Options and Market Making

a thesis presented for the degree of

Doctor of Philosophy of Imperial College London

by

Douglas Machado Vieira

Department of Mathematics

Imperial College

180 Queen's Gate, London SW7 2BZ

May 2022

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Signed: _____

Copyright

Supervisors: Prof Rama Cont
Dr Mikko Pakkanen

Douglas Machado Vieira

Options and Market Making

Abstract

Options and market making are recurring themes in Mathematical Finance. This thesis explores both topics with the ultimate goal of developing an options market making model for exchange-traded vanilla options. We start the derivation of closed-form optimal controls for an asset-agnostic market making model with multiple assets via an ergodic limit. We then investigate the intraday dynamics of options and its connection with spot volatility to gain insights on the high-frequency option price dynamics and on volatility and Greeks estimation. Finally, we develop a market making model for exchange-traded vanilla options that encompasses relevant features that we observe empirically. Closed-form solutions for the options market making model can be obtained via small time-to-horizon asymptotics. The optimal spreads in the small time-to-horizon regime allow us to empirically study options spreads and trading activity.

To the right – but not the obligation – to buy or sell the underlying at a predetermined price.

Acknowledgements

I start by expressing my gratitude to my supervisors Prof Rama Cont and Dr Mikko Pakkanen and my collaborators Prof Olivier Guéant, Dr David Evangelista and Mr Philippe Bergault. I appreciate the active support of the Imperial academic staff, especially Prof Johannes Muhle-Karbe, Prof Damiano Brigo, Dr Antoine Jacquier and Dr Eyal Neuman. I acknowledge the sponsorship from Nomura, and the people that made it possible, namely Dr Eduardo Epperlein, Dr Marc Jeannin, Dr Jiong Zhou, Dr John Sleath and Dr Haakon Skaane. The fruitful discussions from the Mathematical Finance community, including Prof Álvaro Cartea, Prof Jim Gatheral, Prof Mathieu Rosenbaum, Dr Katia Babbar, Dr Paul Bilokon, Dr Andrew Crick, Dr Zoltan Eisler, Dr Claude Martini, Dr Paul McCloud, Dr Roel Oomen, Dr Thomas Grassl, Mr Adnane Ait Omar and Mr Guillaume Bioche were very helpful for this thesis. I finally thank my PhD colleagues at Imperial College London, my family and friends for their support and for making the PhD more enjoyable.

Contents

Introduction

Overview

Options and market making are recurring themes in Mathematical Finance. This thesis explores both topics with the ultimate goal of developing an options market making model for exchange-traded vanilla options.

Market making is a trading strategy by which an institution provides both buy and sell quotes, thus providing liquidity to the market. In the context of financial exchanges and quote-driven markets, the market maker can be designated by the financial exchange or can perform this strategy independently. The source of profit of the market maker is on the spread between the quotes posted to the market. The buy and sell trades have no reason to be balanced, and thus one of the main sources of risk of the market maker is the market risk of the accumulated inventory.

Many asset classes are traded in financial exchanges, including stocks, ETFs and derivatives on commodities, currency, single stocks, indices and volatility. Market making strategies will of course differ depending on the idiosyncrasy of each asset class. In this thesis, we explore the features that make the options asset class unique from the market making perspective.

Structure of the thesis

The thesis is composed of three chapters. Chapter 1 is devoted to the derivation of closed-form optimal controls for an asset-agnostic market making model with multiple assets via an ergodic limit[1]. Chapter 2 is an empirical study of option dynamics at small time scales, with special focus on its relationship with spot volatility. Finally, Chapter 3 derives closed-form solutions for the options market making model via a small time-to-horizon limit and empirically studies options spreads and trading activity under the lens

[1] Chapter 1 corresponds to Bergault et al. (2021).

of the market making model.

In summary, Chapter 1 studies an ergodic regime for a general multi-asset market making model that can be applied for options. Chapter 3 studies the opposite regime in which the market maker horizon is short. To motivate the modeling assumptions for market making under the small time-to-horizon regime, Chapter 2 studies the intraday dynamics of options to find the leading-order drivers, which includes spot volatility.

Summary of Chapter 1

A large proportion of market making models derive from the seminal model of Avellaneda and Stoikov. The numerical approximation of the value function and the optimal quotes in these models remains a challenge when the number of assets is large. In this article, we propose heuristic closed-form approximations for the value functions of many multi-asset extensions of the Avellaneda-Stoikov model. These approximations or proxies can be used (i) as heuristic evaluation functions, (ii) as initial value functions in reinforcement learning algorithms, and/or (iii) directly to design quoting strategies through a greedy approach. Regarding the latter, our results lead to new and easily interpretable closed-form approximations for the optimal quotes, both in the finite-horizon case and in the asymptotic (ergodic) regime.

Summary of Chapter 2

Spot volatility is commonly modelled in option pricing models. Motivated by small time asymptotics of option prices, we empirically investigate the effect of spot volatility on 1-second option price changes. We develop a novel approach for spot volatility estimation which employs option pricing models, which enables the study of spot volatility at fine granularities but at the cost of introducing model dependency. We identify and quantify the effect of the estimated spot volatility on option price changes and find that up to 30% of the option price variation can be solely attributed to the estimated spot volatility changes.

Summary of Chapter 3

We develop a market making model for exchange-traded vanilla options that encompasses relevant features: (i) stochastic volatility, (ii) driving factors for the implied volatility surface, (iii) trade activity driven by moneyness, (iv) friction in the underlying market and (v) end-of-day horizon. The end-of-day horizon motivates the small time-to-horizon

asymptotics, from which we formally derive compact formulas for optimal quotes. In light of the model, we perform an empirical analysis. The observed market bid-ask spreads are remarkably consistent with the optimal spreads provided by the calibrated model. From the structure of the bid-ask spread in moneyness and time-to-expiry, we are then able to explain the structure of the trading activity.

A primer in stochastic control theory

For the optimal market making models, we need to consider controlled stochastic processes with jumps. We provide a formal introduction to the topic and the reader is referred to the books by Fleming and Soner (2006) and Øksendal and Sulem (2019) for rigorous definitions and theorems involved in this primer.

Let $(X_t^{x,u})_{t\in[0,T]}$ be a controlled Markov process of the form

$$dX_t^{x,u} = b(t, X_t^{x,u}, u_t)dt + \sigma(t, X_t^{x,u}, u_t)dW_t + \int_{\mathbb{R}^n} \gamma(X_{t-}^{x,u}, u_{t-}, z)\tilde{N}(dt, dz), \quad X_0 = x,$$

where $u = (u_t)_{t\in[0,T]}$ denotes the control, b, σ and γ are given functions and $\tilde{N}$ is a compensated random measure[2]. Denote by $\mathcal{L}^y$ the infinitesimal generator of $(X_t^{x,u})_{t\in[0,T]}$, where the superscript y denotes that $\mathcal{L}^y$ varies on y with $u_t = y$. Then, the Dynkin's formula on a given function ϕ states that

$$\mathbb{E}\left[\phi\left(X_t^{x,u}\right)\right] = \phi(x) + \mathbb{E}\left[\int_0^t \mathcal{L}^{u_s}\phi(X_s^{x,u})ds\right].$$

In practice, Dynkin's formula allows us to find the infinitesimal generator $\mathcal{L}^y$ by applying Itô's formula on ϕ. The process $(X_t)_{t\in[0,T]}$ is allowed to have jumps, which means that the generator $\mathcal{L}^y$ can also contain a difference operator[3].

In this thesis, we are interested in finite-horizon optimisation problems of the form

$$\mathbb{E}\left[\int_0^T f(s, X_s^{x,u}, u_s)ds + g(X_T^{x,u})\right],$$

where f is called the running profit function and g is called a terminal reward function.

[2]A compensated random measure suffices for this thesis, however a more general setting on controlled Markov process with jumps can be found in Chapter 3 in Øksendal and Sulem (2019).

[3]The explicit generator is provided in Chapter 3 in Øksendal and Sulem (2019).

For such problems, we define the performance criterion

$$J(t,x,u) = \mathbb{E}\left[\int_0^{T-t} f(s, X_s^{x,u}, u_s)ds + g(X_{T-t}^{x,u})\right].$$

The dynamic programming principle provides the so-called Hamilton-Jacobi Bellman equation associated to the optimisation of the above performance criterion, which is

$$\sup_y \left(\left(\partial_t + \mathcal{L}^y\right) v(t,x) + f(t,x,y)\right) = 0,$$

with terminal condition $v(T,x) = g(x)$, where v is known as the value function.

In this Markov setting, the optimal control u^* is of the so-called feedback form $u_t^* = u^*(t, X_t)$, which satisfies

$$u^*(t,x) = \arg\sup_y \left(\left(\partial_t + \mathcal{L}^y\right) v(t,x) + f(t,x,y)\right).$$

Therefore, the optimal control u^* can be found by solving the Hamilton-Jacobi Bellman equation associated with the optimisation criterion. The optimality of u_t^*, in the sense that

$$v(t, X_t^{x,u^*}) = J(t, X_t^{x,u^*}, u_t^*),$$

is made rigorous by a Verification Theorem such as Theorem III.8.1 in Fleming and Soner (2006), which encompasses our setting.

Chapter 1

Closed-form approximations in multi-asset market making

1.1 Introduction

Since the publication of Avellaneda and Stoikov (2008), who revisited Ho and Stoll (1981) (see also Ho and Stoll (1983)), there has been an extensive literature on optimal market making.[1] Guéant et al. (2013) provided a rigorous analysis of the stochastic optimal control problem introduced by Avellaneda and Stoikov (2008) and proved that, under inventory constraints, the problem reduces to a system of linear ordinary differential equations in the case of exponential intensity functions suggested by Avellaneda and Stoikov (2008). They also studied the asymptotics when the time horizon T tends to $+\infty$, proposed closed-form approximations, and introduced extensions to include a drift in the price dynamics and market impact / adverse selection. Cartea and Jaimungal, along with their various coauthors, contributed substantially to the literature and added many features to the initial models: alpha signals, ambiguity aversion, etc. (see Cartea et al. (2017), Cartea et al. (2014), Cartea et al. (2018) – see also their book Cartea et al. (2015)). They also considered a different objective function: the expected PnL minus a running penalty to avoid holding a large inventory instead of the Von Neumann-Morgenstern expected CARA (constant absolute risk aversion) utility of Avellaneda and Stoikov (2008) and Guéant et al. (2013). Many features have also been added by various authors: general dynamics for the price in Fodra and Labadie (2013), general intensities and partial information in Campi and Zabaljauregui (2020), persistence of the order flow

[1] There is an economic literature on market making, for instance the seminal paper by Grossman and Miller (1988). The results in this literature are, however, more interesting for understanding the price formation process than for building market making algorithms.

in Jusselin (2020), several requested sizes in Bergault and Guéant (2019), client tiering and access to a liquidity pool in Barzykin et al. (2020), etc.

In spite of the focus of initial papers on stock markets,[2] the models derived from that of Avellaneda and Stoikov (2008) have been more useful to build market making algorithms in quote-driven markets: corporate bond markets based on requests for quotes, FX markets based on requests for quotes and requests for stream, etc. For stock markets or, more generally, order-driven markets with relatively low bid-ask spread to tick size ratio, many models have been proposed that depart from the original framework of Avellaneda and Stoikov (2008) in that the limit order book is modeled. Instances of papers proposing this type of models include those of Guilbaud and Pham (2013), Guilbaud and Pham (2015), that of Kühn and Muhle-Karbe (2015), that of Fodra and Pham (2015) or the more recent papers by Lu and Abergel (2018) and Baradel et al. (2018).

Most of the literature on optimal market making deals with single-asset models. However, because market making algorithms are typically built for entire portfolios, single-asset models are not sufficient to build operable algorithms, except under the unrealistic assumption that asset prices are uncorrelated. Multi-asset extensions of the Avellaneda-Stoikov model have been proposed. A paper by Gueant and Lehalle (2015) touches upon this extension and a complete analysis for the various objective functions present in the literature can be found in Guéant (2017) (see also the book Guéant (2016)) or in Bergault and Guéant (2019) in which multiple trade sizes are also considered.

Although their mathematical characterization has been known for years, computing the value function and the optimal quotes is complicated in the multi-asset case whenever the prices of the assets are correlated. The grid methods that are classically used to tackle the single-asset case suffer indeed from the curse of dimensionality and do not scale up to many practical multi-asset cases. Bergault and Guéant (2019) proposed a factor method to reduce the dimensionality of the problem. Guéant and Manziuk (2019) proposed a numerical method based on reinforcement learning techniques (an actor-critic approach). In spite of these recent advances, the computational cost of most numerical schemes will still be prohibitive for practical use for some asset classes.

Instead of computing a numerical approximation of the value function (from which one traditionally deduces a numerical approximation of the optimal quotes), we propose in this chapter a method for building a closed-form proxy for the value function. The idea behind the approach is that the value function associated with many market making problems is the solution of a Hamilton-Jacobi equation that can be "approximated" by another Hamilton-Jacobi equation for which the solution can be computed in closed-form.

[2]There was also from the very beginning a focus on options markets – see for instance Stoikov and Sağlam (2009) (cf. Baldacci et al. (2021) and El Aoud and Abergel (2015) for more recent papers).

Of course, such closed-form formula does not define a solution to the initial Hamilton-Jacobi equation, but it has similar properties and should capture most of the relevant financial effects.

Having a proxy of a value function is known to be useful in the community of reinforcement learning (see Sutton and Barto (2018) and Szepesvári (2010) for references to the reinforcement learning terminology). An important use of a closed-form proxy of a value function is as a heuristic evaluation function. Heuristic evaluation functions are mainly used in game-playing computer programs to evaluate the probability to win the game given the current state – usually the current board in board games – but they can be used as terminal values in many Monte-Carlo-based reinforcement learning techniques. Also, such a proxy can be used as a starting point for many iterative algorithms based on value functions: value iteration algorithm, actor-critic approaches, etc. The last application we highlight – which was also our initial motivation – is that one can build from a proxy of a value function a quoting strategy by using what is called in the reinforcement literature the greedy strategy associated with that proxy (i.e. the strategy that makes the locally optimal choice if at each time step the value function associated with the tail problem is replaced by its proxy in the dynamic programming equation). Having such a strategy in closed-form has numerous advantages. First, it can be used directly by market practitioners as a quoting strategy. Second, it can be used as a starting point in iterative algorithms based on policy functions: policy iteration algorithm, actor-critic approaches, etc. Third, it has the advantage of being easily interpretable and gives insights on the true optimal strategy such as the identification of the leading factors and the sensitivity to changes in model parameters.[3]

The method we propose is first applied to the multi-asset market making models of Guéant (2017). Then we generalize the framework in several directions to cover many important practical cases: (i) drift in prices, (ii) client tiering, (iii) several request sizes for each asset and each tier, and (iv) fixed transaction costs for each asset and each tier. The drift in prices models the views of the market maker. Client tiering is a common practice in OTC markets, justified by the large spectrum of needs and behaviors in the set of clients to be served. The introduction of several request sizes for each asset and each tier reflects the reality that request sizes are not in control of the market makers, but rather of their clients. The fixed transaction costs can model extra costs associated with the market making business, for instance related to trading platforms.

We end this introduction by outlining this chapter. In Section 1.2 we recall the multi-asset extensions of the Avellaneda-Stoikov model proposed in Guéant (2017), present the sys-

[3]For market making, the influence of the parameters has already been studied in Guéant et al. (2013) (one-asset case) and Guéant (2017) (multi-asset case).

tem of ordinary differential equations (the Hamilton-Jacobi equation) characterizing the value function, and state the main results regarding the optimal quotes. In Section 1.3, we present our approach and compute a closed-form proxy for the value function. We deduce from that proxy an approximation of the optimal quotes in closed-form. In Section 1.4, we use a perturbation approach to propose a correction term that can easily be computed thanks to Monte-Carlo simulations. In Section 1.5, we extend our results to a more general multi-asset market making model with drift in prices, client tiering, several requested sizes for each asset and each tier, and fixed transaction costs for each asset and each tier. Numerical examples are presented in Section 1.6. They illustrate the quality of our closed-form approximations.

1.2 The multi-asset market making model

1.2.1 Model setup

We fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ equipped with a filtration $(\mathcal{F}_t)_{t\in\mathbb{R}_+}$ satisfying the usual conditions. In what follows, we assume that all stochastic processes are defined on $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathbb{R}_+}, \mathbb{P})$. In all this chapter, $\mathbb{R}_+$ denotes the set of nonnegative real numbers, and $\mathbb{R}_+^*$ denotes the set of positive real numbers.

For $i \in \{1, \ldots, d\}$, the reference price of asset i is modeled by a process $(S_t^i)_{t\in\mathbb{R}_+}$ with dynamics

$$dS_t^i = \sigma^i dW_t^i, \quad S_0^i \text{ given},$$

where $(W_t^1, \ldots, W_t^d)_{t\in\mathbb{R}_+}$ is a d-dimensional Brownian motion with correlation matrix $(\rho^{i,j})_{1\leqslant i,j\leqslant d}$ adapted to the filtration $(\mathcal{F}_t)_{t\in\mathbb{R}_+}$ – hereafter we denote by $\Sigma = (\rho^{i,j}\sigma^i\sigma^j)_{1\leqslant i,j\leqslant d}$ the variance-covariance matrix associated with the process $(S_t)_{t\in\mathbb{R}_+} = (S_t^1, \ldots, S_t^d)_{t\in\mathbb{R}_+}$.

The market maker chooses at each point in time the price at which she is ready to buy/sell each asset: for $i \in \{1, \ldots, d\}$, we let her bid and ask quotes for asset i be modeled by two stochastic processes, respectively denoted by $(S_t^{i,b})_{t\in\mathbb{R}_+}$ and $(S_t^{i,a})_{t\in\mathbb{R}_+}$.

For $i \in \{1, \ldots, d\}$, we denote by $(N_t^{i,b})_{t\in\mathbb{R}_+}$ and $(N_t^{i,a})_{t\in\mathbb{R}_+}$ the two point processes modeling the number of transactions at the bid and at the ask, respectively, for asset i. We assume in this section that the transaction size for asset i is constant and denoted by z^i. The inventory process of the market maker for asset i, denoted by $(q_t^i)_{t\in\mathbb{R}_+}$, has therefore the dynamics

$$dq_t^i = z^i dN_t^{i,b} - z^i dN_t^{i,a}, \quad q_0^i \text{ given},$$

and we denote by $(q_t)_{t\in\mathbb{R}_+}$ the (column) vector process $\left(q_t^1, \ldots, q_t^d\right)^\intercal_{t\in\mathbb{R}_+}$.

For each $i \in \{1, \dots, d\}$, we denote by $(\lambda_t^{i,b})_{t\in\mathbb{R}_+}$ and $(\lambda_t^{i,a})_{t\in\mathbb{R}_+}$ the intensity processes of $(N_t^{i,b})_{t\in\mathbb{R}_+}$ and $(N_t^{i,a})_{t\in\mathbb{R}_+}$, respectively. We assume that the market maker stops proposing a bid (respectively ask) price for asset i when her position in asset i following the transaction would exceed a given threshold Q^i (respectively $-Q^i$).[4]

Formally, we assume that the intensities verify

$$\lambda_t^{i,b} = \Lambda^{i,b}(\delta_t^{i,b})\mathbb{1}_{\{q_{t-}^i + z^i \leq Q^i\}} \quad \text{and} \quad \lambda_t^{i,a} = \Lambda^{i,a}(\delta_t^{i,a})\mathbb{1}_{\{q_{t-}^i - z^i \geq -Q^i\}},$$

where the processes $(\delta_t^{i,b})_{t\in\mathbb{R}_+}$ and $(\delta_t^{i,a})_{t\in\mathbb{R}_+}$ are defined by[5]

$$\delta_t^{i,b} = S_t^i - S_t^{i,b} \quad \text{and} \quad \delta_t^{i,a} = S_t^{i,a} - S_t^i, \quad \forall t \in \mathbb{R}_+.$$

Moreover, we assume that the functions $\Lambda^{i,b}$ and $\Lambda^{i,a}$ satisfy the following properties:

- $\Lambda^{i,b}$ and $\Lambda^{i,a}$ are twice continuously differentiable,
- $\Lambda^{i,b}$ and $\Lambda^{i,a}$ are decreasing, with $\forall \delta \in \mathbb{R}$, $\Lambda^{i,b\prime}(\delta) < 0$ and $\Lambda^{i,a\prime}(\delta) < 0$,
- $\lim_{\delta\to+\infty} \Lambda^{i,b}(\delta) = \lim_{\delta\to+\infty} \Lambda^{i,a}(\delta) = 0$,
- $\sup_\delta \frac{\Lambda^{i,b}(\delta)\Lambda^{i,b\prime\prime}(\delta)}{\left(\Lambda^{i,b\prime}(\delta)\right)^2} < 2 \quad \text{and} \quad \sup_\delta \frac{\Lambda^{i,a}(\delta)\Lambda^{i,a\prime\prime}(\delta)}{\left(\Lambda^{i,a\prime}(\delta)\right)^2} < 2.$

Finally, the process $(X_t)_{t\in\mathbb{R}_+}$ modelling the amount of cash on the market maker's cash account has the following dynamics:

$$\begin{aligned} dX_t &= \sum_{i=1}^d S_t^{i,a} z^i dN_t^{i,a} - S_t^{i,b} z^i dN_t^{i,b} \\ &= \sum_{i=1}^d (S_t^i + \delta_t^{i,a}) z^i dN_t^{i,a} - (S_t^i - \delta_t^{i,b}) z^i dN_t^{i,b} \\ &= \sum_{i=1}^d \left(\delta_t^{i,b} z^i dN_t^{i,b} + \delta_t^{i,a} z^i dN_t^{i,a}\right) - \sum_{i=1}^d S_t^i dq_t^i. \end{aligned}$$

[4] Q^i is assumed to be a multiple of z^i. It corresponds to the inventory risk limit of the market maker for asset i.

[5] It is often assumed in the literature that the point processes are independent of the Brownian motions. In that case, the quote processes $(\delta_t^{i,b})_{t\in\mathbb{R}_+}$ and $(\delta_t^{i,a})_{t\in\mathbb{R}_+}$ have to be independent of prices. In fact, the optimal control problem can be written in a weak form to show that this assumption is not necessary – see 1.7 for more details on the construction of the processes in that case.

1.2.2 The optimization problems

We can consider two different optimization problems for the market maker. Following the initial model proposed by Avellaneda and Stoikov in Avellaneda and Stoikov (2008), we can assume that she maximizes the expected value of a CARA utility function (with risk aversion parameter $\gamma > 0$) applied to the mark-to-market value of her portfolio at a given time T. This mark-to-market value is the sum of the amount X_T on the cash account and the mark-to-market value $\sum_{i=1}^{d} q_T^i S_T^i$ of the assets remaining in the portfolio at date T.[6] More precisely, her optimization problem writes

$$\sup_{\substack{(\delta_t^{1,b})_t,\dots,(\delta_t^{d,b})_t\in\mathcal{A} \\ (\delta_t^{1,a})_t,\dots,(\delta_t^{d,a})_t\in\mathcal{A}}} \mathbb{E}\left[-\exp\left(-\gamma\left(X_T+\sum_{i=1}^{d} q_T^i S_T^i\right)\right)\right],$$

where $\mathcal{A}$ is the set of predictable processes bounded from below. We call Model A our model with this first objective function.

Alternatively, as proposed by Cartea *et al.* in Cartea et al. (2014), we can consider a risk-adjusted expectation for the objective function of the market maker. In that case, the optimization problem writes

$$\sup_{\substack{(\delta_t^{1,b})_t,\dots,(\delta_t^{d,b})_t\in\mathcal{A} \\ (\delta_t^{1,a})_t,\dots,(\delta_t^{d,a})_t\in\mathcal{A}}} \mathbb{E}\left[X_T+\sum_{i=1}^{d} q_T^i S_T^i - \frac{1}{2}\gamma\int_0^T q_t^\intercal \Sigma q_t dt\right].$$

We call Model B our model with this second objective function.

[6] In the literature there is sometimes a penalty function applied to the inventory at terminal time T to "force" liquidation. Here, as we shall focus on the asymptotic regime of the optimal quotes, there is no point considering such a penalty. However, it is noteworthy that most of our non-asymptotic results could be generalized to the case of a quadratic terminal penalty.

1.2.3 The Hamilton-Jacobi-Bellman and Hamilton-Jacobi equations

Let $\{e^i\}_{i=1}^d$ be the canonical basis of $\mathbb{R}^d$. The Hamilton-Jacobi-Bellman equation associated with Model A is

$$\begin{aligned}
0 = \partial_t u(t,x,q,S) &+ \frac{1}{2}\sum_{i,j=1}^{d} \rho^{i,j}\sigma^i\sigma^j \partial^2_{S^iS^j} u(t,x,q,S) \\
&+ \sum_{i=1}^{d} \mathbb{1}_{\{q^i+z^i\leq Q^i\}} \sup_{\delta^{i,b}} \Lambda^{i,b}(\delta^{i,b}) \left(u(t,x-z^iS^i+z^i\delta^{i,b},q+z^ie^i,S)-u(t,x,q,S)\right) \\
&+ \sum_{i=1}^{d} \mathbb{1}_{\{q^i-z^i\geq -Q^i\}} \sup_{\delta^{i,a}} \Lambda^{i,a}(\delta^{i,a}) \left(u(t,x+z^iS^i+z^i\delta^{i,a},q-z^ie^i,S)-u(t,x,q,S)\right),
\end{aligned} \tag{1.1}$$

for all $(t,x,q,S) \in [0,T)\times\mathbb{R}\times\prod_{i=1}^d (z^i\mathbb{Z}\cap[-Q^i,Q^i])\times\mathbb{R}^d$,[7] with terminal condition

$$u(T,x,q,S) = -\exp\left(-\gamma\left(x+\sum_{i=1}^{d} q^iS^i\right)\right) \quad \forall (x,q,S)\in\mathbb{R}\times\prod_{i=1}^{d}\left(z^i\mathbb{Z}\cap[-Q^i,Q^i]\right)\times\mathbb{R}^d.$$

The Hamilton-Jacobi-Bellman equation associated with Model B is

$$\begin{aligned}
0 = \partial_t v(t,x,q,S) &- \frac{1}{2}\gamma q^\intercal\Sigma q + \frac{1}{2}\sum_{i,j=1}^{d} \rho^{i,j}\sigma^i\sigma^j \partial^2_{S^iS^j} v(t,x,q,S) \\
&+ \sum_{i=1}^{d} \mathbb{1}_{\{q^i+z^i\leq Q^i\}} \sup_{\delta^{i,b}} \Lambda^{i,b}(\delta^{i,b}) \left(v(t,x-z^iS^i+z^i\delta^{i,b},q+z^ie^i,S)-v(t,x,q,S)\right) \\
&+ \sum_{i=1}^{d} \mathbb{1}_{\{q^i-z^i\geq -Q^i\}} \sup_{\delta^{i,a}} \Lambda^{i,a}(\delta^{i,a}) \left(v(t,x+z^iS^i+z^i\delta^{i,a},q-z^ie^i,S)-v(t,x,q,S)\right),
\end{aligned} \tag{1.2}$$

for all $(t,x,q,S) \in [0,T)\times\mathbb{R}\times\prod_{i=1}^d (z^i\mathbb{Z}\cap[-Q^i,Q^i])\times\mathbb{R}^d$ with terminal condition

$$v(T,x,q,S) = x+\sum_{i=1}^{d} q^iS^i \quad \forall (x,q,S)\in\mathbb{R}\times\prod_{i=1}^{d}\left(z^i\mathbb{Z}\cap[-Q^i,Q^i]\right)\times\mathbb{R}^d.$$

[7] Given a positive number $z\in\mathbb{R}_+^*$, $z\mathbb{Z}$ denotes the set of multiples of z, i.e. $z\mathbb{Z} = \{\ldots,-2z,-z,0,z,2z,\ldots\}$.

For each $i \in \{1, \dots, d\}$ and $\xi \geq 0$, let us define two Hamiltonian functions[8] $H_\xi^{i,b}$ and $H_\xi^{i,a}$ by

$$H_\xi^{i,b}(p) = \begin{cases} \sup_\delta \frac{\Lambda^{i,b}(\delta)}{\xi z^i}(1 - \exp(-\xi z^i(\delta - p))) & \text{if } \xi > 0, \\ \sup_\delta \Lambda^{i,b}(\delta)(\delta - p) & \text{if } \xi = 0, \end{cases} \tag{1.3}$$

and

$$H_\xi^{i,a}(p) = \begin{cases} \sup_\delta \frac{\Lambda^{i,a}(\delta)}{\xi z^i}(1 - \exp(-\xi z^i(\delta - p))) & \text{if } \xi > 0, \\ \sup_\delta \Lambda^{i,a}(\delta)(\delta - p) & \text{if } \xi = 0. \end{cases} \tag{1.4}$$

Using the ansatz introduced in Guéant (2017) for the two functions $u : [0,T] \times \mathbb{R} \times \prod_{i=1}^d (z^i\mathbb{Z} \cap [-Q^i, Q^i]) \times \mathbb{R}^d \to \mathbb{R}$ and $v : [0,T] \times \mathbb{R} \times \prod_{i=1}^d (z^i\mathbb{Z} \cap [-Q^i, Q^i]) \times \mathbb{R}^d \to \mathbb{R}$, i.e.

$$u(t,x,q,S) = -\exp\left(-\gamma\left(x + \sum_{i=1}^d q^i S^i + \theta(t,q)\right)\right)$$

$$\text{and } v(t,x,q,S) = x + \sum_{i=1}^d q^i S^i + \theta(t,q),$$

we see that solving the Hamilton-Jacobi-Bellman equations (1.1) and (1.2) boils down to finding the solution $\theta : [0,T] \times \prod_{i=1}^d (z^i\mathbb{Z} \cap [-Q^i, Q^i]) \to \mathbb{R}$ of the following Hamilton-Jacobi equation with $\xi = \gamma$ in the case of Model A and $\xi = 0$ in the case of Model B:

$$\begin{aligned} 0 \;=\; & \partial_t \theta(t,q) - \frac{1}{2}\gamma q^\intercal \Sigma q \\ & + \sum_{i=1}^d \mathbb{1}_{\{q^i + z^i \leq Q^i\}} z^i H_\xi^{i,b}\left(\frac{\theta(t,q) - \theta(t,q + z^i e^i)}{z^i}\right) \\ & + \sum_{i=1}^d \mathbb{1}_{\{q^i - z^i \geq -Q^i\}} z^i H_\xi^{i,a}\left(\frac{\theta(t,q) - \theta(t,q - z^i e^i)}{z^i}\right). \end{aligned} \tag{1.5}$$

In both cases, the terminal condition simply boils down to

$$\theta(T,q) = 0. \tag{1.6}$$

[8] It is noteworthy that our definition of $H_\xi^{i,b}$ and $H_\xi^{i,a}$ differs from that of Guéant (2017) (by a factor z^i). The alternative definition we use in this chapter is also that of Bergault and Guéant (2019) for $\xi = 0$.

1.2.4 Existing theoretical results

From (Guéant, 2017, Theorem 5.1), for a given $\xi \geq 0$, there exists a unique $\theta : [0,T] \times \prod_{i=1}^{d} \left(z^i \mathbb{Z} \cap [-Q^i, Q^i]\right) \to \mathbb{R}$, C^1 in time, solution of Eq. (1.5) with terminal condition (1.6). Moreover (see (Guéant, 2017, Theorems 5.2 and 5.3)), a classical verification argument enables to go from θ to optimal controls for both Model A and Model B. The optimal quotes as functions of θ are recalled in the following theorems (for details, see Guéant (2017)).

In the case of Model A, the result is the following:

Theorem 1. *Let us consider the solution θ of Eq. (1.5) with terminal condition (1.6) for $\xi = \gamma$.*

Then, for $i \in \{1, \dots, d\}$, the optimal bid and ask quotes $S_t^{i,b} = S_t^i - \delta_t^{i,b*}$ and $S_t^{i,a} = S_t^i + \delta_t^{i,a*}$ in Model A are characterized by

$$\begin{aligned}
\delta_t^{i,b*} &= \tilde{\delta}_\gamma^{i,b*}\left(\frac{\theta(t,q_{t-}) - \theta(t,q_{t-} + z^i e^i)}{z^i}\right) \quad \text{for } q_{t-} + z^i e^i \in \prod_{j=1}^{d} \left(z^j \mathbb{Z} \cap [-Q^j, Q^j]\right), \\
\delta_t^{i,a*} &= \tilde{\delta}_\gamma^{i,a*}\left(\frac{\theta(t,q_{t-}) - \theta(t,q_{t-} - z^i e^i)}{z^i}\right) \quad \text{for } q_{t-} - z^i e^i \in \prod_{j=1}^{d} \left(z^j \mathbb{Z} \cap [-Q^j, Q^j]\right),
\end{aligned} \tag{1.7}$$

where the functions $\tilde{\delta}_\gamma^{i,b*}(\cdot)$ and $\tilde{\delta}_\gamma^{i,a*}(\cdot)$ are defined by

$$\begin{aligned}
\tilde{\delta}_\gamma^{i,b*}(p) &= {\Lambda^{i,b}}^{-1}\left(\gamma z^i H_\gamma^{i,b}(p) - H_\gamma^{i,b\prime}(p)\right), \\
\tilde{\delta}_\gamma^{i,a*}(p) &= {\Lambda^{i,a}}^{-1}\left(\gamma z^i H_\gamma^{i,a}(p) - H_\gamma^{i,a\prime}(p)\right),
\end{aligned}$$

where for all $i \in \{1, \dots, d\}$, $H_\gamma^{i,b\prime}$ and $H_\gamma^{i,a\prime}$ denote the first derivative of $H_\gamma^{i,b}$ and $H_\gamma^{i,a}$, respectively.

For Model B, the result is the following:

Theorem 2. *Let us consider the solution θ of Eq. (1.5) with terminal condition (1.6) for $\xi = 0$.*

Then, for $i \in \{1, \dots, d\}$, the optimal bid and ask quotes $S_t^{i,b} = S_t^i - \delta_t^{i,b*}$ and $S_t^{i,a} = S_t^i + \delta_t^{i,a*}$

in Model B are characterized by

$$\begin{aligned}
\delta_t^{i,b*} &= \tilde{\delta}_0^{i,b*}\left(\frac{\theta(t,q_{t-})-\theta(t,q_{t-}+z^ie^i)}{z^i}\right) \quad \text{for } q_{t-}+z^ie^i \in \prod_{j=1}^{d}\left(z^j\mathbb{Z}\cap[-Q^j,Q^j]\right), \\
\delta_t^{i,a*} &= \tilde{\delta}_0^{i,a*}\left(\frac{\theta(t,q_{t-})-\theta(t,q_{t-}-z^ie^i)}{z^i}\right) \quad \text{for } q_{t-}-z^ie^i \in \prod_{j=1}^{d}\left(z^j\mathbb{Z}\cap[-Q^j,Q^j]\right),
\end{aligned} \tag{1.8}$$

where the functions $\tilde{\delta}_0^{i,b*}(\cdot)$ *and* $\tilde{\delta}_0^{i,a*}(\cdot)$ *are defined by*

$$\tilde{\delta}_0^{i,b*}(p) = \Lambda^{i,b^{-1}}\left(-H_0^{i,b'}(p)\right) \text{ and } \tilde{\delta}_0^{i,a*}(p) = \Lambda^{i,a^{-1}}\left(-H_0^{i,a'}(p)\right)$$

where for all $i \in \{1,\dots,d\}$, $H_0^{i,b'}$ *and* $H_0^{i,a'}$ *denote the first derivative of* $H_0^{i,b}$ *and* $H_0^{i,a}$, *respectively.*

In the following two sections, we propose new methods to find approximations of the solution to the system of ordinary differential equations (ODEs) (1.5) with terminal condition (1.6). Eqs. (1.7) and (1.8) can then serve to go from approximations of θ (hereafter called – slightly abusively – the value function) to approximations of the optimal quotes. The resulting quotes correspond to what the reinforcement learning community calls the greedy quoting strategy associated with the proxy of the value function.[9]

1.3 A quadratic approximation of the value function and its applications

1.3.1 Introduction

In the field of (stochastic) optimal control, finding value functions and optimal controls in closed-form is the exception rather than the rule. One important exception goes with the class of Linear-Quadratic (LQ) and Linear-Quadratic-Gaussian (LQG) problems. Of course, the above market making problem does not belong to this class of control problems, for instance because the control of point processes is nonlinear by nature. Nevertheless, we see that price risk appears in both Model A and Model B through the quadratic term $\frac{1}{2}\gamma q^\intercal \Sigma q$ in the Hamilton-Jacobi equation (1.5). The main idea of this chapter consists in replacing the Hamiltonian functions associated with our market making problem by quadratic functions that approximate them. The interest of quadratic

[9] The true optimal quotes correspond to the greedy strategy with respect to the value function u (in Model A) or v (in Model B) deduced from the true θ.

Hamiltonian functions lies in that the resulting Hamilton-Jacobi equations can be solved in closed-form using the same tools as for LQ/LQG problems, i.e. Riccati equations.

At first sight, approximating the Hamiltonian functions involved in Eq. (1.5) by quadratic functions seems inappropriate. For all $i \in \{1, \dots, d\}$, the functions $H_\xi^{i,b}$ and $H_\xi^{i,a}$ are indeed positive and decreasing and approximating them with U-shaped functions can only be valid locally. However, one has to bear in mind that our goal is to approximate the solution of the Hamilton-Jacobi equations and not the Hamiltonian functions. This remark is particularly important because the Hamiltonian terms involved in the Hamilton-Jacobi equations are (up to the indicator functions that we shall discard in what follows by considering the limit case where $\forall i \in \{1, \dots, d\}, Q^i = +\infty$) of the form

$$H_\xi^{i,b}\left(\frac{\theta(t,q) - \theta(t, q + z^i e^i)}{z^i}\right) + H_\xi^{i,a}\left(\frac{\theta(t,q) - \theta(t, q - z^i e^i)}{z^i}\right),$$

Assuming that $\frac{\theta(t,q)-\theta(t,q+z^i e^i)}{z^i} \simeq -\frac{\theta(t,q)-\theta(t,q-z^i e^i)}{z^i}$, we clearly see that, with respect to asset i, the function we need to approximate is $p \mapsto H_\xi^{i,b}(p) + H_\xi^{i,a}(-p)$ rather than $H_\xi^{i,b}$ and $H_\xi^{i,a}$ themselves, and it is natural to approximate the former function with a U-shaped one!

Let us formally replace for all $i \in \{1, \dots, d\}$ the Hamiltonian functions $H_\xi^{i,b}$ and $H_\xi^{i,a}$ by the quadratic functions[10]

$$\check{H}^{i,b} : p \mapsto \alpha_0^{i,b} + \alpha_1^{i,b} p + \frac{1}{2}\alpha_2^{i,b} p^2 \quad \text{and} \quad \check{H}^{i,a} : p \mapsto \alpha_0^{i,a} + \alpha_1^{i,a} p + \frac{1}{2}\alpha_2^{i,a} p^2.$$

Remark 1. A natural choice for the functions $(\check{H}^{i,b})_{i\in\{1,\dots,d\}}$ and $(\check{H}^{i,a})_{i\in\{1,\dots,d\}}$ derives from Taylor expansions around $p = 0$. In that case,

$$\forall i \in \{1, \dots, d\}, \forall j \in \{0, 1, 2\}, \quad \alpha_j^{i,b} = H_\xi^{i,b(j)}(0) \quad \text{and} \quad \alpha_j^{i,a} = H_\xi^{i,a(j)}(0).$$

We denote by $\check{\theta}$ the approximation of θ associated with the functions $(\check{H}^{i,b})_{i\in\{1,\dots,d\}}$ and

[10] We omit the subscript ξ in the definition of $\check{H}^{i,b}$ and $\check{H}^{i,a}$. In particular, although the subscript ξ is not written, the coefficients $\alpha_0^{i,b}$, $\alpha_1^{i,b}$, $\alpha_2^{i,b}$, $\alpha_0^{i,a}$, $\alpha_1^{i,a}$, and $\alpha_2^{i,a}$ do depend on ξ.

$(\check{H}^{i,a})_{i\in\{1,\dots,d\}}$, i.e. if we consider the limit case where $\forall i \in \{1,\dots,d\}, Q^i = +\infty$, $\check{\theta}$ verifies

$$\begin{aligned}
0 &= \partial_t\check{\theta}(t,q) - \frac{1}{2}\gamma q^\intercal \Sigma q + \sum_{i=1}^{d} z^i \left(\alpha_0^{i,b} + \alpha_0^{i,a}\right) \\
&\quad + \sum_{i=1}^{d} \left(\alpha_1^{i,b}\left(\check{\theta}(t,q) - \check{\theta}(t,q+z^ie^i)\right) + \alpha_1^{i,a}\left(\check{\theta}(t,q) - \check{\theta}(t,q-z^ie^i)\right)\right) \\
&\quad + \frac{1}{2}\sum_{i=1}^{d} \frac{1}{z^i}\left(\alpha_2^{i,b}\left(\check{\theta}(t,q) - \check{\theta}(t,q+z^ie^i)\right)^2 + \alpha_2^{i,a}\left(\check{\theta}(t,q) - \check{\theta}(t,q-z^ie^i)\right)^2\right) \quad (1.9)
\end{aligned}$$

and of course we consider the terminal condition

$$\check{\theta}(T,q) = 0. \quad (1.10)$$

1.3.2 An approximation of the value function in closed-form

Eq. (1.9) with terminal condition (1.10) can be solved in closed-form. To prove this point, we start with the following proposition:

Proposition 1. Let us introduce for $i \in \{1,\dots,d\}, j \in \{0,1,2\}, k \in \mathbb{N}$,

$$\Delta_{j,k}^{i,b} = \alpha_j^{i,b}(z^i)^k \quad \text{and} \quad \Delta_{j,k}^{i,a} = \alpha_j^{i,a}(z^i)^k,$$

$$V_{j,k}^{b} = \left(\Delta_{j,k}^{1,b},\dots,\Delta_{j,k}^{d,b}\right)^\intercal \quad \text{and} \quad V_{j,k}^{a} = \left(\Delta_{j,k}^{1,a},\dots,\Delta_{j,k}^{d,a}\right)^\intercal,$$

$$D_{j,k}^{b} = \mathrm{diag}(\Delta_{j,k}^{1,b},\dots,\Delta_{j,k}^{d,b}) \quad \text{and} \quad D_{j,k}^{a} = \mathrm{diag}(\Delta_{j,k}^{1,a},\dots,\Delta_{j,k}^{d,a}).$$

Let us consider three differentiable functions $A : [0,T] \to S_d^+$, $B : [0,T] \to \mathbb{R}^d$, and $C : [0,T] \to \mathbb{R}$ solutions of the system of ordinary differential equations[11]

$$\begin{cases}
A'(t) = & 2A(t)\left(D_{2,1}^b + D_{2,1}^a\right)A(t) - \frac{1}{2}\gamma\Sigma \\
B'(t) = & 2A(t)\left(V_{1,1}^b - V_{1,1}^a\right) + 2A(t)\left(D_{2,2}^b - D_{2,2}^a\right)\mathcal{D}(A(t)) + 2A(t)\left(D_{2,1}^b + D_{2,1}^a\right)B(t) \\
C'(t) = & \mathrm{Tr}\left(D_{0,1}^b + D_{0,1}^a\right) + \mathrm{Tr}\left(\left(D_{1,2}^b + D_{1,2}^a\right)A(t)\right) + \left(V_{1,1}^b - V_{1,1}^a\right)^\intercal B(t) \\
& + \frac{1}{2}\mathcal{D}(A(t))^\intercal\left(D_{2,3}^b + D_{2,3}^a\right)\mathcal{D}(A(t)) + \frac{1}{2}B(t)^\intercal\left(D_{2,1}^b + D_{2,1}^a\right)B(t) \\
& + B(t)^\intercal\left(D_{2,2}^b - D_{2,2}^a\right)\mathcal{D}(A(t)),
\end{cases} \quad (1.11)$$

[11] S_d^+ (resp. S_d^++) stands throughout this chapter the set of positive semi-definite (resp. definite) symmetric d-by-d matrices.

with terminal conditions

$$A(T) = 0, B(T) = 0, \text{ and } C(T) = 0, \tag{1.12}$$

where $\mathcal{D}$ is the linear operator mapping a matrix onto the vector of its diagonal coefficients.

Then $\check{\theta} : (t, q) \in [0, T] \times \prod_{i=1}^{d} z^i \mathbb{Z} \mapsto -q^\intercal A(t) q - q^\intercal B(t) - C(t)$ is solution of Eq. (1.9) with terminal condition (1.10).

Proof. We have

$$\begin{aligned}
& \partial_t \check{\theta}(t,q) - \frac{1}{2}\gamma q^\intercal \Sigma q + \sum_{i=1}^{d} z^i \left(\alpha_0^{i,b} + \alpha_0^{i,a}\right) \\
& + \sum_{i=1}^{d} \left(\alpha_1^{i,b}\left(\check{\theta}(t,q) - \check{\theta}(t,q+z^i e^i)\right) + \alpha_1^{i,a}\left(\check{\theta}(t,q) - \check{\theta}(t,q-z^i e^i)\right)\right) \\
& + \frac{1}{2}\sum_{i=1}^{d} \frac{1}{z^i}\left(\alpha_2^{i,b}\left(\check{\theta}(t,q) - \check{\theta}(t,q+z^i e^i)\right)^2 + \alpha_2^{i,a}\left(\check{\theta}(t,q) - \check{\theta}(t,q-z^i e^i)\right)^2\right) \\
= \; & -q^\intercal A'(t) q - q^\intercal B'(t) - C'(t) - \frac{1}{2}\gamma q^\intercal \Sigma q + \sum_{i=1}^{d} z^i(\alpha_0^{i,b} + \alpha_0^{i,a}) \\
& + \sum_{i=1}^{d} \alpha_1^{i,b}\left(2z^i q^\intercal A(t) e^i + (z^i)^2 e^{i\intercal} A(t) e^i + z^i e^{i\intercal} B(t)\right) \\
& + \sum_{i=1}^{d} \alpha_1^{i,a}\left(-2z^i q^\intercal A(t) e^i + (z^i)^2 e^{i\intercal} A(t) e^i - z^i e^{i\intercal} B(t)\right) \\
& + \frac{1}{2}\sum_{i=1}^{d} \frac{1}{z^i}\alpha_2^{i,b}\left(2z^i q^\intercal A(t) e^i + (z^i)^2 e^{i\intercal} A(t) e^i + z^i e^{i\intercal} B(t)\right)^2 \\
& + \frac{1}{2}\sum_{i=1}^{d} \frac{1}{z^i}\alpha_2^{i,a}\left(-2z^i q^\intercal A(t) e^i + (z^i)^2 e^{i\intercal} A(t) e^i - z^i e^{i\intercal} B(t)\right)^2 \\
= \; & -q^\intercal A'(t) q - q^\intercal B'(t) - C'(t) - \frac{1}{2}\gamma q^\intercal \Sigma q + \mathrm{Tr}\left(D_{0,1}^b + D_{0,1}^a\right) \\
& + 2q^\intercal A(t)\left(V_{1,1}^b - V_{1,1}^a\right) + \mathrm{Tr}\left(\left(D_{1,2}^b + D_{1,2}^a\right) A(t)\right) + \left(V_{1,1}^b - V_{1,1}^a\right)^\intercal B(t) \\
& + 2q^\intercal A(t)\left(D_{2,1}^b + D_{2,1}^a\right) A(t) q + \frac{1}{2}\mathcal{D}(A(t))^\intercal \left(D_{2,3}^b + D_{2,3}^a\right)\mathcal{D}(A(t)) \\
& + \frac{1}{2} B(t)^\intercal \left(D_{2,1}^b + D_{2,1}^a\right) B(t) + 2q^\intercal A(t)\left(D_{2,2}^b - D_{2,2}^a\right)\mathcal{D}(A(t)) \\
& + 2q^\intercal A(t)\left(D_{2,1}^b + D_{2,1}^a\right) B(t) + B(t)^\intercal \left(D_{2,2}^b - D_{2,2}^a\right)\mathcal{D}(A(t)) \\
= \; & 0,
\end{aligned}$$

where the last equality comes from the definitions of (A, B, C) and the identification of

the terms of degree 0, 1, and 2 in q.

As the terminal conditions are satisfied, the result is proved. □

Proposition 2. Assume $\alpha_2^{i,b} + \alpha_2^{i,a} > 0$ for all $i \in \{1, \dots, d\}$[12]. The system of ODEs (1.11) with terminal conditions (1.12) admits the unique solution

$$A(t) = \frac{1}{2} D_+^{-\frac{1}{2}} \widehat{A} \left(e^{\widehat{A}(T-t)} - e^{-\widehat{A}(T-t)} \right) \left(e^{\widehat{A}(T-t)} + e^{-\widehat{A}(T-t)} \right)^{-1} D_+^{-\frac{1}{2}}, \tag{1.13}$$

$$B(t) = -2e^{-2\int_t^T A(u)D_+ \, du} \int_t^T e^{2\int_s^T A(u)D_+ \, du} A(s) \left(V_- + D_- \mathcal{D}(A(s)) \right) ds, \tag{1.14}$$

$$\begin{aligned} C(t) = &-\mathrm{Tr}\left(D_{0,1}^b + D_{0,1}^a \right)(T-t) - \mathrm{Tr}\left(\left(D_{1,2}^b + D_{1,2}^a \right) \int_t^T A(s) ds \right) - V_-^\intercal \int_t^T B(s) ds \\ &- \frac{1}{2} \int_t^T \mathcal{D}(A(s))^\intercal \left(D_{2,3}^b + D_{2,3}^a \right) \mathcal{D}(A(s)) ds - \frac{1}{2} \int_t^T B(s)^\intercal D_+ B(s) ds \\ &- \int_t^T B(s)^\intercal D_- \mathcal{D}(A(s)) ds. \end{aligned} \tag{1.15}$$

where

$$D_+ = D_{2,1}^b + D_{2,1}^a, \quad D_- = D_{2,2}^b - D_{2,2}^a, \quad V_- = V_{1,1}^b - V_{1,1}^a, \quad \text{and} \quad \widehat{A} = \sqrt{\gamma} \left(D_+^{\frac{1}{2}} \Sigma D_+^{\frac{1}{2}} \right)^{\frac{1}{2}}.$$

Proof. The system of ODEs (1.11) being triangular – though not linear – we tackle the equations one by one, in order.

Solution for A First, we observe that $D_+ = \mathrm{diag}((\alpha_2^{1,b} + \alpha_2^{1,a})z^1, \dots, (\alpha_2^{d,b} + \alpha_2^{d,a})z^d)$ is a positive diagonal matrix. Therefore $D_+^{\frac{1}{2}}$ is well defined. Then, since $D_+^{\frac{1}{2}} \Sigma D_+^{\frac{1}{2}} \in S_d^+$, $\widehat{A}$ is well defined and in S_d^+.

Now, by introducing the change of variables

$$\mathbf{a}(t) = 2D_+^{\frac{1}{2}} A(t) D_+^{\frac{1}{2}},$$

the terminal value problem for A in (1.11) becomes

$$\begin{cases} \mathbf{a}'(t) = \mathbf{a}(t)^2 - \widehat{A}^2 \\ \mathbf{a}(T) = 0. \end{cases} \tag{1.16}$$

[12]This condition is line with our intuition that $p \mapsto H_\xi^{i,b}(p) + H_\xi^{i,a}(-p)$ should be U-shaped, see Section 1.3.

To solve (1.16) let us introduce the function z defined by

$$z(t) = e^{\widehat{A}(T-t)} + e^{-\widehat{A}(T-t)},$$

that is a $C^2([0,T], S_d^{++})$ function verifying $z''(t) = \widehat{A}^2 z(t)$ and $z'(T) = 0$.

We have

$$\frac{d}{dt}\left(-z'(t)z(t)^{-1}\right) = -z''(t)z(t)^{-1} + z'(t)z(t)^{-1}z'(t)z(t)^{-1} = \left(z'(t)z(t)^{-1}\right)^2 - \widehat{A}^2$$

and $-z'(T)z(T)^{-1} = 0$. Therefore, by Cauchy-Lipschitz theorem, we have $\mathbf{a} = -z'z^{-1}$.

Wrapping up, we obtain

$$\begin{aligned} A(t) &= \frac{1}{2} D_+^{-\frac{1}{2}} \mathbf{a}(t) D_+^{-\frac{1}{2}} \\ &= -\frac{1}{2} D_+^{-\frac{1}{2}} z'(t) z(t)^{-1} D_+^{-\frac{1}{2}} \\ &= \frac{1}{2} D_+^{-\frac{1}{2}} \widehat{A} \left(e^{\widehat{A}(T-t)} - e^{-\widehat{A}(T-t)}\right) \left(e^{\widehat{A}(T-t)} + e^{-\widehat{A}(T-t)}\right)^{-1} D_+^{-\frac{1}{2}}. \end{aligned}$$

Solution for B Let us notice that, by definition of the exponential of a matrix, for all $s, t \in [0,T]$, the matrices

$$\widehat{A},\ \left(e^{\widehat{A}(T-s)} - e^{-\widehat{A}(T-s)}\right),\ \left(e^{\widehat{A}(T-s)} + e^{-\widehat{A}(T-s)}\right)^{-1},$$

$$\left(e^{\widehat{A}(T-t)} - e^{-\widehat{A}(T-t)}\right), \text{ and } \left(e^{\widehat{A}(T-t)} + e^{-\widehat{A}(T-t)}\right)^{-1}$$

commute. Therefore

$$\begin{aligned} & A(s)D_+A(t)D_+ \\ = \ & \frac{1}{4} D_+^{-\frac{1}{2}} \widehat{A} \left(e^{\widehat{A}(T-s)} - e^{-\widehat{A}(T-s)}\right) \left(e^{\widehat{A}(T-s)} + e^{-\widehat{A}(T-s)}\right)^{-1} \\ & \times \widehat{A} \left(e^{\widehat{A}(T-t)} - e^{-\widehat{A}(T-t)}\right) \left(e^{\widehat{A}(T-t)} + e^{-\widehat{A}(T-t)}\right)^{-1} D_+^{\frac{1}{2}} \\ = \ & \frac{1}{4} D_+^{-\frac{1}{2}} \widehat{A} \left(e^{\widehat{A}(T-t)} - e^{-\widehat{A}(T-t)}\right) \left(e^{\widehat{A}(T-t)} + e^{-\widehat{A}(T-t)}\right)^{-1} \\ & \times \widehat{A} \left(e^{\widehat{A}(T-s)} - e^{-\widehat{A}(T-s)}\right) \left(e^{\widehat{A}(T-s)} + e^{-\widehat{A}(T-s)}\right)^{-1} D_+^{\frac{1}{2}} \\ = \ & A(t)D_+A(s)D_+. \end{aligned}$$

Therefore, we can apply the method of Variation of Parameters to the linear ODE char-

acterizing B to obtain

$$B(t) = -2e^{-2\int_t^T A(u)D_+\,du}\int_t^T e^{2\int_s^T A(u)D_+\,du}A(s)\left(V_- + D_-\mathcal{D}(A(s))\right)ds.$$

Solution for C We simply integrate the ODE characterizing C to obtain (1.15).

□

From Eqs. (1.13), (1.14), and (1.15), we can deduce the asymptotic behaviour of (A, B, C) when T goes to infinity.

Proposition 3. Let (A, B, C) be the solution of the system of ODEs (1.11) with terminal conditions (1.12).

Then,

$$\begin{aligned}
A(0) &\overset{T\to+\infty}{\longrightarrow} \frac{1}{2}\sqrt{\gamma}\Gamma,\\
B(0) &\overset{T\to+\infty}{\longrightarrow} -D_+^{-\frac{1}{2}}\widehat{A}\widehat{A}^+D_+^{-\frac{1}{2}}\left(V_- + \frac{1}{2}\sqrt{\gamma}D_-\mathcal{D}(\Gamma)\right),\\
\frac{C(0)}{T} &\overset{T\to+\infty}{\longrightarrow} -\mathrm{Tr}\left(D^b_{0,1}+D^a_{0,1}\right) - \frac{1}{2}\sqrt{\gamma}\mathrm{Tr}\left(\left(D^b_{1,2}+D^a_{1,2}\right)\Gamma\right)\\
&\quad + V_-^\intercal D_+^{-\frac{1}{2}}\widehat{A}\widehat{A}^+D_+^{-\frac{1}{2}}\left(V_- + \frac{1}{2}\sqrt{\gamma}D_-\mathcal{D}(\Gamma)\right) - \frac{1}{8}\gamma\mathcal{D}(\Gamma)^\intercal\left(D^b_{2,3}+D^a_{2,3}\right)\mathcal{D}(\Gamma)\\
&\quad - \frac{1}{2}\left(V_- + \frac{1}{2}\sqrt{\gamma}D_-\mathcal{D}(\Gamma)\right)^\intercal D_+^{-\frac{1}{2}}\widehat{A}\widehat{A}^+D_+^{-\frac{1}{2}}\left(V_- + \frac{1}{2}\sqrt{\gamma}D_-\mathcal{D}(\Gamma)\right)\\
&\quad + \frac{1}{2}\sqrt{\gamma}\left(V_- + \frac{1}{2}\sqrt{\gamma}D_-\mathcal{D}(\Gamma)\right)^\intercal D_+^{-\frac{1}{2}}\widehat{A}\widehat{A}^+D_+^{-\frac{1}{2}}D_-\mathcal{D}(\Gamma),
\end{aligned}$$

where $\Gamma = D_+^{-\frac{1}{2}}\left(D_+^{\frac{1}{2}}\Sigma D_+^{\frac{1}{2}}\right)^{\frac{1}{2}}D_+^{-\frac{1}{2}}$ and $\widehat{A}^+$ is the Moore-Penrose generalized inverse of $\widehat{A}$.

Proof. This proof is divided into three parts corresponding to the derivation of the asymptotic expression for A, B, and C, respectively.

Asymptotics for A Let us recall first that $\widehat{A} = \sqrt{\gamma}\left(D_+^{\frac{1}{2}}\Sigma D_+^{\frac{1}{2}}\right)^{\frac{1}{2}} \in S_d^+$. Therefore, there exists an orthogonal matrix P and there exists a diagonal matrix with nonnegative entries $\mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ such that $\widehat{A} = P\mathrm{diag}(\lambda_1, \ldots, \lambda_d)P^\intercal$. From Eq. (1.13) we have

$$A(0) = \frac{1}{2}D_+^{-\frac{1}{2}}P\mathrm{diag}\left(\lambda_1\tanh(\lambda_1 T), \ldots, \lambda_d\tanh(\lambda_d T)\right)P^\intercal D_+^{-\frac{1}{2}}.$$

As $\lambda \tanh(\lambda T) \overset{T\to+\infty}{\longrightarrow} \begin{cases} 0, & \text{if } \lambda = 0 \\ \lambda, & \text{if } \lambda > 0 \end{cases}$, we clearly have

$$A(0) \overset{T\to+\infty}{\longrightarrow} \frac{1}{2} D_+^{-\frac{1}{2}} P \mathrm{diag}(\lambda_1, \dots, \lambda_d) P^\intercal D_+^{-\frac{1}{2}} = \frac{1}{2} D_+^{-\frac{1}{2}} \widehat{A} D_+^{-\frac{1}{2}} = \frac{1}{2}\sqrt{\gamma}\Gamma.$$

Asymptotics for B From Eq. (1.14), we have

$$\begin{aligned} B(0) &= -2e^{-2\int_0^T A(u)D_+\,du} \int_0^T e^{2\int_s^T A(u)D_+\,du} A(s)\left(V_- + D_-\mathcal{D}(A(s))\right) ds \\ &= -2e^{-2\int_0^T \tilde{A}(u)D_+\,du} \int_0^T e^{2\int_0^s \tilde{A}(u)D_+\,du} \tilde{A}(s)\left(V_- + D_-\mathcal{D}(\tilde{A}(s))\right) ds \end{aligned}$$

where $\tilde{A} : t \mapsto \frac{1}{2} D_+^{-\frac{1}{2}} \widehat{A} \left(e^{\widehat{A}t} - e^{-\widehat{A}t}\right)\left(e^{\widehat{A}t} + e^{-\widehat{A}t}\right)^{-1} D_+^{-\frac{1}{2}}$.

Using the spectral decomposition of $\widehat{A}$ introduced in the above paragraph, we see that

$$2\tilde{A}(u)D_+ = D_+^{-\frac{1}{2}} P \mathrm{diag}\left(\lambda_1 \tanh(\lambda_1 u), \dots, \lambda_d \tanh(\lambda_d u)\right) P^\intercal D_+^{\frac{1}{2}}$$

and therefore, after integration,

$$e^{2\int_0^T \tilde{A}(u)D_+\,du} = D_+^{-\frac{1}{2}} P \mathrm{diag}\left(\cosh(\lambda_1 T), \dots, \cosh(\lambda_d T)\right) P^\intercal D_+^{\frac{1}{2}}$$

and

$$e^{2\int_0^s \tilde{A}(u)D_+\,du}\tilde{A}(s) = \frac{1}{2} D_+^{-\frac{1}{2}} P \mathrm{diag}\left(\lambda_1 \sinh(\lambda_1 s), \dots, \lambda_d \sinh(\lambda_d s)\right) P^\intercal D_+^{-\frac{1}{2}}$$

Wrapping up, we get that $B(0)$ is equal to

$$\begin{aligned} & -\int_0^T D_+^{-\frac{1}{2}} P \mathrm{diag}\left(\lambda_1 \frac{\sinh(\lambda_1 s)}{\cosh(\lambda_1 T)}, \dots, \lambda_d \frac{\sinh(\lambda_d s)}{\cosh(\lambda_d T)}\right) P^\intercal D_+^{-\frac{1}{2}} \left(V_- + D_-\mathcal{D}(\tilde{A}(s))\right) ds \\ = & -\int_0^T D_+^{-\frac{1}{2}} P \mathrm{diag}\left(\lambda_1 \frac{\sinh(\lambda_1 s)}{\cosh(\lambda_1 T)}, \dots, \lambda_d \frac{\sinh(\lambda_d s)}{\cosh(\lambda_d T)}\right) P^\intercal D_+^{-\frac{1}{2}} \left(V_- + \frac{1}{2}\sqrt{\gamma} D_-\mathcal{D}(\Gamma)\right) ds \\ & + \int_0^T D_+^{-\frac{1}{2}} P \mathrm{diag}\left(\lambda_1 \frac{\sinh(\lambda_1 s)}{\cosh(\lambda_1 T)}, \dots, \lambda_d \frac{\sinh(\lambda_d s)}{\cosh(\lambda_d T)}\right) P^\intercal D_+^{-\frac{1}{2}} \left(\frac{1}{2}\sqrt{\gamma} D_-\mathcal{D}(\Gamma) - D_-\mathcal{D}(\tilde{A}(s))\right) ds \\ = & D_+^{-\frac{1}{2}} P \mathrm{diag}\left(1 - \frac{1}{\cosh(\lambda_1 T)}, \dots, 1 - \frac{1}{\cosh(\lambda_d T)}\right) P^\intercal D_+^{-\frac{1}{2}} \left(V_- + \frac{1}{2}\sqrt{\gamma} D_-\mathcal{D}(\Gamma)\right) + J(T), \end{aligned}$$

where

$$J(T) = \int_0^T D_+^{-\frac{1}{2}} P \mathrm{diag}\left(\lambda_1 \frac{\sinh(\lambda_1 s)}{\cosh(\lambda_1 T)}, \dots, \lambda_d \frac{\sinh(\lambda_d s)}{\cosh(\lambda_d T)}\right) P^\intercal D_+^{-\frac{1}{2}} \left(\frac{1}{2}\sqrt{\gamma} D_-\mathcal{D}(\Gamma) - D_-\mathcal{D}(\tilde{A}(s))\right) ds.$$

Let us prove that $J(T) \overset{T\to+\infty}{\longrightarrow} 0$. For that purpose, let us consider $\epsilon > 0$ and let us notice

that there exists $\tau > 0$ such that $\forall s > \tau, \|\frac{1}{2}\sqrt{\gamma}D_+^{-1}D_-\mathcal{D}(\Gamma) - D_+^{-1}D_-\mathcal{D}(\tilde{A}(s))\| \leq \epsilon$, where the norm used is the Euclidian norm on $\mathbb{R}^d$. Let us also denote by M the quantity $\sup_{s\geq 0}\|\frac{1}{2}\sqrt{\gamma}D_-\mathcal{D}(\Gamma) - D_-\mathcal{D}(\tilde{A}(s))\|$.

