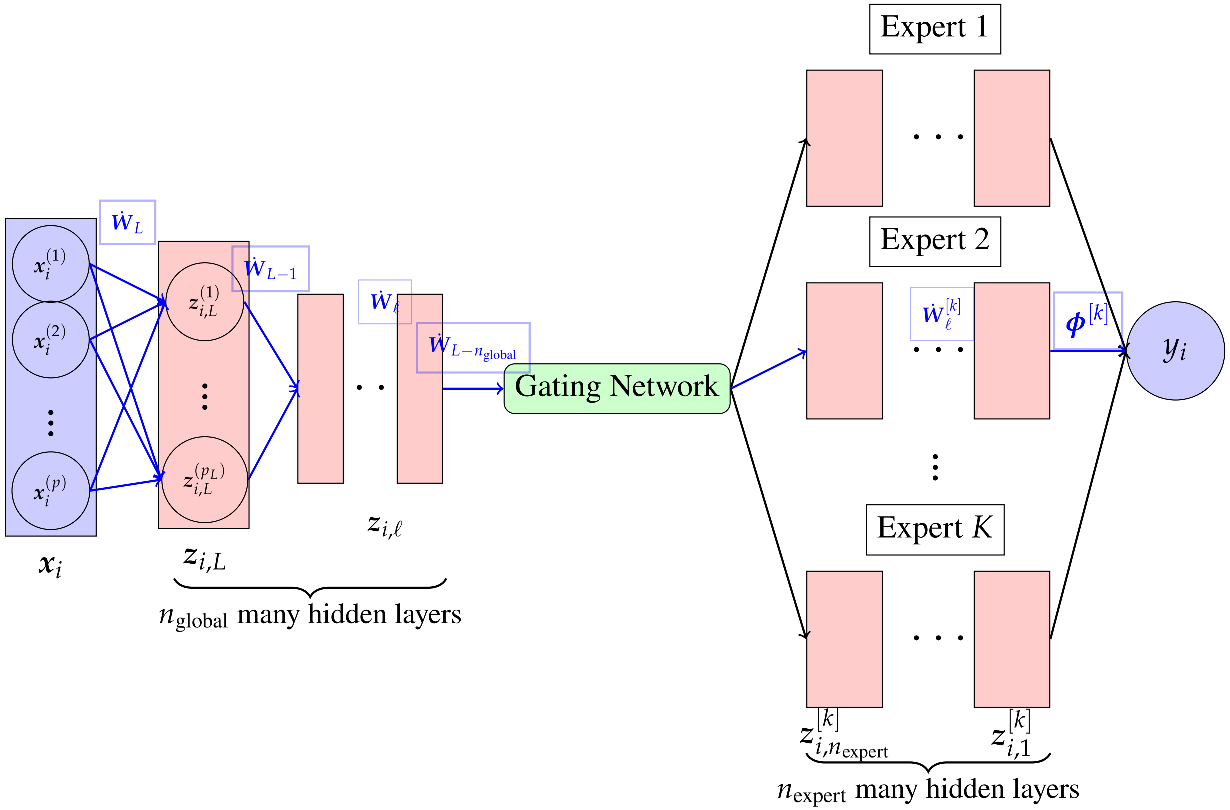


FIGURE 4.2: Our Mixture of Experts (MoEs) model structure.

some new input $\boldsymbol{x}^{\text{new}} \in \mathbb{R}^p$ applying the following procedure.

(i) Sample from the variational approximations to the marginal posterior density functions of the parameters and augmented variables, that is $\mathfrak{q}^\star(\boldsymbol{W}_\ell), \mathfrak{q}^\star(\boldsymbol{b}_\ell)$, and $\mathfrak{q}^\star(\boldsymbol{Z}_\ell)$, in a feed-forward manner for all global hidden layers $\ell = (L - n_{\text{global}} + 1), \ldots, L$.

(ii) Select an expert $k^{\star,\text{new}}$ in the gating network $\boldsymbol{T}$ based on the last hidden layer before it. Specifically, calculate the vector $\boldsymbol{\pi}^{\text{new}} = (\pi_1^{\text{new}}, \ldots, \pi_K^{\text{new}})^\top$ based on the new input for the gating network $\boldsymbol{W}_{L-n_{\text{global}}} \boldsymbol{z}^{\star,\text{new}}_{L-n_{\text{global}}+1} + \boldsymbol{b}_{L-n_{\text{global}}}$, that is,

$$\pi_k^{\text{new}} = \mathbb{P}\{\boldsymbol{z}^{\star,\text{new}}_{n_{\text{expert}}} \in C_k \mid \boldsymbol{T}(\boldsymbol{W}_{L-n_{\text{global}}} \boldsymbol{z}^{\star,\text{new}}_{L-n_{\text{global}}+1} + \boldsymbol{b}_{L-n_{\text{global}}}), \mathcal{D}_n\}, \quad k = 1, \ldots, K,$$

where π_k^{new} is the probability that the new input is assigned to the k^{th} expert. The new input is then assigned to the expert with the highest assignment probability,

$$k^{\star,\text{new}} = \arg\max_{k=1,\ldots,K} \pi_k^{\text{new}}.$$

(iii) Generate m predictions $y^{\text{new},1}, \ldots, y^{\text{new},m}$ approximating the posterior predictive density $\mathfrak{p}(y^{\text{new}} \mid \boldsymbol{x}^{\text{new}}, \mathcal{D}_n)$ via the following steps.

(iii)-1 Find the optimal variational density functions for the selected expert $k^{\star,\text{new}}$ in all expert hidden layer $\ell = 1, \ldots, n_{\text{expert}}$ and sample from the optimized variational densities in a feed-forward manner. If semiparametric MFVB is used to learn the parameters the hidden layer ℓ, find the optimal variational parameters $\boldsymbol{\xi}_\ell^{[k^{\star,\text{new}}]}$. Because a parametric variational approximation is used for the marginal posterior density of the expert hidden layer $\boldsymbol{z}_\ell^{[k^\star]}$, the expert variable $\boldsymbol{z}_\ell^{[k^{\star,\text{new}}]}$ can be directly sampled from the optimized variational density $\tilde{q}^\star(z_\ell; \boldsymbol{\xi}_\ell^{[k^{\star,\text{new}}]})$. On the other hand, if MFVB approximation is used to learn the parameters in the hidden layer, one can directly sample from the corresponding optimized variational density.

(iii)-2 Sample the variational approximations of the posterior density functions for the top layer $\boldsymbol{\phi}^{[k^{\star,\text{new}}]}$, $\mathfrak{q}^\star(\boldsymbol{\phi}^{[k^{\star,\text{new}}]})$.

(iii)-3 Sample the output $y^{\text{new},j}$, $j = 1, \ldots, m$, from the likelihood $\mathfrak{p}(y^{\text{new},j} \mid \boldsymbol{z}_1^{[k^{\star,\text{new}}]}, \boldsymbol{\phi})$, which depends on the probabilistic setting of the top layer.

The resulting posterior predictive distribution is a complete probability distribution that offers a thorough representation of uncertainty in prediction, and allows, for example, for calculation of credible intervals and frequency coverages.

In the following, we demonstrate the application of our methodology to two common tasks, namely regression and classification, focusing on inference and prediction.

4.4 Regression

In the following argument for regression, we argue with the decision tree classifier as the gating network, but the gating network is not necessarily to be decision tree classifier.

For the regression task, we consider a training dataset $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i) : \boldsymbol{x}_i \in \mathbb{R}^p, y_i \in \mathbb{R}, i = 1, \ldots, n\}$, where each $\mathbf{x}_i$ is a p-dimensional vector of explanatory variables, and y_i denotes the corresponding scalar response. The objective is to learn a function $f : \mathbb{R}^p \mapsto \mathbb{R}$ mapping the input to the output. To simplify the exposition, we consider a BNN with a single neuron in the expert hidden layer, employing a probabilistic ReLU activation function. Using the results in Section 4.2, we can formulate the regression model as follows:

$$\begin{aligned} y_i = z_i + \epsilon_i, \quad \epsilon_i \overset{\text{ind.}}{\sim} \mathsf{Normal}(0, \sigma^2), \\ \mathfrak{p}(z_i | \boldsymbol{x}_i, \boldsymbol{w}, b) \propto \tfrac{1}{\sqrt{2\pi z_i}} \exp\left\{ -\tfrac{1}{2z_i} (\boldsymbol{w}^\top \boldsymbol{x}_i + b + z_i)^2 \right\}, \quad i = 1, \ldots, n, \end{aligned} \tag{4.4.1}$$

where $\boldsymbol{w} \in \mathbb{R}^p$ and $b \in \mathbb{R}$. Similarly to Section 4.2, let $\mathfrak{p}(\boldsymbol{w}) \propto 1$, $\mathfrak{p}(b) \propto 1$ and $\mathfrak{p}(\sigma^2) \propto (\sigma^2)^{-1/2}$. It follows that $\mathfrak{p}(\boldsymbol{w}, b; \boldsymbol{x}_i) \propto \exp\{-2\max(\boldsymbol{w}^\top \boldsymbol{x}_i + b, 0)\}$, and the top layer can be viewed as a Gaussian regression with a variance parameter $\sigma^2 > 0$. Note that no additional augmented variables are needed in the top layer.

Assuming that the z_i's are independent, and letting $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{z} = (z_1, \ldots, z_n)^\top$ and $\boldsymbol{X}$ be a matrix obtained by stacking the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ in a row fashion, the joint likelihood function corresponding to model (4.4.1) is:

$$\begin{aligned} \mathfrak{p}(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{w}, b, \sigma^2, \boldsymbol{z}) &= \mathfrak{p}(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{w}, b, \sigma, \boldsymbol{z}) \mathfrak{p}(\boldsymbol{w}) \mathfrak{p}(b) \mathfrak{p}(\sigma^2) \mathfrak{p}(\boldsymbol{z}) \\ &\propto \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left\{ -\frac{\|\boldsymbol{y} - \boldsymbol{z}\|_2^2}{2\sigma^2} \right\} \\ &\quad \times \prod_{i=1}^n \frac{1}{\sqrt{2\pi z_i}} \exp\left\{ -\frac{(\boldsymbol{w}^\top \boldsymbol{x}_i + b + z_i)^2}{2z_i} \right\} \mathfrak{p}(\boldsymbol{w}) \mathfrak{p}(b) \mathfrak{p}(\sigma^2) \mathfrak{p}(\boldsymbol{z}). \end{aligned} \tag{4.4.2}$$

We aim to fit model (4.4.1) by applying the following mean field assumption on the variational approximating density for each cluster k, $k = 1, \dots, K$:

$$\mathsf{q}(\boldsymbol{w}^{[k]}, b^{[k]}, (\sigma^2)^{[k]}, \boldsymbol{z}^{[k]}) \approx \mathsf{q}(\boldsymbol{w}^{[k]}, b^{[k]})\mathsf{q}((\sigma^2)^{[k]})\tilde{\mathsf{q}}(\boldsymbol{z}^{[k]}; \boldsymbol{\xi}^{[k]}), \tag{4.4.3}$$

where $\tilde{\mathsf{q}}(\boldsymbol{z}^{[k]}; \boldsymbol{\xi}^{[k]}) = \prod_{i \in C_k} \tilde{\mathsf{q}}(z_i; \boldsymbol{\xi}^{[k]})$.

4.4.1 Inference in regression

For each expert k, let $\dot{\boldsymbol{w}}^{[k]} = (b^{[k]}, (\boldsymbol{w}^{[k]})^\top)^\top \in \mathbb{R}^{p+1}$ and $\dot{\boldsymbol{X}}^{[k]} = (\mathbf{1}_{n_k}, \boldsymbol{X}^{[k]}) \in \mathbb{R}^{n_k \times (p+1)}$, where $(\boldsymbol{X}^{[k]}, \boldsymbol{y}^{[k]})$ denotes all the data that is assigned to the k^{th} expert, n_k is the number of observations of the k^{th} expert, and $\mathbf{1}_{n_k}$ is a vector of length n_k whose elements are all equal to 1. Applying (4.3.3) to the logarithm of (4.4.2) gives

$$\mathsf{q}^\star(\dot{\boldsymbol{w}}^{[k]}) \propto \exp\left[-\frac{1}{2}\mathsf{E}_{\mathsf{q}(\boldsymbol{z}^{[k]})}\left\{(\dot{\boldsymbol{X}}^{[k]}\dot{\boldsymbol{w}}^{[k]} + \boldsymbol{z}^{[k]})^\top \mathsf{diag}(\boldsymbol{z}^{[k]})^{-1}(\dot{\boldsymbol{X}}^{[k]}\dot{\boldsymbol{w}}^{[k]} + \boldsymbol{z}^{[k]})\right\}\right],$$

where $\mathsf{diag}(\boldsymbol{z}^{[k]})$ is a diagonal matrix with the vector $\boldsymbol{z}^{[k]}$ on its main diagonal. The previous expression implies that $\mathsf{q}^\star(\dot{\boldsymbol{w}}^{[k]})$ is $\mathsf{Normal}(\boldsymbol{\mu}_{\mathsf{q}(\dot{\boldsymbol{w}}^{[k]})}, \boldsymbol{\Sigma}_{\mathsf{q}(\dot{\boldsymbol{w}}^{[k]})})$ with

$$\begin{aligned} \boldsymbol{\mu}_{\mathsf{q}(\dot{\boldsymbol{w}}^{[k]})} &= -\left((\dot{\boldsymbol{X}}^{[k]})^\top \mathrm{diag}\left(\mathsf{E}_{\mathsf{q}(z_1)}[z_1^{-1}], \dots, \mathsf{E}_{\mathsf{q}(z_{n_k})}[z_{n_k}^{-1}]\right)\dot{\boldsymbol{X}}^{[k]}\right)^{-1}(\dot{\boldsymbol{X}}^{[k]})^\top \mathbf{1}_{n_k}, \\ \boldsymbol{\Sigma}_{\mathsf{q}(\dot{\boldsymbol{w}}^{[k]})} &= \left((\dot{\boldsymbol{X}}^{[k]})^\top \mathrm{diag}\left(\mathsf{E}_{\mathsf{q}(z_1)}[z_1^{-1}], \dots, \mathsf{E}_{\mathsf{q}(z_{n_k})}[z_{n_k}^{-1}]\right)\dot{\boldsymbol{X}}^{[k]}\right)^{-1}. \end{aligned} \tag{4.4.4}$$

Applying (4.3.3) to the logarithm of (4.4.2) also gives

$$\mathsf{q}^\star((\sigma^2)^{[k]}) \propto \exp\left[\mathsf{E}_{\mathsf{q}(\boldsymbol{z}^{[k]})\mathsf{q}(\dot{\boldsymbol{w}}^{[k]})}\left\{(-\frac{n_k}{2} - \frac{1}{2})\log((\sigma^2)^{[k]}) - \frac{1}{2(\sigma^2)^{[k]}}(\boldsymbol{y}^{[k]} - \boldsymbol{z}^{[k]})^\top(\boldsymbol{y}^{[k]} - \boldsymbol{z}^{[k]})\right\}\right],$$

from which it follows that $\mathsf{q}^\star((\sigma^2)^{[k]})$ is $\mathsf{Inv\text{-}Gamma}(\alpha^{[k]}_{\mathsf{q}(\sigma^2)}, \beta^{[k]}_{\mathsf{q}(\sigma^2)})$, with

$$\alpha^{[k]}_{\mathsf{q}(\sigma^2)} = \frac{n_k - 1}{2}, \qquad \beta^{[k]}_{\mathsf{q}(\sigma^2)} = \frac{1}{2}\left((\boldsymbol{y}^{[k]})^\top \boldsymbol{y}^{[k]} - 2(\boldsymbol{y}^{[k]})^\top \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{z}^{[k]})} + \boldsymbol{\mu}^\top_{\mathsf{q}(\boldsymbol{z}^{[k]})}\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{z}^{[k]})}\right). \tag{4.4.5}$$

