Risk publications

REPRINTED FROM

Risk.net

RISK MANAGEMENT • DERIVATIVES • REGULATION

Risk.net September 2022

Market-making by a foreign exchange dealer

Market making by a foreign exchange dealer

Dealers make money by providing liquidity to clients but face flow uncertainty and thus price risk. They can efficiently skew their prices and wait for clients to mitigate risk (internalisation), or trade with other dealers in the open market to hedge their position and reduce their inventory (externalisation). Alexander Barzykin, Philippe Bergault and Olivier Guéant propose an optimal control framework for market making that tackles both pricing and hedging, thus answering a question well known to dealers: to hedge or not to hedge?

With more than US$6 trillion of trading turnover per day, the foreign exchange market is the largest financial market, far ahead of that of bonds and stocks. In spite of its size and the concentration of trading in a few financial global hubs, forex remains a highly fragmented over-the-counter (OTC) market with, on one side, a dealer-to-client (D2C) segment where dealers/market makers provide liquidity to clients and, on the other, a dealer-to-dealer (D2D) or interdealer segment where dealers trade together, mainly for hedging purpose.

Market makers in forex cash markets provide liquidity to customers by proposing prices at which they are ready to buy and sell currency pairs through electronic price streams, single-bank platforms, multi-bank platforms, etc. As a consequence of the trading flow from their clients, they have to manage risky positions. They can have two different behaviours: holding the risk until other clients come to offset it (internalisation) or hedging the risk out by trading on the D2D segment (externalisation). Externalisation allows market makers to get rid of the risk, but it usually comes at a cost, that of crossing the bid-ask spread and sometimes walking the book on platforms such as EBS (part of CME Group) or Refinitiv (depending on the currency pair). Furthermore, externalisation usually induces a market impact because trades become visible to more market participants. Internalisation allows market makers to avoid market impact, or at least reduce it, but this is of course risky for the market maker because the price might evolve adversely before the trading flow compensates the current position. The risk can be reduced by skewing prices to attract the flow in the required direction, but the flow is by no means guaranteed. In practice, most market makers both internalise and externalise, depending on the market conditions (an increase in volatility increases the propensity to externalise) and their positions (dealers usually externalise beyond a certain position limit).

The latest Bank for International Settlements (BIS) Triennial Survey documented the growing prevalence of internalisation and the resulting decline of the D2D segment (see Schrimpf & Sushko 2019). However, the trade-off between internalisation and externalisation has attracted little academic interest until recently. In fact, most of the models proposed in the literature on optimal OTC market making have assumed no way to hedge out the risk through the interdealer segment of the market. In the paper by Ho & Stoll (1981) and in the recent literature (see Cartea *et al* (2015) and Guéant (2016) for an overview) on optimal market making that followed the publication of Avellaneda & Stoikov (2008), the market maker is indeed 'only' proposing bid and ask quotes.[1] One of the rare references to the internalisation versus externalisation dilemma is Butz & Oomen (2019), which discusses internalisation on the basis of queuing theory and derives typical internalisation horizons.

In Barzykin *et al* (2021), we proposed a model of algorithmic market making with pricing and hedging that constitutes an important and natural encounter between two problems that have attracted a lot of academic and practitioner interest in the last decade: optimal market making and optimal execution.[2] The model allows us to set an optimal pricing ladder and determine optimal hedging rate in external liquidity pools as functions of the inventory, risk aversion and market-driven parameters. In particular, we proved the existence of a pure flow internalisation area, or equivalently, an inventory threshold below which it is optimal for the dealer not to externalise. This threshold is derived from a subtle balance between uncertainty, execution costs and market impact.

In this paper, we generalise our algorithmic market-making model further to better model the trading flow. In particular, we introduce tiers used by market makers to distinguish both the different sources of the trading flow and the natural diversity of the clients. We describe below our modelling approach to the trading flow and show how to estimate the intensity parameters. We demonstrate that tiers can be conveniently defined using clustering techniques on intensity parameters. We then present our algorithmic market-making model and look into the differences in the optimal strategies for different tiers. By analysing the typical risk-neutralisation time and internalisation ratio derived from the model as functions of the dealer's risk aversion we recover figures consistent with those of Butz & Oomen (2019) and Schrimpf & Sushko (2019). Finally, we discuss the dealer's efficient frontier and comment on the choice of the risk aversion parameter.

Understanding trading flow

One of the central issues for a dealer is inventory management. Indeed, a dealer must, at all times, decide whether they wish to warehouse the risk while waiting for the arrival of future customer flows or if they wish to hedge part of it by trading on the D2D market. This decision obviously depends on price risk but also on customer flow. Furthermore, when the market maker decides to hold risk (internalisation), they skew prices in order to increase or decrease the flow of buying or selling customers, depending on the sign of the inventory. Understanding trading flow and customers' sensitivity to streamed prices is therefore essential.