Using the operator norm (still denoted by $\|\cdot\|$) associated with the Euclidian norm on $\mathbb{R}^d$ and its well-known link with the spectral radius, we see that for $T > \tau$,

$$\begin{aligned}
&\|J(T)\| \\
&\leq \int_0^T \left\| D_+^{-\frac{1}{2}} P \mathrm{diag}\left(\lambda_1 \frac{\sinh(\lambda_1 s)}{\cosh(\lambda_1 T)}, \ldots, \lambda_d \frac{\sinh(\lambda_d s)}{\cosh(\lambda_d T)}\right) P^\intercal D_+^{\frac{1}{2}} \right\| \\
&\qquad\qquad \left\| D_+^{-1}\frac{1}{2}\sqrt{\gamma}D_-\mathcal{D}(\Gamma) - D_+^{-1}D_-\mathcal{D}(\tilde{A}(s)) \right\| ds \\
&\leq \int_0^T \left(\max_i \lambda_i \frac{\sinh(\lambda_i s)}{\cosh(\lambda_i T)}\right) \left\| \frac{1}{2}\sqrt{\gamma}D_+^{-1}D_-\mathcal{D}(\Gamma) - D_+^{-1}D_-\mathcal{D}(\tilde{A}(s)) \right\| ds \\
&\leq M \int_0^\tau \max_i \lambda_i \frac{\sinh(\lambda_i s)}{\cosh(\lambda_i T)} ds + \epsilon \int_\tau^T \max_i \lambda_i \frac{\sinh(\lambda_i s)}{\cosh(\lambda_i T)} ds.
\end{aligned}$$

By defining $\overline{\lambda} = \max\{\lambda_1, \ldots, \lambda_d\}$ and $\underline{\lambda} = \min\{\lambda_i | \forall i \in \{1, \ldots, d\}, \lambda_i > 0\}$, we have

$$\max_{i\in\{1,\ldots,d\}} \lambda_i \frac{\sinh(\lambda_i s)}{\cosh(\lambda_i T)} \leq \max_{i\in\{1,\ldots,d\}} \lambda_i \frac{e^{\lambda_i s}}{e^{\lambda_i T}} = \max_{i\in\{1,\ldots,d\}, \lambda_i > 0} \lambda_i e^{-\lambda_i (T-s)} \leq \overline{\lambda} e^{-\underline{\lambda}(T-s)}.$$

Therefore,

$$\begin{aligned}
& \limsup_{T\to\infty} \|J(T)\| \\
\leq\ & M \limsup_{T\to\infty} \overline{\lambda}\left(e^{-\underline{\lambda}(T-\tau)} - e^{-\underline{\lambda}T}\right) + \epsilon \limsup_{T\to\infty} \overline{\lambda}\left(1 - e^{-\underline{\lambda}(T-\tau)}\right) \\
\leq\ & \epsilon
\end{aligned}$$

which allows to conclude that $J(T) \overset{T\to+\infty}{\longrightarrow} 0$.

Now, as $P\mathrm{diag}\left(1 - \frac{1}{\cosh(\lambda_1 T)}, \ldots, 1 - \frac{1}{\cosh(\lambda_d T)}\right) P^\intercal$ converges toward the orthogonal projector on $\mathrm{Im}(\widehat{A})$, which is also given by $\widehat{A}\widehat{A}^+$, we conclude that

$$B(0) \overset{T\to+\infty}{\longrightarrow} -D_+^{-\frac{1}{2}}\widehat{A}\widehat{A}^+ D_+^{-\frac{1}{2}}\left(V_- + \frac{1}{2}\sqrt{\gamma}D_-\mathcal{D}(\Gamma)\right).$$

Asymptotics for C The asymptotic behavior of C is a straightforward consequence of that of A and B. □

1.3.3 From value functions to heuristics and quotes

Motivation for closed-form approximations

An approximation in closed-form of the value function can be motivated by its numerous applications. In the following, we highlight three of them.

First, it can serve as a heuristic evaluation function in reinforcement learning algorithms. Indeed, in problems where the time horizon is too far away to consider full exploration in time, it is often useful, when using Monte-Carlo-based reinforcement learning techniques, to proxy the value of states in a tractable way – analogous to algorithms such as Deep Blue. The above closed-form approximations can be used for that purpose. Moreover, because the value of $C(t)$ is irrelevant for comparing two states (it vanishes when computing the difference in the value function between two points), it is sometimes possible, especially when T is large, to consider the asymptotic expression

$$-\frac{1}{2}\sqrt{\gamma}q^\intercal\Gamma q + q^\intercal D_+^{-\frac{1}{2}}\widehat{A}\widehat{A}^+D_+^{-\frac{1}{2}}\left(V_- + \frac{1}{2}\sqrt{\gamma}D_-\mathcal{D}(\Gamma)\right)$$

instead of $\check{\theta}(t,q)$.

Second, a closed-form approximation of the value function can be used as a starting point in iterative methods designed to compute the value function (value iteration algorithm, actor-critic algorithms, etc.). Unlike for the above use, the value of $C(t)$ matters in that case.

A third important application, and the one that initially motivated this chapter, is for computing policies (quotes, in our case). Indeed, a policy can be deduced from an approximation of the value function by computing the greedy strategy associated with that approximation. In our market making problem, the quotes obtained in this way are not only easy to compute, but also have the advantage of being easily interpretable.

Quotes: the general case

The greedy quoting strategy associated with our closed-form proxy of the value function leads to the following quotes for all $i \in \{1,\dots,d\}$:

$$
\begin{aligned}
\check{\delta}_t^{i,b} &= \tilde{\delta}_\xi^{i,b*}\left(\frac{\check{\theta}(t,q_{t-})-\check{\theta}(t,q_{t-}+z^ie^i)}{z^i}\right)\\
&= \tilde{\delta}_\xi^{i,b*}\left(2q_{t-}^\intercal A(t)e^i + z^i e^{i\intercal}A(t)e^i + e^{i\intercal}B(t)\right),\\
\check{\delta}_t^{i,a} &= \tilde{\delta}_\xi^{i,a*}\left(\frac{\check{\theta}(t,q_{t-})-\check{\theta}(t,q_{t-}-z^ie^i)}{z^i}\right)\\
&= \tilde{\delta}_\xi^{i,a*}\left(-2q_{t-}^\intercal A(t)e^i + z^i e^{i\intercal}A(t)e^i - e^{i\intercal}B(t)\right),
\end{aligned}
$$

where $\tilde{\delta}_\xi^{i,b*}$ and $\tilde{\delta}_\xi^{i,a*}$ are given in Theorems 1 and 2 for $\xi=\gamma$ and $\xi=0$ respectively (depending on whether one considers Model A or Model B).

The asymptotic regime exhibited in the above paragraphs can then serve to obtain the following simple closed-form approximations:

$$
\check{\delta}_t^{i,b} = \tilde{\delta}_\xi^{i,b*}\left(\sqrt{\gamma}q_{t-}^\intercal\Gamma e^i + \frac{1}{2}\sqrt{\gamma}z^ie^{i\intercal}\Gamma e^i - e^{i\intercal}D_+^{-\frac{1}{2}}\widehat{A}\widehat{A}^+D_+^{-\frac{1}{2}}\left(V_- + \frac{1}{2}\sqrt{\gamma}D_-\mathcal{D}(\Gamma)\right)\right), \tag{1.17}
$$

$$
\check{\delta}_t^{i,a} = \tilde{\delta}_\xi^{i,a*}\left(-\sqrt{\gamma}q_{t-}^\intercal\Gamma e^i + \frac{1}{2}\sqrt{\gamma}z^ie^{i\intercal}\Gamma e^i + e^{i\intercal}D_+^{-\frac{1}{2}}\widehat{A}\widehat{A}^+D_+^{-\frac{1}{2}}\left(V_- + \frac{1}{2}\sqrt{\gamma}D_-\mathcal{D}(\Gamma)\right)\right). \tag{1.18}
$$

It is interesting to notice here that the closed-form approximation of the optimal bid and ask quotes for asset i depend on the current value of the inventory through the term $q_{t-}^\intercal\Gamma e^i$. Since $\Gamma\in S_d^+$ and the functions $\tilde{\delta}_\xi^{i,b*}$ and $\tilde{\delta}_\xi^{i,a*}$ are monotone, we have that, all else equal, the quotes for asset i depend monotonically on the inventory in asset i (the bid and ask prices decrease (resp. increase) when the inventory is positive (resp. negative)). The dependence on the inventory in other assets is more subtle as it is linked to the matrix $\Gamma = D_+^{-\frac{1}{2}}(D_+^{\frac{1}{2}}\Sigma D_+^{\frac{1}{2}})^{\frac{1}{2}}D_+^{-\frac{1}{2}}$ which models the complex interplay between price risk (via Σ) and liquidity (via D_+). Recall that D_+ depends on the trade size and α_2, which in turn can be viewed as $H''(0)$ (see Section 1.3) that is an expression involving the trading intensity function, as shown in Lemma 3.1 in Guéant (2017) for a particular case of our model. Also, as already noted in Guéant et al. (2013) the influence of the risk aversion parameter γ is ambiguous and depends on the value of inventories.

In the case of symmetric intensities, i.e. when $\Lambda^{i,b}=\Lambda^{i,a}$ for all $i\in\{1,\dots,d\}$, the Hamiltonian functions $H_\xi^{i,b}$ and $H_\xi^{i,a}$ given in Eqs. (1.3) and (1.4) are identical and thus it is natural to set $\check{H}^{i,b}=\check{H}^{i,a}$ for all $i\in\{1,\dots,d\}$. In that case, (1.17) and (1.18)

simplify into

$$\breve{\delta}_t^{i,b} = \tilde{\delta}_\xi^{i,b*}\left(\sqrt{\gamma} q_{t-}^\intercal \Gamma e^i + \frac{1}{2}\sqrt{\gamma} z^i e^{i\intercal}\Gamma e^i\right), \tag{1.19}$$

$$\breve{\delta}_t^{i,a} = \tilde{\delta}_\xi^{i,a*}\left(-\sqrt{\gamma} q_{t-}^\intercal \Gamma e^i + \frac{1}{2}\sqrt{\gamma} z^i e^{i\intercal}\Gamma e^i\right). \tag{1.20}$$

All these approximations of the optimal quotes can be used directly or as starting points in iterative methods designed to compute the optimal quotes (policy iteration algorithm, actor-critic algorithms, etc.).

Quotes: the case of symmetric exponential intensities

Exponential intensity functions play an important role in the optimal market making literature and more generally in the algorithmic trading literature. This shape of intensity functions, initially proposed by Avellaneda and Stoikov (2008), leads indeed to simplification because of the form of the associated Hamiltonian functions.

If we assume that the intensity functions are given, for all $i \in \{1, \ldots, d\}$, by

$$\Lambda^{i,b}(\delta) = \Lambda^{i,a}(\delta) = A^i e^{-k^i \delta}, \quad A^i, k^i > 0,$$

then (see Guéant (2017)) the Hamiltonian functions are given, for all $i \in \{1, \ldots, d\}$, by

$$H_\xi^{i,b}(p) = H_\xi^{i,a}(p) = \frac{A^i}{k^i} C_\xi^i \exp(-k^i p),$$

where

$$C_\xi^i = \begin{cases} \left(1 + \frac{\xi z^i}{k^i}\right)^{-\left(1+\frac{k^i}{\xi z^i}\right)} & \text{if } \xi > 0 \\ e^{-1} & \text{if } \xi = 0, \end{cases}$$

and the functions $\tilde{\delta}_\xi^{i,b*}$ and $\tilde{\delta}_\xi^{i,a*}$ are given, for all $i \in \{1, \ldots, d\}$, by

$$\tilde{\delta}_\xi^{i,b*}(p) = \tilde{\delta}_\xi^{i,a*}(p) = \begin{cases} p + \frac{1}{\xi z^i} \log\left(1 + \frac{\xi z^i}{k^i}\right) & \text{if } \xi > 0 \\ p + \frac{1}{k^i} & \text{if } \xi = 0. \end{cases}$$

Therefore, if we consider the quadratic approximation of the Hamiltonian functions based on their Taylor expansion around $p = 0$ (see Remark 1), then (1.19) and (1.20) become

$$\breve{\delta}_t^{i,b} = \begin{cases} \sqrt{\gamma}\left(q_{t-}^\intercal \Gamma e^i + \frac{1}{2} z^i e^{i\intercal} \Gamma e^i\right) + \frac{1}{\gamma z^i} \log\left(1 + \frac{\gamma z^i}{k^i}\right) & \text{in Model A,} \\ \sqrt{\gamma}\left(q_{t-}^\intercal \Gamma e^i + \frac{1}{2} z^i e^{i\intercal} \Gamma e^i\right) + \frac{1}{k^i} & \text{in Model B.} \end{cases}$$

$$\breve{\delta}_t^{i,a} = \begin{cases} -\sqrt{\gamma}\left(q_{t-}^\intercal \Gamma e^i - \frac{1}{2} z^i e^{i\intercal} \Gamma e^i\right) + \frac{1}{\gamma z^i} \log\left(1 + \frac{\gamma z^i}{k^i}\right) & \text{in Model A,} \\ -\sqrt{\gamma}\left(q_{t-}^\intercal \Gamma e^i - \frac{1}{2} z^i e^{i\intercal} \Gamma e^i\right) + \frac{1}{k^i} & \text{in Model B.} \end{cases}$$

where $\Gamma = D_+^{-\frac{1}{2}} (D_+^{\frac{1}{2}} \Sigma D_+^{\frac{1}{2}})^{\frac{1}{2}} D_+^{-\frac{1}{2}}$ and $D_+ = \mathrm{diag}(2A^1 C_\xi^1 k^1 z^1, \ldots, 2A^d C_\xi^d k^d z^d)$.

It is noteworthy that these approximations of the optimal quotes are affine in the current inventory. In particular, in the case of Model A, when the number of assets is reduced to one (with unitary transaction size), they coincide with the affine closed-form approximations obtained in the paper by Guéant et al. (2013). Their approximations, however, are obtained in a fundamentally different manner, by using spectral arguments and a continuous approximation of the initial discrete problem.

Another useful point of view on the above quoting strategy is by observing the resulting approximations of the optimal (half) bid-ask spread and skew. The approximations of the optimal (half) bid-ask spread and skew for asset i are respectively given by

$$\frac{\breve{\delta}_t^{i,a} + \breve{\delta}_t^{i,b}}{2} = \begin{cases} \frac{1}{2}\sqrt{\gamma} z^i e^{i\intercal} \Gamma e^i + \frac{1}{\gamma z^i} \log\left(1 + \frac{\gamma z^i}{k^i}\right) & \text{in Model A,} \\ \frac{1}{2}\sqrt{\gamma} z^i e^{i\intercal} \Gamma e^i + \frac{1}{k^i} & \text{in Model B,} \end{cases}$$

and

$$\frac{\breve{\delta}_t^{i,a} - \breve{\delta}_t^{i,b}}{2} = -\sqrt{\gamma} q_{t-}^\intercal \Gamma e^i \text{ in both Model A and Model B.}$$

These approximations give us a constant bid-ask spread and a skew linear in the inventory. This translates well the intuition that the skew has the role of inventory risk management, whereas the spread balances the trade-off between frequency of transactions and profit per round-trip trade (the term $\frac{1}{\gamma z^i} \log\left(1 + \frac{\gamma z^i}{k^i}\right)$ in Model A, which reduces to $\frac{1}{k^i}$ in the case of Model B[13]), plus an additional risk aversion buffer (the term $\frac{1}{2}\sqrt{\gamma} z^i e^{i\intercal} \Gamma e^i$).

On the parameter sensitivity analysis, beyond the above remarks, the reader is referred to Guéant et al. (2013) for a comprehensive analysis in the single-asset case. The complex interplay between price risk and liquidity expressed through Γ, as mentioned in Section 1.3.3, makes the sensitivity analysis less obvious in the multi-asset case.

[13]It is noteworthy that in the case of Model B the bid-ask spread is a nondecreasing function of the risk aversion parameter γ.

1.4 Beyond the quadratic approximation: towards a correction term

In Section 1.3, we approximated the Hamiltonian functions by quadratic functions in order to "approximate" the Hamilton-Jacobi equation characterizing the value function and then approximate the value function itself. To go further, we can consider a perturbative approach around our quadratic approximation. This means that we regard the real Hamiltonian functions as small perturbations of the quadratic functions used to approximate them and consider then a first order approximation (the zero-th order approximation being then that obtained in Section 1.3).

Formally, writing

$$H_\xi^{i,b}(p) = \check{H}^{i,b}(p) + \epsilon h^{i,b}(p), \quad H_\xi^{i,a}(p) = \check{H}^{i,a}(p) + \epsilon h^{i,a}(p), \quad \text{and} \quad \theta(t,q) = \check{\theta}(t,q) + \epsilon \eta(t,q),$$

and plugging these expressions in Eq. (1.5) in the limit case where $Q^i = +\infty$ for all $i \in \{1, \dots, d\}$, we obtain

$$\begin{aligned}
0 &= \partial_t \theta(t,q) - \frac{1}{2}\gamma q^\intercal \Sigma q + \sum_{i=1}^d z^i H_\xi^{i,b}\left(\frac{\theta(t,q) - \theta(t,q+z^i e^i)}{z^i}\right) + \sum_{i=1}^d z^i H_\xi^{i,a}\left(\frac{\theta(t,q) - \theta(t,q-z^i e^i)}{z^i}\right) \\
&= \partial_t \check{\theta}(t,q) - \frac{1}{2}\gamma q^\intercal \Sigma q + \sum_{i=1}^d z^i \check{H}^{i,b}\left(\frac{\check{\theta}(t,q) - \check{\theta}(t,q+z^i e^i)}{z^i}\right) + \sum_{i=1}^d z^i \check{H}^{i,a}\left(\frac{\check{\theta}(t,q) - \check{\theta}(t,q-z^i e^i)}{z^i}\right) \\
&\quad + \epsilon \left(\partial_t \eta(t,q) + \sum_{i=1}^d z^i h^{i,b}\left(\frac{\check{\theta}(t,q) - \check{\theta}(t,q+z^i e^i)}{z^i}\right) + \sum_{i=1}^d z^i h^{i,a}\left(\frac{\check{\theta}(t,q) - \check{\theta}(t,q-z^i e^i)}{z^i}\right) \right. \\
&\qquad + \sum_{i=1}^d \check{H}^{i,b\prime}\left(\frac{\check{\theta}(t,q) - \check{\theta}(t,q+z^i e^i)}{z^i}\right)\left(\eta(t,q) - \eta(t,q+z^i e^i)\right) \\
&\qquad \left. + \sum_{i=1}^d \check{H}^{i,a\prime}\left(\frac{\check{\theta}(t,q) - \check{\theta}(t,q-z^i e^i)}{z^i}\right)\left(\eta(t,q) - \eta(t,q-z^i e^i)\right) \right) + o(\epsilon) \\
&= \epsilon \left(\partial_t \eta(t,q) + \sum_{i=1}^d z^i h^{i,b}\left(\frac{\check{\theta}(t,q) - \check{\theta}(t,q+z^i e^i)}{z^i}\right) + \sum_{i=1}^d z^i h^{i,a}\left(\frac{\check{\theta}(t,q) - \check{\theta}(t,q-z^i e^i)}{z^i}\right) \right. \\
&\qquad + \sum_{i=1}^d \left(-\check{H}^{i,b\prime}\left(\frac{\check{\theta}(t,q) - \check{\theta}(t,q+z^i e^i)}{z^i}\right)\right)\left(\eta(t,q+z^i e^i) - \eta(t,q)\right) \\
&\qquad \left. + \sum_{i=1}^d \left(-\check{H}^{i,a\prime}\left(\frac{\check{\theta}(t,q) - \check{\theta}(t,q-z^i e^i)}{z^i}\right)\right)\left(\eta(t,q-z^i e^i) - \eta(t,q)\right) \right) + o(\epsilon).
\end{aligned}$$

Therefore,

$$
\begin{aligned}
0 = \partial_t \eta(t,q) &+ \sum_{i=1}^{d} z^i h^{i,b} \left(\frac{\check{\theta}(t,q) - \check{\theta}(t,q+z^i e^i)}{z^i} \right) + \sum_{i=1}^{d} z^i h^{i,a} \left(\frac{\check{\theta}(t,q) - \check{\theta}(t,q-z^i e^i)}{z^i} \right) \\
&+ \sum_{i=1}^{d} \left(-\check{H}^{i,b\prime} \left(\frac{\check{\theta}(t,q) - \check{\theta}(t,q+z^i e^i)}{z^i} \right) \right) \left(\eta(t,q+z^i e^i) - \eta(t,q) \right) \\
&+ \sum_{i=1}^{d} \left(-\check{H}^{i,a\prime} \left(\frac{\check{\theta}(t,q) - \check{\theta}(t,q-z^i e^i)}{z^i} \right) \right) \left(\eta(t,q-z^i e^i) - \eta(t,q) \right),
\end{aligned}
$$

and we have the terminal condition $\eta(T,q) = 0$.

By Feynman-Kac representation theorem, we have

$$
\begin{aligned}
\eta(t,q) = \mathbb{E}^{\check{\mathbb{P}}} \Bigg[\int_t^T \Bigg(& \sum_{i=1}^{d} z^i h^{i,b} \left(\frac{\check{\theta}(s, q_{s-}^{t,q}) - \check{\theta}(s, q_{s-}^{t,q} + z^i e^i)}{z^i} \right) \\
& + \sum_{i=1}^{d} z^i h^{i,a} \left(\frac{\check{\theta}(s, q_{s-}^{t,q}) - \check{\theta}(s, q_{s-}^{t,q} - z^i e^i)}{z^i} \right) \Bigg) ds \Bigg],
\end{aligned}
$$

where under $\check{\mathbb{P}}$ the process $(q_s^{t,q})_{s\in[t,T]}$ satisfies

$$
dq_s^{t,q} = \sum_{i=1}^{d} z^i (d\check{N}_s^{i,b} - d\check{N}_s^{i,a}) e^i \quad \text{and} \quad q_t^{t,q} = q,
$$

with, for each $i \in \{1, \dots, d\}$, $\check{N}^{i,b}$ and $\check{N}^{i,a}$ constructed like $N^{i,b}$ and $N^{i,a}$ but with respective intensities given at time s by

$$
-\check{H}^{i,b\prime} \left(\frac{\check{\theta}(s, q_{s-}^{t,q}) - \check{\theta}(s, q_{s-}^{t,q} + z^i e^i)}{z^i} \right) \quad \text{and} \quad -\check{H}^{i,a\prime} \left(\frac{\check{\theta}(s, q_{s-}^{t,q}) - \check{\theta}(s, q_{s-}^{t,q} - z^i e^i)}{z^i} \right).
$$

In practice, it means that we can approximate $\theta(t,q)$ by $\check{\theta}(t,q)$ plus a correction term that takes the form of an expectation:

$$
\begin{aligned}
\theta(t,q) \simeq \check{\theta}(t,q) + \mathbb{E}^{\check{\mathbb{P}}} \Bigg[\int_t^T \Bigg(& \sum_{i=1}^{d} z^i \left(H_\xi^{i,b} - \check{H}^{i,b} \right) \left(\frac{\check{\theta}(s, q_{s-}^{t,q}) - \check{\theta}(s, q_{s-}^{t,q} + z^i e^i)}{z^i} \right) \\
& + \sum_{i=1}^{d} z^i \left(H_\xi^{i,a} - \check{H}^{i,a} \right) \left(\frac{\check{\theta}(s, q_{s-}^{t,q}) - \check{\theta}(s, q_{s-}^{t,q} - z^i e^i)}{z^i} \right) \Bigg) ds \Bigg].
\end{aligned}
$$

Of course this new approximation is not a closed-form one. However, the correction term can be computed using a Monte-Carlo simulation for a specific (t,q). In particular, it

means that upon receiving a request for quote from a client and if time permits (which depends on asset class and market conditions), a market maker can perform a Monte-Carlo simulation to obtain an approximation of the value function at the relevant points to compute a quote that might account more accurately for the liquidity of the requested asset than a quote computed using the closed-forms of Section 1.3.3.

1.5 A multi-asset market making model with additional features

1.5.1 A more general model

In Section 1.2.1 we presented a multi-asset extension to the classical single-asset market making model of Avellaneda and Stoikov. This extension can itself be extended to encompass important features of OTC markets. In this section we extend our results to a more general multi-asset market making model with drift in prices to model the views of the market maker, client tiering, distributed requested sizes for each asset and each tier, and fixed transaction costs for each asset and each tier.

In terms of modeling, the addition of drifts to the price processes is straightforward. Formally, we assume that for each $i \in \{1, \dots, d\}$, the dynamics of the price process $(S^i_t)_{t\in\mathbb{R}_+}$ of asset i is now given by

$$dS^i_t = \mu^i dt + \sigma^i dW^i_t,$$

where σ^i and $(W^i_t)_{t\in\mathbb{R}_+}$ are defined as in Section 1.2.1 and where μ^i is a real constant. In what follows, we denote by μ the vector $\mu = \left(\mu^1, \dots, \mu^d\right)^\intercal$.

In OTC markets, market makers often divide their clients into groups, called tiers, for instance because they do not have the same commercial relationship with all clients or because the propensity to transact given a quote differs across clients. In particular, they can answer/stream different quotes to clients from different tiers.[14] Let us denote here by $N \in \mathbb{N}^*$ the number of such tiers.

In addition to introducing tiers, we can drop the assumption of constant request size per asset and consider instead that, for each asset and each tier, the size of the requests at the bid and at the ask are distributed according to known distributions.

[14]There can also be tiers to proxy the existence of trading platforms with different clients and/or different costs.

Mathematically, the bid and ask quotes that the market maker propose are then modeled by the maps

$$S^{i,n,b} : (\omega, t, z) \in \Omega \times [0,T] \times \mathbb{R}_+^* \mapsto S_t^{i,n,b}(\omega, z) \in \mathbb{R} \text{ and}$$
$$S^{i,n,a} : (\omega, t, z) \in \Omega \times [0,T] \times \mathbb{R}_+^* \mapsto S_t^{i,n,a}(\omega, z) \in \mathbb{R},$$

where $i \in \{1, \dots, d\}$ is the index of the asset, $n \in \{1, \dots, N\}$ is the index of the tier, and $z \in \mathbb{R}_+^*$ is the size of the request (in number of assets). In the same vein as in Section 1.2.1, we introduce[15]

$$\delta_t^{i,n,b}(z) = S_t^i - S_t^{i,n,b}(z) \quad \text{and} \quad \delta_t^{i,n,a}(z) = S_t^{i,n,a}(z) - S_t^i,$$

and the maps $(\delta_t^{i,n,b}(.))_{t\in\mathbb{R}_+}$ and $(\delta_t^{i,n,a}(.))_{t\in\mathbb{R}_+}$ are assumed to be $\mathbb{F}$-predictable and bounded from below by a given constant $-\delta_\infty < 0$.[16]

With these new features in mind, we introduce for each asset $i \in \{1, \dots, d\}$ and for each tier $n \in \{1, \dots, N\}$ the processes $(N_t^{i,n,b})_{t\in\mathbb{R}_+}$ and $(N_t^{i,n,a})_{t\in\mathbb{R}_+}$ modeling the number of transactions in asset i with clients from tier n at the bid and at the ask, respectively. They are $\mathbb{R}_+^*$-marked point processes, with respective intensity kernels $(\lambda_t^{i,n,b}(dz))_{t\in\mathbb{R}_+^*}$ and $(\lambda_t^{i,n,a}(dz))_{t\in\mathbb{R}_+^*}$ given by

$$\lambda_t^{i,n,b}(dz) = \Lambda^{i,n,b}(\delta_t^{i,n,b}(z)) \mathbb{1}_{\{q_{t-}^i + z \le Q^i\}} \nu^{i,n,b}(dz) \text{ and}$$
$$\lambda_t^{i,n,a}(dz) = \Lambda^{i,n,a}(\delta_t^{i,n,a}(z)) \mathbb{1}_{\{q_{t-}^i - z \ge -Q^i\}} \nu^{i,n,a}(dz),$$

where $\nu^{i,n,b}$ and $\nu^{i,n,a}$ are the two probability measures representing the distribution of the requested sizes at the bid and at the ask respectively, for asset i and tier n, and where $\Lambda^{i,n,b}$ and $\Lambda^{i,n,a}$ satisfy the same assumptions as those satisfied by the intensity functions of Section 1.2.1.

For asset $i \in \{1, \dots, d\}$, the resulting inventory $(q_t^i)_{t\in\mathbb{R}_+}$ has dynamics

$$dq_t^i = \sum_{n=1}^N \int_{\mathbb{R}_+^*} z N^{i,n,b}(dt, dz) - \sum_{n=1}^N \int_{\mathbb{R}_+^*} z N^{i,n,a}(dt, dz),$$

where for each tier $n \in \{1, \dots, N\}$, $N^{i,n,b}(dt, dz)$ and $N^{i,n,a}(dt, dz)$ are the random measures associated with the processes $(N_t^{i,n,b})_{t\in\mathbb{R}_+}$ and $(N_t^{i,n,a})_{t\in\mathbb{R}_+}$, respectively.

Finally, we consider the addition of fixed transaction costs.[17] For that purpose, we

[15] ω is omitted in what follows.

[16] This additional constraint of a fixed lower bound is just a technical one to be able to state theorems in the general case where request sizes are distributed (see Bergault and Guéant (2019)).

[17] Proportional transaction costs can be considered in the initial model through shifts in the intensity

introduce for each asset $i \in \{1, \dots, d\}$ and for each tier $n \in \{1, \dots, N\}$ two real numbers $c^{i,n,b}$ and $c^{i,n,a}$ modelling the fixed cost of a transaction in asset i with a client from tier n, at the bid and at the ask, respectively.

The resulting cash process $(X_t)_{t \in \mathbb{R}_+}$ has, consequently, the following dynamics:

$$dX_t = \sum_{i=1}^{d} \sum_{n=1}^{N} \int_{\mathbb{R}_+^*} \Big[\Big(\delta_t^{i,n,b}(z) z - c^{i,n,b} \Big) N^{i,n,b}(dt, dz) \\ + \left(\delta_t^{i,n,a}(z) z - c^{i,n,a} \right) N^{i,n,a}(dt, dz) \Big] - \sum_{i=1}^{d} S_t^i dq_t^i.$$

1.5.2 The Hamilton-Jacobi equation

In this new setting, one can again show that the two optimization problems introduced in Section 1.2.1 boil down to the resolution of a Hamilton-Jacobi equation of the form

$$\begin{aligned} 0 = \partial_t \theta(t,q) + \mu^\intercal q - \frac{\gamma}{2} q^\intercal \Sigma q \\ + \sum_{i=1}^{d} \sum_{n=1}^{N} \int_{\mathbb{R}_+^*} \mathbb{1}_{\{q^i + z \leq Q^i\}} z H_\xi^{i,n,b} \left(z, \frac{\theta(t,q) - \theta(t, q + z e^i) + c^{i,n,b}}{z} \right) \nu^{i,n,b}(dz) \\ + \sum_{i=1}^{d} \sum_{n=1}^{N} \int_{\mathbb{R}_+^*} \mathbb{1}_{\{q^i - z \geq -Q^i\}} z H_\xi^{i,n,a} \left(z, \frac{\theta(t,q) - \theta(t, q - z e^i) + c^{i,n,a}}{z} \right) \nu^{i,n,a}(dz), \end{aligned} \tag{1.21}$$

with terminal condition

$$\theta(T, q) = 0, \tag{1.22}$$

where $\xi = \gamma$ in the case of Model A and $\xi = 0$ in the case of Model B, and where the functions $H_\xi^{i,n,b}$ and $H_\xi^{i,n,a}$ are defined by

$$H_\xi^{i,n,b}(z,p) := \begin{cases} \sup\limits_{\delta > -\delta_\infty} \frac{\Lambda^{i,n,b}(\delta)}{\xi z} (1 - \exp(-\xi z(\delta - p))) & \text{if } \xi > 0, \\ \sup\limits_{\delta > -\delta_\infty} \Lambda^{i,n,b}(\delta)(\delta - p) & \text{if } \xi = 0 \end{cases}$$

and

$$H_\xi^{i,n,a}(z,p) := \begin{cases} \sup\limits_{\delta > -\delta_\infty} \frac{\Lambda^{i,n,a}(\delta)}{\xi z} (1 - \exp(-\xi z(\delta - p))) & \text{if } \xi > 0, \\ \sup\limits_{\delta > -\delta_\infty} \Lambda^{i,n,a}(\delta)(\delta - p) & \text{if } \xi = 0. \end{cases}$$

functions.

Remark 2. When $\xi = 0$, the dependence in z of the Hamiltonian functions vanishes. Nevertheless, we keep the variable z for the sake of consistency.

Following a method similar to that developed in Bergault and Guéant (2019), we can show that, for a given $\xi \geq 0$, there exists a unique bounded function $\theta : [0,T] \times \prod_{i=1}^{d} [-Q^i, Q^i] \to \mathbb{R}$, C^1 in time, solution of Eq. (1.21) with terminal condition (1.22).

Moreover, under the (mild) assumption that the measures $\nu^{i,n,b}$ and $\nu^{i,n,a}$ have moments of order 2, a classical verification argument enables to go from θ to optimal controls for both Model A and Model B. The optimal quotes as functions of θ are given by the two theorems that follow.

In the case of Model A, the result is the following:

Theorem 3. *Let us consider the solution θ of Eq. (1.21) with terminal condition (1.22), for $\xi = \gamma$.*

Then, for $i \in \{1, \dots, d\}$ and $n \in \{1, \dots, N\}$, the optimal bid and ask quotes as functions of the trade size z, $S_t^{i,n,b}(z) = S_t^i - \delta_t^{i,n,b*}(z)$ and $S_t^{i,n,a}(z) = S_t^i + \delta_t^{i,n,a*}(z)$ in Model A are characterized by

$$\delta_t^{i,n,b*}(z) = \tilde{\delta}_\gamma^{i,n,b*}\left(z, \frac{\theta(t,q_{t-}) - \theta(t,q_{t-} + ze^i) + c^{i,n,b}}{z}\right) \quad \text{for } q_{t-} + ze^i \in \prod_{j=1}^{d} [-Q^j, Q^j],$$

$$\delta_t^{i,n,a*}(z) = \tilde{\delta}_\gamma^{i,n,a*}\left(z, \frac{\theta(t,q_{t-}) - \theta(t,q_{t-} - ze^i) + c^{i,n,a}}{z}\right) \quad \text{for } q_{t-} - ze^i \in \prod_{j=1}^{d} [-Q^j, Q^j],$$

where the functions $\tilde{\delta}_\gamma^{i,n,b*}(\cdot,\cdot)$ and $\tilde{\delta}_\gamma^{i,n,a*}(\cdot,\cdot)$ are defined by

$$\tilde{\delta}_\gamma^{i,n,b*}(z,p) = {\Lambda^{i,n,b}}^{-1}\left(\gamma z H_\gamma^{i,n,b}(z,p) - \partial_p H_\gamma^{i,n,b}(z,p)\right) \vee (-\delta_\infty),$$
$$\tilde{\delta}_\gamma^{i,n,a*}(z,p) = {\Lambda^{i,n,a}}^{-1}\left(\gamma z H_\gamma^{i,n,a}(z,p) - \partial_p H_\gamma^{i,n,a}(z,p)\right) \vee (-\delta_\infty).$$

For Model B, the result is the following:

Theorem 4. *Let us consider the solution θ of Eq. (1.21) with terminal condition (1.22), for $\xi = 0$.*

Then, for $i \in \{1, \dots, d\}$ and $n \in \{1, \dots, N\}$, the optimal bid and ask quotes as functions of the trade size z, $S_t^{i,n,b}(z) = S_t^i - \delta_t^{i,n,b*}(z)$ and $S_t^{i,n,a}(z) = S_t^i + \delta_t^{i,n,a*}(z)$ in Model B are

characterized by

$$\delta_t^{i,n,b*}(z) = \tilde{\delta}_0^{i,n,b*}\left(z, \frac{\theta(t,q_{t-}) - \theta(t,q_{t-}+ze^i) + c^{i,n,b}}{z}\right) \quad \textit{for } q_{t-}+ze^i \in \prod_{j=1}^{d}[-Q^j,Q^j],$$

$$\delta_t^{i,n,a*}(z) = \tilde{\delta}_0^{i,n,a*}\left(z, \frac{\theta(t,q_{t-}) - \theta(t,q_{t-}-ze^i) + c^{i,n,a}}{z}\right) \quad \textit{for } q_{t-}-ze^i \in \prod_{j=1}^{d}[-Q^j,Q^j],$$

where the functions $\tilde{\delta}_0^{i,n,b*}(\cdot,\cdot)$ *and* $\tilde{\delta}_0^{i,n,a*}(\cdot,\cdot)$ *are defined by*

$$\tilde{\delta}_0^{i,n,b*}(z,p) = {\Lambda^{i,n,b}}^{-1}\left(-\partial_p H_0^{i,n,b}(z,p)\right) \vee (-\delta_\infty) \textit{ and}$$
$$\tilde{\delta}_0^{i,n,a*}(z,p) = {\Lambda^{i,n,a}}^{-1}\left(-\partial_p H_0^{i,n,a}(z,p)\right) \vee (-\delta_\infty).$$

1.5.3 Quadratic approximation

As before, let us replace for all $i \in \{1,\dots,d\}$ and $n \in \{1,\dots,N\}$, the Hamiltonian functions $H_\xi^{i,n,b}$ and $H_\xi^{i,n,a}$ by the functions

$$\check{H}^{i,n,b} : (z,p) \mapsto \alpha_0^{i,n,b}(z) + \alpha_1^{i,n,b}(z)p + \frac{1}{2}\alpha_2^{i,n,b}(z)p^2 \text{ and}$$
$$\check{H}^{i,n,a} : (z,p) \mapsto \alpha_0^{i,n,a}(z) + \alpha_1^{i,n,a}(z)p + \frac{1}{2}\alpha_2^{i,n,a}(z)p^2.$$

Remark 3. Here, $\alpha_j^{i,n,b}$ and $\alpha_j^{i,n,a}$ (for $j \in \{0,1,2\}$) are functions of z. A natural choice for the functions $(\check{H}^{i,n,b})_{(i,n)\in\{1,\dots,d\}\times\{1,\dots,N\}}$ and $(\check{H}^{i,n,a})_{(i,n)\in\{1,\dots,d\}\times\{1,\dots,N\}}$ derives from Taylor expansions around $p=0$. In that case,

$$\forall i \in \{1,\dots,d\}, \forall n \in \{1,\dots,N\}, \forall j \in \{0,1,2\},$$
$$\alpha_j^{i,n,b}(z) = \partial_p^j H_\xi^{i,n,b}(z,0) \quad \text{and} \quad \alpha_j^{i,n,a} = \partial_p^j H_\xi^{i,n,a}(z,0).$$

If we consider the limit case where $Q^i = +\infty$ for all $i \in \{1,\dots,d\}$, Eq. (1.21) then

becomes

$$
\begin{aligned}
0 = \partial_t \check{\theta}(t,q) + \mu^\intercal q - \frac{\gamma}{2} q^\intercal \Sigma q + \sum_{i=1}^{d}\sum_{n=1}^{N} \left(\int_{\mathbb{R}_+^*} z\alpha_0^{i,n,b}(z)\nu^{i,n,b}(dz) + \int_{\mathbb{R}_+^*} z\alpha_0^{i,n,a}(z)\nu^{i,n,a}(dz) \right) \\
+ \sum_{i=1}^{d}\sum_{n=1}^{N} \left(\int_{\mathbb{R}_+^*} \alpha_1^{i,n,b}(z) \left(\check{\theta}(t,q) - \check{\theta}(t,q+ze^i) + c^{i,n,b} \right) \nu^{i,n,b}(dz) \right. \\
\left. + \int_{\mathbb{R}_+^*} \alpha_1^{i,n,a}(z) \left(\check{\theta}(t,q) - \check{\theta}(t,q-ze^i) + c^{i,n,a} \right) \nu^{i,n,a}(dz) \right) \\
+ \frac{1}{2} \sum_{i=1}^{d}\sum_{n=1}^{N} \left(\int_{\mathbb{R}_+^*} \frac{1}{z} \alpha_2^{i,n,b}(z) \left(\check{\theta}(t,q) - \check{\theta}(t,q+ze^i) + c^{i,n,b} \right)^2 \nu^{i,n,b}(dz) \right. \\
\left. + \int_{\mathbb{R}_+^*} \frac{1}{z} \alpha_2^{i,n,a}(z) \left(\check{\theta}(t,q) - \check{\theta}(t,q-ze^i) + c^{i,n,a} \right)^2 \nu^{i,n,a}(dz) \right),
\end{aligned}
\tag{1.23}
$$

with terminal condition

$$
\check{\theta}(T,q) = 0. \tag{1.24}
$$

Using the same ansatz as in Section 1.3, we obtain the following result (we omit the proof as it follows the same logic as for that of Proposition 1):

Proposition 4. Let us introduce for all $i \in \{1,\dots,d\}, n \in \{1,\dots,N\}, j \in \{0,1,2\}, k \in \mathbb{N}$, the following constants:

$$
\Delta_{j,k}^{i,n,b} = \int_{\mathbb{R}_+^*} z^k \alpha_j^{i,n,b}(z)\nu^{i,n,b}(dz) \quad \text{and} \quad \Delta_{j,k}^{i,n,a} = \int_{\mathbb{R}_+^*} z^k \alpha_j^{i,n,a}(z)\nu^{i,n,a}(dz),
$$

$$
V_{j,k}^b = \left(\sum_{n=1}^{N} \Delta_{j,k}^{1,n,b}, \dots, \sum_{n=1}^{N} \Delta_{j,k}^{d,n,b} \right)^\intercal \quad \text{and} \quad V_{j,k}^a = \left(\sum_{n=1}^{N} \Delta_{j,k}^{1,n,a}, \dots, \sum_{n=1}^{N} \Delta_{j,k}^{d,n,a} \right)^\intercal,
$$

$$
\tilde{V}_{j,k}^b = \left(\sum_{n=1}^{N} c^{1,n,b}\Delta_{j,k}^{1,n,b}, \dots, \sum_{n=1}^{N} c^{d,n,b}\Delta_{j,k}^{d,n,b} \right)^\intercal
$$

$$
\text{and} \quad \tilde{V}_{j,k}^a = \left(\sum_{n=1}^{N} c^{1,n,a}\Delta_{j,k}^{1,n,a}, \dots, \sum_{n=1}^{N} c^{d,n,a}\Delta_{j,k}^{d,n,a} \right)^\intercal,
$$

$$
\chi_{j,k}^b = \sum_{i=1}^{d}\sum_{n=1}^{N} \Delta_{j,k}^{i,n,b} \quad \text{and} \quad \chi_{j,k}^a = \sum_{i=1}^{d}\sum_{n=1}^{N} \Delta_{j,k}^{i,n,a},
$$

$$\tilde{\chi}^b_{j,k} = \sum_{i=1}^{d}\sum_{n=1}^{N} c^{i,n,b}\Delta^{i,n,b}_{j,k} \quad \text{and} \quad \tilde{\chi}^a_{j,k} = \sum_{i=1}^{d}\sum_{n=1}^{N} c^{i,n,a}\Delta^{i,n,a}_{j,k},$$
$$\hat{\chi}^b_{j,k} = \sum_{i=1}^{d}\sum_{n=1}^{N} (c^{i,n,b})^2\Delta^{i,n,b}_{j,k} \quad \text{and} \quad \hat{\chi}^a_{j,k} = \sum_{i=1}^{d}\sum_{n=1}^{N} (c^{i,n,a})^2\Delta^{i,n,a}_{j,k},$$

and

$$D^b_{j,k} = \operatorname{diag}\left(\sum_{n=1}^{N}\Delta^{1,n,b}_{j,k},\ldots,\sum_{n=1}^{N}\Delta^{d,n,b}_{j,k}\right) \quad \text{and} \quad D^a_{j,k} = \operatorname{diag}\left(\sum_{n=1}^{N}\Delta^{1,n,a}_{j,k},\ldots,\sum_{n=1}^{N}\Delta^{d,n,a}_{j,k}\right).$$

Let us consider three differentiable functions $A : [0,T] \to S_d^+$, $B : [0,T] \to \mathbb{R}^d$, and $C : [0,T] \to \mathbb{R}$ solutions of the system of ordinary differential equations

$$\begin{cases} A'(t) = & 2A(t)\left(D^b_{2,1} + D^a_{2,1}\right)A(t) - \frac{1}{2}\gamma\Sigma, \\ B'(t) = & \mu + 2A(t)\left(V^b_{1,1} - V^a_{1,1}\right) + 2A(t)\left(D^b_{2,2} - D^a_{2,2}\right)\mathcal{D}(A(t)) \\ & +2A(t)\left(D^b_{2,1} + D^a_{2,1}\right)B(t) + 2A(t)\left(\tilde{V}^b_{2,0} - \tilde{V}^a_{2,0}\right), \\ C'(t) = & \operatorname{Tr}\left(D^b_{0,1} + D^a_{0,1}\right) + \operatorname{Tr}\left(\left(D^b_{1,2} + D^a_{1,2}\right)A(t)\right) + \left(V^b_{1,1} - V^a_{1,1}\right)^\intercal B(t) \\ & + \left(\tilde{\chi}^b_{1,0} + \tilde{\chi}^a_{1,0}\right) + \frac{1}{2}\mathcal{D}(A(t))^\intercal\left(D^b_{2,3} + D^a_{2,3}\right)\mathcal{D}(A(t)) \\ & +B(t)^\intercal\left(D^b_{2,2} - D^a_{2,2}\right)\mathcal{D}(A(t)) + \left(\tilde{V}^b_{2,1} + \tilde{V}^a_{2,1}\right)^\intercal\mathcal{D}(A(t)) \\ & +\frac{1}{2}B(t)^\intercal\left(D^b_{2,1} + D^a_{2,1}\right)B(t) + \left(\tilde{V}^b_{2,0} - \tilde{V}^a_{2,0}\right)^\intercal B(t) \\ & +\frac{1}{2}\left(\hat{\chi}^b_{2,0} + \hat{\chi}^a_{2,0}\right). \end{cases} \tag{1.25}$$

with terminal conditions

$$A(T) = 0, B(T) = 0, \text{ and } C(T) = 0. \tag{1.26}$$

Then $\check{\theta} : (t,q) \in [0,T]\times\mathbb{R}^d \mapsto -q^\intercal A(t)q - q^\intercal B(t) - C(t)$ is solution of Eq. (1.23) with terminal condition (1.24).

We can now solve (1.25) with terminal conditions (1.26) in closed-form. This is the purpose of the following proposition whose proof is omitted (see Proposition 2 for a similar proof).

Proposition 5. Assume $\sum_{n=1}^{N}\Delta^{i,n,b}_{2,1} + \Delta^{i,n,a}_{2,1} > 0$ for all $i \in \{1,\ldots,d\}$. The system of

ODEs (1.25) with terminal conditions (1.26) admits the unique solution

$$
\begin{aligned}
A(t) &= \frac{1}{2}D_+^{-\frac{1}{2}}\widehat{A}\left(e^{\widehat{A}(T-t)}-e^{-\widehat{A}(T-t)}\right)\left(e^{\widehat{A}(T-t)}+e^{-\widehat{A}(T-t)}\right)^{-1}D_+^{-\frac{1}{2}},\\
B(t) &= -2e^{-2\int_t^T A(u)D_+\,du}\int_t^T e^{2\int_s^T A(u)D_+\,du}\left(\frac{1}{2}\mu+A(s)\left(V_-+\tilde{V}_-+D_-\mathcal{D}(A(s))\right)\right)ds,\\
C(t) &= -\mathrm{Tr}\left(D_{0,1}^b+D_{0,1}^a\right)(T-t)-\mathrm{Tr}\left(\left(D_{1,2}^b+D_{1,2}^a\right)\int_t^T A(s)ds\right)-V_-^\intercal\int_t^T B(s)ds\\
&\quad-\frac{1}{2}\int_t^T\mathcal{D}(A(s))^\intercal\left(D_{2,3}^b+D_{2,3}^a\right)\mathcal{D}(A(s))ds-\frac{1}{2}\int_t^T B(s)^\intercal D_+B(s)ds\\
&\quad-\int_t^T B(s)^\intercal D_-\mathcal{D}(A(s))ds-\left(\tilde{\chi}_{1,0}^b+\tilde{\chi}_{1,0}^a\right)(T-t)-\frac{1}{2}\left(\hat{\chi}_{2,0}^b+\hat{\chi}_{2,0}^a\right)(T-t)\\
&\quad-\left(\tilde{V}_{2,1}^b+\tilde{V}_{2,1}^a\right)^\intercal\int_t^T\mathcal{D}(A(s))ds,
\end{aligned}
$$

where

$$
D_+=D_{2,1}^b+D_{2,1}^a,\quad D_-=D_{2,2}^b-D_{2,2}^a,\quad V_-=V_{1,1}^b-V_{1,1}^a,\quad \tilde{V}_-=\tilde{V}_{2,0}^b-\tilde{V}_{2,0}^a
$$

$$
\text{and } \widehat{A}=\sqrt{\gamma}\left(D_+^{\frac{1}{2}}\Sigma D_+^{\frac{1}{2}}\right)^{\frac{1}{2}}.
$$

Now, using the same method as in Section 1.3, we get the following asymptotic results:

Proposition 6. Let (A,B,C) be the solution of the system of ODEs (1.25) with terminal conditions (1.26).

If $D_+^{\frac{1}{2}}\mu\in\mathrm{Im}(\widehat{A})$, then,

$$
\begin{aligned}
A(0) &\overset{T\to+\infty}{\longrightarrow} \frac{1}{2}\sqrt{\gamma}\Gamma,\\
B(0) &\overset{T\to+\infty}{\longrightarrow} -D_+^{-\frac{1}{2}}\widehat{A}^+D_+^{\frac{1}{2}}\mu-D_+^{-\frac{1}{2}}\widehat{A}\widehat{A}^+D_+^{-\frac{1}{2}}\left(V_-+\tilde{V}_-+\frac{1}{2}\sqrt{\gamma}D_-\mathcal{D}(\Gamma)\right),\\
\frac{C(0)}{T} &\overset{T\to+\infty}{\longrightarrow} -\mathrm{Tr}\left(D_{0,1}^b+D_{0,1}^a\right)-\frac{1}{2}\sqrt{\gamma}\mathrm{Tr}\left(\left(D_{1,2}^b+D_{1,2}^a\right)\Gamma\right)+V_-^\intercal D_+^{-\frac{1}{2}}\widehat{A}^+D_+^{\frac{1}{2}}\mu\\
&\quad+V_-^\intercal D_+^{-\frac{1}{2}}\widehat{A}\widehat{A}^+D_+^{-\frac{1}{2}}\left(V_-+\tilde{V}_-+\frac{1}{2}\sqrt{\gamma}D_-\mathcal{D}(\Gamma)\right)-\frac{1}{8}\gamma\mathcal{D}(\Gamma)^\intercal\left(D_{2,3}^b+D_{2,3}^a\right)\mathcal{D}(\Gamma)\\
&\quad-\frac{1}{2}\mu^\intercal D_+^{\frac{1}{2}}\widehat{A}^+\widehat{A}^+D_+^{\frac{1}{2}}\mu-\mu^\intercal D_+^{\frac{1}{2}}\widehat{A}^+D_+^{-\frac{1}{2}}\left(V_-+\tilde{V}_-+\frac{1}{2}\sqrt{\gamma}D_-\mathcal{D}(\Gamma)\right)\\
&\quad-\frac{1}{2}\left(V_-+\tilde{V}_-+\frac{1}{2}\sqrt{\gamma}D_-\mathcal{D}(\Gamma)\right)^\intercal D_+^{-\frac{1}{2}}\widehat{A}\widehat{A}^+D_+^{-\frac{1}{2}}\left(V_-+\tilde{V}_-+\frac{1}{2}\sqrt{\gamma}D_-\mathcal{D}(\Gamma)\right)\\
&\quad+\frac{1}{2}\sqrt{\gamma}\mu^\intercal D_+^{\frac{1}{2}}\widehat{A}^+D_+^{-\frac{1}{2}}D_-\mathcal{D}(\Gamma)\\
&\quad+\frac{1}{2}\sqrt{\gamma}\left(V_-+\tilde{V}_-+\frac{1}{2}\sqrt{\gamma}D_-\mathcal{D}(\Gamma)\right)^\intercal D_+^{-\frac{1}{2}}\widehat{A}\widehat{A}^+D_+^{-\frac{1}{2}}D_-\mathcal{D}(\Gamma)
\end{aligned}
$$

$$- \left(\hat{\chi}^b_{2,0} + \hat{\chi}^a_{2,0}\right) - \frac{1}{2}\left(\hat{\chi}^b_{2,0} + \hat{\chi}^a_{2,0}\right) - \frac{1}{2}\sqrt{\gamma}\left(\tilde{V}^b_{2,1} + \tilde{V}^a_{2,1}\right)^\intercal \mathcal{D}(\Gamma),$$

where $\Gamma = D_+^{-\frac{1}{2}}\left(D_+^{\frac{1}{2}}\Sigma D_+^{\frac{1}{2}}\right)^{\frac{1}{2}} D_+^{-\frac{1}{2}}$ and $\widehat{A}^+$ is the Moore-Penrose generalized inverse of $\widehat{A}$.

Remark 4. The assumption $D_+^{\frac{1}{2}}\mu \in \mathrm{Im}(\widehat{A})$ is satisfied when $\mu = 0$ or when Σ is invertible. If this assumption is not satisfied, then it can be shown that $\frac{B(0)}{T} \overset{T\to+\infty}{\longrightarrow} -D_+^{-\frac{1}{2}}\widehat{A}^+\widehat{A}D_+^{\frac{1}{2}}\mu$. In particular, there is no constant asymptotic approximation of the quotes. In fact, if the assumption $D_+^{\frac{1}{2}}\mu \in \mathrm{Im}(\widehat{A})$ is not satisfied, the market maker may have an incentive to propose very good quotes to clients in order to build portfolios bearing positive return at no risk.

1.5.4 From value functions to heuristics and quotes

Quotes: the general case

The greedy quoting strategy associated with our closed-form proxy of the value function leads to the following quotes for all $i \in \{1,\dots,d\}$ and $n \in \{1,\dots,N\}$:

$$\begin{aligned}
\check{\delta}_t^{i,n,b}(z) &= \tilde{\delta}_\xi^{i,n,b*}\left(z, \frac{\check{\theta}(t,q_{t-}) - \check{\theta}(t,q_{t-}+ze^i) + c^{i,n,b}}{z}\right) \\
&= \tilde{\delta}_\xi^{i,n,b*}\left(z, 2q_{t-}^\intercal A(t)e^i + ze^{i\intercal}A(t)e^i + e^{i\intercal}B(t) + \frac{c^{i,n,b}}{z}\right), \\
\check{\delta}_t^{i,n,a}(z) &= \tilde{\delta}_\xi^{i,n,a*}\left(z, \frac{\check{\theta}(t,q_{t-}) - \check{\theta}(t,q_{t-}-ze^i) + c^{i,n,a}}{z}\right) \\
&= \tilde{\delta}_\xi^{i,n,a*}\left(z, -2q_{t-}^\intercal A(t)e^i + ze^{i\intercal}A(t)e^i - e^{i\intercal}B(t) + \frac{c^{i,n,a}}{z}\right),
\end{aligned}$$

where $\tilde{\delta}_\xi^{i,n,b*}$ and $\tilde{\delta}_\xi^{i,n,a*}$ are given in Theorems 3 and 4 for $\xi = \gamma$ and $\xi = 0$ respectively (depending on whether one considers Model A or Model B).

The asymptotic regime exhibited in the above paragraphs can then serve to obtain the

following simple closed-form approximations:

$$\breve{\delta}_t^{i,n,b}(z) = \tilde{\delta}_\xi^{i,n,b*}\bigg(z, \sqrt{\gamma} q_{t-}^\intercal \Gamma e^i + \frac{1}{2}\sqrt{\gamma} z e^{i\intercal}\Gamma e^i - e^{i\intercal} D_+^{-\frac{1}{2}} \widehat{A}^+ D_+^{\frac{1}{2}}\mu \tag{1.27}$$
$$- e^{i\intercal} D_+^{-\frac{1}{2}} \widehat{A}\widehat{A}^+ D_+^{-\frac{1}{2}} \left(V_- + \tilde{V}_- + \frac{1}{2}\sqrt{\gamma} D_- \mathcal{D}(\Gamma)\right) + \frac{c^{i,n,b}}{z}\bigg),$$

$$\breve{\delta}_t^{i,n,a}(z) = \tilde{\delta}_\xi^{i,n,a*}\bigg(z, -\sqrt{\gamma} q_{t-}^\intercal \Gamma e^i + \frac{1}{2}\sqrt{\gamma} z e^{i\intercal}\Gamma e^i + e^{i\intercal} D_+^{-\frac{1}{2}} \widehat{A}^+ D_+^{\frac{1}{2}}\mu \tag{1.28}$$
$$+ e^{i\intercal} D_+^{-\frac{1}{2}} \widehat{A}\widehat{A}^+ D_+^{-\frac{1}{2}} \left(V_- + \tilde{V}_- + \frac{1}{2}\sqrt{\gamma} D_- \mathcal{D}(\Gamma)\right) + \frac{c^{i,n,a}}{z}\bigg).$$

If we assume that for all $i \in \{1, \dots, d\}$ and for all $n \in \{1, \dots, N\}$ we have $\nu^{i,n,b} = \nu^{i,n,a}$ and $\Lambda^{i,n,b} = \Lambda^{i,n,a}$, then $\forall i \in \{1, \dots, d\}, \forall n \in \{1, \dots, N\}, H^{i,n,b} = H^{i,n,a}$, and it is thus natural to chose symmetric approximations of the Hamiltonian functions, i.e. $\forall i \in \{1, \dots, d\}, \forall n \in \{1, \dots, N\}, \breve{H}^{i,n,b} = \breve{H}^{i,n,a}$. In that case, (1.27) and (1.28) simplify into

$$\breve{\delta}_t^{i,n,b}(z) = \tilde{\delta}_\xi^{i,n,b*}\left(z, \sqrt{\gamma} q_{t-}^\intercal \Gamma e^i + \frac{1}{2}\sqrt{\gamma} z e^{i\intercal}\Gamma e^i - e^{i\intercal} D_+^{-\frac{1}{2}} \widehat{A}^+ D_+^{\frac{1}{2}}\mu + \frac{c^{i,n,b}}{z}\right), \tag{1.29}$$

$$\breve{\delta}_t^{i,n,a}(z) = \tilde{\delta}_\xi^{i,n,a*}\left(z, -\sqrt{\gamma} q_{t-}^\intercal \Gamma e^i + \frac{1}{2}\sqrt{\gamma} z e^{i\intercal}\Gamma e^i + e^{i\intercal} D_+^{-\frac{1}{2}} \widehat{A}^+ D_+^{\frac{1}{2}}\mu + \frac{c^{i,n,a}}{z}\right). \tag{1.30}$$

All these approximations of the quotes can be used directly or as a starting point in iterative methods designed to compute the optimal quotes (policy iteration algorithms, actor-critic algorithms, etc.).

Quotes: the case of symmetric exponential intensities

If we assume that for all $i \in \{1, \dots, d\}$ and for all $n \in \{1, \dots, N\}$ we have $\nu^{i,n,b} = \nu^{i,n,a} =: \nu^{i,n}$ and intensity functions given by

$$\Lambda^{i,n,b}(\delta) = \Lambda^{i,n,a}(\delta) = A^{i,n} e^{-k^{i,n}\delta}, \quad A^{i,n}, k^{i,n} > 0,$$

then (see Guéant (2017)), in the limit case where $\delta_\infty = +\infty$ the Hamiltonian functions are given, for all $i \in \{1, \dots, d\}$ and $n \in \{1, \dots, N\}$, by

$$H_\xi^{i,n,b}(z,p) = H_\xi^{i,n,a}(z,p) = \frac{A^{i,n}}{k^{i,n}} C_\xi^{i,n}(z) \exp(-k^{i,n} p),$$

where

$$C_\xi^{i,n}(z) = \begin{cases} \left(1+\frac{\xi z}{k^{i,n}}\right)^{-\left(1+\frac{k^{i,n}}{\xi z}\right)} & \text{if } \xi > 0 \\ e^{-1} & \text{if } \xi = 0, \end{cases}$$

and the functions $\tilde{\delta}_\xi^{i,n,b*}$ and $\tilde{\delta}_\xi^{i,n,a*}$ are given, for all $i \in \{1,\dots,d\}$ and $n \in \{1,\dots,N\}$, by

$$\tilde{\delta}_\xi^{i,n,b*}(z,p) = \tilde{\delta}_\xi^{i,n,a*}(z,p) = \begin{cases} p + \frac{1}{\xi z}\log\left(1+\frac{\xi z}{k^{i,n}}\right) & \text{if } \xi > 0 \\ p + \frac{1}{k^{i,n}} & \text{if } \xi = 0. \end{cases}$$

Therefore, if we consider the quadratic approximation of the Hamiltonian functions based on their Taylor expansion around $p = 0$ (see Remark 3), then (1.29) and (1.30) become

$$\breve{\delta}_t^{i,n,b}(z) = \begin{cases} \sqrt{\gamma}\left(q_{t-}^\intercal \Gamma e^i + \frac{1}{2} z^i e^{i\intercal}\Gamma e^i - \frac{1}{\gamma} e^{i\intercal} D_+^{-\frac{1}{2}} \widehat{A}^+ D_+^{\frac{1}{2}}\mu\right) + \frac{c^{i,n,b}}{z} + \frac{1}{\gamma z}\log\left(1+\frac{\gamma z}{k^{i,n}}\right) & \text{in Model A,} \\ \sqrt{\gamma}\left(q_{t-}^\intercal \Gamma e^i + \frac{1}{2} z^i e^{i\intercal}\Gamma e^i - \frac{1}{\gamma} e^{i\intercal} D_+^{-\frac{1}{2}} \widehat{A}^+ D_+^{\frac{1}{2}}\mu\right) + \frac{c^{i,n,b}}{z} + \frac{1}{k^{i,n}} & \text{in Model B.} \end{cases}$$

$$\breve{\delta}_t^{i,n,a}(z) = \begin{cases} \sqrt{\gamma}\left(q_{t-}^\intercal \Gamma e^i - \frac{1}{2} z^i e^{i\intercal}\Gamma e^i + \frac{1}{\gamma} e^{i\intercal} D_+^{-\frac{1}{2}} \widehat{A}^+ D_+^{\frac{1}{2}}\mu\right) + \frac{c^{i,n,a}}{z} + \frac{1}{\gamma z}\log\left(1+\frac{\gamma z}{k^{i,n}}\right) & \text{in Model A,} \\ \sqrt{\gamma}\left(q_{t-}^\intercal \Gamma e^i - \frac{1}{2} z^i e^{i\intercal}\Gamma e^i + \frac{1}{\gamma} e^{i\intercal} D_+^{-\frac{1}{2}} \widehat{A}^+ D_+^{\frac{1}{2}}\mu\right) + \frac{c^{i,n,a}}{z} + \frac{1}{k^{i,n}} & \text{in Model B.} \end{cases}$$

where $\Gamma = D_+^{-\frac{1}{2}}(D_+^{\frac{1}{2}}\Sigma D_+^{\frac{1}{2}})^{\frac{1}{2}} D_+^{-\frac{1}{2}}$ and

$$D_+ = \operatorname{diag}\left(2\sum_{n=1}^N A^{1,n}k^{1,n}\int_{\mathbb{R}_+^*} C_\xi^{1,n}(z) z \nu^{1,n}(dz), \dots, 2\sum_{n=1}^N A^{d,n}k^{d,n}\int_{\mathbb{R}_+^*} C_\xi^{d,n}(z) z \nu^{d,n}(dz)\right).$$

1.6 Numerical results

To assess the quality of our approximations, we consider a two-asset example for which we can approximate numerically the true function θ and the optimal quotes. By using a Euler scheme in dimension 3 (one dimension for time and two dimensions for inventory) it is indeed possible to approximate numerically the solution of Hamilton-Jacobi equations on a grid.