Instead, for augmented variable z_i,

$$\begin{aligned}
\mathsf{q}^{\star}(z_i) &\propto \exp\left[\mathsf{E}_{\mathsf{q}(\dot{\boldsymbol{w}}^{[k]})\mathsf{q}((\sigma^2)^{[k]})} \log\{\mathfrak{p}(z_i \mid \text{rest})\}\right] \\
&\propto \exp\Bigg[-\frac{(y_i-z_i)^2}{2}\mathsf{E}_{\mathsf{q}((\sigma^2)^{[k]})}\left(\frac{1}{\sigma^2}\right) - \frac{1}{2}\log(z_i) \\
&\quad -\frac{1}{2z_i}\left\{\boldsymbol{\mu}^{\top}_{\mathsf{q}(\dot{\boldsymbol{w}}^{[k]})}\dot{\boldsymbol{x}}_i^{\top}\dot{\boldsymbol{x}}_i\boldsymbol{\mu}_{\mathsf{q}(\dot{\boldsymbol{w}}^{[k]})} + \mathsf{Trace}\left(\dot{\boldsymbol{x}}_i^{\top}\dot{\boldsymbol{x}}_i\boldsymbol{\Sigma}_{\mathsf{q}(\dot{\boldsymbol{w}}^{[k]})}\right)\right\} - \dot{\boldsymbol{x}}_i\boldsymbol{\mu}_{\mathsf{q}(\dot{\boldsymbol{w}}^{[k]})} - \frac{1}{2}z_i\Bigg],
\end{aligned}$$

which does not correspond to any standard density function. To circumvent the need for computationally intensive numerical methods, we adopt a semiparametric mean-field variational Bayes (MFVB) framework. Specifically, we replace $\mathsf{q}(z_i)$ with a density belonging to a pre-specified family of distributions $\mathcal{Q}$, parameterized by $\boldsymbol{\xi} \in \Xi$. This approach imposes a parametric structure on the variational approximations of the z_i's while retaining the flexibility of the semiparametric formulation. In our framework, because of the presence of the expert hidden layer, we need to solve the optimization problem in Equation (4.3.9), where the $\mathsf{ELBO}_{\mathsf{q}}$ is abstractly defined in Equation (4.3.8), and it is given by :

$$\begin{aligned}
\mathsf{ELBO}_{\mathsf{q}}(\mathcal{D}_{C_k}; \dot{\boldsymbol{w}}^{[k]}, (\sigma^2)^{[k]}, \boldsymbol{z}^{[k]}, \boldsymbol{\xi}^{[k]}) &= \sum_{i \in C_k} \mathsf{ELBO}_{\mathsf{q}}(\boldsymbol{x}_i, y_i; \dot{\boldsymbol{w}}^{[k]}, (\sigma^2)^{[k]}, z_i, \boldsymbol{\xi}^{[k]}) \\
&= \sum_{i \in C_k} \mathsf{E}_{\mathsf{q}}\left\{\frac{\mathfrak{p}(\boldsymbol{x}_i, y_i, \dot{\boldsymbol{w}}^{[k]}, (\sigma^2)^{[k]}, z_i)}{\mathsf{q}(z_i; \boldsymbol{\xi}^{[k]})\mathsf{q}(\dot{\boldsymbol{w}}^{[k]})\mathsf{q}((\sigma^2)^{[k]})}\right\} \\
&= \sum_{i \in C_k} \mathsf{E}_{\mathsf{q}}[\log\{\mathfrak{p}(\boldsymbol{x}_i, y_i, \dot{\boldsymbol{w}}^{[k]}, (\sigma^2)^{[k]}, z_i)\}] \\
&\quad + \sum_{i \in C_k} \mathsf{Entropy}\{\mathsf{q}(z_i; \boldsymbol{\xi}^{[k]})\} + \mathsf{Entropy}\{\mathsf{q}(\dot{\boldsymbol{w}}^{[k]})\} \\
&\quad + \mathsf{Entropy}\{\mathsf{q}((\sigma^2)^{[k]})\}.
\end{aligned}$$

The non-entropy part is

$$\begin{aligned}
&\mathsf{E}_{\mathsf{q}}\left[\log\{\mathfrak{p}(\boldsymbol{x}_i, y_i, \dot{\boldsymbol{w}}^{[k]}, (\sigma^2)^{[k]}, z_i)\}\right] \\
&= -\log(2\pi) - \mathsf{E}_{\mathsf{q}((\sigma^2)^{[k]})}\left\{\log((\sigma^2)^{[k]})\right\} - \frac{1}{2}\mathsf{E}_{\mathsf{q}((\sigma^2)^{[k]})}\left(\frac{1}{\sigma^2}\right)\left\{y_i^2 - 2y_i\mathsf{E}_{\mathsf{q}(z_i;\boldsymbol{\xi}^{[k]})}(z_i) + \mathsf{E}_{\mathsf{q}(z_i;\boldsymbol{\xi}^{[k]})}(z_i^2)\right\} \\
&\quad - \frac{1}{2}\mathsf{E}_{\mathsf{q}(z_i;\boldsymbol{\xi}^{[k]})}\{\log(z_i)\} \\
&\quad - \frac{1}{2}\left[\mathsf{E}_{\mathsf{q}(z_i;\boldsymbol{\xi}^{[k]})}(z_i) + 2\dot{\boldsymbol{x}}_i\boldsymbol{\mu}_{\mathsf{q}(\dot{\boldsymbol{w}}^{[k]})} + \mathsf{E}_{\mathsf{q}(z_i;\boldsymbol{\xi}^{[k]})}\left(\frac{1}{z_i}\right)\{\boldsymbol{\mu}^{\top}_{\mathsf{q}(\dot{\boldsymbol{w}}^{[k]})}\dot{\boldsymbol{x}}_i^{\top}\dot{\boldsymbol{x}}_i\boldsymbol{\mu}_{\mathsf{q}(\dot{\boldsymbol{w}}^{[k]})} + \mathsf{Trace}\left(\dot{\boldsymbol{x}}_i^{\top}\dot{\boldsymbol{x}}_i\boldsymbol{\Sigma}_{\mathsf{q}(\dot{\boldsymbol{w}}^{[k]})}\right)\right]
\end{aligned}$$

and the entropy parts are

$$\begin{aligned}\text{Entropy}\{\mathfrak{q}(\dot{\boldsymbol{w}})\} =& \frac{1}{2}(p+1)\{1+\log(2\pi)\} + \frac{1}{2}\log(\boldsymbol{\Sigma}_{\mathfrak{q}(\dot{\boldsymbol{w}}^{[k]})})\\ \text{Entropy}\{\mathfrak{q}((\sigma^2)^{[k]})\} =& \alpha^{[k]}_{\mathfrak{q}(\sigma^2)} + \log(\beta^{[k]}_{\mathfrak{q}(\sigma^2)})\Gamma(\alpha^{[k]}_{\mathfrak{q}(\sigma^2)}) - (1+\alpha^{[k]}_{\mathfrak{q}(\sigma^2)})\psi(\alpha^{[k]}_{\mathfrak{q}(\sigma^2)}),\end{aligned}$$

where $\Gamma(\cdot)$ and $\psi(\cdot)$ are the gamma and digamma functions, respectively.

We now proceed to specify the class of density functions $\tilde{\mathfrak{q}}(z_i;\boldsymbol{\xi}^{[k]})$ under consideration. We assume that for each $i \in C_k$, $\tilde{\mathfrak{q}}(z_i;\boldsymbol{\xi}^{[k]}) \sim \mathsf{Inv\text{-}Gamma}(\alpha^{[k]},\beta^{[k]})$, where $\boldsymbol{\xi}^{[k]} = (\alpha^{[k]},\beta^{[k]})$ and $\alpha^{[k]}$, $\beta^{[k]} > 0$. Closed form expressions for the expectations involving functions of z_i and the entropy of an Inverse Gamma can be found in Appendix 4.A, Equations (4.A.7) and (4.A.8), respectively.

Whence, we need to solve the following optimization problem

$$((\alpha^{[k]})^\star,\ (\beta^{[k]})^\star) = \arg \max_{\alpha^{[k]}>2,\beta^{[k]}>0} \mathsf{ELBO}_{\mathfrak{q}}(\boldsymbol{X}_{C_k}, \boldsymbol{y}_{C_k}; \dot{\boldsymbol{w}}^{[k]}, (\sigma^2)^{[k]}, \boldsymbol{z}^{[k]}, \alpha^{[k]}, \beta^{[k]}), \tag{4.4.6}$$

where

$$\begin{aligned}&\mathsf{ELBO}_{\mathfrak{q}}(\mathcal{D}_{C_k}; \dot{\boldsymbol{w}}^{[k]}, (\sigma^2)^{[k]}, \boldsymbol{z}^{[k]}, \alpha^{[k]}, \beta^{[k]})\\ &= \sum_{i\in C_k} \mathsf{ELBO}_{\tilde{\mathfrak{q}}(z_i;\dot{\boldsymbol{w}}^{[k]},(\sigma^2)^{[k]},\boldsymbol{z}^{[k]},\alpha^{[k]},\beta^{[k]})}(\boldsymbol{x}_i, y_i; \alpha^{[k]}, \beta^{[k]})\\ &= \text{const} + \sum_{i\in C_k}\Bigg(\alpha^{[k]} + \log(\beta^{[k]}) + \log\{\Gamma(\alpha^{[k]})\} - (\alpha^{[k]}+1)\psi(\alpha^{[k]}) + \frac{1}{2}\log(|\boldsymbol{\Sigma}_{\mathfrak{q}(\dot{\boldsymbol{w}}^{[k]})}|)\\ &-\frac{\alpha^{[k]}_{\mathfrak{q}(\sigma^2)}}{2\beta^{[k]}_{\mathfrak{q}(\sigma^2)}}\left\{y_i^2 - 2y_i\left(\frac{\beta^{[k]}}{\alpha^{[k]}-1}\right) + \frac{(\beta^{[k]})^2}{(\alpha^{[k]}-1)(\alpha^{[k]}-2)}\right\} - \frac{1}{2}\{\log(\beta^{[k]}) - \psi(\alpha^{[k]})\}\\ &-\frac{1}{2}\left[\frac{\beta^{[k]}}{\alpha^{[k]}-1} + 2\dot{\boldsymbol{x}}_i\boldsymbol{\mu}_{\mathfrak{q}(\dot{\boldsymbol{w}}^{[k]})} + \frac{\alpha^{[k]}}{\beta^{[k]}}\left\{\boldsymbol{\mu}^\top_{\mathfrak{q}(\dot{\boldsymbol{w}}^{[k]})}\dot{\boldsymbol{x}}_i^\top\dot{\boldsymbol{x}}_i\boldsymbol{\mu}_{\mathfrak{q}(\dot{\boldsymbol{w}}^{[k]})} + \mathsf{Trace}(\dot{\boldsymbol{x}}_i^\top\dot{\boldsymbol{x}}_i\boldsymbol{\Sigma}_{\mathfrak{q}(\dot{\boldsymbol{w}}^{[k]})})\right\}\right]\Bigg).\end{aligned} \tag{4.4.7}$$

The inference procedure for the BNN with a single neuron in an expert hidden layer with a ReLU probabilistic activation function is summarised in Algorithm 2.

In initializing the assignment of training datasets to various experts, we first employ a K-means clustering method that considers both input and output data. This approach not only enhances the initial allocation by effectively partitioning the data based on the relationship between inputs and outputs but also surpasses traditional methods that focus solely on the input space. The decision tree classifier, acting as the hard gating network, adeptly learns the splitting rules necessary to assign input data to the most suitable expert. Throughout each iteration of the algorithm, we continuously update the variational densities of all parameters, the decision tree classifier, and the training dataset assignments. This inference process persists until the proportional changes in the ELBO for all experts become insignificant. A compelling advantage of incorporating a probabilistic ReLU activation function, as previously shown, is that the posterior distribution of weights and biases, denoted as $\dot{w}$, remains manageable. This tractability significantly enhances our ability to conduct Bayesian inference with greater efficiency.

4.4.2 Prediction in regression

We present the prediction procedure for a BNN with a single neuron in an expert hidden layer with a ReLU probabilistic activation function, for the regression model in Equation (4.4.1). We proceed as detailed below.

(i) We assign the new input $\boldsymbol{x}^{\text{new}}$ to the expert $k^{\star,\text{new}}$ by the gating network $\boldsymbol{T}$, which corresponds to the highest mixing coefficient. That is, $k^{\star,\text{new}} = \arg\max_{k=1,\ldots,K} \mathbb{P}(\boldsymbol{x}^{\text{new}}, y^{\text{new}}) \in C_k \mid \boldsymbol{T}(\boldsymbol{x}^{\text{new}}), \mathcal{D}_n)$.

(iii) We generate N predictions $(y_i^{\text{new},j})_{j=1}^{N}$ approximating the posterior predictive density $\mathfrak{p}(y_i^{\text{new}} \mid \boldsymbol{x}_i^{\text{new}}, \mathcal{D}_n)$ via the following steps:

(iii)-1 Find the optimal variational density function for the expert $k^{\star,\text{new}}$ characterized by the optimal variational parameters $\xi^{[k^\star,\text{new}]} = ((\alpha^{[k^\star]})^\star, (\beta^{[k^\star]})^\star)$; $\tilde{\mathfrak{q}}(z^{[k^\star,\text{new}]}; ((\alpha^{[k]})^\star, (\beta^{[k]})^\star) \sim \mathsf{Inv\text{-}Gamma}((\alpha^{[k]})^\star, (\beta^{[k]})^\star)$. We directly sample the expert variable $z^{[k^\star,\text{new}]}$ from the latter optimized variational density.

(iii)-2 Sample the variance $(\sigma^2)^{[k^\star]}$ of the Gaussian top layer according to Equation (4.4.5).

(iii)-3 Generate the output $y_i^{\text{new},j} = z^{[k^\star,\text{new}]} + \epsilon$ by sampling $\epsilon \sim \mathsf{Normal}(0, (\sigma^2)^{[k^\star]})$.

The N predictions, $(y_i^{\text{new},j})_{j=1}^{N}$ are used to approximate the predictive posterior density $\mathfrak{p}(y_i^{\text{new}} \mid \boldsymbol{x}_i^{\text{new}}, \mathcal{D}_n)$. To obtain a mean point estimate, we compute $\widehat{y}_i^{\text{new}} = \frac{1}{N}\sum_{j=1}^{N} y_i^{\text{new},j}$. Alternatively, the median point estimate is given by

$$\mathsf{median} = \begin{cases} y_i^{\text{new},(N/2)}, & \text{if } N \text{ is even,} \\ y_i^{\text{new},((N+1)/2)}, & \text{if } N \text{ is odd.} \end{cases}$$

To quantify the uncertainty, one can construct, for example, a 95% empirical credible interval.

