Received: 13 June 2019 | Accepted: 19 October 2020

DOI: 10.1111/mafi.12295

MATHEMATICAL FINANCE An International Journal of Mathematics, Statistics and Financial Economics

WILEY

ORIGINAL ARTICLE

Optimal make–take fees for market making regulation

Omar El Euch | **Thibaut Mastrolia** | **Mathieu Rosenbaum** | **Nizar Touzi**

CMAP, Ecole Polytechnique, Paris, Palaiseau, France

Correspondence
Thibaut Mastrolia, CMAP, Ecole Polytechnique, Palaiseau, France.
Email: thibaut.mastrolia@polytechnique.edu

Funding information
Chaires *Analytics, and Models for Regulation*, *Financial Risk*, and *Finance, and Sustainable Development*; ERC Grant 679836 Staqamof; ERC Grant 321111 RoFiRM; ANR PACMAN.

Abstract

We address the mechanism design problem of an exchange setting suitable make– take fees to attract liquidity on its platform. Using a principal–agent approach, we provide the optimal compensation scheme of a market maker in quasi-explicit form. This contract depends essentially on the market maker inventory trajectory and on the volatility of the asset. We also provide the optimal quotes that should be displayed by the market maker. The simplicity of our formulas allows us to analyze in details the effects of optimal contracting with an exchange, compared to a situation without contract. We show in particular that it improves liquidity and reduces trading costs for investors. We extend our study to an oligopoly of symmetric exchanges and we study the impact of such common agency policy on the system.

KEYWORDS

financial regulation, high-frequency trading, market making, make–take fees, principal–agent problem, stochastic control

1 | INTRODUCTION

Due to the fragmentation of financial markets, exchanges are nowadays in competition. The traditional international exchanges are challenged by alternative trading venues (see Laruelle & Lehalle, 2018). Consequently, they have to find innovative ways to attract liquidity on their platforms. One solution is to use a maker–taker fees system, that is a rule enabling them to charge in an asymmetric way liquidity provision and liquidity consumption. The most classical setting, used by many exchanges (such as Nasdaq, Euronext, BATS Chi-X, etc.), is of course to subsidize the former while taxing the latter. In practice, this means associating a fee rebate to executed limit orders and applying a transaction cost for market orders.

In the recent years, the topic of make–take fees has been quite controversial. Make–take fees policies are seen as a major facilitating factor to the emergence of a new type of market makers aiming at collecting fee rebates: the high-frequency traders. As stated by the Securities and Exchanges commission in S.E.C. (2010): "Highly automated exchange systems and liquidity rebates have helped establish a business model for a new type of professional liquidity provider that is distinct from the more traditional exchange specialist and over-the-counter market maker." The concern with high-frequency traders becoming the new liquidity providers is twofold. First, their presence implies that slower traders no longer have access to the limit order book, or only in unfavorable situations when high-frequency traders do not wish to support liquidity. This leads to the second classical criticism against high-frequency market makers: they tend to leave the market in time of stress (see Bellia, 2017; Megarbane, Saliba, Lehalle, & Rosenbaum, 2017; Menkveld, 2013; Riordan & Park, 2012, for detailed investigations about high-frequency market making activity).

From an academic viewpoint, studies of make–take fees structures and their impact on the welfare of the markets have been mostly empirical, or carried out in rather stylized models. An interesting theory, suggested in Angel, Harris, and Spatt (2011) and developed in Colliard and Foucault (2012) is that make–take fees have actually no impact on trading costs in the sense that the *cum fee* bid–ask spread should not depend on the make–take fees policy. This result is consistent with the empirical findings in Lutat (2010) and Malinova and Park (2015). Nevertheless, it is clearly shown in these works that many important trading parameters such as depths, volumes, or price impact do depend on the make–take fees structure (see also Harris, 2013). Furthermore, the idea of the neutrality of the make–take fees schedule is also tempered in Foucault, Kadan, and Kandel (2013) where the authors show theoretically that make–take fees may increase welfare of markets provided the tick size is not equal to zero (see also Brolley & Malinova, 2013). More importantly, the results above are obtained in tractable but rather simple discrete-time models that one may want to revisit to be closer to market reality.

In this work, our goal is to provide a quantitative and operational answer to the question of relevant make–take fees. To do so, we take the position of an exchange (or of the regulator) aiming at attracting liquidity. The exchange is looking for the best make–take fees policy to offer to market makers in order to maximize its utility. In other words, it wants to design an optimal contract with the market marker to create an incentive to increase liquidity. We solve the problem the exchange is facing and we do not consider the more involved question of global social welfare. Nevertheless, we have in mind that increasing the liquidity (what the exchange is aiming at) should be beneficial for the welfare of all the agents, what is confirmed in our empirical results. This paper is to our knowledge the first addressing the issue of make–take fees in a realistic continuous-time framework. As a first step, we consider a single market maker in a nonfragmented market, such as, for example, many fixed-income markets which represent some of the most liquid assets in the world. We next consider the case of multiple symmetric exchanges. Actually, our framework can also be interpreted as an extension of the existing works on optimal split of maker–taker fees, in the spirit of Colliard and Foucault (2012) and Foucault et al. (2013). Let c_t and c_m be the take and make fees, respectively. As a matter of fact, for any given constant taker fee c_t, we derive the optimal maker fee c_m. However, we do not restrict ourselves to the case where c_m is a constant, and allow it to be a general contract. It is indeed very natural to consider the possibility of signing a contract between a market maker and an exchange as they are used to do so. For example, obtaining the market maker status on a given exchange implies quantitative trading requirements that are specified in a technical contract between the market maker and the exchange. This is the reason why we allow for a complex fee schedule on the maker side. So for a given c_t, we deduce the optimal c_m. One can then expect some explanations on how to pick the optimal couple (c_m, c_t). However, one of our important results is that from the viewpoint of the exchange, all this couples are equivalent in the sense that the value function of the exchange does not depend on c_t. In other words, the fee c_t paid by the market taker is immediately transferred to the market maker in our framework. This is why we suggest in Section 4.2.3 a choice of c_t based (for example) on the spread the exchange wishes to display on his platform.

Incentive theory has emerged in the 1970s in economics to model how a financial agent can delegate the management of an output process to another agent. Let us recall the formalism of principal–agent problems from the seminal works of Mirrlees (1974) and Hölmstrom (1979). A principal aims at contracting with an agent who provides efforts to manage an output process impacting the wealth of the principal. The principal is not able to control directly the output process since she cannot decide the efforts made by the agent. In our case, the principal is the exchange, the agent is the market maker, the effort corresponds to the quality of the liquidity provided by the market maker (essentially the size of the bid–ask spread proposed by the market maker), the output process is the transactions flow on the platform and the contract depends on the realized transactions flow. Several economics papers have investigated this kind of problems by identifying it with a Stackelberg equilibrium between the two parties. More precisely, since the principal cannot control the work of the agent, she anticipates his best response to a given compensation. We follow the stream of literature initiated in Holmstrom and Milgrom (1987). Then in Sannikov (2008), the author recasts such issue into a stochastic control problem which has been further developed using backward stochastic differential equation theory in Cvitanić, Possamaï, and Touzi (2018). See also Cvitanić and Zhang (2012) for related literature.

This paper provides a quasi-explicit expression for the optimal contract between the exchange and the market maker, and for the market maker optimal quotes. The optimal contract depends essentially on the market maker inventory trajectory and on the volatility of the market. These simple formulas enable us to analyze in details the effects for the welfare of the market of optimal contracting with an exchange, compared to a situation without contract as in Avellaneda and Stoikov (2008) and Guéant, Lehalle, and Fernandez-Tapia (2008). We show that such contracts lead to reduced spreads and lower trading costs for investors. We also propose an extension of this work to an oligopoly of symmetric exchanges aiming at hiring a single market marker. We show in particular that there exists a unique Markovian symmetric Nash equilibrium for the exchanges.

The paper is organized as follows. Our modeling approach is presented in Section 2. In particular, we define the market maker's as well as the exchange's optimization framework. In Section 3, we compute the best response of the market maker for a given contract. Optimal contracts are designed in Section 4 where we solve the exchange's problem. Then, in Section 5, we assess the benefits for market quality of the presence of an exchange contracting optimally with a market maker. In Section 6, we extend our study to an oligopoly of symmetric exchanges. Finally, useful technical results are gathered in an appendix together with the so-called first-best case (see Appendix A.7) which provides different solutions.

2 | THE MODEL

Our starting point is the seminal work of Avellaneda and Stoikov (2008). Our objective is to derive optimal make–take fees in order to monitor the behavior of a market maker on a platform acting according to the optimal market making model of Avellaneda and Stoikov (2008).

2.1 | Contractible and observable variables

Let $T > 0$ be a final horizon time, Ω_c the set of continuous functions from $[0,T]$ into $\mathbb{R}$, Ω_d the set of piecewise constant càdlàg functions from $[0,T]$ into $\mathbb{N}$, and $\Omega = \Omega_c \times (\Omega_d)^2$ with corresponding Borel algebra $\mathcal{F}$. The observable state is the canonical process $(\chi_t)_{t\in[0,T]} = (S_t, N_t^a, N_t^b)_{t\in[0,T]}$ of the measurable space $(\Omega, \mathcal{F})$:

$$S_t(\omega) := s(t),\;\; N_t^a(\omega) := n^a(t),\;\; N_t^b(\omega) := n^b(t), \quad \text{for all } t \in [0,T],\; \omega = (s, n^a, n^b) \in \Omega,$$

with canonical completed filtration $\mathbb{F} = (\mathcal{F}_t)_{t\in[0,T]} = (\mathcal{F}_t^c \otimes (\mathcal{F}_t^d)^{\otimes 2})_{t\in[0,T]}$.

The trading activity is reduced to a single risky asset with observable efficient price S defined by[1]

$$S_t := S_0 + \sigma W_t, \quad t \in [0,T] \tag{1}$$

for some Brownian motion W, initial price $S_0 > 0$, and constant volatility $\sigma > 0$. The market maker chooses processes denoted by δ^b and δ^a, respectively, so as to fix publicly available bid and ask offer prices:

$$P_t^b := S_t - \delta_t^b \quad \text{and} \quad P_t^a := S_t + \delta_t^a,\;\; t \in [0,T].$$

The arrival of bid and ask market orders is modeled by a counting process (N^b, N^a) with unit jumps, so that no more than one market order can occur at each time. We introduce the inventory process of the market maker Q:

$$Q_t := N_t^b - N_t^a \;\in\; \mathbb{N} \cap [-\bar{q}, \bar{q}],\; t \in [0,T],$$

where $N_0^b = N_0^a = 0$ and, as in Guéant et al. (2008), we impose a critical absolute inventory $\bar{q} \in \mathbb{N}$ above which the market maker stops quoting on the ask or bid side.

Let $c > 0$ be the fee collected by the exchange (see Section 2.3). In order to illustrate the impact of the posted prices on the transactions arrival process (N^b, N^a), the corresponding intensity process depends on the departure of the transaction price from the efficient price, that is, $c + \delta_t^i$, $i \in \{b, a\}$, as follows:

$$\lambda_t^{i,\delta} := \lambda(\delta_t^i)\mathbb{1}_{\{\varepsilon_i Q_t > -\bar{q}\}},\; i \in \{b, a\},\;\; (\varepsilon_b, \varepsilon_a) = (-1, 1), \text{ with } \lambda(x) := Ae^{-k\frac{(x+c)}{\sigma}} \tag{2}$$

for some fixed positive constants A and k. In most of the literature (see, for instance, Avellaneda & Stoikov, 2008; Cartea, Jaimungal, & Penalva, 2015; Guéant et al., 2008) order flows are modeled using Poisson processes. This is mainly for technical reasons: they are the easiest nontrivial arrival time processes to work with from a mathematical point of view. For us, it is crucial to have a reasonable modeling of the connection between the behavior of the market maker and that of the arrival flow so that a large spread means that quoted prices by the market maker are far from the efficient price. Hence, they are not very attractive for the market taker. That is why we consider the intensity of the order flow decreasing with respect to the spread. This simple intuition can be made more rigorous from solid financial economics arguments as in Madhavan, Richardson, and Roomans (1997) for example. The dependence on the volatility parameter σ reproduces the stylized fact that the average number of trades per unit of time is a decreasing function of the ratio between spread and volatility (see Dayri & Rosenbaum, 2015; Madhavan et al., 1997; Wyart, Bouchaud, Kockelkoren, Potters, & Vettorazzo, 2008).

Our canonical variables being S, N^a, and N^b, the contracts are allowed to depend on the trajectories of these quantities only: these are our three contractible variables. This is actually very reasonable: the efficient price is a quantity any market participant is used to, whether the chosen proxy for it is the midprice, the last traded price or some volume weighted price. The processes N^a and N^b encode the arrivals of market orders and therefore actual transactions, which are clearly recorded on any exchange and accessible to most participants. So the contracts will be designed on standard, unarguable, and easily obtained financial variables.

Note that the spreads δ^a and δ^b are here observable by market takers, but not contractible. From the exchange viewpoint, it would not be reasonable to introduce the spread variable in a contract. First, quotes are typically not recorded with the same degree of accuracy as transactions since they are evolving on a much higher frequency, which can sometimes be of the order of the millisecond. Second, quoted prices do not in general lead to transactions and it would be probably hard to justify taxing or subsidizing agents based on offer prices having no tangible counterparts. Finally, a quote-based contract would certainly encourage high-frequency traders to attempt exploiting possible flaws in the contract, using, for example, a very high rate of cancellations of their orders, leading to possible market disruptions (see Abergel, Lehalle, & Rosenbaum, 2014; Megarbane et al., 2017, for studies about high-frequency traders behavior).

2.2 | Admissible controls and market maker's problem

The set of admissible controls $\mathcal{A}$ is the collection of all predictable processes $\delta = (\delta^a, \delta^b)$ uniformly bounded by δ_∞, some sufficiently large positive constant to be fixed later. Controlling δ is equivalent to control the arrival intensity of market orders since it is a deterministic function of the spread. Viewing the market maker optimization problem this way, we see that the intensity plays the very same role as the drift in standard principal–agent problems where the agent controls the drift of a diffusion process, this drift being unobserved by the principal (see Sannikov, 2008; Cvitanić et al., 2018). A particular feature of our modeling is that the intensity is observable (although noncontractible) because of its connection with the spread. However, the spread is in some sense an artificial variable here enabling us to fit with market reality (the market maker has access in practice to the spread but not to the intensity of market takers arrivals). Each control process δ induces

- the market maker profit and loss (P&L) process:

$$\mathrm{PL}_t^\delta \;:=\; X_t^\delta + Q_t S_t, \text{ where } X_t^\delta \;:=\; \int_0^t P_r^a dN_r^a - \int_0^t P_r^b dN_r^b,\;\; t \in [0,T] \tag{3}$$

  as the sum of the cash flow X^δ and the inventory risk[2] QS;
- and a probability measure $\mathbb{P}^\delta$ under which S is driven by (1), and

$$\widetilde{N}_t^{i,\delta} \;:= N_t^i - \int_0^t \lambda_r^{i,\delta} dr,\; t \in [0,T],\;\; i \in \{b,a\}, \text{ are martingales.}$$

Then, $\mathbb{P}^\delta$ is defined by the density $\frac{d\mathbb{P}^\delta}{d\mathbb{P}^0}|_{\mathcal{F}_T} = L_T^\delta$, induced by the Doléans-Dade exponential martingale $dL_t^\delta = L_{t-}^\delta \sum_{i=b,a} \mathbb{1}_{\{\varepsilon_i Q_{t-} > -\bar{q}\}} \frac{\lambda(\delta_t^i)-A}{A} d\widetilde{N}_t^{i,0}$, that is,[3] again with $(\varepsilon_b, \varepsilon_a) = (-1, 1)$,

$$L_t^\delta := \exp \sum_{i=b,a} \int_0^t \mathbb{1}_{\{\varepsilon_i Q_{r-} > -\bar{q}\}} \left[\log\left(\frac{\lambda(\delta_r^i)}{A}\right) dN_r^i - (\lambda(\delta_r^i) - A)dr \right], \quad t \in [0, T]. \tag{4}$$

In particular, all probability measures $\mathbb{P}^\delta$ are equivalent. We therefore use the notation $a.s$ for almost surely without ambiguity. We shall write $\mathbb{E}_t^\delta$ for the conditional expectation with respect to $\mathcal{F}_t$ with probability measure $\mathbb{P}^\delta$.

The exchange aims at encouraging the market maker to reduce his spread so as to enhance market liquidity on the platform. This is achieved by setting the terms of an incentive contract defined by an $\mathcal{F}_T$-measurable random variable ξ. In other words, the compensation ξ may depend on the whole paths of the contractible variables N^a, N^b, and S. Given this additional revenue, the market maker's objective is defined by the utility maximization problem

$$\begin{aligned} V_{\mathrm{MM}}(\xi) \; &:= \; \sup_{\delta \in \mathcal{A}} J_{\mathrm{MM}}(\delta, \xi), \;\text{ where }\; J_{\mathrm{MM}}(\delta, \xi) := \mathbb{E}^\delta\left[-e^{-\gamma(\xi + \mathrm{PL}_T^\delta)}\right] \\ &= \mathbb{E}^\delta\left[-e^{-\gamma(\xi + \int_0^T \delta_t^a dN_t^a + \delta_t^b dN_t^b + Q_t dS_t)}\right]. \end{aligned} \tag{5}$$

Here, $\gamma > 0$ is the absolute risk aversion parameter of the CARA market maker. For each compensation ξ, we shall prove below that there exists a unique optimal response $\hat{\delta}(\xi) = (\hat{\delta}^b(\xi), \hat{\delta}^a(\xi)) \in \mathcal{A}$ of the market marker, that is, $V_{\mathrm{MM}}(\xi) = J_{\mathrm{MM}}(\hat{\delta}(\xi), \xi)$.

Remark 2.1. When there is no incentive payment $\xi = 0$, the utility maximization problem (5) reduces to the Avellaneda and Stoikov (2008) and Guéant et al. (2008) optimal market making problem.