1.6.1 Characteristics of our example with two assets

We consider the following characteristics for the two assets:

- Asset prices: $S_0^1 = S_0^2 = 100$ €.
- Drifts: $\mu^1 = 0.1$ € $\cdot\, \text{day}^{-1}$, $\mu^2 = -0.1$ € $\cdot\, \text{day}^{-1}$.

- Volatilities: $\sigma^1 = 1.2$ €$\cdot \text{day}^{-\frac{1}{2}}$, $\sigma^2 = 0.6$ €$\cdot \text{day}^{-\frac{1}{2}}$.
- Correlation: $\rho = 0.5$.

This corresponds to a covariance matrix Σ given by

$$\Sigma = \begin{pmatrix} (\sigma^1)^2 & \rho\sigma^1\sigma^2 \\ \rho\sigma^1\sigma^2 & (\sigma^2)^2 \end{pmatrix} = \begin{pmatrix} 1.44 & 0.36 \\ 0.36 & 0.36 \end{pmatrix}.$$

We consider Model B[18] with time horizon $T = 7$ days and risk aversion parameter $\gamma = 8 \cdot 10^{-6}$ €$^{-1}$.

We consider a framework with one tier only and no transaction costs.

The intensity functions are given for all $i \in \{1, 2\}$ by:

$$\Lambda^{i,b}(\delta) = \Lambda^{i,a}(\delta) = \lambda_{RFQ} \frac{1}{1 + e^{\alpha_\Lambda + \beta_\Lambda \delta}},$$

with $\lambda_{RFQ} = 30$ day^{-1}, $\alpha_\Lambda = 0.7$, and $\beta_\Lambda = 30$ €$^{-1}$. This corresponds to 30 requests per day, a probability of $\frac{1}{1+e^{0.7}} \simeq 33\%$ to trade when the answered quote is the reference price and a probability of $\frac{1}{1+e^{0.4}} \simeq 40\%$ to trade when the answered quote is the reference price improved by 1 cent.

Request sizes are distributed according to a Gamma distribution $\Gamma(\alpha_\mu, \beta_\mu)$ with $\alpha_\mu = 4$ and $\beta_\mu = 4 \cdot 10^{-4}$. This corresponds to an average request size of 10000 assets (i.e. approximately 1000000€) and a standard deviation equal to half the average.

1.6.2 Value function and optimal quotes

In order to discretize the problem, we first approximate the Gamma distribution of sizes with 4 sizes: $z^1 = 6250$, $z^2 = 12500$, $z^3 = 18750$, and $z^4 = 25000$ assets – thereafter refered to by very small, small, large, and very large size – with respective probability $p^1 = 0.534$, $p^2 = 0.350$, $p^3 = 0.097$ and $p^4 = 0.019$. We impose risk limits $Q^1 = 75000$ and $Q^2 = 300000$, i.e. no trade that would result in an inventory q^1 for asset 1 such that $|q^1| > 75000$ is accepted, and similarly no trade that would result in an inventory q^2 for asset 2 such that $|q^2| > 300000$ is accepted.

The solution θ to Eq. (1.21) with terminal condition (1.22) can then be approximated numerically using a monotone implicit Euler scheme on a grid of size $101 \times 25 \times 97$ (101

[18] The results would be similar for Model A.

points in time, 25 points for the inventory of asset 1, and 97 points for the inventory of asset 2).

Because we are mainly interested in asymptotic quotes, it is important to check that the time horizon chosen is sufficienty long. For that purpose, we plot in Figure 1.1 the optimal bid quotes for asset 1 and asset 2 at time $t = 0$ for different values of the inventory. We see that the asymptotic regime is clearly reached and, from now on, we will only consider the value function and the optimal quotes at time $t = 0$.

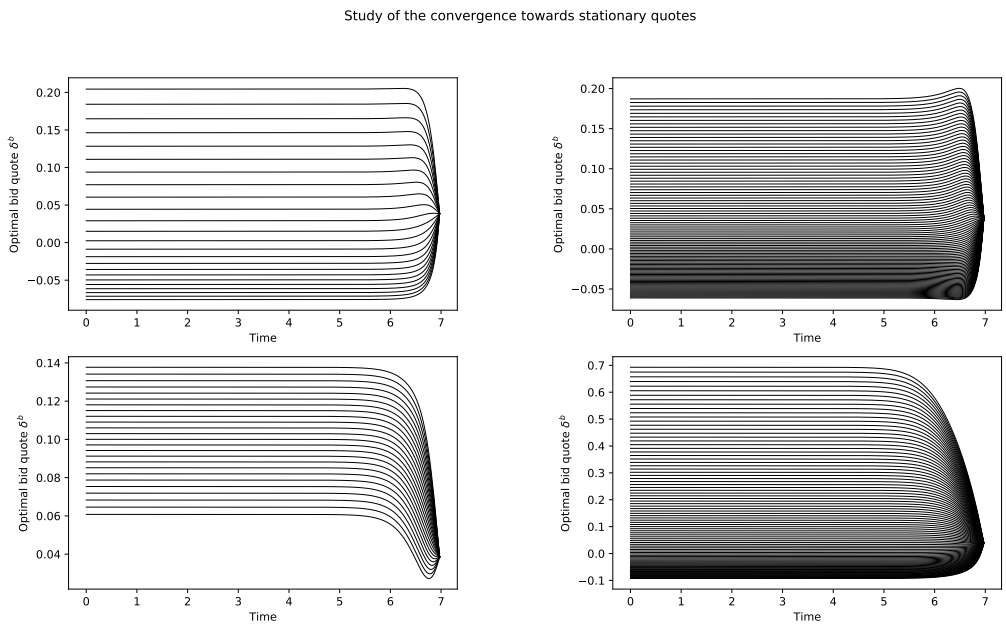


Figure 1.1: Optimal bid quotes as a function of time for different values of the inventory (very small trades) – top left: bid quotes of asset 1 for different values of inventory q^2 ($q^1 = 0$), top right: bid quotes of asset 1 for different values of inventory q^1 ($q^2 = 0$), bottom left: bid quotes of asset 2 for different values of inventory q^2 ($q^1 = 0$), bottom right: bid quotes of asset 2 for different values of inventory q^1 ($q^2 = 0$).

The numerical approximation of the value function θ (at time $t = 0$) is plotted in Figure 1.2. The shape of the function θ is as expected given the risk penalty, the positive drift in the prices of asset 1 and the negative drift in the prices of asset 2. The associated bid quotes are plotted in Figures 1.3 and 1.4 respectively. The shape of the quote surfaces is as expected given the positive correlation coefficient (see Bergault and Guéant (2019) or Guéant (2017) for a deeper discussion about the effect of the different parameters).

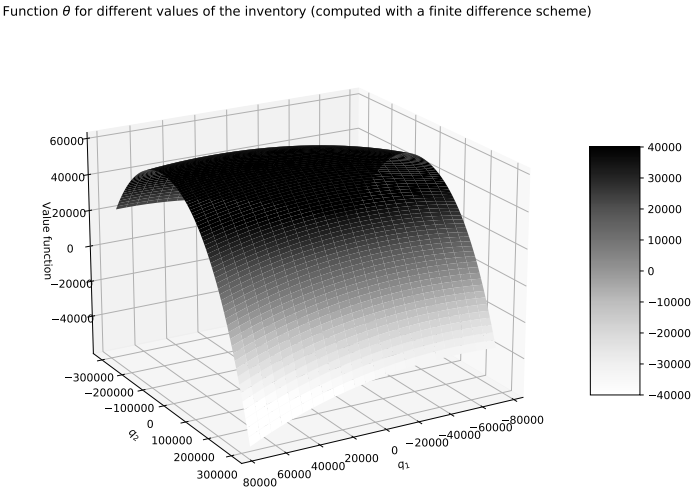


Figure 1.2: Function θ at time $t = 0$ for different values of the inventory.

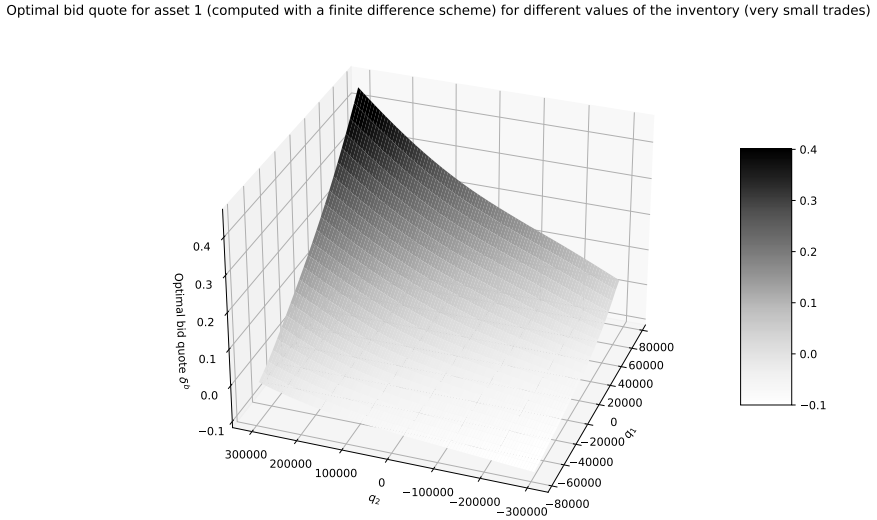


Figure 1.3: Optimal bid quote at $t = 0$ for asset 1 as a function of the inventory (very small trades).

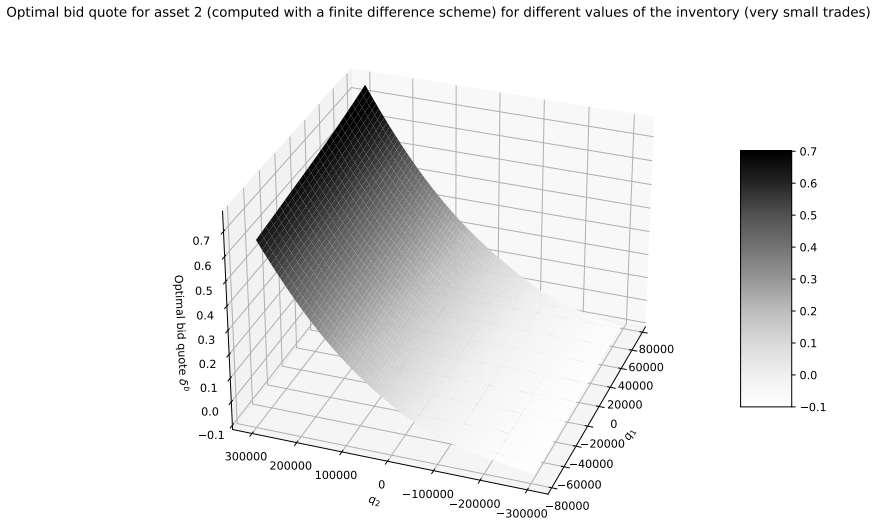


Figure 1.4: Optimal bid quote at $t = 0$ for asset 2 as a function of the inventory (very small trades).

1.6.3 Comparison with closed-form approximations

We now move on to our closed-form approximations. We first plot in Figure 1.5 the closed-form approximation $\check{\theta}$ given by Proposition 4.

We clearly see that, in spite of differences in level between the true value function aproximated numerically and the closed-form approximation, the shape is the same. Therefore, the finite differences involved in the computation of the associated quotes should be similar. This is roughly confirmed in Figures 1.6 and 1.7 that are the closed-form counterparts of Figures 1.3 and 1.4.

In order to assess more precisely the quality of our closed-form approximations, we plot in Figures 1.8, 1.9, 1.10, and 1.11 the functions

$$q^1 \mapsto \bar{\delta}^{1,b}(q^1, 0, z^k),\ k \in \{1, \dots, 4\} \qquad q^1 \mapsto \hat{\delta}^{1,b}(q^1, 0, z^k),\ k \in \{1, \dots, 4\}$$

$$q^2 \mapsto \bar{\delta}^{1,b}(0, q^2, z^k),\ k \in \{1, \dots, 4\} \qquad q^2 \mapsto \hat{\delta}^{1,b}(0, q^2, z^k),\ k \in \{1, \dots, 4\}$$

$$q^1 \mapsto \bar{\delta}^{2,b}(q^1, 0, z^k),\ k \in \{1, \dots, 4\} \qquad q^1 \mapsto \hat{\delta}^{2,b}(q^1, 0, z^k),\ k \in \{1, \dots, 4\}$$

$$q^2 \mapsto \bar{\delta}^{2,b}(0, q^2, z^k),\ k \in \{1, \dots, 4\} \qquad q^2 \mapsto \hat{\delta}^{2,b}(0, q^2, z^k),\ k \in \{1, \dots, 4\}$$

where $\bar{\delta}^{i,b}$ is the optimal bid quote for asset i as a function of time, inventory, and size of

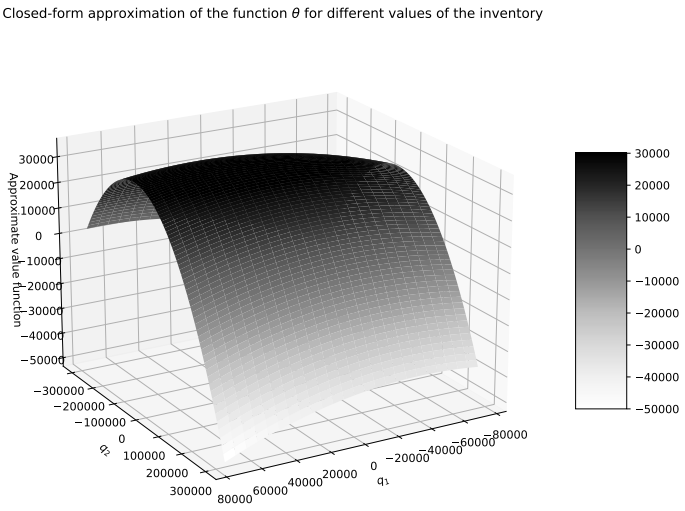


Figure 1.5: Function $\check{\theta}$ at $t = 0$ for different values of the inventory.

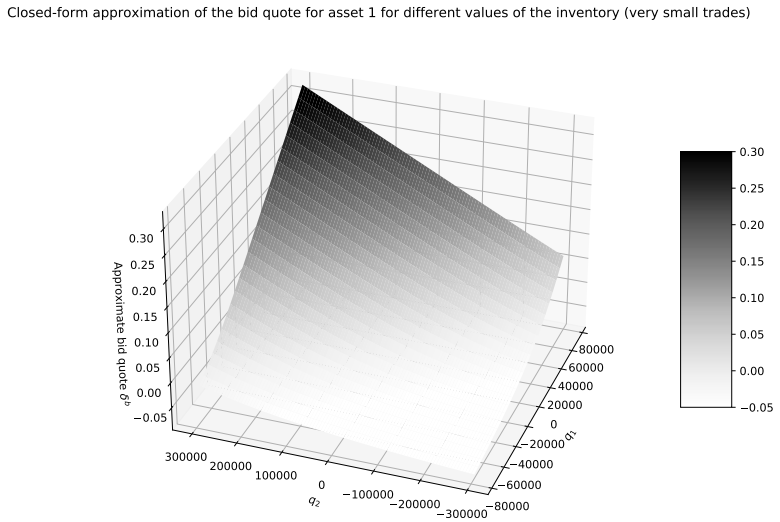


Figure 1.6: Closed-form approximation for the optimal bid quote at $t = 0$ for asset 1 as a function of the inventory (very small trades).

request and $\hat{\delta}^{i,b}$ is the closed-form approximation of the optimal bid quote for asset i as a function of inventory and size of request.

We clearly see that our closed-form approximations are close to the true optimal quotes,

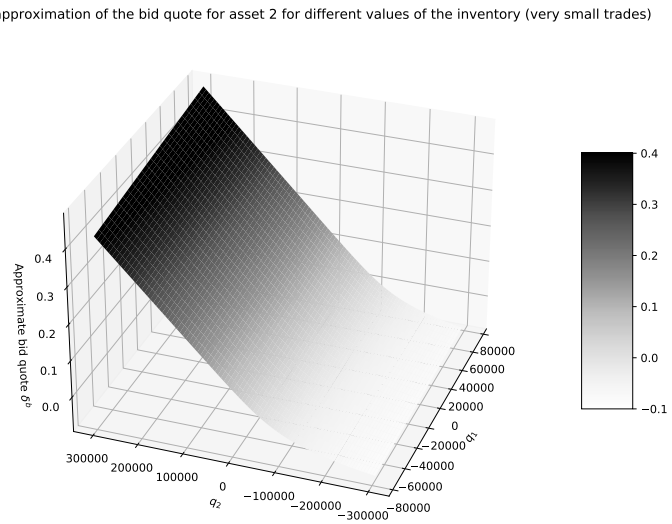


Figure 1.7: Closed-form approximation for the optimal bid quote at $t = 0$ for asset 2 as a function of the inventory (very small trades).

except for large values of the inventory in asset 2.

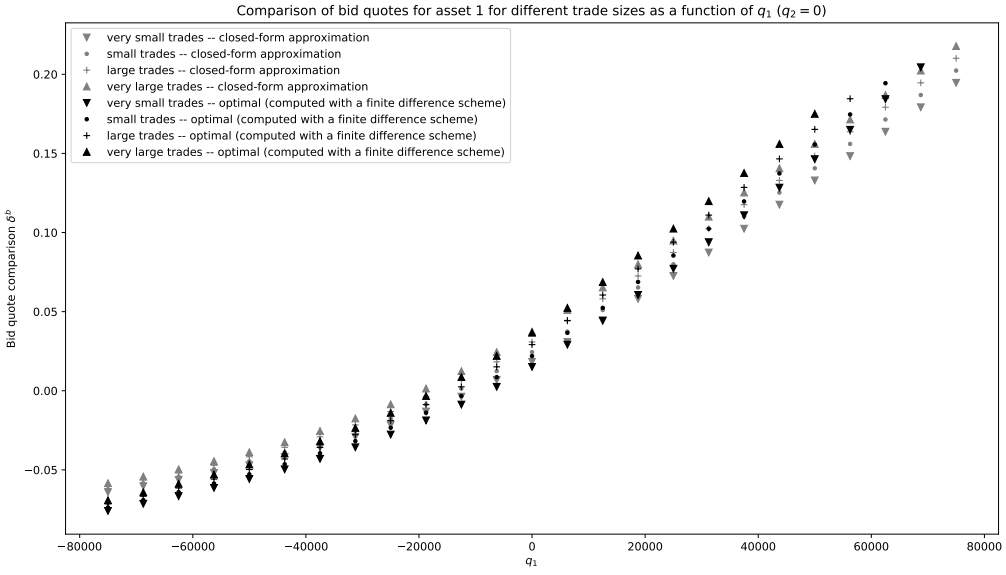


Figure 1.8: Comparison between optimal bid quote for asset 1 and its closed-form approximation for different trade sizes as a function of q^1 ($q^2 = 0$).

In order to confirm the quality of our closed-form approximations, we compare the per-

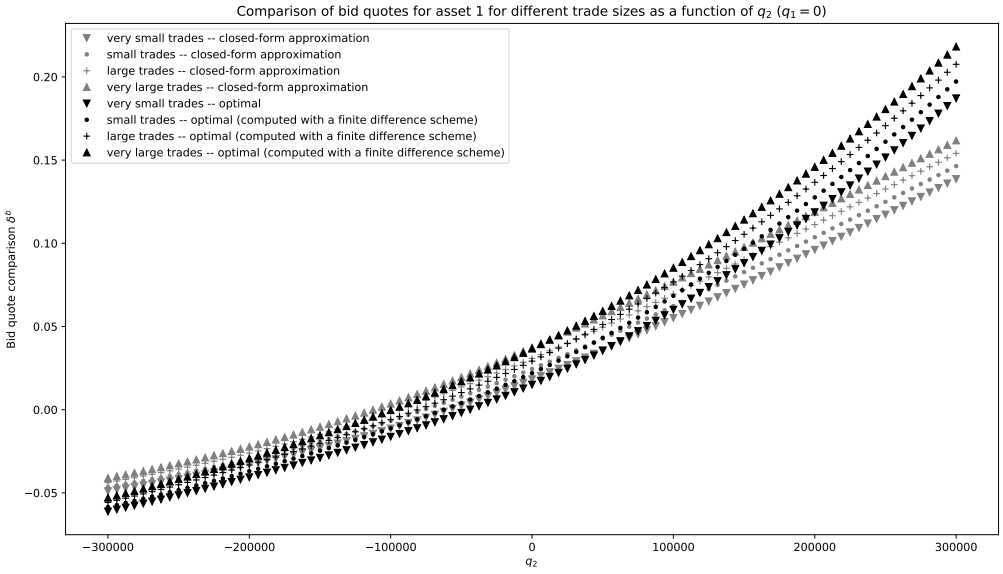


Figure 1.9: Comparison between optimal bid quote for asset 1 and its closed-form approximation for different trade sizes as a function of q^2 ($q^1 = 0$).

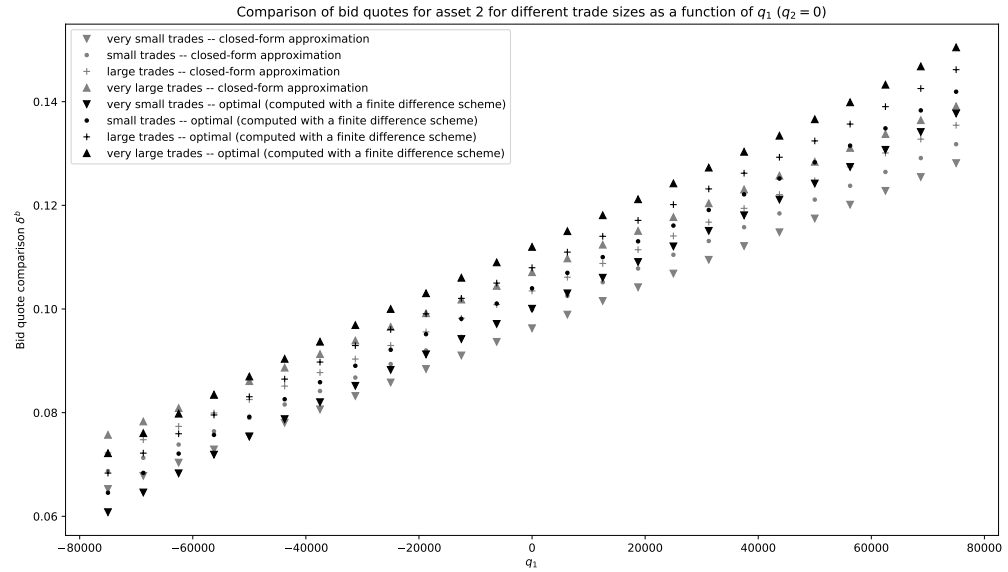


Figure 1.10: Comparison between optimal bid quote for asset 2 and its closed-form approximation for different trade sizes as a function of q^1 ($q^2 = 0$).

formance of a market maker, when quoting the true optimal quotes versus their closed-form approximations. The respective distributions of PnL after 4000 Monte-Carlo simu-

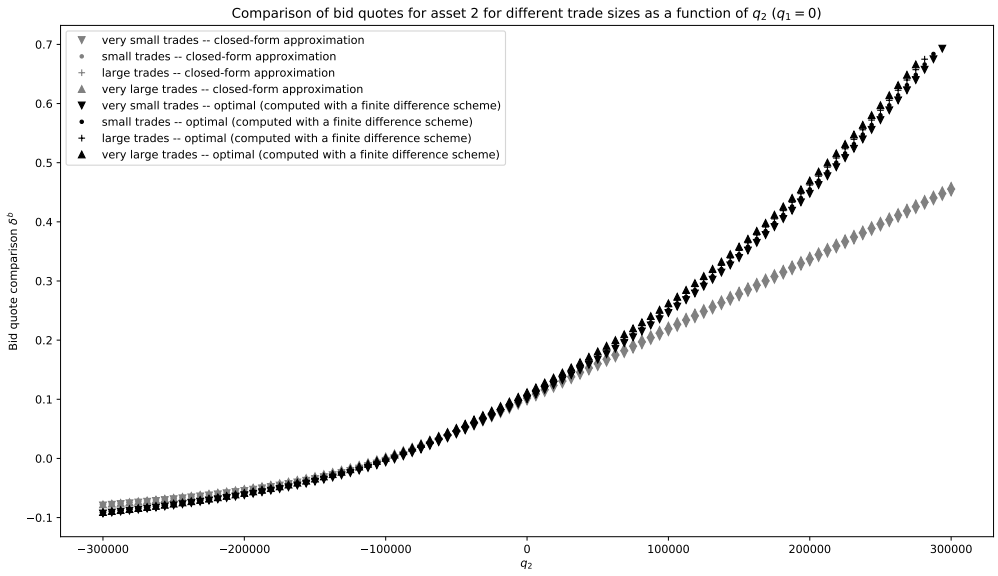


Figure 1.11: Comparison between optimal bid quote for asset 2 and its closed-form approximation for different trade sizes as a function of q^2 ($q^1 = 0$).

lations are plotted in Figures 1.12 and 1.13.

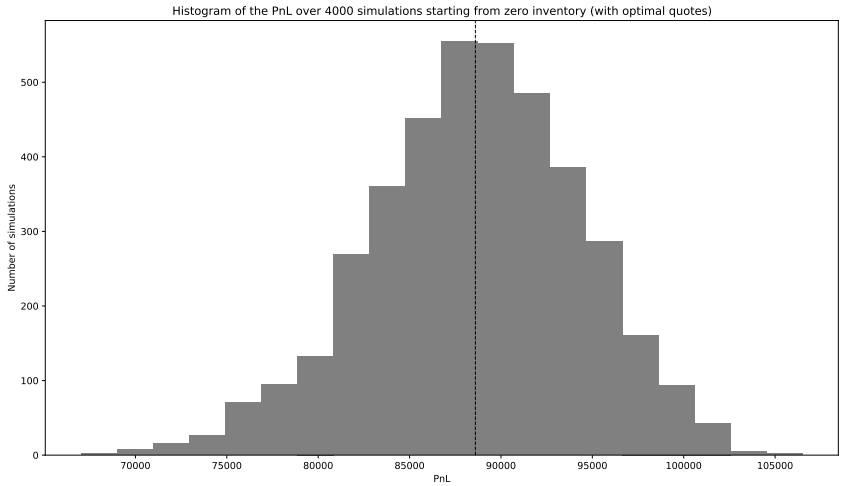


Figure 1.12: Distribution of the PnL of a market maker using the optimal quotes.

When using the optimal quotes, the market maker gets an average PnL of 88600€ with a standard deviation of 86900€. When using the closed-form approximation, the performance is very similar, as she gets an average PnL of 89000€ with a standard deviation of 87500€.

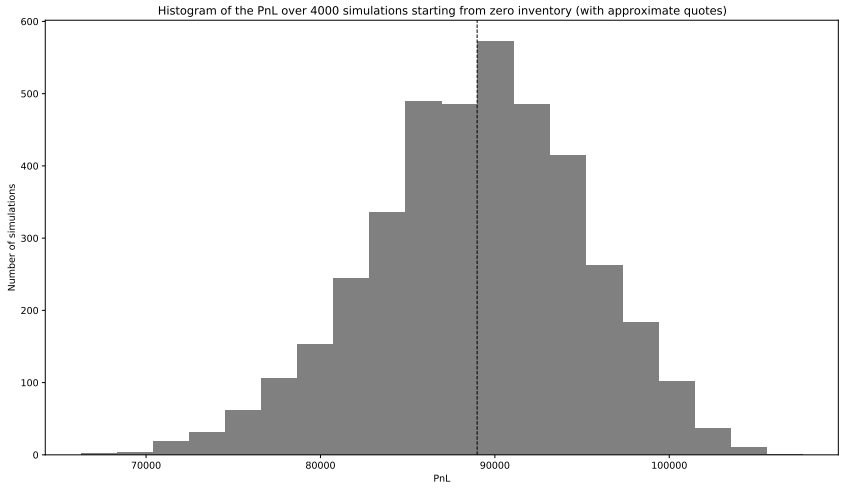


Figure 1.13: Distribution of the PnL of a market maker using the closed-form approximations.

These results are really satisfying in terms of performance. We see that, although the closed-form approximation of optimal quotes may be inaccurate for large values of the inventory, such large inventory are seldom reached and therefore the gap between quotes does not really impact the distribution of the PnL.

We believe that what we observe here in this two-asset example is general. In particular, the results in higher dimensions should be just as good.

Conclusion

We proposed closed-form approximations for the value functions associated with many multi-asset extensions of the market making models available in the literature. These closed-form approximations have been obtained through the "approximation" of a Hamilton-Jacobi equation by another Hamilton-Jacobi equation that can be simplified into a Riccati equation and two linear ordinary differential equations, all solvable in closed-form. These closed-form approximations can be used for various purposes, in particular to design quoting strategies through a greedy approach. The resulting closed-form approximations of the optimal quotes generalize those obtained by Guéant, Lehalle, and Fernandez-Tapia in Guéant et al. (2013) to a general framework suitable for practical use.

1.7 Appendix: On the construction of the processes $N^{i,b}$ and $N^{i,a}$

Let us consider a new filtered probability space $\big(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathbb{R}_+}, \tilde{\mathbb{P}}\big)$. For the sake of simplicity, assume that there is only one asset with size of transactions denoted by z and risk limit Q (the generalization is straightforward). Let us assume that the reference price of that asset is driven by a Brownian motion W as in Section 1.2.1. Let us introduce $\bar{N}^b$ and $\bar{N}^a$ two independent Poisson processes of intensity 1, independent of W. Let N^b and N^a be two processes, starting at 0, solutions of the coupled stochastic differential equation:

$$dN_t^b = \mathbb{1}_{\left\{zN_{t-}^b - zN_{t-}^a + z \le Q\right\}} d\bar{N}_t^b,$$
$$dN_t^a = \mathbb{1}_{\left\{zN_{t-}^b - zN_{t-}^a - z \ge -Q\right\}} d\bar{N}_t^a.$$

Then, under $\tilde{\mathbb{P}}$, N^b and N^a are two point processes with respective intensities

$$\lambda_t^b = \mathbb{1}_{\{q_{t-}+z\le Q\}} \quad \text{and} \quad \lambda_t^a = \mathbb{1}_{\{q_{t-}-z\ge -Q\}},$$

where $q_t = zN_t^b - zN_t^a$. For each $\delta \in \mathcal{A}$, we introduce the probability measure $\tilde{\mathbb{P}}^\delta$ given by the Radon-Nikodym derivative

$$\frac{d\tilde{\mathbb{P}}^\delta}{d\tilde{\mathbb{P}}}\Big|_{\mathcal{F}_t} = L_t^\delta, \tag{1.31}$$

where $\big(L_t^\delta\big)_{t\in\mathbb{R}^+}$ is the unique solution of the stochastic differential equation

$$dL_t^\delta = L_{t-}^\delta \left(\int_{\mathbb{R}_+^*} \big(\Lambda^b(\delta_t^b) - 1\big)\, d\tilde{N}_t^b + \int_{\mathbb{R}_+^*} \big(\Lambda^a(\delta_t^a) - 1\big)\, d\tilde{N}_t^a \right),$$

with $L_0^\delta = 1$, where $\tilde{N}^b$ and $\tilde{N}^a$ are the compensated processes associated with N^b and N^a, respectively.

We then know from Brémaud and Jacod (1977) that under $\tilde{\mathbb{P}}^\delta$, the jump processes N^b and N^a have respective intensities

$$\lambda_t^{\delta,b} = \Lambda^b(\delta_t^b)\mathbb{1}_{\{q_{t-}+z\le Q\}} \quad \text{and} \quad \lambda_t^{\delta,a} = \Lambda^a(\delta_t^a)\mathbb{1}_{\{q_{t-}-z\ge -Q\}}$$

as in Section 1.2.1. Since W is still a Brownian motion under $\tilde{\mathbb{P}}^\delta$, our optimal control problem can be seen as the choice of an optimal probability measure $\tilde{\mathbb{P}}^\delta$.

Chapter 2

Intraday option price dynamics and spot volatility

2.1 Introduction

2.1.1 Classical stochastic volatility literature

Stochastic volatility has been a staple of mathematical finance and econometrics for decades. From a time-series perspective, stochastic volatility can be present in the data in the form of heteroskedasticity. The celebrated ARCH model by Engle (1982) can be interpreted as a discrete-time stochastic volatility model for assets log returns. On this front, we also highlight an empirical research by Andersen et al. (2001) that compares daily log returns before and after normalising by volatility. More especifically, consider the model[1]

$$d\log(S_t) = \mu_t dt + \sigma_t dW_t, \tag{2.1}$$

where $(S_t)_{t\geqslant 0}$, $(\mu_t)_{t\geqslant 0}$, $(\sigma_t)_{t\geqslant 0}$ and $(W_t)_{t\geqslant 0}$ denote the asset price, drift, volatility and Brownian motion processes, respectively, with t measured in days. Motivated by (2.1), they estimate, among other quantities, the empirical distributions of daily log returns $\log(S_{t+1}/S_t)$ and daily standardised log returns $\log(S_{t+1}/S_t)/\sigma_{[t,t+1]}$ for 30 Dow Jones stocks, where $\sigma_{[t,t+1]}$ denotes the integrated volatility over the interval $[t, t+1]$, i.e.

$$\sigma^2_{[t,t+1]} = \int_t^{t+1} \sigma_s^2 ds.$$

[1]For brevity, we adapted the model for log prices so our equations differ slightly from the original ones in Andersen et al. (2001).

The integrated volatility $\sigma_{[t,t+1]}$ is estimated with realised volatility on 5-minute samples of last traded prices controlled for microstructure effects via a MA(1) filter. They have reported that although the empirical distribution of daily log returns exhibit fatter tails than normal distribution – in line with the literature –, the empirical distribution of daily standardised log returns are remarkably close to the normal distribution, with median kurtosis reducing from 5.416 (non-standardised log returns) to 3.129 (standardised log returns).

Stochastic volatility models are also ubiquitous in the option pricing literature. An early, yet still influential model is the Heston (1993) model. Risk-neutral pricing implies that European-type option prices are an expectation of the risk-neutral distribution of the underlying price at expiry. Stochastic volatility models are able to replicate the stylised skewness and kurtosis of such price distributions. In Heston (1993)'s own words:

> Conceptually, one can characterize the option models in terms of the first four moments of the spot return (under the risk-neutral probabilities). The Black and Scholes (1973) model shows that the mean spot return does not affect option prices at all, while variance has a substantial effect. Therefore, the pricing analysis of this article controls for the variance when comparing option models with different skewness and kurtosis. [...] Correlation between volatility and the spot price is necessary to generate skewness. Skewness in the distribution of spot returns affects the pricing of in-the-money options relative to out-of-the money options. Without this correlation, stochastic volatility only changes the kurtosis. Kurtosis affects the pricing of near-the-money versus far-from-the-money options.

The author also remarks on the impact of the "volatility of volatility" parameter to the kurtosis of the distribution.

Of course, when calibrating the model to an entire implied volatility surface, we need to fit multiple marginal distributions from a single set of parameters. In particular, a challenge for stochastic volatility models is to fit the so-called at-the-money volatility skew. The introduction of jumps was an early attempt to model the reproduce statistical properties of the at-the-money volatility skew. More recently, rough volatility models were also found to be able to replicate such properties. We have mentioned this fact to illustrate some limitations of early stochastic volatility models, and we refer the reader to a more in-depth discussion in Gatheral et al. (2018).

Despite its limitations, we will employ the Heston model in the analyses of this chapter, due to its simplicity and ease of implementation. For further discussion on stochastic volatility models, including a modern perspective of the Heston model, we refer the

reader to the books Gatheral (2011) and Bergomi (2015).

2.1.2 Small time option price dynamics

Another aspect of an option pricing model – apart from describing a snapshot of the implied volatility surface via marginal probability distributions – is its ability to accurately describe how such implied volatility surface evolves over time, i.e. its dynamics. These two facets of an option pricing model are well illustrated when comparing local and stochastic volatility models. Local volatility models can be calibrated to tightly match observed implied volatility surfaces, which makes it useful for option pricing consistently across liquid and illiquid markets. However, the only risk factor in local volatility models is the underlying price, which limits its applications to hedging and risk management. Stochastic volatility models, on the other hand, are more restrictive on the shape of their implied volatility surfaces it can generate but a sufficiently rich structure of stochastic factors can describe the implied volatility dynamics more realistically. For a detailed discussion on this topic, we refer the reader to Hagan et al. (2002) and Bergomi (2015).

In our regime of interest, small time option dynamics has been extensively studied in the academic literature. From leading-order asymptotics of at-the-money call options Muhle-Karbe and Nutz (2011), to small time functional central limit theorem of functions of semimartingales Gerhold et al. (2015), to higher-order small time asymptotics for rough volatility models Friz et al. (2021). The topic is usually studied in the point of view of options near expiry, which presents a wide range of applications, such as: to identify the leading order on which option prices converge to their payoff, (ii) to approximate the implied volatility surface – thus providing in particular a useful parametrisation – (iii) to approximate option prices for calibration purposes.

In this chapter, our focus is on the modeling of intraday dynamics of option prices, possibly far away from expiry. In particular, given the empirical nature of this chapter, we are interested in the observed $\mathbb{P}$ dynamics of option prices, in which case the martingale property of option price processes cannot be assumed. We describe precisely this point of view in the following Proposition 7.

We consider a setting in which the market can be incomplete. From the fundamental theorems of asset pricing the market is arbitrage-free if and only if there exists an equivalent martingale measure (ELMM) and is only unique for complete markets. Given that we are in the incomplete market setting, we take the approach in which we are given an ELMM $\mathbb{Q}$ and derive the dynamics of an arbitrage-free price process for an option $(C_t)_{t\in[0,T]}$ with payoff function f. For what follows, we also assume interest rates equal to zero – the non-zero interest rate case can be achieved by using the T-forward measure

as the $\mathbb{Q}$ measure.

Proposition 7. Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}, \mathbb{P})$ be a filtered probability space that satisfies the usual conditions, let $\mathbb{Q}$ be probability measure equivalent to $\mathbb{P}$ and let the state process $(X_t)_{t\in[0,T]}$ be an $\mathbb{R}^d$-valued continuous process of the form

$$\begin{aligned} dX_t &= \mu_t dt + \sigma_t dW_t, && (2.2)\\ &= \mu^{\mathbb{Q}}(t, X_t)dt + \sigma^{\mathbb{Q}}(t, X_t)dW_t^{\mathbb{Q}}, \quad \forall t \in [0,T], && (2.3)\\ X_0 &= x_0 \in \mathbb{R}^d \text{ a.s.},\\ dW_t^{\mathbb{Q}} &= \theta_t dt + dW_t, \quad \forall t \in [0,T], && (2.4) \end{aligned}$$

where $(\mu_t)_{t\in[0,T]}$, $(\sigma_t)_{t\in[0,T]}$ and $(\theta_t)_{t\in[0,T]}$ are adapted $\mathbb{R}^d$, $\mathbb{R}^{d\times n}$ and $\mathbb{R}^d$-valued processes, respectively, that are continuous and square-integrable, i.e.

$$\mathbb{E}\left[\int_0^T \left(\|\mu_t\|^2 + \|\sigma_t\|_F^2 + \|\theta_t\|^2\right) dt\right] < \infty, \tag{2.5}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $\mu^{\mathbb{Q}} : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma^{\mathbb{Q}} : [0,T] \times \mathbb{R}^d \to \mathbb{R}^{d\times n}$ are continuous functions, and $(W_t)_{t\in[0,T]}$, and $(W_t^{\mathbb{Q}})_{t\in[0,T]}$ are adapted n-dimensional vector independent Brownian motions under $\mathbb{P}$ and $\mathbb{Q}$, respectively.

Given an $\mathcal{F}_T$-measurable and bounded function $f : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}^{\mathbb{Q}}[|f(X_T)|] < \infty$, define the option price process $(C_t)_{t\in[0,T]}$ with

$$C_t \coloneqq \mathbb{E}^{\mathbb{Q}}\left[f(X_T) \mid \mathcal{F}_t\right], \quad \forall t \in [0,T].$$

Then, there exists a function $\varphi : [0,T] \times \mathbb{R}^d \to \mathbb{R}$ such that

$$C_t = \varphi(t, X_t), \quad \forall t \in [0,T]. \tag{2.6}$$

Furthermore, if φ is of class $\mathcal{C}^{1,2}([0,T) \times D)$, where $D \subseteq \mathbb{R}^d$ is a set that contains an open neighbourhood of x_0 and

$$\mathbb{P}\left(X_t \in D,\ \forall t \in [0,T)\right) = 1,$$

then the processes $(X_t)_{t\in[0,T]}$ and $(C_t)_{t\in[0,T]}$ admit the small time asymptotics

$$\begin{aligned} \frac{1}{\sqrt{t}} \left\| X_t^i - \tilde{X}_t^i \right\|_{L^2(\mathbb{P})} &\xrightarrow{t\to 0} 0, \quad \forall i \in \{1, \dots, d\} && (2.7)\\ \frac{1}{\sqrt{t}} \left\| C_t - \tilde{C}_t \right\|_{L^2(\mathbb{P})} &\xrightarrow{t\to 0} 0, && (2.8) \end{aligned}$$

where ∇_x denotes the gradient operator with respect to x and

$$\tilde{X}_t = X_0 + \sigma_0 W_t, \quad \tilde{C}_t = C_0 + \nabla_x \varphi(0, X_0) \cdot \left(\tilde{X}_t - \tilde{X}_0 \right), \quad \forall t \in [0, T].$$

Proof. See Appendix 2.7.1. □

To illustrate Proposition 7, we use the Heston model as an example. Under $\mathbb{Q}$, the (discounted) state process is the pair $(X_t = (\log S_t, V_t))_{t \in [0,T]}$ with

$$\begin{aligned} d(\log S_t) &= -\frac{1}{2} V_t dt + \sqrt{V_t} dW_t, \\ dV_t &= \kappa (\theta - V_t)\, dt + \nu \sqrt{V_t} dZ_t, \\ d[W, Z]_t &= \rho dt, \end{aligned} \tag{2.9}$$

where $(W_t, Z_t)_{t \in [0,T]}$ is a pair of correlated $\mathbb{Q}$-Brownian motions. Let L be a lower-diagonal matrix of the Cholesky decomposition of the correlation matrix

$$\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} = L L^\top,$$

then we can identify the coefficients in (2.3) with

$$\mu^{\mathbb{Q}}(t, S_t, V_t) = \begin{bmatrix} -\frac{1}{2} V_t \\ \kappa (\theta - V_t) \end{bmatrix}, \quad \sigma^{\mathbb{Q}}(t, S_t, V_t) = \begin{bmatrix} \sqrt{V_t} & 0 \\ 0 & \nu \sqrt{V_t} \end{bmatrix} L.$$

As for the process $(\theta)_{t \in [0,T]}$, even though we emphasize the importance of the $\mathbb{P}$ dynamics – as it is the measure in which we observe the underlying and option prices move –, we only need to know the precise parametrisation under $\mathbb{Q}$ and assume that the conditions on $(\theta)_{t \in [0,T]}$ are relaxed enough so that in encompasses a realistic choice of $(\theta_t)_{t \in [0,T]}$ for the model under $\mathbb{P}$. We thus illustrate this example for the particular case of $\theta \equiv 0$.

With $\theta \equiv 0$, the square-integrability conditions (2.5) are met when $V_t^2 < \infty$ for all $t \in [0, T]$. We remark that $(V_t)_{t \in [0,T]}$ is the CIR process. Under the Feller condition $2\kappa\theta \geqslant \nu^2$, it is well known that all moments of the CIR process are finite due to its link to the noncentral χ^2 distribution.

The boundedness of the function f can be achieved with the put option payoff function. Besides, a call option would also enjoy the small time asymptotics via put-call parity and the small time asymptotics of the underlying price in (2.7) and put option price in (2.8).

Finally, the assumption that put option formula φ is of class $\mathcal{C}^{1,2}([0, T] \times D)$ under the Heston model deserves attention. The classical conditions to ensure that the PDE that

$\varphi \in \mathcal{C}^{1,2}([0,T] \times D)$, such as that the PDE is uniformly elliptic, does not hold. However, Heath and Schweizer (2000) provide a new set of sufficient conditions[2] that can be applied to popular option pricing models, including the Heston model with the Feller condition.

The small time asymptotics can thus be applied to the Heston state process and vanilla option price process, namely

$$\tilde{S}_t = S_0 + \sqrt{V}_0 S_0 W_t, \quad \tilde{V}_t = \nu\sqrt{V}_0 Z_t,$$

$$\tilde{C}_t = C_0 + \partial_s\varphi(0, S_0, V_0)\left(\tilde{S}_t - \tilde{S}_0\right) + \partial_v\varphi(0, S_0, V_0)\left(\tilde{V}_t - \tilde{V}_0\right), \quad \forall t \in [0,T].$$

Proposition 7 adapts existing results into the context we are interested in. For instance, the small time asymptotics of the state process (2.7) is very similar to Proposition 2.1 in Muhle-Karbe and Nutz (2011) and in particular its Heston model example in Corollary 2.1. Muhle-Karbe and Nutz (2011), however, require the martingale assumption, so the small time asymptotics is only applicable for the $\mathbb{Q}$ dynamics of the underlying price and option price processes, whereas Proposition 7 encompasses the full state process (including the underlying volatility) and the option price process under $\mathbb{P}$.

Another related result is Theorem 3 in Gerhold et al. (2015), which is a functional central limit theorem for the small time process of the form $(f(X_t))_{t\in[0,T]}$, where $(X_t)_{t\in[0,T]}$ is an Itô semimartingale with mild regularity assumptions and f is a twice differentiable multivariate function. The functional central limit theorem is a weak convergence result. In contrast, our Proposition 7 requires strong assumptions, which enables a stronger mode of convergence. The $L^2(\mathbb{P})$ convergence in (2.7) and (2.8) show that Brownian motion in the asymptotic process is the same as in the original state process – a minor detail, which nonetheless helps in building intuition for the convergence result.

The goal of Proposition 7 is to state with precision how the option price moves at small time scales under the $\mathbb{P}$ measure in tandem with its driving factors in the same regime. We discuss this result further in Section 2.3.

2.1.3 Empirical studies on options and volatility

Volatility is usually studied as one of the following incarnations: spot volatility, historical volatility and implied volatility. Under the lens of a stochastic volatility model, spot volatility is typically part of the state process and thus a direct driver of option prices. Other more complex stochastic volatility models might model the forward volatility curves, such as the N-factor Bergomi model and rough volatility models. The spot

[2]See also Theorem 4.7 in Ruf (2013) which requires slightly weaker conditions.

volatility, however, is a latent variable, and as such it is usually estimated via historical or even implied volatility, taking advantage of the theoretical relationship among these three concepts – see Lee (2005) for a review.

Estimating historical volatility using high-frequency data is subject to another difficulty: microstructure noise. Several methods have been proposed to address microstructure noise, see e.g. Robert and Rosenbaum (2011) and the book Aït-Sahalia and Jacod (2014). Furthermore, the approximation of spot volatility by integrated volatility is another source of error. Some methodologies to estimate spot volatility, such as in Kristensen (2010), provide such error estimates. A recent study using transactional tick data from S&P 500 E-mini futures have reported a maximum granularity of approximately 30 minutes for spot volatility estimations to stay within a reasonable noise-to-signal ratio range – see Section 5 in Bennedsen et al. (2021).

Estimating spot volatility via implied volatility seems more difficult, but there are empirical studies that take this approach. For instance, Livieri et al. (2018) have used daily at-the-money implied volatilities near expiry as proxies for spot volatility. They were able to find evidence of rough volatility, although they have also reported that the estimate is biased due the applied methodology. For intraday spot volatility estimates, however, we are not aware of any research pursuing this approach.

On the topic of intraday option price dynamics and spot volatility, the only empirical study that we are aware of is Abergel and Zaatour (2012). They have used tick data on options and futures on Eurostoxx 50, Dax 30 and Kospi 200 with the goal of finding whether quadratic variation is a driver of option prices. They have used the Epps effect on correlations between option and futures price to find a reasonable sampling frequency of 5-minutes. Quadratic variation was then estimated via realised volatility every 5 minutes on a 25-minute sliding window of log-prices sampled every second. Alternatively to realised variance, they have also used the method in Garman and Klass (1980) to estimate quadratic variation which they claim to be less sensitive to microstructure noise. Then, they have compared two linear regression analyses: a single linear regression of option log returns against the underlying future log returns and a multiple linear regression with the addition of the realised variance changes. They have reported that, between the two regressions, both the estimated regression coefficient on the underlying future component and the R^2 between the two regressions has remained unchanged. In conclusion, they have found no evidence that quadratic variation is a driver of option prices.

Given the aforementioned difficulties in estimating intraday spot volatility, it would not be unreasonable to believe that the multiple linear regression analysis failed to capture the effect of spot volatility changes on option log returns due to noise. Indeed, we can identify several sources of noise: the approximation of spot volatility by realised volatility,

microstructure noise and the overlapping sliding windows of consecutive realised volatility estimates. As for microstructure noise, it would be more suitable to use an estimator designed to filter out or be robust against microstructure noise. In Garman and Klass (1980), the estimator seems to be developed as a method to make full use of open-high-low-close data, rather than addressing microstructure noise specifically.

Some related studies are also worth mentioning. In Cont et al. (2002), the implied volatility surface dynamics is empirically studied endogenously via spectral decomposition. The high-frequency dynamics of the implied volatility surface is studied mathematically in Baldacci (2020) linking Hawkes processes to factors such as level, slope and curvature.

2.1.4 Main contributions

Our main goal is to improve the understanding of the role of stochastic volatility in intraday option price dynamics. More specifically, we empirically assess its first-order status as is suggested by the small-time asymptotic result in Proposition 7. Given that option prices are a nonlinear function with respect to its underlying price, one could also expect that second-order factors such as the underlying squared log-returns could have a stronger effect than spot volatility even at small time scales.

Having described the difficulties in estimating spot volatility at fine granularity using time series data, we propose an alternative method that does not rely on the underlying price time series but instead uses snapshots of option prices. Given a stochastic volatility model – here, we employ the Heston model –, the spot volatility can be recovered by the recalibration of the spot volatility variable while keeping the remaining model parameters fixed. Indeed, if we follow the Heston model by the letter, spot volatility V_0 is a latent variable which is part of the state process and all other parameters (κ, θ, ν and ρ) are constants.

To the best of our knowledge, the recalibration of an option pricing model to estimate spot volatility is novel. We highlight that model recalibration itself is a well-studied topic. In Buehler (2006), for example, arbitrage-free model recalibration is studied in the context of recalibrating not only the state variables but also model parameters that would be assumed to be constant by the original model – thus the act of recalibrating such parameters would imply the extension of the original model to a meta-model. We instead follow the original model for which we know precisely the governing dynamics of the model and, in the case of the Heston model, that it fulfills the conditions of Proposition 7. Our contribution is in having model recalibration as the means for estimating spot volatility.

The reliance of the method in the choice of a particular option pricing model comes,

of course, at the cost of model bias. This is a tradeoff that we are willing to take in order to obtain a much finer granularity on spot volatility estimates at the timescale of seconds, which does not seem to be possible via the time-series approach. We discuss this limitation on the time-series approach in more details in Section 2.3.

Equipped with our granular spot volatility estimation and motivated by the linear approximation obtained with the small time asymptotics, we perform a linear regression on option price changes with respect to its underlying price and volatility changes in Section 2.4. The regression coefficients are effectively estimates of the Heston first-order Greeks, with which we compare. Our results indicate that the regression coefficients are aligned with the Heston first-order Greeks and that the second-order effects from the underlying price changes squared are negligible. We further conduct an analysis to quantify the contribution of spot volatility changes that cannot be explained by the underlying price changes.

In the context of the thesis, in the previous chapter we have investigated multi-asset market making in the ergodic regime. In the next chapter, we propose options market making in a small horizon regime. The latter regime is in resonance with the small time asymptotics that we study in this chapter. Therefore, this chapter provides a theoretical and empirical motivation for the market making model proposed in the next chapter.

2.1.5 Dataset and source code

The empirical analysis is done on Euro Stoxx 50 options tick quotes data resampled at 1-second granularity. The dataset spans from 22 November 2021 to 16 December 20211, which starts on the business day following the 19 November 2021 option expiry date and ends the day before the 17 December option expiry date. A summary description of the dataset is shown in Table 2.1. The underlying forward price time series and interest rate term structure are obtained via put-call parity as explained in Section 2.2.

The source code for the data analysis is available online[3]. It is written in Python and uses several open-source libraries. We list them by their role in this paper.

- Heston implementation
  - Fyne (Vieira, 2020)
- Statistical
  - ARCH (Sheppard et al., 2020)

[3]See `https://github.com/dougmvieira/phd-thesis-data-analysis`.

	Active options				Mid-price changes			
	min	median	mean	max	min	median	mean	max
expiry								
2021-12-17	63	96	184.4	368	502,289	1,438,384	5,131,799.3	21,726,774
2022-01-21	71	98	155.9	302	473,453	1,846,570	5,235,255.0	19,131,413
2022-02-18	40	89	142.4	260	609,294	2,418,363	6,527,861.7	20,983,439
2022-03-18	42	78	171.9	366	1,373,316	4,413,055	10,641,834.7	30,627,963
2022-04-14	2	15	50.4	257	54,660	946,325	5,735,217.8	18,696,776
2022-05-20	0	3	29.3	91	0	149,987	4,397,388.5	13,885,170
2022-06-17	28	39	91.7	190	1,552,464	3,133,436	7,528,210.9	19,061,507
2022-09-16	8	25	73.2	168	361,657	1,310,148	6,873,333.4	21,875,047
2022-12-16	13	49	100.5	214	810,932	3,236,768	10,058,118.7	29,411,099
2023-03-17	0	2	60.6	162	0	130,175	5,939,196.1	20,565,618
2023-06-16	0	6	95.2	252	0	437,789	12,229,296.8	44,285,754
2023-09-15	0	1	33.0	92	0	100,761	5,528,575.2	18,882,990
2023-12-15	2	6	35.4	95	86,344	498,257	4,792,903.3	16,748,657
2024-06-21	0	0	24.2	66	0	0	2,246,601.5	8,704,599
2024-12-20	0	1	28.5	78	0	61,661	2,399,767.1	9,227,876
2025-12-19	0	1	29.1	79	0	39,180	1,895,997.4	7,212,144
2026-12-18	0	0	0.4	1	0	0	0.4	1
2027-12-17	0	0	0.1	1	0	0	0.1	2
2028-12-15	0	0	0.2	1	0	0	0.4	4
2029-12-21	0	0	1.6	5	0	0	2.7	11
2030-12-20	0	0	0.4	1	0	0	2.4	29
total	359	474	1,308.2	2,993	6,825,749	18,406,413	97,161,363.4	287,101,368

Table 2.1: Summary daily statistics of the options quotes dataset

  - Statsmodels (Seabold and Perktold, 2020)
- Data structures
  - xarray (Hoyer and Hamman, 2020)

2.1.6 Structure of the chapter

In Section 2.2, we present the bootstrap methodology to extract the underlying forward price time series and interest rate term structure from options data, and show the Heston model calibration results. In Section 2.3, we discuss the small time option asymptotic result and its implications on spot volatility estimation, and then proceed to estimate spot volatility from options data. In Section 2.4, we perform the linear regression analysis on options prices and volatilities, we compare the estimated Greeks with what is expected from the calibrated Heston model, and then quantify the effect of volatility on option price changes.

2.2 Bootstrap methodology

2.2.1 Overview

In this section, we aim at recovering the underlying forward prices and the interest rate term structure using only option quotes. For short expiries, data for the corresponding futures are readily available and could be used as discounted forward prices, however at long expiries option prices present tighter spreads than the corresponding futures. We also need to untangle the effects of dividend yields and interest rates to price options consistently. Having a methodology to extract the forward prices and interest rates using only options quotes is worthwhile because it provides a systematic approach in obtaining such quantities across all expiries that is accurate enough to feed into the option pricing model. Furthermore, even though options might not be traded as much as its corresponding future, their quotes are frequently updated as we can observe from Table 2.1.

The bootstrapping methodology we present relies only on put-call parity and on the structure of dividend yields and bonds, hence it does not require specific assumptions on the stochastic dynamics of the underlying price process. The overall methodology can be summarised in the following steps:

1. Construct synthetic forwards as a combinations of calls and puts

2. Extract a normalised underlying price process via PCA

3. Assuming an affine form for the dividends, perform a linear regression on the normalised underlying price process using the synthetic forwards and strikes of each expiry

4. The forwards and interest rate term structures are obtained from the regression coefficients.

Relying on the regression coefficients after assuming a particular form for the dividends could seem arbitrary at first glance, however, as we see later, these seem sensible steps. Ideally, one would only rely on no-arbitrage arguments. Unfortunately, using no-arbitrage arguments alone provide no-arbitrage bounds for such quantities which are too wide for practical use and that are sensitive to too few options – see Appendix 2.6 for more details. Thus, the extra assumptions allows us to obtain sensible quantities that are robust to fauly price changes on individual options.

The proposed bootstrapping technique is presented by introducing and showing the results of each step. In Section 2.2.2, we recover the normalised underlying price process via PCA. In Section 2.2.3, we perform the linear regression that extracts the forward price time series and the interest rate term structure. Finally, we calibrate the Heston model using the bootstrapped forwards and interest rates in Section 2.2.4.

2.2.2 Normalised underlying price process via PCA

We start by formally stating the put-call parity result that is used for the proposed methodology. Let $C^{\text{bid}}_{t,T,K}$, $C^{\text{ask}}_{t,T,K}$, $P^{\text{bid}}_{t,T,K}$ and $P^{\text{ask}}_{t,T,K}$ denote the bid and ask prices of European call and put options at time t with expiry T and strike K. Let S_t denote the price of its underlying at time t. Let $B_{t,T}$ denote the price of a T-maturity zero-coupon bond at time t with $B_{T,T} = 1$. Let $F_{t,T}$ be the forward price[4] at time t on the same underlying with expiry T. We assume that: (i) there are no arbitrage opportunities and, (ii) for a given strike K and expiry T, that both the call and put options with strike K and expiry T are tradeable, as well as the corresponding forward contract with expiry T. Then, $B_{t,T}$ and $F_{t,T}$ must follow the put-call parity inequalities

$$C^{\text{bid}}_{t,T,K} - P^{\text{ask}}_{t,T,K} \leqslant B_{t,T}F_{t,T} - B_{t,T}K \leqslant C^{\text{ask}}_{t,T,K} - P^{\text{bid}}_{t,T,K}. \tag{2.10}$$

Note that we have stated the inequalities with the forward price instead of the underlying

[4]For clarity, we denote by 'forward price' the strike of a forward contract such that its present value is zero.

price so that the result holds whether or not the underlying pays dividends. In particular, if the underlying is dividend-free, then $S_t = B_{t,T}F_{t,T}$. For completeness, we provide a formal no-arbitrage argument for the above inequalities in Appendix 2.6.1.

The main idea is to use the synthetic forward prices $C_{t,T,K} - P_{t,T,K}$ to find a common driver that is equal to the underlying price up to an affine transformation. On the other hand, the two variables we need to estimate $B_{t,T}$ and $F_{t,T}$ are time and strike dependent, so we need to assume some structure to obtain sensible estimates of the underlying price process at each time t. Considering we are in the intraday regime, we can afford to freeze some variables in time – i.e. assume they are approximately constant in the day. In this respect, we assume that the bond prices are constant in time, i.e. $B_{t,T} = B_T$, and assume an affine dividend structure of the form

$$B_{t,T}F_{t,T} = D_T^0 + D_T^1 S_t, \tag{2.11}$$

where $D_T^0, D_T^1 \in \mathbb{R}$ are constants that for each expiry T. Intuitively, we assume that the dividends issued between t and T have an unconditional component D_T^0 and a component D_T^1 that is proportional to the company's equity. Then, (2.10) becomes

$$C_{t,T,K}^{\text{bid}} - P_{t,T,K}^{\text{ask}} \leqslant D_T^0 + D_T^1 S_t - B_T K \leqslant C_{t,T,K}^{\text{ask}} - P_{t,T,K}^{\text{bid}}.$$

We heuristically approximate the middle term by the mid-point of the bounds i.e. the mid-price of the synthetic futures, so as to obtain

$$D_T^0 + D_T^1 S_t - B_T K \approx F_{t,T,K}^{\text{synthetic}} := \frac{\left(C_{t,T,K}^{\text{bid}} - P_{t,T,K}^{\text{ask}}\right) + \left(C_{t,T,K}^{\text{ask}} - P_{t,T,K}^{\text{bid}}\right)}{2}. \tag{2.12}$$

If we take the difference in time at every interval of size h of the left-hand side, then $B_T K$ and D_T^0 are canceled out and we obtain $D_T^1 S_{t+h} - D_T^1 S_t$. We can further cancel out D_T^1 by normalising by the standard deviation, so we obtain

$$\frac{S_{t+h} - S_t}{\sqrt{\text{Var}\left(S_{t+h} - S_t\right)}} \approx \frac{F_{t+h,T,K}^{\text{synthetic}} - F_{t,T,K}^{\text{synthetic}}}{\sqrt{\text{Var}\left(F_{t+h,T,K}^{\text{synthetic}} - F_{t,T,K}^{\text{synthetic}}\right)}} \tag{2.13}$$

from which we can recover the normalised underlying price process by integrating in time. Later, in Section 2.2.3, we revert the normalisation to obtain the forward price processes.

Note that the left-hand side in (2.13) does not depend on neither strike nor expiry. Therefore, each synthetic future can be used to estimate the normalised underlying price process. It remains to aggregate the estimates in a sensible way. We do so by performing PCA on all normalised synthetic forward prices and collecting the first component. In

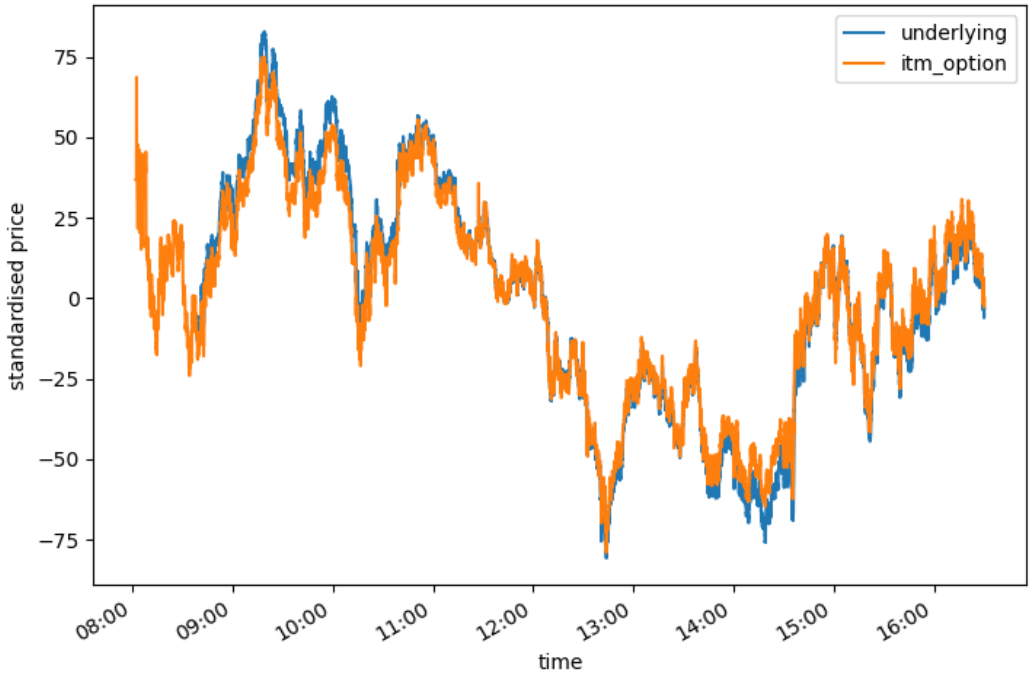


Figure 2.1: Normalised underlying price time series as obtained via the bootstrap methodology compared with the normalised mid-price of an in-the-money option with strike 4000 and shortest expiry – for reference, Euro Stoxx 50 closed at 4,108 on the same day. The normalisation of the option price is done via removing the average mid-price and dividing by standard deviation of the mid-price changes.

this manner, we also hope to average out microstructure effects from our underlying price process estimate, and thus should belong to the second and higher principal components.