4.5 Classification

For the classification task, we consider a training dataset $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i) : \boldsymbol{x}_i \in \mathbb{R}^p, y_i \in \{0,1\},\ i = 1,\ldots,n\}$, where each $\mathbf{x}_i$ is a vector of feature variables, and y_i indicates the binary class membership of $\boldsymbol{x}_i$. The aim is to develop a decision function $f : \mathbb{R}^p \mapsto \{0,1\}$ that predicts the class label y_i based on the feature vector.

For clarity and simplicity, we consider a BNN with a single neuron in the expert hidden layer, equipped with a probabilistic ReLU activation function. Additionally, we adopt the logit link function; however, alternative link functions can be incorporated depending on the response type, as discussed at the end of Section 4.2.1.2. Notably, in classification tasks, we can take advantage of the data augmentation techniques proposed by Polson et al. (2013), which help make the logistic regression model in the top layer manageable. Instead, for regression tasks, we do not require this additional augmented variable. By using the results in Section 4.2, we write the classification model in the following way:

$$\begin{aligned} \mathfrak{p}(y_i|z_i,\alpha_0,\alpha_1) &= \pi_i^{y_i}(1-\pi_i)^{1-y_i} = \frac{\exp\{y_i(\alpha_0+\alpha_1 z_i)\}}{1+\exp\{(\alpha_0+\alpha_1 z_i)\}} \\ \mathfrak{p}(z_i|\boldsymbol{x}_i,\boldsymbol{w},b) &\propto \frac{1}{\sqrt{2\pi z_i}}\exp\left\{-\frac{1}{2z_i}(\boldsymbol{w}^\top \boldsymbol{x}_i + b + z_i)^2\right\}, \quad i = 1,\ldots,n, \end{aligned} \tag{4.5.1}$$

where $\boldsymbol{w} \in \mathbb{R}^p$, $b \in \mathbb{R}$, $y_i = \{0,1\}$ and $\pi_i = \mathbb{P}(y_i = 1 \mid z_i, \alpha_0, \alpha_1) = \frac{1}{1+\exp\{-(\alpha_0+\alpha_1 z_i)\}}$, according to the simple logistic model.

4.5.1 Inference in classification

Let $\dot{\boldsymbol{\alpha}} = (\alpha_0,\alpha_1)^\top \in \mathbb{R}^2$, $\dot{\boldsymbol{w}} = (b, \boldsymbol{w}^\top)^\top \in \mathbb{R}^{1+p}$, $\dot{\boldsymbol{z}}_i^\top = (1, z_i)^\top \in \mathbb{R}^2$, $\dot{\boldsymbol{x}}_i = (1, \boldsymbol{x}_i^\top)^\top \in \mathbb{R}^{1+p}$. We can apply Pólya-Gamma data augmentation trick as in Polson et al. (2013). Let $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_n)^\top \in \mathbb{R}^n$ the corresponding vector of the auxiliary variable in the top layer, and $\boldsymbol{z} = (z_1,\ldots,z_n)^\top \in \mathbb{R}^n$ the vector of the expert variable. By using Equation (4.B.3), we have

$$\frac{\exp(y_i\dot{\boldsymbol{\alpha}}^\top\dot{\boldsymbol{z}}_i\}}{1+\exp\{\dot{\boldsymbol{\alpha}}^\top\dot{\boldsymbol{z}}_i\}} = \frac{1}{2}\exp(\kappa_i\dot{\boldsymbol{\alpha}}^\top\dot{\boldsymbol{z}}_i\}\int_0^\infty \exp\left\{-\frac{\omega_i(\dot{\boldsymbol{\alpha}}^\top\dot{\boldsymbol{z}}_i)^2}{2}\right\}\mathfrak{p}(\omega_i \mid 1,0)\, d\omega_i$$

where $\kappa_i = y_i - \frac{1}{2}$ and $\mathfrak{p}(\omega_i \mid 1, 0) \sim \mathsf{PG}(1, 0)$; see Appendix 4.A.2. In other words, conditioned on the introduced augmented variable ω_i,

$$\mathfrak{p}(y_i | z_i, \dot{\boldsymbol{\alpha}}, \omega_i) \propto \exp\left\{\kappa_i \dot{\boldsymbol{\alpha}}^\top \dot{\boldsymbol{z}}_i - \frac{\omega_i (\dot{\boldsymbol{\alpha}}^\top \dot{\boldsymbol{z}}_i)^2}{2}\right\}.$$

The joint likelihood function corresponding to the model in Equation (4.5.1) is given by the following equation

$$\begin{aligned}
&\mathfrak{p}(y_i, \boldsymbol{x}_i, \dot{\boldsymbol{w}}, \boldsymbol{\alpha}, z_i, \omega_i) \propto \mathfrak{p}(y_i \mid z_i, \boldsymbol{\alpha}, \boldsymbol{\omega}_i) \mathfrak{p}(\omega_i) \mathfrak{p}(z_i | \boldsymbol{x}_i, \dot{\boldsymbol{w}}) \mathfrak{p}(\dot{\boldsymbol{w}}) \mathfrak{p}(\dot{\boldsymbol{\alpha}}) \\
&\propto \exp\left\{\kappa_i \dot{\boldsymbol{\alpha}}^\top \dot{\boldsymbol{z}}_i - \frac{\omega_i (\dot{\boldsymbol{\alpha}}^\top \dot{\boldsymbol{z}}_i)^2}{2}\right\} \mathfrak{p}(\omega_i \mid 1, 0) \frac{1}{\sqrt{2\pi z_i}} \exp\left\{-\frac{1}{2z_i}\left(\dot{\boldsymbol{w}}^\top \dot{\boldsymbol{x}}_i + z_i\right)^2\right\}.
\end{aligned}$$

To develop a semiparametric mean-field variational Bayes (MFVB) algorithm, we begin by assuming the following factorization of the variational distribution $\mathfrak{q}(\boldsymbol{\omega}, \dot{\boldsymbol{\alpha}}, \dot{\boldsymbol{w}}, \dot{\boldsymbol{z}})$:

$$\mathfrak{q}(\boldsymbol{\omega}, \dot{\boldsymbol{\alpha}}, \dot{\boldsymbol{w}}, \dot{\boldsymbol{z}}) \approx \mathfrak{q}(\boldsymbol{\omega}) \mathfrak{q}(\dot{\boldsymbol{\alpha}}) \mathfrak{q}(\dot{\boldsymbol{w}}) \tilde{\mathfrak{q}}(\boldsymbol{z}; \boldsymbol{\xi}),$$

where $\tilde{\mathfrak{q}}(\boldsymbol{z}; \boldsymbol{\xi}) = \prod_{i=1}^n \tilde{\mathfrak{q}}(z_i; \boldsymbol{\xi}^{[k_i]})$ with k_i representing the observation $i \in C_k$. And as usual, $\{\tilde{\mathfrak{q}}(z; \boldsymbol{\xi}) : \boldsymbol{\xi} \in \Xi)\}$ is a collection of pre-specified parametric family. The optimal variational densities of $\omega_i, \boldsymbol{\alpha}$ and $\dot{\boldsymbol{w}}$ are obtained by performing the following computations:

$$\begin{aligned}
\mathfrak{q}^\star(\omega_i) &\propto \exp\left\{\mathsf{E}_{\mathfrak{q}(\dot{\boldsymbol{w}})\mathfrak{q}(\dot{\boldsymbol{\alpha}})\mathfrak{q}(\boldsymbol{z})}\left[\log\{\mathfrak{p}(\omega_i \mid \text{rest})\}\right]\right\}, \quad i = 1, \ldots, n. \\
\mathfrak{q}^\star(\dot{\boldsymbol{\alpha}}) &\propto \exp\left\{\mathsf{E}_{\mathfrak{q}(\boldsymbol{\omega})\mathfrak{q}(\dot{\boldsymbol{w}})\mathfrak{q}(\boldsymbol{z})}\left[\log\{\mathfrak{p}(\dot{\boldsymbol{\alpha}} \mid \text{rest})\}\right]\right\}, \\
\mathfrak{q}^\star(\dot{\boldsymbol{w}}) &\propto \exp\left\{\mathsf{E}_{\mathfrak{q}(\boldsymbol{\omega})\mathfrak{q}(\dot{\boldsymbol{\alpha}})\mathfrak{q}(\boldsymbol{z})}\left[\log\{\mathfrak{p}(\dot{\boldsymbol{w}} \mid \text{rest})\}\right]\right\}.
\end{aligned}$$

On the other hand, for each expert k, $k = 1, \ldots, K$, we have

$$\tilde{\mathfrak{q}}(\boldsymbol{z}; \boldsymbol{\xi}_k) = \arg \max_{\tilde{\mathfrak{q}} \in \{\tilde{\mathfrak{q}}(\boldsymbol{z}; \boldsymbol{\xi}^{[k]}) : \boldsymbol{\xi}^{[k]} \in \Xi)\}} \text{ELBO}_{\mathfrak{q}}(\boldsymbol{X}_{C_k}, \boldsymbol{y}_{C_k}; \dot{\boldsymbol{w}}, \alpha, \boldsymbol{\omega}, \boldsymbol{z}, \boldsymbol{\xi}^{[k]}).$$

Notably, the optimal variational densities for the ω_is are tractable and follow Pólya-Gamma distributions, due to the data augmentation technique presented in Theorem 1 of Polson et al. (2013). The gating network is then updated similarly in each iteration.

4.5.2 Prediction in classification

We present the prediction procedure for a BNN with a single neuron in an expert hidden layer with a ReLU probabilistic activation function, for the classification model in (4.5.1). We proceed as detailed below.

(i) We assign the new input $\boldsymbol{x}_i^{\text{new}}$ to the expert $k^{\star,\text{new}}$ by the gating network $\boldsymbol{T}$, which corresponds to the highest mixing coefficient. That is, $k^{\star,\text{new}} = \arg\max_{k=1,\dots,K} \mathbb{P}((\boldsymbol{x}_i^{\text{new}}, y_i^{\text{new}}) \in C_k \mid \boldsymbol{T}(\boldsymbol{x}_i^{\text{new}}), \mathcal{D}_n)$.

(ii) We generate N predictions $(y_i^{\text{new},j})_{j=1}^N$ approximating the posterior predictive density $\mathfrak{p}(y_i^{\text{new}} \mid \boldsymbol{x}_i^{\text{new}}, \mathcal{D}_n)$ via the following steps:

(ii)-1 Find the optimal variational density function for the expert $k^{\star,\text{new}}$ characterized by the optimal variational parameters $\boldsymbol{\xi}^{[k^{\star,\text{new}}]}$. We directly sample the expert variable $\boldsymbol{z}^{\star,\text{new}}$ from the latter optimized variational density.

(ii)-2 Sample $\boldsymbol{\alpha}$ in the top layer from $\mathfrak{q}^{\star}(\boldsymbol{\alpha})$, and define:

$$p^{\star,\text{new}} = \mathbb{P}(y_i^{\star,\text{new}} | x_i^{\text{new}}, \boldsymbol{\alpha}) = \frac{1}{1 + \exp(\dot{\boldsymbol{\alpha}}^{\top} \dot{\boldsymbol{z}}^{\star,\text{new}})},$$

where $\dot{\boldsymbol{z}}^{\star,\text{new}} = (1, \boldsymbol{z}^{\star,\text{new}})$.

(ii)-3 Generate the output $y_i^{\text{new},j} \sim \mathsf{Bernoulli}(p^{\star,\text{new}})$.

The N predictions, $(y_i^{\text{new},j})_{j=1}^N$ are used to approximate the predictive posterior density $\mathfrak{p}(y_i^{\text{new}} \mid \boldsymbol{x}_i^{\text{new}}, \mathcal{D}_n)$, similarly to the regression case.

4.6 Numerical implementation and results

To illustrate the effectiveness of our proposed fast mixture-of-experts Bayesian learning method for deep neural networks, we present three numerical experiments using real-world datasets: the Concrete Compressive Strength data(regression task; see Subsection 4.6.1), the Energy Efficiency data(regression task; see Subsection 4.6.2), the Boston Housing dataset (regression task; see Subsection 4.6.3), and the Wine Quality dataset (classification task; see Subsection 4.6.4).

4.6.1 Regression task on Concrete Compressive Strength data

The Concrete Compressive Strength data is introduced from Yeh (1998), where the author investigated the predictive modelling of high-performance concrete compressive strength using artificial neural networks. It contains $n = 1030$ concrete mix designs with corresponding 28-day compressive strength values, compiled from multiple laboratory batches prepared under controlled conditions. For each observation, $p = 8$ covariates are recorded, characterizing the composition and curing regime of the mix. The compositional variables include cement content (Cement), blast furnace slag content (Slag), fly ash content (Fly Ash), water content (Water), superplasticizer dosage (Superplasticizer), coarse aggregate mass (Coarse Aggregate), and fine aggregate mass (Fine Aggregate), all measured in kilograms per cubic meter (kg/m^3). The curing variable is the age of the concrete specimen in days (Age) at the time of compressive strength testing. The response variable is the uniaxial compressive strength of the concrete, measured in megapascals (MPa). All covariates are standardized by centering to zero mean and scaling to unit variance. The transformation is performed using the mean and variance estimated from the training data, which are subsequently applied to standardize the validation dataset.

Choice of hyper-parameters and architecture. We use the decision tree classifier as the gating network, and we implement our method with a number of experts K either equals to 16. Two models structure are implemented, one comprising one Gaussian expert hidden layer ($n_{\text{expert}} = 1$) with the number of neurons $p_1 = 4$, while the other one comprising two expert hidden layers ($n_{\text{expert}} = 2$), the number of neurons are $p_2 = 4$ and $p_1 = 4$, respectively. For each expert k, we use a Gaussian hidden layer $\boldsymbol{z}_{i,2}^{[k]} \in \mathbb{R}^{p_2}$ for the first layer and a probabilistic ReLU layer $\boldsymbol{z}_{i,1}^{[k]} \in \mathbb{R}^{p_1}$ for the second one. Weights and biases are denoted as usual by $\dot{\boldsymbol{W}}_1^{[k]} = (\boldsymbol{b}_1^{[k]}, \boldsymbol{W}_1^{[k]}) \in \mathbb{R}^{p_1 \times (1+p)}$ and $\dot{\boldsymbol{W}}_2^k = (\boldsymbol{b}_2^{[k]}, \boldsymbol{W}_2^{[k]}) \in \mathbb{R}^{p_2 \times (1+p_1)}$ and learned via MFVB; Gaussian hidden layer neurons $\boldsymbol{z}_{i,2}^{[k]}$ and the variance parameter $(\sigma_2^2)^{[k]} \in \mathbb{R}$ are learned via MFVB; the augmented variables in probabilistic ReLU layer $\boldsymbol{z}_2$ are updated via semiparametric MFVB. The top layer is a Gaussian regression with parameters $(b_0^{[k]}, \boldsymbol{w}_0^{[k]}, (\sigma_0^2)^{[k]})$ learned via MFVB. $n_{\text{global}} = 0$. We split out 30% of the dataset as the validation set, and we perform 20 simulation runs by using different sub-sampling of the dataset.