[1] *Some of these papers did not address OTC markets specifically but the models they proposed are more adapted to OTC markets than stock markets, which are mainly organised around all-to-all limit order books.*

[2] *For an introduction to optimal execution problems, we refer the reader to Almgren & Chriss (2001) as well as Cartea* et al *(2015) and Guéant (2016).*

In order to model customer flow as a function of streamed prices, let us introduce first a reference price process $(S_t)_t$. Following the industry standard, we take the firm primary mid-price as the reference price at any point in time (EBS for our examples with EUR/USD).[3] Given a streamed pricing ladder at the bid (respectively, ask/offer) modelled by $S^{\mathrm{b}}(t,z) = S_t(1-\delta^{\mathrm{b}}(t,z))$ (respectively, $S^{\mathrm{a}}(t,z) = S_t(1+\delta^{\mathrm{a}}(t,z))$), where $z>0$ is the size of the trade,[4] we assume that buy (respectively, sell) trades[5] of size in $[z, z+\mathrm{d}z]$ arrive over the infinitesimal interval $[t, t+\mathrm{d}t]$ with probability $\Lambda^{\mathrm{b}}(z,\delta^{\mathrm{b}}(t,z))\,\mathrm{d}z\,\mathrm{d}t$ (respectively, $\Lambda^{\mathrm{a}}(z,\delta^{\mathrm{a}}(t,z))\,\mathrm{d}z\,\mathrm{d}t$). In practice, dealers propose pricing ladders for a set $\{z_k,\ 1\leqslant k\leqslant K\}$ of sizes. In what follows, $K=6$ sizes are considered, corresponding to 1, 2, 5, 10, 20 and 50 million euros, respectively. As a consequence, the measures $\Lambda^{\mathrm{b/a}}(z,\delta)\,\mathrm{d}z$ are approximated by discrete measures $\sum_{k=1}^{K}\Lambda_k^{\mathrm{b/a}}(\delta)1_{z_k}(\mathrm{d}z)$ (where $1_{z_k}(\mathrm{d}z)$ is the Dirac measure in z_k). Hereafter, the functions $\Lambda_k^{\mathrm{b/a}}$ are called intensity functions or simply intensities.

For an anonymised sample of HSBC's forex streaming clients[6] trading EUR/USD we obtained access to tables of trades and quotes over the period from January to April 2021.[7] For the purpose of our study, quotes on the bid and ask sides can be summed up, for each size z_k, by a list of couples $((\delta_j,\tau_j))_{j\in\mathcal{J}_k^{\mathrm{b/a}}}$, where δ_j is a streamed quote for size z_k and τ_j is the associated duration of that quote. Trades are not all of sizes $\{z_k,\ 1\leqslant k\leqslant K\}$ but we can associate each trade with the closest z_k and the trade data can then be aggregated, for the bid and ask sides and for each size z_k, by a list of quotes $(\delta_i)_{i\in\mathcal{I}_k^{\mathrm{b/a}}}$.

It is easy to show that the loglikelihoods $\mathrm{LL}_k^{\mathrm{b/a}}$ associated with the bid and ask sides for size z_k are (up to an additive constant):

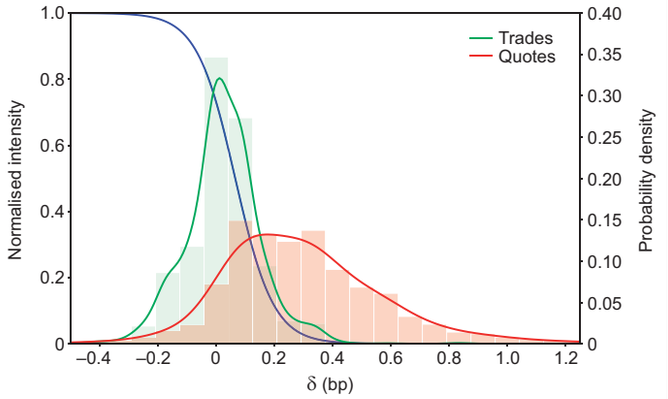
$$\begin{aligned}\mathrm{LL}_k^{\mathrm{b/a}} &= \sum_{i\in\mathcal{I}_k^{\mathrm{b/a}}}\log(\Lambda_k^{\mathrm{b/a}}(\delta^i)) - \sum_{j\in\mathcal{J}_k^{\mathrm{b/a}}}\Lambda_k^{\mathrm{b/a}}(\delta^j)\tau^j\\ &= I_k^{\mathrm{b/a}}\int_\delta \log(\Lambda_k^{\mathrm{b/a}}(\delta))\,f_k^{b/a,\mathcal{T}}(\mathrm{d}\delta) - \bar{\tau}\int_\delta \Lambda_k^{\mathrm{b/a}}(\delta)\,f_k^{b/a,\mathcal{Q}}(\mathrm{d}\delta)\end{aligned}$$

where $f_k^{b/a,\mathcal{T}}(\mathrm{d}\delta)$ is the probability measure of bid/ask trades bucketed with size z_k, $f_k^{b/a,\mathcal{Q}}(\mathrm{d}\delta)$ is the probability measure of streamed quotes (weighted with durations) at the bid/ask for size z_k and $\bar{\tau}=\sum_{j\in\mathcal{J}_k^{\mathrm{b/a}}}\tau^j$ is the total duration of the time window.