The methodology for extracting the underlying price process up to an affine transformation is then summarised as

1. Collect all the synthetic forward contract bid and ask quotes,
2. For each synthetic forward, compute the mid-price changes and normalise by the standard deviation,
3. Perform PCA on the normalised mid-price changes,
4. Integrate the first principal component in time to obtain the normalised underlying price process.

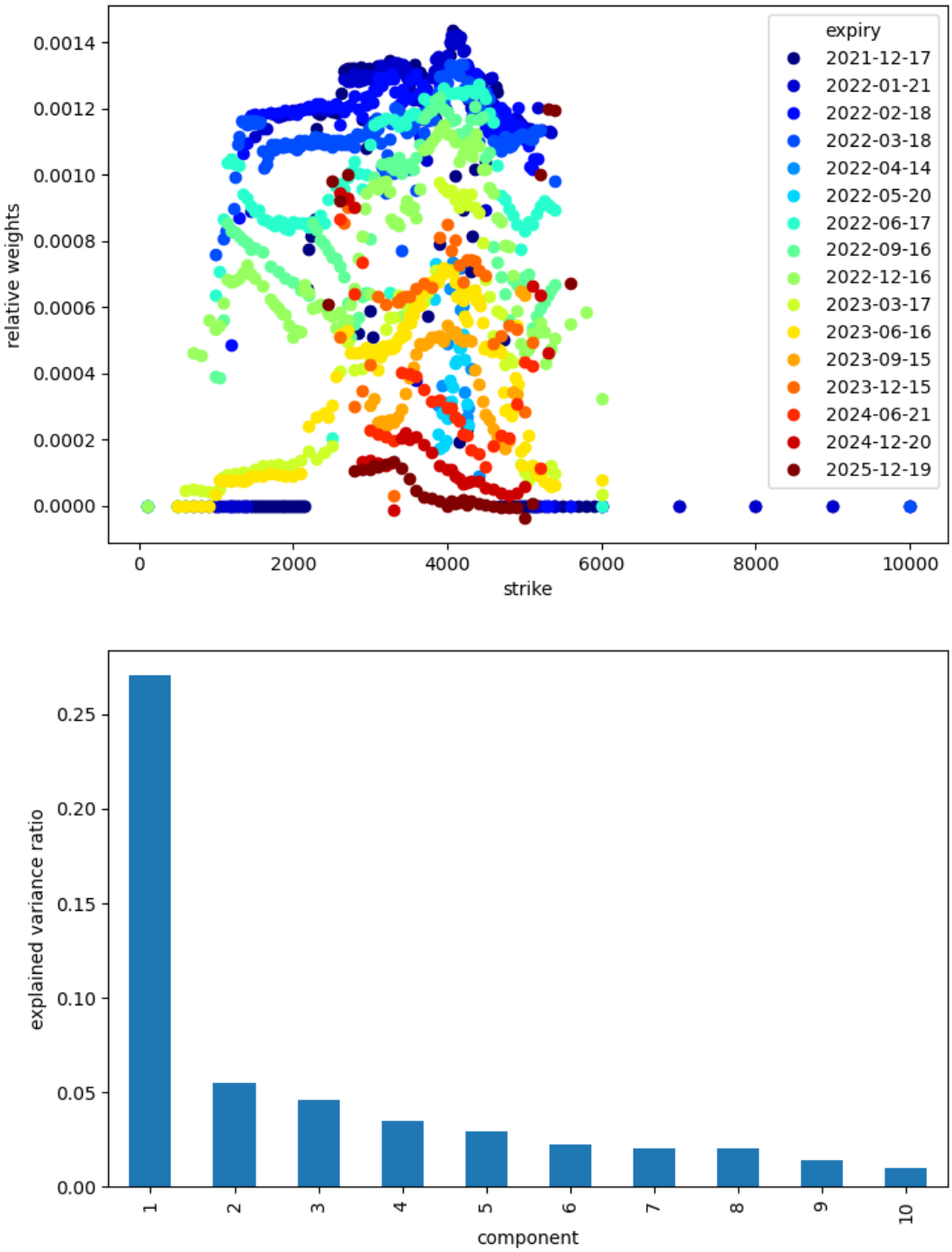


Figure 2.2: The upper plot shows the weights of the first principal component across the strikes and expiries of the synthetic forwards. The lower plot shows the relative variance explained by the first ten principal components.

By applying this methodology on the Euro Stoxx 50 index options, we obtain the time series depicted in Figure 2.1. The corresponding PCA results are depicted in Figure 2.2. The recovered normalised underlying price process passes the visual sanity check by being similar to the in-the-money option. As for the PCA results, we can observe that the weights are higher for options close to expiry and near the money – this is likely due to their tighter spreads.

2.2.3 Forwards and bonds prices regression

Revisiting the dividend form (2.11) and the approximation in (2.12), we have just estimated the normalised version of S_t and so we are ready to find the forward prices $F_{t,T}$. Denote by $\tilde{S}_t$ the normalised underlying price, then we perform, for each expiry T, a linear regression based on the following equation

$$F^{\text{synthetic}}_{t,T,K} = \tilde{D}^0_T + \tilde{D}^1_T \tilde{S}_t - B_T K. \tag{2.14}$$

In the context of deriving no-arbitrage bounds for bonds and interest rates, Figure 2.32 depicts the snapshot of synthetic forwards quotes versus the strike for various expiries.

As for the choice of the loss function of the linear regression, ideally we would like to find estimates for $\tilde{D}^0_T$, $\tilde{D}^1_T$ and B_T such that the fitted time series would lie inside within the bid-ask spread of the synthetic futures. To find such loss function, we consider a family of power loss functions and properly normalise the residuals by the half-spread. Proposition 8 guides our choice for the exponent of the power loss function.

Proposition 8. Let $\{\underline{y}_i, \bar{y}_i, x_i\}_{i=1}^N$ be a dataset with $\underline{y}_i, \bar{y}_i \in \mathbb{R}$, $\underline{y}_i < \bar{y}_i$ and $x_i \in \mathbb{R}^d$ for each $i \in \{1, \dots, N\}$ and define the feasible region of linear constraints

$$F \coloneqq \left\{ \beta \in \mathbb{R}^d : \quad \underline{y}_i < \beta^\top x_i < \bar{y}_i, \quad \forall i \in \{1, \dots, N\} \right\}. \tag{2.15}$$

If F is non-empty, then there exists $p_0 \in [1, \infty)$ such that, for every $p \in (p_0, \infty]$, $\beta^* \in F$, where β^* is a solution to the optimisation problem

$$\inf_{\beta \in \mathbb{R}^d} \left\| \left(\tilde{y}_1 - \beta^\top \tilde{x}_1, \dots, \tilde{y}_N - \beta^\top \tilde{x}_N \right) \right\|_p, \tag{2.16}$$

and

$$\tilde{y}_i = \frac{(\underline{y}_i + \bar{y}_i)/2}{(\bar{y}_i - \underline{y}_i)/2}, \quad \tilde{x}_i = \frac{x_i}{(\bar{y}_i - \underline{y}_i)/2}, \quad \forall i \in \{1, \dots, N\}.$$

Proof. See Appendix 2.7.2. □

Proposition 8 states that if the linear constraints (2.15) is satisfied, then by choosing sufficiently high power $p > p_0$ then our linear regression will find coefficients β that satisfy the constraint. Notice that if $p = 2$, the loss function (2.16) is the familiar ordinary least squares. Hence, we perform the linear regression on the synthetic forwards bid and ask quotes and, if $p_0 < 2$, Proposition 8 states that the fitted prices lie within the bid-ask spread.

Therefore, the complete methodology for recovering the forwards and interest rate term structure is summarised in the following steps:

1. Recover the normalised underlying price time series as in Section 2.2.2,
2. Group the synthetic forward contract bid and ask quotes for each expiry,
3. Perform the linear regression in (2.14) via ordinary least squares,
4. Collect the coefficient B_T as in (2.14) to obtain the interest rate term structure for each expiry,
5. Revert the affine structure of dividends with $\tilde{D}_T^0 + \tilde{D}_T^1 \tilde{S}_t$ as in (2.14) to obtain The forward price time series for each expiry.

Figure 2.3 shows the estimated forward price time series and interest rate term structure, respectively. We clearly see the effect of dividends on the discounted forward prices across the different expiries as the longer expiries the more heavily the discount due to accumulated dividend. The estimated interest rate term structure should reflect the interest rate term structure for the Euro. For reference, the Euro Stoxx 50 Index was at 4,108[5] and the Euro short-term rate €STR was at -0.575%[6] on 2 December 2022. The discontinuity around the expiry 20 May 2022 can be attributed to small sample size – see Table 2.1.

The goodness of fit of the regression can be assessed by whether the fitted synthetic forwards are indeed within the observed bid-ask spreads of the synthetic forwards. Table 2.2 shows the goodness of fit in this point of view. We can observe that the fitted synthetic forwards for most expiries remain within the discounted forward bid-ask spread – the exception being the 14 April and 20 May 2022 expiries which suffer from the small sample size. The maximum relative spread – i.e. the spread between the fitted price and the mid-price divided by the observed half-spread – is mostly tight for the closest expiries,

[5]Qontigo. *EURO STOXX 50*. Retrieved from `https://www.stoxx.com/index-details?symbol=SX5E`.

[6]European Central Bank, Statistical Data Warehouse. *Euro short-term rate data – Volume-weighted trimmed mean rate*. Retrieved from `https://sdw.ecb.europa.eu/quickview.do?SERIES_KEY=438.EST.B.EU000A2X2A25.WT`.

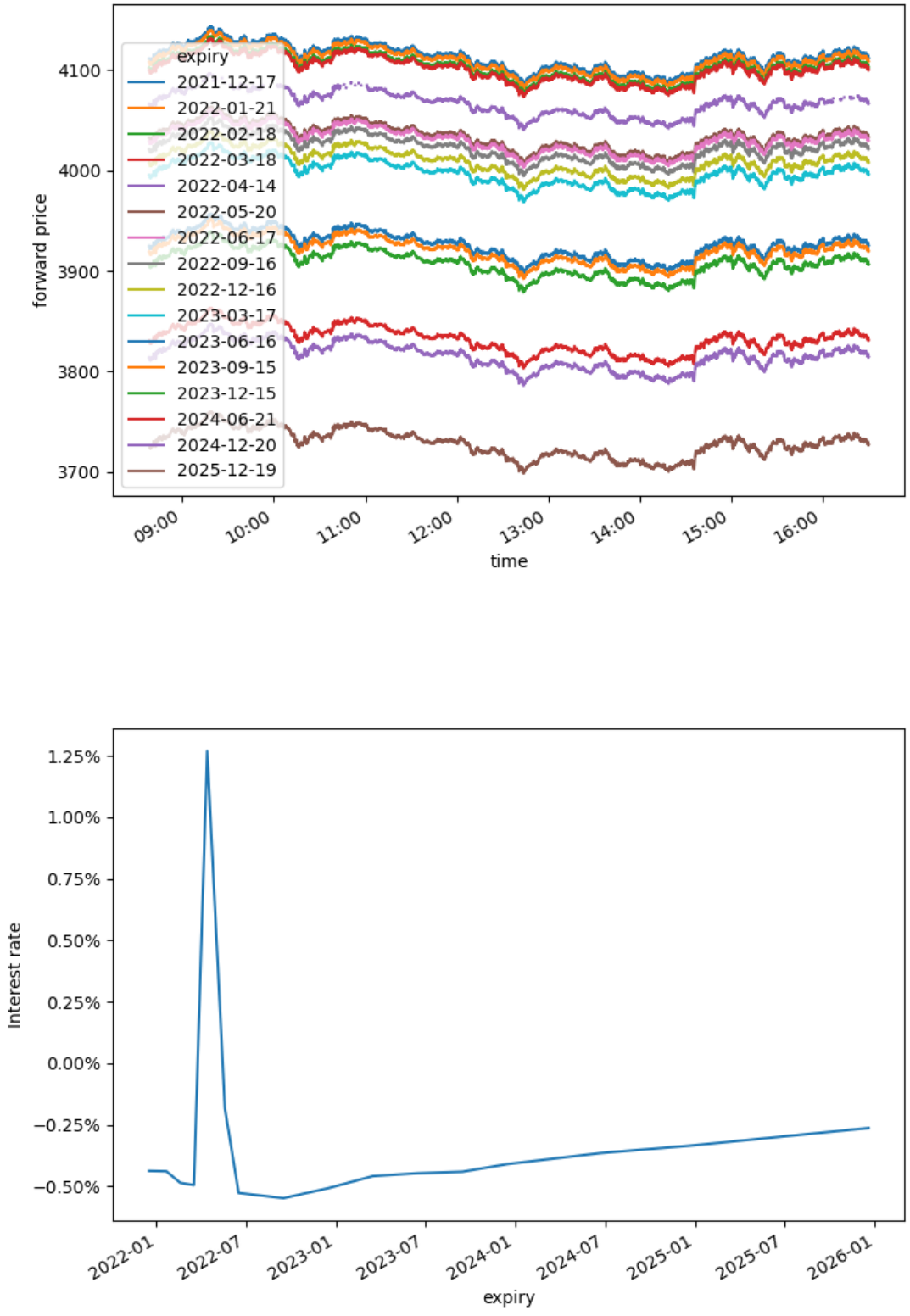


Figure 2.3: Estimated discounted forward price time series for each expiry (upper) and estimated interest rate term structure (lower) on 2 December 2022.

expiry	Maximum spread	Estimates within spreads
2021-12-17	94.202%	100.000%
2022-01-21	93.176%	100.000%
2022-02-18	81.701%	100.000%
2022-03-18	74.356%	100.000%
2022-04-14	189.162%	98.896%
2022-05-20	110.718%	99.968%
2022-06-17	68.093%	100.000%
2022-09-16	55.318%	100.000%
2022-12-16	72.110%	100.000%
2023-03-17	59.482%	100.000%
2023-06-16	73.228%	100.000%
2023-09-15	64.657%	100.000%
2023-12-15	63.101%	100.000%
2024-06-21	61.340%	100.000%
2024-12-20	70.869%	100.000%
2025-12-19	71.317%	100.000%

Table 2.2: Goodness of fit of the linear regression as measured by the relative spread of the fitted synthetic forwards compared to the observed synthetic forwards quotes.

V_0	0.047471
κ	3.646890
θ	0.048739
ν	0.596065
ρ	-0.809805

Table 2.3: Calibrated Heston parameters.

but the gap widens on longer, less liquid expiries. We conclude that the assumptions of the affine structure for dividends and constant intraday bond prices are flexible enough to obtain fitted values within the synthetic forwards bid-ask spread and, by using ordinary linear squares, we find that $p_0 < 2$.

2.2.4 Model calibration

Having explained the bootstrapping methodology we are now ready to calibrate the Heston model parameters with option prices. We recall that we use the dynamics (2.9) for the discounted forward price dynamics. On the option pricing side, given a risk-neutral measure $\mathbb{Q}$ and a short rate process $(r_t)_{t\in[0,T]}$, we price a call option with expiry T and

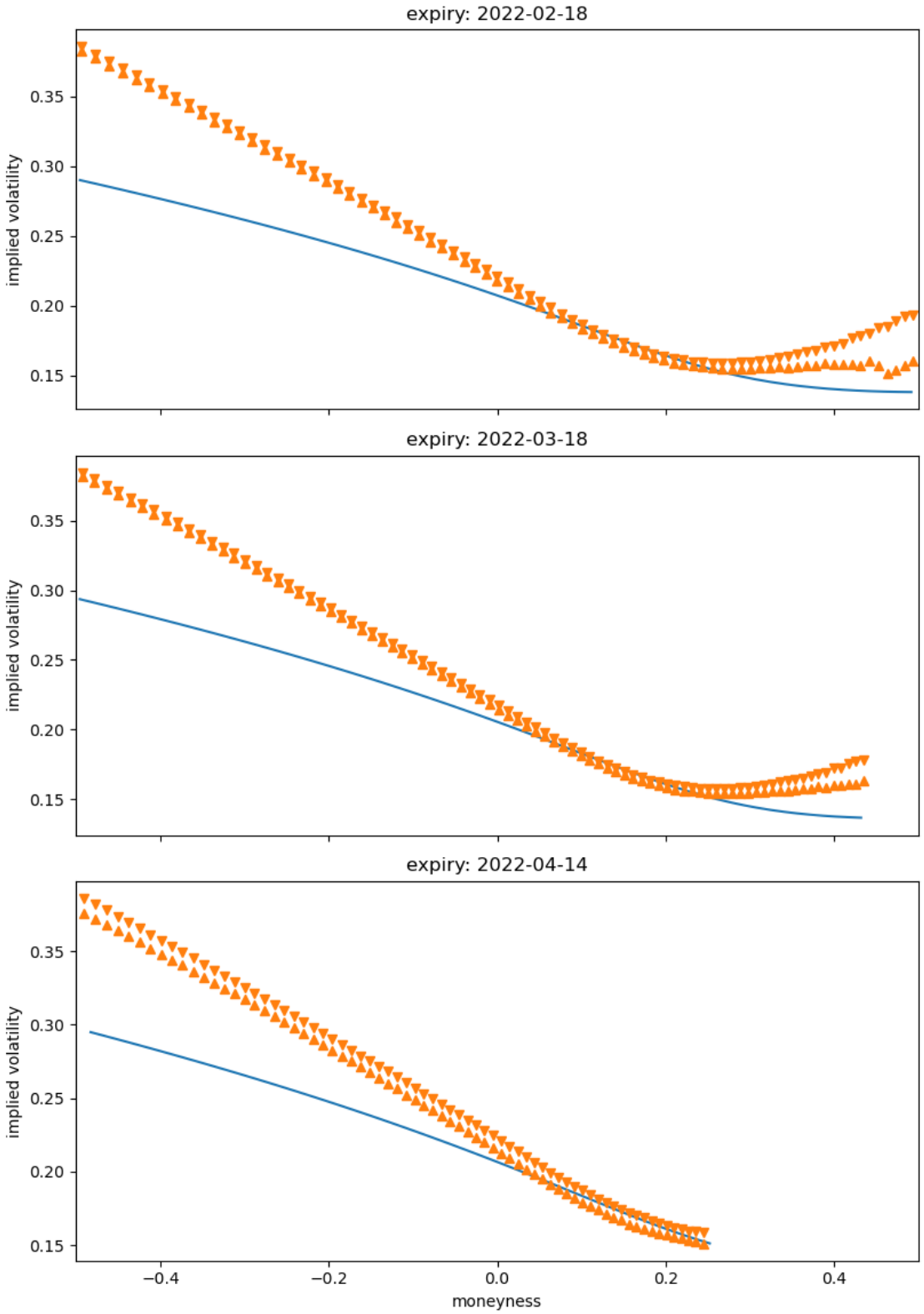


Figure 2.4: Calibrated Heston implied volatilities (solid blue line) versus market bid and ask implied volatilities (orange markers) on 2 December 2022 at 12pm. Implied volatilities with positive log-moneyness are from call options and implied volatilities with negative log-moneyness are from put options. Log-moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

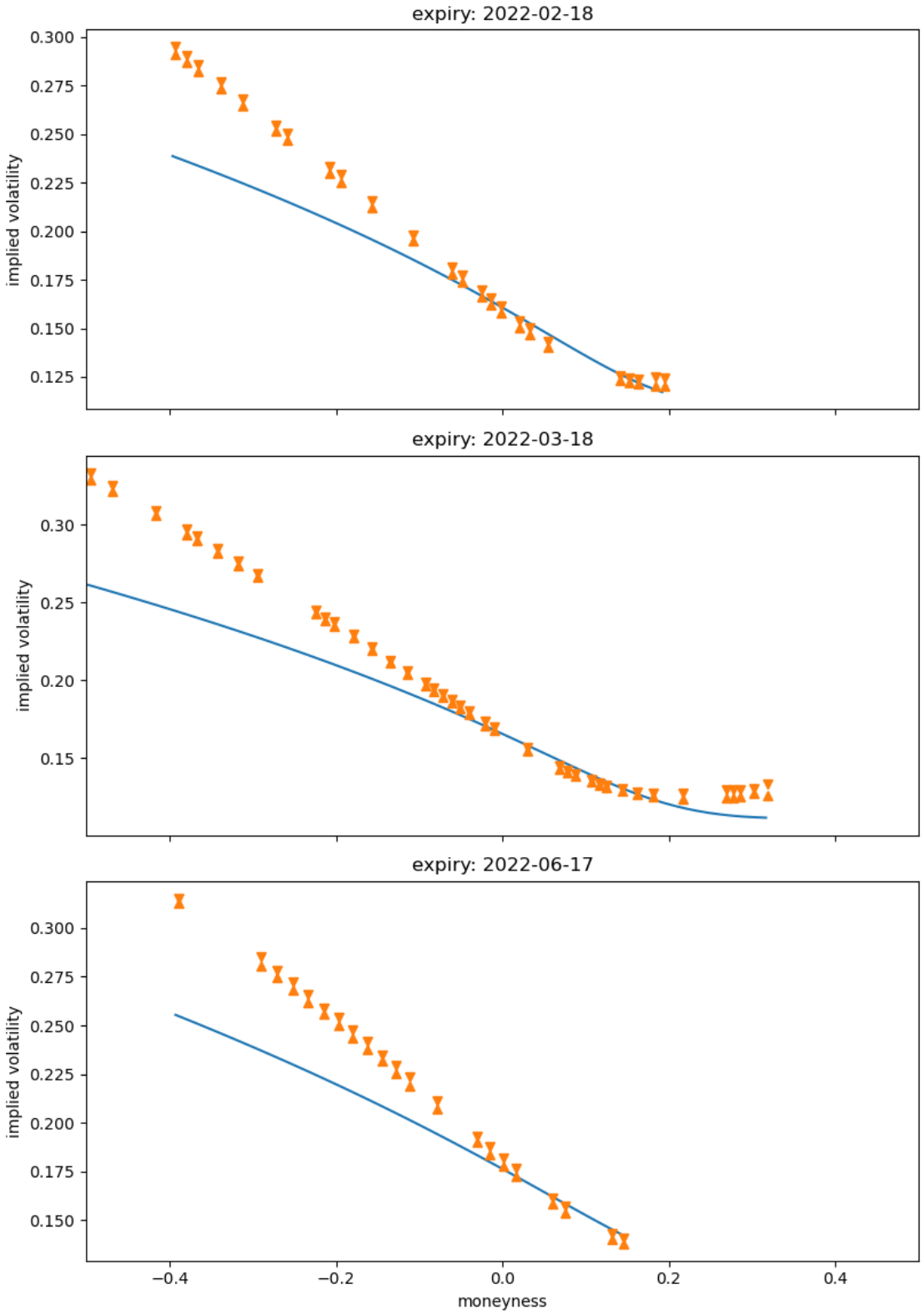


Figure 2.5: Calibrated Heston implied volatilities (solid blue line) versus market bid and ask implied volatilities (orange markers) on 22 November 2022 at 12pm. Implied volatilities with positive log-moneyness are from call options and implied volatilities with negative log-moneyness are from put options. Log-moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

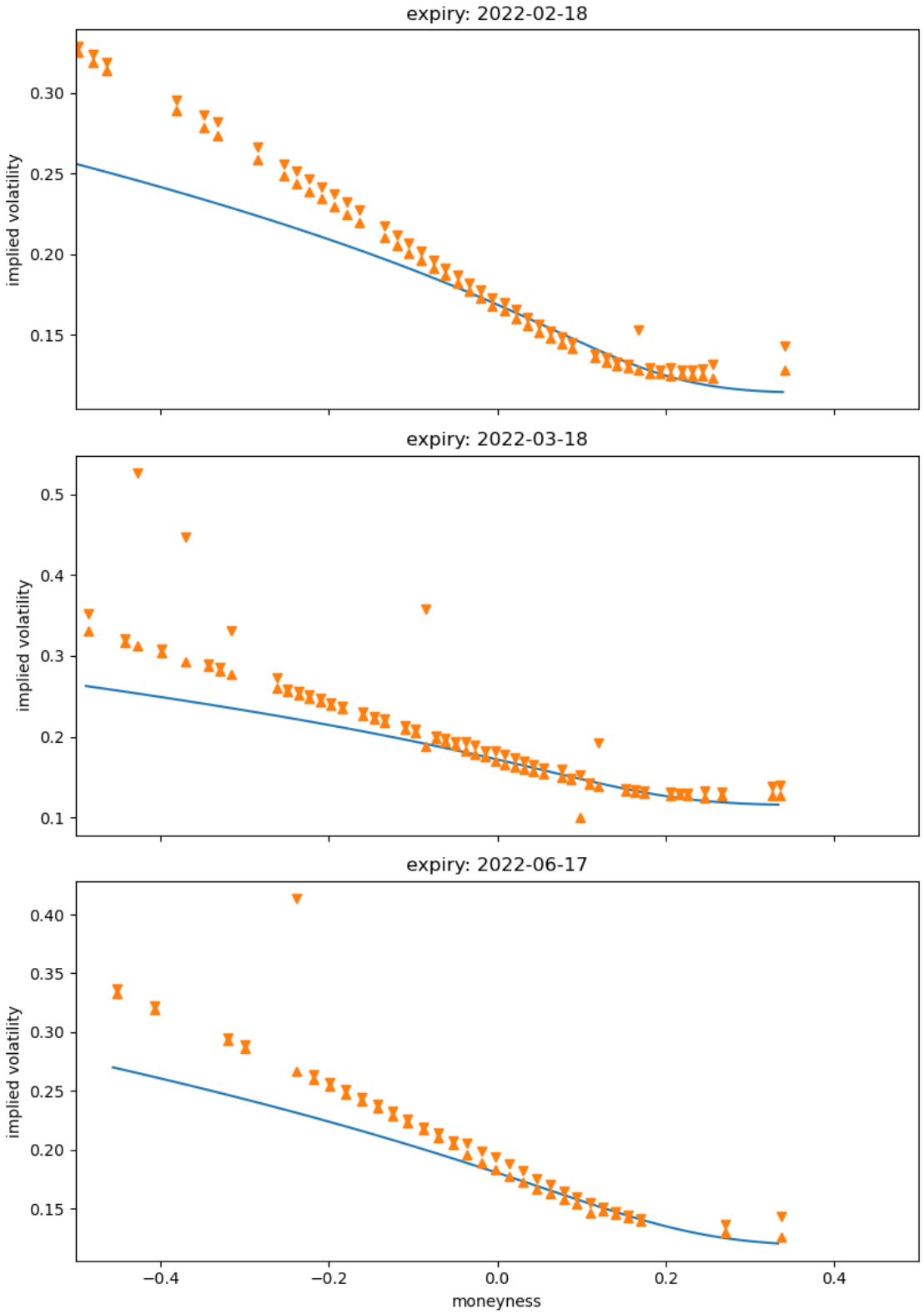


Figure 2.6: Calibrated Heston implied volatilities (solid blue line) versus market bid and ask implied volatilities (orange markers) on 16 December 2022 at 12pm. Implied volatilities with positive log-moneyness are from call options and implied volatilities with negative log-moneyness are from put options. Log-moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

strike K with

$$C_t = \mathbb{E}^{\mathbb{Q}}\left[e^{-\int_t^T r_s ds}(F_{T,T} - K)_+ \middle| \mathcal{F}_t\right], \quad \forall t \in [0, T].$$

If we further assume that the short-rate process is independent of the forward price process we have that

$$\begin{aligned} C_t &= \mathbb{E}^{\mathbb{Q}}\left[e^{-\int_t^T r_s ds} \middle| \mathcal{F}_t\right] \mathbb{E}^{\mathbb{Q}}\left[(F_{T,T} - K)_+ | \mathcal{F}_t\right] \\ &= B_{t,T}\mathbb{E}^{\mathbb{Q}}\left[(F_{T,T} - K)_+ | \mathcal{F}_t\right] \\ &= \mathbb{E}^{\mathbb{Q}}\left[(B_{t,T}F_{T,T} - B_{t,T}K)_+ | \mathcal{F}_t\right], \quad \forall t \in [0, T]. \end{aligned}$$

Of course, the same argument also follows for put options. Therefore, we can apply the discounting directly to the data and avoid embedding interest rates to the model. We use the discounted forward price and discounted strike, but one could alternatively inflate the option prices – i.e. divide the option prices by the bond price – and use the forward price and strikes with no discounting.

In reality, interest rates have an inverse relationship with equity indices. Given that interest rates here are used only for discounting, we believe this effect to be minimised. Also note that the independence assumption is weaker than the assumption of deterministic interest rates.

For the assumptions of Proposition 7 to be satisfied, we need to enforce the Feller condition. It is often mentioned in the literature, see e.g. Da Fonseca and Grasselli (2011), that the Feller condition is violated when calibrating the Heston model – in fact, even the analogous Gindikin condition for Wishart model is reported to be violated in Da Fonseca and Grasselli (2011). This means that enforcing the Feller condition to the model calibration is a considerable toll on goodness of fit. Intuitively, it restricts the how heavy the tails of the marginal distributions can be via the vol-of-vol parameter ν relative to the mean-reversion parameter κ.

On another hand, our focus is on the impact of spot volatility on option prices, which should be stronger for options that are near the money. Therefore, even though the model might not fit well at the tails, the central part of the distribution is more likely to be relevant. One could argue that the Heston model by construction does not realistically model the tail distribution of log returns, which is asymptotically exponential in the model and asymptotically power-law in empirical studies – see Dragulescu and Yakovenko (2002). The consistent violation of the Feller condition could be a reflection of the mismatch on the tail behaviour between the model and data, i.e. when considering a strip of options in a finite range of strikes, in order to achieve a tail decay akin to a power-law, the exponential decay rate must be small.

In summary, the Heston model, in particular with the Feller condition enforced, suffers from two limitations. In Section 2.1.1, we have stated that the Heston model is unable to replicate the at-the-money volatility skew with fidelity. In this section, we also highlight the limitations with respect to the tails of marginal distributions. Taking these limitations into account, we restrict the model calibration to just three expiries (second to fourth expiries) and we use as the loss function the squared relative deviation of option prices, namely

$$\frac{\left(C^{\text{Heston}}_{t,K,T} - C_{t,K,T}\right)^2}{C^2_{t,K,T}}.$$

Using the squared relative deviation as opposed to the squared absolute deviation is common in the literature – see Da Fonseca and Grasselli (2011) for a discussion. We have noticed, however, that our calibration results were quite sensitive as to how the loss function is normalised. This is likely due to the lack of flexibility imposed by the Feller condition. We opt for the price normalisation so that we put more weight on options near and out of the money.

We calibrate the full Heston model once on the snapshot of 12 pm on 2 December 2022, which is in the middle of the dataset. Then, we recalibrate for the spot volatility parameter for all other time slices. The calibrated model parameters for the 2 December snapshot is shown in Table 2.3. The smile fit for 2 December is depicted in Figure 2.4. The smile fit for the first and last day of data are also shown in Figures 2.5 and 2.6.

Looking at the parameters in Table 2.3, we see that the optimal parameters are at the constraint border imposed by the Feller condition. We can see this reflected in Figure 2.4, where the model smiles are considerably flat compared to the market smile, especially when close to expiry. The Figures 2.5 and 2.6 are a sanity check that the model calibration does not look too different when using the same κ, θ, ν and ρ for other dates and we also notice the difference in the number of active options among these three dates.

2.3 Volatility estimation

2.3.1 Overview

In this section, we estimate spot volatility using the method we propose based on the Heston model recalibration. In Section 2.3.2, we revisit the small time asymptotics in the context of spot volatility estimation to motivate our proposed method. In Section 2.3.3, we estimate spot volatility in a controlled study using simulated option quotes with the Heston parameters we found in the previous section. Finally, in Section 2.3.4, we estimate

spot volatility using real options data and compare with the findings of the controlled experiment.

2.3.2 Limitations in spot volatility estimation via realised volatility

In Section 2.1.3, we have mentioned that estimating spot volatility via realised volatility is subject to two types of errors: microstructure noise and approximation of spot volatility by integrated volatility. To make this more precise, let us revisit the continuous SDE model (2.1). We wish to estimate the spot volatility σ_t. For simplicity, we consider the historical volatility estimate as realised volatility, i.e.

$$\hat{\sigma}^2_{[t,t+Nh]} = \sum_{i=1}^{N} \log\left(\frac{S_{t+(i+1)h}}{S_{t+ih}}\right),$$

where $N \in \mathbb{N}$ is the number of samples and h measures the sampling frequency. Realised volatility is a consistent estimator of integrated volatility $\sigma_{[t,t+Nh]}$ as $N \to \infty$ with constant horizon Nh. If the horizon Nh is small, then spot volatility can be approximated by integrated volatility. Therefore, we can identify three sources of error: (i) measurement error $\hat{\sigma}_{[t,t+Nh]} - \sigma_{[t,t+Nh]}$, (ii) "localisation" error $\sigma_{[t,t+Nh]} - \sigma_t$ and (iii) microstructure noise, which is exogenous and include effects such as the bid-ask bounce and rounding error.

The typical scenario for integrated volatility estimation is for daily volatility estimates. In this case, sampling prices at finer granularity is motivated by the goal of reducing measurement error. Then, if the sampling granularity is too fine, e.g. by using tick data, then estimates are subject to microstructure noise. Therefore, several methods have been developed to this end – as those mentioned in Section 2.1.3. In the context of the above example, the horizon Nh is constant, so better estimates are obtained by increasing the number of samples N and, consequently, reducing h. However, when the object of study is intraday spot volatility, we need to control both the measurement error, but also the localisation error, and for the latter we need a small horizon $Nh \to 0$.

For the regime in which $Nh \to 0$, it is useful to analyse measurement errors from the point of view of small time asymptotics. On the Heston example for Proposition 7, we have seen that the volatility of $(S_t)_{t\in[0,T]}$ is asymptotically constant at small time scales. This intuitively means that the finer the granularity the more difficult it is to estimate volatility from sample paths of $(S_t)_{t\in[0,T]}$. This difficulty corresponds to the measurement error.

To illustrate this difficulty, suppose that we are estimating spot variance for a process $(s_t := \sigma_t W_t)_{t\in[-T,T]}$ at two consecutive time intervals $[-\tau, 0]$ and $[0, \tau]$ using realised variance and fix the number of samples N for each time interval. If τ is small enough, then $s_t \approx \sigma_0 W_t$ for $t \in [-\tau, \tau]$, so that the both estimates $\hat{\sigma}_{[-\tau,0]}$ and $\hat{\sigma}_{[0,\tau]}$ will be σ_0^2 plus the measurement error – which follows the χ^2 distribution under the small time regime. Therefore, any change from one estimate to the other would not reflect the change from, say, $\sigma^2_{[-\tau,0]}$ to $\sigma^2_{[0,\tau]}$, but it would be purely χ^2-distributed noise.

Notice that, by keeping N constant and decreasing the time step τ, the sampling frequency N/τ increases. However, it does not increase quickly enough to yield sensible granular estimates. A natural question to ask is how frequently do we need to sample in order to control the measurement error. The following proposition answers this question for a specific toy model.

Proposition 9. Let $(s_t, \sigma_t)_{t\geqslant 0}$ be stochastic processes that follows the SDEs

$$\begin{aligned} ds_t &= \sigma_t dW_t, \\ d\sigma_t &= \nu dZ_t, \end{aligned}$$

where $\nu > 0$, $s_0, \sigma_0 \in \mathbb{R}$ and $(W_t, Z_t)_{t\geqslant 0}$ is a vector of independent Brownian motions. Define the realised variances on $[-T, 0]$ and $[0, T]$:

$$\hat{\sigma}^2_{\pm} = \frac{1}{T}\sum_{i=1}^{N} \left(s_{\pm iT/N} - s_{\pm(i-1)T/N}\right)^2.$$

Then, $\hat{\sigma}^2_+$ and $\hat{\sigma}^2_-$ are unbiased estimators of integrated variance, i.e.

$$\mathbb{E}\left[\hat{\sigma}^2_+\right] = \frac{1}{T}\mathbb{E}\left[\int_0^T \sigma_t^2 dt\right], \quad \mathbb{E}\left[\hat{\sigma}^2_-\right] = \frac{1}{T}\mathbb{E}\left[\int_{-T}^0 \sigma_t^2 dt\right].$$

Furthermore, if $N^\alpha T \to 1$ as $T \to 0$, then the signal-to-noise ratio admits the following subcritical, critical and supercritical regimes.

$$\frac{\mathbb{E}\left[\hat{\sigma}^2_+ - \hat{\sigma}^2_-\right]^2}{\mathrm{Var}\left(\hat{\sigma}^2_+ - \hat{\sigma}^2_-\right)} \xrightarrow{T\to 0} \begin{cases} 0, & \alpha < 1/2, \\ \frac{\nu^4}{2\sigma_0^4}, & \alpha = 1/2, \\ \infty, & \alpha > 1/2. \end{cases}$$

Proof. See Appendix 2.7.4. □

We can interpret Proposition 9 as follows. We start with sufficiently small intervals $[-T, 0]$ and $[0, T]$ and realised variances for each interval based on N equidistant samples. Then

if we wish to shrink the intervals to $[-T/\beta, 0]$ and $[0, T/\beta]$ for $\beta > 1$ with no detriment in signal-to-noise ratio, we require at least $\beta^2 N$ equidistant samples in each smaller interval.

The toy model in Proposition 9 deserves some comments. The small-time asymptotics of a process $(S_t, \sigma_t)_{t\in[0,T]}$ that satisfies the assumptions of Proposition 7 would be

$$dS_t = \sigma_0 dW_t,$$
$$d\sigma_t = \nu_0 dZ_t,$$

where the Brownian motions $(W_t)_{t\in[0,T]}$ and $(Z_t)_{t\in[0,T]}$ could be correlated. To capture the effect of stochastic volatility, we let $dS_t = \sigma_t dW_t$ and to simplify the computations, we assume $(W_t)_{t\in[0,T]}$ and $(Z_t)_{t\in[0,T]}$ are independent. Therefore, the toy model is not entirely artificial: it captures the essential dynamics of more complicated stochastic models at small time scales. It is worth noting that one could alternatively consider a toy model in which the Brownian motions are perfectly (negatively) correlated but this would not be very useful because then option prices would ultimately be driven only by underlying price changes.

When using real data, the granularity of the samples is a given rather than a property that we can control. Hence, how granular we can estimate spot volatility via historical volatility is fundamentally constrained. We have seen in Section 2.1.3 a useful reference that a reasonable granularity for spot volatility estimates considering the measurement and localisation errors would be in the order of minutes. On the other hand, from what we have discussed, it seems that the measurement error is a limiting obstacle when estimating spot volatility with fine granularity via historical volatility. This is the motivation for a method that relies instead on options data.

We propose the use of a stochastic volatility model to recover the spot volatility from option prices, rather than relying on the observation of the asset price process itself. For each snapshot of the implied volatility surface, we recalibrate the pricing model only for the latent state variables – that is the spot volatility in the Heston model – while keeping the model parameters constant. As we have mentioned in Section 2.1.4, this method follows the stochastic volatility model by the letter, because the state variables are dynamic while the model parameters are constants. It is also worth noting that, similarly to the Black-Scholes implied volatility, the inverse mapping from option prices to spot volatility is one-to-one in the Heston model – one can simply observe that the derivative with respect to the spot volatility is positive.

The advantage of using a stochastic volatility model to recover the spot volatility is that the estimation is cross-sectional and as such is free of localisation error. Besides, the use of multiple option prices would intuitively average out rounding error, and in this sense it

would be robust to microstructure noise. The tradeoff, however, is that the spot volatility estimates are model dependent and are thus subject to model bias.

2.3.3 Volatility estimation on simulated data

We perform the volatility estimation first on a controlled setting with simulated option prices. Using the parameters from Table 2.3, we simulate the underlying price and volatility processes by discretising the Heston model as follows

$$S_{i+1} = S_i \exp\left(-\frac{1}{2}V_i h + \sqrt{V}_i Z_i h\right),$$
$$V_{i+1} = \kappa(\theta - V_i)h + \nu\sqrt{V}_i\left(\rho Z_i\sqrt{h} + \sqrt{1-\rho^2}Z_i^{\perp}\sqrt{h}\right),$$

where h is the time scaling factor, Z_i and $Z_i^{\perp}$ are independent standard normal random variables. The step size is one second.

With the underlying price and volatility processes, we compute option prices with the Heston pricing formula for the first 7 expiries as in Table 2.1 with the "today" date being 2 December 2021. For each expiry, we simulate 100 calls and 100 puts for log-moneyness ranging from -0.5 to 0.5, where we define log-moneyness as $\log(K/S_t)/\sqrt{T-t}$. We obtain in total 200 options for each expiry, which is approximately what we have for the most liquid expiries on average according to Table 2.1. Prices are computed for each second between 8am and 4.30pm in line with the real trading session.

As an effort to add more realism to the simulated prices, we round the option prices by the nearest tick and recover the underlying using the bootstrap methodology explained in Section 2.2 and also calibrate the Heston model at 12pm. The calibration result on simulated data is depicted in Figure 2.7.

We finally perform the spot volatility estimation by recalibrating the Heston model only for V_t for each second of the data. We compare the estimated volatility $\sqrt{V}_t$ with the true spot volatility that we have simulated. Additionally, we also compute realised volatilities at each 5-minute time interval by annualising the sample standard deviation on 1-second log returns. We provide confidence intervals for the realised volatility estimates via stationary bootstrap – see Politis and Romano (1994). We have not provided confidence intervals for the spot volatility estimates, but it could be computed using spatial bootstrap techniques. The estimation results are depicted in Figure 2.8. We highlight that the realised volatility is computed for reference and is not used elsewhere in the analysis.

In Figure 2.8, we observe that our proposed method accurately estimated the spot volatility from options data, despite the rounding of option prices. Of course, in our simulated

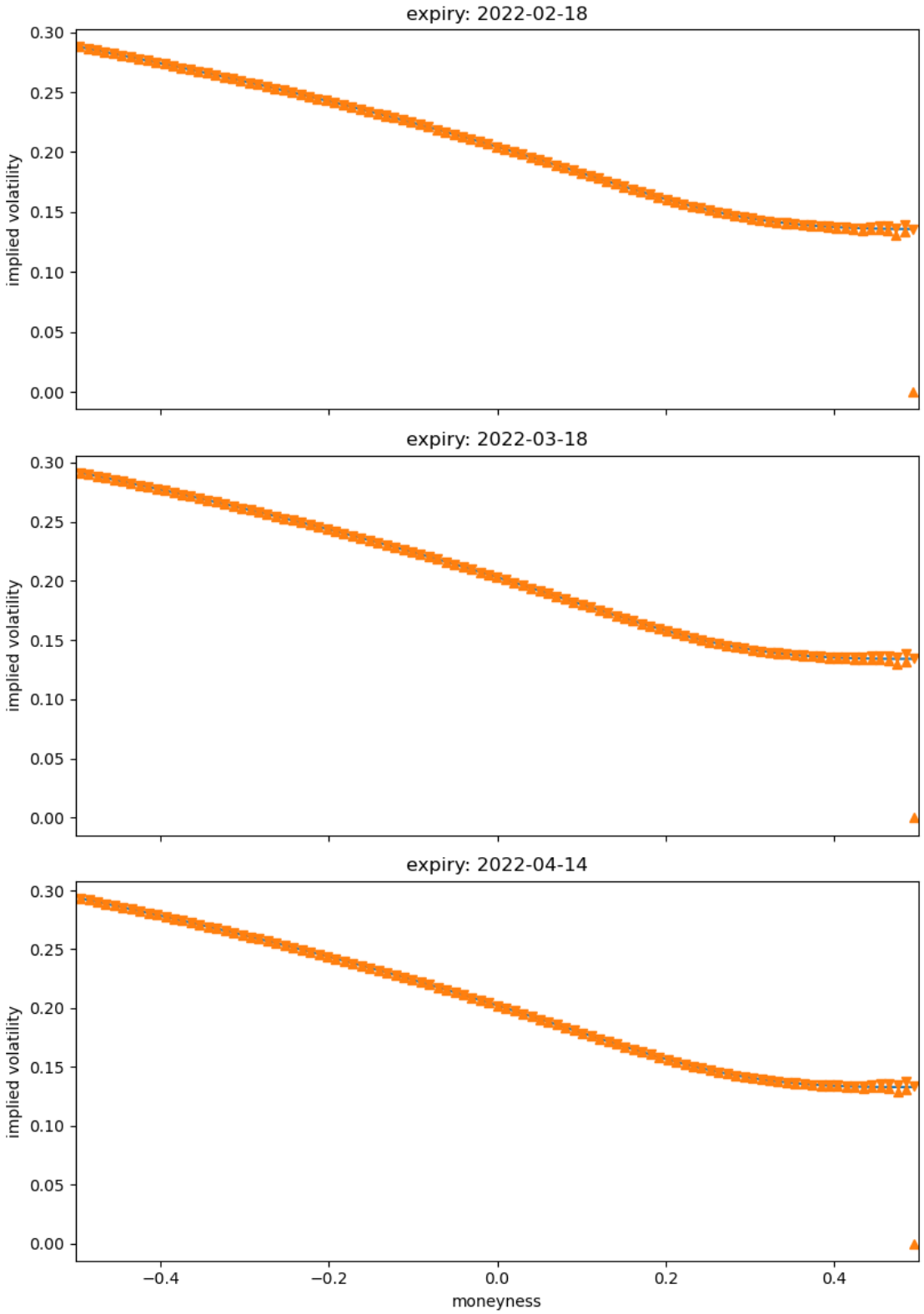


Figure 2.7: Calibrated Heston implied volatilities (solid blue line) versus market bid and ask implied volatilities (orange markers) on 2 December 2021 at 12pm. Implied volatilities with positive log-moneyness are from call options and implied volatilities with negative log-moneyness are from put options. Log-moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

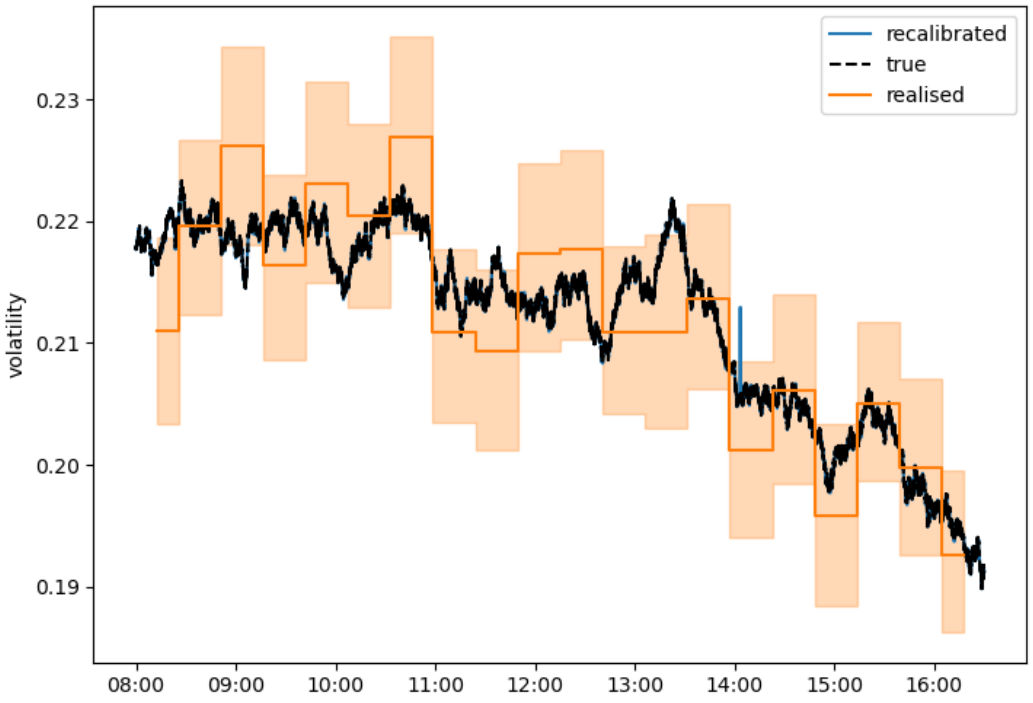


Figure 2.8: Comparison among true spot volatility, spot volatility estimated by recalibrating the Heston model and 5-minute realised volatility with confidence bounds on the simulated data.

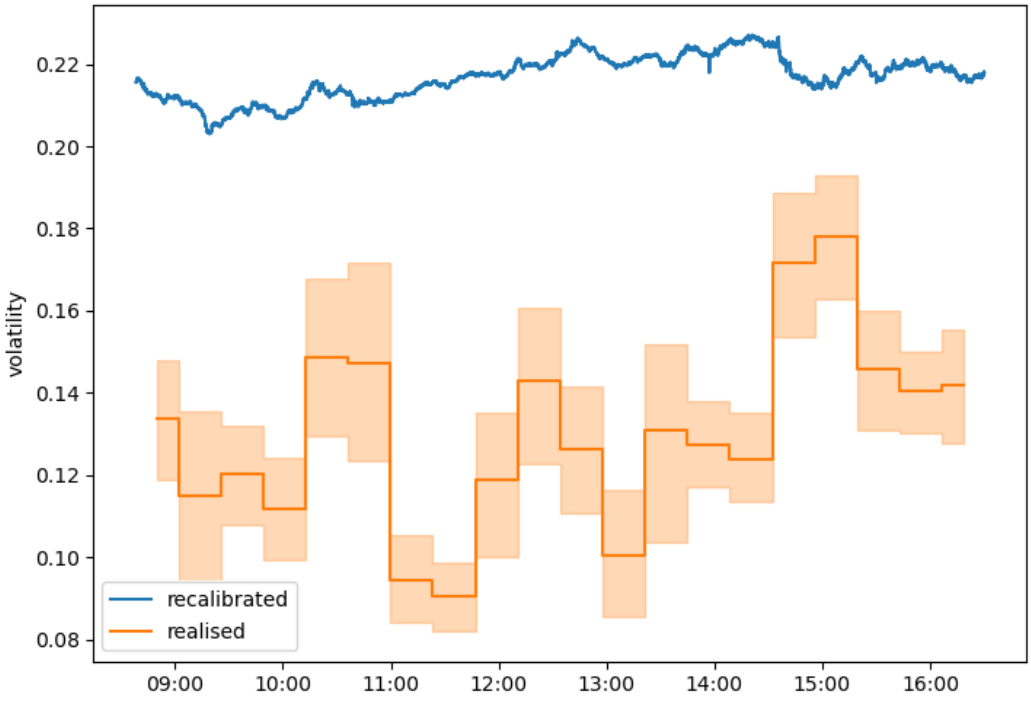


Figure 2.9: Comparison between spot volatility estimated by recalibrating the Heston model and 5-minute realised volatility with confidence bounds on 2 December 2021.

study, there is no model bias because the models used for simulation and for spot volatility estimation are the same. On the other hand, we see that the 5-minute realised volatilities are quite noisy and, as we postulated, we cannot reliably infer the true spot volatility changes from the estimated changes of the realised variance. In fact, we observe that consecutive confidence intervals overlap too often, which indicates that most apparent changes in realised volatility are subject to measurement error. On the other hand, the true spot volatility often lies within the 95% confidence bands of the realised volatility, which is reassuring.

2.3.4 Volatility estimation on real data

We now estimate spot volatility in real data. The estimates for spot volatility and realised volatility measurements are computed in the same way as with the numerical experiment. Figure 2.9 depicts the result for 2 December 2021 and Figure 2.10 depicts the results for all available dates.

The spot volatility estimates appear disconnected from the realised volatility measure-

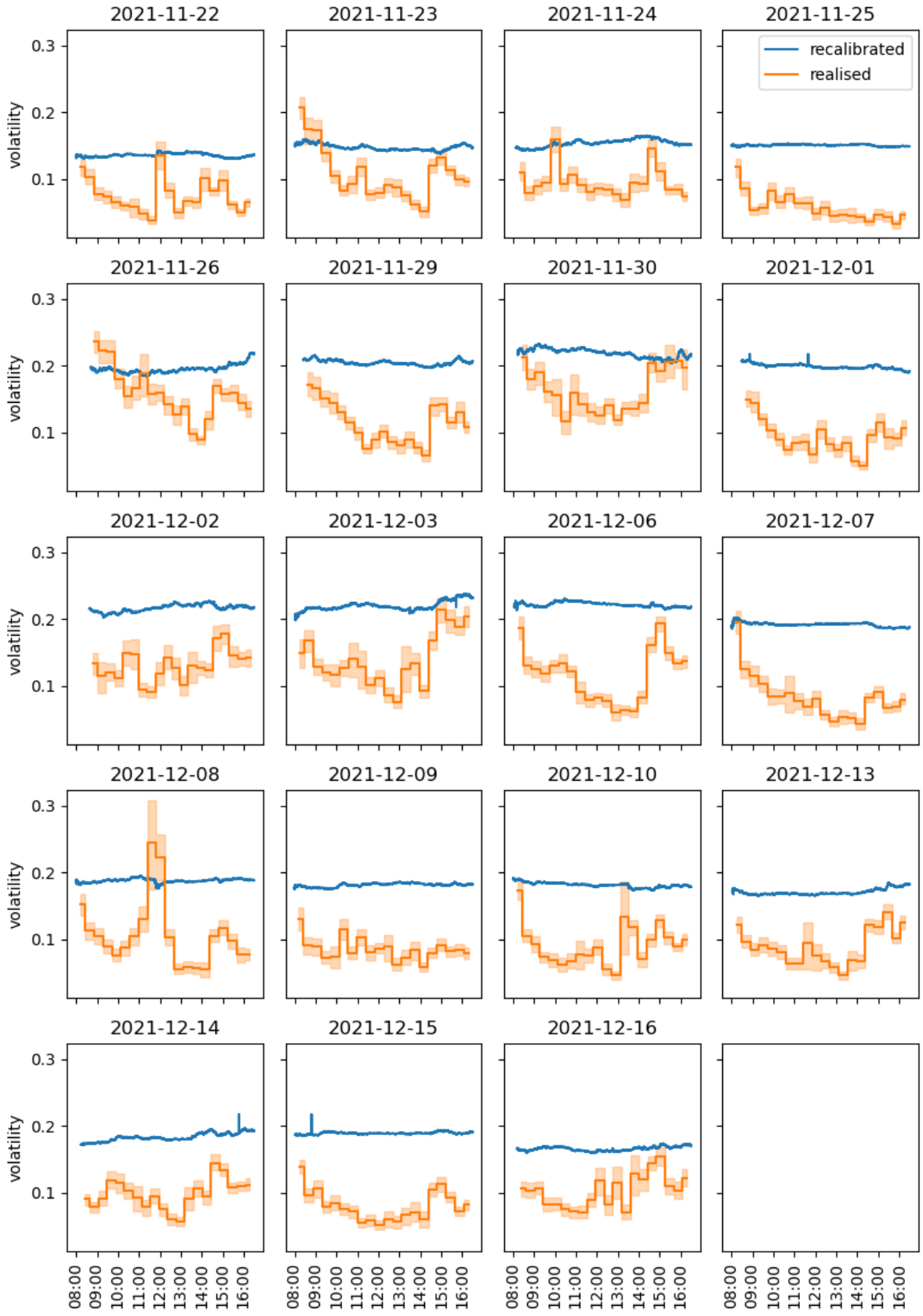


Figure 2.10: Comparison between spot volatility estimated by recalibrating the Heston model and 5-minute realised volatility with confidence bounds for all available dates.

ments across all dates. The estimates of spot volatility do not overlap the confidence bounds of realised volatility and they also appear to vary less throughout the day than the realised volatility measurements. This contrasts with the numerical experiment, which suggests that model bias could be severely affecting the results.

We note that the Feller condition imposes a restriction on θ but not on V_t. Given that both θ and V_t have an effect on the level of the Heston smile, it could be that the Feller condition makes θ artificially small and V_t artificially high, thus systematically overestimating spot volatility. In fact, we have also performed the spot volatility estimation without enforcing the Feller condition, in which case the estimated spot volatility and realised volatility were similar in level. However, even in the absence of the Feller condition, the estimated spot volatilities varied throughout the day compared to the realised volatility.

Another important question is whether changes in the spot volatility estimates are a reflection of changes in true spot volatility. From the numerical study, we have seen that we cannot assess this by naively comparing with changes in realised volatility, especially considering that we are not controlling for microstructure effects. Further investigation would be needed in this regard.

We believe that the spot volatility estimates could be improved by employing a more realistic model, such as the Wishart model. More recent and successful models such as the N-factor Bergomi model and rough volatility models could also, in principle, be applied. However, these models have as an input the full forward volatility term structure in functional form, and thus the choice of the parametrisation of the functional form will directly impact the spot volatility estimates.

In the case of rough volatility models, we also note that Proposition 7 would not apply. It fails due to the rough component in the state process $(X_t)_{t\in[0,T]}$. Taking a fractional Brownian motion with Hurst parameter $H < 1/2$ as an example of a rough process, its self-similarity property tells us that

$$\frac{W^H_{ht}}{h^H} \overset{\mathcal{D}}{=} W^H_t, \quad \forall h > 0,$$

where $\overset{\mathcal{D}}{=}$ means equality in distribution. The scaling factor of a rough fractional Brownian motion is smaller than square root and therefore it dominates over the square root scaling in (2.8).

2.4 Greeks estimation

2.4.1 Overview

In this section, we estimate the Greeks via a linear regression to assess the intraday effect of spot volatility on option prices. In Section 2.4.2, we specify the linear regression under the lens of the small time asymptotics of the Heston model. In Section 2.4.3, we perform the linear regression with the same simulated data as in Section 2.3.3 and perform a semi-partial R^2 analysis. Finally, in Section 2.4.4, we estimate the Greeks using real options data and compare with the findings of the controlled experiment.

2.4.2 Linear regression setup

Even though the option prices are a nonlinear function of the market state parameters (e.g. underlying price and spot volatility), the small-time asymptotics for options in Proposition 7 shows that the function is linear at small enough time scales. This implies that we can estimate the options first-order Greeks by a linear regression of option prices on its state variables. In the context of the Heston model, we perform the following linear regression:

$$\begin{aligned} &C_t^{K,T,P} - C_{t-1}^{K,T,P} = \\ &\Delta_{\text{Heston}}^{K,T,P}\left(S_t - S_{t-1}\right) + \mathcal{V}_{\text{Heston}}^{K,T,P}\left(V_t - V_{t-1}\right) + \Gamma_{\text{Heston}}^{K,T,P}\left(S_t - S_{t-1}\right)^2 + \Theta_{\text{Heston}}^{K,T,P} + \epsilon_t^{K,T,P}, \end{aligned} \quad (2.17)$$

where the time is in seconds and spans over a trading session and the strike-expiry-payoff triples (K, T, P) cover all call and put options, where the payoff component $P \in \{\text{call}, \text{put}\}$ indicates whether the option is a call or a put. Note that neither theta nor gamma are first-order Greeks, and therefore, they should have no contribution to the variance of option price changes under small time scales.

We note that the linear model (2.17) is naturally decoupled for each triple (K, T, P). Therefore, we perform the regression separately for each option. This does not affect the estimates of the regression coefficients and it allows for analysing other regression results, such as R^2, by strike, expiry and payoff. Overall regression results also have their value, however given that our main goal is to assess the effect of spot volatility on option prices, which is highly dependent on the payoff specification of option, especially its moneyness, and thus pooling all options together would dilute the effect of spot volatility.

The linear regression is performed via OLS, and we assume standard conditions that ensure the estimates are consistent and asymptotically normal. Namely, if we put (2.17)

in the form

$$Y_i = \beta \cdot X_i + \epsilon_i, \tag{2.18}$$

for $i \in \{1, \ldots, N\}$, our assumptions are[7]

Linearity The model is of the form (2.18),

Stationarity and weak dependence The process $(Y_i, X_i)_{i=1}^N$ is jointly stationary and weakly dependent,

No Perfect Collinearity No independent variable is a linear combination of the others.

Zero Conditional Mean Conditional on the independent variables, the mean of the errors is zero: $\mathbb{E}[\epsilon_i | X_i] = 0$,

Homoskedasticity Conditional on the independent variables, the variance of the errors is constant, i.e. $\mathrm{Var}(\epsilon_i | X_i) = \sigma^2$,

No Serial Correlation Conditional on the independent variables, the errors are not serially correlated, i.e. $\mathbb{E}[\epsilon_i \epsilon_j | X_i, X_j] = 0$ for all $i \neq j$.

The first four assumptions are sufficient for the OLS estimators $\hat{\beta}$ to be consistent. The last two assumptions are also needed for the OLS estimators $\hat{\beta}$ to be asymptotically normal, which we use when constructing confidence intervals.

2.4.3 Greeks estimation on simulated data

Figures 2.11-2.18 compare, across all strikes of the selected expiries, the 95% confidence intervals of the Greeks estimated via linear regression model in (2.17) with the hypothetical Greeks under the small time asymptotics using the calibrated Heston model. We have selected liquid expiries starting from the third expiry. We have skipped the first two expiries to highlight the effect of spot volatility, which is more pronounced towards longer expiries.

In this controlled study, we observe a tight agreement between the linear regression estimates and the hypothetical values. Having the hypothetical Greeks inside the confidence interval means that we cannot reject the hypothesis that they differ. Furthermore, in the case of the first-order Greeks, having the 2.5% bound above zero means that we reject the hypothesis that the state variables, in particular spot volatility, are not significant. Therefore, in our numerical experiment, we clearly see the small time asymptotics in action and, in particular, the effect of spot volatility on option price changes.

[7] See, for example, Chapter 11 in Wooldridge (2015) for reference.