2.3 | The exchange optimal contracting problem

The exchange receives a fixed fee $c > 0$ for each market order that occurs in the market,[4] and then collects at time T the total revenue $c(N_T^a + N_T^b) - \xi$. The choice of the contract ξ is dictated by the utility maximization problem

$$V_0^E := \sup_{\xi \in C} \mathbb{E}^{\hat{\delta}(\xi)}\left[-e^{-\eta(c(N_T^a + N_T^b) - \xi)}\right], \tag{6}$$

where $\eta > 0$ is the exchange's absolute risk aversion parameter, and the set of admissible contracts C is the collection of all contracts satisfying:

- the participation constraint $V_{\mathrm{MM}}(\xi) \geq R$, where the reservation level $R < 0$ may be chosen to be the utility level without contract,

– together with the integrability conditions:

$$\sup_{\delta\in\mathcal{A}} \mathbb{E}^{\delta}\left[e^{\eta'\xi}\right] < \infty \text{ and } \sup_{\delta\in\mathcal{A}} \mathbb{E}^{\delta}\left[e^{-\gamma'\xi}\right] < \infty, \quad \text{for some } \eta' > \eta,\ \gamma' > \gamma. \tag{7}$$

Since N^a and N^b are point processes with bounded intensities, and the inventory process Q is bounded, it follows from an easy application of the Hölder inequality that the expectations in both problems (5) and (6) are finite.

We assume throughout this paper that the participation level R is so that the set of admissible contracts is nonempty:

$$\mathcal{C} = \{\xi,\ \mathcal{F}_T - \text{measurable such that } V_{\mathrm{MM}}(\xi) \geq R\ and (7) is satisfied\} \neq \emptyset.$$

3 | SOLVING THE MARKET MAKER'S PROBLEM

We start by solving the problem (5) of the market maker facing an arbitrary contract $\xi \in \mathcal{C}$ proposed by the exchange.

3.1 | Market maker's optimal response

For $(\delta, z, q) \in [-\delta_\infty, \delta_\infty]^2 \times \mathbb{R}^3 \times \mathbb{Z}$, with $\delta = (\delta^a, \delta^b)$ and $z = (z^S, z^a, z^b)$, we define

$$h(\delta, z, q) := \sum_{i=b,a} \frac{1 - e^{-\gamma(z^i+\delta^i)}}{\gamma} \lambda(\delta^i)\mathbb{1}_{\{\varepsilon_i q > -\bar{q}\}} \text{ and } H(z, q) := \sup_{|\delta^a|\vee|\delta^b|\leq\delta_\infty} h(\delta, z, q)$$

with $(\varepsilon_b, \varepsilon_a) = (-1, 1)$. For an arbitrary constant $Y_0 \in \mathbb{R}$ and predictable processes $Z = (Z^S, Z^a, Z^b)$, with $\int_0^T (|Z_t^S|^2 + |H(Z_t, Q_t)|)dt < \infty$, we introduce the process

$$Y_t^{Y_0,Z} \;=\; Y_0 + \int_0^t Z_r^a dN_r^a + Z_r^b dN_r^b + Z_r^S dS_r + \left(\frac{1}{2}\gamma\sigma^2(Z_r^S + Q_r)^2 - H(Z_r, Q_r)\right)dr, \tag{8}$$

and we denote by $\mathcal{Z}$ the collection of all such processes Z such that the first integrability condition in (7) is satisfied with $\xi = Y_T^{0,Z}$ and

$$\sup_{\delta\in\mathcal{A}} \sup_{t\in[0,T]} \mathbb{E}^{\delta}[e^{-\gamma' Y_t^{0,Z}}] < \infty, \quad \text{for some } \gamma' > \gamma. \tag{9}$$

Clearly, $\mathcal{Z} \neq \emptyset$ as it contains all bounded predictable processes and

$$\mathcal{C} \supset \Xi \;:=\; \left\{Y_T^{Y_0,Z} \;:\; Y_0 \in \mathbb{R},\ Z \in \mathcal{Z}, \text{ and } V_{\mathrm{MM}}(Y_T^{Y_0,Z}) \geq R\right\}.$$

The next result shows that these sets are in fact equal, and identifies the market maker utility value and the corresponding optimal response. To prove equality of these sets, we are reduced to the problem of representing any contract $\xi \in \mathcal{C}$ as $\xi = Y_T^{Y_0,Z}$ for some $(Y_0, Z) \in \mathbb{R} \times \mathcal{Z}$, which is known in the literature as a problem of backward stochastic differential equation. We refrain from using this terminology, as our analysis does not require any result from this literature.

Theorem 3.1.

(i) Any contract $\xi \in \mathcal{C}$ has a unique representation as $\xi = Y_T^{Y_0,Z}$, for some $(Y_0, Z) \in \mathbb{R} \times \mathcal{Z}$. In particular, $\mathcal{C} = \Xi$.

(ii) Under this representation, the market maker utility value is

$$V_{\mathrm{MM}}(\xi) = -e^{-\gamma Y_0}, \text{ so that } \Xi \ = \ \left\{ Y_T^{Y_0,Z} \ : \ Z \in \mathcal{Z}, \text{ and } Y_0 \geq \hat{Y}_0 \right\}, \ \hat{Y}_0 := -\frac{1}{\gamma} \log(-R),$$

with the following optimal bid–ask policy:

$$\hat{\delta}_t^i(\xi) = \Delta(Z_t^i), \ i \in \{b, a\}, \text{ where } \Delta(z) := (-\delta_\infty) \vee \left\{ -z + \frac{1}{\gamma} \log\left(1 + \frac{\sigma\gamma}{k}\right) \right\} \wedge \delta_\infty. \tag{10}$$

The proof of Part (i) is reported in Section A.4, and is obtained by using the dynamic continuation utility process of the market maker, following the approach of Sannikov (2008).

Proof of Theorem 3.1 (ii). Let $\xi = Y_T^{Y_0,Z}$ with $(Y_0, Z) \in \mathbb{R} \times \mathcal{Z}$. We first prove that $J_{\mathrm{MM}}(\delta, \xi) \leq -e^{-\gamma Y_0}$ for all $\delta \in \mathcal{A}$. Denote $\overline{Y}_t := Y_t^{Y_0,Z} + \sum_{i=a,b} \int_0^t \delta_t^i dN_t^i + Q_t dS_t$. Setting $h^\delta := h(\delta, .)$, it follows from Itô's formula that

$$de^{-\gamma \overline{Y}_t} = \gamma e^{-\gamma \overline{Y}_{t-}} \left[-(Q_t + Z_t^S) dS_t - \sum_{i=b,a} \frac{1 - e^{-\gamma(Z_t^i + \delta_t^i)}}{\gamma} d\widetilde{N}_t^{i,\delta} + \left(H - h^{\delta_t}\right)(Z_t, Q_t) dt \right]$$

implying that $e^{-\gamma \overline{Y}}$ is a $\mathbb{P}^\delta$-local submartingale. By Condition (9), the uniform boundedness of the intensities of N^a and N^b and Hölder inequality, $(e^{-\gamma \overline{Y}_t})_{t \in [0,T]}$ is uniformly integrable. By Doob–Meyer decomposition theorem, we conclude that $\int_0^\cdot \gamma e^{-\gamma \overline{Y}_{t-}} (-(Q_t + Z_t^S) dS_t - \sum_{i=b,a} \frac{1 - e^{-\gamma(Z_t^i + \delta_t^i)}}{\gamma} d\widetilde{N}_t^{i,\delta})$, is a martingale. It follows that

$$J_{\mathrm{MM}}(\delta, \xi) = \mathbb{E}^\delta \left[-e^{-\gamma \overline{Y}_T} \right] = -e^{-\gamma Y_0} - \mathbb{E}^\delta \left[\int_0^T \gamma e^{-\gamma \overline{Y}_t} (H(Z_t, Q_t) - h(\delta_t, Z_t, Q_t)) dt \right] \leq -e^{-\gamma Y_0}.$$

On the other hand, equality holds in the last inequality if and only if δ is chosen as the maximizer of the Hamiltonian H ($dt \times d\mathbb{P}^0$–a.e.), thus leading to the unique maximizer $\hat{\delta}(\xi)$ given by (10), which then induces $J_{\mathrm{MM}}(\hat{\delta}(\xi), \xi) = -e^{-\gamma Y_0}$. This completes the proof that $V_{\mathrm{MM}}(\xi) = -e^{-\gamma Y_0}$ with optimal response $\hat{\delta}(\xi)$. □

4 | DESIGNING THE OPTIMAL CONTRACT

4.1 | Risk-neutral exchange as a toy example

To understand the shape of the optimal contract, we first study the case where the exchange is risk-neutral, corresponding to the limit where η goes to 0 for which we can derive the optimal compensation with explicit computations. In the present setting, we set $\overline{q} = +\infty$, thus relaxing the boundedness restriction on the inventory. By Theorem 3.1, the problem of the exchange reduces to

$$V_0^E = \sup_{Y_0 \geq \hat{Y}_0} \sup_{Z \in \mathcal{Z}} \mathbb{E}^{\hat{\delta}(Y_T^{Y_0,Z})}\left[c\left(N_T^a + N_T^b\right) - Y_T^{Y_0,Z}\right] = \sup_{Z \in \mathcal{Z}} \mathbb{E}^{\hat{\delta}\left(Y_T^{\hat{Y}_0,Z}\right)}\left[c\left(N_T^a + N_T^b\right) - Y_T^{\hat{Y}_0,Z}\right] \tag{11}$$

with $\hat{\delta}_t^i = \Delta(Z_t^i)$, $t \in [0, T]$, $i \in \{a, b\}$, and where the maximization over Y_0 is achieved at $\hat{Y}_0$, due to the fact that the market maker optimal response $\hat{\delta}(Y_T^{Y_0,Z})$, given by (10), does not depend on Y_0 so that the objective function is decreasing in Y_0.

Theorem 4.1. *Consider the risk-neutral exchange case $\eta \searrow 0$, and assume $\delta_\infty \geq \frac{\sigma}{k} - \frac{\sigma}{k+\sigma\gamma} + \frac{1}{\gamma}\log\left(1 + \frac{\sigma\gamma}{k}\right) - c$. Then the optimal contract for the exchange problem (11) is*

$$\hat{\xi} = \hat{Y}_0 + c\left(N_T^a + N_T^b\right) - \int_0^T Q_r dS_r - \frac{\sigma T}{k + \sigma\gamma}, \tag{12}$$

with optimal market maker effort:

$$\hat{\delta}_t^a(\hat{\xi}) = \hat{\delta}_t^b(\hat{\xi}) = \frac{\sigma}{k} - \frac{\sigma}{k + \sigma\gamma} + \frac{1}{\gamma}\log\left(1 + \frac{\sigma\gamma}{k}\right) - c.$$

Proof. By setting $\tilde{c} = c + \frac{\sigma}{k+\sigma\gamma}$, note that

$$\mathbb{E}^{\hat{\delta}\left(Y_T^{\hat{Y}_0,Z}\right)}\left[c\sum_{i=b,a} N_T^i - Y_T^{\hat{Y}_0,Z}\right] = \mathbb{E}^{\hat{\delta}\left(Y_T^{\hat{Y}_0,Z}\right)}\left[\int_0^T \sum_{i=b,a} \lambda\left(\hat{\delta}_t^i\right)\left(\tilde{c} - Z_t^i\right) dt - \hat{Y}_0 - \int_0^T \left(\frac{1}{2}\gamma\sigma^2\left(Z_r^S + Q_r\right)^2\right) dr\right],$$

so that the optimizer in (11) are given by $Z_r^{S,\star} = -Q_r,\ Z_r^{a,\star} = Z_r^{b,\star} = \tilde{c} - \frac{\sigma}{k}$. □

Note that the optimal contract given by (12) emphasizes a risk transfer between the payoff of the market maker and that of the exchange through the term $\int_0^T Q_r dS_r$.

4.2 | Exponential risk-averse exchange

By Theorem 3.1, and solving the maximization with respect to $Y_0 \geq \hat{Y}_0$ as in the previous subsection, the exchange problem (6) reduces to

$$V_0^E = e^{\eta \hat{Y}_0}\, v_0^E, \text{ where } v_0^E := \sup_{Z \in \mathcal{Z}} \mathbb{E}^{\hat{\delta}(Y_T^{\hat{Y}_0,Z})}\left[-e^{-\eta\left(c(N_T^a + N_T^b) - Y_T^{0,Z}\right)}\right]. \tag{13}$$

4.2.1 | The HJB equation for the reduced exchange problem

Our approach for the control problem v_0^E of (13) is to derive a solution v of the corresponding Hamilton-Jacobi-Bellman equation (HJB equation), and to proceed by the standard verification argument in stochastic control to prove that the proposed solution v coincides with the value function v_0^E.

Since all the coefficients in (13) do not depend on S, we guess[5] that the HJB equation associated to (13) reduces to

$$\partial_t v(t,q) + H_E(q, v(t,q), v(t,q+1), v(t,q-1)) = 0, \quad q \in \{-\bar{q}, \dots, \bar{q}\}, \quad t \in [0,T), \tag{14}$$

with boundary condition $v|_{t=T} = -1$, with Hamiltonian $H_E : [-\bar{q}, \bar{q}] \times (-\infty, 0]^3 \to \mathbb{R}$:

$$H_E(q, y, y_+, y_-) = H_E^1(q,y) + \mathbb{1}_{\{q>-\bar{q}\}} H_E^0(y, y_-) + \mathbb{1}_{\{q<\bar{q}\}} H_E^0(y, y_+), \tag{15}$$

and

$$H_E^1(q,y) = \sup_{z_s \in \mathbb{R}} h_E^1(q,y,z_s) \text{ and } h_E^1(q,y,z_s) = \frac{\eta\sigma^2}{2}\, y\big(\gamma(z_s+q)^2 + \eta z_s^2\big),$$

$$H_E^0(y,y') = \sup_{\zeta \in \mathbb{R}} h_E^0(y,y',\zeta) \text{ and } h_E^0(y,y',\zeta) = \lambda(\Delta(\zeta))\left[y' e^{\eta(\zeta-c)} - y\left(1 + \eta\,\frac{1-e^{-\gamma(\zeta+\Delta(\zeta))}}{\gamma}\right)\right].$$

By Lemma A.2, the maximizers $\hat{z} = (\hat{z}^s, \hat{z}^a, \hat{z}^b)$ of H_E are given by

$$\hat{z}^s(t,q) = -\tfrac{\gamma}{\gamma+\eta} q, \;\; \hat{z}^a(t,q) = \hat{\zeta}(v(t,q), v(t,q-1)), \;\; \hat{z}^b(t,q) = \hat{\zeta}(v(t,q), v(t,q+1)), \tag{16}$$

$$\text{with } \hat{\zeta}(y,y') = \zeta_0 + \frac{1}{\eta}\log\left(\frac{y}{y'}\right), \;\; \zeta_0 = c + \frac{1}{\eta}\log\left(1 - \frac{\sigma^2\gamma\eta}{(k+\sigma\gamma)(k+\sigma\eta)}\right).$$

Here, δ_∞ is sufficiently large so that Condition (A.1) of Lemma A.2 is always met, namely,

$$\delta_\infty \;\geq\; C_\infty + \frac{1}{\eta} \sup_{t\in[0,T]} \sup_{q\in[-\bar{q},\bar{q}-1]} \left|\log\left(\frac{v(t,q)}{v(t,q+1)}\right)\right|, \tag{17}$$

with C_∞ given in Lemma A.2, and we shall check in our verification argument that our candidate solution of the HJB equation will verify it. Using again the calculation reported in Lemma A.2, we rewrite the HJB equation (14) as

$$\partial_t v(t,q) + \frac{\gamma\eta^2\sigma^2}{2(\gamma+\eta)} q^2 v(t,q) - C_0 v(t,q)\left[\mathbb{1}_{\{q>-\bar{q}\}}\left(\frac{v(t,q)}{v(t,q-1)}\right)^{\frac{k}{\sigma\eta}} + \mathbb{1}_{\{q<\bar{q}\}}\left(\frac{v(t,q)}{v(t,q+1)}\right)^{\frac{k}{\sigma\eta}}\right] = 0, \tag{18}$$

with boundary condition $v|_{t=T} = -1$, where the constant C_0 is given by

$$C_0 = C_0\left(\frac{\sigma\gamma}{k}, \frac{\sigma\eta}{k}\right), \text{ with } C_0(\alpha,\beta) := A\beta(1+\alpha)^{-\frac{1}{\alpha}}\left(1 - \frac{\alpha\beta}{(1+\alpha)(1+\beta)}\right)^{1+\frac{1}{\beta}}.$$

Inspired by Guéant et al. (2008), we now make the key observation that this equation can be reduced to a linear equation by introducing $u := (-v)^{-\frac{k}{\sigma\eta}}$. By direct substitution, we obtain the following linear differential equation:

$$u\Big|_{t=T} = 1, \text{ and } \partial_t u(t,q) - F_{C_1,C_1'}(q, u(t,q), u(t,q+1), u(t,q-1)) = 0,\ t < T,\ |q| \le \bar{q}, \tag{19}$$

$$F_{m,m'}(q,y,y',y'') := mq^2 y - m'\left(y'\mathbb{1}_{\{q<\bar{q}\}} + y''\mathbb{1}_{\{q>-\bar{q}\}}\right),\ C_1 := \frac{k\gamma\eta\sigma}{2(\gamma+\eta)},\ C_1' := \frac{kC_0}{\sigma\eta}.$$

This equation can be written in terms of the $\mathbb{R}^{2\bar{q}+1}$-valued function $\mathbf{u}(t) = (u(t,q))_{q\in\{-\bar{q},\dots,\bar{q}\}}$, of the variable t only, as the linear ordinary differential equation

$$\partial_t \mathbf{u} = -\mathbf{B}u, \text{ where } \mathbf{B} = \begin{pmatrix} -C_1\bar{q}^2 & C_1' & & & \\ \ddots & \ddots & \ddots & & \\ & C_1' & -C_1 q^2 & C_1' & \\ & & \ddots & \ddots & \ddots \\ & & & C_1' & -C_1\bar{q}^2 \end{pmatrix} \begin{matrix} \\ \\ \leftarrow q\text{th line,} \\ \\ \\ \end{matrix}$$

is a tridiagonal matrix with lines labeled $-\bar{q},\dots,\bar{q}$. Denote by $\mathbf{b}_q$ the vector of $\mathbb{R}^{2\bar{q}+1}$ with zeros everywhere except at the position q, that is, $\mathbf{b}_{q,j} = \mathbb{1}_{\{j=q\}}$ for $j \in \{-\bar{q},\dots,\bar{q}\}$, and $\mathbf{1} = \sum_{q=-\bar{q}}^{\bar{q}} \mathbf{b}_q$. Then, this Ordinary Differential Equation (ODE) has a unique solution

$$\mathbf{u}(t) = e^{(T-t)\mathbf{B}}\mathbf{1}, \text{ so that } u(t,q) = \mathbf{b}_q \cdot e^{(T-t)\mathbf{B}}\mathbf{1}, \text{ and } v(t,q) = -\left(\mathbf{b}_q \cdot e^{(T-t)\mathbf{B}}\mathbf{1}\right)^{-\frac{\sigma\eta}{k}}. \tag{20}$$

In the next section, we shall prove that this solution v of the HJB equation (14) coincides with the value function of the reduced exchange problem (13), with optimal controls $\hat{z}(t,q)$ given in (16), thus inducing the optimal contract $Y_T^{\hat{Y}_0,\hat{Z}}$ with $\hat{Z}_t = \hat{z}(t, Q_{t-})$.

