Agents hétérogènes et formation des prix sur les marchés financiers

Jonathan Donier

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

SPÉCIALITÉ : MATHÉMATIQUES APPLIQUÉES
ECOLE DOCTORALE : « MATHÉMATIQUES PARIS CENTRE »

présentée par

Jonathan DONIER

Agents hétérogènes et formation des prix sur les marchés financiers

---

sous la direction de

Mathieu Rosenbaum, ***Laboratoire de Probabilités et Modèles Aléatoires***
Jean-Philippe Bouchaud, ***Capital Fund Management***

soutenue le 10 Octobre 2016 devant le jury composé de :

Mathieu Rosenbaum	***Directeur de thèse***
Jean-Philippe Bouchaud	***Directeur de thèse***
Emmanuel Bacry	***Rapporteur***
Fabrizio Lillo	***Rapporteur***
Jim Gatheral	***Examinateur***
Rama Cont	***Examinateur***
Thierry Foucault	***Examinateur***

Préface

Peu de sujets traversent autant les communautés que la finance. Peu d'entre eux soulèvent autant de questions scientifiques, philosophiques et comportementales. Si les premiers à avoir épaulé les praticiens furent des mathématiciens (Bachelier, Mandelbrot...) et des économistes (Black, Scholes, Fama...), les dernières décennies ont vu l'arrivée massive de physiciens dans l'arène (Stanley, Bouchaud...). Cela ne devrait guère nous étonner. L'apparition de données en masse depuis plus de 20 ans a fait de la finance un des plus grands champs d'expériences scientifiques et comportementales, à côté duquel la communauté physique ne pouvait pas passer. Cependant, si la richesse et la variété des approches est en théorie un atout, force est de constater que les différentes écoles de pensée sont encore relativement hermétiques entre elles et ne bénéficient que partiellement de cette coexistence. Alors que les économistes sont omniprésents dans la littérature académique et les instances régulatrices, les mathématiciens et physiciens ont quant à eux inondé le marché de la finance quantitative dans les banques et les fonds d'investissement. Est-il pour autant nécessaire que cette disparité des objectifs résulte en une disparité des méthodes, des modèles voire même des croyances ? Je ne le crois pas, et c'est donc sans *a priori* aucun que se présente mon étude, dans l'espoir que réunir des points de vue différents enrichira la compréhension de ce sujet complexe. Si ma scolarité à l'Ecole Polytechnique m'a permis d'aborder la finance avec la perspective de l'économie et des mathématiques financières, c'est entouré de physiciens – et de praticiens – que j'ai réalisé l'essentiel de ces travaux : j'espère ici rendre justice à chacun.

Si chacune des communautés regorge de richesses indéniables, il arrive cependant que celles-ci soient si fortement attachées à certains de leurs principes que toute ouverture vers l'extérieur semble parfois difficile, voire impossible. Le parti-pris qui sous-tend ces travaux, au contraire, est de ne pas se laisser entraver par de telles contraintes, et je n'hésiterai pas, le cas échéant, à confronter certains principes et hypothèses majoritairement acceptés, en lesquels la confiance générale me parait exagérée. Il me semble que dans des sciences aussi inexactes que les sciences sociales, les hypothèses les plus courantes, si attrayantes soient-elles, doivent toujours être remises en cause, et ne pas devenir des préceptes qui guident aveuglément nos tentatives de modélisation.

Enfin, il me semblait difficile de pouvoir bien comprendre ce travail s'il n'était pas replacé dans son contexte général, et positionné par rapport aux grandes questions qui l'ont motivé. Cela m'oblige à commencer par des discussions relativement générales sur la modélisation économique et financière, que de nombreux commentateurs ont déjà engagées avant moi. Si le risque est grand de tomber dans le *déjà-vu*, je crois cependant que le coeur de cette thèse, avec son contenu empirique et théorique, leur procure à l'inverse des éclairages nouveaux. Cet aller-retour entre modélisation et méta-réflexion, s'il fut essentiel dans l'entreprise de ces travaux, me semble tout autant crucial pour leur compréhension.

Remerciements

Ce travail n'aurait jamais vu le jour sans l'aide et le concours de nombreuses personnes, que je souhaiterais maintenant remercier. Le premier de mes remerciements va sans la moindre des hésitations à mon directeur de thèse Jean-Philippe Bouchaud, qui m'a tant appris et tant donné durant ces trois années. Merci pour ta disponibilité, ton investissement et ton encadrement exemplaire qui ont dépassé toutes mes attentes, si élevées furent-elles.

Un grand merci également à Mathieu Rosenbaum, mon directeur de thèse académique à Paris 6, pour m'avoir inspiré et orienté depuis mes débuts à l'Ecole Polytechnique jusqu'à aujourd'hui. Merci pour ton temps, pour tes avis éclairés et pour ta perpétuelle bonne humeur.