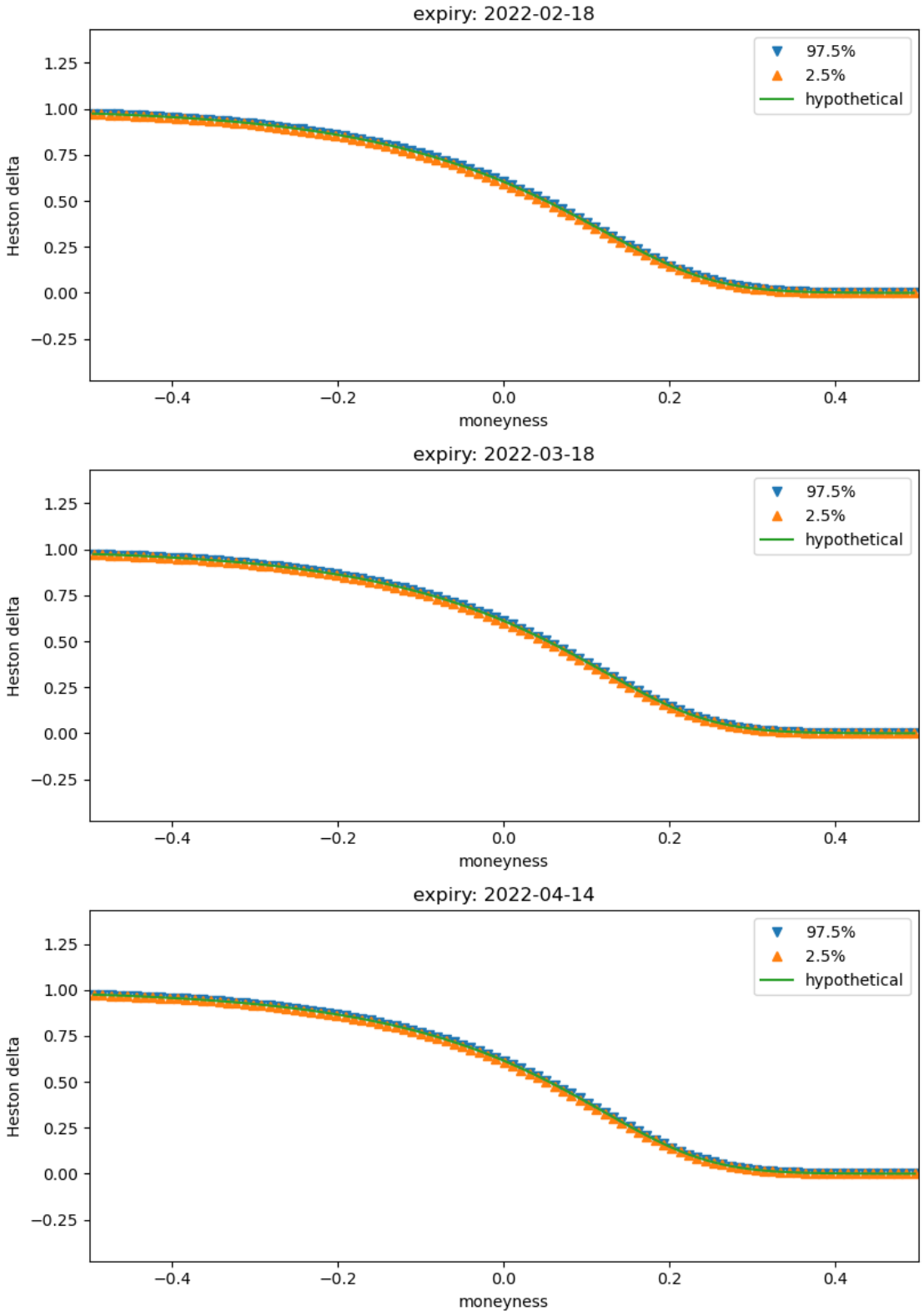


Figure 2.11: Greek delta under the Heston model across call options on the simulated data. The triangles represent the 95% confidence interval of the delta estimated by the linear regression. The solid line is the theoretical delta computed using the calibrated Heston model on the average forward price and average volatility. Moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

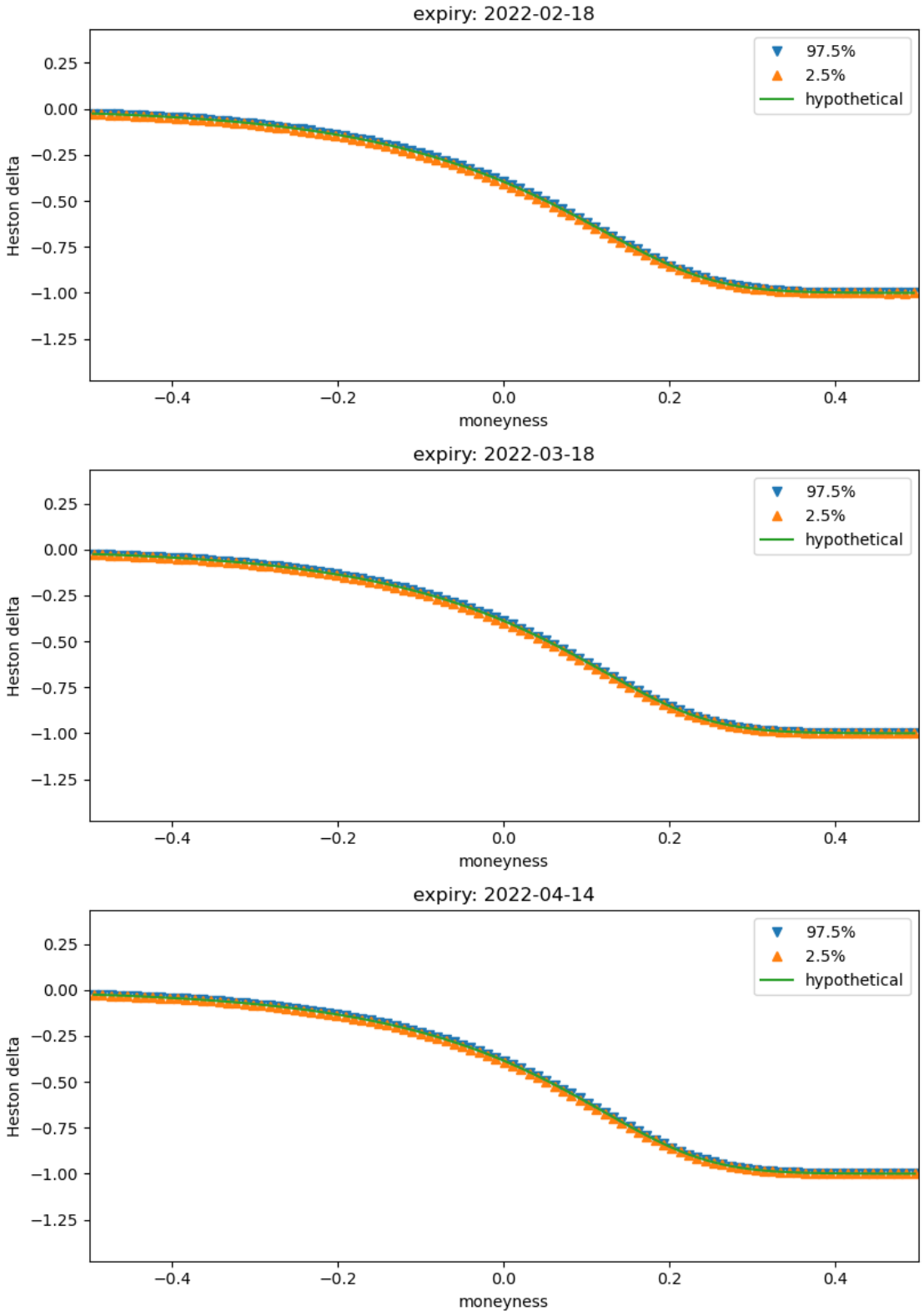


Figure 2.12: Greek delta under the Heston model across put options on the simulated data. The triangles represent the 95% confidence interval of the delta estimated by the linear regression. The solid line is the theoretical delta computed using the calibrated Heston model on the average forward price and average volatility. Moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

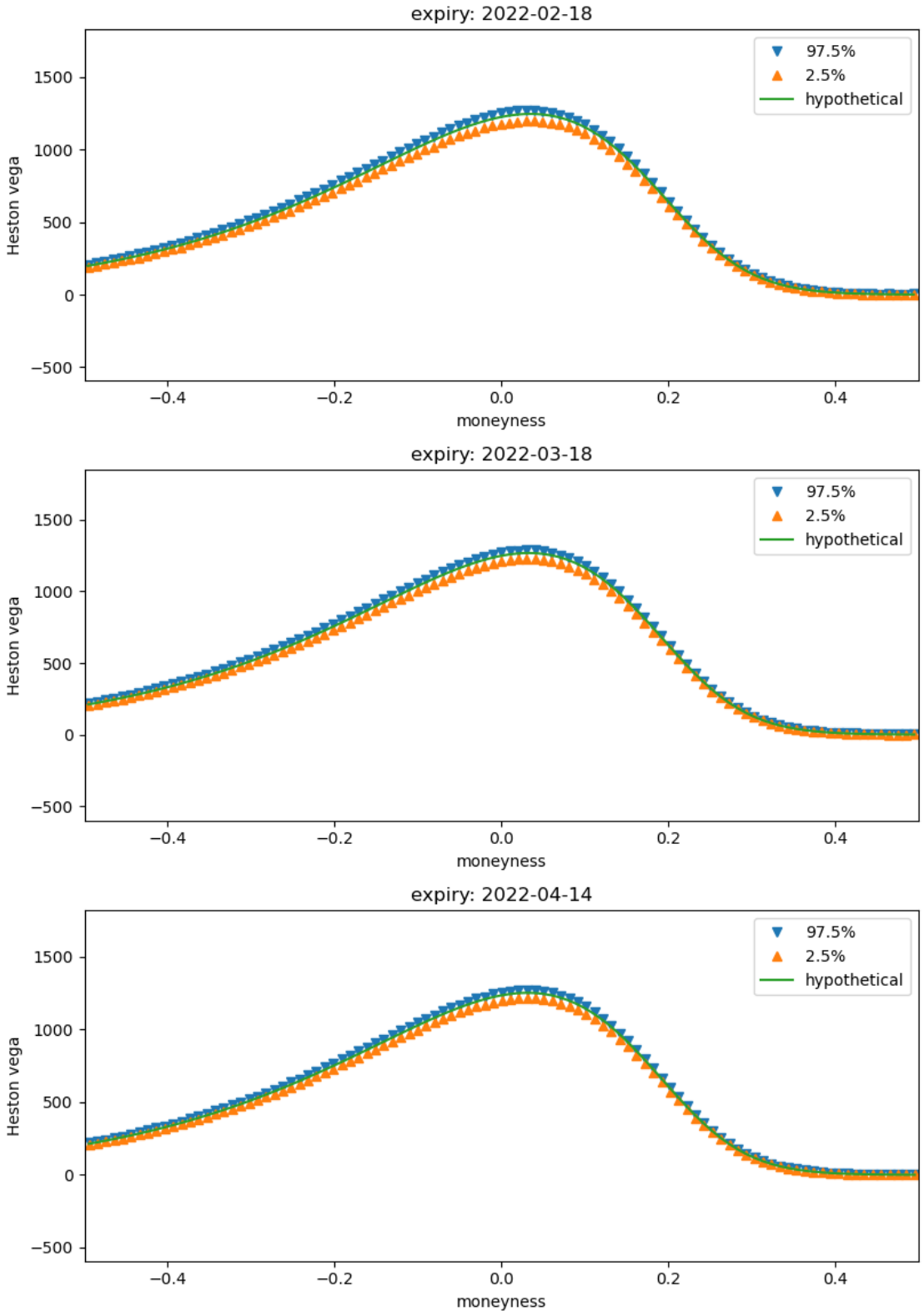


Figure 2.13: Greek vega under the Heston model across call options on the simulated data. The triangles represent the 95% confidence interval of the vega estimated by the linear regression. The solid line is the theoretical vega computed using the calibrated Heston model on the average forward price and average volatility. Moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

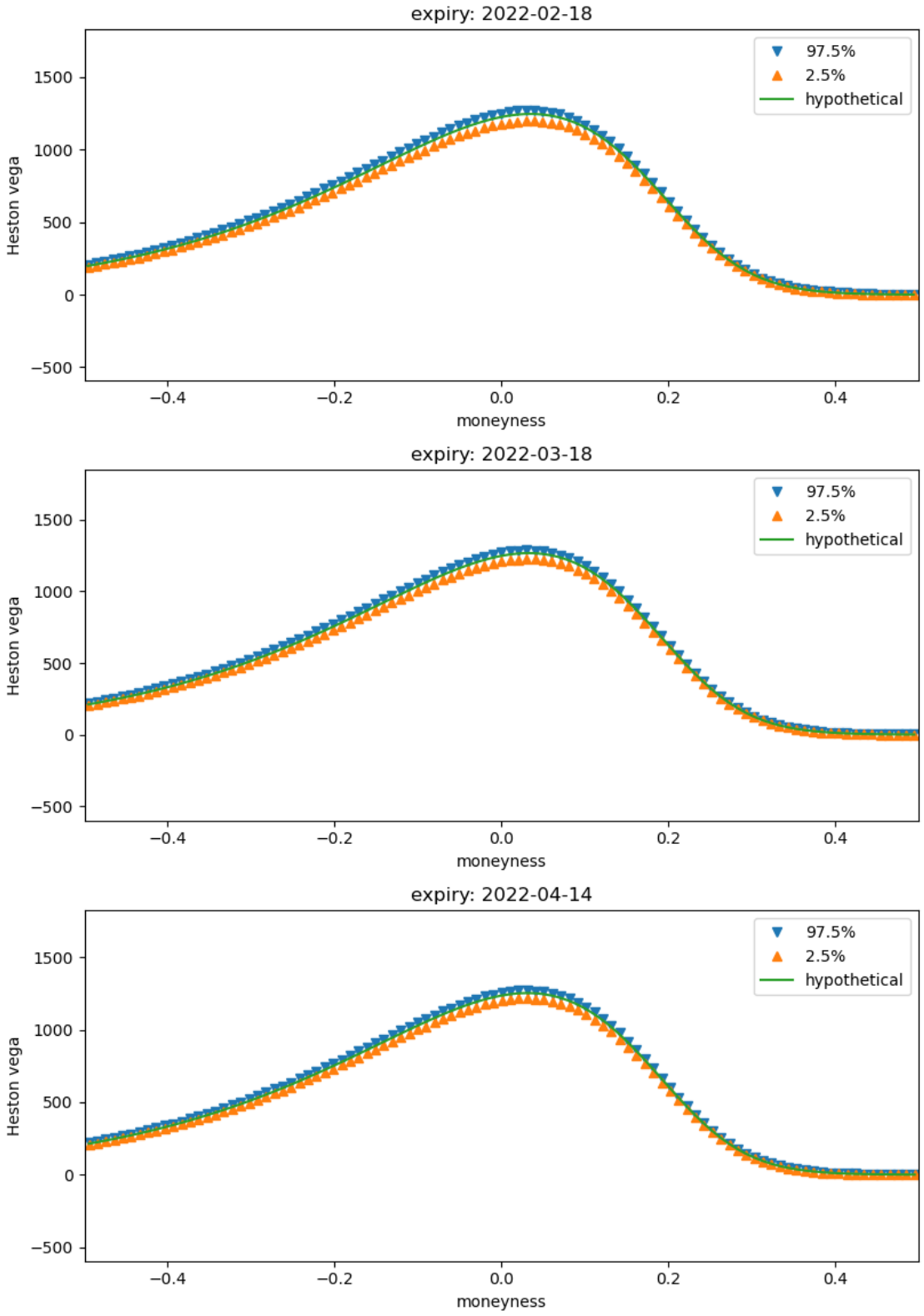


Figure 2.14: Greek vega under the Heston model across put options on the simulated data. The triangles represent the 95% confidence interval of the vega estimated by the linear regression. The solid line is the theoretical vega computed using the calibrated Heston model on the average forward price and average volatility. Moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

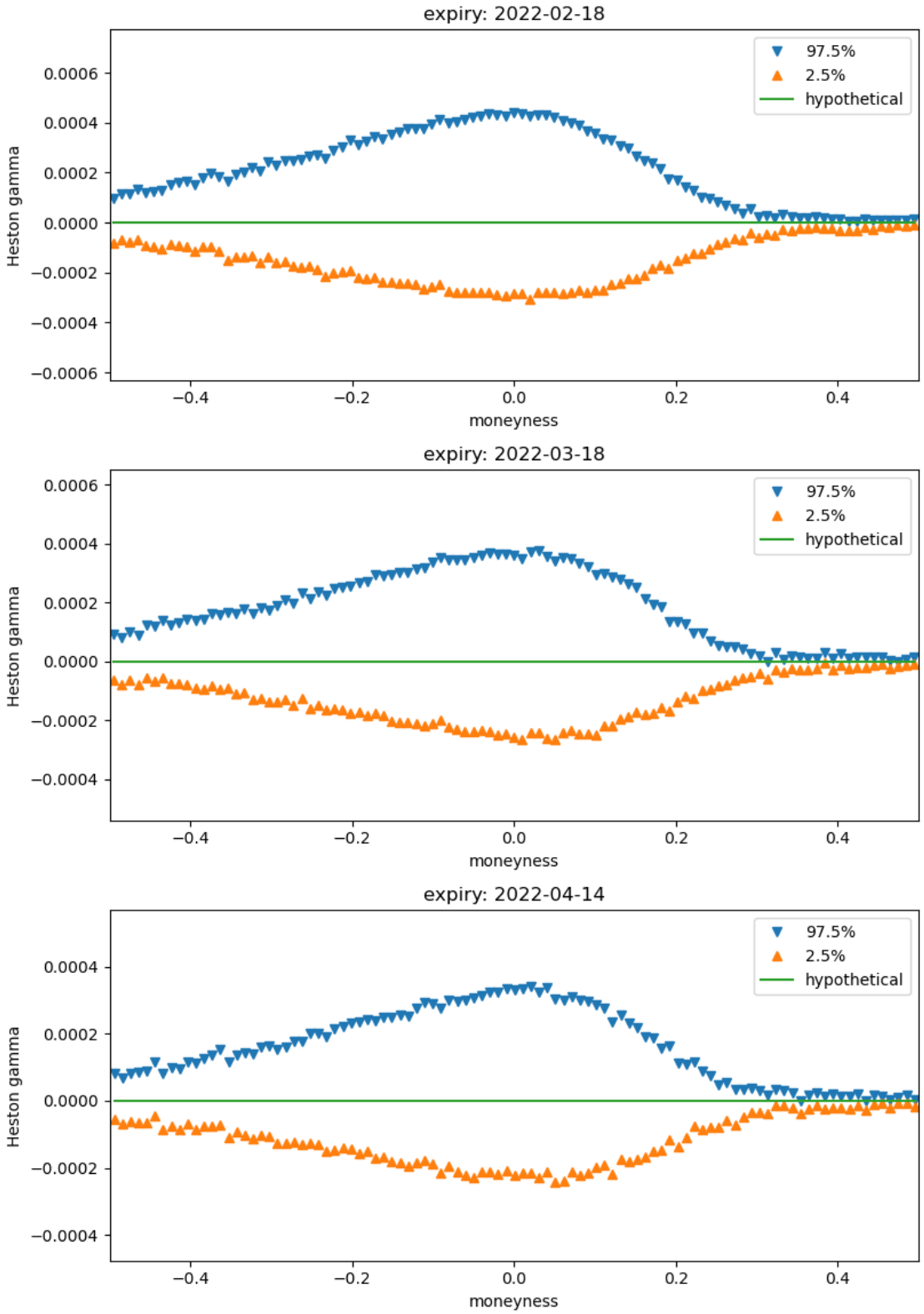


Figure 2.15: Greek gamma under the Heston model across call options on the simulated data. The triangles represent the 95% confidence interval of the gamma estimated by the linear regression. The solid line is the theoretical gamma which, under the small time asymptotics is zero. Moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

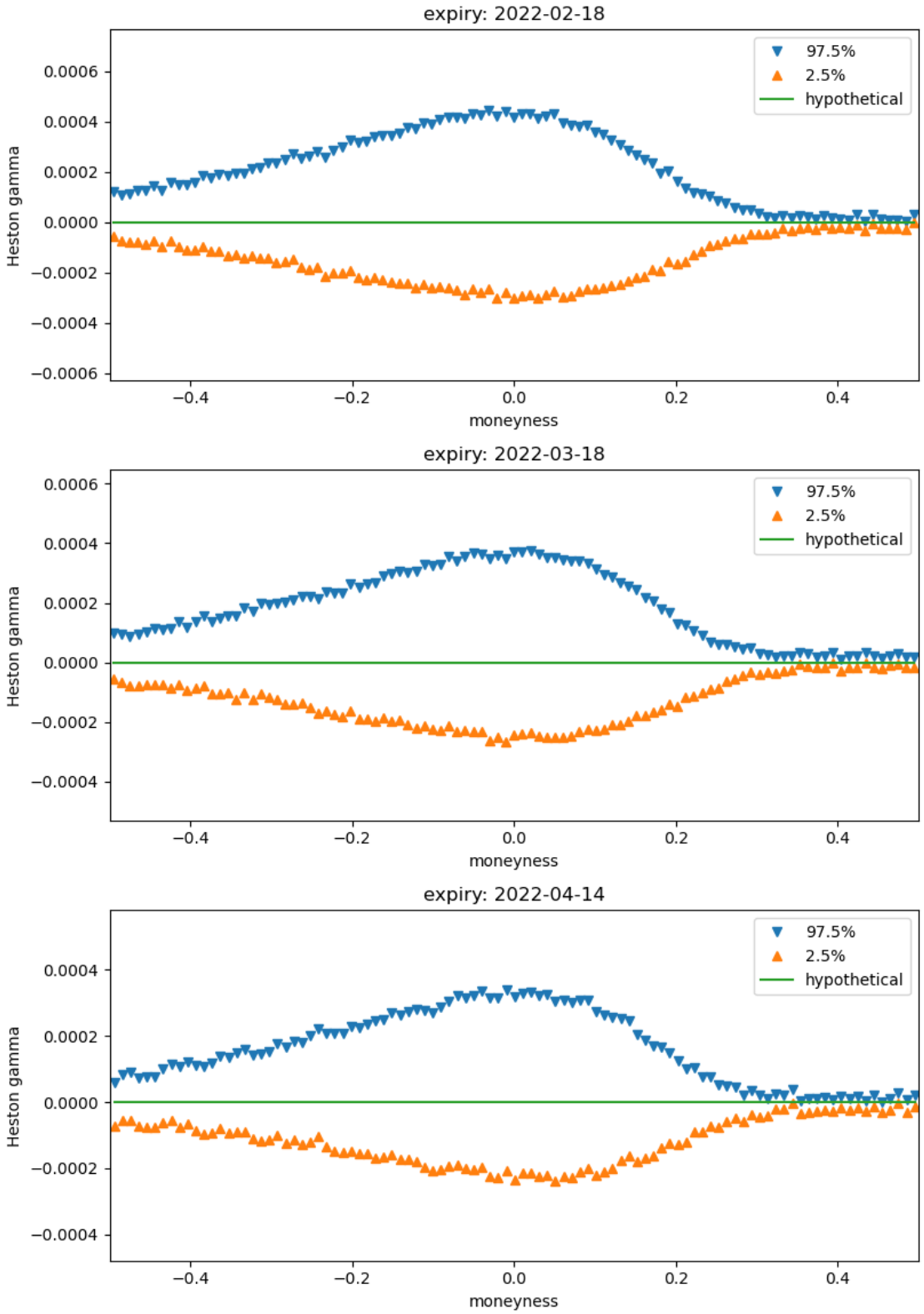


Figure 2.16: Greek gamma under the Heston model across put options on the simulated data. The triangles represent the 95% confidence interval of the gamma estimated by the linear regression. The solid line is the theoretical gamma which, under the small time asymptotics is zero. Moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

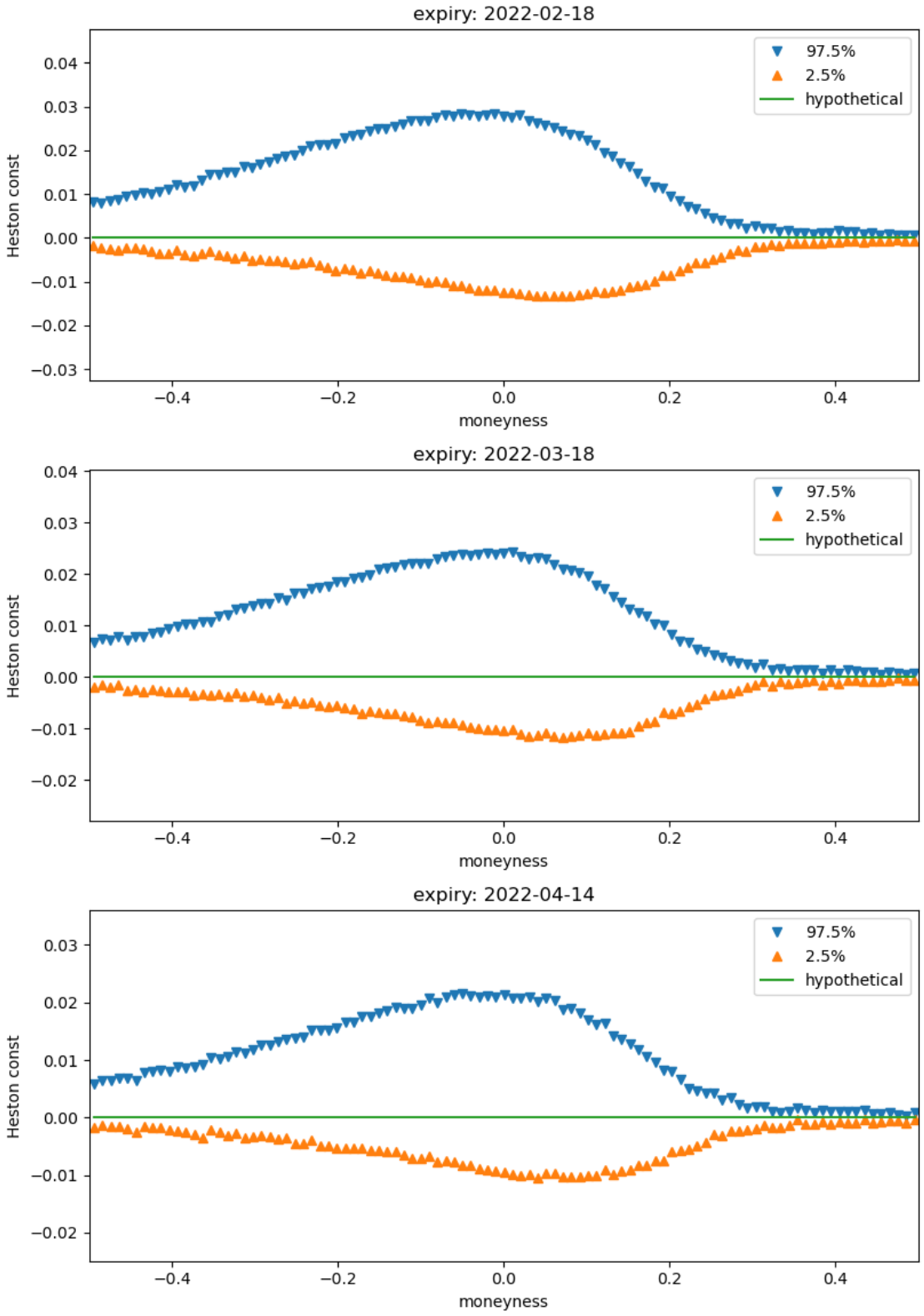


Figure 2.17: Greek theta in seconds under the Heston model across call options on 2 December 2021. The triangles represent the 95% confidence interval of the theta estimated by the linear regression. The solid line is the theoretical gamma which, under the small time asymptotics is zero. Moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

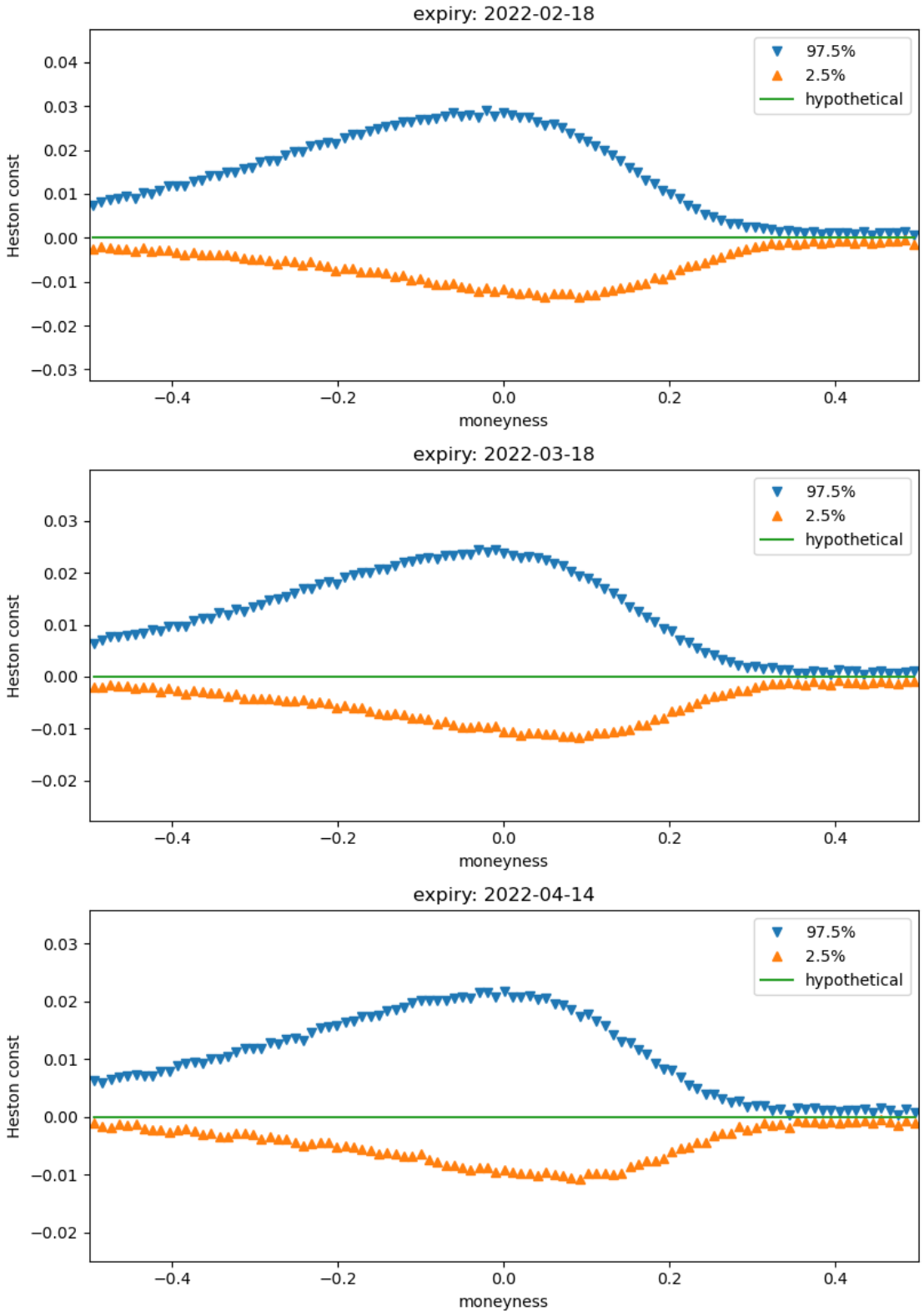


Figure 2.18: Greek theta in seconds under the Heston model across put options on 2 December 2021. The triangles represent the 95% confidence interval of the theta estimated by the linear regression. The solid line is the theoretical gamma which, under the small time asymptotics is zero. Moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

As for the Greeks gamma and theta, we notice the widening of the confidence intervals around the money. This could indicate some residual contribution that arises from longer time scales. In particular, we note the effect of theta, which highlights that options are not martingales in the historical $\mathbb{P}$ measure.

Besides assessing whether or not the spot volatility has a significant effect on option price changes, we also wish to quantify such effect. For this, we employ the semi-partial R^2.

To introduce the semi-partial R^2, we first recall the ANOVA decomposition

$$\text{TSS} = \text{ESS} + \text{RSS},$$

where TSS is the total sum of squares (of the exogenous variables), ESS is the sum of the explained sum of squares (explained by the endogenous variables) and RSS is the residual sum of squares. The coefficient of determination R^2 is a measure of explained variance in the sense that, from its definition we have

$$1 = R^2 + \frac{\text{RSS}}{\text{TSS}} \Rightarrow R^2 = \frac{\text{ESS}}{\text{TSS}}.$$

Now, suppose we would like to isolate the variance contribution of one of the endogenous variables, say X_1. This is not simply its covariance with respect to the exogenous variable because X_1 can be correlated with the other endogenous variables. Instead, consider a submodel in which X_1 is removed, and let ESS' denote the explained sum of squares for this submodel. Then,

$$\text{TSS} = (\text{ESS} - \text{ESS}') + \text{ESS}' + \text{RSS},$$

which implies

$$1 = sR^2 + (R')^2 + \frac{\text{RSS}}{\text{TSS}},$$

where sR^2 is the semi-partial R^2 on X_1. It is a measure of the variance explained by X_1 beyond what is already explained by the other endogenous variables – which is measured by $(R')^2$. For more details, see Chapter 3 in Cohen (2013).

Figures 2.19 and 2.20 display the R^2 and the spot volatility sR^2, respectively, across all strikes of the selected expiries. In Figure 2.19, we see that almost all options attain $R^2 = 1$, which means that the linear approximation at such a small time scale seems appropriate. The exception is with out-of-the-money call options, which could be due to the tick value rounding.

In Figure 2.20, we can observe that out-of-the-money call options at around 0.3-0.4 moneyness achieve very high sR^2 of up to ca. 70%, which shows how strong the effect of spot volatility can be on some options. From the Greeks plots – Figures 2.11 and

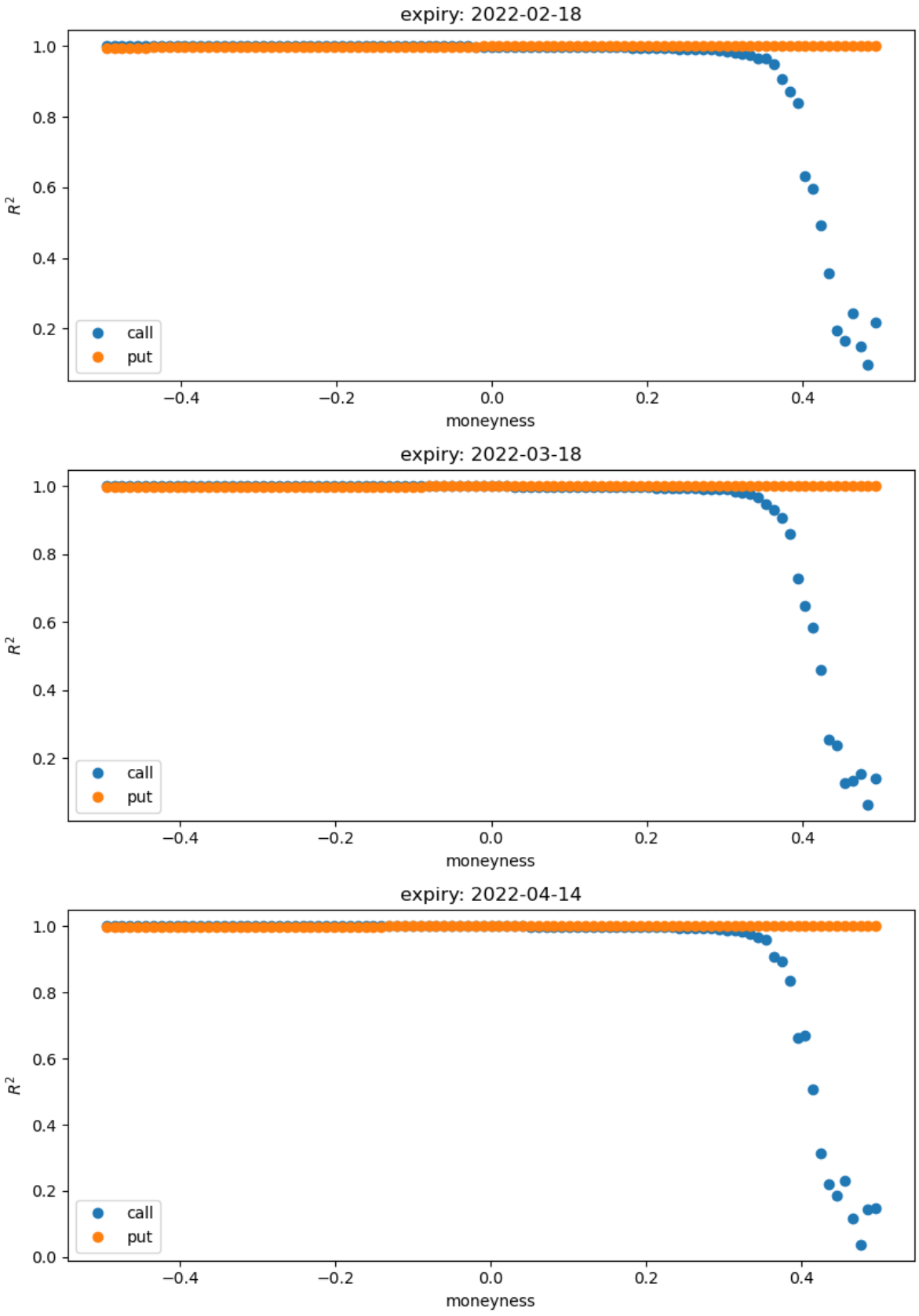


Figure 2.19: The R^2 for the linear regression in (2.17) across put and call options on simulated data. Moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

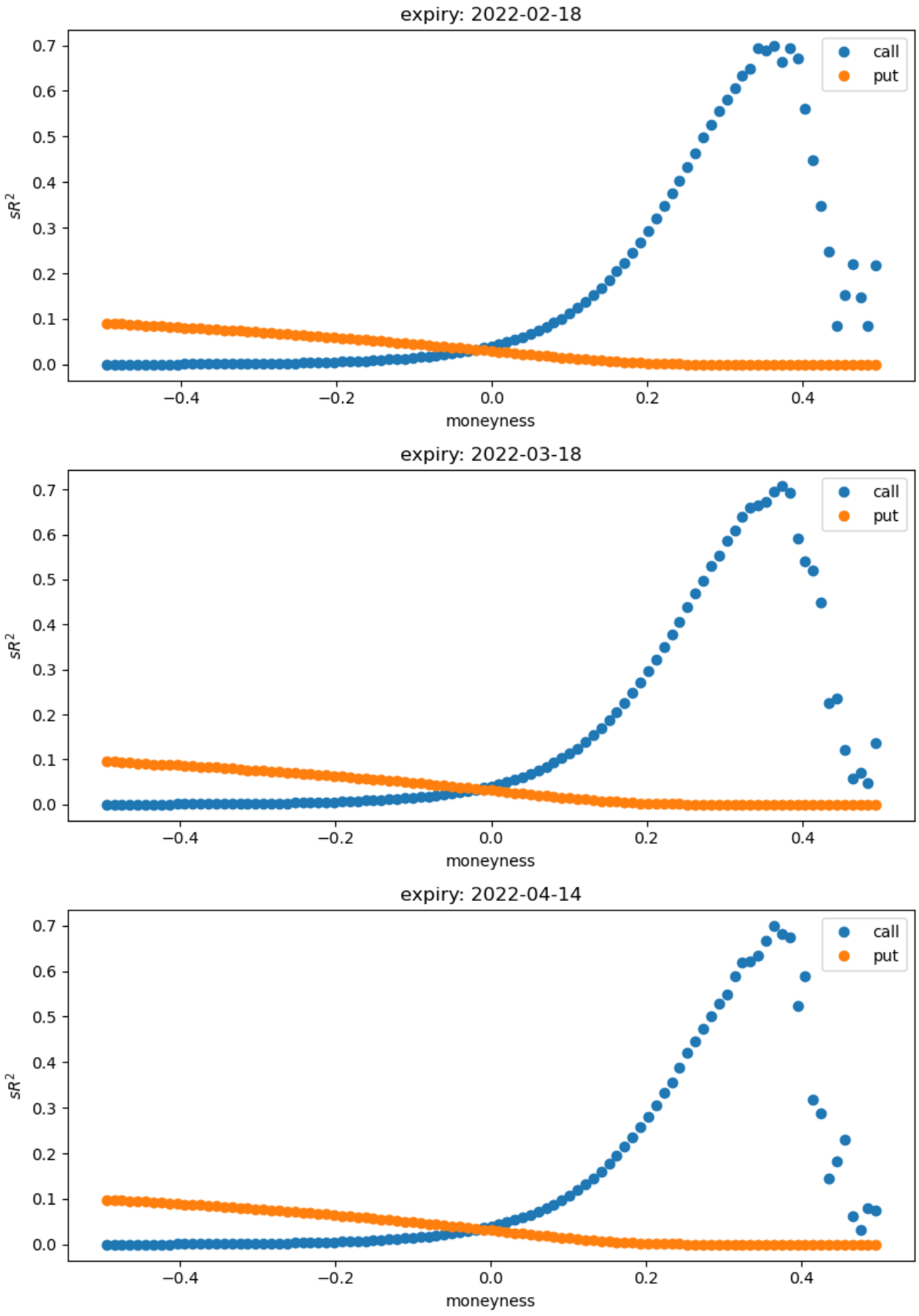


Figure 2.20: The spot volatility semi-partial R^2 for the linear regression in (2.17) across put and call options on simulated data. Moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

2.13 –, we observe that vega seems high relative to delta around this region. In contrast, put options peak at more modest levels of ca. 10%. This asymmetry can also be observed in the corresponding Greeks plots for put options in Figures 2.12 and 2.14. This is likely caused by the leverage effect, encoded in the parameter ρ.

2.4.4 Greeks estimation on real data

We now perform the same analysis in real data on 2 December 2021. This date has been chosen because it sits in the middle of the date range of our dataset and presents a high number of active options. Figures 2.21-2.28 compare, across all strikes of the selected expiries, the 95% confidence intervals of the Greeks estimated via linear regression model in (2.17) with the hypothetical Greeks under the small time asymptotics using the calibrated Heston model. Despite the mismatch between our spot volatility estimates and realised volatility in Section 2.3.4, the Greeks estimation in real data seems overall in line with the numerical experiments.

For the Greek delta, Figures 2.21 and 2.22 shows that the estimates exhibit tight confidence bounds around the hypothetical Heston delta. For the Greek vega, Figures 2.23 and 2.24 show estimates with wider confidence bounds, which indicates that the estimates are noisier, as we would expect given the difficulties in estimating spot volatility. We also notice that the hypothetical vega – and, to some extent, the hypothetical deltas – seem lower than the linear regression estimates. This could be another artefact of the Feller condition, or maybe a limitation of the Heston model itself.

For the Greeks gamma and theta, Figures 2.25-2.28 show that the estimates are indeed mostly not significant, as we would predict with the small time asymptotics, with possibly the exception of gammas for at-the-money call options near expiry. Indeed, as we get closer to expiry, the Greek gamma explodes – a well-known effect in the context of pin risk –, which could explain the higher Gamma sensitivity. Comparing the Greeks vega and gamma, we highlight that we have found an estimate of spot volatility that is indeed of first-order and has higher explanatory power at small time scales than gamma.

We have checked for robustness when varying the resampling frequency. For reference, Figure 2.29 shows the same linear regression but with samples resampled every 5-seconds rather than the 1-second frequency employed elsewhere. We have noticed that decreasing the sampling frequency degrades the Greeks estimates. Besides, when performing the linear regression on dates with a lower number of active options, we also notice a degradation of the Greeks estimates. This indicates that the linear regression estimates benefit from larger sample sizes.

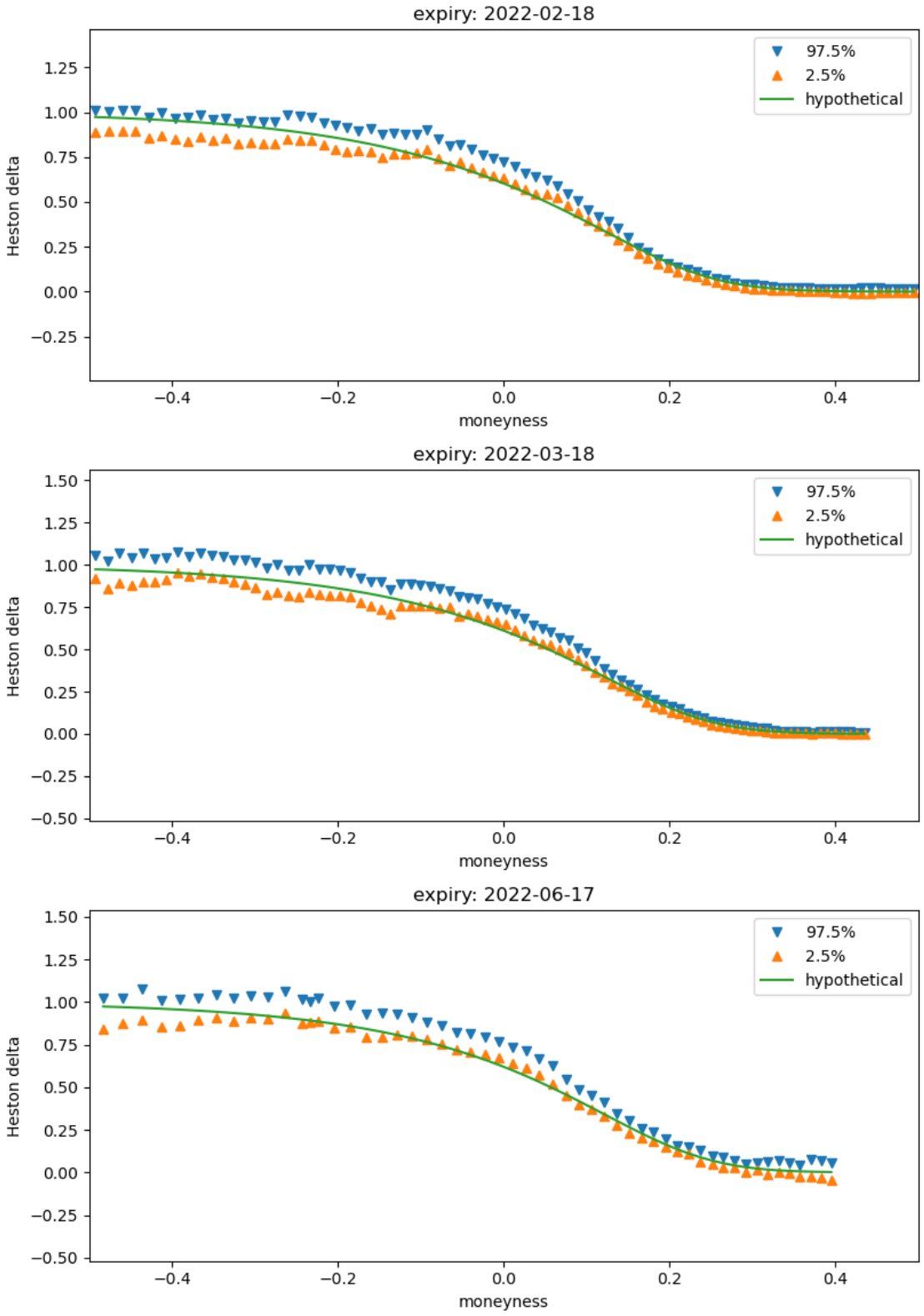


Figure 2.21: Greek delta under the Heston model across call options on 2 December 2021. The triangles represent the 95% confidence interval of the delta estimated by the linear regression. The solid line is the theoretical delta computed using the calibrated Heston model on the average forward price and average volatility. Moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

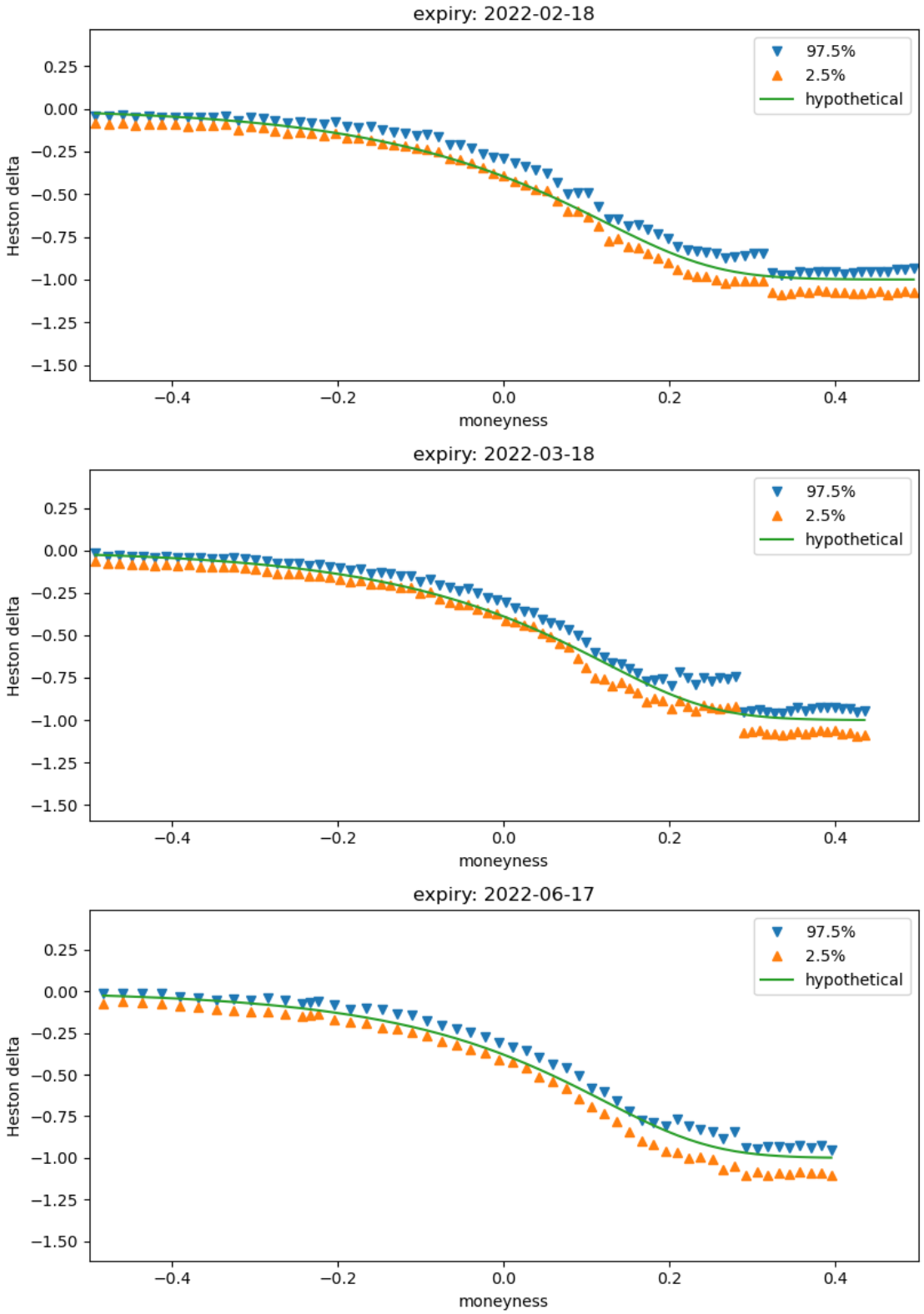


Figure 2.22: Greek delta under the Heston model across put options on 2 December 2021. The triangles represent the 95% confidence interval of the delta estimated by the linear regression. The solid line is the theoretical delta computed using the calibrated Heston model on the average forward price and average volatility. Moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

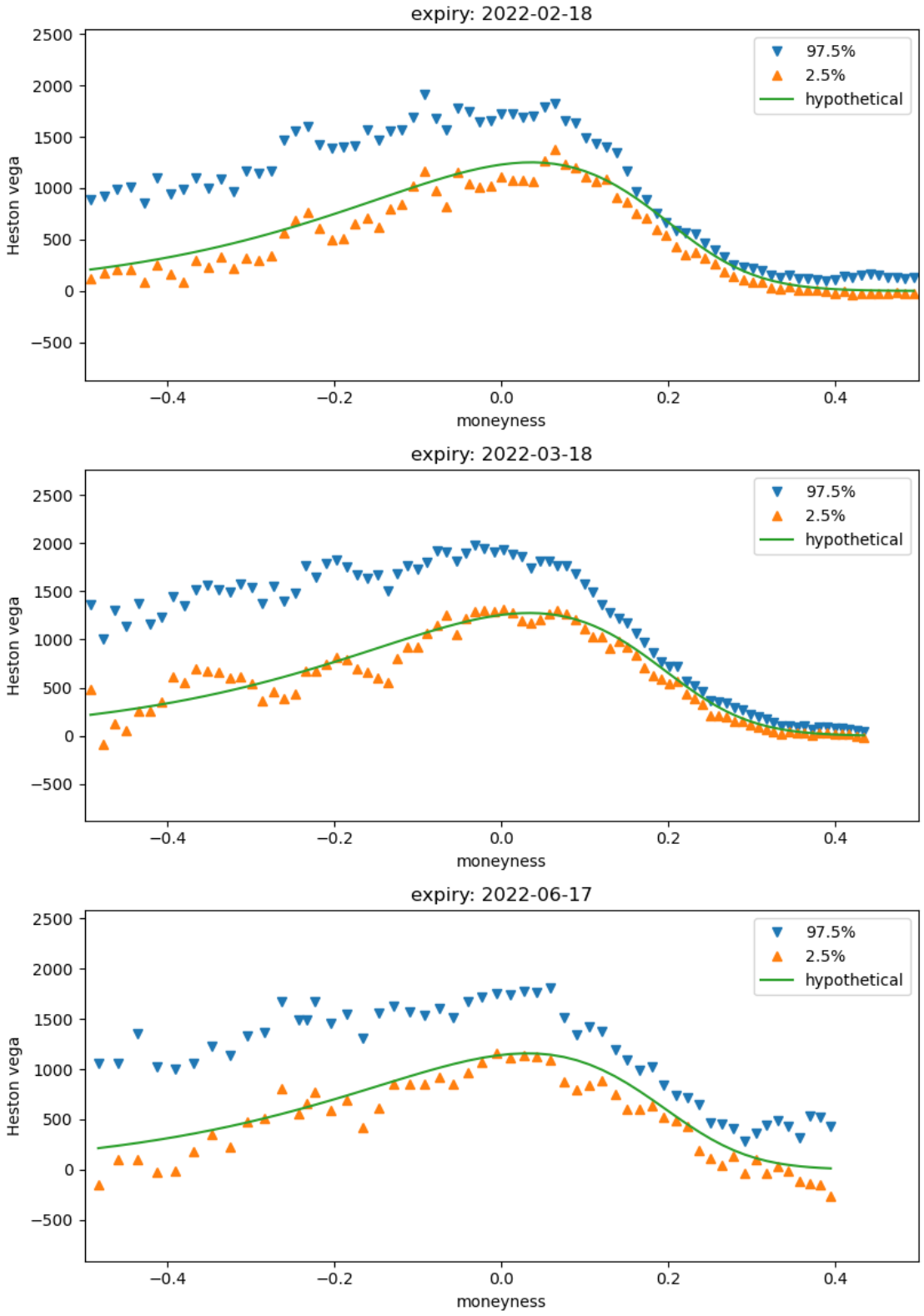


Figure 2.23: Greek vega under the Heston model across call options on 2 December 2021. The triangles represent the 95% confidence interval of the vega estimated by the linear regression. The solid line is the theoretical vega computed using the calibrated Heston model on the average forward price and average volatility. Moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

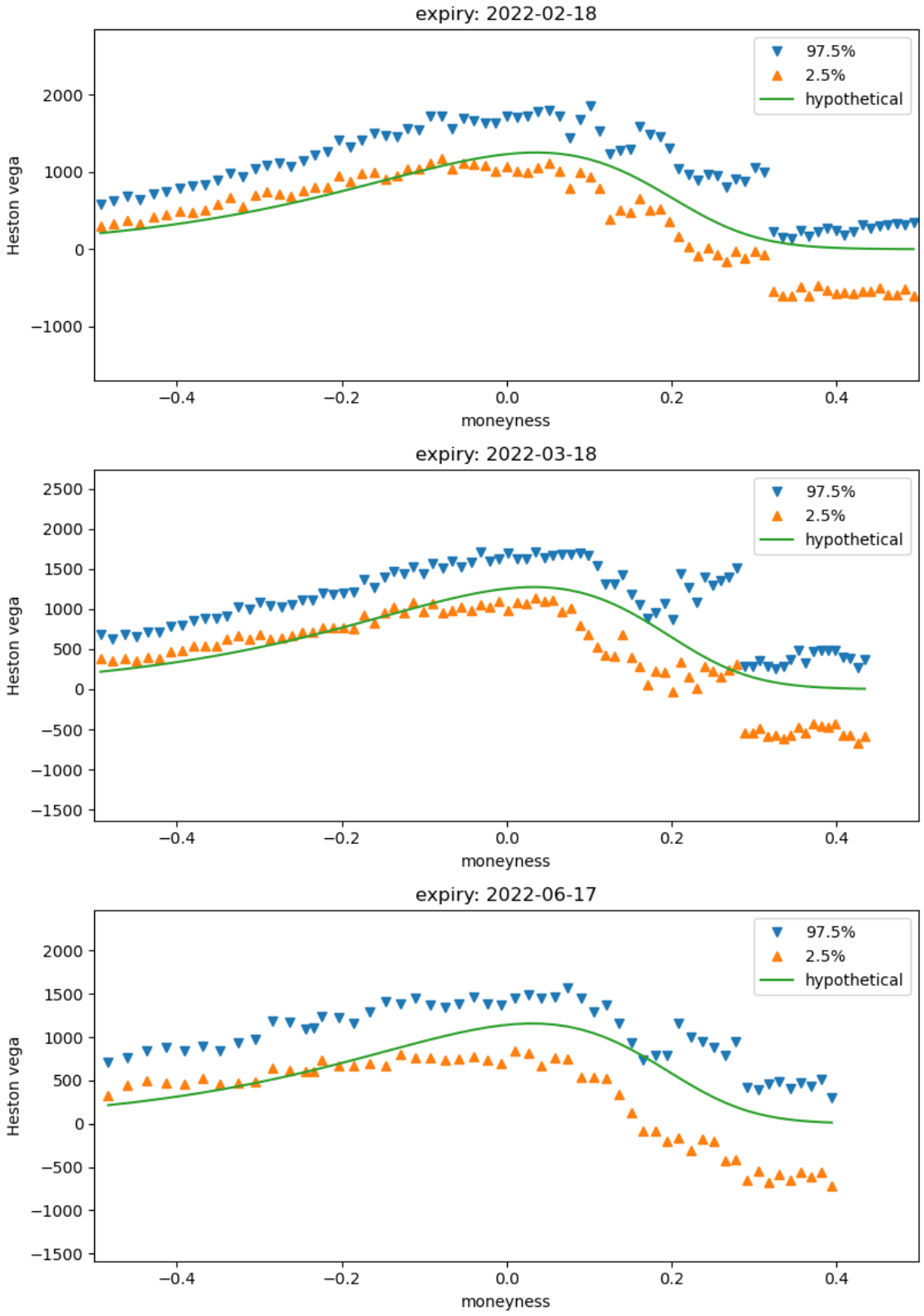


Figure 2.24: Greek vega under the Heston model across put options on 2 December 2021. The triangles represent the 95% confidence interval of the vega estimated by the linear regression. The solid line is the theoretical vega computed using the calibrated Heston model on the average forward price and average volatility. Moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

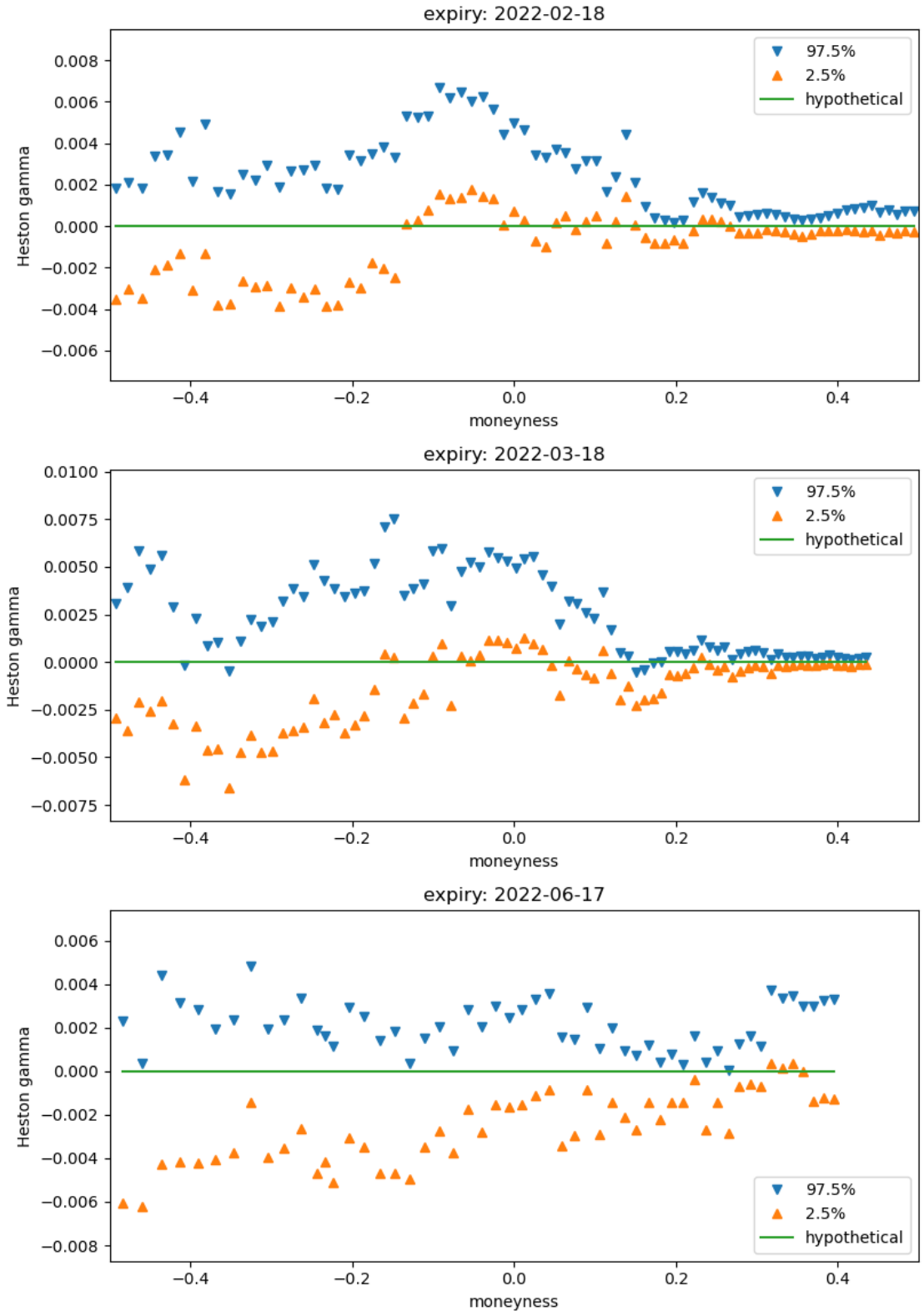


Figure 2.25: Greek gamma under the Heston model across call options on 2 December 2021. The triangles represent the 95% confidence interval of the gamma estimated by the linear regression. The solid line is the theoretical gamma which, under the small time asymptotics is zero. Moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

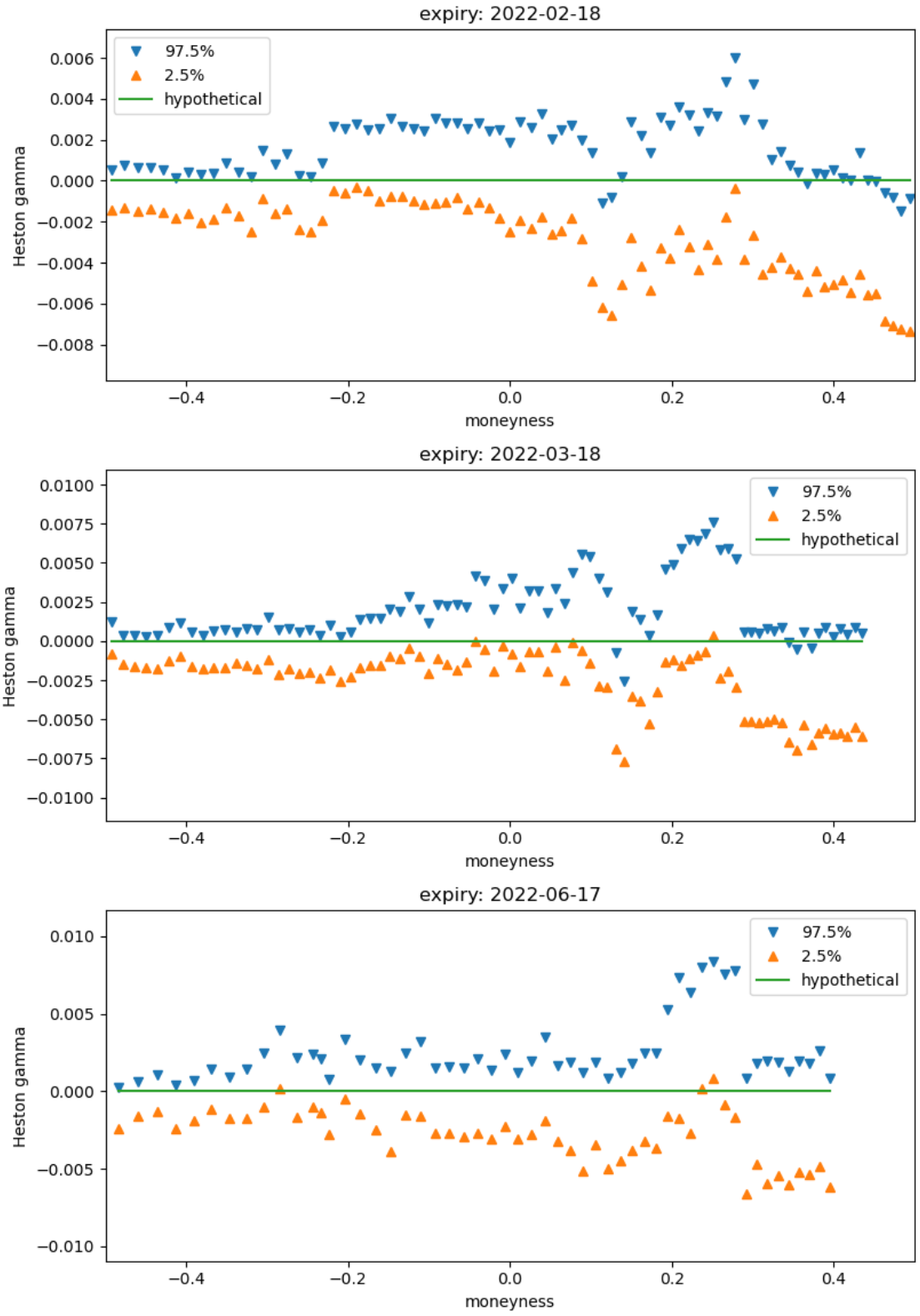


Figure 2.26: Greek gamma under the Heston model across put options on 2 December 2021. The triangles represent the 95% confidence interval of the gamma estimated by the linear regression. The solid line is the theoretical gamma which, under the small time asymptotics is zero. Moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