Let us notice that we may provide a more explicit expression of the above function u:

$$u(t,q) \;=\; \sum_{p\ge 0} \frac{\left[C_1'(T-t)\right]^p}{p!} \sum_{j\ge 0} \frac{\left[C_1'(T-t)\right]^j}{j!} e^{-C_1(T-t)(q+j-p)^2} \mathbb{1}_{\{|q+j-p|\le\bar{q}\}}, \tag{21}$$

see Appendix A.3 for the more general case of N symmetric exchanges in Nash equilibrium. We conclude this section by an (yet one more) alternative representation of the function u, which is convenient for the derivation of some useful properties.

Proposition 4.2. *Let u and v be defined by (20). The function u can be represented as*

$$u(t,q) = \mathbb{E}\left[e^{\int_t^T (-C_1(Q_s^{t,q})^2 + \overline{\lambda}_s + \underline{\lambda}_s)ds}\right],$$

where $Q_s^{t,q} = q + \int_t^s d(\overline{N}_u - \underline{N}_u)$, and $(\overline{N}, \underline{N})$ is a two-dimensional point process with intensity $(\overline{\lambda}_s, \underline{\lambda}_s) = C_1'(\mathbb{1}_{\{Q_{s-}<\bar{q}\}}, \mathbb{1}_{\{Q_{s-}>-\bar{q}\}})$. In particular, we have $e^{-C_1\bar{q}^2T} \le u \le e^{2C_1'T}$, and Condition (17) *is verified whenever*

$$\delta_\infty \;\geq\; \Delta_\infty := C_\infty + \frac{\sigma}{k}(2C'_1 + C_1\bar{q}^2)T. \tag{22}$$

Proof. Notice that u is a smooth bounded function. Denote $f(x) = -C_1x^2 + C'_1(\mathbb{1}_{\{x>-\bar{q}\}} + \mathbb{1}_{\{x<\bar{q}\}})$, and $M_s = e^{\int_t^s f(Q_u^{t,q})du}u(s, Q_s^{t,q})$, $t \leq s \leq T$. We now show that M is a martingale, so that $u(t,q) = M_t = \mathbb{E}[M_T] = \mathbb{E}[e^{-\int_t^T f(Q_s^{t,q})ds}]$, as $u(T,.) = 1$. To see that M is a martingale, we compute by Itô's formula that

$$dM_s = \Big[u(s,Q_s^{t,q})f(Q_s^{t,q}) + \partial_t u(s,Q_s^{t,q})\Big]ds$$
$$+C'_1\Big[u(s,Q_{s-}^{t,q}+1) - u(s,Q_{s-}^{t,q})\Big]d\overline{N}_s + C'_1\Big[u(s,Q_{s-}^{t,q}-1) - u(s,Q_{s-}^{t,q})\Big]d\underline{N}_s.$$

Since u is solution of (19), we get

$$dM_s = C'_1\Big[u(s,Q_{s-}^{t,q}+1) - u(s,Q_{s-}^{t,q})\Big]d\overline{M}_s + C'_1\Big[u(s,Q_{s-}^{t,q}-1) - u(s,Q_{s-}^{t,q})\Big]d\underline{M}_s,$$

where $(\overline{M}, \underline{M}) = (\overline{N} - \int_0^\cdot \overline{\lambda}_s ds, \underline{N} - \int_0^\cdot \underline{\lambda}_s ds)$ is a martingale. The martingale property of M now follows from the boundedness of u as it can be verified from the expression (20). Finally, the bound $|Q_s^{t,q}| \leq \bar{q}$ induces directly the announced bounds on u, which, in turn, imply Condition (17) when (22) is satisfied because $v = -u^{-\frac{\sigma\eta}{k}}$. □

4.2.2 | Main result

We now verify that the function v derived in the previous section is the value function of the exchange, with optimal feedback controls $(\hat{z}^s, \hat{z}^a, \hat{z}^b)$ as given in (16), thus identifying a unique optimal contract to be proposed by the exchange to the market maker. The proof of this theorem is postponed in Appendix A.5.

Theorem 4.3. *Assume that $\delta_\infty \geq \Delta_\infty$, with Δ_∞ given by (22) and define u and v by (20). Then the optimal contract for the problem of the exchange (6) is given by*

$$\hat{\xi} = \hat{Y}_0 + \int_0^T \hat{Z}_r^a dN_r^a + \hat{Z}_r^b dN_r^b + \hat{Z}_r^S dS_r + \left(\frac{1}{2}\gamma\sigma^2\big(\hat{Z}_r^S + Q_r\big)^2 - H\big(\hat{Z}_r, Q_r\big)\right)dr, \tag{23}$$

with $\hat{Z}_r^S = \hat{z}^s(r, Q_{r-})$, $\hat{Z}_r^a = \hat{z}^a(r, Q_{r-})$, and $\hat{Z}_r^b = \hat{z}^b(r, Q_{r-})$ as defined in (16). The market maker's optimal effort is given by

$$\hat{\delta}_t^a = \hat{\delta}_t^a(\hat{\xi}) = -\hat{Z}_t^a + \frac{1}{\gamma}\log\left(1 + \frac{\sigma\gamma}{k}\right), \quad \hat{\delta}_t^b = \hat{\delta}_t^b(\hat{\xi}) = -\hat{Z}_t^b + \frac{1}{\gamma}\log\left(1 + \frac{\sigma\gamma}{k}\right). \tag{24}$$

Moreover, the value function of the exchange is $V_0^E = v(0, Q_0)$ and does not depend on c.

Remark 4.4. Notice that in our model the exchange observes the spread set by the market maker. However, as explained above, the spread cannot be part of the contract. Consequently, the second best exchange problem in Theorem 4.3 does not coincide with the first best where the exchange could use the observe bid–ask policy δ in the contract ξ, under the market maker participation constraint. The corresponding computations are reported in Appendix A.7.

4.2.3 | Discussions and interpretations

The processes $\hat{Z}^a$, $\hat{Z}^b$, and $\hat{Z}^S$ defining the optimal contract have natural interpretations. Based on Proposition 4.2, and since $\hat{Z}^i = \zeta_0 + \frac{1}{\eta}\log\left(\frac{v(t,Q_{t-})}{v(t,Q_{t-}-\varepsilon_i)}\right)$, we can get the intuition that (at least for large inventories)

$$\hat{Z}^i = \zeta_0 - \frac{\sigma}{k}\log\left(\frac{u(t,Q_{t-})}{u(t,Q_{t-}-\varepsilon_i)}\right) \underset{|q|\to+\infty}{\sim} \zeta_0 + \varepsilon_i\frac{\sigma}{2k}\sqrt{\frac{C_1}{C_1'}}(2Q_{t-}-\varepsilon_i),\ i\in\{a,b\}, \tag{25}$$

recalling that $(\varepsilon_b,\varepsilon_a) = (-1,1)$. This is confirmed in our Figure 4 at time $t = 0$ (since $\hat{Z}^b$ and $\hat{Z}^a$ are the opposite of the optimal bid and ask spreads, respectively). This is in fact shown for any time in the numerical simulations and asymptotic expansion in Guéant et al. (2008, section 4) and Avellaneda and Stoikov (2008, section 3.2) where same type of Partial Differential Equation (PDE) as ours is considered. Thus, when the inventory is highly positive, the exchange provides incentives to the market maker so that it attracts buy market orders and discourage him from more sell market orders, and vice versa for a negative inventory. The integral $\int_0^T \hat{Z}_r^S dS_r$ can be understood as a risk-sharing term. More precisely, $\int_0^t Q_r dS_r$ corresponds to the price driven component of the inventory risk $Q_t S_t$. Hence, the exchange supports the proportion $\frac{\gamma}{\gamma+\eta}$ of this risk so that the market maker maintains reasonable quotes despite some inventory.[6]

Notice that for a highly risk-averse exchange, that is, $\eta \nearrow \infty$,

$$\int_0^T \hat{Z}_r^a dN_r^a + \hat{Z}_r^b dN_r^b \approx c(N_T^a + N_T^b),\ \hat{Z}_r^S \approx 0,$$

meaning that the exchange transfers to the market maker the total fee. This is the so-called *selling the firm* effect, as the exchange delegates all benefit to the market maker.[7]

Until now, we have focused on the maker part of the make–take fees problem since we have considered that the taker cost c is fixed. Nevertheless, our approach also enables us to suggest the exchange a relevant value for c. Actually, we see that when acting optimally, the exchange transfers the totality of the fixed taker fee c to the market maker. It is therefore neutral to the value of c as its optimal utility function $v_0^E = v(0,Q_0)$ is independent of the taker cost (see (18)). However, c plays an important role in the optimal spread offered by the market maker given by

$$-2c + \frac{\sigma}{k}\log\left(\frac{u(t,Q_{t-})^2}{u(t,Q_{t-}-1)u(t,Q_{t-}+1)}\right) - \frac{2}{\eta}\log\left(1 - \frac{\sigma^2\gamma\eta}{(k+\sigma\gamma)(k+\sigma\eta)}\right) + \frac{2}{\gamma}\log\left(1+\frac{\sigma\gamma}{k}\right).$$

Furthermore, from numerical computations[8] or the asymptotic development (25), we remark that $\frac{u(t,q)^2}{u(t,q-1)u(t,q+1)}$ is close to unity for any t and q. Hence if, for example, the exchange targets a spread close to one tick (see Dayri & Rosenbaum, 2015; Huang, Lehalle, & Rosenbaum, 2016 for details on optimal tick sizes and spreads), it can be obtained by setting

$$c \approx -\frac{1}{2}\text{Tick} - \frac{1}{\eta}\log\left(1 - \frac{\sigma^2\gamma\eta}{(k+\sigma\gamma)(k+\sigma\eta)}\right) + \frac{1}{\gamma}\log\left(1 + \frac{\sigma\gamma}{k}\right).$$

For $\sigma\gamma/k$ small enough, this equation reduces to

$$c \approx \frac{\sigma}{k} - \frac{1}{2}\text{Tick}. \tag{26}$$

This is a particularly simple formula for setting the taker constant fee c, as the parameters σ and k can be easily estimated from market data. We see that the higher the volatility, the larger the taker cost should be. The decrease in k is also natural: If k is large, the liquidity vanishes rapidly when the spread becomes wide, meaning that market takers are sensitive to extra costs relative to the efficient price. Therefore, the taker cost has to be small if the exchange wants to maintain a reasonable market order flow.

5 | EXCHANGE IMPACT ON MARKET QUALITY

In this section, we compare our setting with the situation without incentive policy from an exchange toward market making activities which corresponds to the problem of optimal market making considered in Avellaneda and Stoikov (2008) and Guéant et al. (2008). The results in Avellaneda and Stoikov (2008) are taken as benchmark for our investigation to emphasize the impact of the incentive policy on market quality. We will refer to this case as the neutral exchange case.

Let us first recall the results in Avellaneda and Stoikov (2008) and Guéant et al. (2008). The optimal controls of the market maker denoted by $\widetilde{\delta}^a$ and $\widetilde{\delta}^b$ are given as a function of the inventory Q_t by

$$\widetilde{\delta}_t^i = \frac{\sigma}{k}\log\left(\frac{\widetilde{u}(t,Q_{t-})}{\widetilde{u}(t,Q_{t-}-\varepsilon_i)}\right) + \frac{1}{\gamma}\log\left(1+\frac{\sigma\gamma}{k}\right),\ \ i \in \{b,a\},\ (\varepsilon_b,\varepsilon_a) = (-1,1),$$

where $\widetilde{u}$ is the unique solution of the linear differential equation

$$\widetilde{u}\Big|_{t=T} = 1 \text{ and } \partial_t\widetilde{u}(t,q) - F_{\widetilde{C}_1,\widetilde{C}_1'}(q,\widetilde{u}(t,q),\widetilde{u}(t,q+1),\widetilde{u}(t,q-1)) = 0,\ \ t<T, |q| \le \bar{q},$$

with $\widetilde{C}_1 = \frac{\sigma\gamma k}{2}$ and $\widetilde{C}_1' = A\left(1+\frac{\sigma\gamma}{k}\right)^{-(1+\frac{\sigma\gamma}{k})}$. In our case, the optimal quotes $\hat{\delta}^a$ and $\hat{\delta}^b$ are obtained from Theorem 4.3 and satisfy for $i \in \{b,a\}$, and $(\varepsilon_b,\varepsilon_a) = (-1,1)$:

$$\hat{\delta}_t^i = \frac{\sigma}{k}\log\left(\frac{u(t,Q_{t-})}{u(t,Q_{t-}-\varepsilon_i)}\right) + \frac{1}{\gamma}\log(1+\frac{\sigma\gamma}{k}) - c - \frac{1}{\eta}\log\left(1-\frac{\sigma^2\gamma\eta}{(k+\sigma\gamma)(k+\sigma\eta)}\right),$$

where u is solution of the linear equation (19).

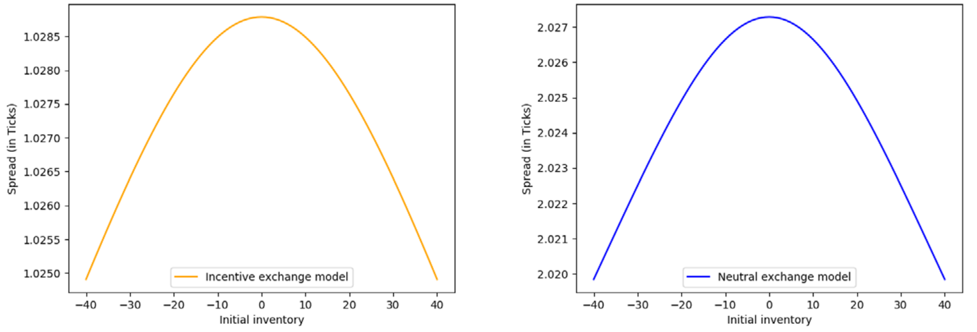

FIGURE 1 Comparison of optimal initial spreads with/without incentive policy from the exchange [Color figure can be viewed at wileyonlinelibrary.com]

Numerical experiments show that u and $\widetilde{u}$ decrease quickly to zero when q becomes large, inducing numerical instabilities in the computation of

$$v_+(t,q) = \log\left(\frac{u(t,q+1)}{u(t,q)}\right), \quad \widetilde{v}_+(t,q) = \log\left(\frac{\widetilde{u}(t,q+1)}{\widetilde{u}(t,q)}\right), \quad q \in \{-\bar{q}, \dots, \bar{q}-1\},$$

which are crucial in the expressions of optimal quotes. To circumvent this numerical difficulty, we remark that v_+ and $\widetilde{v}_+$ are solution of the following integro-differential equations:

$$v_+\Big|_{t=T} = 0 \text{ and } \partial_t v_+(t,q) + \mathcal{F}_{C_1,C_1'}(q, v_+(t,q), v_+(t,q+1), v_-(t,q+1)) = 0, \tag{27}$$

$$\widetilde{v}_+\Big|_{t=T} = 0 \text{ and } \partial_t \widetilde{v}_+(t,q) + \mathcal{F}_{\widetilde{C_1},\widetilde{C_1'}}(q, \widetilde{v}_+(t,q), \widetilde{v}_+(t,q+1), \widetilde{v}_+(t,q-1)) = 0, \tag{28}$$

where, again with $(\varepsilon_b, \varepsilon_a) = (-1, 1)$,

$$\mathcal{F}_{\alpha,\beta}(q, y, y_+, y_-) = \alpha(2q+1) - \beta \sum_{i\in\{a,b\}} \varepsilon_i e^{\varepsilon_i v_+(t,q-\varepsilon_i)} \mathbb{1}_{\{\varepsilon_i q < \bar{q}-1\}} - \varepsilon_i e^{\varepsilon_i v_+(t,q)}.$$

We thus rather apply classical finite difference schemes to (27) and (28).

In the following numerical illustrations, in the spirit of Guéant et al. (2008, section 6), we take T = 600s for an asset with volatility $\sigma = 0.3$ Tick.s$^{-1/2}$ (unless specified differently). Market orders arrive according to the intensities (2) with A = 1.5s^{-1} and k = 0.3s$^{-1/2}$. We assume that the threshold inventory of the market maker is $\bar{\text{q}} = 50$ units and we set his risk aversion parameter to $\gamma = 0.01$. The exchange is taken more risk averse with $\eta = 1$. Finally, we assume that the taker cost c = 0.5 Tick.[9]

5.1 | Impact of the exchange on the spread and market liquidity

We start by comparing the optimal spread $\hat{\delta}_0^a + \hat{\delta}_0^b$ at time 0 obtained when contracting optimally with the optimal spread $\widetilde{\delta}_0^a + \widetilde{\delta}_0^b$ without contracting. The optimal spreads are plotted in Figure 1

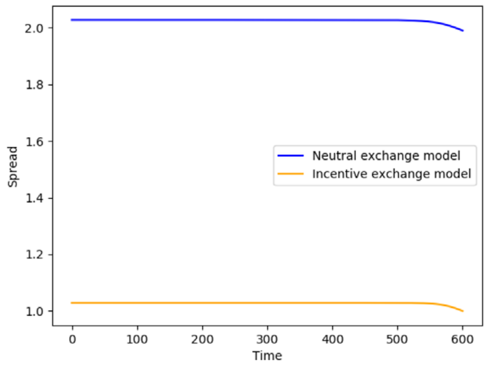

FIGURE 2 Average spread on$[0, T]$with95%confidence interval, with/without incentive [Color figure can be viewed at wileyonlinelibrary.com]

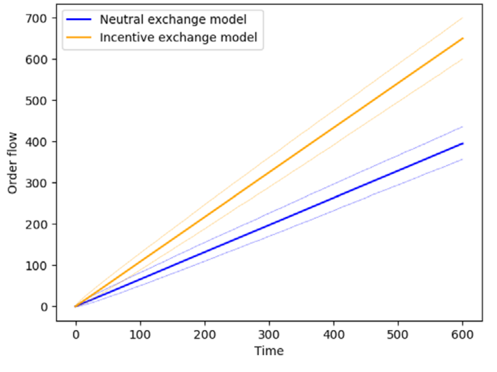

FIGURE 3 Average order flow on$[0, T]$with95%confidence interval, with/without incentive [Color figure can be viewed at wileyonlinelibrary.com]

for different initial inventory values $Q_0 \in \{-\bar{q}, \dots, \bar{q}\}$.