Initialization. We initialize our BNN as follows. Experts assignments are first obtained via K-MEANS clustering using *both* input and output data. A decision tree classifier $\boldsymbol{T}$ is then fitted to reproduce these experts assignments from the input data alone, yielding disjoint expert assignments $C_1, \ldots, C_K$.

In the first model, with a Gaussian expert layer with $K = 16, p_1 = 4$, each expert k applies a PCA with p_1 components to $\boldsymbol{X}_{C_k}$ to obtain initial $\boldsymbol{z}_{i,1}^{[k]} \in \mathbb{R}^{p_1}$, for all $i \in C_k$. The first-layer weights and biases $\dot{\boldsymbol{W}}_1^{[k]}$ are initialized via multivariate multiple linear regression from $\boldsymbol{X}_{C_k}$ to the corresponding $\{\boldsymbol{z}_i^{[k]}\}_{i \in C_k}$, while the second-layer weights and biases $\dot{\boldsymbol{W}}_0^{[k]}$ are set to the coefficients of an OLS regression from $\{\boldsymbol{z}_i^{[k]}\}_{i \in C_k}$ to $\boldsymbol{y}_{C_k}$. The variance parameters for both the hidden and output layers, $(\sigma_1^2)^{[k]}$ and $(\sigma_0^2)^{[k]}$, $k = 1, \ldots, K$, are initialized from the residual sum of squares of their respective regressions.

In the second model, we have a Gaussian expert layer followed by a probabilistic ReLU expert layer, with $K = 16, p_1 = 4, p_2 = 4$. The Gaussian expert layer parameters: $\boldsymbol{z}_{i,2}^{[k]} \in \mathbb{R}^{p_2}$ for all $i \in C_k$, $\dot{\boldsymbol{W}}_2^{[k]}$ are initialized exactly as before. The probabilistic ReLU layer is initialized as follows. $\boldsymbol{z}_{i,1}^{[k]} \in \mathbb{R}^{p_1}$ is initialized by sampling from an inverse-gamma distribution. The weights and biases $\dot{\boldsymbol{W}}_1$ is initialized as the multivariate multiple linear regression coefficients from $\{\boldsymbol{z}_{i,2}^{[k]}\}_{i \in C_k}$ to $\{\boldsymbol{z}_{i,1}^{[k]}\}_{i \in C_k}$, while the variance parameter $(\sigma_1^2)^{[k]}$ is initialized as 1.0. The output layer weights and biases $\dot{\boldsymbol{W}}_0^{[k]}$ are drawn from a zero-mean Gaussian distribution with small variance, and the variance parameter $(\sigma_0^2)^{[k]}$ is set to 1.0.

Based on these initializations, the variation means required in MFVB and semiparametric MFVB are obtained by taking empirical means over the corresponding initialized quantities.

Comparison with other methodologies. We compare the performance of our methodology against several baseline regression frameworks. First, we consider a classical feedforward neural network (FFNN) with ReLU activation functions consisting of a single hidden layer with 64 units and a dropout rate of 0.5 (see Srivastava et al., 2014). Second, we include the data augmentation Gaussian regression (DA-GR) model proposed by Wang et al. (2022) (see Section 4.1). The implementation employs a single hidden-layer ReLU network with 64 neurons and a dropout rate of 0.5 (see Srivastava et al., 2014). We consider numbers of MCMC copies $J \in \{2, 5, 10\}$, with stochastic noise parameters in the top and hidden layers set to $\tau_0 = 0.4$ and $\tau_z = 0.5$, respectively, selected via grid search. The configurations in FFNN and DA-GR are matched, allowing for a consistent comparison.

We also implemented two linear regression models: ordinary least square (OLS) and ridge regression (Ridge) with regularization parameter $\alpha = 3.0$. Lastly, three non-linear regression models are considered: random forest regression (Random Forest) (see Breiman, 2001), an ensemble of $n = 10$ regression trees, aggregated via bagging, with bootstrap sampling and randomized feature selection at each split; extreme gradient boosting (XGBoost) (see Chen and Guestrin, 2016), a gradient-boosted decision tree ensemble with $n = 30$, learning rate $\eta = 0.05$, maximum depth $d = 4$, subsampling ratio 0.8, and column subsampling ratio 0.8

at each tree; Gaussian process regressor (GP) (e.g., MacKay, 1992), using a kernel composed of a constant term and a radial basis function (RBF). The kernal takes the form

$$\mathsf{Kernel}(x, x^i) = C \exp\left(-\frac{1}{2r^2}\|x - x^{'}\|_2^2\right),$$

where C controls the vertical scale of the function, and r, the length scale, controls the smoothness of sample paths; $\|\cdot\|_2^2$ indicates the usual squared euclidean distance between input vectors. Prior to model fitting, the response variable y is normalized to zero mean prior to model fitting.

The OLS regression and ridge regression are implemented via the *LinearRegression* and *Ridge* classes from `scikit-learn`. The random forest regressor is implemented via the *RandomForestRegressor* class from `scikit-learn`, while the extreme gradient boosting (XGBoost) regressor is implemented via the *XGBRegressor* class from `xgboost`. The Gaussian process regressor is implemented via the *GaussianProcessRegressor* class from `scikit-learn`. We evaluate the performance of the different methods by computing the MedAE.

Discussion of the results. Figure 4.3 displays the MedAE (reported in logarithmic scale) as a function of iterations on validation data for all the methods described in the prvious paragraph. The shaded areas indicate the interquartile range ($25^{th} - 75^{th}$ percentile) of the MedAE across repeated runs. The results indicate that DA-GR with $J = \{2, 5, 10\}$ exhibits a strong performance when compared to ReLU-based FFNN, achieving lower validation errors throughout the iterations. In contrast, the figure highlights the excellent performance of our proposed methods. With the aforementioned initializations, both PDL variants attain low MedAE values from the very first iterations, demonstrating fast convergence and stable performance. Moreover, it is worth noted that the inclusion of the probabilistic ReLU expert layer in the Gaussian+ProbReLU configuration yields additional performance gains over the purely Gaussian expert layer. This suggests that introducing a probabilistic nonlinearity in the expert layers can enhance model flexility and prove predictive accuracy. Notably, the PDL models achieve validation errors that are consistently comparable to, and often lower than, those of the best-performing baseline regressors (OLS, Ridge, Random Forest, XGBoost, and GP).

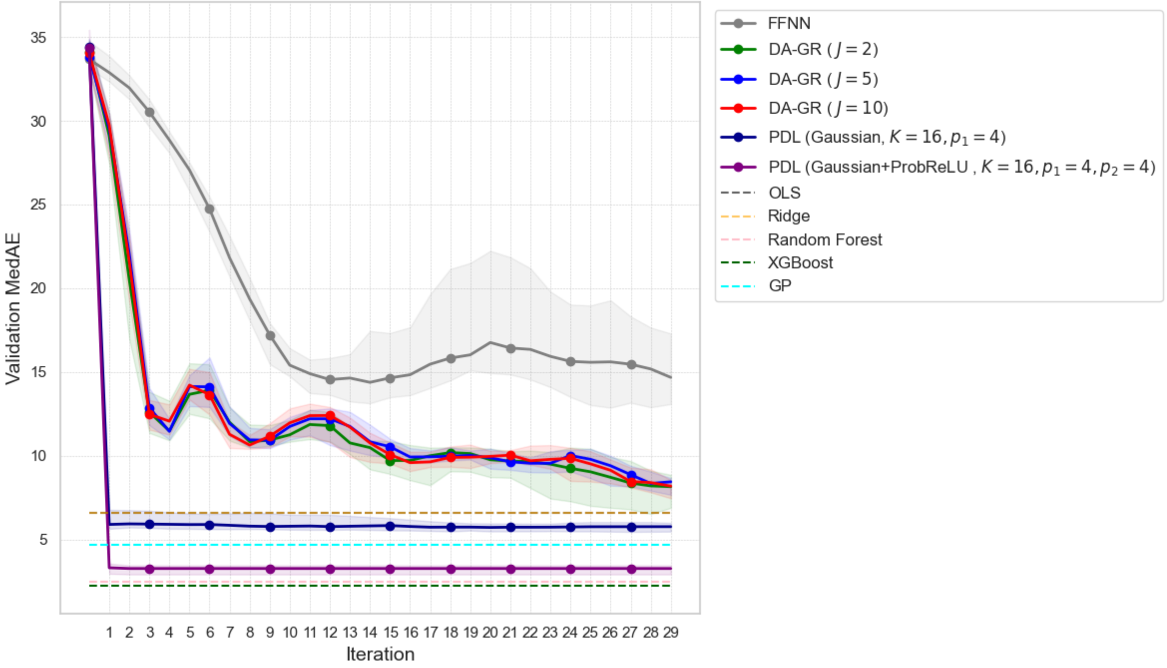


FIGURE 4.3: Regression task on Concrete Compressive Strength Data. Median absolute error (MEDAE; and the corresponding variability) as a function of iterations on validation data for the following methods: feedforward neural network with ReLU activation (FFNN) (gray solid line); DA-GR with $J = 2$ (green solid line), $J = 5$ (blue solid line), $J = 10$ (red solid line); our probabilistic deep learning (PDL) Bayesian model with two different architectures (navy-blue solid line and purple solid line); ordinary least square (OLS) regression (gray dashed line), ridge regression (yellow dashed line), random forest regression (pink dashed line), extreme gradient boosting (XGBoost; green dashed line) and Gaussian process (GP) regressor (sky-blue dashed line).

4.6.2 Regression task on Energy Efficiency data

The Energy Efficient dataset comprises $n = 768$ simulated building designs, each characterized by $p = 8$ features that describe the geometry and physical attributes of the building, as originally introduced by Tsanas and Xifara (2012). The covariates include relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area and glazing area distribution. Orientation and glazing area distribution are represented as integer encodings corresponding to categories, while the remaining covariates are continuous variables. The data contains two continuos target variables: heating load and cooling load, measured in kilowatt-hours per square meter (kWh/m^2). There are no missing values. All covariates are standardized by centering to zero mean and scaling to unit variance based on

the training set, which are then applied to validation dataset.
We conduct two separate regression tasks, one for each target variable: heating load (test1) and cooling load (test2).

Choice of hyper-parameters and architecture. We use the decision tree classifier as the gating network, and we implement our method with a number of experts K either equals to 6 or 10, each one comprising two expert hidden layers ($n_{\text{expert}} = 2$); the number of neurons $p_2 = 6$ and $p_1 = 4$. For each expert k, we use a Gaussian hidden layer $\boldsymbol{z}_{i,2}^{[k]} \in \mathbb{R}^{p_2}$ for the first layer and a probabilistic ReLU layer $\boldsymbol{z}_{i,1}^{[k]} \in \mathbb{R}^{p_1}$ for the second one. Weights and biases are denoted as usual by $\dot{\boldsymbol{W}}_1^{[k]} = (\boldsymbol{b}_1^{[k]}, \boldsymbol{W}_1^{[k]}) \in \mathbb{R}^{p_1 \times (1+p)}$ and $\dot{\boldsymbol{W}}_2^{[k]} = (\boldsymbol{b}_2^{[k]}, \boldsymbol{W}_2^{[k]}) \in \mathbb{R}^{p_2 \times (1+p_1)}$ and learned via MFVB; Gaussian hidden layer neurons $\boldsymbol{z}_{i,2}^{[k]}$ and the variance parameter $(\sigma_2^2)^{[k]} \in \mathbb{R}$ are learned via MFVB; the augmented variables in probabilistic ReLU layer $\boldsymbol{z}_2^{[k]}$ are updated via semiparametric MFVB. The top layer is a Gaussian regression with parameters $(b_0^{[k]}, \boldsymbol{w}_0^{[k]}, (\sigma_0^2)^{[k]})$ learned via MFVB. $n_{\text{global}} = 0$. We split out 30% of the dataset as the validation set, and we perform 20 simulation runs by using different sub-sampling of the dataset.

Initialization. We initialize our BNN in the following way. Experts assignments are obtained via K-means clustering using *both* input and output data; the decision tree classifier $\boldsymbol{T}$ is then fitted to these experts assignments, yielding disjoint expert assignments $C_1, \ldots, C_K$. The number of experts K equals either 10 or 6, each expert k has a Gaussian expert layer followed by a probabilistic ReLU expert layer, with $p_1 = 6, p_2 = 4$. The Gaussian expert layer parameters: $\boldsymbol{z}_{i,2}^{[k]} \in \mathbb{R}^{p_2}$ for all $i \in C_k$, $\dot{\boldsymbol{W}}_2^{[k]}$ are initialized exactly as before. The probabilistic ReLU layer is initialized as follows. $\boldsymbol{z}_{i,1}^{[k]} \in \mathbb{R}^{p_1}$ is initialized by sampling from an inverse-gamma distribution. The weights and biases $\dot{\boldsymbol{W}}_1{}^{[k]}$ is initialized as the multivariate multiple linear regression coefficients from $\{\boldsymbol{z}_{i,2}^{[k]}\}_{i \in C_k}$ to $\{\boldsymbol{z}_{i,1}^{[k]}\}_{i \in C_k}$, while the variance parameter $(\sigma_1^2)^{[k]}$ is initialized as 1.0. The output layer weights and biases $\dot{\boldsymbol{W}}_0^{[k]}$ are drawn from a zero-mean Gaussian distribution with small variance, and the variance parameter $(\sigma_0^2)^{[k]}$ is set to 1.0.
Based on these initializations, the variation means required in MFVB and semiparametric MFVB are obtained by taking empirical means over the corresponding initialized quantities.

Comparison with other methodologies. We compare the performance of our methodology against several baseline regression frameworks. First, we consider a classical feedforward neural network (FFNN) with ReLU activation functions consisting of a single hidden layer with 64 units and a dropout rate of 0.5 (see Srivastava et al., 2014). Second, we include the data augmentation Gaussian regression (DA-GR) model proposed by Wang et al. (2022)

(see Section 4.1). The implementation employs a single hidden-layer ReLU network with 64 neurons and a dropout rate of 0.5 (see Srivastava et al., 2014). We consider numbers of MCMC copies $J \in \{2, 5, 10\}$, with stochastic noise parameters in the top and hidden layers set to $\tau_0 = 0.4$ and $\tau_z = 0.5$, respectively, selected via grid search. The configurations in FFNN and DA-GR are matched, allowing for a consistent comparison.

We also implemented two linear regression models: ordinary least square (OLS) and ridge regression (Ridge) with regularization parameter $\alpha = 3.0$. Lastly, three non-linear regression models are considered: random forest regression (Random Forest) (see Breiman, 2001), an ensemble of $n = 10$ regression trees, aggregated via bagging, with bootstrap sampling and randomized feature selection at each split; extreme gradient boosting (XGBoost) (see Chen and Guestrin, 2016), a gradient-boosted decision tree ensemble with $n = 30$, learning rate $\eta = 0.05$, maximum depth $d = 4$, subsampling ratio 0.8, and column subsampling ratio 0.8 at each tree; Gaussian process regressor (GP) (e.g., MacKay, 1992), using a kernel composed of a constant term and a radial basis function (RBF). The kernal takes the form

$$\mathsf{Kernel}(x, x^{i}) = C \exp\left(-\frac{1}{2r^2}\|x - x^{'}\|_2^2\right),$$

where C controls the vertical scale of the function, and r, the length scale, controls the smoothness of sample paths; $\|\cdot\|_2^2$ indicates the usual squared euclidean distance between input vectors. Prior to model fitting, the response variable y is normalized to zero mean prior to model fitting. We evaluate the performance of the different methods by computing the MedAE.