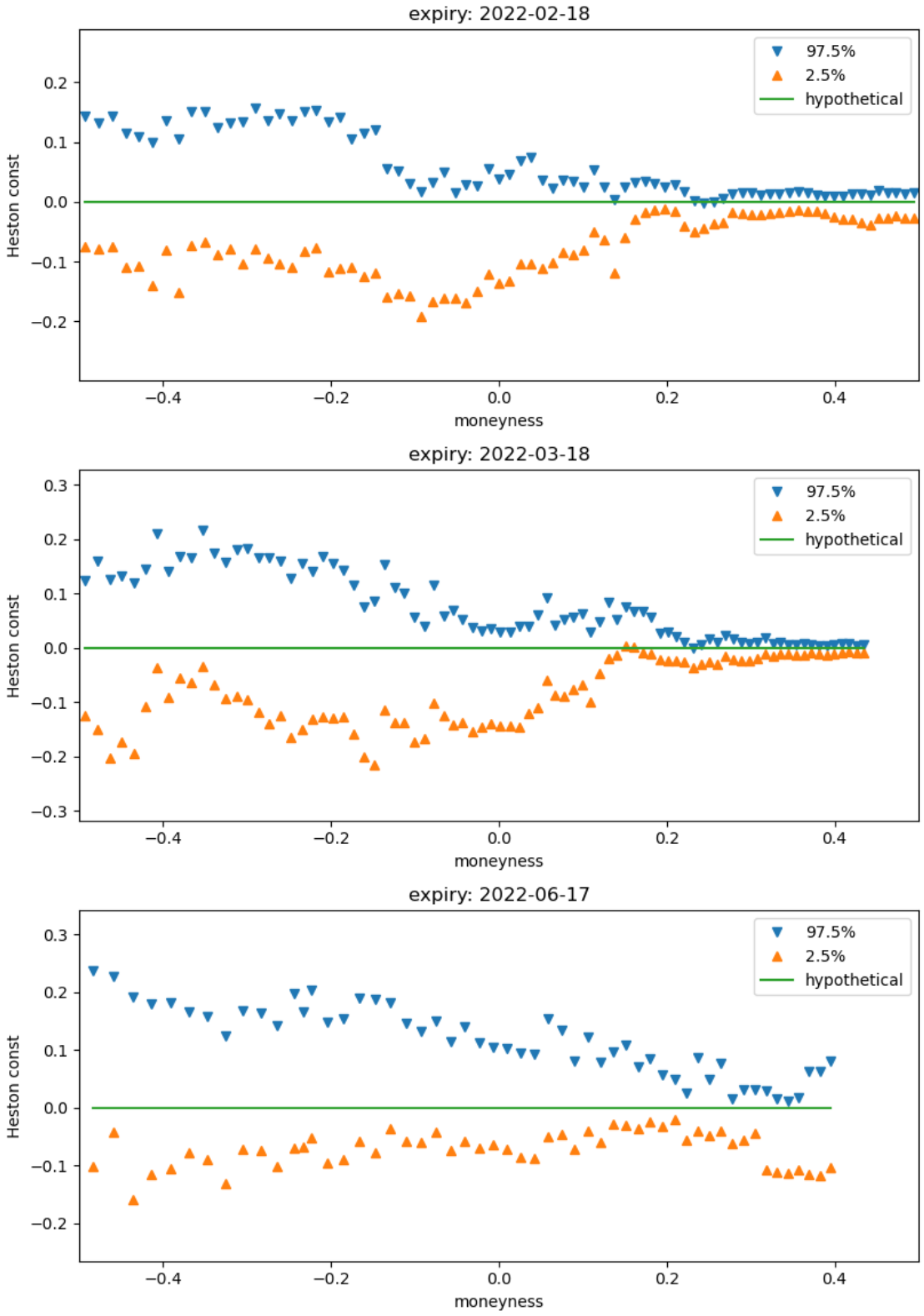


Figure 2.27: Greek theta in seconds under the Heston model across call options on 2 December 2021. The triangles represent the 95% confidence interval of the theta estimated by the linear regression. The solid line is the theoretical gamma which, under the small time asymptotics is zero. Moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

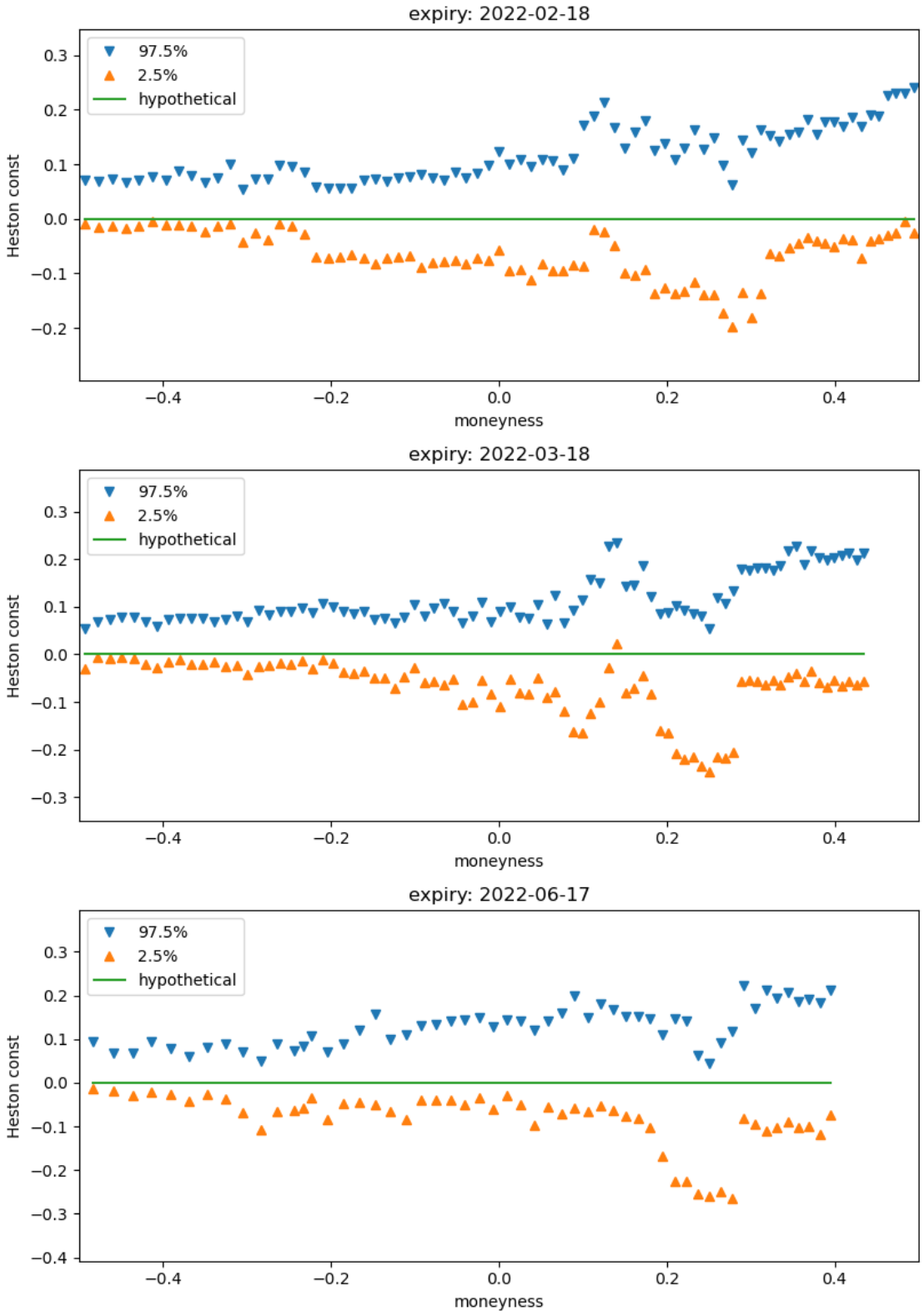


Figure 2.28: Greek theta in seconds under the Heston model across put options on 2 December 2021. The triangles represent the 95% confidence interval of the theta estimated by the linear regression. The solid line is the theoretical gamma which, under the small time asymptotics is zero. Moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

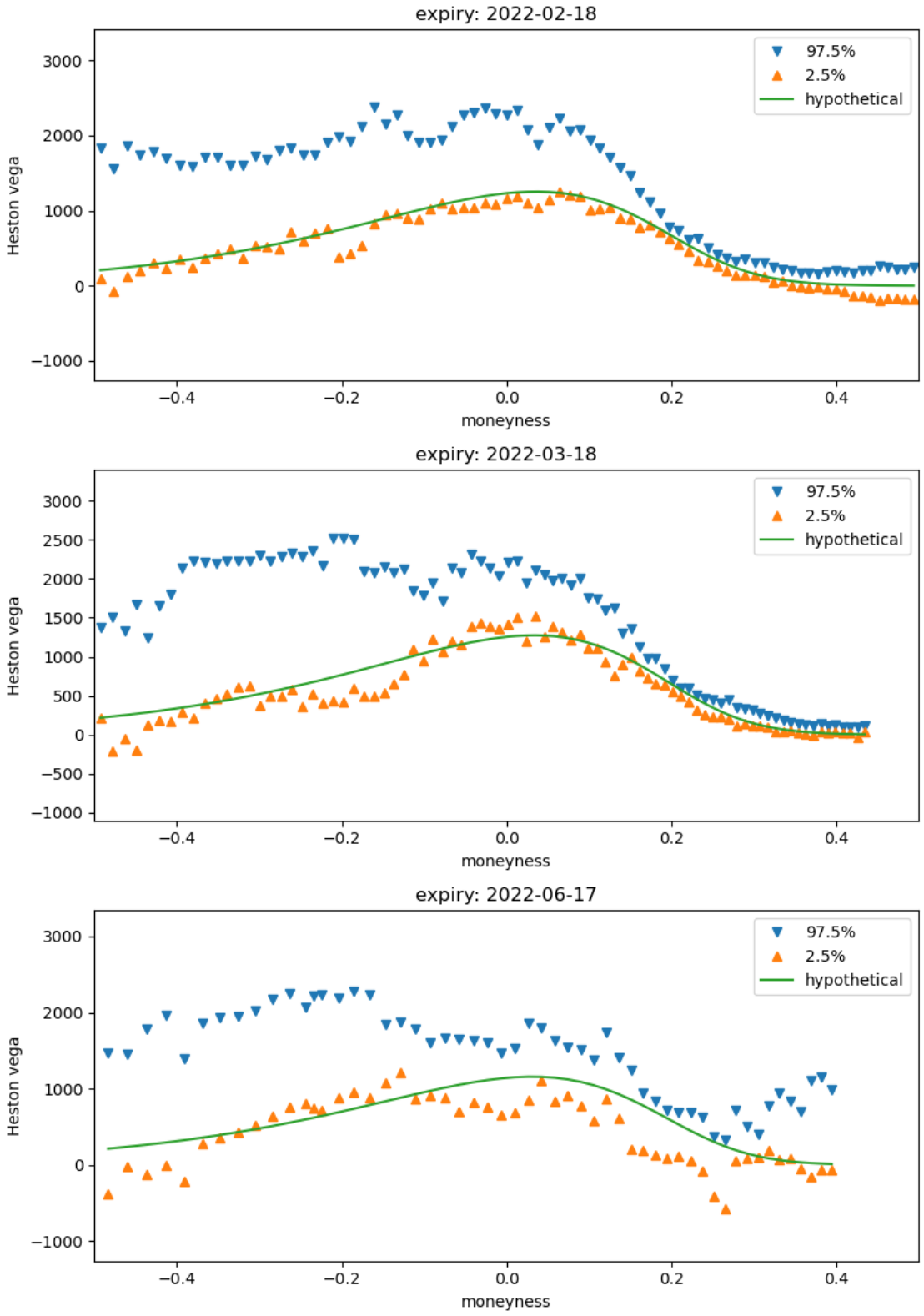


Figure 2.29: Greek vega under the Heston model across call options on 2 December 2021 on 5-second samples rather than 1-second samples. The triangles represent the 95% confidence interval of the vega estimated by the linear regression. The solid line is the theoretical vega computed using the calibrated Heston model on the average forward price and average volatility. Moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

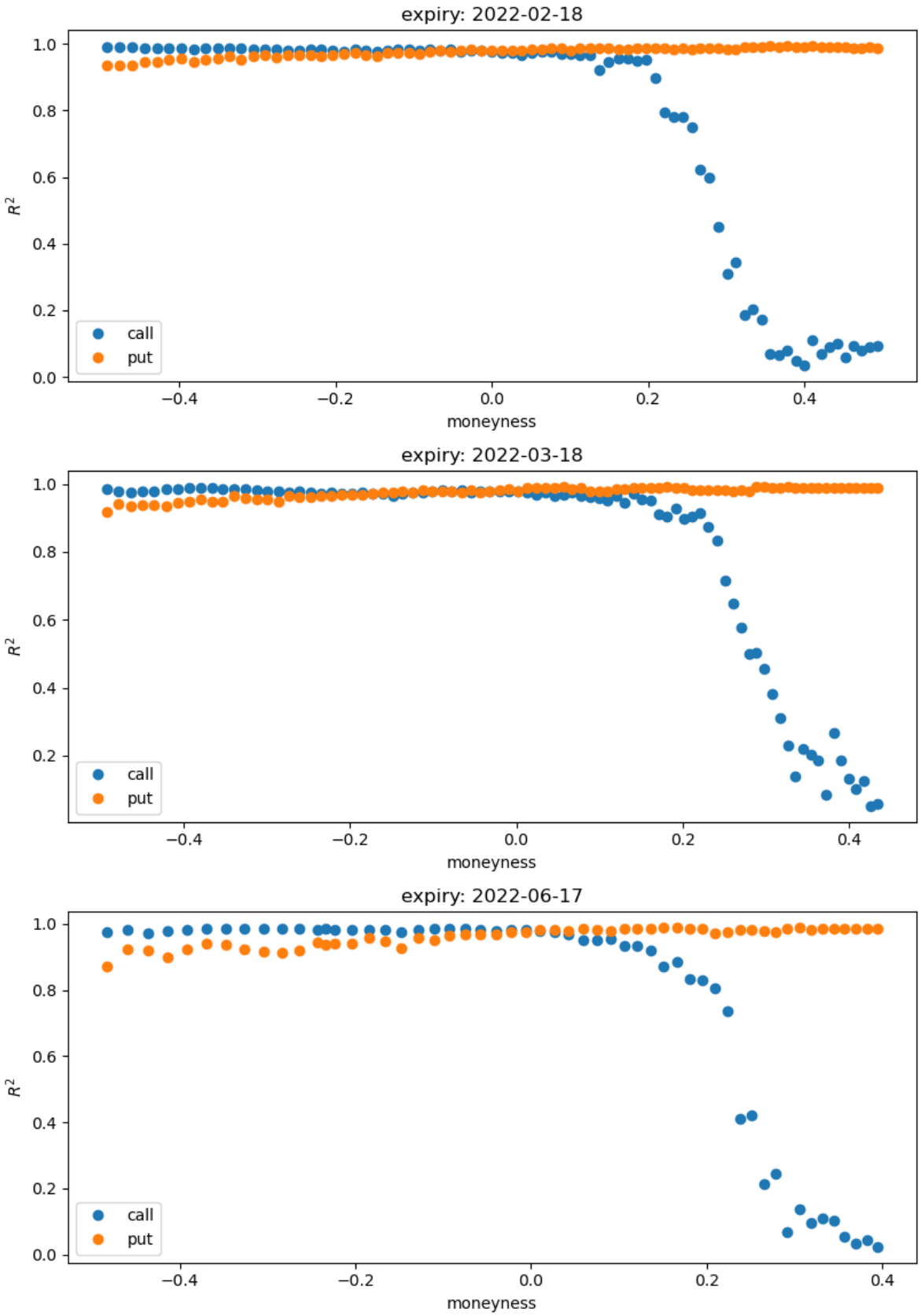


Figure 2.30: The R^2 for the linear regression in (2.17) across put and call options on 2 December 2021. Moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

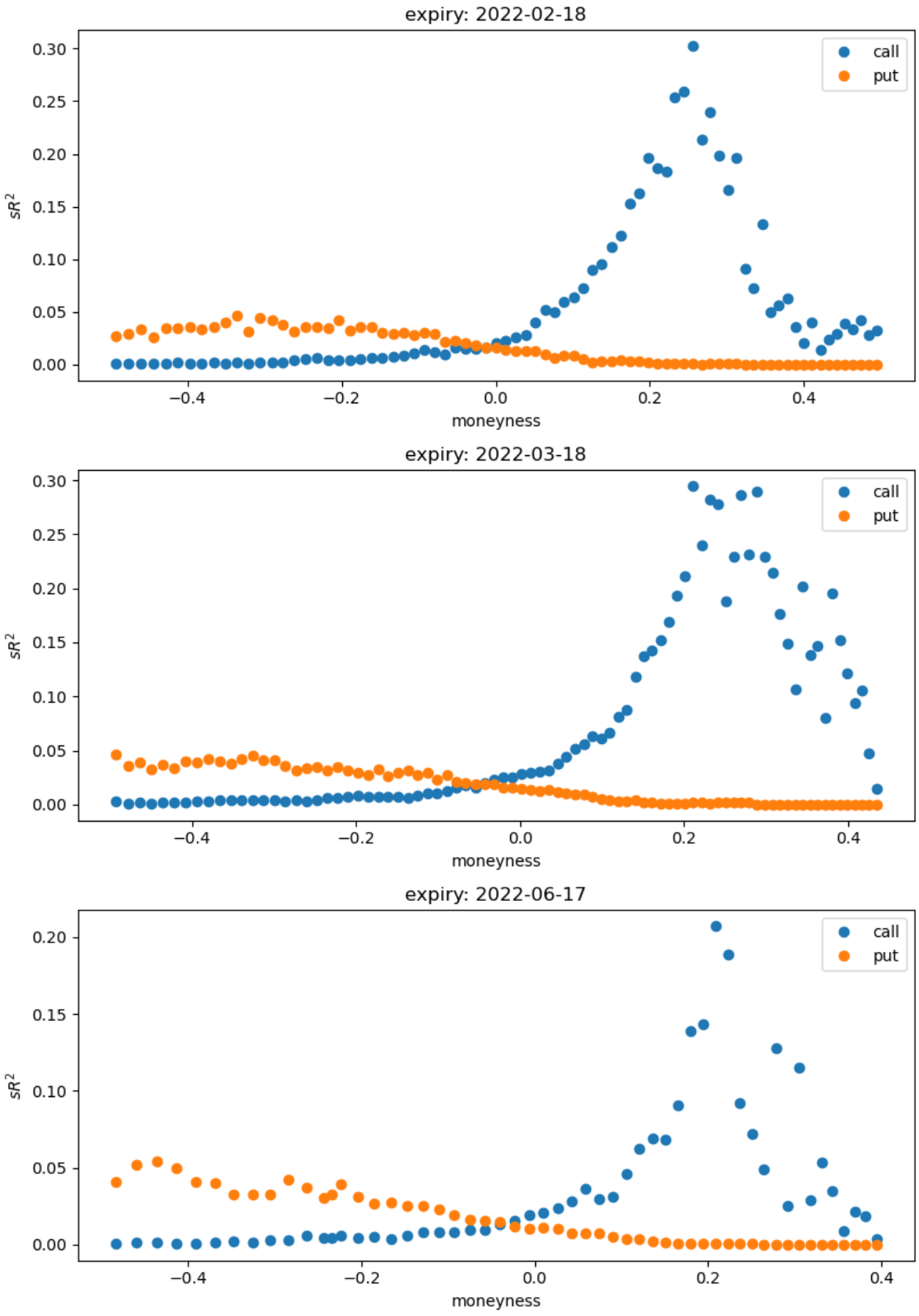


Figure 2.31: The spot volatility semi-partial R^2 for the linear regression in (2.17) across put and call options on 2 December 2021. Moneyness is computed as $\log(K/S_t)/\sqrt{T-t}$.

Finally, the semi-partial R^2 analysis on 2 December 2021 is depicted in Figures 2.30 and 2.31. We observe the same patterns as in the numerical experiment, although the R^2 decays faster for out-of-the-money calls and puts and the sR^2 peaks at most half as high as in the numerical experiment. This could indicate that microstructure noise is more present in real data or that we are missing other factors that drive option price changes. The latter hypothesis is plausible given the empirical findings in Cont et al. (2002) which have found that at least 2 factors (eigenmodes) are required to explain more than 95% of variance of the implied volatility surface of S&P 500, whereas we treat spot volatility as its own factor – following the Heston model.

2.5 Conclusion

We have shown that the measurement error in estimation of spot volatility via historical volatility limits the granularity of its intraday estimates. This motivated us to propose an alternative estimation method for spot volatility using options data. Such method relies on inverting a stochastic volatility model, which enables us to circumvent the measurement error limitation but at the cost of introducing model dependency.

Our spot volatility estimates from the Heston model are indeed biased and its link to the true spot volatility dynamics requires further investigation. Nevertheless, our spot volatility estimates were able to estimate Greeks with 1-second option price changes which were consistent with the Greeks that the calibrated Heston model predicted. Our Greeks estimates reject the hypothesis that the spot volatility has no effect on option prices at a 2.5% significance level for at-the-money calls and puts. We have further quantified the effect of spot volatility using semi-partial R^2 and have identified that up to 30% of the variance of option price changes can be uniquely attributed to changes in spot volatility. This peak was identified for out-of-the-money call options and we have also noticed the asymmetry with respect to out-of-the-money put options, which indicated a modest peak of ca. 5%.

We suggest some directions for further research. The use of a more realistic stochastic volatility model such as the Wishart model could reduce model bias and more accurately estimate spot volatility with our proposed method. Besides, mathematically or empirically assess the model bias of the proposed method and its ability to estimate spot volatility changes with fidelity.

Finally, given our positive result our spot volatility estimates driving option prices, it would be interesting to revisit the approach in Abergel and Zaatour (2012) in which spot volatility is estimated from historical volatility. Given the difficulties mentioned

in Section 2.1.3 and further expanded in Section 2.3.2, special attention is required on measurement error of the estimates so as to choose the appropriate estimation granularity and on microstructure noise so that an appropriate estimator is used and possibly adjusted with respect to intraday patterns, as is done in e.g. Bennedsen et al. (2021).

2.6 Appendix: No-arbitrage bounds for forwards and bonds

In this section, we start by providing a no-arbitrage argument for (2.10) and then apply it to obtain no-arbitrage bounds on forwards and bonds.

2.6.1 No-arbitrage argument for (2.10)

Consider a portfolio with the following positions: long in a call option, short in a put option, short in a forward contract and long in $K - F_{t,T}$ units of a T-maturity zero-coupon bond. At expiry, the value of the portfolio is

$$(S_T - K)_+ - (K - S_T)_+ - (S_T - F_{t,T}) + (F_{t,T} + K) = 0$$

Therefore, if there is no arbitrage, the value of this portfolio at time t must not be positive, i.e.

$$C^{\text{bid}}_{t,T,K} - P^{\text{ask}}_{t,T,K} - B_{t,T}F_{t,T} + B_{t,T}K \leqslant 0. \tag{2.19}$$

If we consider a portfolio with reverse positions, by the same arguments, we obtain that

$$-C^{\text{ask}}_{t,T,K} + P^{\text{bid}}_{t,T,K} + B_{t,T}F_{t,T} - B_{t,T}K \leqslant 0. \tag{2.20}$$

The inequalities (2.19) and (2.20) imply (2.10).

2.6.2 Bounds for forwards and bonds

We analyse no-arbitrage bounds for the forward prices and the interest rate term structure. The no-arbitrage bounds in this section are derived from the put-call parity inequality (2.10). From it, we define the bid and ask quotes for the bond and the discounted forward contract via linear programming.

Definition 1. Given a time t, expiry T, a finite set of strikes $\mathcal{K}^{\text{bid}}$ for which bid European calls and ask European puts with expiry T are available, and a finite set $\mathcal{K}^{\text{ask}}$ for which

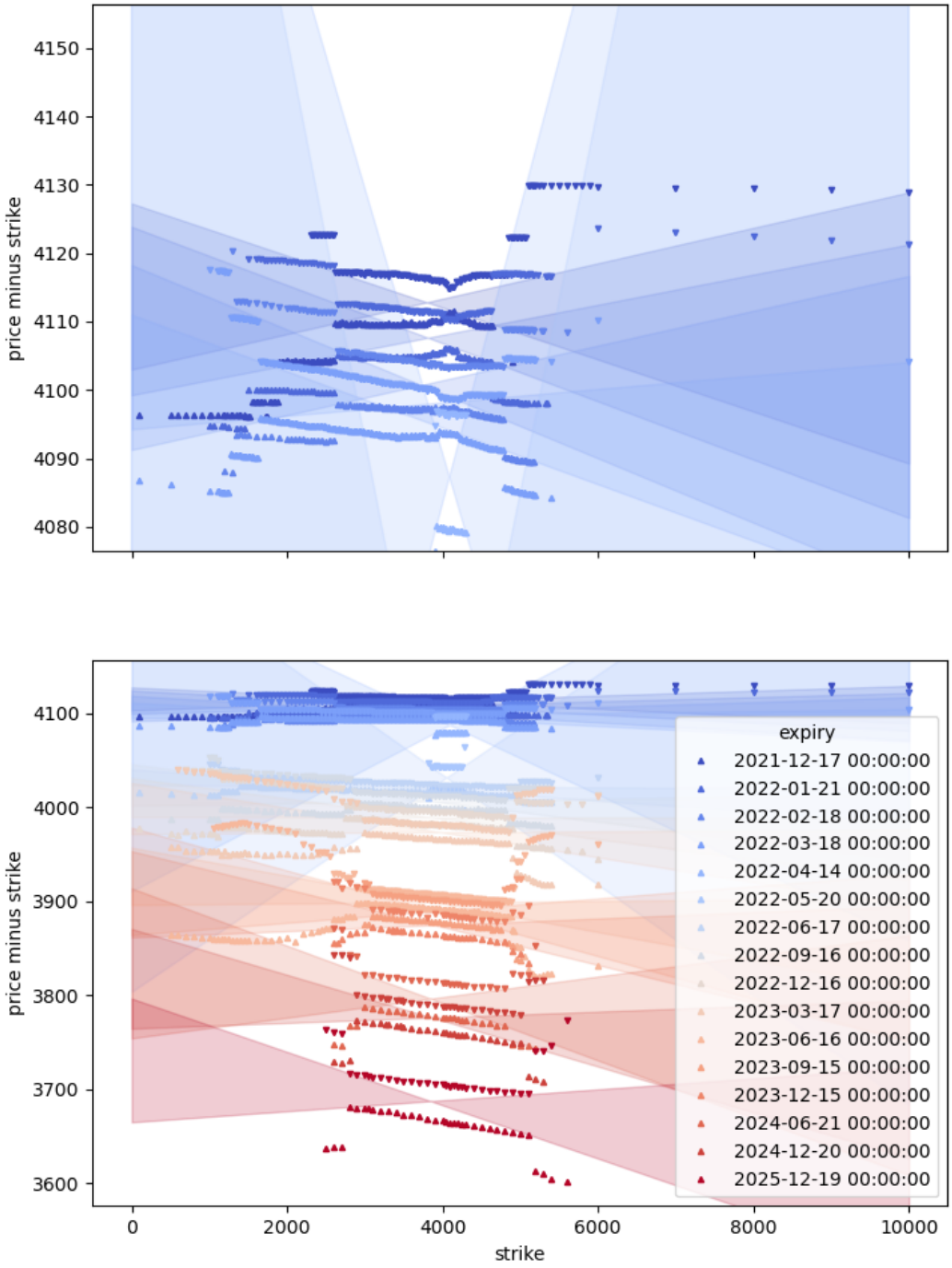
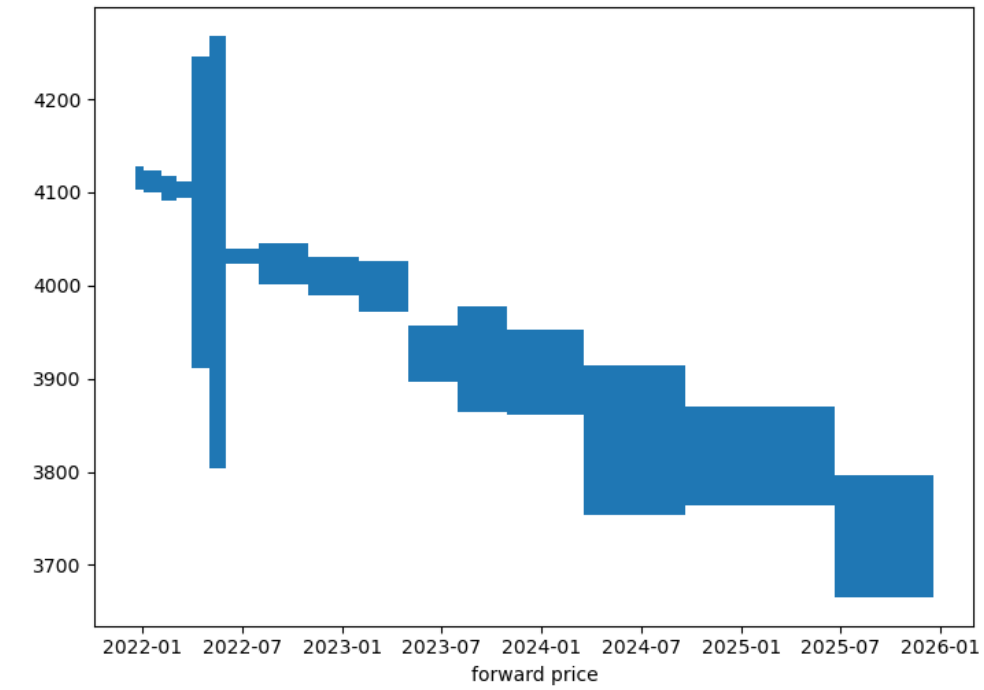


Figure 2.32: Synthetic forward bid and ask for all strikes and expiries plus their corresponding strikes – i.e. $C^{\text{bid}}_{t,T,K} - P^{\text{ask}}_{t,T,K} + K$ and $\leqslant C^{\text{ask}}_{t,T,K} - P^{\text{bid}}_{t,T,K} + K$. The shaded areas are the feasible set for no-arbitrage bonds and discounted forwards – i.e. for $\tilde{F}_{t,T} - B_{t,T}K$ as in Definition 1.



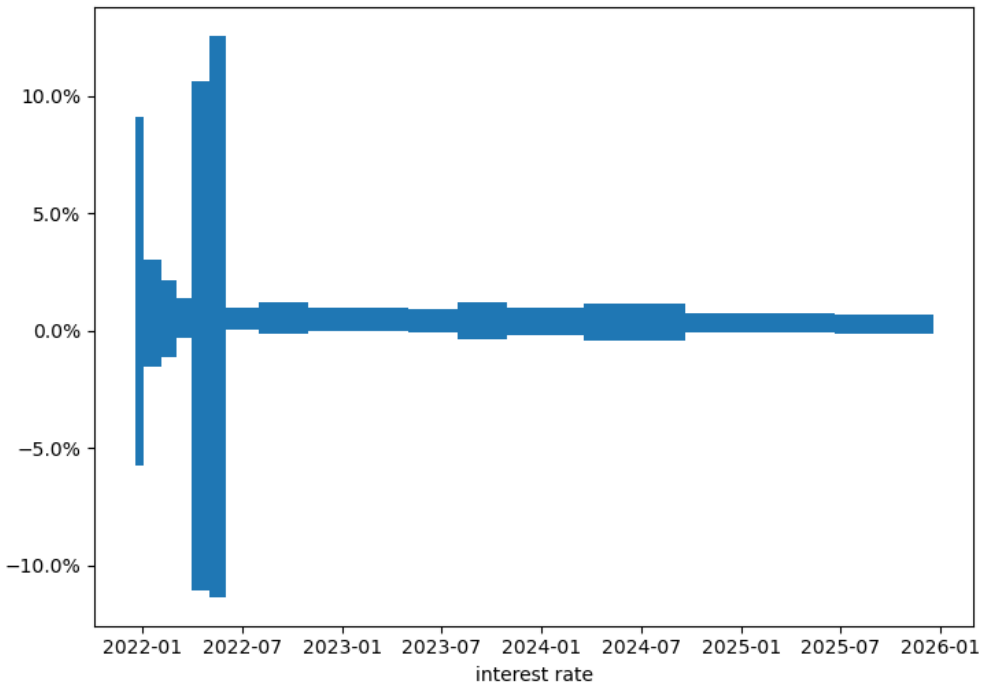


Figure 2.33: No arbitrage interval for the forward prices (top plot) and interest rate term structure (bottom plot).

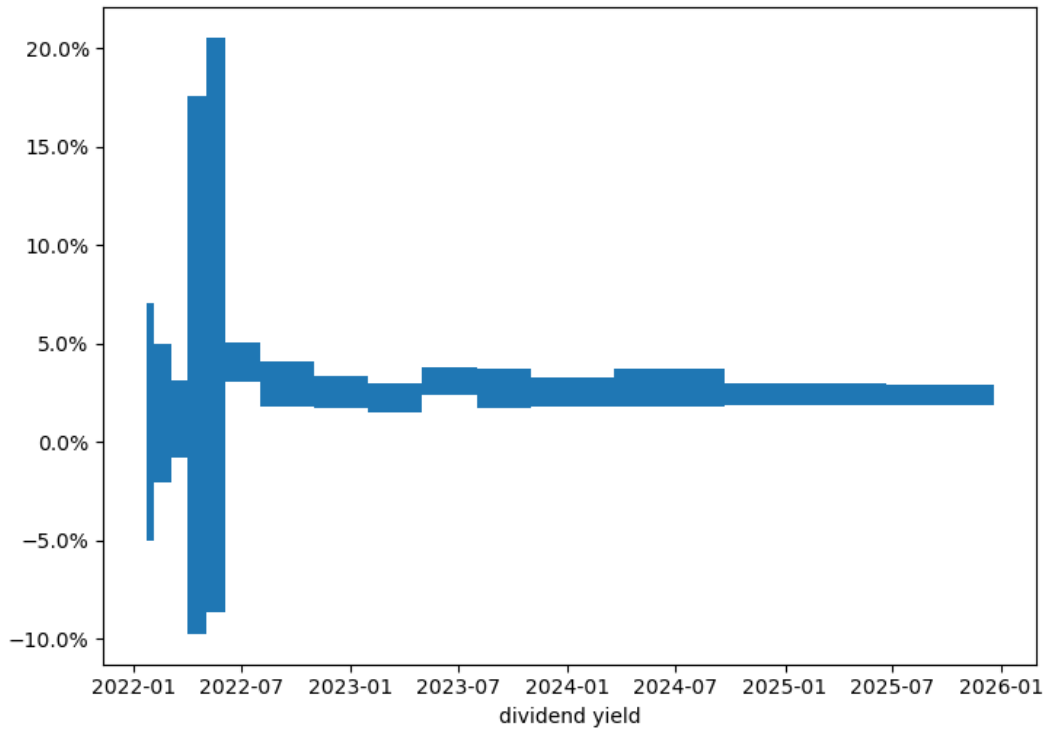


Figure 2.34: No arbitrage intervals for dividend yields of the underlying for the first expiry.

bid European calls and ask European puts with expiry T are available, we define the no-arbitrage bond prices $B^{\text{bid}}_{t,T}$ and $B^{\text{ask}}_{t,T}$ and the no-arbitrage discounted forward price $\tilde{F}^{\text{ask}}_{t,T}$ and $\tilde{F}^{\text{bid}}_{t,T}$ by

$$\begin{aligned}
\tilde{F}^{\text{ask}}_{t,T} &:= \max_{\tilde{F}_{t,T}\in\mathbb{R}, B_{t,T}\in\mathbb{R}^+} \Big\{ \tilde{F}_{t,T} : C^{\text{bid}}_{t,T,K} - P^{\text{ask}}_{t,T,K} \leqslant \tilde{F}_{t,T} - B_{t,T}K,\ \forall K \in \mathcal{K}^{\text{bid}}, \\
&\qquad\qquad \tilde{F}_{t,T} - B_{t,T}K \leqslant C^{\text{ask}}_{t,T,K} - P^{\text{bid}}_{t,T,K},\ \forall K \in \mathcal{K}^{\text{ask}} \Big\}, \\
\tilde{F}^{\text{bid}}_{t,T} &:= \min_{\tilde{F}_{t,T}\in\mathbb{R}, B_{t,T}\in\mathbb{R}^+} \Big\{ \tilde{F}_{t,T} : C^{\text{bid}}_{t,T,K} - P^{\text{ask}}_{t,T,K} \leqslant \tilde{F}_{t,T} - B_{t,T}K,\ \forall K \in \mathcal{K}^{\text{bid}}, \\
&\qquad\qquad \tilde{F}_{t,T} - B_{t,T}K \leqslant C^{\text{ask}}_{t,T,K} - P^{\text{bid}}_{t,T,K},\ \forall K \in \mathcal{K}^{\text{ask}} \Big\}, \\
B^{\text{ask}}_{t,T} &:= \max_{\tilde{F}_{t,T}\in\mathbb{R}, B_{t,T}\in\mathbb{R}^+} \Big\{ B_{t,T} : C^{\text{bid}}_{t,T,K} - P^{\text{ask}}_{t,T,K} \leqslant \tilde{F}_{t,T} - B_{t,T}K,\ \forall K \in \mathcal{K}^{\text{bid}}, \\
&\qquad\qquad \tilde{F}_{t,T} - B_{t,T}K \leqslant C^{\text{ask}}_{t,T,K} - P^{\text{bid}}_{t,T,K},\ \forall K \in \mathcal{K}^{\text{ask}} \Big\}, \\
B^{\text{bid}}_{t,T} &:= \min_{\tilde{F}_{t,T}\in\mathbb{R}, B_{t,T}\in\mathbb{R}^+} \Big\{ B_{t,T} : C^{\text{bid}}_{t,T,K} - P^{\text{ask}}_{t,T,K} \leqslant \tilde{F}_{t,T} - B_{t,T}K,\ \forall K \in \mathcal{K}^{\text{bid}}, \\
&\qquad\qquad \tilde{F}_{t,T} - B_{t,T}K \leqslant C^{\text{ask}}_{t,T,K} - P^{\text{bid}}_{t,T,K},\ \forall K \in \mathcal{K}^{\text{ask}} \Big\}.
\end{aligned}$$

The linear programming method is illustrated in Figure 2.32. For each linear programming problem, the optimal pair $(\tilde{F}_{t,T}, B_{t,T})$ defines a straight line that is constrained by two quotes. The projection of the shaded areas in Figure 2.32 to the y-axis are the no-arbitrage forward price interval and the slope of the lines define the no-arbitrage bond price interval. The actual no-arbitrage forward and bond price intervals are depicted in Figure 2.33.

Additionally, from the forward prices at different expiries, it is also possible to define no-arbitrage dividend yields that are accrued from the first expiry. Such dividend yields are depicted in Figure 2.34.

The no-arbitrage bounds obtained with this methodology has the advantage of being completely model-free and their bounds are defined by static arbitrage trading strategies. As a bootstrapping method, however, it faces some limitations: the no-arbitrage gaps are too wide and most bounds are each defined by four away-from-the-money option prices only – the information on the most liquid options is disregarded.

2.7 Appendix: Proofs

2.7.1 Proof of Proposition 7

Given that $(X_t)_{t\in[0,T]}$ has a SDE form (2.3) under $\mathbb{Q}$ it is, in particular, a Markov process under $\mathbb{Q}$, hence

$$C_t = \mathbb{E}^{\mathbb{Q}}\left[f(X_T)|\mathcal{F}_t\right] = \mathbb{E}^{\mathbb{Q}}\left[f(X_T)|X_t\right] \quad \forall t \in [0,T].$$

Consequently, for each $t \in [0,T]$, there exists a function $x \mapsto \varphi(t,x)$ such that (2.6) is true.

Now, we would like to show the small time asymptotics (2.7) for the state process. For each $i \in \{1,\dots,d\}$, triangle inequality yields

$$\begin{aligned}\frac{1}{\sqrt{t}}\left\|X_t^i - \tilde{X}_t^i\right\|_{L^2(\mathbb{P})} &= \frac{1}{\sqrt{t}}\left\|\int_0^t \mu_s^i ds + \int_0^t \left(\sigma_s^i - \sigma_0^i\right)\cdot dW_s\right\|_{L^2(\mathbb{P})} \\ &\leqslant \frac{1}{\sqrt{t}}\left\|\int_0^t \mu_s^i ds\right\|_{L^2(\mathbb{P})} + \frac{1}{\sqrt{t}}\left\|\int_0^t \left(\sigma_s^i - \sigma_0^i\right)\cdot dW_s\right\|_{L^2(\mathbb{P})},\end{aligned}$$

where X_t^i and μ_t^i denote the i-th element of the vector X_t and μ_t, respectively, and σ_t^i denotes the i-th row of the matrix σ_t. Therefore, we arrive at (2.7) if both

$$\frac{1}{\sqrt{t}}\left\|\int_0^t \mu_s^i ds\right\|_{L^2(\mathbb{P})} \xrightarrow{t\to 0} 0, \text{ and} \tag{2.21}$$

$$\frac{1}{\sqrt{t}}\left\|\int_0^t \left(\sigma_s^i - \sigma_0^i\right)\cdot dW_s\right\|_{L^2(\mathbb{P})} \xrightarrow{t\to 0} 0. \tag{2.22}$$

By the mean value theorem, for each $t \in [0,T]$, there exists $r \in [0,t]$ such that

$$\mu_r^i = \frac{1}{t}\int_0^t \mu_s^i ds,$$

which, $(\mu_t)_{t\in[0,T]}$ being continuous, implies

$$\lim_{t\to 0}\frac{1}{t}\int_0^t \mu_s^i ds = \mu_0^i. \tag{2.23}$$

Jensen's inequality yields, for each $t \in [0,T]$,

$$\left(\frac{1}{t}\int_0^t \mu_s^i ds\right)^2 \leqslant \frac{1}{t}\int_0^t \left(\mu_s^i\right)^2 ds,$$

hence,

$$\left(\frac{1}{\sqrt{t}}\int_0^t \mu_s^i ds\right)^2 = t\left(\frac{1}{t}\int_0^t \mu_s^i ds\right)^2 \leqslant \int_0^t (\mu_s^i)^2 ds \leqslant \int_0^T (\mu_s^i)^2 ds,$$

which is $\mathbb{P}$-integrable by hypothesis. Thus, we can apply dominated convergence theorem and (2.23) to obtain

$$\begin{aligned}
\lim_{t\to 0}\frac{1}{t}\mathbb{E}\left[\left(\int_0^t \mu_s^i ds\right)^2\right] &= \lim_{t\to 0}\mathbb{E}\left[\left(\frac{1}{\sqrt{t}}\int_0^t \mu_s^i ds\right)^2\right] \\
&= \mathbb{E}\left[\lim_{t\to 0}\left(\frac{1}{\sqrt{t}}\int_0^t \mu_s^i ds\right)^2\right] \\
&= \mathbb{E}\left[\left(\lim_{t\to 0} t\right)\left(\lim_{t\to 0}\frac{1}{t}\int_0^t \mu_s^i ds\right)^2\right] \\
&= 0,
\end{aligned}$$

which implies (2.21).

By Itô isometry, we obtain that

$$\mathbb{E}\left[\left(\int_0^t (\sigma_s^i - \sigma_0^i)\cdot dW_s\right)^2\right] = \mathbb{E}\left[\int_0^t (\sigma_s^i - \sigma_0^i)\cdot(\sigma_s^i - \sigma_0^i) ds\right], \quad \forall t \geqslant 0. \tag{2.24}$$

Additionally,

$$\begin{aligned}
\int_0^t (\sigma_s^i - \sigma_0^i)\cdot(\sigma_s^i - \sigma_0^i) ds &= \sigma_0^i\cdot\sigma_0^i t - 2\sigma_0^i\cdot\int_0^t \sigma_s^i ds + \int_0^t \sigma_s^i\cdot\sigma_s^i ds \\
&\leqslant \left|\sigma_0^i\cdot\sigma_0^i\right| T + 2\left\|\sigma_0^i\right\|_1 \int_0^T \left\|\sigma_s^i\right\|_1 ds + \int_0^T \left\|\sigma_s^i\right\|^2 ds, \quad \forall t \geqslant 0.
\end{aligned}$$

Given that $(\sigma_t)_{t\in[0,T]}$ is square-integrable and that $L^2(\mathbb{P}) \subset L^1(\mathbb{P})$, we have that the right-hand side of the inequality is $\mathbb{P}$-integrable. Hence, we can use Fubini theorem to obtain

$$\mathbb{E}\left[\int_0^t (\sigma_s^i - \sigma_0^i)\cdot(\sigma_s^i - \sigma_0^i) ds\right] = \int_0^t \mathbb{E}\left[(\sigma_s^i - \sigma_0^i)\cdot(\sigma_s^i - \sigma_0^i)\right] ds, \quad \forall t \geqslant 0. \tag{2.25}$$

Using the same argument with the mean-value theorem and the assumption that $(\sigma_t)_{t\in[0,T]}$ is continuous, we have that

$$\lim_{t\to 0}\frac{1}{t}\int_0^t \mathbb{E}\left[(\sigma_s^i - \sigma_0^i)\cdot(\sigma_s^i - \sigma_0^i)\right] ds = (\sigma_0^i - \sigma_0^i)\cdot(\sigma_0^i - \sigma_0^i) = 0.$$

This limit combined with (2.25) and (2.24) imply (2.22).

It remains to show the small time asymptotics (2.8) of the option price process. If φ is of class $\mathcal{C}^{1,2}([0,T)\times D)$, then we can apply Itô's formula to obtain

$$d\varphi(t,X_t)=\Theta_t dt+\nabla_x\varphi(t,X_t)\cdot\sigma^{\mathbb{Q}}(t,X_t)dW_t^{\mathbb{Q}}\quad\forall t\in[0,T),$$

where

$$\begin{aligned}\Theta_t=&\partial_t\varphi(t,X_t)+\nabla_x\varphi(t,X_t)\cdot\mu^{\mathbb{Q}}(t,X_t)\\&+\frac{1}{2}\mathrm{Tr}\left(\sigma^{\mathbb{Q}}(t,X_t)^\top\Delta_x\varphi(t,X_t)\sigma^{\mathbb{Q}}(t,X_t)\right),\quad\forall t\in[0,T),\end{aligned}$$

and Δ_x denotes the Laplace operator with respect to x.

By construction, $(C_t)_{t\in[0,T]}$ is a $\mathbb{Q}$-martingale, which implies $\Theta\equiv 0$ and, therefore

$$dC_t=\nabla_x\varphi(t,X_t)\cdot\sigma^{\mathbb{Q}}(t,X_t)dW_t^{\mathbb{Q}},\quad\forall t\in[0,T).$$

From (2.4), we rewrite the dynamics above as

$$dC_t=\nabla_x\varphi(t,X_t)\cdot\sigma^{\mathbb{Q}}(t,X_t)\theta_t dt+\nabla_x\varphi(t,X_t)\cdot\sigma^{\mathbb{Q}}(t,X_t)dW_t,\quad\forall t\in[0,T).\tag{2.26}$$

Let $K\subset D$ be a non-empty compact set containing x_0, which exists due to the assumption that D has non-empty interior containing x_0. Define the stopping time τ as the first exit time of $(X_t)_{t\in[0,T]}$ on K, i.e.

$$\tau=\inf\left\{t\in[0,T]:\ X_t\notin K\right\},$$

where we take the convention that $\inf\emptyset=\infty$.

Define

$$\begin{aligned}C_t'\coloneqq C_0&+\int_0^t\nabla_x\varphi\left(s\wedge\tau,X_{s\wedge\tau}\right)\cdot\sigma^{\mathbb{Q}}\left(s\wedge\tau,X_{s\wedge\tau}\right)\theta_{s\wedge\tau}ds\\&+\int_0^t\nabla_x\varphi\left(s\wedge\tau,X_{s\wedge\tau}\right)\cdot\sigma^{\mathbb{Q}}\left(s\wedge\tau,X_{s\wedge\tau}\right)dW_s,\quad t\in[0,T/2].\end{aligned}$$

In order to prove (2.8), it is sufficient, by triangle inequality, to show that both

$$\frac{1}{\sqrt{t}} \left\| C_t - C'_t \right\|_{L^2(\mathbb{P})} \xrightarrow{t \to 0} 0, \text{ and} \tag{2.27}$$

$$\frac{1}{\sqrt{t}} \left\| C'_t - \tilde{C}_t \right\|_{L^2(\mathbb{P})} \xrightarrow{t \to 0} 0. \tag{2.28}$$

We start with (2.28). By hypothesis, φ is continuously differentiable in $[0,T) \times D$, in particular $\nabla_x \varphi$ is continuous on the compact set $K \times [0, T/2]$ and thus uniformly bounded on $K \times [0, T/2]$ by, say, $M > 0$.

Furthermore, we have that quadratic variation is preserved by change of measure, i.e.

$$\operatorname{Tr}\left(\sigma^{\mathbb{Q}}(t, X_t)^\top \sigma^{\mathbb{Q}}(t, X_t)\right) dt = d[X,X]_t = \operatorname{Tr}\left(\sigma_t^\top \sigma_t\right) dt, \quad \forall t \in [0,T].$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\int_0^{T/2} \left\|\sigma^{\mathbb{Q}}(t \wedge \tau, X_{t\wedge\tau})\, \varphi_x(t \wedge \tau, X_{t\wedge\tau})\right\|^2 dt\right] \\
&\leqslant M^2 \mathbb{E}\left[\int_0^{T/2} \left\|\sigma^{\mathbb{Q}}(t \wedge \tau, X_{t\wedge\tau})\right\|_F^2 dt\right] \\
&= M^2 \mathbb{E}\left[\int_0^{T/2} \left\|\sigma_t\right\|_F^2 dt\right] < \infty,
\end{aligned}$$

and, by Cauchy-Schwarz inequality on the $L^2(\mathbb{P})$ norm,

$$\begin{aligned}
&\mathbb{E}\left[\int_0^{T/2} \left(\varphi_x(t \wedge \tau, X_{t\wedge\tau}) \cdot \sigma^{\mathbb{Q}}(t \wedge \tau, X_{t\wedge\tau})\, \theta_{t\wedge\tau}\right)^2 dt\right] \\
&\leqslant M^2 \mathbb{E}\left[\int_0^{T/2} \left\|\sigma^{\mathbb{Q}}(t \wedge \tau, X_{t\wedge\tau})\, \theta_{t\wedge\tau}\right\|^2 dt\right] \\
&\leqslant M^2 \mathbb{E}\left[\int_0^{T/2} \left\|\sigma^{\mathbb{Q}}(t \wedge \tau, X_{t\wedge\tau})\right\|_F^2 dt\right] \mathbb{E}\left[\int_0^{T/2} \left\|\theta_{t\wedge\tau}\right\|^2 dt\right] < \infty,
\end{aligned}$$

which implies that $(C'_t)_{t\in[0,T/2]}$ fulfills the square-integrability condition (2.5) for its drift and diffusion coefficients. This enables us to use the same arguments as in $(X_t)_{t\in[0,T]}$ to obtain the small time asymptotics for $(C'_t)_{t\in[0,T/2]}$, which implies (2.28).

Because we have assumed f is bounded, we have that C_t is bounded almost surely for all $t \in [0, T/2]$ and thus $\|C_t\|_{L^2(\mathbb{P})}$ is uniformly bounded on $[0, T/2]$. We have just showed that $(C'_t)_{t\in[0,T/2]}$ satisfies the square-integrability conditions, which implies, by Jensen

inequality, that $\|C'_t\|_{L^2(\mathbb{P})}$ is also uniformly bounded on $[0, T/2]$. Therefore $\|C_t - C'_t\|_{L^2(\mathbb{P})}$ is uniformly bounded on $[0, T/2]$.

On the other hand, since $(X_t)_{t\in[0,T]}$ is continuous and $X_0 = x_0$, then $\tau > 0$ a.s. This implies that

$$\frac{C'_t - C_t}{\sqrt{t}} \xrightarrow{t\to 0} 0 \text{ a.s.} \tag{2.29}$$

We have already shown that $\|C_t - C'_t\|_{L^2(\mathbb{P})}$ is uniformly bounded on the interval $[0, T/2]$, thus the dominated convergence theorem lets us go from (2.29) to (2.27).

2.7.2 Proof of Proposition 8

Note that, for each $i \in \{1, \dots, N\}$,

$$\underline{y}_i = \frac{\underline{y}_i + \bar{y}_i}{2} - \frac{\bar{y}_i - \underline{y}_i}{2}, \quad \bar{y}_i = \frac{\underline{y}_i + \bar{y}_i}{2} + \frac{\bar{y}_i - \underline{y}_i}{2}.$$

Replacing this in (2.15), yields

$$\begin{aligned}
F &= \left\{\beta \in \mathbb{R}^d : \quad \frac{\underline{y}_i + \bar{y}_i}{2} - \frac{\bar{y}_i - \underline{y}_i}{2} < \beta^\top x_i < \frac{\underline{y}_i + \bar{y}_i}{2} + \frac{\bar{y}_i - \underline{y}_i}{2}, \quad \forall i \in \{1, \dots, N\}\right\} \\
&= \left\{\beta \in \mathbb{R}^d : \quad -\frac{\bar{y}_i - \underline{y}_i}{2} < \beta^\top x_i - \frac{\underline{y}_i + \bar{y}_i}{2} < \frac{\bar{y}_i - \underline{y}_i}{2}, \quad \forall i \in \{1, \dots, N\}\right\} \\
&= \left\{\beta \in \mathbb{R}^d : \quad \left|\tilde{y}_i - \beta^\top \tilde{x}_i\right| < 1, \quad \forall i \in \{1, \dots, N\}\right\} \\
&= \left\{\beta \in \mathbb{R}^d : \quad \left\|\left(\tilde{y}_1 - \beta^\top \tilde{x}_1, \dots, \tilde{y}_N - \beta^\top \tilde{x}_N\right)\right\|_\infty < 1\right\}
\end{aligned}$$

Since F is non-empty, take an element $\beta_0 \in F$ and let β^*_∞ be the solution to (2.16) with $p = \infty$. Then,

$$\left\|\left(\tilde{y}_1 - \beta_*^\top \tilde{x}_1, \dots, \tilde{y}_N - \beta_*^\top \tilde{x}_N\right)\right\|_\infty \leqslant \left\|\left(\tilde{y}_1 - \beta_0^\top \tilde{x}_1, \dots, \tilde{y}_N - \beta_0^\top \tilde{x}_N\right)\right\|_\infty < 1,$$

which shows $\beta^*_\infty \in F$. Define the functions

$$\ell(\beta, p) = \left\|\left(\tilde{y}_1 - \beta^\top \tilde{x}_1, \dots, \tilde{y}_N - \beta^\top \tilde{x}_N\right)\right\|_p, \quad \ell^*(p) = \inf_{\beta\in\mathbb{R}^d} \ell(\beta, p).$$

We have that $\ell(\beta, p)^p$ is convex for all $\beta \in \mathbb{R}^d$ and $p \in (1, \infty)$. Therefore, $\ell^*(p)^p$ is also convex for all $p \in (1, \infty)$, which implies it is continuous for all $p \in (1, \infty)$. Since $\ell(\beta, p)$ is continuous for all $\beta \in \mathbb{R}^d$ and $p \in (1, \infty]$, then

$$\lim_{p\to\infty} \ell^*(p) = \lim_{p\to\infty} \inf_{\beta\in\mathbb{R}^d} \ell(\beta, p) = \inf_{\beta\in\mathbb{R}^d} \ell(\beta, \infty) = \ell^*(\infty),$$

which shows $\ell^*(p)$ is also continuous at $p = \infty$.

Consider the open interval $O = (\ell^*(\infty), 1)$. Since $\beta^*_\infty \in F$, then $O \subset F$. Furthermore, because $\ell^*(p)$ is continuous for all $p \in (1, \infty]$, then $(\ell^*)^{-1}(O)$ is an open interval (p_0, ∞) for some $p_0 \in [1, \infty)$, which concludes the proof.

2.7.3 Lemma for Proposition 9

Lemma 1. *Consider the SDE*

$$\begin{aligned} dS_t &= \sigma_t dW_t, \quad S_0 = 0, \\ d\sigma_t &= \nu dZ_t, \quad \sigma_0 \in \mathbb{R}, \end{aligned}$$

where $(W_t, Z_t)_{t \geqslant 0}$ is a vector of independent Brownian motions. Then,

$$\mathbb{E}\left[S_t^2\right] = \mathbb{E}\left[\int_0^t \sigma_s^2 ds\right] = \sigma_0^2 t + \frac{1}{2}\nu^2 t^2, \quad \mathrm{Var}(S_t^2) = 2\sigma_0^4 t^2 + 6\nu^2\sigma_0^2 t^3 + \frac{3}{2}\nu^4 t^4.$$

Proof. We start by computing the following.

$$\begin{aligned} \mathbb{E}\left[\sigma_t^2\right] &= \sigma_0^2 + \nu^2 t, \\ \mathbb{E}\left[\sigma_t^4\right] &= \mathbb{E}\left[(\sigma_0 + \nu\sqrt{t}W_1)^4\right] \\ &= \sigma_0^4 + 6\nu^2\sigma_0^2 t + 3\nu^4 t^2, \\ \mathbb{E}\left[S_t^2\right] &= \int_0^t \mathbb{E}\left[\sigma_s^2\right] ds \\ &= \sigma_0^2 t + \frac{1}{2}\nu^2 t^2, \\ \mathbb{E}\left[S_t^2\right]^2 &= \sigma_0^4 t^2 + \nu^2\sigma_0^2 t^3 + \frac{1}{4}\nu^4 t^4, \\ \mathbb{E}\left[\sigma_s^2 S_s^2\right] &= \mathbb{E}\left[\int_0^t S_s^2 d[\sigma,\sigma]_s\right] + \mathbb{E}\left[\int_0^t \sigma_s^2 d[S,S]_s\right] \\ &= \nu^2 \int_0^t \mathbb{E}\left[S_s^2\right] ds + \int_0^t \mathbb{E}\left[\sigma_s^4\right] ds \\ &= \frac{1}{2}\nu^2\sigma_0^2 t^2 + \frac{1}{6}\nu^4 t^3 + \sigma_0^4 t + 3\nu^2\sigma_0^2 t^2 + \nu^4 t^3 \\ &= \sigma_0^4 t + \frac{7}{2}\nu^2\sigma_0^2 t^2 + \frac{7}{6}\nu^4 t^3, \\ \mathbb{E}\left[S_t^4\right] &= 6\mathbb{E}\left[\int_0^t S_s^2 d[S,S]_s\right] \\ &= 6\int_0^t \mathbb{E}\left[S_s^2\sigma_s^2\right] ds \end{aligned}$$

$$= 3\sigma_0^4 t^2 + 7\nu^2\sigma_0^2 t^3 + \frac{7}{4}\nu^4 t^4.$$

Combining these, we obtain

$$\mathbb{E}\left[S_t^2\right] = \mathbb{E}\left[\int_0^t \sigma_s^2 ds\right] = \int_0^t \mathbb{E}\left[\sigma_s^2\right] ds = \sigma_0^2 t + \frac{1}{2}\nu^2 t^2,$$
$$\mathrm{Var}(S_t^2) = \mathbb{E}\left[S_t^4\right] - \mathbb{E}\left[S_t^2\right]^2 = 2\sigma_0^4 t^2 + 6\nu^2\sigma_0^2 t^3 + \frac{3}{2}\nu^4 t^4.$$

□

2.7.4 Proof of Proposition 9

Define the estimators

$$\hat{\sigma}_\pm^2 = \frac{1}{T}\sum_{i=1}^{N}\left(S_{\pm iT/N} - S_{\pm(i-1)T/N}\right)^2.$$

Using Lemma 1 and the Markov property, we have that

$$\begin{aligned}
\mathbb{E}\left[\hat{\sigma}_+^2\right] &= \frac{1}{T}\sum_{i=1}^{N}\mathbb{E}\left[\left(S_{iT/N} - S_{(i-1)T/N}\right)^2\right] \\
&= \frac{1}{T}\sum_{i=1}^{N}\mathbb{E}\left[\int_{iT/N}^{(i-1)T/N}\sigma_t^2 dt\right] \\
&= \frac{1}{T}\mathbb{E}\left[\int_0^T \sigma_t^2 dt\right],
\end{aligned}$$

and, similarly,

$$\mathbb{E}\left[\hat{\sigma}_-^2\right] = \frac{1}{T}\mathbb{E}\left[\int_{-T}^0 \sigma_t^2 dt\right],$$

from which we conclude that the estimators are unbiased.

From Lemma 1, we also have

$$\begin{aligned}
\mathbb{E}\left[\hat{\sigma}_+^2 - \hat{\sigma}_-^2\right] &= \frac{1}{T}\left(\mathbb{E}\left[\int_0^T \sigma_t^2 dt\right] + \mathbb{E}\left[\int_0^{-T}\sigma_t^2 dt\right]\right) \\
&= \frac{1}{T}\left(\sigma_0^2 T + \frac{1}{2}\nu^2 T^2 + \sigma_0^2(-T) + \frac{1}{2}\nu^2(-T)^2\right) \\
&= \nu^2 T,
\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}\left(\hat{\sigma}_+^2 - \hat{\sigma}_-^2\right) &= 2\mathrm{Var}\left(\hat{\sigma}_+^2\right) \\ &= \frac{1}{T^2}\sum_{i=1}^{N}\mathrm{Var}\left(\left(S_{iT/N} - S_{(i-1)T/N}\right)^2\right) \\ &= \frac{N}{T^2}\left(2\sigma_0^4\frac{T^2}{N^2} + 6\nu^2\sigma_0^2\frac{T^3}{N^3} + \frac{3}{2}\nu^4\frac{T^4}{N^4}\right) \\ &= 2\sigma_0^4\frac{1}{N} + 6\nu^2\sigma_0^2\frac{T}{N^2} + \frac{3}{2}\nu^4\frac{T^2}{N^3}.\end{aligned}$$

Therefore, the signal-to-noise ratio is

$$\frac{\mathbb{E}\left[\hat{\sigma}_+^2 - \hat{\sigma}_-^2\right]^2}{\mathrm{Var}\left(\hat{\sigma}_+^2 - \hat{\sigma}_-^2\right)} = \frac{\nu^4 T^2}{2\sigma_0^4\frac{1}{N} + 6\nu^2\sigma_0^2\frac{T}{N^2} + \frac{3}{2}\nu^4\frac{T^2}{N^3}} = \frac{\nu^4 N T^2}{2\sigma_0^4 + 6\nu^2\sigma_0^2\frac{T}{N} + \frac{3}{2}\nu^4\frac{T^2}{N^2}}$$

If $N^\alpha T \to 1$ as $T \to 0$, then

$$\frac{\mathbb{E}\left[\hat{\sigma}_+^2 - \hat{\sigma}_-^2\right]^2}{\mathrm{Var}\left(\hat{\sigma}_+^2 - \hat{\sigma}_-^2\right)} \overset{T\to 0}{\longrightarrow} \begin{cases} 0, & \alpha < 1/2, \\ \frac{\nu^4}{2\sigma_0^4}, & \alpha = 1/2, \\ \infty, & \alpha > 1/2. \end{cases}$$

$$\mathrm{Var}(S_t^2) = 2\sigma_0^4 t^2 + 6\nu^2\sigma_0^2 t^3 + \frac{3}{2}\nu^4 t^4.$$

$$\hat{\sigma}_{T/2} \overset{T\to 0}{\approx} \frac{\sigma_0^2}{NT}\sum_{i=1}^{N}\left(W_{iT/N} - W_{(i-1)T/N}\right)^2 = \frac{\sigma_0^2}{N^2}\sum_{i=1}^{N} Z_i^2, \quad Z_i \sim N(0,1)$$

$$\hat{\sigma}_{T/2} - \hat{\sigma}_{-T/2} \overset{T\to 0}{\approx} \frac{\sigma_0^2}{N^2}\sum_{i=-N+1}^{N} Z_i^2$$

Chapter 3

High-frequency options market making

3.1 Introduction

3.1.1 Motivation

Optimal market making has gained traction in mathematical finance since the seminal paper by Avellaneda and Stoikov (2008). Shortly after, the first paper on options market making was written by Stoikov and Sağlam (2009). Despite the continuous stream of papers for market making for general assets – see, e.g., the books Guéant et al. (2013), Guéant (2017) and references therein –, there was a hiatus on market making for options. Recently, the topic regained interest and was revisited by El Aoud and Abergel (2015), Baldacci et al. (2021) and Baldacci et al. (2020). Despite the resurgence, we believe the options market making problem is still not well understood.