We observe in Figure 1 that the initial spread does not depend a lot on the initial inventory (because the considered time interval $[0, T]$ is not too small) and that it is reduced thanks to the optimal contract between the market maker and the exchange. This is not surprising since in our case the exchange aims at increasing the market order flow by proposing an incentive contract to the market maker inducing a spread reduction. Actually this phenomenon occurs over the whole trading period $[0, T]$. To see this, we generate 5,000 paths of market scenarios and compute the average spread over $[0, T]$ for an initial inventory $Q_0 = 0$. The results are given in Figure 2. Since the spread is tighter during the trading period under an incentive policy from the exchange, the arrival intensity of market orders is more important and hence the market is more liquid as shown in Figure 3.

We now consider in Figure 4 the bid and ask sides separately. We see that when the inventory is positive and very large, $\hat{\delta}^a$ and $\widetilde{\delta}^a$ are negative, meaning that the market maker is ready to sell at prices lower than the efficient price in order to attract market orders and reduce his inventory risk. On the contrary, if the inventory is negative and very large, in both situations, its ask quotes are well above the efficient price in order to repulse the arrival of buy market orders. However, since in our case the exchange remunerates the market maker for each arrival of market order,

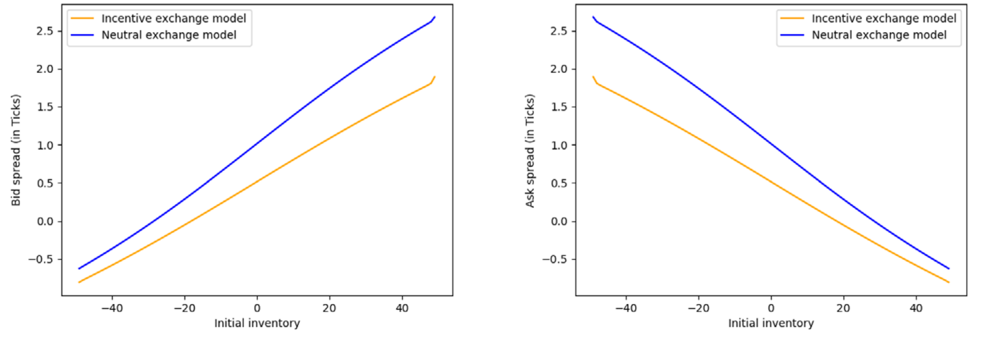

FIGURE 4 Optimal ask and bid spreads, with/without incentive policy [Color figure can be viewed at wileyonlinelibrary.com]

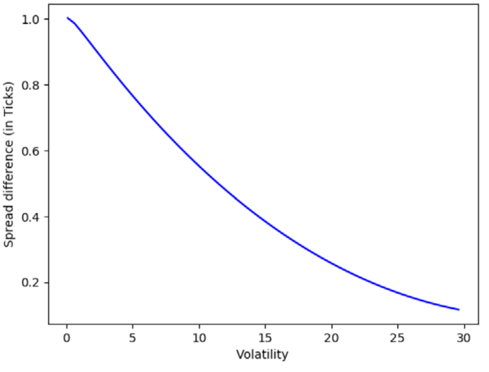

FIGURE 5 Initial optimal spread difference between the situations with and without incentive [Color figure can be viewed at wileyonlinelibrary.com]

we get that the ask spread with contract $\hat{\delta}^a$ is smaller than $\widetilde{\delta}^a$. A symmetric conclusion holds for the bid part of the spread.

We now turn to the impact of the volatility on the spread. The optimal contract obtained in (23) induces an inventory risk-sharing phenomenon through the term $\hat{Z}^S$. Hence, when the volatility increases, the spread difference between situations with/without incentive policy becomes less important, see Figure 5 in which we consider the optimal initial spread difference when the initial inventory is set to zero between both situations with/without incentive policy from the exchange to the market maker for different values of the volatility.

5.2 | Impact on the P&L of the exchange and the market maker

We assume that $Q_0 = 0$. Recall that PL^{δ} defined in (3) denotes the trading part of the P&L of the market maker for a given strategy δ. In our case, the underlying[10] total P&L at time t of a market maker acting optimally, denoted by $\mathrm{PL}_t^{\star}$, is

$$\mathrm{PL}_t^{\star} = \mathrm{PL}_t^{\hat{\delta}} + Y_t^{\hat{Y}_0,\hat{Z}},$$

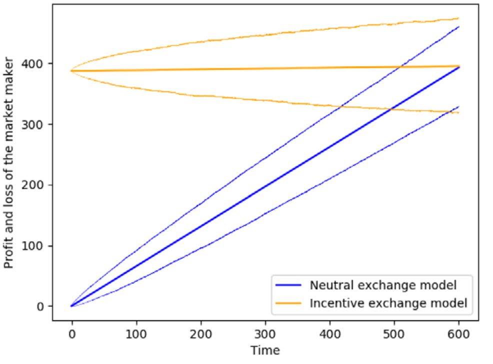

FIGURE 6 Average P&L of market maker with/without incentive, with95%confidence interval [Color figure can be viewed at wileyonlinelibrary.com]

where $Y_t^{\hat{Y}_0,\hat{Z}}$ corresponds to the quantity on the right-hand side of (23) with T replaced by t. We now compare this quantity to the benchmark $\mathrm{PL}_t^{\tilde{\delta}}$ which corresponds to the optimal P&L without intervention of the exchange.

To make $\mathrm{PL}_t^{\star}$ and $\mathrm{PL}_t^{\tilde{\delta}}$ comparable, we choose $\hat{Y}_0$ in (23) so that the market maker gets the same utility in both situations, that is $\hat{Y}_0 = \frac{k}{\sigma}\log(\tilde{u}(0, Q_0))$. Thus, the market maker is indifferent between the situation with or without exchange intervention. We generate 5,000 paths of market scenarios and compare the average of both P&L in Figure 6 with and without incentive policy.

Since $\hat{Y}_0$ is set to obtain the same utility in both cases, the two average P&L are very close at the end of the trading period. The variance of the P&L also seems to be the same in both situations. The only difference from the market maker viewpoint here is that in the case of a contract, the P&L is already made at time 0 thanks to the compensation of the exchange and then fluctuates slightly. This is because he is earning the spread but paying continuous "coupons" $(H(\hat{Z}_t, Q_t) - \frac{\sigma^2\gamma}{2}\left(\hat{Z}_t^S + Q_t\right)^2)dt$ from the contract. In the case without exchange intervention, the market maker increases his P&L over the whole trading period thanks to the spread.

We now compare the P&L of the exchange in the two considered cases. When it applies an incentive policy toward the market maker, the P&L of the exchange is given by $c\left(N_t^a + N_t^b\right) - Y_t^{\hat{Y}_0,\hat{Z}}$. When the exchange is neutral, its P&L is simply $c\left(N_t^a + N_t^b\right)$. We compare these two quantities in Figure 7.

We see that the initial P&L of the contracting exchange is negative because of the initial payment $\hat{Y}_0$. However it finally exceeds, with a smaller standard deviation, the P&L in the situation without incentive policy from the exchange. Hence, the incentive policy of the exchange proves to be successful. Both configurations are indeed equivalent for market makers but the exchange obtains more revenues when contracting optimally. This is due to the fact that the contract triggers more market orders.

Finally, we plot the aggregated average P&L of the market maker and the exchange (independent of the choice of the initial payment) in Figure 8. We observe that it is always greater in the optimal contract case.

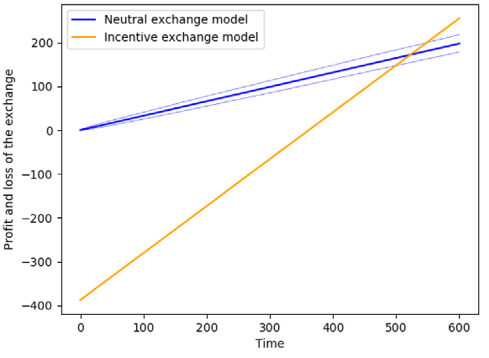

FIGURE 7 Average P&L of the exchange with/without incentive, with95%confidence interval [Color figure can be viewed at wileyonlinelibrary.com]

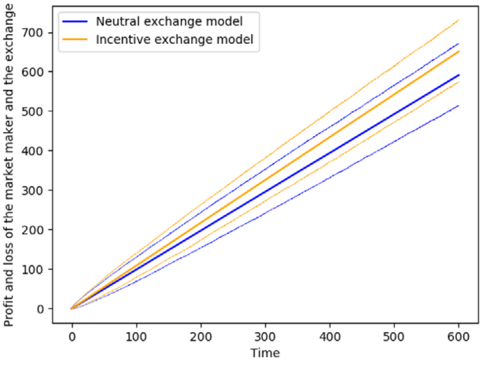

FIGURE 8 P&L of exchange and market maker with/without incentive,95%confidence interval [Color figure can be viewed at wileyonlinelibrary.com]

5.3 | Impact of the incentive policy on the trading cost

We consider one single market taker. In the case without exchange, with the specified parameters and under optimal reaction of the market maker, this investor buys on average 200 shares over $[0, T]$. To make the comparison with the case with exchange intervention, we modify the parameter A appearing in the intensity (2) when simulating a market with optimal contract. This new value is chosen so that the investor buys on average the same number of assets (200) over the time period. This amounts to take $A = 0.9s^{-1}$. We confirm in Figure 9 that the average ask order flows agree in both situations.

Finally, Figure 10 compares the average cost of trading for the market taker $\mathbb{E}^{\delta}[\int_0^T \delta_t^a dN_t^a]$, with and without incentive, and shows that the reduced spreads lead to significantly smaller trading costs for investors.

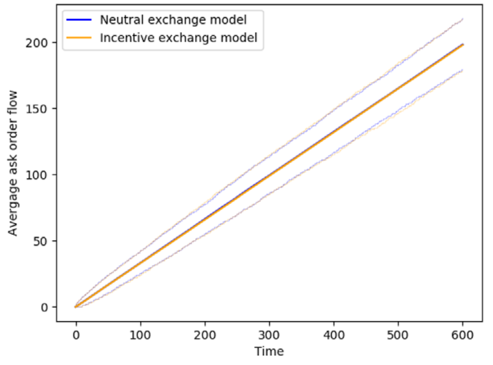

FIGURE 9 Setting similar average ask order flows on$[0, T]$by taking different intensity basis$A$in the case with and in the case without incentive policy;95%confidence interval [Color figure can be viewed at wileyonlinelibrary.com]

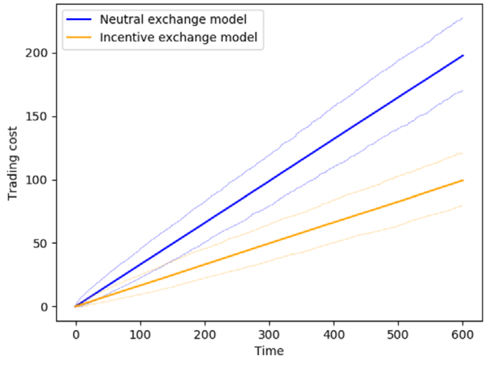

FIGURE 10 Average trading cost on$[0, T]$with95%confidence interval, with/without incentive [Color figure can be viewed at wileyonlinelibrary.com]

6 | EXTENSION: SYMMETRIC EXCHANGES COMPETITION

In this section, we extend the previous study by considering a first step toward the investigation of the case of several exchanges in competition.

6.1 | Symmetric exchanges in Nash equilibrium

We assume here that N identical exchanges display the quotes of one market maker and that the trading flows are split equally between the exchanges. More precisely, each time the market taker acts on the market, his trade of size one is split into N trades of size $1/N$ distributed across all exchanges. This is equivalent to modify the market taker fee received by each exchange from c to c/N.

Remark 6.1. The symmetric splitting should result from a symmetric equilibrium rather than be an assumption. Therefore, our results can be seen as a preliminary analysis of the case of different exchanges with various market makers. Furthermore, as we will see below, the situation considered here is already significantly more intricate than the case of one exchange treated in the previous sections. That is why we focus on the case where the market taker splits his order evenly among N exchanges since we are interested in a symmetric equilibrium among identical exchanges.

The market maker receives the aggregation of the compensation given by the N exchanges denoted by $\overline{\xi} = \xi + \tilde{\xi}$, where ξ and $\tilde{\xi}$ are, respectively, the remuneration given by a representative exchange and the aggregation of the $N-1$ others. Hence, $\overline{\xi}$ inherits all the technical assumptions made previously on ξ (for only one exchange), since the problem of the market maker is similar by considering $\overline{\xi}$ for his compensation. Consequently, the market maker's problem returns an optimal spread $\hat{\delta}(\overline{\xi})$ so that Theorem 3.1 holds by considering $\overline{\xi}$. In view of the symmetry assumption made on the exchanges, any exchange aims at solving

$$V_0^E(\tilde{\xi}) = \sup_{\xi \in \mathcal{C}} \mathbb{E}^{\hat{\delta}(\xi+\tilde{\xi})}\left[-e^{-\eta(\frac{c}{N}(N_T^a+N_T^b)-\xi)}\right], \tag{29}$$

where $\tilde{\xi}$ is fixed, and $\eta > 0$ is the common risk aversion parameter of the N exchanges.

Definition 6.2 (Nash equilibrium and symmetric Nash equilibrium)**.** A N-tuple $(\xi^{\mathrm{e}})_{1\le \mathrm{e}\le N}$ is a Nash equilibrium if for any $\mathrm{e} \in \{1, \dots, N\}$ we have

$$V_0^E(\xi^{\mathrm{e}}) = \mathbb{E}^{\hat{\delta}(\sum_{j=1}^N \xi^j)}\left[-e^{-\eta(\frac{c}{N}(N_T^a+N_T^b)-\xi^{\mathrm{e}})}\right].$$

An N-tuple of contracts $(\xi^{\mathrm{e}})_{1\le \mathrm{e}\le N}$ is a symmetric Nash equilibrium if $(\xi^{\mathrm{e}})_{1\le \mathrm{e}\le N}$ is a Nash equilibrium such that $\xi^1 = \cdots = \xi^N$. We denote by $\mathcal{S}^N := \{\xi^0 : (\xi^0, \dots, \xi^0) \in \mathcal{C}^N\}$ the collection of all such symmetric Nash equilibria.

From Theorem 3.1, it follows that any symmetric Nash equilibrium $\xi^0 \in \mathcal{S}^N$ is induced by a pair $(\tilde{y}_0, \tilde{Z}) \in [\hat{Y}_0, +\infty) \times \mathcal{Z}$ such that

$$\begin{aligned}
\xi^0 &= \frac{1}{N} Y_T^{\tilde{y}_0,\tilde{Z}} = \frac{\tilde{y}_0}{N} + \int_0^T \frac{1}{N}\tilde{Z}_r d\chi_r + \frac{\gamma\sigma^2}{2N}\left(\tilde{Z}_r^S + Q_r\right)^2 dr - \frac{1}{N}\sum_{i=a,b} H^i\left(\tilde{Z}_r^i, Q_r\right) dr,\\
&= \frac{\tilde{y}_0}{N} + \int_0^T \zeta_r^0 d\chi_r + \frac{\gamma\sigma^2}{2N}\left(N\zeta_r^{S,0} + Q_r\right)^2 dr - \frac{1}{N}\sum_{i=a,b} H^i\left(N\zeta_r^{i,0}, Q_r\right) dr,
\end{aligned} \tag{30}$$

with $\zeta^0 = \frac{\tilde{Z}}{N}$, and

$$H^i(z,q) = \lambda(\hat{\delta}(z))\frac{\sigma}{k+\sigma\gamma}\mathbf{1}_{\varepsilon_i q<\overline{Q}}, \text{ with } (\varepsilon_b, \varepsilon_a) = (-1, 1).$$

We now denote by $\xi^{0,N-1}$ the $(N-1)$-tuple of identical contracts ξ^0 defined by (30), and we set $\tilde{Y}_0 := \frac{N-1}{N}\tilde{y}_0$. As $\xi + (N-1)\xi^0 = Y_T^{Y_0+\tilde{Y}_0,Z}$, by setting $\zeta := Z - (N-1)\zeta^0$, for some $(Y_0, Z) \in$

$[\hat{Y}_0, +\infty) \times \mathcal{Z}$ the problem of each exchange reduces to

$$V_0^E(\xi^{0,N-1}) = \sup_{Y_0,\zeta} \mathbb{E}^{\hat{\delta}(\zeta+(N-1)\zeta^0)}\left[-e^{-\eta(\int_0^T(\sum_{i=a,b}(\frac{c}{N}-\zeta_t^i)dN_t^i-\zeta_t^S dS_t)-Y_0-\int_0^T G(\zeta_t,\zeta_t^0,Q_t)-\alpha_t^0 dt)}\right],$$

where Y_0 ranges in $[\hat{Y}_0 - \tilde{Y}_0, +\infty)$, $\zeta \in \mathcal{Z}$, and

$$\begin{aligned} G(\zeta,\zeta^0,q) &= \frac{1}{2}\gamma\sigma^2(\zeta^S + (N-1)\zeta^{S,0} + q)^2 - \sum_{i=a,b} H^i(\zeta^i + (N-1)\zeta^{i,0}, q), \\ \alpha_t^0 &= \frac{N-1}{N}\left(\frac{1}{2}\gamma\sigma^2(N\zeta_t^{S,0} + Q_t)^2 - \sum_{i=a,b} H^i(N\zeta_t^{i,0}, Q_t)\right). \end{aligned}$$

The optimization over Y_0 is immediately solved, leading to

$$V_0^E(\xi^{0,N-1}) = \sup_{\zeta\in\mathcal{Z}} \mathbb{E}^{\hat{\delta}(\zeta+(N-1)\zeta^0)}\left[-e^{-\eta(\int_0^T(\sum_{i=a,b}(\frac{c}{N}-\zeta_t^i)dN_t^i-\zeta_t^S dS_t)-Y_0^\star-\int_0^T G(\zeta_t,\zeta_t^0,Q_t)-\alpha_t^0 dt)}\right], \tag{31}$$

with $Y_0^\star = \hat{Y}_0 - \tilde{Y}_0$.