Intensity functions can be interpreted in the following two-step fashion. First, there is a given flow of prospective customers who look at the prices. The probability that they trade then depends on the quotes proposed by the dealer. Therefore, inspired by logistic regression techniques, a natural functional form is:

$$\Lambda_k^{\mathrm{b/a}}(\delta) = \frac{\lambda_k^{\mathrm{b/a}}}{1+\mathrm{e}^{\alpha_k^{\mathrm{b/a}}+\beta_k^{\mathrm{b/a}}\delta}}$$

1 Trade (green) and streamed quote (red) frequency histograms and smoothed probability density functions (associated with $f_1^{\mathcal{T}}(\mathrm{d}\delta)$ and $f_1^{\mathcal{Q}}(\mathrm{d}\delta)$) for a client chosen at random in our sample and trades of €1M (right axis) along with the corresponding estimated intensity function (blue, left axis) (normalised to a maximum of 1)



where $\lambda_k^{\mathrm{b/a}}$ represents the flow of prospective customers and the term $1/(1+\mathrm{e}^{\alpha_k^{\mathrm{b/a}}+\beta_k^{\mathrm{b/a}}\delta})$ represents the probability of trading given the quotes proposed. By using a maximum likelihood approach, we can easily estimate the parameters, ie, for each k:

$$\begin{aligned}&(\lambda_k^{\mathrm{b/a}},\alpha_k^{\mathrm{b/a}},\beta_k^{\mathrm{b/a}})\\ &\quad\in \operatorname{argmax}\left(I_k^{\mathrm{b/a}}\int_\delta \log\left(\frac{\lambda_k^{\mathrm{b/a}}}{1+\mathrm{e}^{\alpha_k^{\mathrm{b/a}}+\beta_k^{\mathrm{b/a}}\delta}}\right)f_k^{b/a,\mathcal{T}}(\mathrm{d}\delta)\right.\\ &\qquad\left.-\bar{\tau}\int_\delta\frac{\lambda_k^{\mathrm{b/a}}}{1+\mathrm{e}^{\alpha_k^{\mathrm{b/a}}+\beta_k^{\mathrm{b/a}}\delta}}f_k^{b/a,\mathcal{Q}}(\mathrm{d}\delta)\right)\end{aligned}$$

While carrying out the above estimation procedure on individual clients, we noticed that intensities on the bid and ask sides were not significantly different. Therefore, we assumed $\Lambda_k^{\mathrm{b}}(\delta)=\Lambda_k^{\mathrm{a}}(\delta)=\Lambda_k(\delta)$. This assumption enabled us to achieve more precise estimations since bid and ask tables could then be concatenated and the loglikelihoods added. For the examples in this paper, we therefore fitted the functions:

$$\Lambda_k(\delta) = \frac{\lambda_k}{1+\mathrm{e}^{\alpha_k+\beta_k\delta}}$$

by choosing, for each k:

$$\begin{aligned}&(\lambda_k,\alpha_k,\beta_k)\in\operatorname{argmax}\left(I_k\int_\delta\log\left(\frac{\lambda_k}{1+\mathrm{e}^{\alpha_k+\beta_k\delta}}\right)f_k^{\mathcal{T}}(\mathrm{d}\delta)\right.\\ &\qquad\left.-2\bar{\tau}\int_\delta\frac{\lambda_k}{1+\mathrm{e}^{\alpha_k+\beta_k\delta}}f_k^{\mathcal{Q}}(\mathrm{d}\delta)\right)\end{aligned}$$

where:

$$I_k = I_k^{\mathrm{b}}+I_k^{\mathrm{a}},\quad f_k^{\mathcal{T}}=\frac{I_k^{\mathrm{b}}f_k^{\mathrm{b},\mathcal{T}}+I_k^{\mathrm{a}}f_k^{\mathrm{a},\mathcal{T}}}{I_k^{\mathrm{b}}+I_k^{\mathrm{a}}},\quad f_k^{\mathcal{Q}}=\frac{f_k^{\mathrm{b},\mathcal{Q}}+f_k^{\mathrm{a},\mathcal{Q}}}{2}$$

The results for $z_1=$ €1 million are shown in figure 1 for a single client chosen at random in our sample. We do not display the scale (ie, λ_1) as only the shape is important.

[3] *The notion of reference price can be obscure in the case of forex due to the significant geographical delocalisation of liquidity and last look practice (see Oomen 2017). The true market price can only be known with limited accuracy. Nevertheless, the primary venue provides a reliable measurable reference, suitable for the purpose of this analysis.*

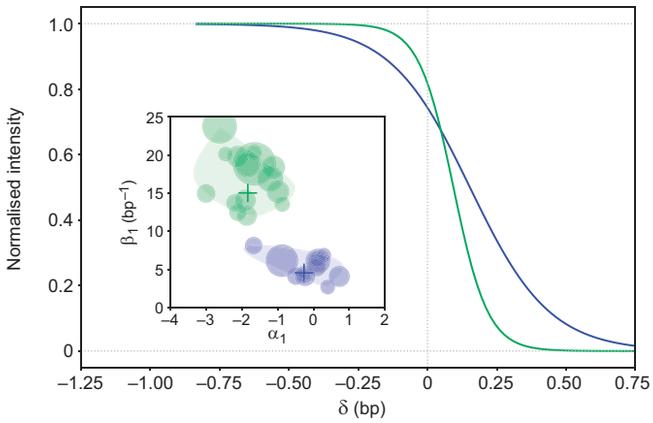
[4] *Throughout, we shall use the term quote for $\delta^{\mathrm{b/a}}$ although it is only a mark-up or a discount with respect to the reference price.*