Emmanuel Bacry et Fabrizio Lillo ont accepté la tâche fastidieuse de relire et de commenter ce manuscrit. J'en suis extrêmement honoré et je souhaite leur exprimer toute ma gratitude pour le temps qu'ils ont accepté d'y consacrer et pour la qualité de leurs remarques. Rama Cont et Thierry Foucault, sans probablement le savoir, ont fait partie de ces personnes dont les travaux et opinions ont grandement contribué à structurer et à diversifier ma pensée, et sans qui cette thèse ne serait pas ce qu'elle est. C'est donc un grand honneur de pouvoir les compter parmi mon jury. Enfin, je souhaiterais addresser un grand merci à Jim Gatheral, qui fut mon premier mentor de stage et sans qui je n'aurais sans doute pas suivi cette voie. Jim, merci pour ton enthousiasme et ta confiance qui ont été déterminants dans ma décision de faire une thèse, et pour ces quatre mois exceptionnels passée à Baruch College. Je suis extrêmement touché de pouvoir te compter aujourd'hui parmi les membres de mon jury.

Je remercie tout particulièrement mes collaborateurs, Iacopo, Julius, Martin, Michael, Pierre et bien sûr Jean-Philippe, pour l'ensemble des travaux réalisés et dont je suis très fier. Travailler avec chacun de vous fut un réel plaisir.

Ces trois années passées au sein de Capital Fund Management (CFM) ont été aussi agréables que riches en enseignements et déterminantes pour cette thèse. J'aimerais remercier l'ensemble des chercheurs pour nos interactions, et notamment Adam, Bence, Charles-Albert, Emmanuel, Joachim, Julianus, Loïc, Marc, Peter, Raphaël, Stephen, Yves – et tous ceux que j'oublie. Un merci tout particulier à Nicolas K. pour nos discussions tour à tour concrètes et philosophiques sur la formation des prix, les automates de trading et les monnaies virtuelles.

Je remercie la fondation CFM pour avoir contribué au financement de cette thèse, et à Nathalie et Erin pour leur aide et leur sourire. Merci également au Corps des Mines pour son soutien, et en particulier à Aurélie et Pierre-Edouard pour leur disponibilité.

Je n'oublierai pas toute la team « stagiaires » CFM, en particulier Marc - mon fidèle partenaire de baby-foot, Joël, Pierre, Yanis - notre coach de crossfit improvisé, Nadir, Alexia, Alexis et Benoit avec lesquels j'ai eu la chance de partager successivement les salles des $5^{\text{ème}}$ et $6^{\text{ème}}$ étages, et qui ont été une source de bonne humeur intarissable durant ces trois années.

Merci à Pierre *aka* « le golgoth » pour avoir été un point de mire tant sur la piste que dans le vie. Un immense merci à Antoine, source d'inspiration perpétuelle, qui m'a initié à la finance, au Bitcoin et à peu près à tout ce qui m'intéresse aujourd'hui, et envers qui je ne formulerai jamais trop d'éloges.

Merci enfin à ma famille pour m'avoir accompagné jusqu'ici – et à Alix, pour tout le reste.

Table des matières

Première partie

Objectif de l'ouvrage et concepts élémentaires

Chapitre 1

Introduction

1.1 Des marchés...

J'ai toujours trouvé dans les marchés quelque chose de fascinant. Rien de plus banal cependant que le geste d'échange : qu'un acheteur s'accorde avec un vendeur, et le tour est joué. Mais le miracle des marchés n'est pas là, ou plutôt, va bien au-delà. Un marché ne rassemble pas seulement *un* acheteur et *un* vendeur. Il fédère *des* acheteurs et *des* vendeurs, une *multitude* d'acheteurs et une *multitude* de vendeurs, autour d'un prix qui semble être « le bon » [1] – mais que personne ne semble fixer, pourtant.

« Le bon », certes, mais en apparence si peu contrôlé, si fragile : si personne ne fixe le prix, qu'est-ce qui nous protège de ses sautes d'humeur ? La théorie économique de nous rassurer sans attendre en répondant : *l'efficience* [2]. C'est-à-dire, les prix valent ce qu'ils *doivent* valoir, étant donné tout ce que tout le monde sait – et s'approchent d'autant plus de la valeur fondamentale du bien, que ce savoir commun s'étend. En notant le savoir commun $\mathcal{F}$ [3] et la valeur intrinsèque du bien V, cela se traduit en termes mathématiques :

$$\text{prix} = \mathbb{E}\left[V \mid \mathcal{F}\right]. \tag{1.1}$$

Même le lecteur sans grandes notions de probabilités comprendra que, de cette identité, résulte – un peu magiquement – une unique valeur du prix. Celui-ci ne peut donc changer, que si la valeur intrinsèque du bien change, ou si la connaissance qu'en ont les protagonistes de marché s'étend [4].

1. Cette capacité du marché à trouver de lui-même le « bon » prix sans y être forcé par quelque interventionnisme est probablement ce qui me fascine le plus dans les marchés financiers.

2. Voir Malkiel and Fama (1970) pour plus de détails sur le concept d'efficience. La notion d'absence d'*arbitrage*, qui lui est intimement reliée, est extrêmement pratique pour les mathématiciens. Ces deux notions peuvent être vues comme les conséquences d'une hypothèse comportementale plus fondamentale : *la concurrence (parfaite)*.