Discussion of the results. Figure 4.4 shows the MedAE (reported on logarithmic scale) as a function of iterations on the validation data for both regression tasks: heating load (test 1) and cooling load (test 2). Shaded areas represent the interquartile range ($25^{th} - 75^{th}$ percentile) of the MedAE across repetitions. For both tasks, DA-GR with $J \in \{2, 5, 10\}$ demonstrates clear performance improvements over the standard ReLU-based FFNN. The proposed methods achieve strong validation performances, characterized by fast convergence and consistently lower prediction errors throughout. Notably, increasing the number of experts K tends to yield smaller validation MedAE, suggesting that a finer partitioning of the input space potentially allows experts to focus more effectively on their specialized regions. This observation motivates careful tuning of K to balance model complexity and predictive accuracy. Across iterations, the proposed methods remain competitive with the best-performing baseline regressors (OLS, Ridge, Random Forest, XGBoost, and GP).

4.6.3 Regression task on Boston Housing data

The Boston Housing dataset comprises $n = 506$ observations, each corresponding to a distinct urban area within the city of Boston. For each observation, $p = 13$ covariates are recorded, capturing a variety of ecological and socio-economic characteristics. The ecological variables include nitrogen oxide concentration (NOX), levels of particulate matter (PART), and proximity to the Charles River (CHAS). Socio-economic indicators include the proportion of Black residents (B), property tax rate (TAX), percentage of lower-status population (LSTAT), crime rate (CRIM), pupil-to-teacher ratio (PTRATIO), proportion of older buildings (AGE), average number of rooms per dwelling (RM), proportion of residential land zoned for large lots (ZN), weighted distance to employment centers (DIS), and an index of accessibility to radial highways (RAD). The target variable is the median housing value in each area. All covariates are standardized by centering to zero mean and scaling to unit variance. The transformation is applied using the mean and variance estimated from the training data, which are then used to standardize the validation sets.

Choice of hyper-parameters and architecture. We use the decision tree classifier as gating network, and we implement our method with a number of experts K either equals to 6 or 10, each one comprising two expert hidden layers ($n_{\text{expert}} = 2$); the number of neurons p_2 is either 6 (when $K = 6$) or 4 (when $K = 10$), whereas $p_1 = 1$. Building on the observed benefits of incorporating a probabilistic ReLU expert layer in the Concrete Compressive Strength experiments in Section 4.6.1, we adopt a Gaussian expert layer as the first hidden layer and a probabilistic ReLU layer as the second. For each expert $k = 1, \dots, K$, weights and biases are denoted as usual by $\dot{\boldsymbol{W}}_1^{[k]} = (\boldsymbol{b}_1^{[k]}, \boldsymbol{W}_1^{[k]}) \in \mathbb{R}^{p_1 \times (1+p)}$ and $\dot{\boldsymbol{W}}_2^{[k]} = (\boldsymbol{b}_2^{[k]}, \boldsymbol{W}_2^{[k]}) \in \mathbb{R}^{p_2 \times (1+p_1)}$ and learned via MFVB; instead, the augmented variables are updated via semiparametric MFVB. The top layer is a Gaussian regression with parameters $(b_0^{[k]}, \boldsymbol{w}_0^{[k]}, (\sigma_0^2)^{[k]})$, learned, again, via MFVB. As for diabetes data, $n_{\text{global}} = 0$. We hold out 30% of the dataset as the validation set, and we perform 20 simulation runs by using different sub-sampling of the dataset.

Initialization. We initialize our BNN in the following way. Experts assignments are first obtained via K-MEANS clustering using *both* input and output data. A decision tree classifier $\boldsymbol{T}$ is then fitted to reproduce these experts assignments from the input data alone, yielding disjoint expert assignments $C_1, \dots, C_K$.

Two configurations $(K = 6, p_2 = 6, p_1 = 4)$ and $(K = 10, p_2 = 6, p_3 = 3)$ are initialized similarly. For each expert k, a Gaussian expert layer with p_2 neurons is followed by a probabilistic ReLU expert layer with p_1 neurons. The Gaussian expert layer parameters:

$\boldsymbol{z}_{i,2}^{[k]} \in \mathbb{R}^{p_2}$ for all $i \in C_k$, $\dot{\boldsymbol{W}}_2^{[k]}$ are initialized exactly as before. The probabilistic ReLU layer is initialized as follows. $\boldsymbol{z}_{i,1}^{[k]} \in \mathbb{R}^{p_1}$ is initialized by sampling from an inverse-gamma distribution. The weights and biases $\dot{\boldsymbol{W}}_1{}^{[k]}$ is initialized as the multivariate multiple linear regression coefficients from $\{\boldsymbol{z}_{i,2}^{[k]}\}_{i \in C_k}$ to $\{\boldsymbol{z}_{i,1}^{[k]}\}_{i \in C_k}$, while the variance parameter $(\sigma_1^2)^{[k]}$ is initialized as 1.0. The output layer weights and biases $\dot{\boldsymbol{W}}_0^{[k]}$ are drawn from a zero-mean Gaussian distribution with small variance, and the variance parameter $(\sigma_0^2)^{[k]}$ is set to 1.0.
Based on these initializations, the variation means required in MFVB and semiparametric MFVB are obtained by taking empirical means over the corresponding initialized quantities. The weight and bias $\dot{\boldsymbol{W}}_2$ are initialized using a multivariate multiple linear regression, whereas the weight and bias $\dot{\boldsymbol{W}}_1$ are set to the coefficients obtained from a standard linear regression.

Comparison with other methodologies. We compare the performance of our methodology against several baseline regression frameworks. First, we consider a classical feedforward neural network (FFNN) with ReLU activation functions consisting of a single hidden layer with 64 units and a dropout rate of 0.5 (see Srivastava et al., 2014). Second, we include the data augmentation Gaussian regression (DA-GR) model proposed by Wang et al. (2022) (see Section 4.1). The implementation employs a single hidden-layer ReLU network with 64 neurons and a dropout rate of 0.5 (see Srivastava et al., 2014). We consider numbers of MCMC copies $J \in \{2, 5, 10\}$, with stochastic noise parameters in the top and hidden layers set to $\tau_0 = 0.4$ and $\tau_z = 0.5$, respectively, selected via grid search. The configurations in FFNN and DA-GR are matched, allowing for a consistent comparison.
We also implemented two linear regression models: ordinary least square (OLS) and ridge regression (Ridge) with regularization parameter $\alpha = 3.0$. Lastly, three non-linear regression models are considered: random forest regression (Random Forest) (see Breiman, 2001), an ensemble of $n = 10$ regression trees, aggregated via bagging, with bootstrap sampling and randomized feature selection at each split; extreme gradient boosting (XGBoost) (see Chen and Guestrin, 2016), a gradient-boosted decision tree ensemble with $n = 30$, learning rate $\eta = 0.05$, maximum depth $d = 4$, subsampling ratio 0.8, and column subsampling ratio 0.8 at each tree; Gaussian process regressor (GP) (e.g., MacKay, 1992), using a kernel composed of a constant term and a radial basis function (RBF). The kernal takes the form

$$\mathsf{Kernel}(\boldsymbol{x}, \boldsymbol{x}^{i}) = C \exp\left(-\frac{1}{2r^2} \|\boldsymbol{x} - \boldsymbol{x}^{'}\|_2^2\right),$$

where C controls the vertical scale of the function, and r, the length scale, controls the smoothness of sample paths; $\| \cdot \|_2^2$ indicates the usual squared euclidean distance between input vectors. Prior to model fitting, the response variable y is normalized to zero mean prior to model fitting.

We evaluate the performance of the different methods by computing the MedAE.

Discussion of the results. Figure 4.5 displays the MedAE on validation data (reported in logarithmic scale) as a function of the iteration number for all the methods described in the previous paragraph. The shaded areas indicate the interquartile range (25th–75th percentile) of the MedAE across repetitions. The figure shows that DA-GR outperforms the ReLU-based FFNN, both in terms of prediction accuracy and convergence speed, particularly during the early iterations. It should be noted that in their study, Wang et al. (2022) compares their DA-GR methodology exclusively to ReLU FFNN. In contrast, our results demonstrate the proposed method achieves excellent performance from the very first iterations. Furthermore, increasing the number of experts leads to lower validation MedAE, reinforcing the potential benefits of employing finer expert segmentations. Finally, our methods remains comparable to the other baseline regressors.

4.6.4 Classification task on Wine Quality data

The Wine Quality data, which encompasses the physicochemical properties of white and red variants of Portuguese “Vinho Verde" wine, consists of $n = 4898$ and $n = 1599$ observations and $p = 11$ continuous covariates. The frequency of each quality is summarized as in Table 4.1. These covariates include measures of fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol. Each covariate is centred to zero mean and scaled to unit variance. The target variable is the wine quality, is a score between 0 and 10 and it is based on sensory data (median of at least 3 evaluations made by wine experts).

In order to run the classification task, we provide four types of classifications based on these two different wine quality data: *(i)* white wine with quality score of 5 or 6 (Test 1); *(ii)* white wine with quality score of ≤ 5 or > 5 (Test 2); *(iii)* white wine with quality score of 6 or 7 (Test 3); *(iv)* red wine with quality score of ≤ 5 or > 5 (Test 4). Test 1 and Test 2 replicate the classification settings from Polson et al. (2013), where the class labels are relatively balanced, providing a standard benchmark to assess classification performance under symmetric class distributions. In contrast, Test 3 introduces a more challenging scenario based on white wine quality scores: it distinguishes between quality levels 6 and 7, with observations $2,198$ and 880, respectively. This results in a significant class imbalance, with approximately 71.4% of the observations belonging to the majority class. The inclusion of Test 3 allows us to evaluate the robustness of each method when handling imbalanced data, an important consideration in many real-world applications where class distributions are inherently skewed. Test 4, based on red wine quality data, defines a binary classification task with more balanced class sizes. It serves to complement Test 3 by providing a second wine-related dataset with a different distributional structure. The inclusion of both Tests 3 and 4 allows us to assess the generalization ability of our method across datasets that differ not only in the domain but also in the balance and structure of the response variable. This contributes to a more comprehensive evaluation of the models' classification performance under both balanced and imbalanced conditions.

Choice of hyper-parameters, architecture and initialization. We use Gaussian mixture models as the gating networks in all four tests. In the first three tests with white wine quality data, we implement our method with a number of experts K either equals to 6 or 14, and a Gaussian expert hidden layer ($n_{\text{expert}} = 1$) with p_1 either equals to 6 or 8 neurons. In Test 4, we implement our method with number of experts K either equals to 6 or 10, and a Gaussian expert hidden layer with p_1 either equals to 9 or 6.

We consider a binary regression model with a logit link function, leveraging data augmentation, as in Polson et al. (2013), to facilitate Bayesian inference. We hold out 20% of the training set as the validation set, and we perform 20 simulation runs by using different sub-sampling of the training set. We initialize our BNN in the following way. First, experts assignments are performed via Gaussian mixture models with the number of clusters equals to the number of experts K. For each expert k, a PCA with p_1 components is used to initialize $\boldsymbol{z}_i^{[k]} \in \mathbb{R}^{p_1}$, $i \in C_k$; in the subsequent iterations, these are updated using semiparametric MFVB. The initial values for the weight and bias $\dot{\boldsymbol{W}}_1^{[k]} = (\boldsymbol{b}_1^{[k]}, \boldsymbol{W}_1^{[k]}) \in \mathbb{R}^{p_1 \times (1+p)}$ are obtained via a multivariate multiple linear regression, and then learned through MFVB. The noise variance $(\sigma_1^2)^{[k]}$ is estimated via MFVB. For the top layer, we estimate a binary regression model with logit link using a few iterations of the Pólya-Gamma data augmentation scheme of Polson et al. (2013) to approximate the posterior mode.

Comparison with other methodologies. We compare the performance of our methodology against several widely used regression frameworks. These include a classical ReLU-based FFNN, tested with both a single hidden layer of 64 units and a deeper architecture with two hidden layers containing 64 and 16 units, respectively. We also consider a Support Vector Machine (SVM) with a linear kernel, and a decision tree classifier with a maximum tree depth set to 16. To evaluate and compare the predictive accuracy of the methods, we compute both the misclassification rate and the F_1 score (Van Rijsbergen, 2004), defined as:

$$\mathrm{F}_1 = \left(\frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \right) \times 100\%,$$

where precision is given by $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$ and recall by $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$. Here, TP, FP, and FN denote the number of true positives, false positives (i.e., incorrectly selected fixed effects), and false negatives (i.e., relevant fixed effects that were not selected), respectively. The F_1 score ranges from 0% to 100%, with higher values indicating better classification performance.

Discussion of the results. Figures 4.6–4.9 display the validation median of misclassification rates and F_1 scores for Test1–Test4. We also report the 25-75% quantile range of the misclassification rates and F_1 scores. Overall, the figures consistently demonstrate that our method outperforms the competing models across multiple settings. Moreover, we notice that in both Test 1 and Test 2, the performance of our model closely matches that of the DA-SVM and DA-logit classification models proposed in Wang et al. (2022) (see Figure 5 in their paper). In terms of F_1 score, our model performs consistently well across all settings; in particular, it achieves strong results in Test 3 in Figure 4.8, where the classification task is notably imbalanced. Furthermore, our MoE framework allows expert-specific performance evaluation, offering the potential to further enhance classification accuracy by refining or updating underperforming experts, as illustrated in Figure 4.10, where we computed each expert's misclassification rate and F_1 score.

4.7 Conclusion and discussion

In this work, we proposed a novel and flexible Bayesian learning framework for deep neural networks, grounded in a Mixture of Experts (MoE) architecture and trained via semiparametric Mean Field Variational Bayes (MFVB). Our model leverages both data augmentation techniques and probabilistic activation functions (e.g., probabilistic ReLU) to offer a computationally efficient alternative to standard Bayesian neural networks, while maintaining interpretability and adaptability across different learning tasks.

Through empirical evaluation on diverse datasets—including regression tasks on the Concrete Compressive Strength data, Energy Efficiency data, and Boston Housing datasets, as well as a classification task on the Wine Quality dataset, we demonstrated the strong predictive performance of our method compared to several well-established baselines. These include traditional regression models (OLS, Lasso, Ridge), advanced ensemble methods (Random

Forest, XGBoost), Bayesian approaches (DA-GR), and deep learning architectures (FFNNs, Gaussian Processes). Across all tasks, our model exhibited faster convergence, competitive or superior accuracy, and improved robustness, especially in early training iterations.

An important takeaway is the versatility of our approach: MoE design combined with efficient Bayesian inference allows our method to scale across tasks of varying complexity while remaining relatively simple to tune and initialize. In particular, our initialization strategies (based on PCA and multivariate regressions) and the use of gating networks (such as k-means clustering and decision trees) contribute to the model's overall efficiency and interpretability.

Future work may explore the extension of this framework to other problem domains such as time series forecasting, structured data (e.g., graphs), or semi-supervised settings. Additionally, incorporating more expressive gating mechanisms or adaptive selection of the number of experts could further enhance performance and flexibility.