[5] *We take the dealer's viewpoint when it comes to trade sides.*

[6] *This sample is sufficiently diverse to provide realistic results, but by no means complete enough to fully represent the HSBC forex market-making franchise.*

[7] *In what follows, we focused only on the most liquid hours.*

2 Normalised intensity functions for the two client tiers (Tier 1 is in blue and Tier 2 in green)



Tiers are identified using a standard k-means procedure on individually fitted intensity parameters for the sample of clients considered in the paper and trades of €1M (inset)

Figure 2 collects (α_1, β_1) parameters for all clients in the sample. Two clusters are clearly visible, justifying the creation of tiers. The intensity functions corresponding to the two tiers (estimated on pooled data for each tier using the same maximum likelihood approach as above) are shown as well. We can observe significantly different price sensitivities across the two tiers. The respective parameters are (after rounding) $\alpha_1^1 = -0.3$ and $\beta_1^1 = 5$ per basis point for Tier 1 and $\alpha_1^2 = -1.9$ and $\beta_1^2 = 15\text{bp}^{-1}$ for Tier 2. This correlates well with recent findings on informativeness and trading behaviour of typical forex OTC market participants, with the different pricing sensitivities of different types of clients driven by significantly different business horizons, risk management requirements and information access (Ranaldo & Somogyi 2021).

For other sizes, the shape parameters were found to be consistent with the results for trades of €1M. $(\lambda_1, \ldots, \lambda_6)$ have been rounded and found proportional to $(0.4, 0.25, 0.19, 0.1, 0.05, 0.01)$ for both tiers. We therefore define throughout the paper the parameters $\alpha^1 = -0.3$ and $\beta^1 = 5\text{bp}^{-1}$ for Tier 1 and $\alpha^2 = -1.9$ and $\beta^2 = 15\text{bp}^{-1}$ for Tier 2.

The market-making model for multiple tiers

Let us now examine the market-making model. In the general case, we denote by N the number of tiers ($N = 2$ in our examples). The market maker streams a pricing ladder for each tier: for Tier $n \in \{1, \ldots, N\}$ they propose a pricing ladder $S^{\text{b},n}(t,z) = S_t(1-\delta^{\text{b},n}(t,z))$ at the bid and $S^{\text{a},n}(t,z) = S_t(1+\delta^{\text{a},n}(t,z))$ at the ask. The associated intensities for Tier n are denoted by $\Lambda^{\text{b},n}$ and $\Lambda^{\text{a},n}$, respectively. Following the above empirical results, we assume that the functions $\Lambda^{\text{b},n}$ and $\Lambda^{\text{a},n}$ have the form:[8]

$$\Lambda^{\text{b},n}(z,\delta) = \Lambda^{\text{a},n}(z,\delta) = \Lambda^n(z,\delta) = \lambda^n(z) f^n(\delta)$$
$$\text{with } f^n(\delta) = \frac{1}{1+\mathrm{e}^{\alpha^n+\beta^n\delta}}$$

The market maker can also trade on a platform to hedge their position. The execution rate of the market maker on this platform is modelled by a process $(v_t)_t$.

We assume the dynamics of the reference price has two parts: an exogenous part with classical lognormal dynamics and an endogenous part corresponding to the permanent market impact of the market maker's trades on the platform (ie, when they externalise). Mathematically, $(S_t)_t$ has the dynamics:

$$\mathrm{d}S_t = \sigma S_t\,\mathrm{d}W_t + k v_t S_t\,\mathrm{d}t$$

where $(W_t)_t$ is a standard Brownian motion, k represents the magnitude of the (linear) permanent impact and σ is the volatility parameter.

We denote by $(q_t)_t$ the inventory process of the market maker resulting from trades with clients and trades on the platform. Mathematically, denoting by $J^{\text{b},n}(\mathrm{d}t,\mathrm{d}z)$ and $J^{\text{a},n}(\mathrm{d}t,\mathrm{d}z)$ the random measures modelling the times and sizes of trades with Tier n on the bid and ask sides, respectively, the dynamics of $(q_t)_t$ is given by

$$\mathrm{d}q_t = \sum_{n=1}^{N}\int_{z=0}^{\infty} z J^{\text{b},n}(\mathrm{d}t,\mathrm{d}z) - \sum_{n=1}^{N}\int_{z=0}^{\infty} z J^{\text{a},n}(\mathrm{d}t,\mathrm{d}z) + v_t\,\mathrm{d}t.$$

The resulting cash process $(X_t)_t$ of the market maker can be written as:

$$\mathrm{d}X_t = \sum_{n=1}^{N}\int_{z=0}^{\infty} S^{\text{a},n}(t,z) z J^{\text{a},n}(\mathrm{d}t,\mathrm{d}z) - \sum_{n=1}^{N}\int_{z=0}^{\infty} S^{\text{b},n}(t,z) z J^{\text{b},n}(\mathrm{d}t,\mathrm{d}z) - v_t S_t\,\mathrm{d}t - L(v_t)S_t\,\mathrm{d}t$$

where the term $L(v_t)S_t$ accounts for the execution costs.[9]

The market maker wants to maximise the expected mark-to-market value of their portfolio at the end of the period $[0,T]$ while managing the risk associated with their inventory. Mathematically, we assume that they want to maximise:

$$\mathbb{E}\left[X_T + q_T S_T - \frac{C}{2}\int_0^T q_t^2\,\mathrm{d}[S]_t\right]$$

by choosing $\delta^{\text{b},n}$, $\delta^{\text{a},n}$ and v, where the respective importance of the expected profit and loss (P&L) and risk management components can be chosen through the coefficient $C \geqslant 0$. This is a standard objective function discussed in the market-making literature.[10] Market share is also often targeted by dealers, and this could be part of a more general multi-objective optimisation problem, but P&L and risk will always remain at the core.