3. Voir encore Malkiel and Fama (1970) pour plus de détails sur ce que peut être ce « savoir commun ».

4. Si $\mathcal{F}$ représente la connaissance de Dieu dans l'équation 1.1, alors elle produit : prix $= V$.

Le premier mécanisme est inévitable quel que soit le système d'échanges, et donc acceptable. Quant au second, il tend à rapprocher le prix de marché de la vraie valeur du bien, ce qui l'un dans l'autre semble nécessaire. De toute manière, il serait difficile de faire mieux, chacun étant fondamentalement limité par son savoir – mais l'exploitant naturellement au mieux [5]. Grâce à l'efficience, la question *comportementale* a été réduite à une question *mathématique* [6]. L'idée est tellement belle en effet – et tellement pratique du point de vue technique – qu'il est difficile de ne pas l'ériger en grand principe de base des mathématiques et de l'économie financière. Ayant trouvé le concept central d'un monde *idéal* – comprendre : *tel qu'il devrait être* –, s'en séparer, ou même seulement feindre de l'ignorer, semblerait de fait peu raisonnable. Plutôt, pourquoi ne pas simplement étudier ce monde idéal, y tirer des conclusions pour les appliquer ensuite au monde réel, justifié par quelque théorème de convergence de l'humanité vers son état idéal [7] ? Cette approche présente en effet deux avantages, non des moindres : *primo*, cela permet de faire des mathématiques, et de l'économie [8]. *Deuxio*, cela permet de ne jamais se soumettre à la *réalité*, et donc de ne jamais avoir tort – tant que les mathématiques sont correctes. Seule la réalité peut être blâmée, si elle ne correspond pas aux prédictions – et alors on attendra que le monde converge un peu plus.

Que l'on ne se méprenne pas : le concept d'efficience mérite bien une place de choix en modélisation financière [9]. La seule critique que l'on pourrait formuler porte sur le caractère transcendant qu'on tend à lui conférer – par idéologie, ou par commodité. Cependant, avant de porter plus loin cette critique, hâtons-nous d'abord de comprendre plus en détail ce qu'est l'efficience, et pourquoi elle devrait prévaloir au sein de la structure de marché organisé.

---

5. La valeur de p donnée par la définition 1.1 peut être vue comme la meilleure estimation du prix sachant ce qui est connu de la valeur du bien. Connu par qui, me direz-vous ? Cette question est si compliquée que pour y remédier on suppose habituellement que tout le monde – ou presque – possède le même savoir. Cela n'est pas très satisfaisant : nous cherchions en effet à comprendre comment le marché pouvait produire un prix unique pour une multitude d'acheteurs et de vendeurs, et si l'on décrète en amont que tous ont des avis identiques *a priori*, il n'y a plus de mystère – mais nous n'avons répondu à rien. Il semblerait donc qu'il faille écrire pour chaque agent i :

$$\text{prix}_i = \mathbb{E}_{\mathbb{P}_i}\left[V \mid \mathcal{F}_i\right], \tag{1.2}$$

où $\mathcal{F}_i$ est l'information qu'il possède et $\mathbb{P}_i$ son modèle, ou ses croyances. Pour en déduire un prix unique, il faut alors comprendre en profondeur ce qu'est un marché – et dépasser la tautologie de l'équation 1.1.

6. Et la question d'accord mutuel sur la valeur du prix, à une question de filtrations (il s'agit là du terme mathématique qui désigne ce qu'on a appelé $\mathcal{F}$) et de mesure de probabilité (sous laquelle on calcule l'espérance $\mathbb{E}$). Il n'en faut pas beaucoup pour se rendre compte que rien n'est résolu. Rien ou presque : car les filtrations possèdent quelques propriétés mathématiques, qui se traduiront sur le prix. A moins que les mesures de probabilités des agents (i.e. leurs modèles, ou leurs croyances) elles-mêmes n'évoluent, ce que l'efficience a tendance à exclure, mais que le bon sens a tendance à soutenir. Nous sommes ici à la frontière entre économie rationnelle et économie comportementale.

7. Cela n'est pas sans rappeler l'idée Hégélienne (puis Marxiste) de « fin de l'histoire » – que Camus commente dans *L'homme révolté* : « Marx a cru que les fins de l'histoire, au moins, se révèleraient morales et rationnelles. C'est là son utopie. » (Camus, 1951)

8. Cela permet en effet d'utiliser des outils mathématiques et économiques existant comme les martingales, les probabilités, la théorie des jeux etc. Cela permet également de satisfaire son désir de *résoudre le jeu*, car c'est comme tels que les problèmes sont formulés en économie – désir qui remonte à l'enfance, où le but était en général de *gagner* le jeu.

9. Essentiellement comme nécessité interne aux modèles quant il s'agit de modèles mathématiques, du moins en ce qui concerne le *non-arbitrage*.

1.2 De l'efficience et de l'utilité

Il serait malvenu d'ignorer l'aspect humain de la finance (et de l'économie en général). Je ne parle pas ici du caractère *imparfait* que l'on lui associe souvent, et sur lequel je m'étendrai plus loin. Je veux parler au contraire de son *intelligence*. Contrairement aux particules d'un système physique, l'être humain n'est pas *totalement* régi par des lois fondamentales qui lui préexistent [10]. Il a au contraire toute une latitude pour agir, le plus souvent – et de préférence – dans son propre intérêt, dans l'espace de liberté dont il dispose [11]. Pour ce faire, il commence par former des *anticipations* sur le futur (en anglais : *expectations*), prenant en compte le cas échéant l'effet de ses propres actions éventuelles [12], en fonction desquelles il va agir ensuite.