Definition 6.3 (Markovian Nash equilibrium). A symmetric Nash equilibrium $\xi^0 \in \mathcal{S}^N$ is Markovian if the coefficients ζ^0 appearing in (30) is given by $\zeta_t^0 = \zeta^0(t, Q_t)$ for some deterministic function ζ^0.

Remark 6.4. Note that if ξ^0 is a symmetric Nash equilibrium with decomposition (30), we necessarily have $\tilde{y}_0 = \hat{Y}_0$, $\hat{\zeta}(\zeta^0) = \zeta^0$, where $\hat{\zeta}(\zeta^0)$ denotes an optimizer of (31). This allows to characterize any symmetric Nash equilibrium if there exists at least one.

6.2 | The main result

Similarly to the one exchange problem studied previously, we introduce the HJB equation

$$v\Big|_{t=T} = -1 \text{ and } \partial_t v(t,q) - \eta v(t,q)\hat{F}(t,q,v(t,q),v(t,q+1),v(t,q-1)) = 0, \tag{32}$$

with

$$\hat{F}(t,q,y,y^+,y^-) = \sup_{\zeta^S} F^S(t,q,\zeta^S) + \sup_{\zeta} F^0(t,q,y,y^+,\zeta)\mathbf{1}_{q<\bar{q}} + \sup_{\zeta} F^0(t,q,y,y^-,\zeta)\mathbf{1}_{q>-\bar{q}},$$

and

$$\begin{aligned} F^S(t,q,z) &= -\hat{H}^S(q,\zeta^{S,0}(q),z) - \frac{\eta}{2}\sigma^2|z|^2, \\ F^0(t,q,y,y',z) &= -\lambda\left(\hat{\delta}(z+(N-1)\widetilde{\zeta^0}(y,y'))\right)\left(\frac{y'}{y}\frac{e^{\eta(z-\frac{c}{N})}}{\eta} - \frac{1}{\eta} - \frac{\sigma}{k+\sigma\gamma}\right) \end{aligned}$$

$$-\lambda\Big(\hat{\delta}(N\widetilde{\zeta}^0(y,y'))\Big)\frac{(N-1)\sigma}{N(k+\sigma\gamma)},$$

$$\hat{H}^S(q,\tilde{z},z) = \frac{1}{2}\sigma^2\gamma\left[(z+(N-1)\tilde{z}+q)^2 - \frac{N-1}{N}(N\tilde{z}+q)^2\right],$$

$$\zeta^{S,0}(q) = -\frac{\gamma}{\eta+N\gamma}q,\ \ \widetilde{\zeta}^0(y,y') \ = \ \hat{\zeta}(y,y') + \frac{1-N}{N}c.$$

Hence, by denoting $\widehat{\zeta}^S$, $\widehat{\zeta}^N$ the optimizers of F^S and F^0, respectively, we get

$$\widehat{\zeta}^S(q) = \zeta^{S,0}(q),\ \ \widehat{\zeta}^N(y,y') = \widetilde{\zeta}^0(y,y').$$

Consequently, the HJB equation (32) reduces to

$$\partial_t v(t,q) + \frac{\gamma\eta^2\sigma^2}{2N(N\gamma+\eta)}q^2 v(t,q) - \hat{C}_N v(t,q)\left[\mathbb{1}_{\{q>-\bar{q}\}}\left(\frac{v(t,q)}{v(t,q-1)}\right)^{\frac{Nk}{\sigma\eta}} + \mathbb{1}_{\{q<\bar{q}\}}\left(\frac{v(t,q)}{v(t,q+1)}\right)^{\frac{Nk}{\sigma\eta}}\right] \ = \ 0, \quad (33)$$

with boundary condition $v|_{t=T} = -1$, where

$$\hat{C}_N = C_0 e^{\frac{(N-1)k}{\sigma\eta}}\ \frac{\sigma\gamma + \frac{1}{N}(\sigma\eta+k)}{\sigma\gamma+(\sigma\eta+k)}.$$

We now set $u = (-v)^{-\frac{kN}{\sigma\eta}}$. By direct substitution, we obtain the following linear equation:

$$u\Big|_{t=T} = 1 \text{ and } \partial_t u(t,q) - F_{C_N,C_N'}(q,u(t,q),u(t,q+1),u(t,q-1)) = 0, \quad t\in[0,T), \qquad (34)$$

with

$$C_N \ = \ \frac{k\gamma\eta\sigma}{2(N\gamma+\eta)} \text{ and } C_N' = \hat{C}_N\frac{kN}{\sigma\eta}.$$

Similarly to Section 4, we deduce that

$$v(t,q) = -(u(t,q))^{-\frac{\sigma\eta}{kN}} \text{ where } u(t,q) = \mathbf{b}_q\cdot e^{(T-t)\mathbf{B}_N}\mathbf{1},$$

and

$$\mathbf{B}_N = \begin{pmatrix} -C_N\bar{q}^2 & C_N' & & & \\ \ddots & \ddots & \ddots & & \\ & C_N' & -C_N q^2 & C_N' & \\ & & \ddots & \ddots & \ddots \\ & & & C_N' & -C_N\bar{q}^2 \end{pmatrix} \leftarrow q\text{th line},$$

Direct calculations reported in Appendix A.3 provide another form for the function u :

$$u(t,q) = \sum_{p\geq 0} \frac{[C_N'(T-t)]^p}{p!} \sum_{j\geq 0} \frac{[C_N'(T-t)]^j}{j!} e^{-C_N(T-t)(q+j-p)^2} \mathbb{1}_{\{|q+j-p|\leq \bar{q}\}}. \tag{35}$$

The following result establishes the existence of a unique symmetric Nash equilibrium which is moreover Markovian. The proof is postponed to Appendix A.6.

Theorem 6.5. *There is a unique symmetric Nash equilibrium $\xi^0 \in \mathcal{S}^N$ defined by*

$$\xi^0 = \frac{\widehat{Y_0}}{N} + \int_0^T \zeta_r^0 d\chi_r + \frac{\gamma\sigma^2}{2N}(N\zeta_r^{S,0} + Q_r)^2 dr - \frac{1}{N}\sum_{i=a,b} H^i(N\zeta_r^{i,0}, Q_r)dr, \tag{36}$$

where, for $i \in \{a,b\}$ and $(\varepsilon_b, \varepsilon_a) = (-1, 1)$,

$$\zeta_r^{S,0} = -\frac{\gamma}{\eta + N\gamma} Q_r, \quad \zeta_r^{i,0} = \zeta_N + \frac{1}{\eta}\log\left(\frac{v(r,Q_r)}{v(r,Q_r-\varepsilon_i)}\right),$$

with

$$\zeta_N = \frac{c}{N} + \frac{1}{\eta}\log\left(1 - \frac{\sigma^2\gamma\eta}{(\sigma\eta+k)(\sigma\gamma+k)}\right).$$

In particular, this unique symmetric Nash equilibrium is Markovian.

Remark 6.6. There exists infinitely many (nonsymmetric) Nash equilibria. For instance, by the contract $(\xi^{\mathrm{e}})_{\mathrm{e}\leq N}$ defined by

$$\xi^{\mathrm{e}} = Y_0^j + \int_0^T \zeta_r^0 d\chi_r + \frac{\gamma\sigma^2}{N}(N\zeta_r^{S,0} + Q_r)^2 dr - \frac{1}{N}\sum_{i=a,b} H^i(N\zeta_r^{i,0}, Q_r)dr$$

is a (nonsymmetric) Nash equilibrium for any Y_0^j satisfying $\sum_{j=1}^N Y_0^j = \hat{Y}_0$.

Remark 6.7. The symmetry of the problem allows us to find a natural candidate for the equilibrium and by using a verification procedure (see the first step of the proof) we prove that it is indeed a (symmetric) Nash equilibrium. This follows from the fact that the integro-differential equation under consideration admits a smooth solution. If now one wants to extend this study to an heterogeneous oligopoly of exchanges hiring one market maker, the solution will be strongly linked to a system of fully coupled HJB equation has explained in Mastrolia and Ren (2018). However, the existence of a smooth solution for such a system is not clear.

6.3 | Economic insights

Notice that the total compensation $N\xi^0$ obtained by the market maker in the N-symmetric exchanges situation differs from the optimal contract $\hat{\xi}$ of the one-exchange situation in (36).

Hence, our result is not trivial and worth of interest even in the simple symmetric exchanges setting in symmetric Nash equilibrium.

Notice that we also have a similar representation of u as in Proposition 4.2:

$$u(t,q) = \mathbb{E}\left[e^{\int_t^T(-C_N(Q_s^{t,q})^2+\overline{\lambda}_s+\underline{\lambda}_s)ds}\right], \tag{37}$$

where $Q_s^{t,q} = q + \int_t^s d(\overline{N}_u - \underline{N}_u)$, and $(\overline{N}, \underline{N})$ is a two-dimensional point process with intensity $(\overline{\lambda}_s, \underline{\lambda}_s) = C'_N(\mathbb{1}_{\{Q_{s-}<\bar{q}\}}, \mathbb{1}_{\{Q_{s-}>-\bar{q}\}})$. By using the same arguments than those in Section 4.2.3 (and based on the asymptotic expansion in Avellaneda & Stoikov, 2008; Guéant et al., 2008), we note that

$$\zeta_t^{i,0} \sim \zeta_N + \varepsilon_i \frac{\sigma}{2kN}\sqrt{\frac{C_N}{C'_N}}(2Q_{t-} - \varepsilon_i),\ i \in \{a,b\}$$

for $i \in \{a,b\}$. Again, when the inventory is highly positive, the exchanges provide incentives to the market maker to attract buy market orders and discourage additional sell market orders, and vice versa for a negative inventory. As ζ_N and $\frac{C_N}{C'_N}$ are decreasing with respect to N, this effect is reduced when the number of platforms increases.

Moreover the optimal spread is not a monotonic function of N for a given c, thus suggesting that an optimal number of exchanges may exist. This is not surprising since we consider a single market maker acting on different exchanges and is therefore in a position to benefit from the competition of exchanges. Furthermore, note that there is no competition between market makers which would naturally reduce the spread. Finally, recall that from the exchanges viewpoint the variable c plays no role in their value function. Therefore, we may expect exchanges to choose c in order to obtain a specific microstructure as explained in Section 4.2.3.

Note now that when N becomes large, $N\zeta_r^{S,0} \sim -Q_r$. In other words, for a large number of platforms, the inventory risk is transferred to the oligopoly of exchanges.

ACKNOWLEDGMENTS

We are grateful to Angélique Begrand, Luxi Chen, and Laurent Fournier from Euronext for helpful comments. This work benefits from the financial support of the Chaires *Analytics, and Models for Regulation*, *Financial Risk*, and *Finance, and Sustainable Development*. Omar El Euch and Mathieu Rosenbaum gratefully acknowledge the financial support of the *ERC Grant 679836 Staqamof*. Nizar Touzi gratefully acknowledges the financial support of the *ERC Grant 321111 RoFiRM*. Thibaut Mastrolia gratefully acknowledges the financial support of the *ANR PACMAN*. The authors also thank Bas Werker for relevant comments.

ORCID

Omar El Euch https://orcid.org/0000-0001-7107-8946
Thibaut Mastrolia https://orcid.org/0000-0003-0578-0168
Mathieu Rosenbaum https://orcid.org/0000-0001-7489-1115

ENDNOTES

[1] In practice, the efficient price can be thought of as the midprice of the asset, see Robert and Rosenbaum (2011) and Delattre, Robert, and Rosenbaum (2013).

[2] As in Avellaneda and Stoikov (2008), for sake of simplicity, we assume that the market maker estimates his inventory risk using the efficient price S.

[3] See, for example, theorem III.3.11 in Jacod and Shiryaev (2013); the uniform boundedness of δ guarantees that L^δ is a martingale (see Sokol, 2013).

[4] In practice, some exchanges add to this fixed fee a component which is proportional to the traded cash amount. Our analysis can be extended to more elaborated fee schedules. Our choice of a constant fee is motivated by the induced simplicity which will be crucial to derive our quasi-explicit solution. Furthermore, we will in fact see that when using the optimal contract, the exchange is somehow indifferent to the value of c (see Section 4.2.3).

[5] See Appendix A.5.1.

[6] Note that in practice, it is unclear whether an exchange would accept sharing risk with (high-frequency) market makers, perhaps for image or commercial reasons. In our approach, we of course do not force the contract to depend on S. We actually show that it is optimal for the exchange that it does depend on S. So in theory the exchange should use the variable S to design his contract as it is the best strategy for his own interest. This is actually in line with the literature on principal–agent problems which deals a lot with risk-sharing issues. What we are doing here is providing the optimal solution for a contract in our idealized setting, that should be seen as a benchmark for building the contract in practice.

[7] We would like to thank an anonymous referee for suggesting this interpretation.

[8] See indeed Figure 2 by noting that u does not depend on the fee c.

[9] Remark that the taker cost is chosen according to Criteria (26). We expect the optimal spread to be close to one tick. Note also that here the tick is just a unit and not a true market parameter.

[10] More precisely, Figure 6 does not display the P&L as a lump-sum payment at terminal time. It rather represents a virtual P&L accumulated during the trading period. However, the form of the optimal contract being additive and adapted in time (this is an integral with respect to time), it makes sense to view the final lump sum payment as the terminal value of a cumulative sum.

[11] From (4), notice that for any $\delta \in \mathcal{A}$, the conditional expectation $\mathbb{E}^\delta_\tau$ depends only on the restriction of δ on $[\tau, T]$. Hence $\mathbb{E}^\mu_\tau$ is defined without ambiguity for $\mu \in \mathcal{A}_\tau$.

[12] Note that $\mathbb{E}^\delta[U^\delta_T] = J_T(0,\delta) > -\infty$ using Hölder inequality together with (7) and the uniform boundedness of the intensities of N^a and N^b. Hence, the process U^δ is integrable.

[13] In view of the class of contracts considered, we know that the principal proposes a contract such that there exists at least one optimal bid–ask spread for the agent denoted by $\tilde{\delta}$. Hence, $U^{\tilde{\delta}}_t$ is a $\mathbb{P}^{\tilde{\delta}}$-martingale and according to Doob regularization result, we know that we can find a càdlàg version of $U^{\tilde{\delta}}_t$ under $\mathbb{P}^{\tilde{\delta}}$. Thus, V_t admits a càdlàg version under $\mathbb{P}^{\tilde{\delta}}$, and since all the measure $\mathbb{P}^\delta$ for $\delta \in \mathcal{A}$ are equivalent, U^δ_t admits a càdlàg version.

REFERENCES

Abergel, F., Lehalle, C. A., & Rosenbaum, M. (2014). Understanding the stakes of high-frequency trading. *Journal of Trading*, *9*(4), 49–73.

Angel, J. J., Harris, L. E., & Spatt, C. S. (2011). Equity trading in the 21st century. *Quarterly Journal of Finance*, *1*(01), 1–53.

Avellaneda, M., & Stoikov, S. (2008). High-frequency trading in a limit order book. *Quantitative Finance*, *8*(3), 217–224.

Bellia, M. (2017). High-frequency market making: Liquidity provision, adverse selection, and competition. SSRN paper.

Brolley, M., & Malinova, K. (2013). Informed trading and maker-taker fees in a low-latency limit order market. Available at SSRN 2178102.

Cartea, À., Jaimungal, S., & Penalva, J. (2015). *Algorithmic and high-frequency trading*. Cambridge: Cambridge University Press.

Colliard, J. E., & Foucault, T. (2012). Trading fees and efficiency in limit order markets. *Review of Financial Studies*, *25*(11), 3389–3421.

Cvitanić, J., & Karatzas, I. (1993). Hedging contingent claims with constrained portfolios. *Annals of Applied Probability*, *3*(3), 652–681.

Cvitanić, J., Possamaï, D., & Touzi, N. (2018). Dynamic programming approach to principal-agent problems. *Finance and Stochastics*, *22*(1), 1–37.

Cvitanić, J., & Zhang, J. (2012). *Contract theory in continuous-time models*. Berlin, Heidelberg: Springer Science & Business Media.

Dayri, K., & Rosenbaum, M. (2015). Large tick assets: Implicit spread and optimal tick size. *Market Microstructure and Liquidity*, *1*(1), 1550003.

Delattre, S., Robert, C. Y., & Rosenbaum, M. (2013). Estimating the efficient price from the order flow: A Brownian Cox process approach. *Stochastic Processes and Their Applications*, *123*(7), 2603–2619.

Foucault, T., Kadan, O., & Kandel, E. (2013). Liquidity cycles and make/take fees in electronic markets. *Journal of Finance*, *68*(1), 299–341.

Guéant, O., Lehalle, C.-A., & Fernandez-Tapia, J. (2008). Dealing with the inventory risk: A solution to the market making problem. *Mathematics and Financial Economics*, *7*(4), 1–31.

Harris, L. (2013). *Maker-taker pricing effects on market quotations*. USC Marshall School of Business Working Paper. Retrieved from http://bschool.huji.ac.il/.upload/hujibusiness/Maker-taker.pdf

Hölmstrom, B. (1979). Moral hazard and observability. *Bell Journal of Economics*, *10*(1), 74–91.

Holmstrom, B., & Milgrom, P. (1987). Aggregation and linearity in the provision of intertemporal incentives. Econometrica. *Journal of the Econometric Society*, *55*(2), 303–328.

Huang, W., Lehalle, C. A., & Rosenbaum, M. (2016). How to predict the consequences of a tick value change? Evidence from the Tokyo Stock Exchange Pilot program. *Market Microstructure and Liquidity*, *2*(03n04), 1750001.

Jacod, J., & Shiryaev, A. (2013). *Limit theorems for stochastic processes* (Vol. 288). Berlin, Heidelberg: Springer Science & Business Media.

Laruelle, S., & Lehalle, C. A. (2018). *Market microstructure in practice*. New Jersey: World Scientific.

Lutat, M. (2010). The effect of maker-taker pricing on market liquidity in electronic trading systems-empirical evidence from European equity trading. Available at SSRN 1752843.

Madhavan, A., Richardson, M., & Roomans, M. (1997). Why do security prices change? A transaction-level analysis of NYSE stocks. *Review of Financial Studies*, *10*(4), 1035–1064.

Malinova, K., & Park, A. (2015). Subsidizing liquidity: The impact of make/take fees on market quality. *Journal of Finance*, *70*(2), 509–536.