Applying Itô's formula to the process $(X_t + q_t S_t)_t$ allows us to see that this problem is equivalent to maximising:

$$\mathbb{E}\left[\int_0^T\left(\sum_{n=1}^{N}\int_0^{\infty}(z\delta^{\text{b},n}(t,z)\Lambda^n(\delta^{\text{b},n}(t,z)) + z\delta^{\text{a},n}(t,z)\Lambda^n(\delta^{\text{a},n}(t,z)))S_t\,\mathrm{d}z + k q_t v_t S_t - L(v_t)S_t - \frac{C}{2}\sigma^2 q_t^2 S_t^2\right)\mathrm{d}t\right]$$

[8] *Generalisations are of course straightforward.*

[9] *L is typically nonnegative, strictly convex and asymptotically superlinear.*

[10] *A terminal penalty can be introduced on the residual inventory at time T.*

As T is chosen small in what follows, it makes sense to approximate S_t by S_0 in the expression above, and the problem becomes that of maximising:

$$\mathbb{E}\Bigg[\int_0^T \Bigg(\sum_{n=1}^N \int_0^\infty (z\delta^{\mathrm{b},n}(t,z)\Lambda^n(\delta^{\mathrm{b},n}(t,z)) + z\delta^{\mathrm{a},n}(t,z)\Lambda^n(\delta^{\mathrm{a},n}(t,z)))\,\mathrm{d}z + kq_t v_t - L(v_t) - \frac{\gamma}{2}\sigma^2 q_t^2\Bigg)\,\mathrm{d}t\Bigg]$$

where $\gamma = CS_0$ is analogous to the risk aversion parameter in most models in the market-making literature.

We denote by $\theta\colon [0,T]\times\mathbb{R}\to\mathbb{R}$ the value function of this stochastic control problem. The Hamilton-Jacobi equation associated with it is:

$$\begin{aligned}0 = {}& \partial_t\theta(t,q) - \frac{\gamma}{2}\sigma^2 q^2 \\ &+ \sum_{n=1}^N \int_0^\infty zH^n\left(\frac{\theta(t,q)-\theta(t,q+z)}{z}\right)\lambda^n(z)\,\mathrm{d}z \\ &+ \sum_{n=1}^N \int_0^\infty zH^n\left(\frac{\theta(t,q)-\theta(t,q-z)}{z}\right)\lambda^n(z)\,\mathrm{d}z \\ &+ \mathcal{H}(\partial_q\theta(t,q)+kq) \quad \forall (t,q)\in[0,T)\times\mathbb{R}\end{aligned}$$

with terminal condition $\theta(T,\cdot)=0$, where:

$$H^n\colon \quad p\in\mathbb{R}\mapsto \sup_\delta f^n(\delta)(\delta-p)$$

$$\mathcal{H}\colon \quad p\in\mathbb{R}\mapsto \sup_v(vp-L(v))$$

Under mild assumptions (see Bergault & Guéant (2021) for similar results), it can be proved that, given a smooth solution to the above Hamilton-Jacobi equation, the optimal controls are given by:[11]

$$\delta^{\mathrm{b},n*}(t,z) = \bar{\delta}^n\left(\frac{\theta(t,q_{t-})-\theta(t,q_{t-}+z)}{z}\right)$$

$$\delta^{\mathrm{a},n*}(t,z) = \bar{\delta}^n\left(\frac{\theta(t,q_{t-})-\theta(t,q_{t-}-z)}{z}\right)$$

and:

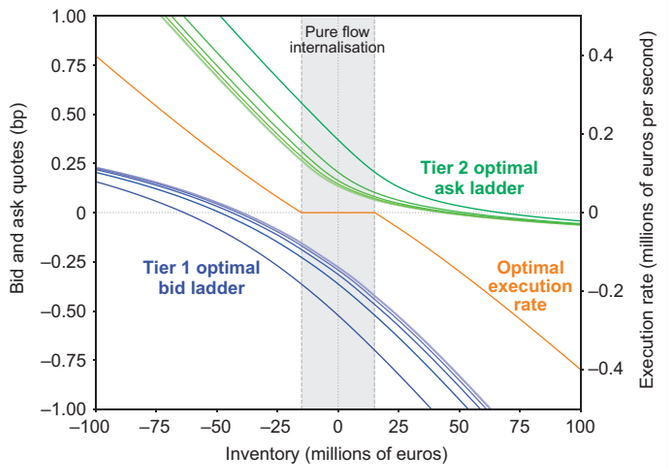
$$v_t^* = \mathcal{H}'(\partial_q\theta(t,q_{t-})+kq_{t-})$$

where $\bar{\delta}^n(p) = (f^n)^{-1}(-H^{n\prime}(p))$.