In conclusion, our contribution lies in bridging the gap between classical variational inference and modern deep learning architectures by developing a scalable, interpretable, and statistically principled method for Bayesian deep learning.

Algorithm 2: *Algorithm for training a BNN with a single neuron in an expert hidden layer and a ReLU probabilistic activation function via a decision tree classifier with hard gating MoE. The algorithm stops when the change in* $\mathsf{ELBO}_{\mathfrak{q}}(\mathcal{D}_{C_k}; \dot{\boldsymbol{w}}^{[k]}, (\sigma^2)^{[k]}, \boldsymbol{z}^{[k]}, \alpha^{[k]}, \beta^{[k]})$ *in* (4.4.7) *is negligible.*

Data Input: Training dataset $\mathcal{D}_n$, number of experts K.

Initialize:

- Experts assignments by K-means clustering using both input and output data.
- The the decision tree classifier $\boldsymbol{T}^{(0)}$.
- The probabilities of belonging to each expert k, i.e., π_{ik} for each $i = 1, \ldots, n$ and $k = 1, \ldots, K$.
- The statistics of the variational densities $\mathfrak{q}^\star(\dot{\boldsymbol{w}}^{[k]}), \mathfrak{q}^\star((\sigma^2)^{[k]})$, $\tilde{\mathfrak{q}}(z_i; \alpha^{[k_i]}, \beta^{[k_i]})$ for all observations $i = 1, \ldots, n$; $i \in C_{k_i}$ and $k_i \in \{1, \ldots, K\}$ is representing that the i^{th} observation is assigned to the k_i^{th} expert.

Cycle until convergence:

- Assign the training dataset $\mathcal{D}_n$ to K experts by the decision tree $\boldsymbol{T}^{(t-1)}$. The k-th clustered data, for each $k = 1, \ldots, K$, is denoted by $\mathcal{D}_{C_k} = \{(\boldsymbol{x}_i, y_i) : i \in C_k\}$, which collects all observations for the k^{th} expert.
- For each cluster C_k, $k = 1, \ldots, K$ update the parameters via semi-parametric mean-field variational Bayes by updating the tractable parameters $(\dot{\boldsymbol{w}}, \sigma^2)$ posteriors by mean-field variational approximations and the k^{th} experts' parameters by parametric variational approximations. First, we solve the following optimization problem:

$$(\alpha^{[k]}, \beta^{[k]})^{(t)} = \arg \max_{\alpha^{[k]} > 2, \beta^{[k]} > 0} \mathsf{ELBO}_{\mathfrak{q}}(\mathcal{D}_{C_k}; \dot{\boldsymbol{w}}^{[k]}, (\sigma^2)^{[k]}, \boldsymbol{z}^{[k]}, \alpha^{[k]}, \beta^{[k]}), \quad k = 1, \ldots, K,$$

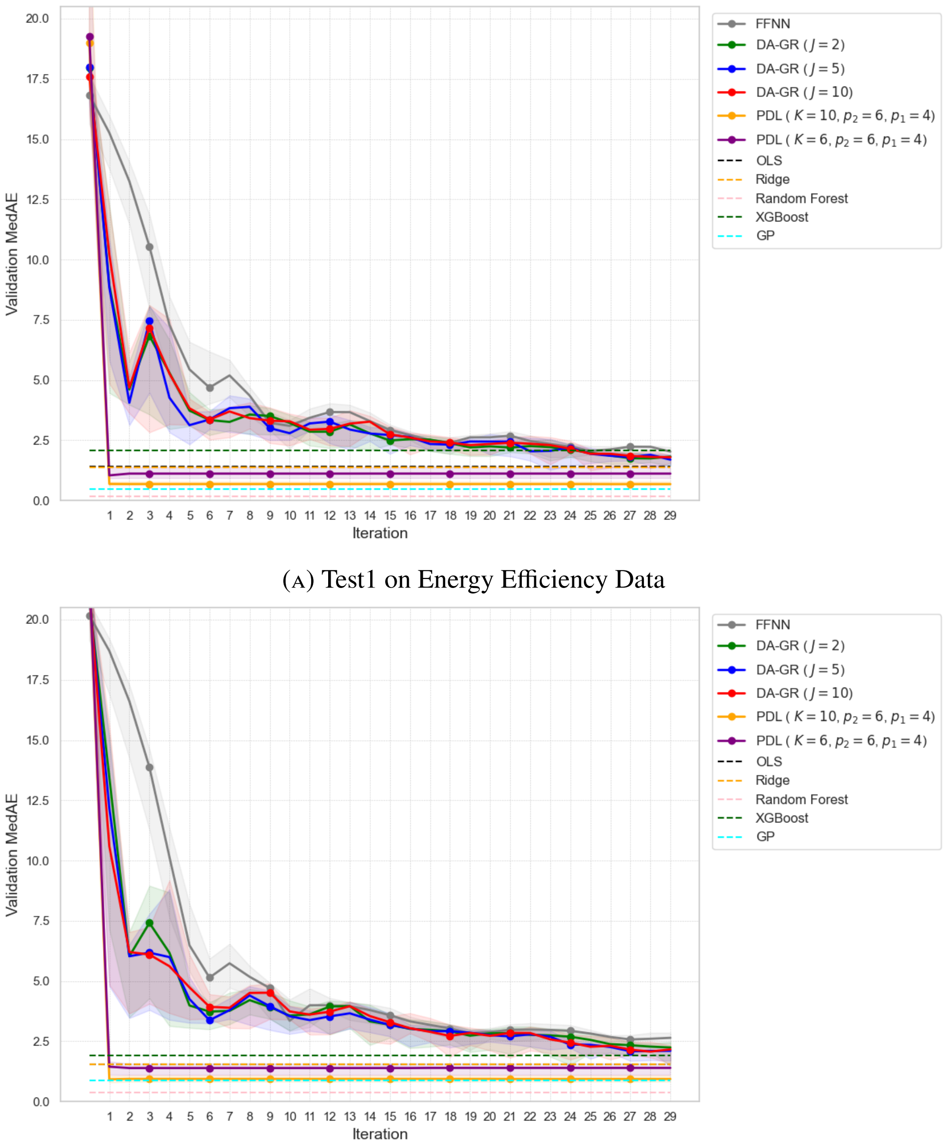
  where $\mathsf{ELBO}_{\mathfrak{q}}(\mathcal{D}_{C_k}; \dot{\boldsymbol{w}}^{[k]}, (\sigma^2)^{[k]}, \boldsymbol{z}^{[k]}, \alpha^{[k]}, \beta^{[k]})$ is defined in Equation (4.4.7). Then, update $\mathfrak{q}^\star(\dot{\boldsymbol{w}}^{[k]})$ according to Equation (4.4.4) and $\mathfrak{q}((\sigma^2)^{[k]})$ according to Equation (4.4.5) by using the updated $\tilde{\mathfrak{q}}(z_i; (\alpha^{[k_i]}, \beta^{[k_i]}))$.
- Update the decision tree classifier gating network. First, update responsibilities $\boldsymbol{\tau}_i = (\tau_{i1}, \tau_{i2}, \ldots, \tau_{iK})$ according to Equation (4.3.10), where

$$\begin{aligned} & f_k(y_i, \boldsymbol{x}_i; z_i, \boldsymbol{w}^{[k]}, b^{[k]}, (\sigma^2)^{[k]}) \\ & \propto \exp\left\{-\frac{1}{2(\sigma^2)^{[k]}}(y_i - z_i)^2\right\} \times \frac{1}{\sqrt{2\pi z_i}} \exp\left\{-\frac{1}{2z_i}((\boldsymbol{w}^{[k]})^\top \boldsymbol{x}_i + b^{[k]} + z_i)^2\right\} \times \left(\frac{1}{(\sigma^2)^{[k]}}\right)^{1/2}. \end{aligned}$$

  Assign the i^{th} observation to the k^{th} category with probability τ_{ik} for each $i = 1, \ldots, n$ and $k = 1, \ldots, K$, if the algorithm is not "stationary". In this case, let $\Gamma_{ik} \sim \mathsf{Categorical}(\tau_{i1}, \tau_{i2}, \ldots, \tau_{iK})$. Otherwise, assign the i^{th} observation to the k^{th} category with the greatest responsibility τ_{ik} in Equation (4.3.10), i.e.,

$$\Gamma_{ik} = \begin{cases} 1 & \text{if } k = \arg\max_{j=1,\ldots,K} \tau_{ij}, \\ 0 & \text{otherwise.} \end{cases} .$$

  Then, refit the decision tree classifier $\boldsymbol{T}^{(t)}$ based on this updated assignment in the training dataset. Finally, update the mixing coefficients $\boldsymbol{\pi}$.



(A) Test1 on Energy Efficiency Data

(B) Test2 on Energy Efficiency Data

FIGURE 4.4: Regression tasks on Energy Efficiency Data. We report the median absolute error (MEDAE; and the corresponding variability) as a function of iterations on validation data for the following methods: feedforward neural network with ReLU activation (FFNN) (gray solid line); DA-GR with $J = 2$ (green solid line), $J = 5$ (blue solid line), $J = 10$ (red solid line); our probabilistic deep learning (PDL) Bayesian model with two different architectures (orange solid line and purple solid line); ordinary least square (OLS) regression (black dashed line), ridge regression (yellow dashed line), random forest regression (pink dashed line), extreme gradient boosting (XGBoost; green dashed line), and Gaussian process (GP) regressor (sky-blue dashed line).

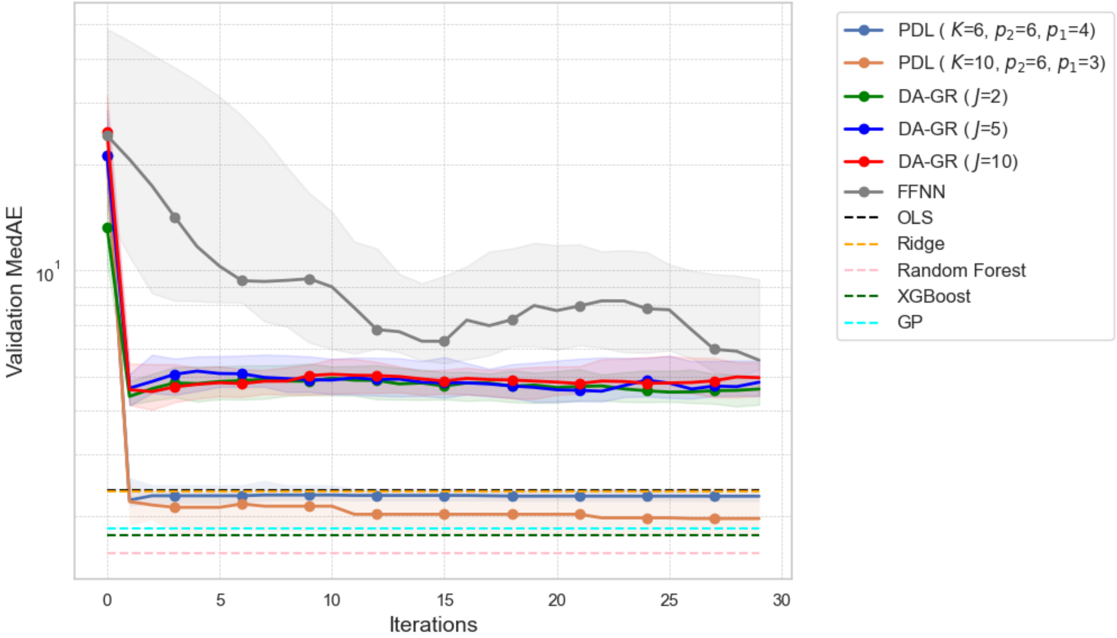


FIGURE 4.5: Regression task on Boston Housing Data. We report the median absolute error (MEDAE; and the corresponding variability) as a function of iterations on validation data for the following methods: feedforward neural network with ReLU activation (FFNN) (gray solid line); DA-GR with $J = 2$ (green solid line), $J = 5$ (blue solid line), $J = 10$ (red solid line); our probabilistic deep learning (PDL) Bayesian model with two different architectures (dark-blue solid line and orange solid line); ordinary least square (OLS) regression (gray dashed line), ridge regression (orange dashed line), random forest regression (pink dashed line), extreme gradient boosting (XGBoost; green dashed line) and Gaussian process (GP) regressor (sky-blue dashed line).

TABLE 4.1: Comparison of white and red wine quality distributions by score.

Score	white wine	red wine
3	20	10
4	163	53
5	1457	681
6	2198	638
7	880	199
8	175	18
9	5	0
Total	4898	1599

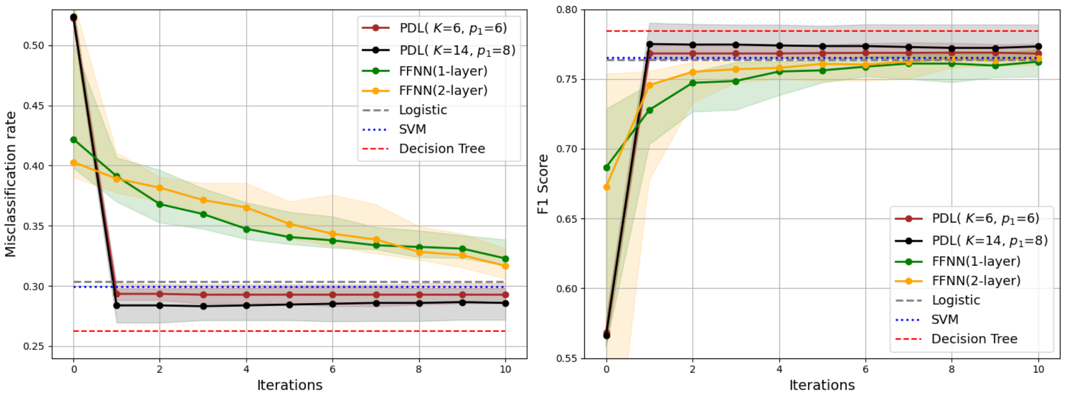


FIGURE 4.6: Misclassification rate (left panel) and F1-score (right panel) comparison for the Wine quality validation performance (Test 1).

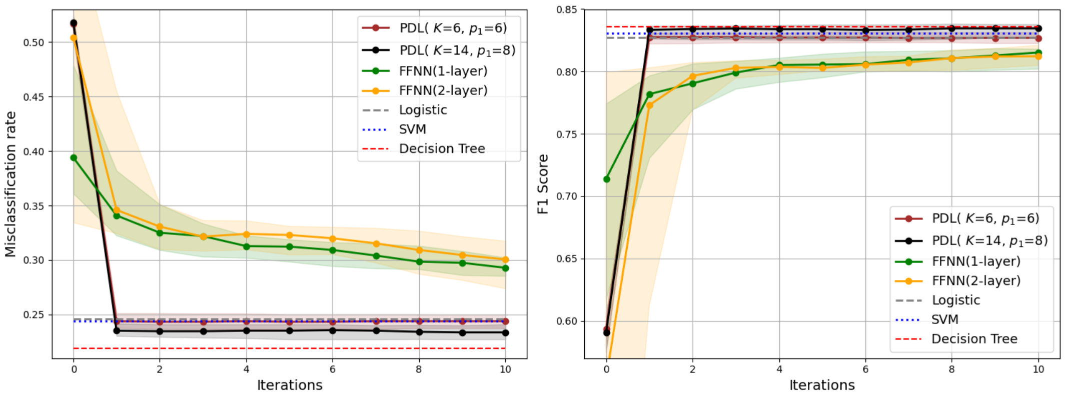


FIGURE 4.7: Misclassification rate (left panel) and F1-score (right panel) comparison for the Wine quality validation performance (Test 2).