Mastrolia, T., & Ren, Z. (2018). Principal-agent problem with common agency without communication. *SIAM Journal on Financial Mathematics*, *9*(2), 775–799.

Megarbane, N., Saliba, P., Lehalle, C. A., & Rosenbaum, M. (2017). The behavior of high-frequency traders under different market stress scenarios. *Market Microstructure and Liquidity*, *3*(03n04), 1850005.

Menkveld, A. J. (2013). High frequency trading and the new market makers. *Journal of Financial Markets*, *16*(4), 712–740.

Mirrlees, J. (1974). Notes on welfare economics, information and uncertainty. In: M. Balch, D. McFadden, & S.-Y. Wu (Eds.), *Essays on economic behavior under uncertainty* (pp. 243–261). Amsterdam: North-Holland.

Neveu, J. (1972). *Martingales à temps discret*. Dunod edition Masson et Cie, éditeurs, Paris.

Riordan, R., & Park, A. (2012). *Maker/taker pricing and high frequency trading*. Foresight report, UK Government Office for Science.

Robert, C. Y., & Rosenbaum, M. (2011). A new approach for the dynamics of ultra-high-frequency data: The model with uncertainty zones. *Journal of Financial Econometrics*, *9*(2), 344–366.

Sannikov, Y. (2008). A continuous-time version of the principal-agent problem. *Review of Economic Studies*, *75*(3), 957–984.

Securities and Exchange Commission. (2010). Concept release on equity market structure. Release No. 34-61358; File No. S7-02-10.

Sokol, A. (2013). Optimal Novikov-type criteria for local martingales with jumps. *Electronic Communications in Probability*, *18*, 1–8.

Wyart, M., Bouchaud, J. P., Kockelkoren, J., Potters, M., & Vettorazzo, M. (2008). Relation between bid-ask spread, impact and volatility in order-driven markets. *Quantitative Finance*, *8*(1), 41–57.

How to cite this article: Euch OE, Mastrolia T, Rosenbaum M, Touzi N. Optimal make–take fees for market making regulation. *Mathematical Finance*. 2021;31:109–148. https://doi.org/10.1111/mafi.12295

APPENDIX

A.1 | Predictable representation

The following result is probably well-known, we report it for completeness as we could not find a precise reference.

Lemma A.1. *Let $(\Omega, \mathcal{F}, \mathbb{P}, \mathbb{F})$ be a filtered probability space where $\mathbb{F} = \mathbb{F}^W \vee \mathbb{F}^N$ is the right continuous completed filtration of a Brownian motion W and a d-dimensional integrable point process $N = (N^1, \dots, N^d)$ with compensator $A = (A^1, \dots, A^d)$. Then, for any $\mathbb{F}$-martingale X there exists a predictable process $Z = (Z^W, Z^1, \dots, Z^d)$ such that*

$$X_t = X_0 + \int_0^t Z_s^W dW_s + \sum_{i=1}^d \int_0^t Z_s^i (dN_s^i - dA_s^i).$$

Proof. For sake of simplicity, we take $d = 1$. Let $\mathbb{P}$ be a solution of the martingale problem associated to $M_t = N_t - A_t$ and W_t. By theorem III.4.29 in Jacod and Shiryaev (2013), to prove Lemma A.1 we need to establish the uniqueness of $\mathbb{P}$.

We denote by $\mathbb{P}^W$ the law $\mathbb{P}$ conditional on W. We first show that M is still a martingale under $\mathbb{P}^W$. To do so, we consider $B_s \in \mathcal{F}_s$ and want to prove that

$$\mathbb{E}^{\mathbb{P}^W}\left[\mathbb{1}_{B_s}(M_t - M_s)\right] = 0$$

for $0 \le s \le t \le T$. Let $C \in \mathcal{F}_T^W$. We aim at showing that

$$\mathbb{E}\left[\mathbb{1}_C \mathbb{E}^{\mathbb{P}^W}\left[\mathbb{1}_{B_s}(M_t - M_s)\right]\right] = \mathbb{E}\left[\mathbb{1}_C \mathbb{1}_{B_s}(M_t - M_s)\right] = 0.$$

By the martingale representation theorem for Brownian martingales, we can write $\mathbb{1}_C = \alpha_s + \int_s^T \phi_u dW_u$, where $\alpha_s = \mathbb{E}[\mathbb{1}_C | \mathcal{F}_s^W]$ and $(\phi_u)_{u \ge 0}$ is $\mathbb{F}^W$ predictable process. Using the martingale property of M, we obtain

$$\mathbb{E}\left[\alpha_s \mathbb{1}_{B_s}(M_t - M_s)\right] = 0.$$

Then W and M being orthogonal martingales, we deduce

$$\mathbb{E}\left[\int_s^T \phi_u dW_u \mathbb{1}_{B_s}(M_t - M_s)\right] = 0.$$

Consequently, using theorem III.1.21 in Jacod and Shiryaev (2013), $\mathbb{P}^W$ is the unique probability measure such that M is an $\mathbb{F}$-martingale conditional on W. Finally, by integration, the uniqueness of $\mathbb{P}^W$ implies that of $\mathbb{P}$. □

A.2 | Exchange Hamiltonian maximization

The following result follows from (tedious) direct calculations:

Lemma A.2. *For all $v_1, v_2 < 0$, define*

$$\varphi(z) := A e^{-k\frac{\Delta(z)+c}{\sigma}} \left[v_1 e^{\eta(z-c)} - v_2 \left(\frac{\eta}{\gamma} \left(1 - e^{-\gamma(z+\Delta(z))}\right) + 1 \right) \right], \ z \in \mathbb{R},$$

with $\Delta(z) := (-\delta_\infty) \vee \Delta^0(z) \wedge \delta_\infty$ *and* $\Delta^0(z) := -z + \log\left(1+\frac{\sigma\gamma}{k}\right)^{\frac{1}{\gamma}}$*, for some parameter* δ_∞ *satisfying*

$$\delta_\infty \geq C_\infty + \left|\log\left(\frac{v_2}{v_1}\right)^{\frac{1}{\eta}}\right|, \text{ with } C_\infty := c + \log\frac{\left(1+\frac{\sigma\gamma}{k}\right)^{\frac{1}{\eta}+\frac{1}{\gamma}}}{\left(1-\frac{\sigma^2\gamma\eta}{(k+\sigma\gamma)(k+\sigma\eta)}\right)^{\frac{1}{\eta}}}. \tag{A.1}$$

Then, φ *has a maximum point* $z^\star$ *given by*

$$z^\star = c + \frac{1}{\eta}\log\left(\frac{v_2}{v_1}\right) + \frac{1}{\eta}\log\left(1-\frac{\sigma^2\gamma\eta}{(k+\sigma\gamma)(k+\sigma\eta)}\right) \text{ with } \varphi(z^\star) = -Cv_2\left(\frac{v_2}{v_1}\right)^{\frac{k}{\sigma\eta}},$$

$$|\Delta^0(z^\star)| \leq \delta_\infty, \text{ and } C = A\frac{\sigma\eta}{k}\left(1+\frac{\sigma\gamma}{k}\right)^{-\frac{k}{\sigma\gamma}}\left(1-\frac{\sigma^2\gamma\eta}{(k+\sigma\gamma)(k+\sigma\eta)}\right)^{1+\frac{k}{\sigma\eta}}.$$

A.3 | Justification of (21) and (35)

Denote by $\mathbf{D}$ and $\mathbf{J}$ the matrices defined by the entries $\mathbf{D}^{q,r} = q^2\mathbf{1}_{q=r}$ and $\mathbf{J}^{q,r} = \mathbb{I}_{r=q+1} + \mathbb{I}_{r=q+1}$, $-\overline{q} \leq p, r \leq \overline{q}$. Notice that the calculations reported in (21) and (35) reduce to

$$U(q) := \mathbf{b}_q \cdot \sum_{|\ell|\leq\bar{q}} e^{\alpha\mathbf{J}-\beta\mathbf{D}}\,\mathbf{b}_\ell,\ |q| \leq \bar{q}, \quad \text{for } -\overline{q} \leq q \leq \overline{q}.$$

We first compute that

$$e^{\alpha\mathbf{J}-\beta\mathbf{D}} = \sum_{k\geq 0}\frac{(\alpha\mathbf{J}-\beta\mathbf{D})^k}{k!} = \sum_{k\geq 0}\frac{1}{k!}\sum_{j=0}^{k}\binom{k}{j}\alpha^j(-\beta)^{k-j}\ell^{2(k-j)}\mathbf{J}^j.$$

As $\mathbf{J}^j \cdot \mathbf{b}_\ell = \sum_{p=0}^{j}\binom{j}{p}\mathbf{b}_{\ell-j+2p}$, and $\mathbf{b}_q \cdot \mathbf{b}_{q'} = \mathbb{I}_{\{q=q'\}}$, this provides

$$\begin{aligned}
\mathbf{b}_q \cdot e^{\alpha\mathbf{J}-\beta\mathbf{D}}\mathbf{b}_\ell &= \sum_{k\geq 0}\frac{1}{k!}\sum_{j=0}^{k}\binom{k}{j}\alpha^j(-\beta)^{k-j}\ell^{2(k-j)}\sum_{p=0}^{j}\binom{j}{p}\mathbb{I}_{\{\ell-j+2p=q\}} \\
&= \sum_{k\geq 0}\sum_{j=0}^{k}\sum_{p=0}^{j}\frac{\alpha^j(-\beta\ell^2)^{k-j}}{p!(k-j)!(j-p)!}\mathbb{I}_{\{\ell-j+2p=q\}} \\
&= \sum_{p\geq 0}\sum_{j\geq p}\frac{\alpha^j}{p!(j-p)!}\,e^{-\beta\ell^2}\mathbb{I}_{\{\ell=q+j-2p\}} = \sum_{p\geq 0}\sum_{j\geq 0}\frac{\alpha^{j+p}}{p!j!}\,e^{-\beta\ell^2}\mathbb{I}_{\{\ell=q+j-p\}},
\end{aligned}$$

and we finally conclude that

$$U(q) = \sum_{p\geq 0}\sum_{j\geq 0}\frac{\alpha^{j+p}}{p!j!}\,e^{-\beta(q+j-p)^2}\mathbb{I}_{\{|q+j-p|\leq\bar{q}\}}, \text{ for } |q| \leq \bar{q}.$$

A.4 | Dynamic programming principle and representation

For all $\mathbb{F}$-stopping time τ with values in $[t, T]$ and for any $\mu \in \mathcal{A}_\tau$, we define[11]

$$J_T(\tau, \mu) = \mathbb{E}_\tau^\mu \left[-e^{-\gamma \int_\tau^T (\mu_u^a dN_u^a + \mu_u^b dN_u^b + Q_u dS_u)} e^{-\gamma \xi} \right], \text{ and } \mathcal{J}_{\tau,T} = (J_T(\tau, \mu))_{\mu \in \mathcal{A}_\tau},$$

where $\mathcal{A}_\tau$ denotes the restriction of $\mathcal{A}$ to controls on $[\tau, T]$. The continuation utility of the market maker is defined for all $\mathbb{F}$-stopping time τ by

$$V_\tau = \operatorname*{ess\,sup}_{\mu \in \mathcal{A}_{\iint}} J_T(\tau, \mu).$$

Lemma A.3. *Let τ be an $\mathbb{F}$-stopping time with values in $[t, T]$. Then, there exists a nondecreasing sequence $(\mu^n)_{n \in \mathbb{N}}$ in $\mathcal{A}_\tau$ such that $V_\tau = \lim_{n \to +\infty} \uparrow J_T(\tau, \mu^n)$.*

Proof. For μ and μ' in $\mathcal{A}_\tau$, define $\hat{\mu} = \mu \mathbb{1}_{\{J_T(\tau,\mu) \geq J_T(\tau,\mu')\}} + \mu' \mathbb{1}_{\{J_T(\tau,\mu) < J_T(\tau,\mu')\}}$. Then $\hat{\mu} \in \mathcal{A}_\tau$ and by definition of $\hat{\mu}$, we have $J_T(\tau, \hat{\mu}) \geq \max(J_T(\tau, \mu), J_T(\tau, \mu'))$. This shows that $\mathcal{J}_{\tau,T}$ is directly upward, and the required result follows from Neveu (1972, proposition VI.I.I, p. 121). □

Lemma A.4. *Let $t \in [0, T]$ and τ be an $\mathbb{F}$-stopping time with values in $[t, T]$. Then,*

$$V_t = \operatorname*{ess\,sup}_{\delta \in \mathcal{A}} \mathbb{E}_t^\delta \left[e^{-\gamma \int_t^\tau (\delta_u \cdot dN_u + Q_u dS_u)} V_\tau \right].$$

Proof. We follow the same argument as in Cvitanić & Karatzas (1993, proof of proposition 6.2). Denote $\tilde{V}_t$ the right-hand side of the required equality. First, by the tower property,

$$V_t = \operatorname*{ess\,sup}_{\delta \in \mathcal{A}} \mathbb{E}_t^\delta \left[e^{-\gamma \int_t^\tau (\delta_u \cdot dN_u + Q_u dS_u)} \mathbb{E}_\tau^\delta \left[-e^{-\gamma \left(\int_\tau^T (\delta_u \cdot dN_u + Q_u dS_u) + \xi \right)} \right] \right].$$

For all $\delta \in \mathcal{A}$, the quotient $\frac{L_T^\delta}{L_\tau^\delta}$ does not depend on the values of δ before time τ. Then,

$$\begin{aligned} \mathbb{E}_\tau^\delta \left[-e^{-\gamma \left(\int_\tau^T (\delta_u \cdot dN_u + Q_u dS_u) + \xi \right)} \right] &= \mathbb{E}_\tau^0 \left[-\frac{L_T^\delta}{L_\tau^\delta} e^{-\gamma \left(\int_\tau^T (\delta_u \cdot dN_u + Q_u dS_u) + \xi \right)} \right] \\ &\leq \operatorname*{ess\,sup}_{\mu \in \mathcal{A}_\tau} \mathbb{E}_\tau^\mu \left[-e^{-\gamma \left(\int_\tau^T (\mu_u \cdot dN_u + Q_u dS_u) + \xi \right)} \right] = V_\tau, \end{aligned}$$

which implies that $V_t \leq \tilde{V}_t$.

We next prove the reverse inequality. Let $\delta \in \mathcal{A}$ and $\mu \in \mathcal{A}_\tau$. We define $(\delta \otimes_\tau \mu)_u = \delta_u \mathbf{1}_{0 \leq u < \tau} + \mu_u \mathbf{1}_{\tau \leq u \leq T}$. Then $\delta \otimes_\tau \mu \in \mathcal{A}$ and

$$\begin{aligned} V_t &\geq \mathbb{E}_t^{\delta \otimes_\tau \mu} \left[-e^{-\gamma \left(\int_t^\tau (\delta_u \cdot dN_u + Q_u dS_u) + \int_\tau^T (\mu_u \cdot dN_u + Q_u dS_u) \right)} e^{-\gamma \xi} \right] \\ &= \mathbb{E}_t^{\delta \otimes_\tau \mu} \left[e^{-\gamma \int_t^\tau (\delta_u \cdot dN_u + Q_u dS_u)} \mathbb{E}_\tau^{\delta \otimes_\tau \mu} \left[-e^{-\gamma \int_\tau^T (\mu_u \cdot dN_u + Q_u dS_u)} e^{-\gamma \xi} \right] \right]. \end{aligned} \tag{A.2}$$

From Bayes' Formula and by noticing that $\frac{L_T^{\delta\otimes_\tau\mu}}{L_\tau^{\delta\otimes_\tau\mu}} = \frac{L_T^\mu}{L_\tau^\mu}$, we get

$$\mathbb{E}_\tau^{\delta\otimes_\tau\mu}\left[-e^{-\gamma\int_\tau^T(\mu_u\cdot dN_u+Q_udS_u)}e^{-\gamma\xi}\right] = \mathbb{E}_\tau^0\left[\frac{L_T^{\delta\otimes_\tau\mu}}{L_\tau^{\delta\otimes_\tau\mu}}\left(-e^{-\gamma\int_\tau^T(\mu_u\cdot dN_u+Q_udS_u)}e^{-\gamma\xi}\right)\right] = J_T(\tau,\mu).$$

Thus, Inequality (A.2) becomes $V_t \geq \mathbb{E}_t^{\delta\otimes_\tau\mu}[e^{-\gamma\int_t^\tau(\delta_u\cdot dN_u+Q_udS_u)}J_T(\tau,\mu)]$, and by using again Bayes' formula and by noticing that $\frac{L_\tau^{\delta\otimes_\tau\mu}}{L_t^{\delta\otimes_\tau\mu}} = \frac{L_\tau^\delta}{L_t^\delta}$, we have

$$\begin{aligned}
V_t &\geq \frac{\mathbb{E}_t^0\left[L_T^{\delta\otimes_\tau\mu}e^{-\gamma\int_t^\tau(\delta_u\cdot dN_u+Q_udS_u)}J_T(\tau,\mu)\right]}{L_t^{\delta\otimes_\tau\mu}} \\
&= \mathbb{E}_t^0\left[\mathbb{E}_\tau^0\left[\frac{L_T^{\delta\otimes_\tau\mu}}{L_\tau^{\delta\otimes_\tau\mu}}\frac{L_\tau^{\delta\otimes_\tau\mu}}{L_t^{\delta\otimes_\tau\mu}}e^{-\gamma\int_t^\tau(\delta_u\cdot dN_u+Q_udS_u)}J_T(\tau,\mu)\right]\right] \\
&= \mathbb{E}_t^0\left[\mathbb{E}_\tau^0\left[\frac{L_T^{\delta\otimes_\tau\mu}}{L_\tau^{\delta\otimes_\tau\mu}}\right]\frac{L_\tau^{\delta\otimes_\tau\mu}}{L_t^{\delta\otimes_\tau\mu}}e^{-\gamma\int_t^\tau(\delta_u\cdot dN_u+Q_udS_u)}J_T(\tau,\mu)\right] \\
&= \mathbb{E}_t^0\left[\frac{L_\tau^{\delta\otimes_\tau\mu}}{L_t^{\delta\otimes_\tau\mu}}e^{-\gamma\int_t^\tau(\delta_u\cdot dN_u+Q_udS_u)}J_T(\tau,\mu)\right] = \mathbb{E}_t^\delta\left[e^{-\gamma\int_t^\tau(\delta_u\cdot dN_u+Q_udS_u)}J_T(\tau,\mu)\right].
\end{aligned}$$

Since the previous inequality holds for all $\mu \in \mathcal{A}_\tau$ we deduce from monotone convergence theorem together with Lemma A.3 that there exists a sequence $(\mu^n)_{n\in\mathbb{N}}$ in $\mathcal{A}_\tau$ such that

$$\begin{aligned}
V_t &\geq \lim_{n\to+\infty}\uparrow \mathbb{E}_t^\delta\left[e^{-\gamma\int_t^\tau(\delta_u\cdot dN_u+Q_udS_u)}J_T(\tau,\mu^n)\right] \\
&= \mathbb{E}_t^\delta\left[e^{-\gamma\int_t^\tau(\delta_u\cdot dN_u+Q_udS_u)}\lim_{n\to+\infty}\uparrow J_T(\tau,\mu^n)\right] = \mathbb{E}_t^\delta\left[e^{-\gamma\int_t^\tau(\delta_u\cdot dN_u+Q_udS_u)}V_\tau\right],
\end{aligned}$$

thus concluding the proof. □

Proof of Theorem 3.1 (i). We proceed in several steps.