Numerical results and discussion

To illustrate the optimal market-making strategy, let us focus on the case of a typical top-tier bank dealer on EUR/USD. Regarding size buckets, client tiering and the shape of intensities, we used the same parameters as in the above study of the trading flow on a sample of HSBC clients. Regarding intensity amplitudes, we set $(\lambda_1,\dots,\lambda_6) = \lambda\cdot(0.4, 0.25, 0.19, 0.1, 0.05, 0.01)$ for both tiers, with $\lambda = 1{,}800\ \mathrm{day}^{-1}$. This figure was chosen so that, using the optimal strategy, the trading flow is of the same order of magnitude as the estimation proposed in Butz & Oomen (2019) (see also the BIS data (Schrimpf & Sushko 2019)). We obtained approximately €10 billion of daily turnover (the estimation in Butz & Oomen (2019) was US$7.32M/min).[12]

3 Optimal bid ladder for Tier 1 (blue): $-\delta^{\mathrm{b},1*}(0,z)$ as a function of q_{0-} for $z\in\{z_1,\dots,z_6\}$



Optimal ask ladder for Tier 2 (green): $\delta^{\mathrm{a},2*}(0,z)$ as a function of q_{0-} for $z\in\{z_1,\dots,z_6\}$. Optimal external hedging rate (orange): v_0^* as a function of q_{0-}. Risk aversion parameter: $\gamma = 2\cdot 10^{-3}\mathrm{bp}^{-1}\cdot(\mathrm{M}€)^{-1}$

As far as the execution cost and market impact parameters are concerned, we used standard estimation techniques on a sample of HSBC execution data and chose (after rounding):

- $L\colon v\in\mathbb{R}\mapsto \eta v^2+\phi|v|$ with $\eta = 10^{-5}\mathrm{bp}\cdot\mathrm{day}\cdot(\mathrm{M}€)^{-1}$ and $\phi = 0.1\mathrm{bp}$.
- Permanent market impact: $k = 5\cdot 10^{-3}\mathrm{bp}\cdot(\mathrm{M}€)^{-1}$.

We set the volatility to $\sigma = 50\mathrm{bp}\cdot\mathrm{day}^{-1/2}$ and considered a time horizon $T = 0.05$ days (72 minutes), which ensures convergence towards stationary quotes and hedging rates at time $t=0$ (see more on convergence in Barzykin *et al* (2021)). In order to approximate the value function θ, we added boundary conditions by imposing that no trade that would result in an inventory $|q| > €250$ million is admitted, and used a monotone implicit Euler scheme on a grid with 501 points for the inventory.

Figure 3 summarises optimal pricing and hedging strategies for the above set of parameters and risk aversion of $\gamma = 2\cdot 10^{-3}\mathrm{bp}^{-1}\cdot(\mathrm{M}€)^{-1}$.[13] There are several interesting features worth noting. First, we observe a range of inventory around zero, where the dealer will only internalise by skewing the quotes, ie, no hedging. We call this interval the pure flow internalisation area. Since $L(v) = \eta v^2+\phi|v|$ implies that:

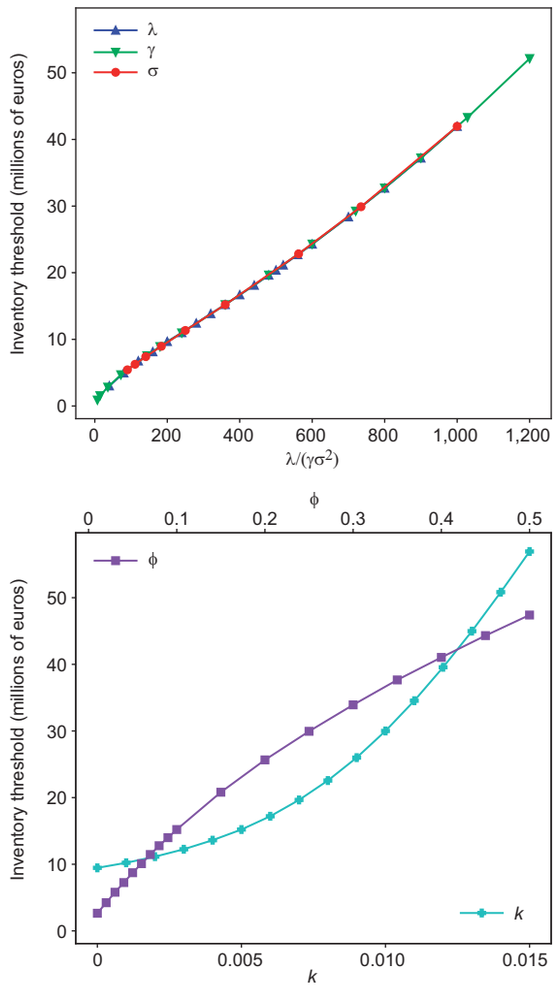
$$\mathcal{H}'(p) = \frac{1}{2\eta}\,\mathrm{sgn}(p)\max(0,|p|-\phi)$$