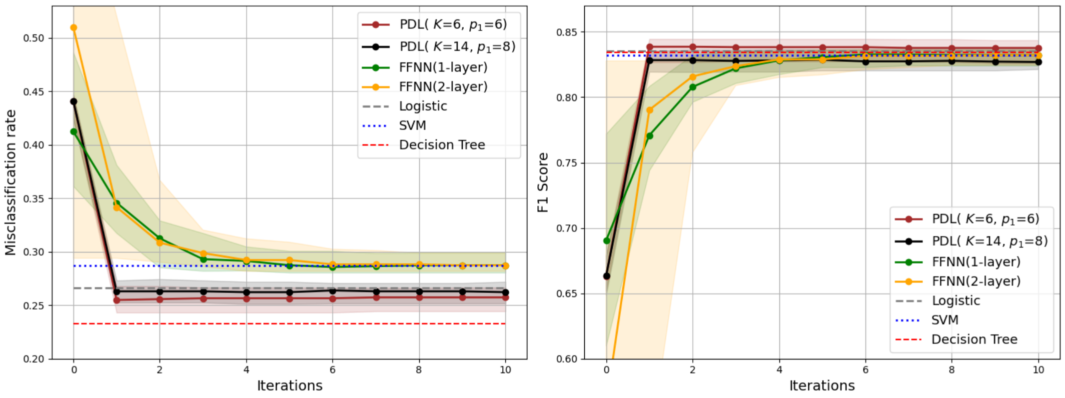


FIGURE 4.8: Misclassification rate (left panel) and F1-score (right panel) comparison for the Wine quality validation performance (Test 3).

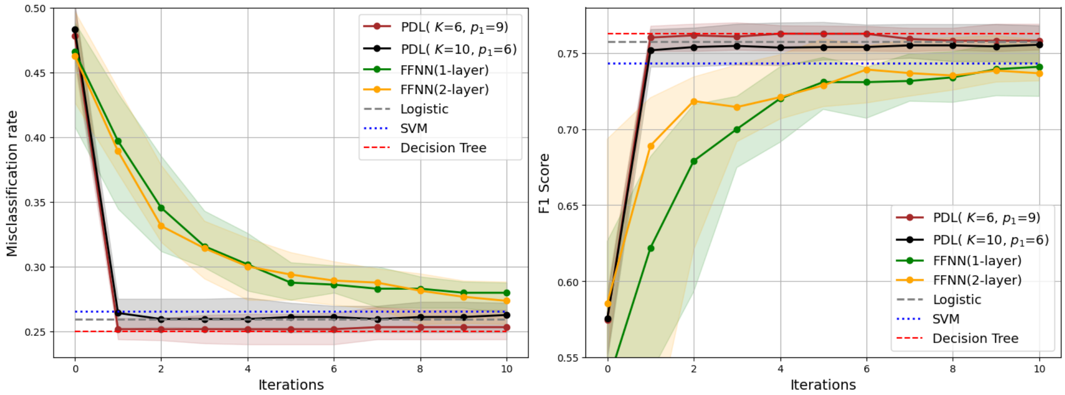


FIGURE 4.9: Misclassification rate (left panel) and F1-score (right panel) comparison for the Wine quality validation performance (Test 4).

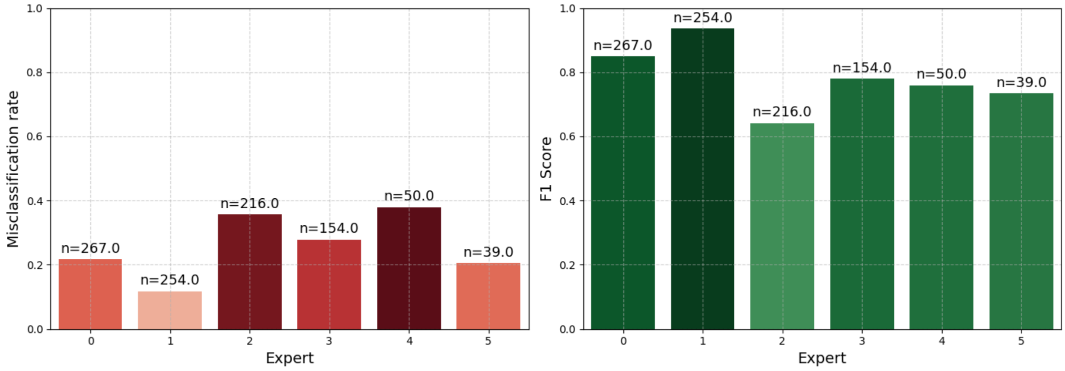


FIGURE 4.10: Per expert misclassification rate and F_1 score in Test 2 for $K = 6$ and $p_1 = 6$.

Appendix

4.A List of probability distributions and their properties

This section describes some probability distributions used in the main text and their properties.

4.A.1 Generalized Inverse Gaussian Distribution and Inverse Gaussian Distribution

A continuous random variable x has a Generalized Inverse Gaussian distribution with parameters $p \in \mathbb{R}, a > 0$ and $b > 0$, written $x \overset{\text{d.}}{\sim} \mathsf{GIG}(p, a, b)$, if the density function of x is

$$\mathfrak{p}(x \mid p, a, b) \propto x^{p-1} \exp\left\{-\frac{1}{2}\left(\frac{b}{x} + ax\right)\right\}, \quad x > 0.$$

A continuous random variable x has a Inverse Gaussian distribution with parameters $\mu > 0, \lambda > 0$, written $x \overset{\text{d.}}{\sim} \mathsf{IG}(\mu, \lambda)$, if the density function of x is

$$\mathfrak{p}(x \mid \mu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} x^{p-1} \exp\left\{-\frac{1}{2}\left(\frac{\lambda(x-\mu)^2}{\mu^2 x}\right)\right\}, \quad x > 0.$$

When $p = 1/2$, $a = \lambda/\mu^2$ and $b = \lambda$, $x \overset{\text{d.}}{\sim} \mathsf{GIG}(p, a, b)$ iff $x^{-1} \overset{\text{d.}}{\sim} \mathsf{IG}(\mu, \lambda)$. Moreover, using results in Devroye (1986, Page 479, Lemma 7.4 C) it is possible to show these additional equivalences between distributions: $x \overset{\text{d.}}{\sim} \mathsf{GIG}(p, a, b)$ iff $x^{-1} \overset{\text{d.}}{\sim} \mathsf{GIG}(-p, b, a)$ and

$$x \overset{\text{d.}}{\sim} \mathsf{GIG}\left(\frac{1}{2}, a, b\right) \text{ iff } x^{-1} \overset{\text{d.}}{\sim} \mathsf{IG}\left(\sqrt{\frac{a}{b}}, a\right). \tag{4.A.1}$$

4.A.2 Pólya-Gamma

A random variable x has a Pólya-Gamma distribution (Polson et al., 2013, Definition 1) with parameters $b > 0$ and $c \in \mathbb{R}$, written $x \overset{\text{d.}}{\sim} \mathsf{PG}(b, c)$, if

$$x \overset{\text{d.}}{\sim} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k-1/2)^2 + c^2/(4\pi^2)}, \tag{4.A.2}$$

where the $g_k \overset{\text{ind.}}{\sim} \mathsf{Gamma}(b, 1)$ are independent Gamma random variables.

By setting $c = 0$, $g_k \overset{\text{ind.}}{\sim} \mathsf{Gamma}(b, 1)$ still holds and the Laplace transform of $x \overset{\text{d.}}{\sim} \mathsf{PG}(b, 0)$ is given by

$$\mathsf{E}\left(e^{-xt}\right) = \prod_{k=1}^{\infty} \left(1 + \frac{t}{2\pi^2(k - 1/2)^2}\right)^{-b} = \cosh^{-b}(\sqrt{t/2}), \tag{4.A.3}$$

where the last equivalence follows from application of the Weierstrass factorisation of $\cosh(x)$, that is

$$\cosh(x) = \prod_{k=1}^{\infty} \left(1 + \frac{x^2}{\pi^2(k - 1/2)^2}\right).$$

Also, notice that when $b = 1$ and $c = 0$, $\mathsf{PG}(1, 0)$ is the subset of the class of infinite convolutions of exponentials, since $g_k \overset{\text{ind.}}{\sim} \mathsf{Gamma}(1, 1)$.

For general values of b and c, the $\mathsf{PG}(b, c)$ distribution can be written as an exponential tilting of the $\mathsf{PG}(b, 0)$ distribution with density function

$$\mathfrak{p}(x \mid b, c, w) = \frac{\exp\left(-\frac{c^2}{2}x\right) \mathfrak{p}(x \mid b, 0)}{\mathsf{E}\left\{\exp\left(-\frac{c^2}{2}\omega\right)\right\}}, \quad x > 0, \tag{4.A.4}$$

where $\omega \overset{\text{d.}}{\sim} \mathsf{PG}(b, 0)$. The Laplace transform of $x \overset{\text{d.}}{\sim} \mathsf{PG}(b, c)$ can be computed using (4.A.3) and given by

$$\mathsf{E}\left(e^{-tx}\right) = \prod_{k=1}^{\infty} (1 + d_k^{-1} t)^{-b} \quad \text{with} \quad d_k = 2\left(k - \frac{1}{2}\right)^2 \pi^2 + \frac{c^2}{2}. \tag{4.A.5}$$

Each term in the product equals the Laplace transform of a Gamma distribution with shape parameter b and scale parameter d_k^{-1}. From the scaling property of the Gamma distribution it follows that

$$x \overset{\text{d.}}{\sim} \sum_{k=1}^{\infty} \mathsf{Gamma}(b, d_k^{-1}) = \sum_{k=1}^{\infty} \frac{\mathsf{Gamma}(b, 1)}{d_k} = \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{\mathsf{Gamma}(b, 1)}{(k - 1/2)^2 + c^2/(4\pi^2)},$$

where the right-hand side corresponds to (4.A.2).

Finally, the expectation of a Pólya-Gamma random variable is given by

$$\mathsf{E}(x) = \frac{b}{2c} \tanh\left(\frac{c}{2}\right) = \frac{b}{2c}\left(\frac{e^c - 1}{1 + e^c}\right), \tag{4.A.6}$$

which is obtained by differentiating the left-hand side of (4.A.5) with respect to t and evaluating the derivative at zero.

4.A.3 Inverse Gamma

A continuous random variable x has an Inverse Gamma distribution with parameters $\kappa > 0$ and $\lambda > 0$, written $x \overset{\text{d.}}{\sim} \mathsf{Inv\text{-}Gamma}(\lambda, \kappa)$, if the density function of x is

$$\mathfrak{p}(x \mid \lambda, \kappa) = \frac{\kappa^{\lambda}}{\Gamma(\lambda)} x^{-(\lambda+1)} \exp\{-\frac{\kappa}{x}\}, \quad x > 0.$$

The inverse Gamma distribution is a special case of the generalized inverse Gaussian (GIG) family. In particular, under the standard GIG parametrisation, $\mathsf{Inv\text{-}Gamma}(\lambda, \kappa)$ corresponds to a $\mathsf{GIG}(-\lambda,\, 2\kappa,\, 0)$ distribution.

Also,

$$\begin{gathered} \mathsf{E}(x) = \tfrac{\kappa}{\lambda-1}, \quad \mathsf{E}\left(\tfrac{1}{x}\right) = \tfrac{\lambda}{\kappa}, \\ \mathsf{E}(x^2) = \kappa^2 \tfrac{\Gamma(\lambda-2)}{\Gamma(\lambda)} = \tfrac{\kappa^2}{(\lambda-1)(\lambda-2)}, \quad \mathsf{E}\{\log(x)\} = \log(\kappa) - \psi(\lambda), \end{gathered} \tag{4.A.7}$$

and

$$\mathsf{Entropy}(x; \lambda, \kappa) = \lambda + \log(\kappa) + \log\{\Gamma(\lambda)\} - (\lambda+1)\psi(\lambda), \tag{4.A.8}$$

where $\psi(\cdot)$ is the digamma function.

4.B Useful Integral Identities

This section presents some integral identities that are used in our Bayesian learning procedure of DNNs.

The first integral identity (see Andrews and Mallows, 1974, Equation (2.2)) is useful in data augmentation strategies involving the ReLu activation function:

$$\int_0^\infty \exp\left\{-\frac{1}{2}\left(a^2u^2+\frac{b^2}{u^2}\right)\right\}\,\mathrm{d}u = \sqrt{\frac{\pi}{2a^2}}\exp(-|ab|).$$

Introducing the change of variable $\omega = u^2$, the integral above becomes

$$\sqrt{\frac{\pi}{2a^2}}\int_0^\infty \frac{a}{\sqrt{2\pi\omega}}\exp\left\{-\frac{1}{2}\left(a^2\omega+\frac{b^2}{\omega}\right)\right\}\,\mathrm{d}\omega = \sqrt{\frac{\pi}{2a^2}}\exp(-|ab|). \tag{4.B.1}$$

This can be related to the family of ReLu activation functions as follows:

$$\begin{aligned}
&\int_0^\infty \frac{1}{\sqrt{2\pi\omega}}\exp\left\{-\frac{1}{2\omega}(x-1-\omega)^2\right\}\,\mathrm{d}\omega\\
&\quad=\left[\int_0^\infty \frac{1}{\sqrt{2\pi\omega}}\exp\left\{-\frac{1}{2}\left(\omega+\frac{(x-1)^2}{\omega}\right)\right\}\,\mathrm{d}\omega\right]\exp(x-1)\\
&\quad=\exp(-|x-1|)\exp(x-1)\\
&\quad=\exp\left[-\{|1-x|+(1-x)\}\right]\\
&\quad=\exp\{-2\max(1-x,0)\},
\end{aligned}$$

where the second equality is obtained by setting $a = 1$ and $b = x - 1$ in equation (4.B.1) and and the last equality follows from the identity $\max(u, 0) = \frac{1}{2}(|u| + u)$. The integral in (4.B.2) can be generalized in the following way:

$$\begin{aligned}
&\int_0^\infty \frac{a}{\sqrt{2\pi c\omega}}\exp\left\{-\frac{1}{2c\omega}(x+a\omega)^2\right\}\,\mathrm{d}\omega\\
&\quad=\left[\int_0^\infty \frac{\frac{a}{\sqrt{c}}}{\sqrt{2\pi\omega}}\exp\left\{-\frac{1}{2}\left(\frac{x^2}{c\omega}+\frac{a^2}{c}\omega\right)\right\}\,d\omega\right]\exp\left(-\frac{ax}{c}\right)\\
&\quad=\exp\left\{-\left(\left|\frac{ax}{c}\right|+\frac{ax}{c}\right)\right\}\\
&\quad=\exp\left\{-\frac{2}{c}\max(ax,0)\right\},
\end{aligned} \tag{4.B.2}$$

where the second equality is obtained by replacing a and b with $a/\sqrt{c}$ and $x/\sqrt{c}$ respectively in equation (4.B.1).