Pour définir un modèle il faut donc stipuler : (i) comment l'agent forme ses anticipations et (ii) ses critères de choix entre telle ou telle action, sachant ses anticipations. D'éventuelles dynamiques exogènes doivent parfois être ajoutées pour clore le modèle : dans le cas d'un match de tennis par exemple, les lois physiques qui transforment le mouvement de la raquette en un mouvement de la balle, etc. Nous ne nous étendrons pas sur ce dernier point, qui est en général soit solidement justifié – dans le cas de lois physiques par exemple – soit décidé arbitrairement lorsqu'on ne sait pas, ou ne souhaite pas, faire autrement [13]. C'est dans la manière d'adresser les points (i) et (ii) que la sensibilité propre au modélisateur entre en compte. L'économie classique y répond par : (i) les anticipations rationnelles et (ii) la fonction d'utilité de l'agent, notée $\mathcal{U}(.)$ (ou plus précisément, son optimisation). Les deux sont, bien entendu, des *hypothèses*. La première stipule que les agents effectuent la meilleure anticipation possible du futur avec l'information dont ils disposent (souvent, une information publique commune, voir la note de bas de page 5). En particulier, ils n'effectuent pas d'erreur systématique dans leurs anticipations et ont en moyenne raison (pour plus de détails, voir par exemple Hommes (2006)). La seconde suppose que chaque individu est ensuite rationnel dans chacun de ses choix (l'*homo economicus* !), et vise à maximiser son bien-être global espéré, mesuré par $\mathcal{U}$. [14]

Arrêtons-nous un instant sur ces deux hypothèses. Il est évident qu'elles proviennent essentiellement de la nécessité de *faire des calculs* : la rationalité en permettant de changer le problème d'anticipations en un problème de point fixe – car anticipations et réalisations doivent coïncider

10. L'être humain étant lui-même régi par les lois de la nature, on pourrait en fait dire que si. La compatibilité entre déterminisme et libre arbitre est une question philosophique que nous n'aborderons pas ici : nous nous contenterons de considérer que déduire les comportements humains des lois fondamentales de la nature est impossible, et nous intéresserons uniquement à des lois « effectives ».

11. Les instances de régulation ayant pour but, justement, de définir le bon espace de liberté de manière à ce que le système dans son ensemble se comporte « bien ».

12. Une discipline des mathématiques est associée à ce type de processus de décision : le contrôle optimal.

13. Par exemple car on souhaite un modèle *effectif* pour faire des calculs ou se procurer des intuitions. Les modèles à propagateur en sont un exemple en finance, voir e.g. Almgren and Chriss (2001).

14. En particulier, $\mathcal{U}$ est souvent donné à l'avance pour un individu et ne varie pas dans le temps, et la tractabilité mathématique – toujours elle – tend à lui imposer des formes simples et peu d'arguments, souvent seulement le profit.

si les anticipations sont rationnelles [15] – et l'utilité en réduisant un ensemble d'événements à un simple scalaire pour un agent donné, que celui-ci peut donc tenter de maximiser. Supposant fondées les utilités choisies, on comprend bien que l'hypothèse d'anticipations rationnelles est à peu près la seule – ou la moins arbitraire – dont les chances de produire un unique équilibre sont conséquentes : s'il doit y avoir un équilibre, qu'au moins ce soit celui-là ! On voit bien cependant qu'on a opéré une réduction drastique du champ des possibles – trop drastique, peut-être.

1.3 Les critiques de l'efficience

Les critiques de ce mode de pensée – et de l'intolérance qu'ont développé ses adeptes vis-à-vis de toute tentative un peu hétérodoxe – sont à ce jour nombreuses, soutenues par diverses études empiriques [16]. L'hypothèse que les agents ont non seulement une connaissance parfaite des lois fondamentales de l'économie et des conditions d'équilibre, mais soient ensuite en mesure de calculer précisément leur décision optimale, semble en effet forte. Dès 1900, Poincaré écrivait dans une lettre à l'économiste Léon Walras :

> Vous regardez les hommes comme infiniment égoïstes et infiniment clairvoyants. La première hypothèse peut être admise dans une première approximation, mais la deuxième nécessiterait peut-être quelques réserves.

Dans les années 1950, l'économiste Herbert Simon propose une alternative aux anticipations rationnelles, en introduisant la *rationalité limitée*. Les agents ne cherchent plus l'issue *optimale*, mais une issue *satisfaisante*, étant incapables de faire mieux, soit parce qu'ils n'ont qu'une connaissance imparfaite du système – qui est un système *humain*, rappelons-le – soit parce que sachant précisément ses lois ils ne sont pas en mesure de calculer à tout moment leur décision optimale [17]. Pour cela, au lieu de calculer l'équilibre de l'ensemble du système économique, chaque agent utilise une règle empirique (en anglais, *rule of thumb*) qui lui semble décrire le système de manière effective, et qu'il déduit de ses observations passées du système. Un telle règle peut être simple – en finance : trend-following, mean-reversion etc. – ou beaucoup plus complexe – lorsque l'agent *apprend* sa règle de décision en temps réel, e.g. par *renforcement* (voir par exemple Sargent (1993); Evans and Honkapohja (2001) pour des discussions sur la rationalité limitée et l'apprentissage). Notons que ces pratiques correspondent à celles pratiquées dans l'industrie financière quantitative, par exemple par les *hedge funds*, pour décider de leurs investissements.