[11] *$(q_s)_s$ is a càdlàg (right continuous with left limits) process and we write the left limit of the process q at time t as $q_{t-} = \lim_{s\uparrow t} q_s$.*

[12] *Note that the maximum daily turnover corresponding to these parameters is approximately €31 billion. However, the dealer can only hypothetically reach this level by quoting far better prices than the mid-price and losing money.*

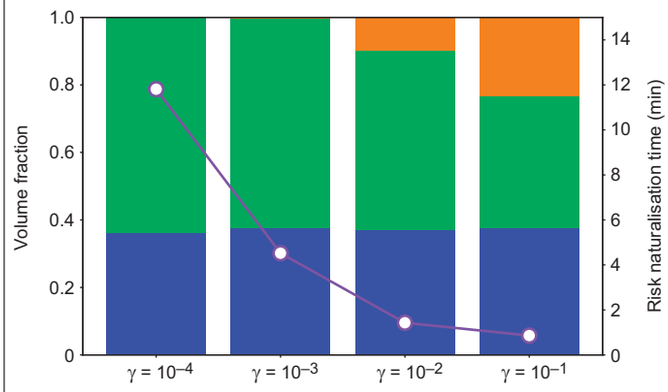
[13] *Due to the assumption of flow symmetry, it suffices to plot only bid or ask quotes for each tier, as the other side would be a mirror image. We decided to plot $-\delta^{\mathrm{b},1*}(0,z)$ as a function of q_{0-} for $z\in\{z_1,\dots,z_6\}$ for Tier 1 and $\delta^{\mathrm{a},2*}(0,z)$ as a function of q_{0-} for $z\in\{z_1,\dots,z_6\}$ for Tier 2.*

4 Inventory threshold of the pure flow internalisation area for different levels of risk aversion (γ), volatility (σ), franchise size (λ), execution costs (ϕ) and permanent market impact (k)



The top panel scales against $\lambda/(\gamma\sigma^2)$ with the varying parameter colour coded, with others being fixed at default values

5 Traded volume fraction executed with Tier 1 and Tier 2 clients and externally for hedging purpose (bars: Tier 1 is in blue, Tier 2 is in green, external trading is in orange)



Risk-neutralisation time (line and dots) for different values of γ. Results were obtained by Monte Carlo simulation of 10^5 trajectories over a time horizon $T = 10$ days for a market maker following the stationary optimal strategy

we have $\mathcal{H}'(p) = 0 \iff |p| \leqslant \phi$. Given the expression for the optimal controls, the pure flow internalisation area corresponds to the set of inventories q verifying:

$$|\partial_q \theta(0, q) + kq| \leqslant \phi$$

which contains 0 plus an interval around 0 (as soon as $\theta(0, \cdot)$ is continuously differentiable). In terms of sensitivity to the parameters, we noticed empirically (in line with intuition) a wider pure flow internalisation area for a less risk-averse market maker with a larger franchise, exposed to higher execution costs and market impact and in a less volatile market (see figure 4). We also note that the optimal execution rate curve is almost linear with respect to inventory outside of the pure flow internalisation area. Second, the bid-ask spread is driven by the flow signature, leading to different pricing strategies for the two tiers we considered. Our estimation of the inventory-neutral top-of-book bid-ask spread (ie, the difference between the ask and bid prices for a notional of €1M) for price-sensitive clients is 0.26bp, while for less sensitive clients it is 0.55bp. This compares well with an average composite[14] bid-ask spread of 0.23bp and average primary venue bid-ask spread of 0.65bp at New York open at the time of writing in early July 2021. This is particularly interesting, as no market bid-ask spread was introduced into the model. We note that the market maker's OTC spread is mainly driven by the empirical shape of the intensity function.

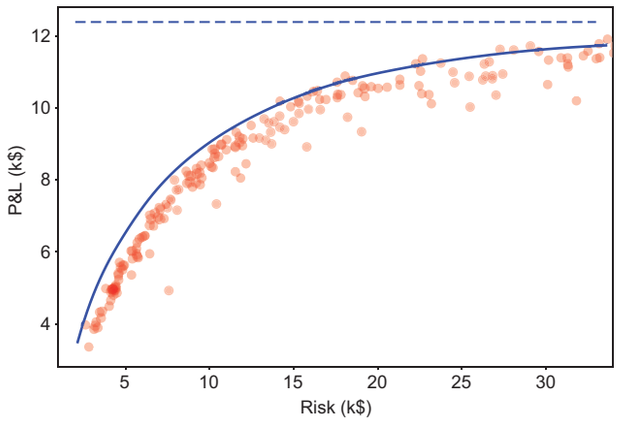
Once the optimal strategy has been computed, we can simulate the behaviour of our market maker and assess the volume share of external hedging for different levels of dealer's risk aversion. Figure 5 shows a span of four orders of magnitude in γ, illustrating the crossover from pure internalisation to significant externalisation. Note that the volume share of less price-sensitive clients (Tier 1) remains basically the same while the dealer will prefer to sacrifice price-sensitive flow (Tier 2) for the certainty of inventory management when risk aversion increases. The level of internalisation for a risk-aware dealer is in line with BIS reporting around 80% internalisation in G10 currencies by top-tier banks.