Step 1. For $\delta \in \mathcal{A}$, it follows from the dynamic programming principle of Lemma A.4 that the process $U_t^\delta := V_t\, e^{-\gamma\int_0^t(\delta_u^a dN_u^a+\delta_u^b dN_u^b+Q_udS_u)}$, $t \in [0,T]$, defines a $\mathbb{P}^\delta$-supermartingale[12] for all $\delta \in \mathcal{A}$. By standard analysis,[13] we may then consider it in its càdlàg version (by taking right limits along rationals). By the Doob–Meyer decomposition, we write

$$U_t^\delta = M_t^\delta - A_t^{\delta,c} - A_t^{\delta,d}, \tag{A.3}$$

where M^δ is a $\mathbb{P}^\delta$-martingale and $A^\delta = A^{\delta,c} + A^{\delta,d}$ is an integrable nondecreasing predictable process such that $A_0^{\delta,c} = A_0^{\delta,d} = 0$, with pathwise continuous component $A^{\delta,c}$, and a piecewise constant predictable process $A^{\delta,d}$. By the martingale representation theorem under $\mathbb{P}^\delta$, see Lemma A.1, we have

$$M_t^\delta = V_0 + \int_0^t \widetilde{Z}_r^\delta . d\chi_r - \int_0^t \widetilde{Z}_r^{\delta,a}\lambda(\delta_r^a)\mathbb{1}_{\{Q_r>-\bar{q}\}}dr - \int_0^t \widetilde{Z}_r^{\delta,b}\lambda(\delta_r^b)\mathbb{1}_{\{Q_r<\bar{q}\}}dr, \quad \text{(A.4)}$$

predictable process $\widetilde{Z}^\delta = (\widetilde{Z}^{\delta,S}, \widetilde{Z}^{\delta,a}, \widetilde{Z}^{\delta,b})$, where we recall that $\chi = (S, N^a, N^b)$.

Step 2. We show that V is a negative process. In fact, thanks to the uniform boundedness of $\delta \in \mathcal{A}$, we show that

$$\frac{L_T^\delta}{L_t^\delta} \geq \alpha_{t,T} = e^{-\frac{k\delta_\infty}{\sigma}(N_T^a - N_t^a + N_T^b - N_t^b) - 2Ae^{-\frac{kc}{\sigma}}(e^{\frac{k\delta_\infty}{\sigma}}+1)(T-t)} > 0, \quad \text{(A.5)}$$

which implies that $V_t \leq \mathbb{E}^0[-\alpha_{t,T}e^{-\gamma(\delta_\infty(N_T^a-N_t^a+N_T^b-N_t^b)+\int_t^T Q_u dS_u)}e^{-\gamma\xi}] < 0$.

Step 3. Let Y be the process defined by $V_t = -e^{-\gamma Y_t}$ for all $t \in [0,T]$. As $A^{\delta,d}$ is a predictable point process and the jumps of (N^a, N^b) are totally inaccessible stopping times under $\mathbb{P}^0$, we have $[N^a, A^{\delta,d}] = 0$ and $[N^b, A^{\delta,d}] = 0$ a.s, see proposition I.2.24 in Jacod and Shiryaev (2013). Using Itô's formula, we obtain from (A.3) and (A.4) that

$$Y_T = \xi, \text{ and } dY_t = Z_t^a dN_t^a + Z_t^b dN_t^b + Z_t^S dS_t - dI_t - d\widetilde{A}_t^d,$$

where $Z^a, Z^b, Z^S, I, \widetilde{A}^d$ are independent of δ, as they may be expressed as $Z_t^i dN_t^i = d[Y, N^i]_t$, $i \in \{a,b\}$, $Z_t^S\sigma^2 dt = d\langle Y_t, S_t\rangle_t$, $\widetilde{A}^d$ the predictable pure jumps of Y. Moreover, Itô's formula yields

$$Z_t^a = -\frac{1}{\gamma}\log\left(1 + \frac{\widetilde{Z}_t^{\delta,a}}{U_{t-}^\delta}\right) - \delta_t^a, \quad Z_t^b = -\frac{1}{\gamma}\log\left(1 + \frac{\widetilde{Z}_t^{\delta,b}}{U_{t-}^\delta}\right) - \delta_t^b, \quad Z_t^S = -\frac{\widetilde{Z}_t^{\delta,b}}{\gamma U_{t-}^\delta} - Q_{t-},$$

and

$$I_t = \int_0^t \left(\overline{h}(\delta_r, Z_r, Q_r)dr - \frac{1}{\gamma U_r^\delta}dA_r^{\delta,c}\right), \quad \widetilde{A}_t^d = \frac{1}{\gamma}\sum_{s\leq t}\log\left(1 - \frac{\Delta A_t^{\delta,d}}{U_{t-}^\delta}\right)$$

with $\overline{h}(\delta, z, q) = h(\delta, z, q) - \frac{1}{2}\gamma\sigma^2(z^s)^2$. In particular, the last relation between $\widetilde{A}^d$ and $A^{\delta,d}$ shows that $\Delta a_t = \frac{-\Delta A_t^{\delta,d}}{U_{t-}^\delta} \geq 0$ is independent of $\delta \in \mathcal{A}$; recall that $U^\delta < 0$.

In the subsequent steps, we argue that $Z = (Z^S, Z^a, Z^b) \in \mathcal{Z}$, and

$$A_t^{\delta,d} = -\sum_{s\leq t} U_{s-}^\delta \Delta a_s = 0, \text{ (so that } \widetilde{A}_t^d = 0\text{), and } I_t = \int_0^t \overline{H}(Z_r, Q_r)dr, \; t \in [0,T], \quad \text{(A.6)}$$

where $\overline{H}(z,q) = H(z,q) - \frac{1}{2}\gamma\sigma^2(z^s)^2$.

Step 4. Since $V_T = -1$, notice that

$$0 = \sup_{\delta\in\mathcal{A}} \mathbb{E}^\delta \left[U_T^\delta\right] - V_0 = \sup_{\delta\in\mathcal{A}} \mathbb{E}^\delta \left[U_T^\delta - M_T^\delta\right]$$
$$= \gamma \sup_{\delta\in\mathcal{A}} \mathbb{E}^0 \left[L_T^\delta \int_0^T U_{r-}^\delta \left(dI_r - \overline{h}(\delta_r, Z_r, Q_r)dr + \frac{da_r}{\gamma}\right)\right]. \tag{A.7}$$

Moreover, since the controls are uniformly bounded, we have

$$U_t^\delta \leq -\beta_t \ := \ V_t e^{-\gamma\delta_\infty(N_t^a - N_0^a + N_t^b - N_0^b) - \gamma\int_0^t Q_r dS_r} < 0. \tag{A.8}$$

Then, since $A^{\delta,d} \geq 0$, $U^\delta \leq 0$, and $dI_t - \overline{h}(\delta_t, Z_t, Q_t) \geq 0$, it follows from (A.7) together with the inequalities (A.5) and (A.8) that

$$0 \leq \sup_{\delta\in\mathcal{A}} \mathbb{E}^0 \left[\alpha_{0,T} \int_0^T -\beta_{r-} \left(dI_r - \overline{h}(\delta_r, Z_r, Q_r)dr + \frac{da_r}{\gamma}\right)\right]$$
$$= -\mathbb{E}^0 \left[\alpha_{0,T} \int_0^T \beta_{r-} \left(dI_r - \overline{H}(Z_r, Q_r)dr + \frac{da_r}{\gamma}\right)\right].$$

As $\alpha_{0,T} \int_0^T \beta_{r-}(dI_r - \overline{H}(Z_r, Q_r)dr) \geq 0$ and $\alpha_{0,T} \int_0^T \beta_r da_r \geq 0$, this implies (A.6).

Step 5. We now prove that $Z \in \mathcal{Z}$ by showing that

$$\sup_{\delta\in\mathcal{A}} \sup_{t\in[0,T]} \mathbb{E}^\delta[e^{-\gamma(p+1)Y_t}] < \infty \quad \text{for some } p > 0. \tag{A.9}$$

Using Hölder inequality together with Condition (7) and the boundedness of the intensities of N^a and N^b, we have that $\sup_{\delta\in\mathcal{A}} \mathbb{E}^\delta[|U_T^\delta|^{p'+1}] < \infty$ for some $p' > 0$. Hence,

$$\sup_{\delta\in\mathcal{A}} \sup_{t\in[0,T]} \mathbb{E}^\delta[|U_t^\delta|^{p'+1}] = \sup_{\delta\in\mathcal{A}} \mathbb{E}^\delta[|U_T^\delta|^{p'+1}] < \infty,$$

because U^δ is a negative P^δ-supermartingale. This leads to (A.9) using Hölder inequality, the uniform boundedness of the intensities of N^a and N^b and that $e^{-\gamma Y} = U^\delta e^{\gamma\int_0^\cdot(\delta_u^a dN_u^a + \delta_u^b dN_u^b + Q_u dS_u)}$.

Step 6. We finally prove uniqueness of the representation. Let $(Y_0, Z), (Y_0', Z') \in \mathbb{R} \times \mathcal{Z}$ be such that $\xi = Y_T^{Y_0,Z} = Y_T^{Y_0',Z'}$. By following the line of the verification argument in the proof of Theorem 3.1(ii), we obtain the equality $Y_t^{Y_0,Z} = Y_t^{Y_0',Z'}$ by considering the value of the continuation utility of the market maker

$$-\exp\left(-\gamma Y_t^{Y_0,Z}\right) = -\exp\left(-\gamma Y_t^{Y_0',Z'}\right) = \operatorname*{ess\,sup}_{\delta\in\mathcal{A}} \mathbb{E}_t^\delta[-e^{-\gamma(\mathrm{PL}_T^\delta - \mathrm{PL}_t^\delta + \xi)}], \ \ t \in [0,T].$$

This, in turn, implies that $Z_t^i dN_t^i = Z'^i_t dN_t^i = d[Y^{Y_0,Z}, N^i]_t$, $i \in \{a,b\}$, and $Z_t^S \sigma^2 dt = Z'^S_t \sigma^2 dt = d\langle Y, S\rangle_t$, $t \in [0,T]$. Hence, $(Y_0, Z) = (Y_0', Z')$.

□

A.5 | Proof of Theorem 4.3

A.5.1 | Informal verification theorem

We begin to write the HJB equation associated with the principal's problem and we show informally that it reduces to solve (14).

It turns out that the value function of the principal does not depend on s, and this implies that the optimal contract, defined by the optimal choice of Z^a, Z^b, Z^S does not depend on s. The reason why the principal's value function does not depend on s is that the model parameters are independent of s. Indeed, starting from a value function depending on (t, s, y, n^a, n^b), it follows from standard stochastic control theory that the corresponding HJB equation is

$$
(\mathrm{H})\begin{cases}
0 = \partial_t V(t, n^a, n^b, s, y) + \frac{1}{2}\sigma^2\partial_{ss}V \\
\quad + \sup_{z^S}\left\{ +\partial_{sy}Vz^S\sigma + \frac{1}{2}\partial_{yy}V|z^S\sigma|^2 + \partial_y V\frac{1}{2}\gamma\sigma^2(z^S + n^b - n^a)^2 \right\} \\
\quad + \sup_{z^a}\left\{ \lambda(\Delta(z^a))\left(V(t, n^a+1, n^b, s, y+z^a) - V(t, n^a, n^b, s, y) - \partial_y V\frac{1-e^{z^a+\Delta(z^a)}}{\gamma} \right) \right\} \\
\quad + \sup_{z^b}\left\{ \lambda(\Delta(z^b))\left(V(t, n^a, n^b+1, s, y+z^b) - V(t, n^a, n^b, s, y) - \partial_y V\frac{1-e^{z^b+\Delta(z^b)}}{\gamma} \right) \right\} \\
V(T, n^a, n^b, s, y) = -e^{-\eta(c(n^a+n^b)-y)}.
\end{cases}
$$

As none of the coefficients of the last PDE depends on s, we guess that the solution does not depend on s. Moreover, given the form of the objective function, we search for a solution of the form

$$
V(t, n^a, n^b, s, y) := v(t, q)e^{-\eta(c(n^a+n^b)-y)}
$$

for some function v with $q = n^b - n^a$. In this case, $\partial_y V = \eta V,\ \partial_{yy}V = -\eta^2 V$, and (H) reduces to

$$
\begin{cases}
0 = \partial_t v(t, q) + \sup_{z^S}\left\{ \frac{\sigma^2\eta}{2}v(t, q)\big(\gamma(z^S+q)^2 - |z^S|^2\big) \right\} \\
+ \sum_{i\in\{a,b\}} \sup_{z^i}\left\{ \lambda(\Delta(z^i))\left(v(t, q+\epsilon_i)e^{-\eta(c-z^i)} - v(t, q)(1+\eta\frac{1-e^{z^i+\Delta(z^i)}}{\gamma} \right) \right\}. \\
v(T, q) = -1.
\end{cases}
$$

We now know that this PDE coincides with PDE (14) which admits a unique solution (see Proposition 4.2 together with $v = -u^{-\frac{\sigma\eta}{k}}$) inducing optimal controls as the optimizers of the Hamiltonian for the exchange problem.

A.5.2 | Rigorous proof of Theorem 4.3

The proof of the main result of Theorem 4.3 requires the following technical result. We observe that this is the place where the first integrability condition in (7) is needed.

Lemma A.5. *Let $Z \in \mathcal{Z}$. There exists $C > 0$ and $\varepsilon > 0$ such that*

$$
\sup_{t\in[0,T]} \mathbb{E}^{\hat\delta(Y_T^{\hat Y_0, Z})}[|K_t^Z|^{1+\varepsilon}] \le C.
$$

Proof. We use again the notation $K_t^Z := e^{-\eta(c(N_t^a+N_t^b)-Y_t^{0,Z})}$, $t\in[0,T]$, for all $Z\in\mathcal{Z}$. Let $p>1$. By using Hölder's inequality and the uniform boundedness of the intensities of N^a and N^b, we deduce that there exists $C'>0$ such that

$$\mathbb{E}^{\hat{\delta}(Y_T^{\hat{Y}_0,Z})}[|K_t^Z|^p]\leq C'\mathbb{E}^0\left[(e^{-\gamma Y_t^{0,Z}})^{-\frac{p'\eta}{\gamma}}\right]^{\frac{p}{p'}}$$

with any $p'>p$. Thus,

$$\begin{aligned}\mathbb{E}^{\hat{\delta}(Y_T^{\hat{Y}_0,Z})}\left[|K_t^Z|^p\right]&\leq C'\left(1+\mathbb{E}^0\left[(e^{-\gamma Y_t^{0,Z}})^{-\frac{p'\eta}{\gamma}}\right]\right)\\&=C'\left(1+\mathbb{E}^0\left[\left(-\sup_{\delta\in\mathcal{A}}\mathbb{E}_t^\delta\left[-e^{-\gamma(Y_T^{0,Z}+PL_T^\delta-PL_t^\delta)}\right]\right)^{-\frac{p'\eta}{\gamma}}\right]\right).\end{aligned}$$

By Jensen's inequality and Hölder's inequality, we deduce for any $p''>p'$ that

$$\mathbb{E}^{\hat{\delta}(Y_T^{\hat{Y}_0,Z})}\left[|K_t^Z|^p\right]\leq C'\left(1+\mathbb{E}^0\left[\sup_{\delta\in\mathcal{A}}\mathbb{E}_t^\delta[e^{p'\eta(Y_T^{0,Z}+PL_T^\delta-PL_t^\delta)}]\right]\right)\leq C'\left(1+\mathbb{E}^0\left[\sup_{\delta\in\mathcal{A}}\mathbb{E}_t^\delta\left[e^{p''\eta Y_T^{0,Z}}\right]\right]\right).$$

By using a dynamic programming principle, similarly to the proof of Lemma A.4 by noticing that the family $(\widetilde{J}(\mu,\delta)=\mathbb{E}_\tau^\delta[e^{p''\eta Y_T^{0,Z}}])_{\mu\in\mathcal{A}_{\int\!\int}}$ is directly upward, we get

$$\mathbb{E}^{\hat{\delta}(Y_T^{\hat{Y}_0,Z})}\left[|K_t^Z|^p\right]\leq C'\left(1+\sup_{\delta\in\mathcal{A}}\mathbb{E}^\delta\left[e^{p''\eta Y_T^{0,Z}}\right]\right).$$

By setting $\varepsilon=\frac{\eta'-\eta}{3}$, if we take $p=1+\varepsilon$, then $p'=p+\varepsilon$ and $p''=p'+\varepsilon$, we obtain

$$\mathbb{E}^{\hat{\delta}(Y_T^{\hat{Y}_0,Z})}[|K_t^Z|^{1+\varepsilon}]\leq C'\left(1+\sup_{\delta\in\mathcal{A}}\mathbb{E}^\delta\left[e^{\eta' Y_T^{0,Z}}\right]\right).$$

From the definition of $\mathcal{Z}$ (involving the first integrability condition in (7)), we get

$$\mathbb{E}^{\hat{\delta}(Y_T^{\hat{Y}_0,Z})}\left[|K_t^Z|^{1+\varepsilon}\right]\leq C,\ t\in[0,T],\text{ with }C=C'\left(1+\sup_{\delta\in\mathcal{A}}\mathbb{E}^\delta\left[e^{\eta' Y_T^{0,Z}}\right]\right)<+\infty.$$

□

Proof of Theorem 4.3. In order to prove the theorem, we verify that the function v introduced in (20) coincides at $(0,Q_0)$ with the value function of the reduced exchange problem (13), with maximum achieved at the optimal control $\hat{Z}$.