15. Ce qui règle le problème évoqué en note de bas de page 6 sur les mesures de probabilité des agents : on leur impose d'être cohérentes avec la mesure de probabilité « physique » – i.e. celle qui régit réellement le monde.

16. Ainsi un papier récent (Gennaioli et al., 2015) montre que les anticipations des *Chief Financial Officers* sur les revenus de leurs propres entreprises sont systématiquement biaisés.

17. Cela est déjà suffisamment difficile dans des modèles simples, et la réalité est autrement plus complexe – car elle est en principe non linéaire.

Toutefois, rien n'empêche *a priori* un système à rationalité limitée de converger, lorsque le système a évolué suffisamment longtemps, vers le point fixe prédit par l'hypothèse d'anticipations rationnelles[18]. En finance, si tous les signaux de prix prévisibles sont arbitrés, les prix deviennent imprévisibles[19], peu importe la manière dont les agents s'y sont pris pour aboutir à cette convergence. Si tous ne sont pas arbitrés, on s'attend à ce qu'un agent – existant ou extérieur – détecte tôt ou tard l'opportunité de profits, et s'y engouffre : c'est ce qui donne au prix efficient une sorte de *transcendence*[20]. Cependant, tant pour les systèmes sociaux que pour les systèmes financiers, de nombreux exemples ont été trouvés où cette convergence n'a pas lieu, et où le système atteint un état d'équilibre alternatif – on pensera entre autres au modèle de ségrégation de Schelling (1971), et dans un contexte plus financier à Kirman (1993).

1.4 Les critiques de l'utilité

La critique principale que l'on peut formuler à l'encontre de la fonction d'utilité, est qu'elle n'est souvent qu'un moyen de résoudre *arbitrairement* le problème en amont : ne voulant pas introduire d'arbitraire dans le calcul lui-même – en particulier, dans les comportements des agents – on le déplace *avant* le calcul, et de cette manière, les résultats apparaissent plus fondés. En reconnaissant aux manipulations mathématiques intermédiaires leur caractère de *tautologies triviales* (Vapnik, 2006), on se rend rapidement compte que

$$\text{utilité} \leftrightarrow \text{résultat}$$

et donc

$$\text{stipuler une utilité} \leftrightarrow \text{stipuler un résultat.}$$

Je n'écris pas ici une équivalence stricte, et le symbole $\leftrightarrow$ dans cette dernière identité doit être compris plus littérairement que mathématiquement par « revient à ». En effet, une même classe de résultats pouvant rester valable pour toute une classe d'utilités – voire pour l'ensemble des utilités raisonnables – le raisonnement par utilités *peut* garder tout son sens. Dans le cas contraire cependant il ne s'agit que d'un déplacement du problème, en donnant l'illusion d'un problème résolu scientifiquement[21]. Pour citer Poincaré encore une fois :

18. On m'a un jour donné l'image suivante : les feuilles d'un arbre s'organisent de manière à maximiser la quantité de soleil reçue par l'arbre. On pourrait de même imaginer que les optimisations locales des agents d'un système économique fassent tendre le système global vers l'efficience. Cependant, l'arbre ne fait que *tendre* vers l'optimum mais ne l'atteint pas, ce qui pourrait également être le cas en finance (voir Zhang (1999), ou les conclusions du Chapitre 10).

19. Pour les initiés : des *martingales*.

20. *« Le marché à son insu, obéit à une loi qui le domine : la loi de la probabilité. »* (Bachelier, 1900)

21. D'autant plus que de plus l'utilité est souvent choisie par commodité mathématique, ce qui la rend relativement vide d'intérêt.

> Dans vos prémisses vont donc figurer un certain nombre de fonctions arbitraires ; mais une fois ces prémisses posées, vous avez le droit d'en tirer des conséquences par le calcul ; si, dans ces conséquences, les fonctions arbitraires figurent encore, ces conséquences ne seront pas fausses, mais elles seront dénuées de tout intérêt parce qu'elles seront subordonnées aux conventions arbitraires faites au début. Vous devez donc vous efforcer d'éliminer ces fonctions arbitraires [...].

1.5 Les critiques de la valeur intrinsèque

Quel sens donner à la valeur intrinsèque V ? L'existence d'une valeur « vraie » (ou « fondamentale ») des actifs est un éternel débat. Si Marx considère que la valeur d'un bien est déterminée par la quantité de travail nécessaire à sa production, elle se définit pour l'école néoclassique par le désir que les agents en éprouvent – ce qui est subjectif et dépend de l'individu, et ne permet donc pas de répondre à notre question initiale. Dans le cadre d'un marché, les biens peuvent de plus être revendus dans le futur au prix... du marché (!), ce qui rend la question de leur valeur plus subtile encore. Dans ce contexte, la pertinence du concept de valeur intrinsèque est pour moi intimement reliée à la notion d'horizon temporel : pour un produit produisant un *payoff* à une échéance courte (par exemple : jouer au loto du lendemain), la valeur de V est relativement objective [22]. En revanche, qu'en est-il des actions, pour lesquelles aucune échéance n'est définie ? V correspond-elle à la valeur attendue à « $t = +\infty$ » [23] ? Cette remarque soulève une dernière question, plus comportementale. Sachant qu'aucun agent ne pourra bénéficier de l'actif à $t = +\infty$, la valeur de *marché* devient au moins aussi importante que la valeur intrinsèque : ce qui compte alors, c'est surtout de pouvoir revendre l'actif à quelqu'un dans un futur atteignable. Mettre V dans l'équation 1.1 (ou même l'équation 1.2) lorsque l'on est un participant de marché, c'est déjà croire en la volonté des autres de faire de même dans le futur – et en leur capacité à effectivement produire par ce processus un prix raisonnable. Dans le cas contraire, ce n'est plus V que chaque participant doit anticiper, mais seulement le prix de marché lui-même – et l'on tourne en rond, comme souligné par Keynes en 1936 [24].