Options comprise an asset class that requires special treatment for market making strategies. We list some features of options that are relevant for market making:

Stochastic volatility The non-linear payoff of options makes it dependent on the underlying asset's volatility, as empirically shown in Andersen et al. (2001) for daily returns. This dependency is also present at smaller time scales as shown in Chapter 2.

Small number of driving factors Options are driven by the underlying price and the volatility surface. In turn, Cont et al. (2002) has reported that only three principal components are necessary for capturing 95% of the total variance of volatility surfaces for equity markets. This small number of driving factors are important

especially because several distinct options are available for each underlying in a given venue.

Liquidity linked to moneyness At-the-money options are traded more frequently than others. Given that moneyness depends on the underlying price, liquidity is stochastic. We discuss this empirically in Section 3.2.

The existing literature models different combinations of the above features in different manners, which of course has implications on the tractability of the model. Stoikov and Sağlam (2009) employ a discrete-time approach for optimal market making of a single option in three cases: (i) constant volatility with frictionless delta hedging, (ii) constant volatility and market making on both option and underlying (no active delta hedging), and (iii) stochastic volatility version in which vega and gamma risks are modelled. They arrive at recursive formulas for optimal quotes in the multi-period models (ii) and (iii).

The recent papers by El Aoud and Abergel (2015), Baldacci et al. (2021) and Baldacci et al. (2020) are continuous-time models with stochastic volatility that rely on frictionless trading on the underlying asset. El Aoud and Abergel (2015) focus on model misspecification and find analytical formulas for the optimal quotes for a single option. Baldacci et al. (2021) introduce the problem of market making multiple options, for which they provide a low-dimensional PDE to numerically compute the optimal quotes. The dimension reduction is achieved by the assumption of constant vega. Finally, Baldacci et al. (2020) offers an alternative approach to the problem in Baldacci et al. (2021) in which the dimension of the PDE is obtained heuristically via a quadratic ansatz inspired by approach in Chapter 1, thus allowing for dynamic vega. None of the papers, however, model the link between liquidity and moneyness.

In this chapter, we focus on exchange-traded vanilla options. These are typically traded by high-frequency market makers – see Menkveld (2013). The assumption of frictionless delta hedging may be realistic in OTC markets, especially for exotic options, but is unsuitable for our application – the liquidity of exchange-traded vanilla options is comparable with the liquidity of its underlying or their futures. Instead, we do not include any *a priori* hedging strategy – a passive hedging strategy naturally emerges from the optimal quotes – see Section 3.3 for more details. This feature is in contrast with the existing literature that, with the exception of model (ii) in Stoikov and Sağlam (2009), rely on frictionless trading on the underlying asset.

We highlight some extensions to our model that we have not included to preserve tractability. One would be to allow the market maker to send market orders, which could be especially useful for options market making so that hedging could be performed actively – either delta hedging by trading on the underlying or even vega hedging by trading on

liquid at-the-money options. Although rare, there is indeed literature in optimal market making that employs limit and market orders – see Guilbaud and Pham (2013) and Section 2.4 in Ricci (2014). A richer literature is found in the related topic of optimal execution, for which the reader is referred to the aforementioned books Guéant et al. (2013) and Guéant (2017).

Another plausible feature in highly correlated markets such as the options market would be a coupling of liquidity. Intuitively, if a trader seeks delta and vega exposure in a particular ratio, there are many combinations of options and underlying that could achieve the desired exposure. Therefore, if the demand for delta and vega exposure is fixed and a subset of options present spreads that are too high, other options would enjoy higher trading activity. To the best of our knowledge, such coupling of liquidity among different assets has never been done in market making literature. The coupling between assets or options in the aforementioned papers are done exclusively via the price processes only – either via the covariance matrix or the Greeks.

A third extension, which is especially useful in the context of exchange-traded financial products is the modelling of competition. Our proposed model derives from the Avellaneda and Stoikov (2008) model, which does not model competition and, as mentioned in Guéant (2017), is better translated in the context of OTC markets in which market makers post their quotes directly to clients and have no information on the quotes of theirs competitors. Hence, in principle, our model depicts a monopolistic market maker. An extension that could add some effects of competition is adverse selection as is done in Chapter 1, in the sense that, the more competition, the more likely is that the traded quote is an unfavourable quote for the market maker that posted it. Other approaches present in the literature is the modelling of partial information as done in Campi and Zabaljauregui (2020) and game-theoretical approaches as in Oomen (2017) and Bank et al. (2021) – see also a mean-field game approach in Huang et al. (2019).

To obtain tractable optimal quotes, we consider high-frequency market makers with short-term strategies, looking to optimise the end-of-day P&L. This motivates the use of small time-to-horizon asymptotics for which we formally derive explicit asymptotic formulas for the optimal quotes. With this method, we retain a fair amount of flexibility with regards to the option dynamics and the shape of the so-called trading intensity function – which models the trade activity as a function of the market maker controls.

Finally, we perform empirical analysis to understand the structure of option spreads as a function of moneyness and expiry. The optimal spreads are found to fit very well with the observed market spreads. This enables us to obtain insights into the codependency of Greeks, option volatility, trade activity and spreads. The explicit link between liquidity and moneyness has not been yet studied in the

3.1.2 Main contributions

On the theoretical side, we explore an overlooked asymptotic approximation. The small time-to-horizon asymptotics has only been briefly mentioned in footnote 8 in Guéant et al. (2013) on the comment that a Taylor expansion of the optimal quotes from their explicit formula for t close to T coincide with the optimal quotes in Avellaneda and Stoikov (2008) – which are optimal quotes for small inventory. Indeed, our optimal quotes for the CARA optimisation criterion reduces to the optimal quotes in Avellaneda and Stoikov (2008) in the single-asset case.

A related approximation is the small risk aversion, which has been studied by Fodra and Labadie (2013) and turn the market making models with stochastic volatility tractable. For our purposes, the small time-to-horizon asymptotics is more appropriate because: (i) it is compatible with the small time asymptotics as studied in Chapter 2 and (ii) produces compact optimal quotes when the liquidation penalty is zero. Another strength of the small time-to-horizon, in general, is that it allows for arbitrary liquidation penalty functions.

Another asymptotic regime is the ergodic limit performed by Guéant et al. (2013) and subsequent papers – see Chapter 1 and also Baldacci et al. (2020) in the case of options market making. The ergodic regime is complementary to our result and one could use both (each in its own regime) to obtain a rough approximation of the global solution.

On the empirical side, we start by investigating the shape of the trade intensity function – a key ingredient in market making models. For this purpose, we show how trade activity varies across moneyness and expiries. It is known at least since George and Longstaff (1993) that at-the-money and close to expiry options present higher trading activity, but we analyse this fact in the market making context.

We then apply the model to find the codependency among Greeks, options volatility, trade activity and spreads. The options market microstructure literature has found empirical connection among these concepts – see Wei and Zheng (2010) and references therein –, and thus our contribution is to provide a theoretical ground for the empirical findings.

As stated in Wei and Zheng (2010), the options market microstructure literature is quite scarce compared to other assets. For the early literature in this field, we refer the reader to the review by Coughenour and Shastri (1999), which typically studies the relationship between the options market and its underlying market. A reminiscent topic that is still active is on price discovery, see e.g. Patel et al. (2019). Bid-ask option spreads have also been regularly debated. Wei and Zheng (2010) and other authors usually apply econometric models for option spread dynamics.

3.1.3 Dataset and source code

The empirical analysis is done on AEX index options quotes and trades tick data on 4 January 2016[1]. The dataset also included the AEX index itself at each timestamp of the tick data. The dataset was freely available on Euronext's website at `ftp://ftp.eua-data.euronext.com`. The interest rate term structure bootstrapping and the Heston calibration were performed with a methodology similar to the one in Chapter 2.

The empirical analysis was performed in Python and used several open-source libraries. We list them by their role in this chapter.

- Heston implementation
  - Fyne (Vieira, 2020)
- Statistical
  - ARCH (Sheppard et al., 2020)
  - Statsmodels (Seabold and Perktold, 2020)

3.1.4 Structure of this chapter

We start by empirically investigating the structure of trading activity across moneyness and expiries in Section 3.2 to obtain insights about the trading intensity function. Then, in Section 3.3, we introduce the options market making model and derive its optimal quotes under three optimisation criteria: (i) risk-neutral, (ii) quadratic running inventory penalty and (iii) expected CARA utility function, and use the small time-to-horizon asymptotics to obtain the explicit optimal quotes in the latter two criteria. Finally, in Section 3.4, we perform an empirical analysis on options spreads across moneyness and expiries and analyse its dependency on Greeks, options volatility and trading activity.

3.2 Empirical trading intensity

3.2.1 Overview

In this section, we empirically investigate a key ingredient for market making models, which is the trading intensity function. The trade intensity function determines the speed at which trades happen at the bid (resp. ask) quote conditional to the spread

[1]This date was the only one freely available when the data analysis was performed.

between the bid (resp. ask) quote and a reference price, which can be interpreted as an ideal fair price. Since the reference price is not observable, we restrict ourselves to the study of the trading activity conditional on the observed bid-ask spread. More precisely, if we denote by $(N_t)_{t\in[0,T]}$ the point process that counts the number of trades with its associated intensity function $(\lambda_t)_{t\in[0,T]}$ and by $(\delta_t)_{t\in[0,T]}$ the half-spread observed in the market, we would like to estimate a deterministic function Λ such that

$$\lambda_t = \Lambda(\delta_t). \tag{3.1}$$

We start with a high-level analysis of the trades across moneyness and expiry in Section 3.2.2. We then turn into the intraday patterns of the trade activity conditional to spreads in Section 3.2.3. There is rich literature in intraday patterns, including trade activity – see e.g. Bouchaud et al. (2018) –, and therefore our contribution is on conditioning this trade activity to the bid-ask spread. Finally, in Section 3.2.4, we calibrate the exponential trading intensity function and discuss statistical issues as well as the shape of the intensity function itself.

3.2.2 Arrival rates overview

In this section, we seek a sensible methodology to estimate the intensity function Λ as in (3.1). For this purpose, we first turn our attention to the structure of arrival rates across option strikes.

Figure 3.1 shows estimates of order arrival rates across strikes and expiries for call and put options. At first glance, we observe a symmetry between calls and puts, in particular the order rates seem higher for slightly out-of-the-money calls and puts. Although not clearly visible in Figure 3.1, there is actually an asymmetry between calls and puts, which is discussed later in Section 3.4.3, where trade activity is revisited but without conditioning on half-spread. Another pattern, now similar to the bid-ask spread patterns in Figure 3.9 is that, as the strike is deeper in-the-money, the spreads increase. We note that there are two tick sizes in the displayed plots: €0.05 and €0.01, which explain why very out-of-the-money options have a finer grid than the remaining options. The main conclusion of Figure 3.1 is that the arrival rate varies across strikes and expiries.

A more subtle pattern in Figure 3.1 is that the arrival rates are not a monotone function of the half-spread. Intuitively, one would expect that higher spreads translate to lower trade activity. We observe, however, that the highest trade activities do occur where the half-spreads are at the lowest levels, but the lowest trade activities do not occur at the highest spreads. One cause for this unintuitive effect is explained by intraday effects,

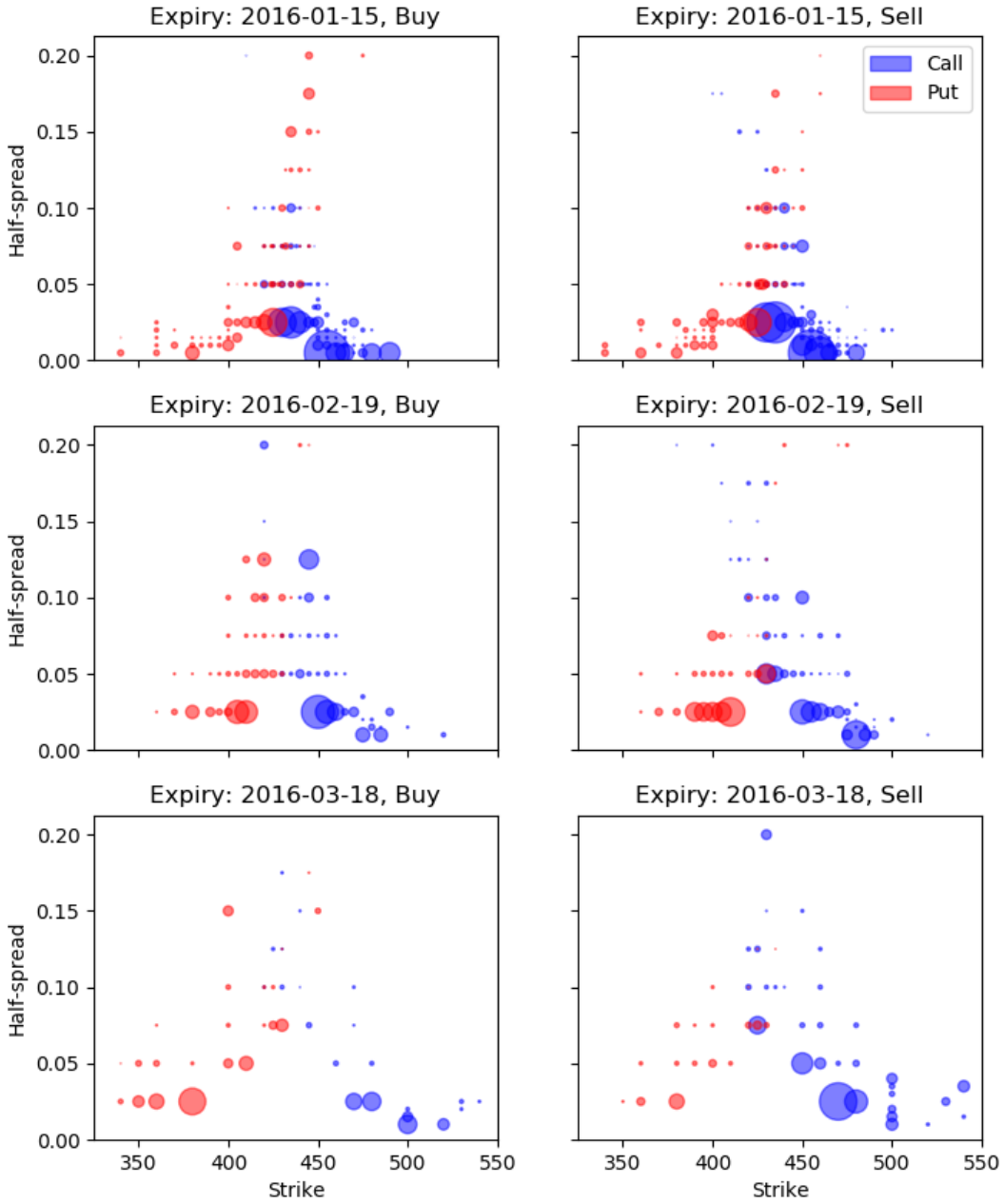


Figure 3.1: Buy and sell order arrival rate estimates conditional on half-spread across strikes and expiries. The larger the bubble, the larger the arrival rate. The estimates are for half-spread that lasted for at least 5 minutes.

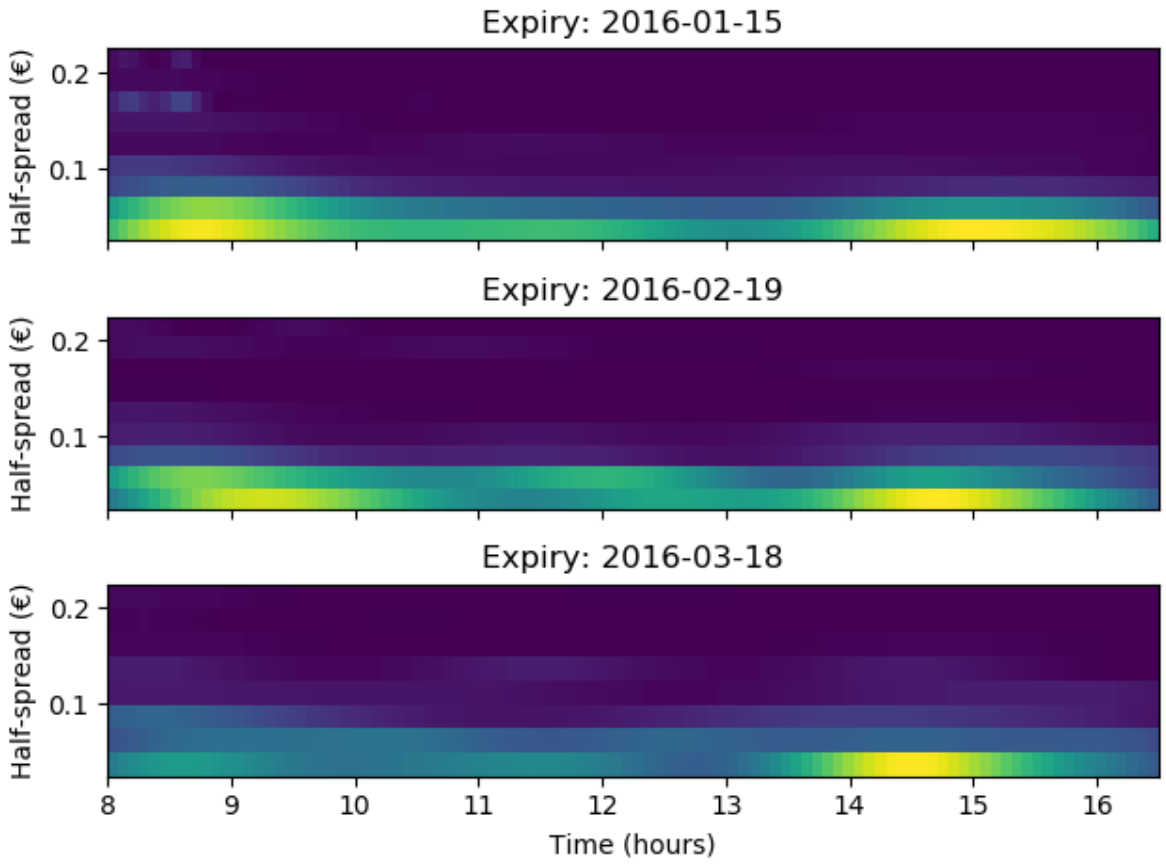


Figure 3.2: Number of trades conditional on the half-spread throughout the day, aggregated among all options. The densities for each half-spread has been computed using kernel density estimation.

given in more detail in Section 3.2.3.

3.2.3 Intraday patterns

As evidenced in Section 3.2.2, the structure of arrival rates is not limited by tick size, strike and expiry. In this section, we look at how the arrival rates change in time.

Figures 3.2 and 3.3 depict the evolution of the number of trades and half-spreads in time, both conditioned on half-spreads. The unconditioned versions of these plots are well studied in the market microstructure literature. Volumes and spreads exhibit the stylised U-shapes. For volumes, Figure 3.2 we indeed observe peaks of volume at the beginning and the end of the trading session. By conditioning on the half-spread, we also observe that the trades at the beginning and the end of the day – but especially at the beginning – happen at higher spread than the rest of the day.

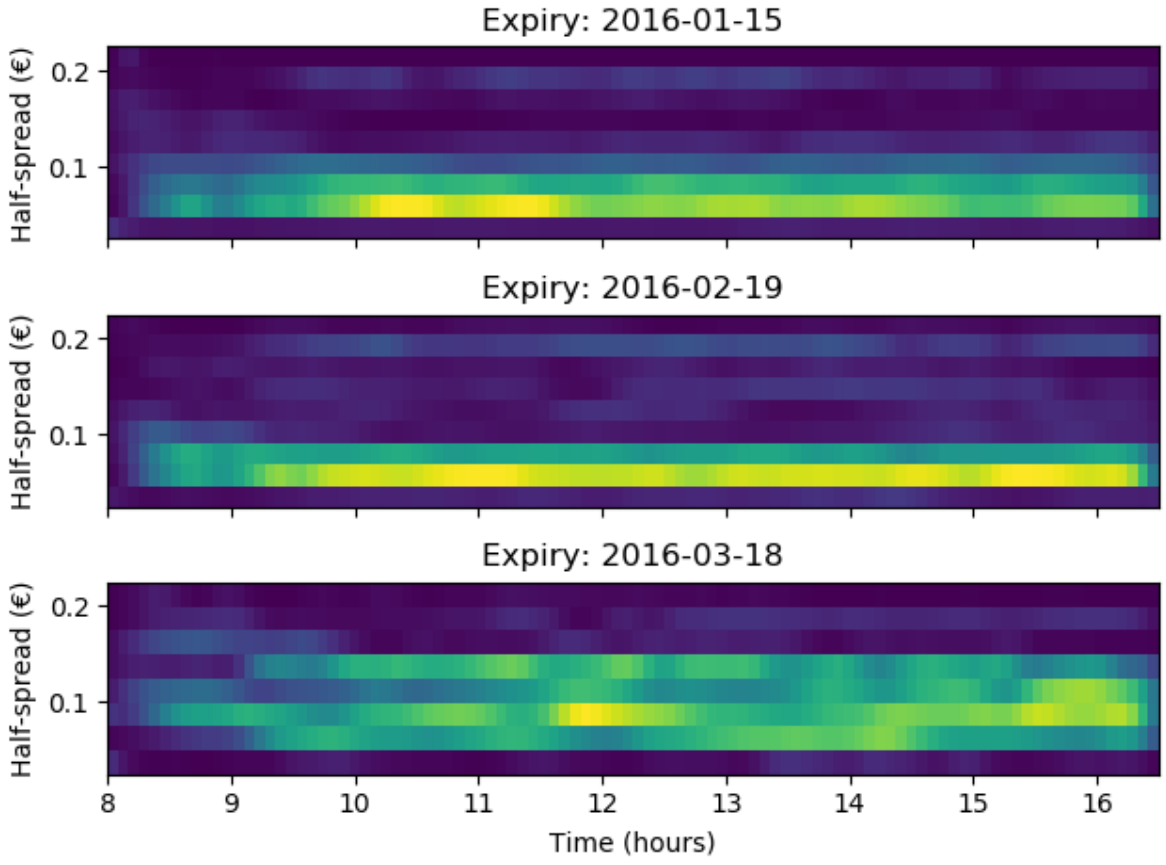


Figure 3.3: Duration of each half-spread throughout the day, aggregated among all options. The densities for each half-spread has been computed using kernel density estimation – the location of each duration is the midpoint of each time interval where the half-spread has been observed continuously.

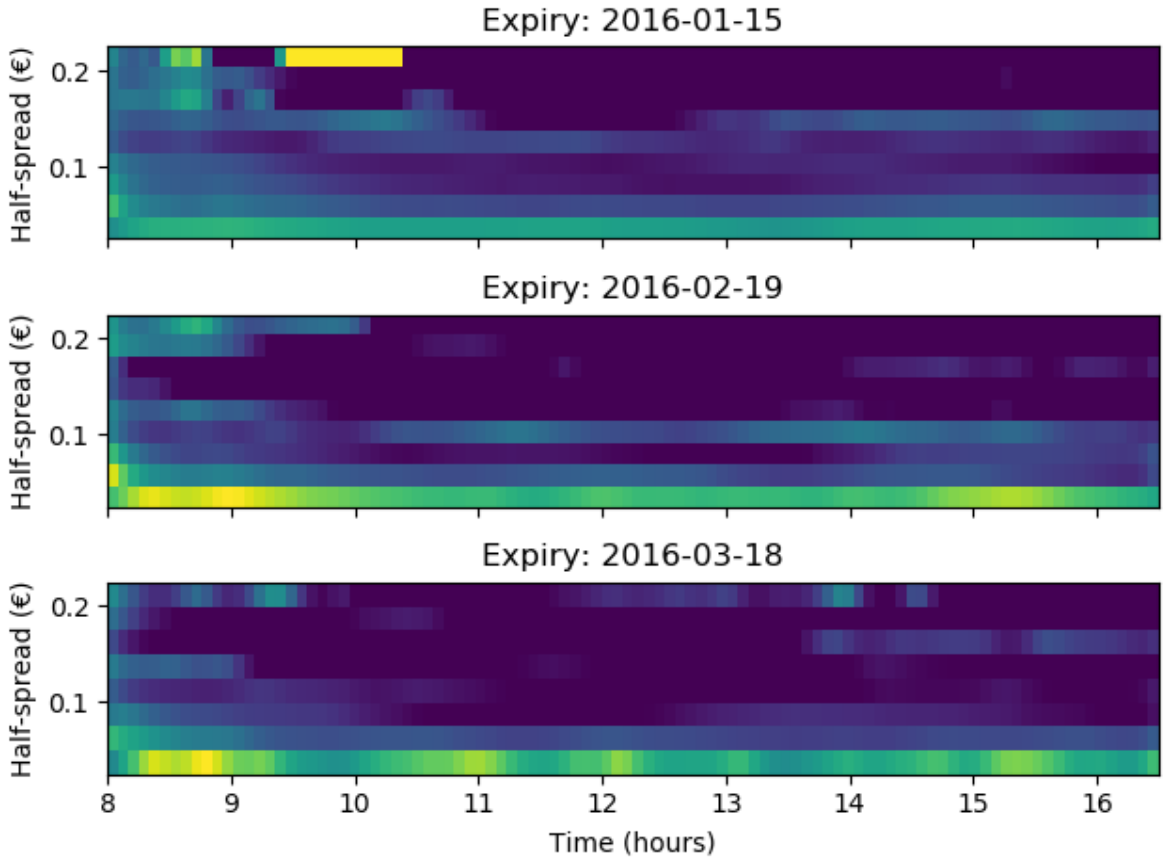


Figure 3.4: Order arrival rate estimates conditional on the half-spread throughout the day, aggregated among all options. The densities are the simple division of the volume densities as in Figure 3.2 and the duration densities as in Figure 3.3. The estimates are shown on a logarithmic scale and estimates shown are clipped between 0.001 and 10 trades per second.

Figure 3.3 also exhibits the stylised U-shape for average spread, being the spread wider at the beginning of the day – this is sometimes referred to as the inverted J-shape. We also observe that the spread is rarely 1 tick, although 2 ticks are the most common spread for options at the two closest expiries.

Finally, Figure 3.4 combines the previous two plots into the arrival rates estimates. From the observations we have made, it is clear that the arrival rates at the beginning and end of the day – especially at the beginning – are skewed towards higher half-spread. As a consequence, arrival rates also have a structure in time, i.e. the trading intensity needs to be modelled as time-dependent. In fact, from the market making perspective, the beginning and end of the trading session would be the most profitable moments in which spreads are wide and trading activity is high. We expect this effect to be present in most asset classes and not only options. As such, given that this chapter focuses on options market making, we henceforth discard the beginning and end of the trading session so that we can assume that the base intensity is constant in time.

We finally revisit the observation in Section 3.2.2 that the arrival rates are not a monotone function of the half-spread and its connection to the intraday patterns. Indeed, the U-shape in arrival rates can explain the higher than expected arrival rates at higher spreads in Figure 3.1.

3.2.4 Exponential fit

With the lessons learned in Sections 3.2.2 and 3.2.3, we are ready to perform a fit to a particular form of the trading intensity function, which is

$$\Lambda_K(\delta) = A_K e^{-\kappa_K \delta}, \quad \forall K \in \{€435, €445, €460\}.$$

The choice of the exponential shape is due to its mathematical tractability – see Section 3.3, and use the option strikes as a measure of moneyness, under the assumption that the underlying price does not change enough during the day to shift too many options from one group to another. From the conclusion of Section 3.2.2, we focus on the second closest expiry 19 Feb 2016 and tick size €0.05 only and group options with similar strikes. Then, to avoid intraday effects on the parameters, we disregard the first hour and the last half hour of the trading session. We can then assume that the base and decay intensity parameters are constant. The calibration problem then reduces to linear regression on the logarithm of arrival rates versus the half-spread. Now, given that some estimates of arrival rates are more accurate than others and since we cannot guarantee that these estimates are independent, we apply Generalised Least Squares to perform the

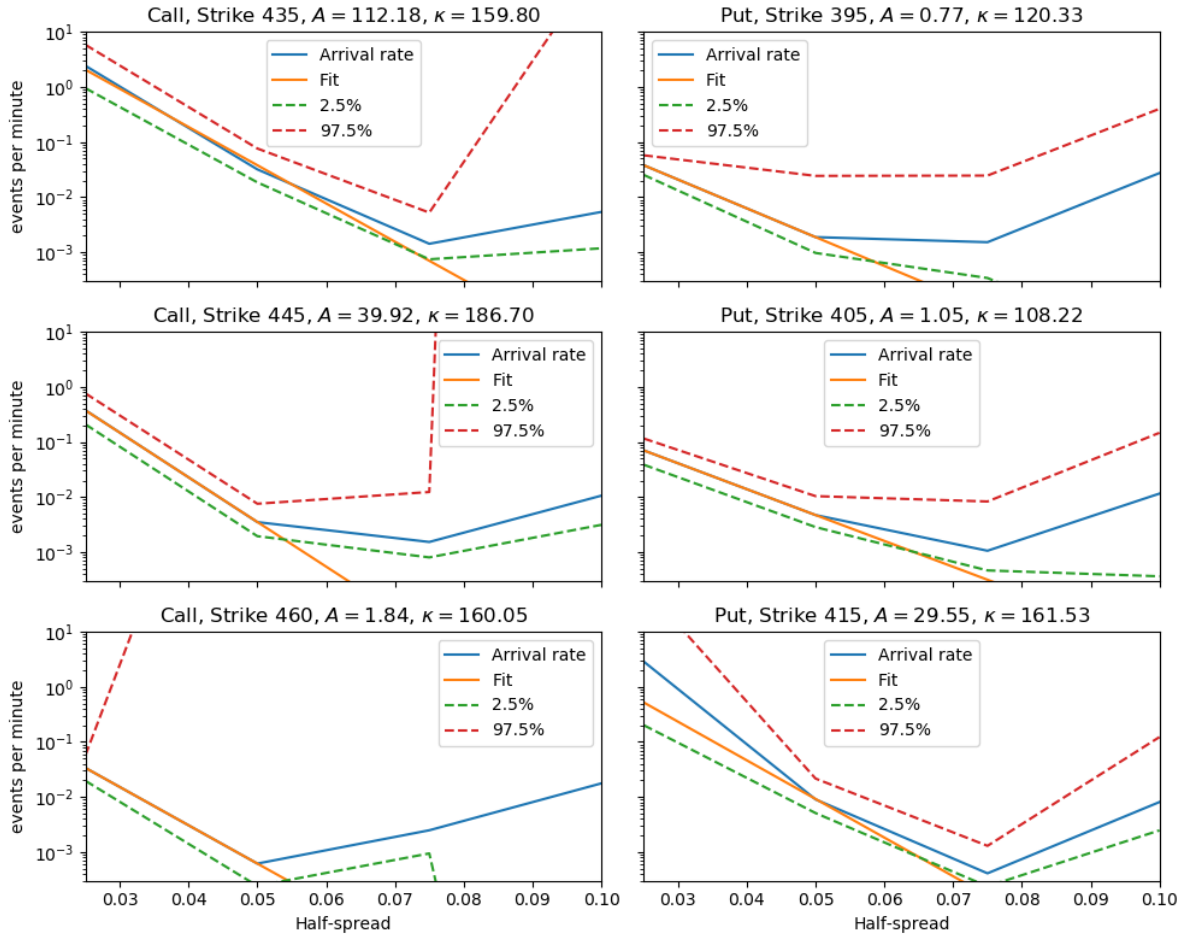


Figure 3.5: GLS fit of the arrival rates for selected strikes of options expiring on 19 Feb 2016 and tick size 0.05, discarding the first hour and the last half hour of the trading session. The selected strikes are strikes corresponding to the number of trades tertiles. The arrival rates are obtained using Nadaraya–Watson kernel regression and both the error bars and the covariance matrix of the errors for the GLS are estimated via stationary bootstrapping.

Class	Strike	A	κ
C	435.0	112.18 (20.96, 600.23)	159.80 (120.08, 199.52)
	445.0	39.92 (25.31, 62.97)	186.70 (174.96, 198.44)
	460.0	1.84 (nan, nan)	160.05 (nan, nan)
P	395.0	0.77 (nan, nan)	120.33 (nan, nan)
	405.0	1.05 (0.80, 1.39)	108.22 (100.97, 115.46)
	415.0	29.55 (0.00, 34895864.26)	161.53 (-119.01, 442.07)

Table 3.1: GLS fit of the arrival rates as in Figure 3.5. The confidence interval for each estimate has 10% significance.

linear regression. The estimates of the covariance matrix of the arrival rate estimates are obtained via stationary bootstrapping – see Politis and Romano (1994).

The resulting calibration can be visualised in Figure 3.5 and Table 3.1. In Figure 3.5, we first observe that even after filtering out the beginning and end of the trading session, the estimates show a slight increase in arrival rates at the highest half-spread. The estimates at high half-spreads are quite inaccurate, though, since this is an unusual regime in the market. Another observation is that the arrival rates present some convexity, which indicates that the arrival rates decay could be better explained with a power law. Looking at the regression fit, on the other hand, shows the effect of GLS, which puts more weight towards the most accurate arrival rate intensity estimates.

Table 3.1 shows that the estimation of the parameters is very noisy. This indicates that either the statistical model is too flexible or that the data is insufficient. We can still conclude, however, that the intensity decay parameter κ appears mostly constant throughout strikes and option types. The base intensity A seems to explain most of the variation of trade activity that we have observed in Figure 3.1.

3.3 Optimal trading strategy

3.3.1 Overview

Our model setup is based on the multi-asset market making model setups of Guéant (2017) and Chapter 1. The model is tailored for options by incorporating three features: (i) a stochastic volatility model for the underlying price, (ii) option dynamics with stochastic Greeks and (iii) a specific trading intensity function motivated by the structure described in Section 3.2. We then formally solve the optimal control problem for three optimisation criteria: (i) risk-neutral, (ii) running quadratic inventory penalty and (iii) the expected

CARA utility function. In each case, we reduce the dimensionality of the problem by applying the ansatz introduced in Fodra and Labadie (2013) for the risk-neutral case and the ansatz introduced in Guéant (2017) for the risk-averse cases.

As in Fodra and Labadie (2013), we solve the risk-neutral problem explicitly via the Feynman-Kac formula, which is useful to compare with the risk-averse case, for which we find explicit formulas for the optimal quotes under the small time-to-horizon approximation. In turn, the small time-to-horizon regime is motivated by our empirical results in the small-time dynamics of option prices in Chapter 2.

We remark that we do not provide verification theorems and as such we do guarantee the optimality of the solutions that we derive. On the other hand, it is reassuring that the solutions we find are compatible with similar results in the literature, namely the risk-neutral case in Fodra and Labadie (2013) and the small time-to-horizon case in footnote 8 in Guéant et al. (2013).

3.3.2 Model setup

Price dynamics

Let $(\Omega, (\mathcal{F}_t)_{t\in[0,T]}, \mathbb{P})$ be a filtered probability measure. The time $T > 0$ is a horizon for the strategy, which must be shorter than all option expiries. The market maker trades on d options, which are driven by n independent Brownian motions. We consider a Markovian setting in which there exists a pricing function $\varphi : [0, T], \mathbb{R}^n \to \mathbb{R}^d$ such that the options reference prices $(C_t)_{t\in[0,T]}$ is given by $C_t = \varphi(t, S_t, V_t)$, where $(S_t)_{t\in[0,T]}$ is the underlying price process and $(V_t)_{t\in[0,T]}$ is the $(n-1)$-vector process of volatility factors.

As such, we model the option price dynamics according to the options Greeks, which arise from formally applying the Itô formula to φ, i.e.

$$dC_t = \Theta_t dt + \Delta_t dS_t + \mathcal{V}_t dV_t = \Theta_t dt + \bar{\Delta}_t d\bar{S}_t.$$

where the drift vector $(\Theta_t)_{t\in[0,T]}$ and the Greek delta vector $(\Delta_t)_{t\in[0,T]}$ are $\mathbb{R}^d$-valued adapted process, the matrix of volatility-related Greeks $(\mathcal{V}_t)_{t\in[0,T]}$ is an $\mathbb{R}^{d\times(n-1)}$-valued adapted process, and underlying price and volatility factors $\big(\bar{S}_t = (S_t, V_t)\big)_{t\in[0,T]}$ are an $\mathbb{R}^n$-valued Itô process given by

$$d\bar{S}_t = \mu_t dt + \sigma_t dW_t,$$

where $(\mu_t)_{t\in[0,T]}$ is an $\mathbb{R}^n$-valued adapted process and $(\sigma_t)_{t\in[0,T]}$ is an $\mathbb{R}^{n\times n}$-valued adapted process, and $(W_t)_{t\in[0,T]}$ is an $\mathbb{R}^n$-vector of independent Brownian motions.

The drift vector $(\Theta_t)_{t\in[0,T]}$ is included in the option price dynamics because the market is incomplete, i.e. the option cannot be fully hedged by the first-order Greeks $(\bar{\Delta}_t)_{t\in[0,T]}$ alone. In fact, in the case where $\mathbb{Q} = \mathbb{P}$, we need $\Theta_t dt$ to cancel the drift from $\mathcal{V}_t dV_t$, because $(V_t)_{t\in[0,T]}$ is not necessarily a $\mathbb{Q}$-martingale. We also note that by the Itô formula argument, we have that $(\Theta_t)_{t\in[0,T]}$ is composed of the Greek theta in the time derivative, and the Greek gamma and second-order volatility Greeks in the Itô term.

The controls are the bid and ask prices $(C_t^{\text{bid}}, C_t^{\text{ask}})_{t\in[0,T]}$ around the reference price via the mark-down and mark-up vectors $(\delta_t^{\text{bid}}, \delta_t^{\text{ask}})_{t\in[0,T]}$ with

$$C_t^{\text{bid}} = C_t - \delta_t^{\text{bid}}, \quad C_t^{\text{ask}} = C_t + \delta_t^{\text{ask}}.$$

Trade dynamics

Define the $\mathbb{N}^d$-valued adapted point processes $(N_t^{\text{bid}}, N_t^{\text{ask}})_{t\in[0,T]}$ as the number of trades that occur at the market maker's bid and ask quotes, respectively. The intensity processeses $(\lambda_t^{\text{bid}}, \lambda_t^{\text{ask}})_{t\in[0,T]}$ associated to $(N_t^{\text{bid}}, N_t^{\text{ask}})_{t\in[0,T]}$ are vector-valued and assumed to be of the form

$$e_i \cdot \lambda_t^{\text{bid}} = \Lambda_i\left(S_t, e_i \cdot \delta_t^{\text{bid}}\right), \quad e_i \cdot \lambda_t^{\text{ask}} = \Lambda_i\left(S_t, e_i \cdot \delta_t^{\text{ask}}\right), \quad \forall i \in \{1, \dots, d\},$$

where $\{e_1, \dots, e_d\}$ denotes the canonical basis of $\mathbb{R}^d$ and $\Lambda_1, \dots, \Lambda_d : [0,T] \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ are exponential intensity functions of the form

$$\Lambda_i(S, \delta) = e^{a_i(S) - b\delta}, \quad \forall i \in \{1, \dots, d\},$$

where $a_1, \dots, a_n : [0,T] \times (0,\infty) \times \mathbb{R} \to \mathbb{R}$ are the base intensity functions and $b > 0$ is the intensity decay parameter.

Define the inventory and cash processes $(q_t)_{t\in[0,T]}$ and $(X_t)_{t\in[0,T]}$ as follows

$$dq_t = dN_t^{\text{bid}} - dN_t^{\text{ask}}, \quad dX_t = C_t^{\text{ask}} dN_t^{\text{ask}} - C_t^{\text{bid}} dN_t^{\text{bid}}.$$

The wealth process $(Y_t)_{t\in[0,T]}$ is defined as the sum of the cash and the mark-to-market value of the inventory, i.e.

$$Y_t = X_t + q_t \cdot C_t.$$

Using the Itô product rule, we rewrite it as

$$\begin{aligned} dY_t &= C_t^{\text{ask}} dN_t^{\text{ask}} - C_t^{\text{bid}} dN_t^{\text{bid}} + q_t \cdot dC_t + C_t \cdot dq_t \\ &= q_t \cdot dC_t + \delta_t^{\text{ask}} dN_t^{\text{ask}} + \delta_t^{\text{bid}} dN_t^{\text{bid}}, \\ &= q_t \cdot \left(\Theta_t + \bar{\Delta}_t \mu_t\right) dt + q_t \cdot \bar{\Delta}_t \sigma_t dW_t + \delta_t^{\text{ask}} dN_t^{\text{ask}} + \delta_t^{\text{bid}} dN_t^{\text{bid}}. \end{aligned}$$

The resulting wealth process can be interpreted as follows. The first two terms $q_t \cdot \left(\Theta_t + \bar{\Delta}_t \mu_t\right) dt$ and $q_t \cdot \bar{\Delta}_t \sigma_t dW_t$ correspond to what we would typically find in self-financing equations – they denote the P&L resulting from the exposure from the Greeks. The last two terms $\delta_t^{\text{ask}} dN_t^{\text{ask}}$ and $\delta_t^{\text{bid}} dN_t^{\text{bid}}$ are unique to market making strategies – they denote the half-spread is immediately earned on every trade.

3.3.3 Risk-neutral case

Optimisation problem and HJB equation

The risk-neutral market maker solves the optimisation problem

$$\sup_{\delta^{\text{bid}}, \delta^{\text{ask}} \in \mathcal{A}} \mathbb{E}\left[Y_T\right],$$

where the set of admissible strategies $\mathcal{A}$ is the set of predictable processes. The associated HJB equation is

$$\begin{aligned} &\left(\partial_t + \mathcal{L}^{S,\Theta,\bar{\Delta},\mu,\sigma}\right) u + \left(q \cdot \left(\Theta + \bar{\Delta}\mu\right)\right) \partial_Y u + \frac{1}{2}\left(q \cdot \bar{\Delta}\sigma\sigma^\top \bar{\Delta}^\top q\right) \partial_{YY} u \\ &+ \sup_{\delta^{\text{bid}}} \sum_{i=1}^{d} \Lambda_i(S, \delta^{\text{bid}}) \left(u(t, q + e_i, Y + \delta_i^{\text{bid}}, S, \Theta, \bar{\Delta}, \mu, \sigma) - u(t, q, Y, S, \Theta, \bar{\Delta}, \mu, \sigma)\right) \\ &+ \sup_{\delta^{\text{ask}}} \sum_{i=1}^{d} \Lambda_i(S, \delta^{\text{ask}}) \left(u(t, q - e_i, Y + \delta_i^{\text{ask}}, S, \Theta, \bar{\Delta}, \mu, \sigma) - u(t, q, Y, S, \Theta, \bar{\Delta}, \mu, \sigma)\right) = 0, \end{aligned} \tag{3.2}$$

with the terminal condition $u(T, q, Y, S, \Theta, \bar{\Delta}, \mu, \sigma) = Y$, where $\mathcal{L}^{S,\Theta,\bar{\Delta},\mu,\sigma}$ denotes the infinitesimal generator of $(S_t, \Theta, \bar{\Delta}_t, \mu_t, \sigma_t)_{t\in[0,T]}$.

Solution

Using the ansatz

$$u(t, q, Y, S, \Theta, \bar{\Delta}, \mu, \sigma) = Y - v_0(t, S, \Theta, \bar{\Delta}, \mu) - q \cdot v_1(t, \Theta, \bar{\Delta}, \mu)$$

on the HJB equation (3.2), we obtain

$$\begin{aligned}\left(\partial_t + \mathcal{L}^{S,\Theta,\bar{\Delta},\mu}\right) v_0 + q \cdot \left(\partial_t + \mathcal{L}^{\Theta,\bar{\Delta},\mu}\right) v_1 - q \cdot \left(\Theta + \bar{\Delta}\mu\right) \\ + \sup_{\delta^{\text{bid}}} \sum_{i=1}^{d} e^{a_i(S) - b\delta_i^{\text{bid}}} \left(-\delta_i^{\text{bid}} + e_i \cdot v_1(t,\Theta,\bar{\Delta},\mu)\right) \\ + \sup_{\delta^{\text{ask}}} \sum_{i=1}^{d} e^{a_i(S) - b\delta_i^{\text{ask}}} \left(-\delta_i^{\text{ask}} - e_i \cdot v_1(t,\Theta,\bar{\Delta},\mu)\right) = 0,\end{aligned}$$

with the terminal condition $v_0(T,S,\Theta,\bar{\Delta},\mu) = v_1(T,\Theta,\bar{\Delta},\mu) = 0$.

From this reduced HJB equation, we obtain that the optimal controls $\delta^{\text{bid}*}$ and $\delta^{\text{ask}*}$ in feedback form are

$$\delta^{\text{bid}*}(t,\Theta,\bar{\Delta},\mu) = \frac{1}{b} + v_1(t,\Theta,\bar{\Delta},\mu), \quad \delta^{\text{ask}*}(t,\Theta,\bar{\Delta},\mu) = \frac{1}{b} - v_1(t,\Theta,\bar{\Delta},\mu),$$

where v_1 is a solution to the PDE

$$\left(\partial_t + \mathcal{L}^{\Theta,\bar{\Delta},\mu}\right) v_1 = \Theta + \bar{\Delta}\mu.$$

The Feynman-Kac formula yields

$$v_1(t,\Theta,\bar{\Delta},\mu) = -\mathbb{E}\left[\int_t^T \left(\Theta_s + \bar{\Delta}_s\mu_s\right) ds \middle| \Theta_t = \Theta, \bar{\Delta}_t = \bar{\Delta}, \mu_t = \mu\right].$$

Therefore, the optimal controls are

$$\begin{aligned}\delta^{\text{bid}*}(t,\Theta,\bar{\Delta},\mu) &= \frac{1}{b} - \mathbb{E}\left[\int_t^T \left(\Theta_s + \bar{\Delta}_s\mu_s\right) \middle| \Theta_t = \Theta, \bar{\Delta}_t = \bar{\Delta}, \mu_t = \mu\right]. \\ \delta^{\text{ask}*}(t,\Theta,\bar{\Delta},\mu), &= \frac{1}{b} + \mathbb{E}\left[\int_t^T \left(\Theta_s + \bar{\Delta}_s\mu_s\right) \middle| \Theta_t = \Theta, \bar{\Delta}_t = \bar{\Delta}, \mu_t = \mu\right].\end{aligned} \tag{3.3}$$

Optimal quotes

From (3.3), we can express the optimal quotes as the optimal mid-price and half-spread as follows:

$$\begin{aligned}\frac{C_t^{\text{ask}*} + C_t^{\text{bid}*}}{2} &= \mathbb{E}\left[C_t + \int_t^T \left(\Theta_s + \bar{\Delta}_s\mu_s\right) ds \middle| \mathcal{F}_t\right], \\ &= \mathbb{E}\left[C_T | \mathcal{F}_t\right], \\ \frac{C_t^{\text{ask}*} - C_t^{\text{bid}*}}{2} &= \frac{1}{b}.\end{aligned}$$

In this setting, we have no liquidation penalty, which can be interpreted either as the case where the market maker liquidates the terminal portfolio at the reference price or that it marks it to market without liquidating. This and the absence of further penalisation in the inventory implies that it is not optimal for the market maker to manage their inventory.

The optimal mid-price is the $\mathbb{P}$ expectation of the option price at the end of the horizon. This raises an interesting distinction between the probability measures $\mathbb{P}$ and $\mathbb{Q}$. The optimal mid-price is a $\mathbb{P}$-martingale, rather than a $\mathbb{Q}$-martingale. The relationship between $\mathbb{P}$ and $\mathbb{Q}$, in this case, is that $\mathbb{Q}$ determines the hedging process and the reference price, whereas the role of $\mathbb{P}$ arises from the fact that the strategy's horizon is shorter than any option's expiry, which prevents an arbitrage-free strategy from taking place. For example, if

$$\left| \int_t^T \left(\Theta_s + \bar{\Delta}_s \mu_s \right) ds \right| > 1/b,$$

then the optimal bid or ask at t could be arbitrageable by a replication strategy – if such strategy exists.

The optimal half-spread is the constant $1/b$ and corresponds to the compromise between small spreads and high flow of trades versus large spreads and low flow – see Cartea et al. (2015) for more details.

Despite b being present in the optimal quotes, the function a is absent. Consequently, the optimal quotes are invariant to changes in liquidity caused by moneyness in the risk-neutral case.

3.3.4 Inventory penalty case

Optimisation problem and HJB equation

We now consider the problem of maximisation of the expected terminal wealth with a running quadratic inventory penalty and liquidation penalty, namely

$$\sup_{\delta^{\text{bid}}, \delta^{\text{ask}} \in \mathcal{A}} \mathbb{E} \left[Y_T - \ell(q_T) - \frac{1}{2} \gamma \int_0^T q_t \cdot d[C, C]_t q_t \right],$$

where $\gamma > 0$ is the inventory penalty parameter, $\ell : \mathbb{R}^d \to \mathbb{R}_+$ is a liquidation function that is assumed to be non-decreasing and the set of admissible strategies $\mathcal{A}$ is the set of

predictable processes. The associated HJB equation is

$$\begin{aligned}
&\left(\partial_t + \mathcal{L}^{S,\Theta,\bar{\Delta},\mu,\sigma}\right) u + \left(q \cdot \left(\Theta + \bar{\Delta}\mu\right)\right) \partial_Y u + \frac{1}{2}\left(q \cdot \bar{\Delta}\sigma\sigma^\top \bar{\Delta}^\top q\right) \partial_{YY} u \\
&+ \sup_{\delta^{\text{bid}}} \sum_{i=1}^{d} \Lambda_i(S, \delta^{\text{bid}}) \left(u(t, q + e_i, Y + \delta_i^{\text{bid}}, S, \Theta, \bar{\Delta}, \mu, \sigma) - u(t, q, Y, S, \Theta, \bar{\Delta}, \mu, \sigma)\right) \\
&+ \sup_{\delta^{\text{ask}}} \sum_{i=1}^{d} \Lambda_i(S, \delta^{\text{ask}}) \left(u(t, q - e_i, Y + \delta_i^{\text{ask}}, S, \Theta, \bar{\Delta}, \mu, \sigma) - u(t, q, Y, S, \Theta, \bar{\Delta}, \mu, \sigma)\right) \\
&= \frac{1}{2}\gamma q \cdot \bar{\Delta}\sigma\sigma^\top \bar{\Delta}^\top q,
\end{aligned} \tag{3.4}$$

with the terminal condition $u(T, q, Y, S, \Theta, \bar{\Delta}, \mu, \sigma) = Y - \ell(q)$, where $\mathcal{L}^{S,\Theta,\bar{\Delta},\mu,\sigma}$ denotes the infinitesimal generator of $(S_t, \Theta_t, \bar{\Delta}_t, \mu_t, \sigma_t)_{t\in[0,T]}$.

Small time-to-horizon solution

Using the ansatz $u(t, q, Y, S, \Theta, \bar{\Delta}, \mu, \sigma) = Y - v(t, q, S, \Theta, \bar{\Delta}, \mu, \sigma)$ on (3.4), we obtain

$$\begin{aligned}
&\left(\partial_t + \mathcal{L}^{S,\Theta,\bar{\Delta},\mu,\sigma}\right) v - q \cdot \left(\Theta + \bar{\Delta}\mu\right) + \frac{1}{2}\gamma q \cdot \bar{\Delta}\sigma\sigma^\top \bar{\Delta}^\top q \\
&+ \sup_{\delta^{\text{bid}}} \sum_{i=1}^{d} e^{a_i(S) - b\delta_i^{\text{bid}}} \left(v(t, q + e_i, S, \Theta, \bar{\Delta}, \mu, \sigma) - v(t, q, S, \Theta, \bar{\Delta}, \mu, \sigma) - \delta_i^{\text{bid}}\right) \\
&+ \sup_{\delta^{\text{ask}}} \sum_{i=1}^{d} e^{a_i(S) - b\delta_i^{\text{ask}}} \left(v(t, q - e_i, S, \Theta, \bar{\Delta}, \mu, \sigma) - v(t, q, S, \Theta, \bar{\Delta}, \mu, \sigma) - \delta_i^{\text{ask}}\right) = 0,
\end{aligned}$$

with the terminal condition $v(T, q, S, \Theta, \bar{\Delta}, \mu, \sigma) = \ell(q)$.

From this reduced HJB equation, we obtain that the optimal controls $\delta^{\text{bid}*}$ and $\delta^{\text{ask}*}$ in feedback form are

$$\begin{aligned}
\delta^{\text{bid}*}(t, q, S, \Theta, \bar{\Delta}, \mu, \sigma) &= \frac{1}{b} + v(t, q + e_i, S, \Theta, \bar{\Delta}, \mu, \sigma) - v(t, q, S, \Theta, \bar{\Delta}, \mu, \sigma), \\
\delta^{\text{ask}*}(t, q, S, \Theta, \bar{\Delta}, \mu, \sigma) &= \frac{1}{b} + v(t, q - e_i, S, \Theta, \bar{\Delta}, \mu, \sigma) - v(t, q, S, \Theta, \bar{\Delta}, \mu, \sigma).
\end{aligned}$$

Substituting this back into the HJB equation, we obtain the PIDE

$$\begin{aligned}\left(\partial_t + \mathcal{L}^{S,\Theta,\bar{\Delta},\mu,\sigma}\right) v - q \cdot \left(\Theta + \bar{\Delta}\mu\right) + \frac{1}{2}\gamma q \cdot \bar{\Delta}\sigma\sigma^\top \bar{\Delta}^\top q \\ + \frac{e^{-1}}{b}\sum_{i=1}^{d} e^{a_i(S) - b\left(v(t,q+e_i,S,\Theta,\bar{\Delta},\mu,\sigma) - v(t,q,S,\Theta,\bar{\Delta},\mu,\sigma)\right)} \\ + \frac{e^{-1}}{b}\sum_{i=1}^{d} e^{a_i(S) - b\left(v(t,q-e_i,S,\Theta,\bar{\Delta},\mu,\sigma) - v(t,q,S,\Theta,\bar{\Delta},\mu,\sigma)\right)} = 0,\end{aligned}$$

with the terminal condition $v(T, q, S, \Theta, \bar{\Delta}, \mu, \sigma) = \ell(q)$.

Now, we switch the direction of time with $\tau = T - t$ and scale $\tau \mapsto \epsilon\tau$

$$\begin{aligned}w^\epsilon(\tau, q, S, \Theta, \bar{\Delta}, \mu, \sigma) &= v(T - \epsilon\tau, q, S, \Theta, \bar{\Delta}, \mu, \sigma), \\ \partial_\tau w^\epsilon(\tau, q, S, \Theta, \bar{\Delta}, \mu, \sigma) &= -\epsilon\partial_t v(T - \epsilon\tau, q, S, \Theta, \bar{\Delta}, \mu, \sigma),\end{aligned}$$

so that the PIDE becomes

$$\begin{aligned}\partial_\tau w^\epsilon = \epsilon\mathcal{L}^{S,\Theta,\bar{\Delta},\mu,\sigma} w^\epsilon - \epsilon q \cdot \left(\Theta + \bar{\Delta}\mu\right) + \frac{1}{2}\epsilon\gamma q \cdot \bar{\Delta}\sigma\sigma^\top \bar{\Delta}^\top q \\ + \epsilon\frac{e^{-1}}{b}\sum_{i=1}^{d} e^{a_i(S) - b\left(w^\epsilon(\tau,q+e_i,S,\Theta,\bar{\Delta},\mu,\sigma) - w^\epsilon(\tau,q,S,\Theta,\bar{\Delta},\mu,\sigma)\right)} \\ + \epsilon\frac{e^{-1}}{b}\sum_{i=1}^{d} e^{a_i(S) - b\left(w^\epsilon(\tau,q-e_i,S,\Theta,\bar{\Delta},\mu,\sigma) - w^\epsilon(\tau,q,S,\Theta,\bar{\Delta},\mu,\sigma)\right)},\end{aligned}$$

with the initial condition $w^\epsilon(0, q, S, \Theta, \bar{\Delta}, \mu, \sigma) = \ell(q)$.

We then propose an ansatz that expands the solution in ϵ

$$w^\epsilon(\tau, q, S, \Theta, \bar{\Delta}, \mu, \sigma) = \ell(q) + \sum_{i=1}^{\infty} \epsilon^i w_i(\tau, q, S, \Theta, \bar{\Delta}, \mu, \sigma),$$

and expand the Taylor series of the exponential functions on ϵ to get

$$\begin{aligned}
&\partial_\tau w_1 + \sum_{n=1}^{\infty} \epsilon^n \partial_\tau w_{n+1} \\
&= -q \cdot \left(\Theta + \bar{\Delta}\mu\right) + \frac{1}{2}\gamma q \cdot \bar{\Delta}\sigma\sigma^\top \bar{\Delta}^\top q + \sum_{n=1}^{\infty} \epsilon^n \mathcal{L}^{S,\Theta,\bar{\Delta},\mu,\sigma} w_n \\
&+ \frac{e^{-1}}{b} \sum_{i=1}^{d} e^{a_i(S) - b(\ell(q+e_i) - \ell(q))} \prod_{n=1}^{\infty} \sum_{m=0}^{\infty} \frac{\epsilon^{nm}(-b)^m}{m!} \left(w_n(\tau, q+e_i, S, \Theta, \bar{\Delta}, \mu, \sigma) - w_n\right)^m \\
&+ \frac{e^{-1}}{b} \sum_{i=1}^{d} e^{a_i(S) - b(\ell(q-e_i) - \ell(q))} \prod_{n=1}^{\infty} \sum_{m=0}^{\infty} \frac{\epsilon^{nm}(-b)^m}{m!} \left(w_n(\tau, q-e_i, S, \Theta, \bar{\Delta}, \mu, \sigma) - w_n\right)^m,
\end{aligned}$$

with the initial conditions

$$w_1(0, q, S, \Theta, \bar{\Delta}, \mu, \sigma) = w_2(0, q, S, \Theta, \bar{\Delta}, \mu, \sigma) = \ldots = 0.$$

Since the PIDE must be true for any $\epsilon > 0$, we equate each coefficient of a power of ϵ. The terms constant in ϵ produces the differential equation

$$\begin{aligned}
\partial_\tau w_1 = -q \cdot \left(\Theta + \bar{\Delta}\mu\right) + \frac{1}{2}\gamma q \cdot \bar{\Delta}\sigma\sigma^\top \bar{\Delta}^\top q + \frac{e^{-1}}{b} \sum_{i=1}^{d} e^{a_i(S) - b(\ell(q+e_i) - \ell(q))} \\
+ \frac{e^{-1}}{b} \sum_{i=1}^{d} e^{a_i(S) - b(\ell(q-e_i) - \ell(q))},
\end{aligned}$$

to which the solution is

$$\begin{aligned}
w_1(\tau, q, S, \Theta, \bar{\Delta}, \mu, \sigma) = &- \tau q \cdot \left(\Theta + \bar{\Delta}\mu\right) + \frac{1}{2}\gamma\tau q \cdot \bar{\Delta}\sigma\sigma^\top \bar{\Delta}^\top q \\
&+ \frac{e^{-1}}{b}\tau \sum_{i=1}^{d} e^{a_i(S) - b(\ell(q+e_i) - \ell(q))} + \frac{e^{-1}}{b}\tau \sum_{i=1}^{d} e^{a_i(S) - b(\ell(q-e_i) - \ell(q))}.
\end{aligned}$$

Therefore, the solution for v is

$$\begin{aligned}
v(T - \epsilon\tau, q, S, \Theta, \bar{\Delta}, \mu, \sigma) &= \ell(q) + \epsilon w_1\left(\tau, q, S, \Theta, \bar{\Delta}, \mu, \sigma\right) + O(\epsilon^2) \\
&= \ell(q) - \epsilon\tau q \cdot \left(\Theta + \bar{\Delta}\mu\right) + \frac{1}{2}\gamma\epsilon\tau q \cdot \bar{\Delta}\sigma\sigma^\top \bar{\Delta}^\top q \\
&+ \frac{e^{-1}}{b}\epsilon\tau \sum_{i=1}^{d} e^{a_i(S) - b(\ell(q+e_i) - \ell(q))} \\
&+ \frac{e^{-1}}{b}\epsilon\tau \sum_{i=1}^{d} e^{a_i(S) - b(\ell(q-e_i) - \ell(q))} + O(\epsilon^2)
\end{aligned}$$

and the optimal controls for each option $i \in \{1, \dots d\}$ are

$$\begin{aligned}
&\delta_i^{\text{bid}*}(T - \epsilon\tau, S, \Theta, \bar{\Delta}, \mu, \sigma) \\
&= \frac{1}{b} - \epsilon\tau e_i \cdot \left(\Theta + \bar{\Delta}\mu\right) + \gamma\epsilon\tau e_i \cdot \bar{\Delta}\sigma\sigma^\top\bar{\Delta}^\top\left(\frac{1}{2}e_i + q\right) + \ell(q + e_i) - \ell(q) \\
&+ \frac{e^{-1}}{b}\epsilon\tau\sum_{j=1}^{d} e^{a_j(S)}\left(e^{-b(\ell(q+e_j+e_i)-\ell(q+e_i))} - e^{-b(\ell(q+e_j)-\ell(q))}\right) \\
&+ \frac{e^{-1}}{b}\epsilon\tau\sum_{j=1}^{d} e^{a_j(S)}\left(e^{-b(\ell(q-e_j+e_i)-\ell(q+e_i))} - e^{-b(\ell(q-e_j)-\ell(q))}\right) + O(\epsilon^2), \\
&\delta_i^{\text{ask}*}(T - \epsilon\tau, S, \Theta, \bar{\Delta}, \mu, \sigma) \\
&= \frac{1}{b} + \epsilon\tau e_i \cdot \left(\Theta + \bar{\Delta}\mu\right) + \gamma\epsilon\tau e_i \cdot \bar{\Delta}\sigma\sigma^\top\bar{\Delta}^\top\left(\frac{1}{2}e_i - q\right) + \ell(q - e_i) - \ell(q) \\
&+ \frac{e^{-1}}{b}\epsilon\tau\sum_{j=1}^{d} e^{a_j(S)}\left(e^{-b(\ell(q+e_j-e_i)-\ell(q-e_i))} - e^{-b(\ell(q+e_j)-\ell(q))}\right) \\
&+ \frac{e^{-1}}{b}\epsilon\tau\sum_{j=1}^{d} e^{a_j(S)}\left(e^{-b(\ell(q-e_j-e_i)-\ell(q-e_i))} - e^{-b(\ell(q-e_j)-\ell(q))}\right) + O(\epsilon^2).
\end{aligned} \tag{3.5}$$

Optimal quotes

From (3.5), if we let $\ell \equiv 0$, we have the optimal mid-price and half-spread in vector form

$$\begin{aligned}
\left(\frac{C^{\text{ask}*} + C^{\text{bid}*}}{2}\right)(T - \epsilon\tau, C, g, \Theta, \bar{\Delta}, \mu, \sigma) &= C + \epsilon\tau\left(\Theta + \bar{\Delta}\left(\mu - \gamma\sigma\sigma^\top g\right)\right) + O(\epsilon^2), \\
\left(\frac{C^{\text{ask}*} - C^{\text{bid}*}}{2}\right)(T - \epsilon\tau, C, g, \Theta, \bar{\Delta}, \mu, \sigma) &= \frac{1}{b} + \frac{1}{2}\gamma\epsilon\tau\mathcal{D}\left(\bar{\Delta}\sigma\sigma^\top\bar{\Delta}^\top\right) + O(\epsilon^2).
\end{aligned}$$

where $\mathcal{D}$ is the linear operator that maps a square matrix to a vector of its diagonal elements and $g = \bar{\Delta}q$, i.e. the risk exposure on the first-order Greeks.