Another useful integral identity is instead related to the Pólya-Gamma distribution (see Subsection 4.A.2) $\mathsf{PG}(b,0)$ and can be found in Polson et al. (2013, Theorem 1). For all $a \in \mathbb{R}$,

$$2^{-b} e^{\kappa\psi} \int_0^\infty \exp\left(-\frac{\omega\psi^2}{2}\right) \mathfrak{p}(\omega \mid b,0)\, \mathrm{d}\omega = \frac{(\exp(\psi))^a}{(1+\exp(\psi))^b}, \tag{4.B.3}$$

where $\kappa = a - \frac{b}{2}$ and $\mathfrak{p}(\omega \mid b,0)$ is the density function of a $\mathsf{PG}(b,0)$. Moreover, since the integrand in Equation (4.B.3) can be seen as a non-normalised joint density in (ω,ψ), the conditional distribution $(\omega \mid \psi)$ with density

$$\mathfrak{p}(\omega \mid \psi) = \frac{\exp\left(-\frac{\omega\psi^2}{2}\right) \mathfrak{p}(\omega|b,0)}{\int_0^\infty \exp\left(-\frac{\omega\psi^2}{2}\right) \mathfrak{p}(\omega|b,0)\, \mathrm{d}\omega},$$

is also in the Pólya-Gamma class. Exactly. From (4.A.4) it follows that $(\omega \mid \psi) \sim \mathsf{PG}(b,\psi)$.

Bibliography

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

Almgren, R., Thum, C., Hauptmann, E., and Li, H. (2005). Direct estimation of equity market impact. *Risk*.

Amewou-Atisso, M., Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (2003). Posterior consistency for semi-parametric regression problems. *Bernoulli*, 9(2):291–312.

Anceschi, N., Fasano, A., Durante, D., and Zanella, G. (2023). Bayesian conjugacy in probit, tobit, multinomial probit and extensions: A review and new results. *Journal of the American Statistical Association*, 118(542):1451–1469.

Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1):99–102.

Ashukha, A., Lyzhov, A., Molchanov, D., and Vetrov, D. (2020). Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. *arXiv preprint arXiv:2002.06470*.

Avery, C. and Zemsky, P. (1998). Multidimensional Uncertainty and Herd Behavior in Financial Markets. *The American Economic Review*, 88(4):724–748.

Back, K. (1992). Insider trading in continuous time. *The Review of Financial Studies*, 5(3):387–409.

Back, K., Crotty, K., and Li, T. (2013). Estimating the Order-Flow Component of Security Returns. *SSRN Electronic Journal*.

Banerjee, S. and Green, B. (2015). Signal or noise? Uncertainty and learning about whether other traders are informed. *Journal of Financial Economics*, 117(2):398–423.

Barber, D. and Bishop, C. (1997). Ensemble learning for multi-layer networks. *Advances in neural information processing systems*, 10.

Barra (1997). *Market Impact Model Handbook*. Barra, Berkeley, California. Proprietary research report.

Barron, A., Schervish, M. J., and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27(2):536–561.

Bershova, N. and and Rakhlin, D. (2013). The non-linear market impact of large trades: Evidence from buy-side order flow. *Quantitative Finance*, 13(11):1759–1778.

Bingham, N. H., Goldie, C. M., and Teugels, J. L. (1987). *Regular Variation*. Encyclopedia of Mathematics and its Applications. Cambridge University Press.

Bishop, C. and Svenskn, M. (2003). Bayesian hierarchical mixtures of experts. In *UAI'03 Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 57–64. Morgan Kaufmann Publishers Inc.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

Bogousslavsky, V., Fos, V., and Muravyev, D. (2024). Informed Trading Intensity. *The Journal of Finance*, 79(2):903–948.

Bouchaud, J.-P., Bonart, J., Donier, J., and Gould, M. (2018). *Trades, quotes and prices: financial markets under the microscope*. Cambridge University Press.

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2018). Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data (with Discussion). *Bayesian Analysis*, 13(1):253 – 310.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brown, D. P. and Jennings, R. H. (1989). On technical analysis. *Review of Financial Studies*, 2(4):527–551.

Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174.

Castiglione, C. and Bernardi, M. (2025). Non-conjugate variational bayes for pseudo-likelihood mixed effect models. *Journal of Computational and Graphical Statistics*, 0(ja):1–18.

Çetin, U. and Waelbroeck, H. (2024). Power laws in market microstructure. In *Peter Carr Gedenkschrift: Research Advances in Mathematical Finance*, pages 753–819. World Scientific.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335.

Choi, T. and Ramamoorthi, R. V. (2008). *Remarks on consistency of posterior distributions*, page 170–186. Institute of Mathematical Statistics.

Choi, T. and Schervish, M. J. (2007). On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis*, 98(10):1969–1987.

Choudhuri, N., Ghosal, S., and Roy, A. (2004). Bayesian estimation of the spectral density of a time series. *Journal of the American Statistical Association*, 99(468):1050–1059.

Cohen, K. J., Maier, S. F., Schwartz, R. A., and Whitcomb, D. K. (1981). Transaction costs, order placement strategy, and the existence of the bid-ask spread. *Journal of Political Economy*, 89(2):287–305.

Collin-Dufresne, P. and Fos, V. (2015). Do Prices Reveal the Presence of Informed Trading? *The Journal of Finance*, 70(4):1555–1582.

Cont, R., Kukanov, A., and Stoikov, S. (2014). The Price Impact of Order Book Events. *Journal of Financial Econometrics*, 12(1):47–88.

Cremer, C., Li, X., and Duvenaud, D. (2018). Inference suboptimality in variational autoencoders.

David M. Blei, A. K. and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, NY, USA.

Diaconis, P. and Freedman, D. A. (1986). On the consistency of bayes estimates. *Annals of Statistics*, 14(1):1–26.

Doob, J. L. (1949). Application of the theory of martingales. *Le Calcul des Probabilités et ses Applications*, pages 23–27.

Durante, D. (2019). Conjugate bayes for probit regression via unified skew-normal distributions. *Biometrika*, 106(4):765–779.

Easley, D. and O'Hara, M. (1987). Price, trade size, and information in securities markets. *Journal of Financial Economics*, 19(1):69–90.

Easley, D., O'Hara, M., and Yang, L. (2014). Opaque Trading, Disclosure, and Asset Prices: Implications for Hedge Fund Regulation. *Review of Financial Studies*, 27(4):1190–1237.

Farmer, J. D. and Lillo, F. (2004). On the origin of power-law tails in price fluctuations. *Quantitative Finance*, 4(1).

Foucault, T. (1999). Order flow composition and trading costs in a dynamic limit order market. *Journal of Financial Markets*.

Freedman, D. A. (1963). On the asymptotic behavior of bayes estimates in the discrete case i. *Annals of Mathematical Statistics*, 34(4):1386–1403.

Freedman, D. A. (1965). On the asymptotic behavior of bayes estimates in the discrete case ii. *Annals of Mathematical Statistics*, 36(2):454–456.

Gabaix, X., Gopikrishnan, P., Plerou, V., and Stanley, H. E. (2003). A theory of power-law distributions in financial market fluctuations. *Nature*, 423:267–270.

Gao, F., Song, F., and Wang, J. (2013). Rational expectations equilibrium with uncertain proportion of informed traders. *Journal of Financial Markets*, 16(3):387–413.

Garman, M. B. (1976). Market microstructure. *Journal of Financial Economics*, 3(3):257–275.

Gershman, S. J. and Goodman, N. D. (2014). Amortized inference in probabilistic reasoning. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, volume 36, pages 517–522, Quebec City, Canada. Cognitive Science Society.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999). Posterior consistency of dirichlet mixtures in density estimation. *Annals of Statistics*, 27(1):143–158.

Ghosal, S. and van der Vaart, A. W. (2007). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223.

Ghosal, S. and van der Vaart, A. W. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Number 44 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer Series in Statistics. Springer New York.

Glosten, L. R. (1994). Is the electronic open limit order book inevitable? *The Journal of Finance*, 49(4):1127–1161.

Glosten, L. R. and Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1):71–100.

Goan, E. and Fookes, C. (2020). *Bayesian Neural Networks: An Introduction and Survey*, pages 45–87. Springer International Publishing, Cham.

Goettler, R. L., Parlour, C. A., and Rajan, U. (2009). Informed traders and limit order markets. *Journal of Financial Economics*, 93(1):67–87.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA.

Gopikrishnan, P., Plerou, V., Gabaix, X., and Stanley, H. E. (2000). Statistical properties of share volume traded in financial markets. *Physical review e*, 62(4):R4493.

Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee.

Grossman, S. J. and Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *The American Economic Review*, 70(3):393–408.

Grundy, B. D. and McNichols, M. (1989). Trade and revelation of information through prices and direct disclosure. *The Review of Financial Studies*, 2(4):485–526.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Hellwig, M. F. (1980). On the aggregation of information in competitive markets. *Journal of Economic Theory*, 22(3):477–498.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97.

Ho, T. and Stoll, H. R. (1981). Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial Economics*, 9(1):47–73.

Holden, C. W. and Subrahmanyam, A. (1992). Long-lived private information and imperfect competition. *The Journal of Finance*, 47(1):247–270.

Honkela, A., Raiko, T., Kuusela, M., Tornio, M., and Karhunen, J. (2010). Approximate riemannian conjugate gradient learning for fixed-form variational bayes. *The Journal of Machine Learning Research*, 11:3235–3268.

Izmailov, P., Vikram, S., Hoffman, M. D., and Wilson, A. G. (2021). What are bayesian neural network posteriors really like?

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214.

Jospin, L. V., Laga, H., Boussaid, F., Buntine, W., and Bennamoun, M. (2022). Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48.

Kingma, D. P. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kingma, D. P., Welling, M., et al. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392.

Knowles, D. and Minka, T. (2011). Non-conjugate variational message passing for multinomial and binary regression. *Advances in neural information processing systems*, 24.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, 53(6):1315–1335.

Lampinen, J. and Vehtari, A. (2001). Bayesian approach for neural networks—review and case studies. *Neural Networks*, 14(3):257–274.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Li, T. (2013). Insider Trading with Uncertain Informed Trading.

Lillo, F., Mike, S., and Farmer, J. D. (2005). Theory for long memory in supply and demand. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 71(6):066122.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

MacKay, D. J. C. (1992). A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc.

Moradi, R., Berangi, R., and Minaei, B. (2020). A survey of regularization strategies for deep models. *Artificial Intelligence Review*, 53(6):3947–3986.

Moro, E., Vicente, J., Moyano, L. G., Gerig, A., Farmer, J. D., Vaglica, G., Lillo, F., and Mantegna, R. N. (2009). Market impact and trading profile of hidden orders in stock markets. *Phys. Rev. E*, 80:066102.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

Nguyen, A., Yosinski, J., and Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436.

O'Hara, M. (1995). *Market microstructure theory / Maureen O'Hara.* Blackwell Publishers, Cambridge, MA.

Onorati, P. and Liseo, B. (2025). An extension of the unified skew-normal family of distributions and its application to bayesian binary regression. *Journal of Computational and Graphical Statistics*, 0(0):1–14.

Opper, M. and Archambeau, C. (2009). The variational gaussian approximation revisited. *Neural Comput*, 21(3):786–792.

Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, 64(2):140–153.

Papadimitriou, D. (2023). Trading under uncertainty about other market participants. *Financial Review*, 58(2):343–367.

Parlour, C. A. (2015). Price dynamics in limit order markets. *The Review of Financial Studies*, 11(4):789–816.

Polson, N. G. and Scott, J. G. (2013). Data augmentation for non-Gaussian regression models using variance-mean mixtures. *Biometrika*, 100(2):459–471.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

Polson, N. G. and Scott, S. L. (2011). Data augmentation for support vector machines. *Bayesian Analysis*, 6(1):1 – 23.

Polson, N. G. and Sokolov, V. (2017). Deep Learning: A Bayesian Perspective. *Bayesian Analysis*, 12(4):1275 – 1304.

Potters, M. and Bouchaud, J.-P. (2003). More statistical properties of order books and price impact. *Physica A: Statistical Mechanics and its Applications*, 324(1):133–140.

Razavi, A., van den Oord, A., Poole, B., and Vinyals, O. (2019). Preventing posterior collapse with delta-vaes.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models.

Rohde, D. and Wand, M. P. (2016). Semiparametric mean field variational bayes: General principles and numerical issues. *Journal of Machine Learning Research*, 17(172):1–47.

Rosu, I. (2009). A Dynamic Model of the Limit Order Book. *The Review of Financial Studies*, 22(11):4601–4641.

Saddier, L. and Marsili, M. (2024). A bayesian theory of market impact. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(8):083404.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897.

Schwartz, L. (1965). On bayes procedures. *Z. Wahrsch. Verw. Gebiete*, 4:10–26.

Shalizi, C. R. (2009). Dynamics of bayesian updating with dependent data and misspecified models. *Electronic Journal of Statistics*, 3:1–36.

Sheinkman, A. and Wade, S. (2024). Variational bayesian bow tie neural networks with shrinkage. *arXiv preprint arXiv:2411.11132*.

Shridhar, K., Lee, J., Hayashi, H., Mehta, P., Iwana, B. K., Kang, S., Uchida, S., Ahmed, S., and Dengel, A. (2019). Probact: A probabilistic activation function for deep neural networks. *arXiv preprint arXiv:1905.10761*.

Smith, J. T. H., Lawson, D., and Linderman, S. W. (2021). Bayesian inference in augmented bow tie networks. In *Bayesian Deep Learning Workshop, NeurIPS 2021*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Tan, L. S. and Nott, D. J. (2013). Variational inference for generalized linear mixed models using partially noncentered parametrizations. *Statistical Science*, 28(2):168–188.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.

Titterington, D. M. (2004). Bayesian Methods for Neural Networks and Related Models. *Statistical Science*, 19(1):128 – 139.

Tóth, B., Eisler, Z., and Bouchaud, J.-P. (2016). The Square-Root Impace Law Also Holds for Option Markets. *Wilmott*, 2016(85):70–73.

Tsanas, A. and Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49:560–567.

Van Rijsbergen, C. J. (2004). *The geometry of information retrieval*. Cambridge University Press.

Wainwright, M. J., Jordan, M. I., et al. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

Walker, S. G. (2004). New approaches to bayesian consistency. *Annals of Statistics*, 32(5):2028–2043.

Wang, Y., Polson, N., and Sokolov, V. O. (2022). Data Augmentation for Bayesian Deep Learning. *Bayesian Analysis*, pages 1 – 29.

Waterhouse, S., MacKay, D., and Robinson, A. (1995). Bayesian methods for mixtures of experts. In Touretzky, D., Mozer, M., and Hasselmo, M., editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press.

Williams, D. (1991). *Probability with Martingales*. Cambridge Mathematical Textbooks. Cambridge University Press.

Yeh, I.-C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, 28(12):1797–1808.

Zarinelli, E., Treccani, M., Farmer, J. D., and Lillo, F. (2015). Beyond the Square Root: Evidence for Logarithmic Dependence of Market Impact on Size and Participation Rate. *Market Microstructure and Liquidity*, 01(02):1550004.

Zhang, C., Bütepage, J., Kjellström, H., and Mandt, S. (2019). Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):2008–2026.