22. Même si en fonction de leur aversion au risque les individus ne donnent pas tous la même valeur à un gain incertain. Ils peuvent également avoir des préférences pour le présent différentes, et ne pas accorder la même valeur immédiate à un gain futur certain. Mais l'on peut en réalité déplacer ces problèmes en aval, en considérant que cela doit rentrer en compte dans la mesure de probabilité de l'agent (et donc dans l'espérance) et non pas dans l'estimation de V : c'est le concept de probabilité *risque-neutre*.

23. Lorsqu'il y a des dividendes, il faut aussi prendre en compte leur somme actualisée jusqu'à $t = +\infty$.

24. C'est la célèbre métaphore du concours de beauté (Keynes, 1936).

1.6 Critique de la critique

Bien entendu, la plupart des modèles alternatifs ne font pas non plus disparaître leurs hypothèses dans le calcul : ainsi de nombreux modèles à agents postulent un certain nombre de *comportements* ou *règles de décision* possibles, évidemment non exhaustifs, qui conditionnent fortement leurs conclusions. Bien que généralement basés sur des hypothèses comportementales considérées raisonnables, ceux-ci nécessitent toutefois d'être confrontés à la réalité, contrairement aux modèles *normatifs* qui pensent la transcender – sinon, à quoi bon introduire des comportements réalistes ? En souhaitant reproduire la réalité pourtant, beaucoup de modèles sont choisis *ad-hoc*, et sont tout autant soumis à la critique de la section précédente. Trouver que la réalité est *compatible* avec tel ou tel type de comportement *ad-hoc* peut vite se révéler vide de sens, et l'on comprend l'exaspération des adeptes de la rationalité qui au moins ne se laissent pas le droit d'exploiter l'infini des comportments possibles. Les adeptes de modèles comportementaux essayent alors de se donner des règles strictes pour ne pas tomber dans cet excès (Hommes, 2006).

De manière générale, il semble qu'un modèle devrait être tel que toute fonction de décision (stipulée ou cachée) disparaisse dans le calcul, comme le demande Poincaré. Cela devient – presque – possible, dans une certaine limite, grâce à l'*agent hétérogène* comme nous le verrons dans la section suivante, et tout au long de cet ouvrage – même s'il fait une hypothèse : les agents sont *hétérogènes*, et généralement *petits*.

1.7 L'agent représentatif contre l'agent hétérogène

L'*agent* est l'élément de base de la modélisation en économie financière, et ce chapitre d'introduction ne serait pas complet sans une discussion sur la manière de l'inclure dans un modèle. Cela se fait en général de l'une des deux façons suivantes :

L'agent *représentatif* : on suppose que l'économie et constituée de plusieurs classes d'agents qui peuvent chacune être réduites à *un* représentant. A chacun une fonction d'utilité est donnée, en fonction de la classe à laquelle il appartient. Le modèle consiste alors à déterminer l'équilibre qui s'instaure entre ces agents représentatifs, et à comprendre comment celui-ci est affecté par des changements dans leurs fonctions d'utilité ou dans les règles qui leur sont imposées – souvent dans une optique de régulation. L'hypothèse de l'agent représentatif est intimement liée à la notion de fonction d'utilité qu'elle permet d'utiliser, dans un contexte de théorie des jeux.

Supposer que l'économie (ou un marché financier) est réductible à un jeu entre un petit nombre d'agents est toutefois une hypothèse forte : cela nécessite en effet une synchronisation et une homogénéité au sein de chaque classe d'agent qui semblent difficiles à défendre. C'est en réponse à cette critique que le concept d'*agent hétérogène* se développe, connaissant aujourd'hui une popularité

croissante (voir par exemple Kirman (1992); Arthur et al. (1997); Hommes (2006)).

L'agent *hétérogène* : chacune des classes d'agents n'est plus réduite à un unique agent, mais contient de nombreux agents (souvent une infinité, par commodité mathématique) dont les actions sont hétérogènes, soit parce que leurs fonctions d'utilité sont quantitativement différentes (bien que qualitativement similaires dans chaque classe), soit parce que leur connaissance du monde est hétérogène. Il s'agit donc d'une partition *quantitative* du système, au lieu de la partition *qualitative* opérée par l'agent représentatif, et une unique classe d'agents est parfois suffisante pour décrire le système, comme nous le verrons par la suite [25].

C'est lorsque chacun des agents hétérogènes est supposé petit (microscopique) que la magie opère, et que les fonctions arbitraires peuvent disparaître : des lois *globales* peuvent alors émerger, indépendamment des hypothèses initialement émises sur les comportements des agents [26] – et en particulier, leur rationalité ou non. Toutefois, la réalité n'est pas composée uniquement d'agents microscopiques, et il convient de s'assurer que les lois limites conviennent toujours en présence d'acteurs plus gros – ce qui devient difficile sans quelques notions d'économie.