Concerning the optimal spread, as in the risk-neutral case, we also have the $1/b$ term. Additionally, there is a second term that is an additional widening of spreads if the market maker is more conservative, which is proportional to the variance of the option itself. It is interesting to notice that the variance of an option is tied to its Greeks.

Concerning the mid-price, as in the risk-neutral case, it also estimates the near future movement via the θ and μ terms. This also adds the possibility of arbitrageable quotes again in the analogous case in which $\Theta + \bar{\Delta}\mu$ is large enough. Additionally, it also manages risk of the market maker via a proportional control in g. This term could also be interpreted as a passive hedging arising naturally as a result of risk aversion. The fact

that the spread is constant in the inventory reflects the separation of roles between spread and mid-price: the spread optimises the profitability of the overall strategy, whereas the mid-price is skewed for dynamic risk management.

The link to moneyness for the inventory penalty criterion, as in the risk neutral criterion, is absent due to the lack of the a function in the optimal quotes. This is due to the small time-to-horizon asymptotics since the term a actually appears in the ergodic limit in Chapter 1. For comparison, the optimal half-spread approximation for the inventory penalty criterion under the ergodic limit under our framework is

$$\frac{1}{b} + \frac{1}{2}\sqrt{\frac{\gamma}{2e^{a-1}b}}\left(\bar{\Delta}\sigma\sigma^{\top}\bar{\Delta}^{\top}\right)^{\frac{1}{2}}.$$

The choice of the exponential intensity function also leads to arbitrageable quotes even when $\theta \equiv 0$ and $\mu \equiv 0$. Indeed, for large enough inventory, the mid-price can be skewed enough to let the bid or ask quotes beyond arbitrage bounds. This could be fixed if the intensity function reaches infinity at origin, i.e. $\Lambda_i(S, \delta) \to \infty$ as $\delta \to 0$ from above. An example of an intensity function with this property is the power law, which has been studied in the context of optimal execution in Bayraktar and Ludkovski (2014). We remark that the spread $C_t^{\text{ask}*} - C_t^{\text{bid}*}$, however, can never be negative because it is a sum of non-negative terms.

We remark that the optimal quotes are invariant to the number of traded options, but tied to the number of factors that drive the option prices. This property allows us to quote the whole volatility surface as a continuous function of strike and expiry, since the dimensionality is tied to the number of driving factors (underlying and volatility) instead of the number of options. This implies, in particular, that including or not the underlying asset in the market making strategy does not change the optimal quotes. Of course, this would be different if we acknowledge the differences in liquidity between the underlying asset and its options.

We should also remark that we have ignored the liquidation penalty. Indeed, the optimal controls in (3.5) can be seen as a perturbation around the terminal condition, which is the liquidation penalty. In liquid markets, however, transaction costs are small – especially at the end of the trading session, where the spread is at its minimum – and the inventory penalty already ensures that the inventory is kept at reasonable levels.

3.3.5 CARA case

Optimisation problem and HJB equation

We now consider the problem of maximisation of the expected CARA utility of the terminal wealth and liquidation penalty, namely

$$\sup_{\delta^{\text{bid}},\delta^{\text{ask}}\in\mathcal{A}} \mathbb{E}\left[e^{-\gamma(Y_T - q_T\cdot L q_t)}\right]$$

where $\gamma > 0$ is the risk aversion parameter, $\ell : \mathbb{R}^d \to \mathbb{R}_+$ is a liquidation function that is assumed to be non-decreasing and the set of admissible strategies $\mathcal{A}$ is the set of predictable processes bounded from below. The associated HJB equation is

$$\begin{aligned}
&\left(\partial_t + \mathcal{L}^{S,\Theta,\bar{\Delta},\mu,\sigma}\right) u + \left(q\cdot\left(\Theta + \bar{\Delta}\mu\right)\right)\partial_Y u + \frac{1}{2}\left(q\cdot\bar{\Delta}\sigma\sigma^\top\bar{\Delta}^\top q\right)\partial_{YY} u \\
&+ \sup_{\delta^{\text{bid}}} \sum_{i=1}^{d} \Lambda_i(S,\delta^{\text{bid}})\left(u(t,q+e_i,Y+\delta_i^{\text{bid}},S,\Theta,\bar{\Delta},\mu,\sigma) - u(t,q,Y,S,\Theta,\bar{\Delta},\mu,\sigma)\right) \\
&+ \sup_{\delta^{\text{ask}}} \sum_{i=1}^{d} \Lambda_i(S,\delta^{\text{ask}})\left(u(t,q-e_i,Y+\delta_i^{\text{ask}},S,\Theta,\bar{\Delta},\mu,\sigma) - u(t,q,Y,S,\Theta,\bar{\Delta},\mu,\sigma)\right) = 0,
\end{aligned} \tag{3.6}$$

with the terminal condition $u(T,q,Y,Z) = e^{-\gamma(Y-\ell(q))}$, where $\mathcal{L}^{S,\Theta,\bar{\Delta},\mu,\sigma}$ denotes the infinitesimal generator of $(S_t,\bar{\Delta}_t,\mu_t,\sigma_t)_{t\in[0,T]}$.

Small time-to-horizon solution

Using the ansatz $u(t,q,Y,S,\Theta,\bar{\Delta},\mu,\sigma) = e^{-\gamma\left(Y - v(t,q,S,\Theta,\bar{\Delta},\mu,\sigma)\right)}$ on (3.6), we obtain

$$\begin{aligned}
&\left(\partial_t + \mathcal{L}^{S,\Theta,\bar{\Delta},\mu,\sigma}\right) v - q\cdot\left(\Theta + \bar{\Delta}\mu\right) + \frac{1}{2}\gamma q\cdot\bar{\Delta}\sigma\sigma^\top\bar{\Delta}^\top q \\
&+ \sup_{\delta^{\text{bid}}} \sum_{i=1}^{d} \frac{e^{a_i(S)-b\delta_i^{\text{bid}}}}{\gamma}\left(e^{-\gamma\left(\delta_i^{\text{bid}} - v(t,q+e_i,S,\Theta,\bar{\Delta},\mu,\sigma) + v(t,q,S,\Theta,\bar{\Delta},\mu,\sigma)\right)} - 1\right) \\
&+ \sup_{\delta^{\text{ask}}} \sum_{i=1}^{d} \frac{e^{a_i(S)-b\delta_i^{\text{ask}}}}{\gamma}\left(e^{-\gamma\left(\delta_i^{\text{ask}} - v(t,q-e_i,S,\Theta,\bar{\Delta},\mu,\sigma) + v(t,q,S,\Theta,\bar{\Delta},\mu,\sigma)\right)} - 1\right) = 0,
\end{aligned}$$

with the terminal condition $v(T,q,S,\Theta,\bar{\Delta},\mu,\sigma) = \ell(q)$.

From this reduced HJB equation, we obtain that the optimal controls $\delta^{\text{bid}*}$ and $\delta^{\text{ask}*}$ in

feedback form are

$$\delta^{\text{bid}*}(t,S,\Theta,\bar{\Delta},\mu,\sigma)=\frac{1}{\gamma}\log\left(1+\frac{\gamma}{b}\right)+v(t,q+e_i,S,\Theta,\bar{\Delta},\mu,\sigma)-v(t,q,S,\Theta,\bar{\Delta},\mu,\sigma),$$
$$\delta^{\text{ask}*}(t,S,\Theta,\bar{\Delta},\mu,\sigma)=\frac{1}{\gamma}\log\left(1+\frac{\gamma}{b}\right)+v(t,q-e_i,S,\Theta,\bar{\Delta},\mu,\sigma)-v(t,q,S,\Theta,\bar{\Delta},\mu,\sigma).$$

Substituting this back into the HJB equation, we obtain the PIDE

$$\begin{aligned}&\left(\partial_t+\mathcal{L}^{S,\Theta,\bar{\Delta},\mu,\sigma}\right)v-q\cdot\left(\Theta+\bar{\Delta}\mu\right)+\frac{1}{2}\gamma q\cdot\bar{\Delta}\sigma\sigma^\top\bar{\Delta}^\top q\\&-\frac{\epsilon}{b}\left(1+\frac{\gamma}{b}\right)^{-\frac{b}{\gamma}-1}\sum_{i=1}^{d}e^{a_i(S)-b\left(v(t,q+e_i,S,\Theta,\bar{\Delta},\mu,\sigma)-v(t,q,S,\Theta,\bar{\Delta},\mu,\sigma)\right)}\\&-\frac{\epsilon}{b}\left(1+\frac{\gamma}{b}\right)^{-\frac{b}{\gamma}-1}\sum_{i=1}^{d}e^{a_i(S)-b\left(v(t,q-e_i,S,\Theta,\bar{\Delta},\mu,\sigma)-v(t,q,S,\Theta,\bar{\Delta},\mu,\sigma)\right)}=0,\end{aligned}$$

with the terminal condition $v(T,q,S,\Theta,\bar{\Delta},\mu,\sigma)=\ell(q)$.

Now, we switch the direction of time with $\tau=T-t$ and scale $\tau\mapsto\epsilon\tau$

$$\begin{aligned}w^\epsilon(\tau,q,S,\Theta,\bar{\Delta},\mu,\sigma)&=v(T-\epsilon\tau,q,S,\Theta,\bar{\Delta},\mu,\sigma),\\\partial_\tau w^\epsilon(\tau,q,S,\Theta,\bar{\Delta},\mu,\sigma)&=-\epsilon\partial_t v(T-\epsilon\tau,q,S,\Theta,\bar{\Delta},\mu,\sigma),\end{aligned}$$

so that the PIDE becomes

$$\begin{aligned}\partial_\tau w^\epsilon=&\epsilon\mathcal{L}^{S,\Theta,\bar{\Delta},\mu,\sigma}w^\epsilon-\epsilon q\cdot\left(\Theta+\bar{\Delta}\mu\right)+\frac{1}{2}\epsilon\gamma q\cdot\bar{\Delta}\sigma\sigma^\top\bar{\Delta}^\top q\\&-\frac{\epsilon}{b}\left(1+\frac{\gamma}{b}\right)^{-\frac{b}{\gamma}-1}\sum_{i=1}^{d}e^{a_i(S)-b\left(v(t,q+e_i,S,\Theta,\bar{\Delta},\mu,\sigma)-v(t,q,S,\Theta,\bar{\Delta},\mu,\sigma)\right)}\\&-\frac{\epsilon}{b}\left(1+\frac{\gamma}{b}\right)^{-\frac{b}{\gamma}-1}\sum_{i=1}^{d}e^{a_i(S)-b\left(v(t,q-e_i,S,\Theta,\bar{\Delta},\mu,\sigma)-v(t,q,S,\Theta,\bar{\Delta},\mu,\sigma)\right)},\end{aligned}$$

with the initial condition $w^\epsilon(0,q,S,\Theta,\bar{\Delta},\mu,\sigma)=\ell(q)$.

We then propose an ansatz that expands the solution in ϵ

$$w^\epsilon(\tau,q,S,\Theta,\bar{\Delta},\mu,\sigma)=\ell(q)+\sum_{i=1}^{\infty}\epsilon^i w_i(\tau,q,S,\Theta,\bar{\Delta},\mu,\sigma),$$

and expand the Taylor series of the exponential functions on ϵ to get

$$\begin{aligned}
&\partial_\tau w_1 + \sum_{n=1}^{\infty} \epsilon^n \partial_\tau w_{n+1} \\
&= -q \cdot \left(\Theta + \bar{\Delta}\mu\right) + \frac{1}{2}\gamma q \cdot \bar{\Delta}\sigma\sigma^\top \bar{\Delta}^\top q + \sum_{n=1}^{\infty} \epsilon^n \mathcal{L}^{S,\Theta,\bar{\Delta},\mu,\sigma} w_n \\
&- \frac{1}{b}\left(1+\frac{\gamma}{b}\right)^{-\frac{b}{\gamma}-1} \sum_{i=1}^{d} e^{a_i(S)-b(\ell(q+e_i)-\ell(q))} \prod_{n=1}^{\infty}\sum_{m=0}^{\infty} \frac{\epsilon^{nm}(-b)^m}{m!} \left(w_n(\tau, q+e_i, S, \Theta, \bar{\Delta}, \mu, \sigma) - w_n\right)^m \\
&- \frac{1}{b}\left(1+\frac{\gamma}{b}\right)^{-\frac{b}{\gamma}-1} \sum_{i=1}^{d} e^{a_i(S)-b(\ell(q-e_i)-\ell(q))} \prod_{n=1}^{\infty}\sum_{m=0}^{\infty} \frac{\epsilon^{nm}(-b)^m}{m!} \left(w_n(\tau, q-e_i, S, \Theta, \bar{\Delta}, \mu, \sigma) - w_n\right)^m,
\end{aligned}$$

with the initial conditions

$$w_1(0, q, S, \Theta, \bar{\Delta}, \mu, \sigma) = w_2(0, q, S, \Theta, \bar{\Delta}, \mu, \sigma) = \ldots = 0.$$

Since the PIDE must be true for any $\epsilon > 0$, we equate each factor of ϵ^n to zero. The terms constant in ϵ produces the differential equation

$$\begin{aligned}
\partial_\tau w_1 = -q \cdot \left(\Theta + \bar{\Delta}\mu\right) + \frac{1}{2}\gamma q \cdot \bar{\Delta}\sigma\sigma^\top \bar{\Delta}^\top q &- \frac{1}{b}\left(1+\frac{\gamma}{b}\right)^{-\frac{b}{\gamma}-1} \sum_{i=1}^{d} e^{a_i(S)-b(\ell(q+e_i)-\ell(q))} \\
&- \frac{1}{b}\left(1+\frac{\gamma}{b}\right)^{-\frac{b}{\gamma}-1} \sum_{i=1}^{d} e^{a_i(S)-b(\ell(q-e_i)-\ell(q))},
\end{aligned}$$

to which the solution is

$$\begin{aligned}
w_1(\tau, q, S, \Theta, \bar{\Delta}, \mu, \sigma) = -\tau q \cdot \left(\Theta + \bar{\Delta}\mu\right) + \frac{1}{2}\gamma\tau q \cdot \bar{\Delta}\sigma\sigma^\top \bar{\Delta}^\top q \\
- \frac{\tau}{b}\left(1+\frac{\gamma}{b}\right)^{-\frac{b}{\gamma}-1} \sum_{i=1}^{d} e^{a_i(S)-b(\ell(q+e_i)-\ell(q))} \\
- \frac{\tau}{b}\left(1+\frac{\gamma}{b}\right)^{-\frac{b}{\gamma}-1} \sum_{i=1}^{d} e^{a_i(S)-b(\ell(q-e_i)-\ell(q))}.
\end{aligned}$$

Therefore, the solution for v is

$$\begin{aligned}
v(T-\epsilon\tau, q, S, \Theta, \bar{\Delta}, \mu, \sigma) &= \ell(q) + \epsilon w_1\left(\tau, q, S, \Theta, \bar{\Delta}, \mu, \sigma\right) + O(\epsilon^2) \\
&= \ell(q) - \epsilon\tau q \cdot \left(\Theta + \bar{\Delta}\mu\right) + \frac{1}{2}\gamma\epsilon\tau q \cdot \bar{\Delta}\sigma\sigma^\top\bar{\Delta}^\top q \\
&- \frac{\epsilon\tau}{b}\left(1+\frac{\gamma}{b}\right)^{-\frac{b}{\gamma}-1}\sum_{i=1}^{d} e^{a_i(S)-b(\ell(q+e_i)-\ell(q))} \\
&- \frac{\epsilon\tau}{b}\left(1+\frac{\gamma}{b}\right)^{-\frac{b}{\gamma}-1}\sum_{i=1}^{d} e^{a_i(S)-b(\ell(q-e_i)-\ell(q))} + O(\epsilon^2)
\end{aligned}$$

and the optimal controls for each option $i \in \{1, \dots d\}$ are

$$\begin{aligned}
&\delta_i^{\text{bid}*}(T-\epsilon\tau, S, \Theta, \bar{\Delta}, \mu, \sigma) \\
&= \frac{1}{\gamma}\log\left(1+\frac{\gamma}{b}\right) - \epsilon\tau e_i \cdot \left(\Theta + \bar{\Delta}\mu\right) + \gamma\epsilon\tau e_i \cdot \bar{\Delta}\sigma\sigma^\top \bar{\Delta}^\top\left(\frac{1}{2}e_i + q\right) + \ell(q+e_i) - \ell(q) \\
&- \frac{\epsilon\tau}{b}\left(1+\frac{\gamma}{b}\right)^{-\frac{b}{\gamma}-1}\sum_{j=1}^{d} e^{a_j(S)}\left(e^{-b(\ell(q+e_j+e_i)-\ell(q+e_i))} - e^{-b(\ell(q+e_j)-\ell(q))}\right) \\
&- \frac{\epsilon\tau}{b}\left(1+\frac{\gamma}{b}\right)^{-\frac{b}{\gamma}-1}\sum_{j=1}^{d} e^{a_j(S)}\left(e^{-b(\ell(q-e_j+e_i)-\ell(q+e_i))} - e^{-b(\ell(q-e_j)-\ell(q))}\right) + O(\epsilon^2), \\
&\delta_i^{\text{ask}*}(T-\epsilon\tau, S, \Theta, \bar{\Delta}, \mu, \sigma) \\
&= \frac{1}{\gamma}\log\left(1+\frac{\gamma}{b}\right) + \epsilon\tau e_i \cdot \left(\Theta + \bar{\Delta}\mu\right) + \gamma\epsilon\tau e_i \cdot \bar{\Delta}\sigma\sigma^\top \bar{\Delta}^\top\left(\frac{1}{2}e_i - q\right) + \ell(q-e_i) - \ell(q) \\
&- \frac{\epsilon\tau}{b}\left(1+\frac{\gamma}{b}\right)^{-\frac{b}{\gamma}-1}\sum_{j=1}^{d} e^{a_j(S)}\left(e^{-b(\ell(q+e_j-e_i)-\ell(q-e_i))} - e^{-b(\ell(q+e_j)-\ell(q))}\right) \\
&- \frac{\epsilon\tau}{b}\left(1+\frac{\gamma}{b}\right)^{-\frac{b}{\gamma}-1}\sum_{j=1}^{d} e^{a_j(S)}\left(e^{-b(\ell(q-e_j-e_i)-\ell(q-e_i))} - e^{-b(\ell(q-e_j)-\ell(q))}\right) + O(\epsilon^2).
\end{aligned} \tag{3.7}$$

Optimal quotes

From (3.7), if we let $\ell \equiv 0$, we have the optimal mid-price and half-spread in vector form

$$\begin{aligned}
\left(\frac{C^{\text{ask}*}+C^{\text{bid}*}}{2}\right)(T-\epsilon\tau, C, g, \Theta, \bar{\Delta}, \mu, \sigma) &= C + \epsilon\tau\bar{\Delta}\left(\mu - \gamma\sigma\sigma^\top g\right) + O(\epsilon^2), \\
\left(\frac{C^{\text{ask}*}-C^{\text{bid}*}}{2}\right)(T-\epsilon\tau, C, g, \Theta, \bar{\Delta}, \mu, \sigma) &= \frac{1}{\gamma}\log\left(1+\frac{\gamma}{b}\right) + \frac{1}{2}\gamma\epsilon\tau\mathcal{D}\left(\bar{\Delta}\sigma\sigma^\top\bar{\Delta}^\top\right) + O(\epsilon^2).
\end{aligned}$$

where $\mathcal{D}$ is the linear operator that maps a square matrix to a vector of its diagonal elements and $g = \bar{\Delta} q$, i.e. the risk exposure on the first-order Greeks.

Compared to the inventory penalty case, we observe that the only difference is on the first term of the optimal spread – we have $(1/\gamma)\log(1+\gamma/b)$ instead of $1/b$. These terms come from the optimal controls in feedback form, and therefore they are also present in other asymptotic approximations, such as in Chapter 1, where the limit is taken for $T \to \infty$ for both the inventory penalty and the CARA optimisation problems. For reference, the optimal half-spread approximation for the CARA criterion under the ergodic limit under our framework is

$$\frac{1}{b} + \frac{1}{2}\sqrt{\frac{\gamma}{2e^a b}\left(1+\frac{\gamma}{b}\right)^{1+\frac{b}{\gamma}}}\left(\bar{\Delta}\sigma\sigma^\top\bar{\Delta}^\top\right)^{\frac{1}{2}}.$$

The difference between the CARA and the inventory penalty optimisation criteria is that the risk aversion in the inventory penalty only considers the market risk, whereas the CARA criterion also encompasses the positive jumps on the wealth process, namely $\delta_t^{\text{ask}} dN_t^{\text{ask}} + \delta_t^{\text{bid}} dN_t^{\text{bid}}$. This is then reflected in the first term of the optimal spread and indeed we have that in the limit $\gamma \to 0^+$, the CARA optimal spread converges to the inventory penalty optimal spread:

$$\lim_{\gamma\to 0^+} \frac{1}{\gamma}\log\left(1+\frac{\gamma}{b}\right) = \lim_{\gamma\to 0^+} \log\left(1+\frac{1/b}{1/\gamma}\right)^{1/\gamma} = \frac{1}{b}.$$

As with the other two criteria, we also observe the lack of the a function in the optimal quotes.

A remarkable fact of the optimal controls for the CARA criterion is that it reduces to the original approximation by Avellaneda and Stoikov (2008) for the single-asset case, even though their approximation is for small inventory. The link between the small time-to-horizon and small inventory approximations, however, has already been noticed in (Guéant et al., 2013, footnote 8) by a Taylor expansion of their exact expression for optimal quotes – found under the additional assumption of hard inventory constraints. Nevertheless, the approximations differ when the liquidation penalty is considered.

3.3.6 Numerical illustration

We perform a numerical illustration of the trading strategy with running quadratic inventory penalty under the small time-to-horizon regime. For this, we use on the calibrated Heston model and exponential intensity function parameters in Section 3.2. For risk aversion, we use the value found later in Section 3.4. We trade 3 calls and 3 puts with the

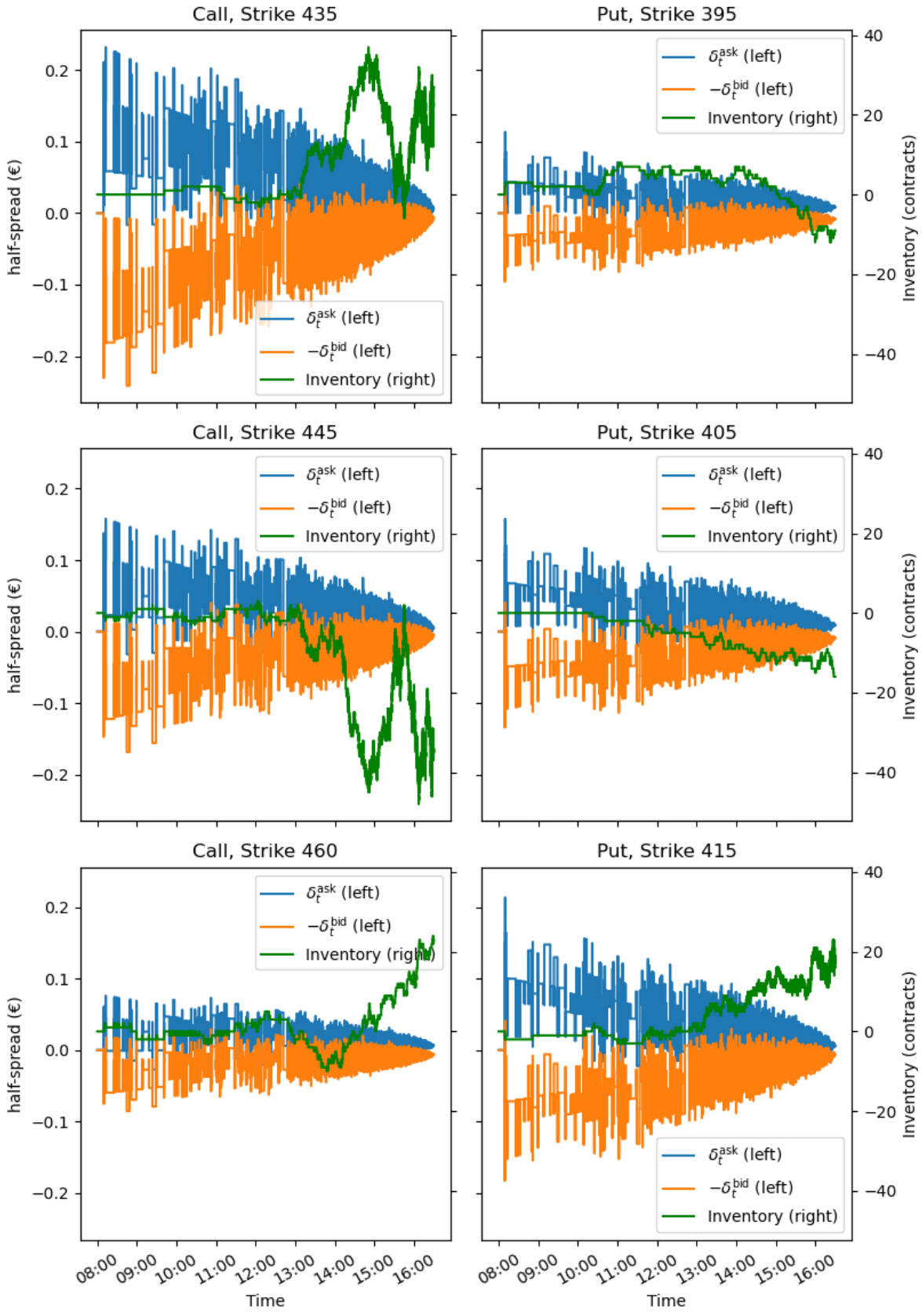


Figure 3.6: Simulation of one day of trading. We display the controls and inventory for the six traded options.

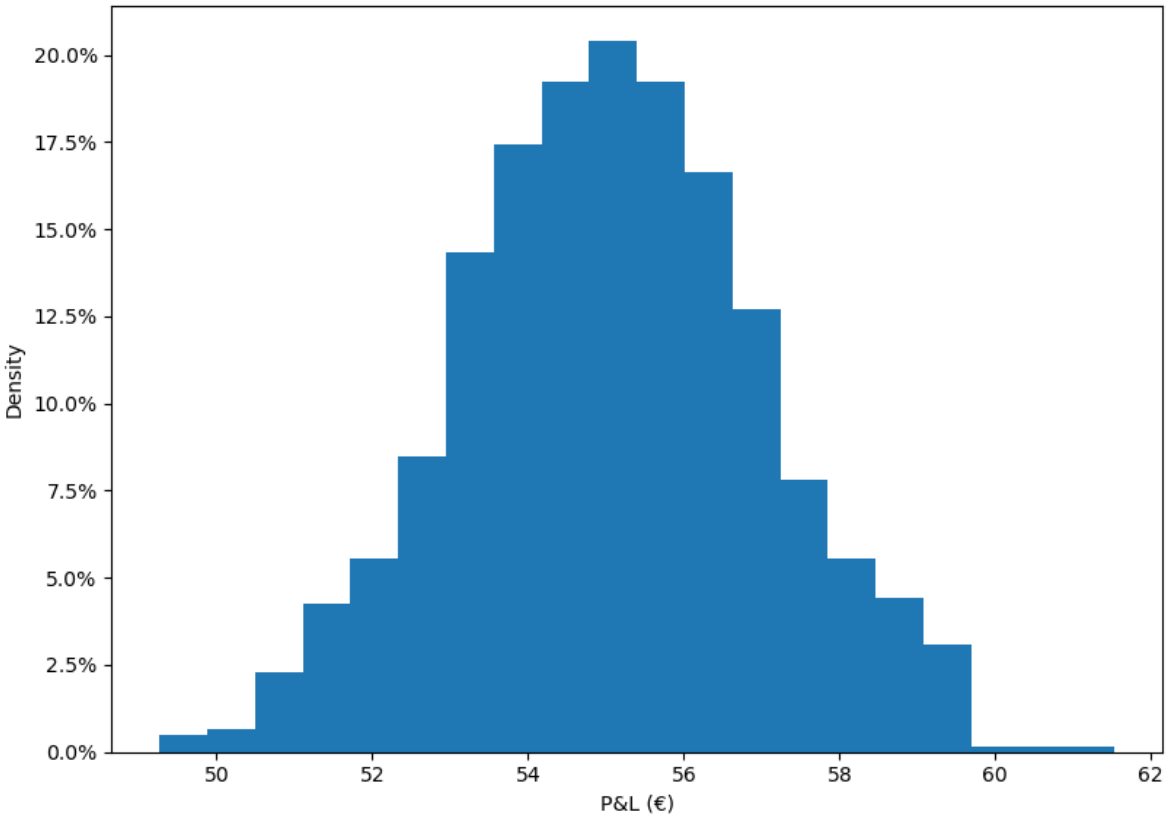


Figure 3.7: Histogram of daily P&Ls for 1,000 simulations.

strikes specified in Table 3.1. For simplicity, we take $\mathbb{P} = \mathbb{Q}$, in which case option prices are martingales, so that Θ cancels the μ term.

In Figure 3.6, we can see some features of the optimal quotes and its effect on the inventory. First, we notice that the spread of the optimal quotes decreases almost linearly in time, as we would expect from the asymptotic formula, which causes a large number of trades by the end of the day. We also note the effect of passive hedging by having inventories with different signs, most notably the almost opposite inventories between the call options with strikes 435 and 445.

In Figure 3.7, we depict the histogram of daily P&Ls, which has not been adjusted for the lot size. The positive mean is the result of the accumulation of the half-spreads at each trade and the dispersion should be mostly due market risk while holding inventory. It is reassuring that all daily P&Ls are positive, which indicates that the passive hedging effect is effective at managing Greeks risk.

3.4 Empirical structure of spreads

3.4.1 Overview

Endowed with the optimal spreads from Section 3.3, we assess how these optimal spreads fit market data and analyse the relationship among Greeks, spreads and trading activity. We do so by analysing two forms of bid-ask spread: the bid-ask spread of option prices and the bid-ask spread of implied volatilities.

In Section 3.4.2, we fit the optimal spreads to the observed bid-ask spreads of option prices, and make the relation between spreads and Greeks. Then, in Section 3.4.3, we analyse the bid-ask spread of implied volatilities and show how spreads and Greeks impact the trading activity of options across moneyness and expiries.

3.4.2 Structure of optimal spreads

We first analyse the structure of spreads of option prices. Figure 3.8 shows the density of the spreads along with the fit from the optimal spreads. The density is obtained using a Gaussian kernel density estimator to obtain the empirical distribution of the spreads as a function of log-moneyness. At first glance, a clear pattern is that the spreads tend to a constant when absolute moneyness goes to infinity.

On the optimal spreads fit, we use the small time-to-horizon asymptotics of the optimal

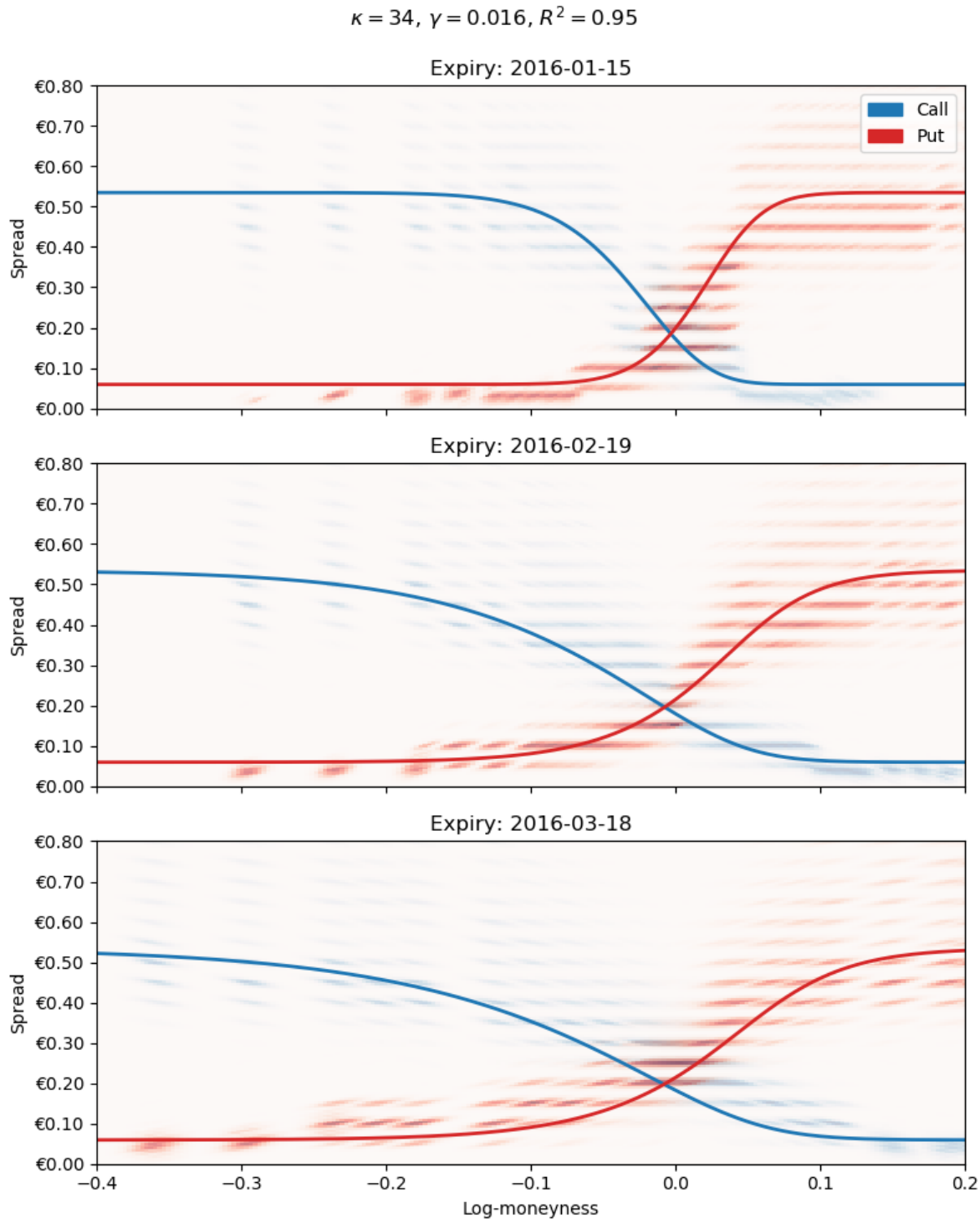


Figure 3.8: The background is a density plot in which the stronger the colour, the longer an option with given moneyness exhibits the given spread. Quotes for all options are aggregated for this plot. The solid lines are the optimal spreads given the regression parameters. The linear regression is done on the average spread conditioned on moneyness. The estimates for the model parameters are $b = 34$ and $\gamma = 0.016$, and the coefficient of determination is $R^2 = 0.95$. The log-moneyness is computed with $\log(K/S_t)$.

spreads from the inventory penalty criterion in Section 3.3. For convenience, we transcribe the optimal spreads here:

$$\left(\frac{C^{\text{ask}*} - C^{\text{bid}*}}{2}\right)(T - \epsilon\tau, C, g, \bar{\Delta}, \mu, \sigma) = \frac{1}{b} + \frac{1}{2}\gamma\epsilon\tau\mathcal{D}\left(\bar{\Delta}\sigma\sigma^{\top}\bar{\Delta}^{\top}\right) + O(\epsilon^2),$$

where we recall $\mathcal{D}$ maps a square matrix to a vector of its diagonal elements.

The above formula enables us to perform a linear regression on the market spread, where the intercept is $1/b$ and the linear parameter is γ – both parameters being fixed across all options. The Greeks are computed with the Heston model calibrated to option prices. We regress the optimal spread against the average spread conditional in log-moneyness and time to expiry from the empirical distribution. We highlight that the regression does not satisfy the usual assumption of an OLS linear regression, so this regression is to be interpreted as an L^2 fit.

Figure 3.8 thus shows that the model fit is very convincing, especially considering that only two parameters were fitted against data from all options of the first three expiries. The R^2 despite being very high, does not convey much information since the regression is on non-stationary variables. Given the optimal formula, we conclude that the shape of the optimal spreads follows the shape of the Delta function, which explains why the spreads converge to a constant when log-moneyness is large in either direction. The model also contains the Vega parameter, but as we have seen in Section 2.4, the spot volatility contribution to the variance of options at small time scales is relatively small even with the Heston simulation study, especially at the tails, hence we indeed expect the shape of Vega to be less perceptible. This shape, however, is only possible in the presence of positive risk aversion, which is found to be $\gamma = 0.016$.

Furthermore, the lack of the base intensity parameter a in the regression model highlights the fact that indeed it has little effect in the high-frequency regime.

3.4.3 Spreads on implied volatility and trading activity

We now turn our attention to the spread of implied volatilities and its connection to trading activity. Figure 3.9 depicts the density of the implied volatility spread and Figure 3.10 depicts the density of trading activity. At first glance, we observe an inverse relationship between the two plots – which is explained later in this section.

In Figure 3.9 shows many features of the implied volatility spread. We again observe symmetry between calls and puts. We have a convex shape whose minimum is slightly out-of-the-money. The shape also disperses with time to expiry.

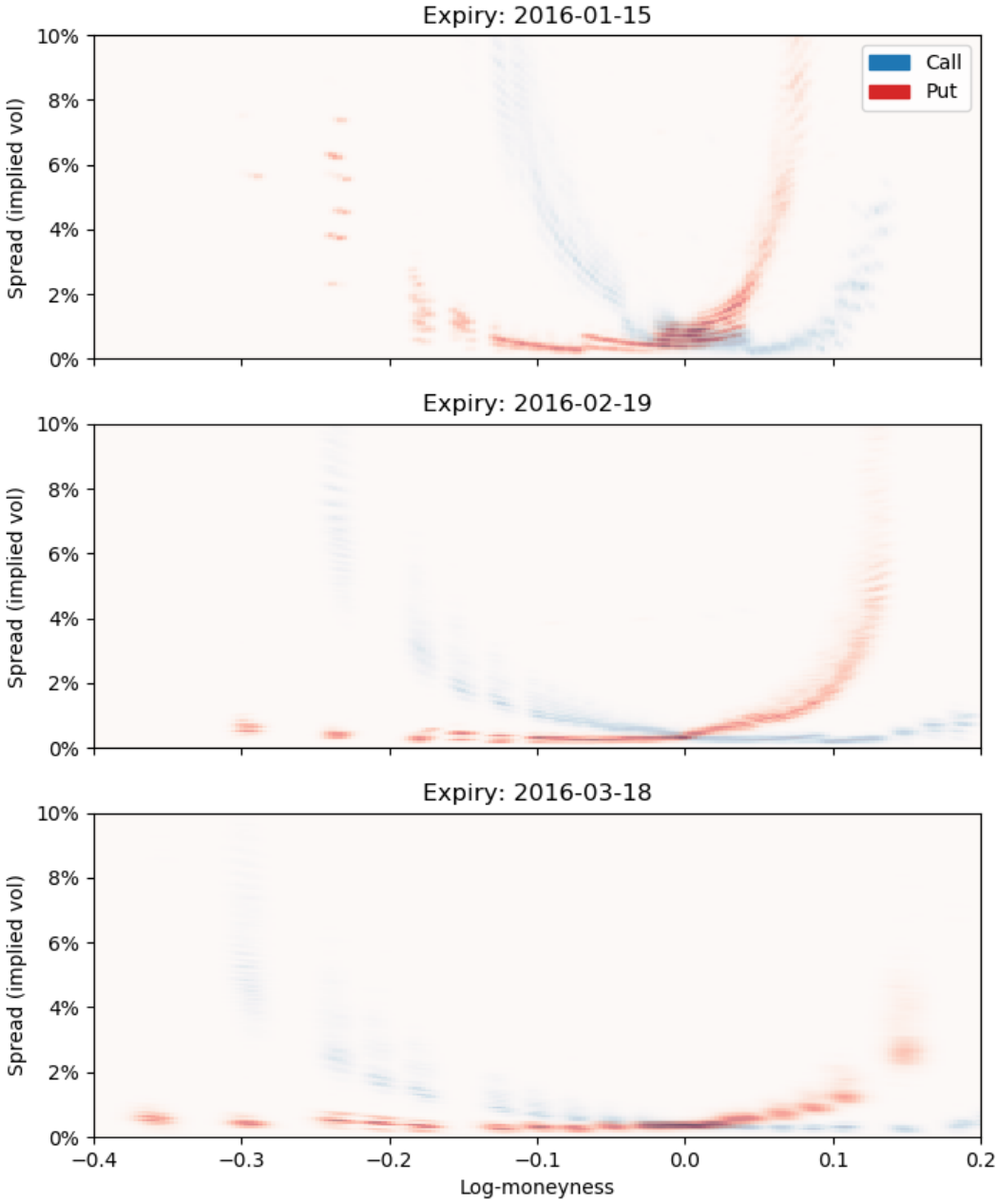


Figure 3.9: Distribution of relative implied volatility spreads versus log-moneyness – i.e. $\log(K/S_t)$ – for different expiries. This relative spread is the difference between the bid and ask implied volatilities divided by their midpoint implied volatility. The stronger the colour, the longer the bid-ask spread at the given level and moneyness.

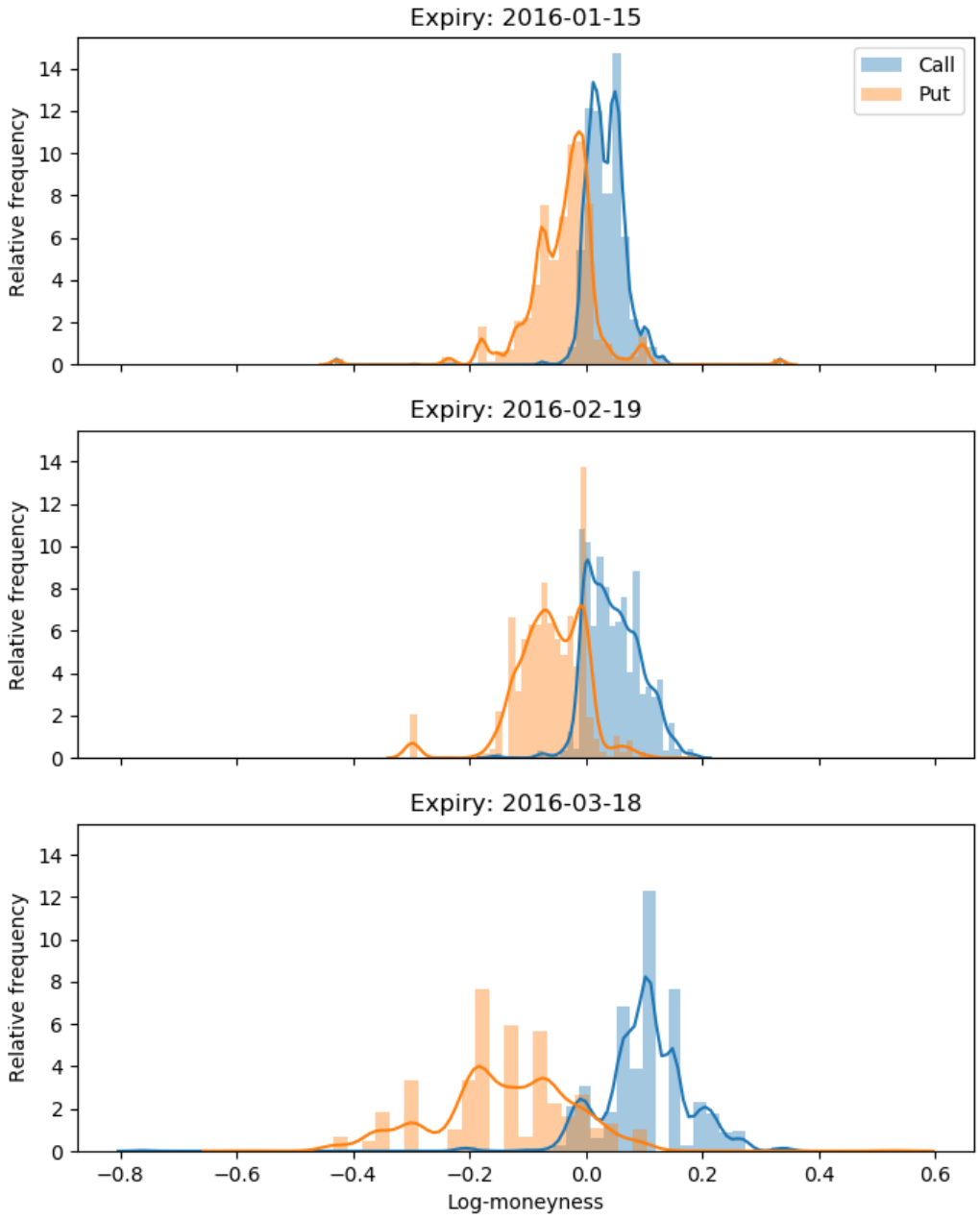


Figure 3.10: Relative trade volumes across log-moneyness – i.e. $\log(K/S_t)$ – on all options belonging to the three closest to expiry dates.

A key property of the implied volatility spread is its relation to the Greek Vega. Given a fixed underlying price, let f be a function from implied volatility to option price. The Greek vega is then $\mathcal{V} = df/d\sigma$. From the mean value theorem, we have that

$$\frac{f(\sigma_2) - f(\sigma_1)}{\sigma_2 - \sigma_1} = \mathcal{V}(\bar{\sigma}),$$

for a $\bar{\sigma} \in [\sigma_1, \sigma_2]$. Therefore,

$$\sigma_2 - \sigma_1 = \frac{f(\sigma_2) - f(\sigma_1)}{\mathcal{V}(\bar{\sigma})}. \tag{3.8}$$

This implies that the implied volatility spread is the bid-ask spread of the option divided by the Greek Vega evaluated at an intermediate point. Combining with the spreads on option prices in Figure 3.8, we have that, although the Greek vega has a peak at the money, the minimum implied volatility spread is offset to slightly out-of-the-money because of lower spreads on option prices. Therefore, although the spread of option prices seem to be more natural from the perspective of our model, the implied volatility spread can be seen as a useful normalisation, especially when comparing spreads with trading activity.

The trading activity in Figure 3.10, as mentioned before, looks like the inverse of the implied volatility spreads in Figure 3.9. We observe symmetry between calls and puts and a concave shape with a peak slightly out-of-the-money. The relation to the Greek Vega in (3.8) shines light to this inverse relation: trading activity is higher on options where the Greek Vega is cheaper.

Upon a closer inspection of Figure 3.10, however, we notice that the peak of the trade activity for put options is smaller than the corresponding peak for call options. This is the asymmetry we have hinted in Figure 3.2.2. The lower activity for put options also contributes to the argument that trading activity is higher on options where the Greek Vega is cheaper. Recall that, in Figure 2.31 from the Greeks estimation study in Chapter 2, the peak on the volatility semi-partial R^2 was higher for calls than for puts, thus indicating that the slightly out-of-the-money call options can offer vega exposure that is less mixed with delta exposure than what out-of-the-money put options can offer.

Another feature in Figure 3.10 is its 'bumpy' shape. Given a range in which the underlying has moved during the day, each bump reflects the existence of an option that covers a range of moneyness. It is also worth noting that some options exhibit proportionally more activity than their neighbouring options. We believe this is connected to the fact that new options are issued to refine the strike grid as they get more liquid. Therefore, older options have an accumulated open interest that likely reflects in more trading activity.

Expiry	Total volume	Relative volume
2016-01-15	18145	60.7%
2016-02-19	6015	20.1%
2016-03-18	3290	11.0%
2016-06-17	1168	3.9%
2016-09-16	206	0.7%
2016-12-16	637	2.1%
2017-06-16	15	0.1%
2017-12-15	83	0.3%
2018-12-21	122	0.4%
2019-12-20	104	0.3%
2020-12-18	84	0.3%

Table 3.2: Total volume of trades for all options grouped by expiry date.

Table 3.2 shows the trade activity across expiries. It is clear that options closer to expiry show higher activity. Similarly to Figure 3.10, however, we also see that some expiries are favoured relative to their neighbouring expiries. We offer an analogous explanation here: the expiry grid gets finer with time and thus options of the sparser grid have accumulated more open interest which reflects in trade activity.

In summary, we observe that our model fits the spreads data well. The lack of the base intensity parameter in the optimal quotes suggests that the shape of the observed spreads is mainly attributed to the risk aversion of the market maker with little or no compromise the trade activity of each option. Instead, the trading activity seems to be motivated by liquidity takers seeking to trade vega cheaply – i.e. with lower transaction costs.

3.5 Conclusion

In Section 3.2, we have fitted the exponential trading intensity function and found that the trading intensity features: (i) a base intensity that is sensitive to the moneyness of the option and (ii) an exponential decay parameter that appears constant. With this shape of intensity function and assuming that the market maker quotes under small time-to-horizon, we find in Section 3.3 compact optimal that quotes, under the inventory penalty and CARA criteria, are insensitive to the base intensity and, thus, only dependent on the exponential decay parameter, which is invariant to the moneyness of options. Besides, the optimal quotes are invariant to the number of options traded, in the sense that optimal quotes only track the exposure to the different Greeks rather than the inventory of each option individually. Furthermore, we learn that the optimal spread across different

options can be summarised in the first-order Greeks of the options – i.e. delta and volatility-related Greeks. This is in contrast to the literature that assumes continuous delta-hedging – as in Baldacci et al. (2020) –, in which the Greek delta is absent from the HJB equation and, consequently, absent from the optimal spreads. Finally, in Section 3.4, we find that the optimal spread indeed fits well with the observed market spreads, which is evidence that the choice of the trade intensity function and the small time-to-horizon regime were reasonable. Furthermore, we find that the shape of market spreads are mainly due to the effect of options Greeks and that the trade activity is higher where Greek vega is cheaper to trade – cheaper in the sense that the spread is tighter. In other words, when varying moneyness and expiries, it seems that lower spreads attract trading activity, but it is not optimal to lower the spreads where the trading activity is higher.

For further research, we suggest four directions: (i) power-law intensity function, (ii) time-dependent base intensity, (iii) stochastic trade sizes and (iv) self-excitation of trades. As noted in Section 3.2, the exponential function does not capture the excess convexity that is observed in the log-log plot. Furthermore, as observed in Section 3.3, the exponential intensity function allows for arbitrage if the inventory is large enough (in either positive or negative directions). We have observed in Section 3.2.3 that the beginning and end of the trading sessions present wide spreads and high trading activity, which are both favourable for market makers. As such, modelling time-dependent base intensity could incorporate this effect and can be relevant to market making on other asset classes. Following the approach by Bergault and Guéant (2019), introducing stochastic trade sizes for the optimal market making model is not a difficult task, and it could be more realistic since the Greeks of the at-the-money and out-of-the-money options can quickly change upon a change in the underlying price and thus the size of the trades would likely follow. Finally, the self-excitation of trades, as done in Cartea et al. (2014), could be another interesting extension. In their model, the self-excitation affects the base intensity only, and thus we would conjecture that this does not change the optimal quotes under our framework.

Concluding remarks

In Chapter 1, we have derived closed-form solutions for the ergodic limit of an asset-agnostic market making model thanks to a quadratic approximation of its Hamiltonian. With this closed-form solution, we could observe the effect of the base intensity for the optimal quotes. In Chapter 2, thanks to a novel methodology to estimate spot volatility changes at small time scales, were able to empirically recover volatility-related Greeks from option price changes. We concluded, therefore, that, even locally, spot volatility drives option prices. In Chapter 3, we have further noticed that the trading activity of options is linked to its moneyness but only via the base intensity. From this observation and the need to model stochastic volatility, we have derived an options market making model suitable for exchange-traded vanilla options. Contrary to the optimal quotes obtained via the ergodic limit in Chapter 1, the optimal quotes in the small time-to-horizon limit are invariant to the base intensity, which implies that the optimal quotes are independent of moneyness in this regime. We then used the optimal spread to link the Greeks of the options to the observed spread and identified the inverse relationship between trading activity and the transaction costs – measured as the bid-ask spread in euros – per unit of vega.

We highlight that our results rely heavily on three assumptions: (i) the exponential shape of the trading intensity, (ii) frictions in trading the underlying and (iii) the short horizon of the market maker. As shown in Chapter 3, the exponential shape of the intensity function has two drawbacks: (i) it does not fit well with empirical data and (ii) it allows the market maker quotes to be arbitrageable when the inventory is large. Base intensity is ignored for equity index options, having trade intensity depend on moneyness under a different regime is still an open problem. The choice of exponential intensity was due to its tractability, but we believe that an intensity function with a power-law shape could be more appropriate. Whether the option is liquid relative to its underlying is important when considering whether the assumption of perfect delta hedge makes sense. In the case of exchange-traded vanilla options, we believe that the liquidity of the options is comparable to their underlying, thus we do not assume the perfect delta-hedging strategy. For exotic options or even for vanilla options of other asset classes –

e.g. options in currency –, the perfect delta hedge assumption could be more realistic than not allowing for active delta hedging. The assumption that the market maker has a short horizon makes sense for liquid exchange-traded options but is less appropriate in markets where trades occur less frequently, such as for swaptions.

Possible extensions on the empirical analysis in Chapter 2 and on the options market making model in Chapter 3 are: (i) using a multi-factor stochastic volatility model – such as the Bergomi model or even a model based on the evolution of the implied volatility surface directly –, and (ii) consider American puts and calls instead of European vanilla options.

References

Abergel, F. and Zaatour, R. (2012). What drives option prices? *Journal of Trading*, 7(3):12–28.

Aït-Sahalia, Y. and Jacod, J. (2014). *High-frequency financial econometrics*. Princeton University Press.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics*, 61(1):43–76.

Avellaneda, M. and Stoikov, S. (2008). High-frequency trading in a limit order book. *Quantitative Finance*, 8(3):217–224.

Baldacci, B. (2020). High-frequency dynamics of the implied volatility surface. *arXiv preprint arXiv:2012.10875*.

Baldacci, B., Bergault, P., and Guéant, O. (2021). Algorithmic market making for options. *Quantitative Finance*, 21(1):85–97.

Baldacci, B., Derchu, J., and Manziuk, I. (2020). An approximate solution for options market-making in high dimension. *arXiv preprint arXiv:2009.00907*.

Bank, P., Ekren, I., and Muhle-Karbe, J. (2021). Liquidity in competitive dealer markets. *Mathematical Finance*, 31(3):827–856.

Baradel, N., Bouchard, B., and Dang, N. M. (2018). Optimal control under uncertainty and bayesian parameters adjustments. *SIAM Journal on Control and Optimization*, 56(2):1038–1057.

Barzykin, A., Bergault, P., and Guéant, O. (2020). Algorithmic market making in FX cash markets. *Working paper*.

Bayraktar, E. and Ludkovski, M. (2014). Liquidation in limit order books with controlled intensity. *Mathematical Finance*, 24(4):627–650.

Bennedsen, M., Lunde, A., and Pakkanen, M. S. (2021). Decoupling the Short- and Long-Term Behavior of Stochastic Volatility. *Journal of Financial Econometrics*. nbaa049.

Bergault, P., Evangelista, D., Guéant, O., and Vieira, D. (2021). Closed-form approximations in multi-asset market making. *Applied Mathematical Finance*, 28(2):101–142.

Bergault, P. and Guéant, O. (2019). Size matters for otc market makers: General results and dimensionality reduction techniques. *Mathematical Finance*.

Bergomi, L. (2015). *Stochastic volatility modeling*. CRC Press.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654.

Bouchaud, J.-P., Bonart, J., Donier, J., and Gould, M. (2018). *Trades, quotes and prices: financial markets under the microscope*. Cambridge University Press.

Brémaud, P. and Jacod, J. (1977). Processus ponctuels et martingales: résultats récents sur la modélisation et le filtrage. *Advances in Applied Probability*, 9(2):362–416.

Buehler, H. (2006). Consistent variance curve models. *Finance and Stochastics*, 10(2):178–203.

Campi, L. and Zabaljauregui, D. (2020). Optimal market making under partial information with general intensities. *Applied Mathematical Finance*, 27(1-2):1–45.

Cartea, Á., Donnelly, R., and Jaimungal, S. (2017). Algorithmic trading with model uncertainty. *SIAM Journal on Financial Mathematics*, 8(1):635–671.

Cartea, Á., Jaimungal, S., and Penalva, J. (2015). *Algorithmic and high-frequency trading*. Cambridge University Press.

Cartea, Á., Jaimungal, S., and Ricci, J. (2014). Buy low, sell high: A high frequency trading perspective. *SIAM Journal on Financial Mathematics*, 5(1):415–444.

Cartea, Á., Jaimungal, S., and Ricci, J. (2018). Algorithmic trading, stochastic control, and mutually exciting processes. *SIAM Review*, 60(3):673–703.

Cohen, J. (2013). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Routledge Ltd - M.U.A., 3rd ed. edition.

Cont, R., Da Fonseca, J., et al. (2002). Dynamics of implied volatility surfaces. *Quantitative Finance*, 2(1):45–60.

Coughenour, J. and Shastri, K. (1999). Symposium on market microstructure: A review of empirical research. *Financial Review*, 34(4):1–27.

Da Fonseca, J. and Grasselli, M. (2011). Riding on the smiles. *Quantitative Finance*, 11(11):1609–1632.

Dragulescu, A. A. and Yakovenko, V. M. (2002). Probability distribution of returns in the heston model with stochastic volatility. *Quantitative Finance*, 2(6):443.

El Aoud, S. and Abergel, F. (2015). A stochastic control approach to option market making. *Market Microstructure and Liquidity*, 1(01):1550006.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007.

Fleming, W. H. and Soner, H. M. (2006). *Controlled Markov processes and viscosity solutions*, volume 25. Springer Science & Business Media.

Fodra, P. and Labadie, M. (2013). High-frequency market-making for multi-dimensional markov processes. *arXiv preprint arXiv:1303.7177*.

Fodra, P. and Pham, H. (2015). High frequency trading and asymptotics for small risk aversion in a markov renewal model. *SIAM Journal on Financial Mathematics*, 6(1):656–684.

Friz, P. K., Gassiat, P., and Pigato, P. (2021). Short-dated smile under rough volatility: asymptotics and numerics. *Quantitative Finance*, pages 1–18.

Garman, M. B. and Klass, M. J. (1980). On the estimation of security price volatilities from historical data. *Journal of Business*, pages 67–78.

Gatheral, J. (2011). *The volatility surface: a practitioner's guide*, volume 357. John Wiley & Sons.

Gatheral, J., Jaisson, T., and Rosenbaum, M. (2018). Volatility is rough. *Quantitative Finance*, 0(0):1–17.

George, T. J. and Longstaff, F. A. (1993). Bid-ask spreads and trading activity in the s&p 100 index options market. *Journal of Financial and Quantitative Analysis*, 28(3):381–397.

Gerhold, S., Kleinert, M., Porkert, P., and Shkolnikov, M. (2015). Small time central limit theorems for semimartingales with applications. *Stochastics*, 87(5):723–746.

Grossman, S. J. and Miller, M. H. (1988). Liquidity and market structure. *the Journal of Finance*, 43(3):617–633.

Guéant, O. (2016). *The Financial Mathematics of Market Liquidity: From optimal execution to market making*, volume 33. CRC Press.

Guéant, O. (2017). Optimal market making. *Applied Mathematical Finance*, 24(2):112–154.

Gueant, O. and Lehalle, C.-A. (2015). General intensity shapes in optimal liquidation. *Mathematical Finance*, 25(3):457–495.

Guéant, O., Lehalle, C.-A., and Fernandez-Tapia, J. (2013). Dealing with inventory risk. a solution to the market making problem. *Mathematics and Financial Economics*, 7(4):477–507.

Guéant, O. and Manziuk, I. (2019). Deep reinforcement learning for market making in corporate bonds: beating the curse of dimensionality. *Applied Mathematical Finance*, 26(5):387–452.

Guilbaud, F. and Pham, H. (2013). Optimal high-frequency trading with limit and market orders. *Quantitative Finance*, 13(1):79–94.

Guilbaud, F. and Pham, H. (2015). Optimal high-frequency trading in a pro rata microstructure with predictive information. *Mathematical Finance*, 25(3):545–575.

Hagan, P. S., Kumar, D., Lesniewski, A. S., and Woodward, D. E. (2002). Managing smile risk. *The Best of Wilmott*, 1:249–296.

Heath, D. and Schweizer, M. (2000). Martingales versus PDEs in finance: an equivalence result with examples. *J. Appl. Probab.*, 37(4):947–957.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6(2):327–343.

Ho, T. and Stoll, H. R. (1981). Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial Economics*, 9(1):47–73.

Ho, T. S. and Stoll, H. R. (1983). The dynamics of dealer markets under competition. *The Journal of Finance*, 38(4):1053–1074.

Hoyer, S. and Hamman, J. (2020). xarray: Nd labeled arrays and datasets in python. Available at `http://xarray.pydata.org`. Version 0.15.0.

Huang, X., Jaimungal, S., and Nourian, M. (2019). Mean-field game strategies for optimal execution. *Applied Mathematical Finance*, 26(2):153–185.

Jusselin, P. (2020). Optimal market making with persistent order flow. *arXiv preprint arXiv:2003.05958*.

Kristensen, D. (2010). Nonparametric filtering of the realized spot volatility: A kernel-based approach. *Econometric Theory*, 26(1):60–93.

Kühn, C. and Muhle-Karbe, J. (2015). Optimal liquidity provision. *Stochastic Processes and their Applications*, 125(7):2493–2515.

Lee, R. W. (2005). Implied volatility: Statics, dynamics, and probabilistic interpretation. In *Recent Advances in Applied Probability*, pages 241–268. Springer.

Livieri, G., Mouti, S., Pallavicini, A., and Rosenbaum, M. (2018). Rough volatility: evidence from option prices. *IISE Transactions*, 50(9):767–776.

Lu, X. and Abergel, F. (2018). Order-book modeling and market making strategies. *Market Microstructure and Liquidity*, 4(01n02):1950003.

Menkveld, A. J. (2013). High frequency trading and the new market makers. *Journal of financial Markets*, 16(4):712–740.

Muhle-Karbe, J. and Nutz, M. (2011). Small-time asymptotics of option prices and first absolute moments. *Journal of Applied Probability*, 48(4):1003–1020.

Øksendal, B. and Sulem, A. (2019). *Applied Stochastic Control of Jump Diffusions*. Springer.

Oomen, R. (2017). Execution in an aggregator. *Quantitative Finance*, 17(3):383–404.

Patel, V., Putniņš, T. J., Michayluk, D., and Foley, S. (2019). Price discovery in stock and options markets. *Journal of Financial Markets*, page 100524.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313.

Ricci, J. (2014). *Applied stochastic control in high frequency and algorithmic trading*. University of Toronto (Canada).

Robert, C. Y. and Rosenbaum, M. (2011). The model with uncertainty zones for ultra high frequency prices and durations: applications to statistical estimation and mathematical finance. In *Econophysics of order-driven markets*, New Econ. Windows, pages 203–224. Springer, Milan.

Ruf, J. (2013). Hedging under arbitrage. *Math. Finance*, 23(2):297–317.

Seabold, S. and Perktold, J. (2020). statsmodels: Econometric and statistical modeling with python. Available at `https://www.statsmodels.org`. Version 0.11.0.

Sheppard, K., Khrapov, S., Lipták, G., mikedeltalima, Capellini, R., Hugle, esvhd, Fortin, A., JPN, Adams, A., jbrockmendel, Rabba, M., Rose, M. E., Rochette, T., RENE-CORAIL, X., and syncoding (2020). bashtage/arch: Release 4.15.

Stoikov, S. and Sağlam, M. (2009). Option market making under inventory risk. *Review of Derivatives Research*, 12(1):55–79.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Szepesvári, C. (2010). Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103.

Vieira, D. (2020). fyne: the Python module for option pricing with affine models. Available at `https://fyne.readthedocs.io`. Version 0.5.4.

Wei, J. and Zheng, J. (2010). Trading activity and bid–ask spreads of individual equity options. *Journal of Banking & Finance*, 34(12):2897–2916.

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Cengage learning.