The function v is negative bounded and has bounded gradient. Moreover, since $\delta_\infty\geq\Delta_\infty$, it follows that v is a solution of the HJB equation (14) of the exchange reduced problem (see

Lemma A.2). For $Z \in \mathcal{Z}$, denote

$$K_t^Z = e^{-\eta\left(c(N_t^a - N_0^a + N_t^b - N_0^b) - Y_t^{0,Z}\right)}, \; t \in [0,T].$$

By direct application of Itô's formula, and substitution of $\partial_t v$ from the HJB equation satisfied by v, we see that

$$\begin{aligned} d\left[v(t,Q_t)K_t^Z\right] = K_{t-}^Z\Big(\left(h_t^Z - \mathcal{H}_t\right)dt + \eta v(t,Q_t)Z_t^s dS_t \\ + \sum_{i=a,b}\left[v(t,Q_{t-} + \Delta Q_t)e^{-\eta(c - Z_t^i)} - v(t,Q_{t-})\right] d\widetilde{N}_t^{\hat{\delta}(Y^{\hat{Y}_0,Z}),i}\Big), \end{aligned} \tag{A.10}$$

where, using the notations of (15) and the subsequent equations,

$$\begin{aligned} \mathcal{H}_t &:= H_E(Q_t, v(t,Q_t), v(t,Q_t+1), v(t,Q_t-1)), \\ h_t^Z &:= h_E^1(Q_t, v(t,Q_t), Z_t^S) + \mathbb{1}_{\{Q_t > -\bar{q}\}} h_E^0\left(v(t,Q_t), v(t,Q_t-1), Z_t^a\right) \\ &\quad + \mathbb{1}_{\{Q_t < \bar{q}\}} h_E^0\left(v(t,Q_t), v(t,Q_t+1), Z_t^b\right). \end{aligned}$$

Exploiting the fact that v is bounded and that K^Z is uniformly integrable, see Lemma A.5, we get that $(v(t,Q_t)K_t^Z)_{t\in[0,T]}$ is a $\mathbb{P}^{\hat{\delta}(Y_T^{\hat{Y}_0,Z})}$-supermartingale and by Doob–Meyer decomposition theorem, the local martingale term in (A.10) is a true martingale. Hence,

$$\begin{aligned} v(0,Q_0) &= \mathbb{E}^{\hat{\delta}(Y_T^{\hat{Y}_0,Z})}\left[v(T,Q_T)K_T^Z + \int_0^T K_t^Z(\mathcal{H}_t - h_t^Z)dt\right] \\ &\geq \mathbb{E}^{\hat{\delta}(Y_T^{\hat{Y}_0,Z})}\left[v(T,Q_T)K_T^Z\right] = \mathbb{E}^{\hat{\delta}(Y_T^{\hat{Y}_0,Z})}[-K_T^Z], \end{aligned}$$

by the boundary condition $v(T,.) = -1$. By arbitrariness of $Z \in \mathcal{Z}$, this provides the inequality $v(0,Q_0) \geq \sup_{Z\in\mathcal{Z}} \mathbb{E}^{\hat{\delta}(Y_T^{Y^0,Z})}[-K_T^Z] = v_0^E$. □

On the other hand, consider the maximizer $\hat{Z}$ of the reduced exchange problem, induced by the feedback controls $\hat{z}$ in (16). As $\hat{Z}$ is bounded, it follows that $\hat{Z} \in \mathcal{Z}$. Moreover, $h^{\hat{Z}} - \mathcal{H} = 0$, by definition, so that the last argument leads to the equality $v(0,Q_0) = \mathbb{E}^{\hat{\delta}(Y_T^{\hat{Y}_0,\hat{Z}})}[-K_T^{\hat{Z}}]$, instead of the inequality. This shows that $v(0,Q_0) = v_0^E$, the reduced exchange problem of (13), with optimal control $\hat{Z}$. From Theorem 3.1, the corresponding optimal market maker response of the market maker is given by (10) with $\xi = Y_T^{\hat{Y}_0,\hat{Z}}$. Moreover, Condition (17) implies that $|-Z_t^i + \frac{1}{\gamma}\log(1 + \frac{\sigma k}{k})| \leq \delta_\infty, \quad i = a,b$. Hence, the optimal effort can be reduced to (24).

A.6 | Proof of Theorem 6.5

The proof of Theorem 6.5 is divided in two steps. First, we show that there exists a symmetric and Markovian Nash equilibrium for the problem (29). Then, we show that this Nash equilibrium is unique among the class of symmetric Nash equilibria.

A.6.1 | Existence of a symmetric Markovian Nash equilibrium

We denote by v a smooth solution to (32) or equivalently (33). Note that ζ^0 as defined in Theorem 6.5 is a deterministic function of t, Q_t. Denote by $\mathcal{K}_t^\zeta$ the process defined for any $\zeta \in \mathcal{Z}$ by

$$\mathcal{K}_t^\zeta := e^{\eta\left\{\int_0^T \sum_{i=a,b}(\zeta_t^i - \frac{c}{N})dN_t^i + \zeta_t^S dS_t + [\hat{H}^S(Q_t, \zeta_t^{S,0}, \zeta_t^S) - \sum_{i=a,b}\hat{H}^i(Q_t, \zeta_t^{i,0}, \zeta_t^i)) dt\right\}},$$

with

$$\hat{H}^i(q, \tilde{z}, z) = H^i(z + (N-1)\tilde{z}, q) - \frac{N-1}{N} H^i(N\tilde{z}, q). \tag{A.11}$$

Note that

$$\begin{aligned} d\mathcal{K}_t^\zeta &= \eta \mathcal{K}_t^\zeta \zeta_t^S dS_t + \sum_{i=a,b} \mathcal{K}_{t-}^\zeta (e^{\eta(\zeta_t^i - \frac{c}{N})} - 1) dN_t^i \\ &\quad + \eta K_t^\zeta \left(\hat{H}^S\left(Q_t, \zeta_t^{S,0}, \zeta_t^S\right) - \sum_{i=a,b} \hat{H}^i\left(Q_t, \zeta_t^{i,0}, \zeta_t^i\right) + \frac{\eta}{2}\sigma^2 |\zeta_t^S|^2 \right) dt. \end{aligned}$$

Applying Ito's formula, we obtain

$$\begin{aligned} d(v(t,Q_t)\mathcal{K}_t^\zeta) &= \mathcal{K}_t^\zeta \partial_t v(t,Q_t) dt \\ &\quad + \mathcal{K}_t^\zeta \left(\eta v(t,Q_t) \left(\hat{H}^S(Q_t, \zeta^{S,0}(Q_t), \zeta_t^S) - \sum_{i=a,b} \hat{H}^i(Q_t, \zeta_t^{i,0}, \zeta_t^i) + \frac{\eta}{2}\sigma^2 |\zeta_t^S|^2 \right) \right) dt \\ &\quad + \eta \mathcal{K}_t^\zeta v(t,Q_t) \zeta_t^S dS_t + \mathcal{K}_{t-}^\zeta \sum_{i=a,b} (v(t, Q_t + \Delta Q_t) - v(t,Q_t)) e^{\eta(\zeta_t^i - \frac{c}{N})} dN_t^i \\ &= \mathcal{K}_t^\zeta \left(\partial_t v(t,Q_t) + \eta v(t,Q_t) F_t^\zeta \right) dt + \eta \mathcal{K}_t^\zeta v(t,Q_t) \zeta_t^S dS_t \\ &\quad + \mathcal{K}_{t-}^\zeta \sum_{i=a,b} (v(t, Q_t + \Delta Q_t) - v(t,Q_t)) e^{\eta(\zeta_t^i - \frac{c}{N})} d\tilde{N}_t^i, \end{aligned}$$

where

$$\begin{aligned} F_t^\zeta &= F^S(t, Q_t, \zeta_t^S) + F^0(t, Q_t, v(t,Q_t), v(t, Q_t+1), \zeta_t^b) \mathbf{1}_{Q_t < \overline{Q}} \\ &\quad + F^0(t, Q_t, v(t,Q_t), v(t, Q_t - 1), \zeta_t^a) \mathbf{1}_{Q_t > -\overline{Q}}. \end{aligned}$$

Since v satisfies HJB equation (32), we deduce that $\mathbb{E}^{\hat{\delta}(\zeta + (N-1)\zeta^0)}[-\mathcal{K}_T^\zeta] \leq v(0, Q_0)$, with equality for $\zeta^S = \zeta^{S,0}$ and $\zeta^i = \zeta^{i,0}$.

A.6.2 | Uniqueness among the set of general symmetric Nash equilibria

We now prove that if there exists a symmetric Nash equilibrium, it is unique and given by the Markovian equilibrium ξ^0 defined by (36). Let ξ^0 characterized by (30) for general ζ^0. We consider the dynamic value function of any exchange given ξ^0 fixed by the other, denoted by $V_t(\xi^0)$ and defined in view of Remark 6.4 by

$$V_t(\xi^0) = e^{\eta \hat{Y}_0} \operatorname*{ess\,sup}_{\zeta \in \mathcal{Z}} \mathbb{E}_t^{\hat{\delta}(\zeta+(N-1)\zeta^0)} \left[-e^{-\eta\left(\int_t^T \sum_{i=a,b} (\frac{c}{N} - \zeta_r^i) dN_r^i - \int_t^T \zeta_r^S dS_r - \int_t^T \hat{H}_r^S(\zeta_r^S) - \sum_{i=a,b} \hat{H}_r^i(\zeta_r^i) dr \right)} \right],$$

with $\hat{H}_r^S(z) = \hat{H}_r^S\left(Q_r, \zeta_r^{S,0}, z\right)$ and $\hat{H}_r^i(z) = \hat{H}_r^i\left(Q_r, \zeta_r^{i,0}, z\right)$. By using a DPP similarly to A.4, we prove that $\left(V_t(\xi^0)\mathcal{K}_t^\zeta\right)_{t\in[0,T]}$ is a $\mathbb{P}^{\zeta+(N-1)\zeta^0}$-super martingale. The martingale condition thus provides the optimal ζ played by the representative exchange given that the others choose ζ^0. By using a Doob–Meyer decomposition, the martingale property leads to the solution of the following Backward Stochastic Differential Equation (BSDE):

$$dR_t = U_t^S dS_t + \sum_{i=a,b} U_t^i d\widetilde{N_t^i} - F_t(U_t, Q_t)dt, \; R_T = 0 \tag{A.12}$$

with

$$F_t(u,q) := \sup_{\zeta^S} \left(-\hat{H}_t^S(\zeta^S) - \frac{\sigma^2\eta}{2}|\zeta^S - u^S|^2) \right) + \sum_{i=a,b} \sup_{\zeta^i} \left(\hat{H}_t^i(\zeta^i) - \lambda_t^{i,\zeta}\left(u^i + \frac{1-e^{-\eta(u^i-\zeta^i+\frac{c}{N})}}{\eta} \right) \right),$$

with $\lambda_t^{i,\zeta} = \lambda(\hat{\delta}(\zeta^i + (N-1)\zeta^{i,0}))$. We directly derive the maximizers

$$\zeta_t^{S,\star} = -\frac{\gamma(N-1)}{\eta+\gamma}\zeta_t^{S,0} - \frac{\gamma}{\eta+\gamma}Q_t - \frac{\eta}{\eta+\gamma}U_t^S,$$

$$\zeta_t^{i,\star} = U_t^i + \frac{c}{N} + \frac{1}{\eta}\log\left(\frac{k(k+\sigma(\gamma-\eta))}{(k+\sigma\eta)(k+\sigma\gamma)} + \frac{U_t^i k\eta}{k+\sigma\eta} \right).$$

Since ζ^0 is assumed to be a symmetric Nash equilibrium, we obtain from Definition 6.2 that $\zeta_t^{S,0}$ and $\zeta_t^{i,0}$ are necessarily uniquely determined as function of Q_t and U_t by

$$\zeta_t^{S,0} = -\frac{\gamma}{\eta+N\gamma}Q_t - \frac{\eta}{\eta+N\gamma}U_t^S,$$

and

$$\zeta_t^{i,0} = U_t^i + \frac{c}{N} + \frac{1}{\eta}\log\left(\frac{k(k+\sigma(\gamma-\eta))}{(k+\sigma\eta)(k+\sigma\gamma)} + \frac{U_t^i k\eta}{k+\sigma\eta} \right).$$

Hence, we note that the BSDE (A.12) is Markovian. The integro-partial differential equation associated with this BSDE remains to solve (32) for which we know that there exists a continuous solution given by $v(t,q)$. We thus deduce that if there is a symmetric Nash equilibrium, it is Markovian in the sense of Definition 6.3 and the first step of the proof shows that it is unique.

A.7 | First-best exchange problem

In this section, we analyze the first-best problem of the exchange. In this setting, the exchange optimally chooses the contract ξ and the optimal bid–ask posting policy of the market maker under her participation constraint. Introducing a Lagrange multiplier $\lambda > 0$ to penalize for this constraint, we reduced the first-best exchange value function to the unconstrained problem:

$$V_0^{FB} = \inf_{\lambda>0} \sup_{\xi\in\mathfrak{C},\delta\in\mathcal{A}} \mathbb{E}^{\mathbb{P}^\delta}\Big[-e^{-\eta(c(N_T^a-N_T^b)-\xi)} - \lambda e^{-\gamma(\xi+X_T+Q_TS_T)} - \lambda R\Big],$$

with $\mathfrak{C} = \{\xi,\ \mathcal{F}_T - \text{measurable such that (7) is satisfied}\}$.

We first compute the supremum on ξ by fixing λ, δ. The first-order condition in ξ is $e^{-\eta(c(N_T^a-N_T^b)-\xi_\lambda^\star)} = \frac{\lambda\gamma}{\eta} e^{-\gamma(\xi_\lambda^\star+X_T+Q_TS_T)}$, implying

$$\xi_\lambda^\star = \frac{1}{\eta+\gamma}\left(\log\left(\frac{\lambda\gamma}{\eta}\right) - \gamma(X_T+Q_TS_T) + \eta c\left(N_T^a+N_T^b\right)\right).$$

Substituting this expression, we see that

$$\begin{aligned}
V_0^{FB} &= \inf_{\lambda>0}\sup_{\delta\in\mathcal{A}} \mathbb{E}^{\mathbb{P}^\delta}\left[-\lambda\frac{\eta+\gamma}{\eta}e^{-\gamma(\xi_\lambda^\star+X_T+Q_TS_T)} - \lambda R\right] \\
&= \inf_{\lambda>0}\sup_{\delta\in\mathcal{A}} \mathbb{E}^{\mathbb{P}^\delta}\left[-\lambda\frac{\eta+\gamma}{\eta}\left(\frac{\eta}{\lambda\gamma}\right)^{\frac{\gamma}{\eta+\gamma}} e^{-\frac{\gamma\eta}{\gamma+\eta}(X_T+Q_TS_T+c(N_T^a+N_T^b))} - \lambda R\right] \\
&= \inf_{\lambda>0}\lambda\left[\frac{\eta+\gamma}{\eta}\left(\frac{\eta}{\lambda\gamma}\right)^{\frac{\gamma}{\eta+\gamma}}\widetilde{V_0} - R\right], \text{ with } \widetilde{V}_0 = \sup_{\delta\in\mathcal{A}}\mathbb{E}^{\mathbb{P}^\delta}\left[-e^{-\frac{\gamma\eta}{\gamma+\eta}(X_T+Q_TS_T+c(N_T^a+N_T^b))}\right].
\end{aligned}$$

As $\widetilde{V}_0$ is independent of λ, we obtain the optimal Lagrange multiplier $\lambda^\star = \frac{\eta}{\gamma}\left(\frac{\widetilde{V}_0}{R}\right)^{1+\frac{\eta}{\gamma}}$, and we deduce the optimal first-best contract:

$$\xi^\star = \xi_{\lambda^*}^\star = \frac{1}{\eta+\gamma}\left(\log\left(\frac{\lambda^\star\gamma}{\eta}\right) - \gamma(X_T+Q_TS_T) + \eta c\left(N_T^a+N_T^b\right)\right).$$

We finally solve the problem $\widetilde{V}_0$. Note that by setting $\tilde{\delta} := \delta + c$ in view of the definition of $\mathbb{P}^\delta$ given by (4) together with (2) we get

$$\widetilde{V}_0 = \sup_{\tilde{\delta}\in\tilde{\mathcal{A}}}\mathbb{E}^{\mathbb{P}^{\tilde{\delta}-c}}\left[-e^{-\Gamma(X_T+Q_TS_T)}\right], \text{ with } \Gamma = \frac{\gamma\eta}{\gamma+\eta},$$

and where $\tilde{\mathcal{A}}$ is defined similarly to $\mathcal{A}$ with bound $\delta_\infty + c$. We are then reduced to the framework of Avellaneda and Stoikov (2008) and Guéant et al. (2008) so that the optimal bid–ask spreads are given by

$$\widetilde{\delta}_t^i = -c + \frac{1}{\Gamma}\log\left(1+\frac{\sigma\Gamma}{k}\right) + \frac{\sigma}{k}\log\left(\frac{\widetilde{u}^{FB}(t,Q_{t-})}{\widetilde{u}^{FB}(t,Q_{t-}+\varepsilon_i)}\right),\ \ i\in\{b,a\},\ (\varepsilon_a,\varepsilon_b)=(-1,1),$$

where $\widetilde{u}^{FB}$ is the unique solution of the linear differential equation

$$\widetilde{u}^{FB}\Big|_{t=T} = 1,\ \ \partial_t\widetilde{u}^{FB}(t,q) - F_{C_1^{FB},\widetilde{C_1}^{FB}}(q,\widetilde{u}^{FB}(t,q),\widetilde{u}^{FB}(t,q+1),\widetilde{u}^{FB}(t,q-1)) = 0, \quad \text{(A.13)}$$

with constants $C_1^{FB} = \frac{\sigma\Gamma k}{2}$ and $\widetilde{C_1}^{FB} = A\left(1+\frac{\sigma\Gamma}{k}\right)^{-(1+\frac{\sigma\Gamma}{k})}$, and so that

$$v^{FB}(0,0) = \widetilde{V_0}^{FB},\ \text{with}\ \widetilde{u}^{FB} = (-\widetilde{v}^{FB})^{-\frac{k}{\sigma\eta}}.$$

Since the solution of PDE (A.13) is different from the solution of (19), we deduce that the value function of the exchange in the first-best case does not coincide with his value function in the second-best model.