1.8 Lien avec ces travaux

En me lançant dans ce travail, je n'ai jamais eu pour ambition de répondre à des questions aussi générales de modélisation économique, et ma recherche tout au long de cet ouvrage sera subordonnée à la question particulière à laquelle je m'intéresse, que je vais introduire dans le chapitre suivant. Il m'est toutefois apparu que les questions les plus fondamentales n'étaient jamais très loin – il suffit simplement de creuser suffisamment en profondeur et de ne rien accepter tel quel, quelle que soit la question initiale. Et en effet il me semble maintenant *a posteriori* que les conclusions auxquelles ce travail m'a permis d'aboutir donnent certaines leçons sur la manière d'aborder la modélisation en micro-économie, et certains éléments de réponse aux questions posées dans les paragraphes précédents. En particulier, nous nous rendrons compte que la clé n'est parfois pas dans l'efficience mais dans la statistique des agents, et qu'une approche *physique* n'est parfois pas dénuée de sens en économie, et peut permettre d'expliquer certains faits stylisés non-triviaux que les approches classiques ne prédisent pas, justifiant encore une fois le mot d'ordre très physicien : « *Find the essential in the non-obvious* » [27].

25. Chose impossible avec des agents représentatifs : un agent ne peut pas jouer tout seul !

26. Deux remarques sur ce point. (i) Cela rappelle la renormalisation en physique statistique : souvent, la structure microscopique des systèmes disparaît dans le dézoom, faisant apparaître des lois macroscopiques universelles. (ii) Les lois globales (« stables ») peuvent dépendre des caractéristiques microscopique : par exemple, pour des variables aléatoires i.i.d., le théorème centrale limite gaussien ne s'applique que lorsque le moment d'ordre deux existe. Dans le cas contraire, la loi limite est une loi de Lévy stable. Il en va de même ici – néanmoins, le nombre de lois possibles s'en voit drastiquement réduit, et la réalité peut permettre de trancher.

27. *Trouver l'essentiel dans ce qui n'est pas évident.* (Vapnik, 2006). En physique, de nombreuses théories majeures sont en effet nées grâce à des observations inattendues : on peut penser à la théorie de la relativité, à la physique

quantique... – donnant tort à Lord Kelvin, si convaincu par la physique classique qu'il annonçait alors : « *There is nothing new to be discovered in physics now. All that remains is more and more precise measurement* » (Il n'y a plus rien à découvrir en physique aujourd'hui, tout ce qui reste est d'améliorer la précision des mesures).

Chapitre 2

Objectif de l'ouvrage

2.1 Problématique et enjeux

Après autant de considérations générales, il est maintenant grand temps d'introduire la problématique qui a motivé ce travail. La question fondamentale est la suivante : comment les prix se forment-ils sur les marchés financiers ? Et plus généralement, qu'est-ce qu'un prix de marché réellement ? Cette dernière question n'est pas aussi simple qu'elle en a l'air, comme de nombreuses études empiriques l'ont montré avant moi, et comme nous allons le voir. Pour comprendre pourquoi la notion de prix de marché peut être subtile, commençons par poser la question suivante :

Question 1 : Si je souhaite acheter une certaine quantité d'un bien ou d'un actif sur un marché à un moment donné, quel prix obtiendrai-je ?

Il s'avère qu'il n'existe pas de valeur du prix qui réponde universellement à cette question, ce qui met au défi de nombreuses idées reçues sur l'existence d'un unique prix de marché à chaque instant (comme supposé par une écrasante majorité des modèles en mathématiques financières [1]). Plus généralement, cette question est contenue dans la suivante – celle de l'*impact* :

Question 2 : Si je souhaite acheter une certaine quantité d'un bien ou d'un actif sur un marché à un moment donné, comment affecterai-je le marché (immédiatement, et dans le futur) ?

Seulement une fois une réponse donnée à cette dernière question, saura-t-on ce qu'est un prix de marché, et aura-t-on des pistes pour comprendre la question plus générale de la *formation des prix avec l'offre et de la demande* : d'une certaine manière, c'est l'interaction entre acheteurs et vendeurs seule qui détermine les prix, et l'impact est la fonction de transfert entre les actions des agent et l'état du système. Même sans adhérer à cette vision causale de la formation des prix [2], la question

1. Comme les célèbres modèles de Black and Scholes (1973), de Heston (1993), etc.

2. Selon Hasbrouck (2006) : « *Orders do not impact prices. It is more accurate to say that orders forecast prices.* » (Les ordres n'impactent pas les prix. Il est plus juste de dire que les ordres prédisent les prix). Mais alors, comment les prix bougent-ils ?

de l'impact est *fondamentale* dans la compréhension de l'articulation entre l'offre et de la demande, puisqu'en sa qualité de *sonde* de l'état de l'offre et de la demande à un moment donné, il révèle leur *structure* autour du « prix » ainsi que leur *dynamique*.