Figure 5 also illustrates the dependence of the characteristic risk-neutralisation time τ_R on γ, where τ_R is defined as the integral of the inventory autocorrelation function. It appears that pure internalisation clearly comes with a significantly higher risk. It is noteworthy that the value of τ_R for $\gamma = 0.01$ compares very well with the EUR/USD internalisation time (1.39 minutes) estimated in Butz & Oomen (2019).

Figure 6 explores the optimal risk-reward trade-off. In order to obtain the dealer's efficient frontier by analogy with Markowitz modern portfolio theory, we chose different values of the risk aversion parameter γ and perturbed the optimal strategy by randomly shifting bid and ask quotes for both tiers and randomly choosing the width of the pure internalisation area and the slope of the hedging rate curve around their optimal values.

[14] *An aggregated order book of multiple Electronic Communication Networks was used.*

6 Expected P&L versus standard deviation of the P&L over time horizon $T = 0.05$ days of a market maker following the stationary optimal strategy with different values of the risk aversion parameter (solid line) and 20 randomly perturbed strategies for each value of γ (circles)



Maximum expected P&L without risk management (dashed line). Results were obtained by Monte Carlo simulation of 10^5 trajectories for several values of γ ranging from 10^{-4} to 10^{-1}. The curve has been obtained with cubic splines

The resulting outcomes are almost all below the curve built using the optimal stationary pricing and hedging strategy although our objective function is not exactly a mean-variance one. Our penalty for inventory risk indeed ignores part of the variance (see the discussion on objective functions for market making in Guéant (2016)), and random perturbations could occasionally end up being above the curve,[15] but our approach appears to be a very good one from a risk-reward perspective.

It must be noted that there is a significant difference between the efficient frontier of Markowitz modern portfolio theory and ours, in that the expected P&L is bounded from above in our case. Figure 6 shows the maximum expected P&L with no risk management. The simulated optimal curve saturates at a lower value when $\gamma \to 0$ because of the inventory limit we imposed to build a grid-based finite difference scheme.

The choice of γ ultimately rests with the dealer and it is clear that the optimal risk-reward curve can be useful in making the decision if we want to optimise a risk-adjusted financial performance measure such as Sharpe ratio. Forex dealers often have other objectives than those purely based on risk-adjusted financial performance. For instance, they often care about market share. Although our model does not include such a criterion, simulations similar to those carried out above could help in choosing strategies that provide good results even when additional criteria are taken into account.

Concluding remarks

We introduced and analysed numerically a model of optimal market making incorporating fundamental risk controls: pricing ladders over a distribution of sizes and client tiers as well as the rate of hedging in external markets. The model has immediate practical application to foreign exchange where the marketplace is significantly fragmented and dealers must continuously solve the dilemma of whether to internalise or externalise their flow. We described the relevant features of client flow, taking a typical EUR/USD franchise as an example and showed how tiers and pricing ladders as a function of size naturally arise from this analysis. The results obtained regarding bid-ask spreads, risk-neutralisation times and internalisation ratios are consistent with empirical data and publicly reported figures. ■

Alexander Barzykin is a director at HSBC GFX and commodities, specialising in algorithmic execution and market making. He is based in London. Philippe Bergault is a postdoctoral researcher in applied mathematics at École Polytechnique. Olivier Guéant is full professor of applied mathematics at Université Paris 1 Panthéon-Sorbonne and adjunct professor of quantitative finance at ENSAE - IP Paris. The results presented in this paper are part of the research carried out within the HSBC FX Research Initiative. The views expressed are those of the authors and do not necessarily reflect the views or the practices at HSBC. The authors are grateful to Richard Anthony (HSBC) and Paris Pennesi (HSBC) for helpful discussions and support throughout the project.
Email: alexander.barzykin@hsbc.com,
bergault.philippe@protonmail.com,
olivier.gueant@univ-paris1.fr.

[15] *This may also be linked to finite sample statistics and to the use of the optimal stationary strategy rather than the time-dependent one close to time T.*

REFERENCES

Almgren R and N Chriss, 2001
Optimal execution of portfolio transactions
Journal of Risk 3, pages 5–40

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

Barzykin A, P Bergault and O Guéant, 2021
Algorithmic market making in foreign exchange cash markets with hedging and market impact
Preprint, arXiv:2106.06974

Bergault P and O Guéant, 2021
Size matters for OTC market makers: general results and dimensionality reduction techniques
Mathematical Finance 31(1), pages 279–322

Butz M and R Oomen, 2019
Internalisation by electronic FX spot dealers
Quantitative Finance 19(1), pages 35–56

Cartea Á, S Jaimungal and J Penalva, 2015
Algorithmic and High-Frequency Trading
Cambridge University Press

Guéant O, 2016
The Financial Mathematics of Market Liquidity: From Optimal Execution to Market Making (volume 33)
CRC Press

Ho T and HR Stoll, 1981
Optimal dealer pricing under transactions and return uncertainty
Journal of Financial economics 9(1), pages 47–73

Oomen R, 2017
Last look
Quantitative Finance 17(7), pages 1,057–1,070

Ranaldo A and F Somogyi, 2021
Asymmetric information risk in FX markets
Journal of Financial Economics 140(2), pages 391–411

Schrimpf A and V Sushko, 2019
FX trade execution: complex and highly fragmented
BIS Quarterly Review, December