Une fois une telle compréhension acquise, on peut alors espérer tirer des leçons plus générales sur la modélisation des systèmes économiques. Nous verrons en particulier que les courbes d'offre et de demande Walrasiennes classiques (formalisées par Arrow and Debreu (1954)), encore communément utilisées de nos jours, perdent tout leur sens dans le contexte des marchés financiers, ou du moins y acquièrent des propriétés si singulières que notre manière de les aborder doit profondément changer. Nous verrons également quel rôle joue l'efficience dans la question de l'impact et de la formation des prix – moins fondamental que l'on pourrait le penser. Le rôle de l'efficience en finance, qui est un des seuls domaines où celle-ci peut être testée grâce à la quantité astronomique de données disponibles et au grand nombre d'expériences réalisées – chaque transaction effectuée peut être considérée comme une expérience ! – et à leur caractère standard, me semble être une question cruciale : elle pose en effet la question du bien-fondé des hypothèses d'efficience en général, dans des systèmes souvent moins optimisés et professionnels que la finance mais tout aussi hétérogènes (lorsque par exemple on s'intéresse aux ménages et à leur consommation, aux petites entreprises, etc).

2.2 Ambition globale et approche

Pour répondre à des questions aussi vastes et complexes sur le fonctionnement des marchés, il me semble qu'il faut commencer par s'y confronter [3]. Il m'a donc paru indispensable, pour accomplir cette tâche avec sens, de repartir objectivement du matériau le plus fondamental, c'est-à-dire des données, *sans a priori* sur le type de modèle susceptible de les produire ni sur ses hypothèses de départ ou même son contenu technique. Une fois ce travail effectué seulement, se posera-t-on la question de la modélisation.

Un remarquable travail a été réalisé récemment par Hommes (2006) sur l'utilisation des modèles à agents hétérogènes et de la théorie du chaos en économie. Plus que présenter des modèles précis, son but a été de donner des outils et des pistes de modélisation, et de mettre en évidence les conséquences riches et complexes que des comportements simples peuvent produire. Récemment, Lasry and Lions (2007) ont introduit les *Mean Field Games* (en français, *jeux à champs moyen*) qui introduisent un aspect d'optimisation et de théorie des jeux dans les modèles à agents hétérogènes. De manière similaire, je ne prétendrai pas ici proposer un modèle ultime, et cette notion n'a d'ailleurs pas vraiment de sens lorsque l'on parle de systèmes humains. Il me semble important en revanche d'orienter la réflexion et la modélisation dans des directions pertinentes, et d'introduire les outils qui permettront ensuite à d'autres de développer des modèles plus appropriés aux situations particulières. Les conclusions des présents travaux constituent de convaincants soutiens à ces deux

3. Et idéalement, y agir activement – pas seulement regarder des données de manière passive.

courants de modélisation, et montrent à quel point ils peuvent produire des résultats sur lesquels les modèles à agents représentatifs butent.

Pour y parvenir (ou plutôt pour *tenter* d'y parvenir), je n'adopte pas une « simple » approche de théoricien, dans le sens où je ne m'intéresse pas aux *modèles* en eux-mêmes mais à leur aptitude à représenter la réalité. En effet, il me semble que l'économie et la finance n'ont particulièrement pas manqué de théoriciens au cours de ces dernières décennies, et face à l'impressionnante capacité de l'esprit humain à proposer des explications théoriques à tous genres d'observations [4] ainsi qu'à son étonnante motivation à proposer et résoudre l'ensemble des modèles théoriques possibles et imaginables, j'ai plutôt envie de prendre la direction opposée – ou plutôt, complémentaire. La tâche que je me donne est donc plutôt de juger et de discriminer, *de manière constructive*, c'est-à-dire en convergeant à la fin vers une classe de modèles pertinente – et en ne me contentant pas de confirmer ou d'infirmer empiriquement certaines hypothèses spécifiques, comme cela est souvent fait.

2.3 Plan de l'ouvrage

Pour bien comprendre l'ensemble de ce travail, il est d'abord nécessaire de familiariser le lecteur avec le fonctionnement des marchés modernes, ce que le Chapitre 3 tâchera de réaliser. Le Chapitre 4 s'étendra plus en détail sur l'offre, la demande et la formation des prix, et introduira l'*impact* et les questions qui y sont liés. Après ces chapitres d'introduction je pourrai enfin exposer le coeur de mes travaux. Ceux-ci consistent essentiellement en quatre principaux articles qui aboutiront à une théorie dynamique générale de l'offre et de la demande et fourniront des éléments de réponse aux questions conceptuelles et philosophiques posées dans l'introduction.

Ces quatre articles ont été laissés dans leur langue de publication, c'est-à-dire en anglais. Toutefois, ils seront systématiquement précédés par une préface en français expliquant les enjeux qu'ils contiennent et les questions auxquelles ils tentent de répondre. Une discussion en français est également présentée après chaque article, résumant leurs résultats principaux et les mettant en perspective à l'intérieur du grand objectif fixé au début de cette thèse.

La lecture intégrale des articles n'est pas nécessaire pour qui ne souhaite pas lire en anglais ou investir trop de temps dans leur analyse. Les lecteurs pressés sont donc invités à lire pour chacun leur préface et leur conclusion, qui renverront si nécessaire à leurs passages les plus importants. Les deux premiers articles, plus techniques, présentent tour à tour une étude d'impact inédite effectuée sur le Bitcoin (Chapitre 5) et le modèle d'impact vers lequel les résultats pointent (Chapitre 6). Si le lecteur devait lire en détail un article, je conseillerais la lecture de celui présenté au Chapitre 7 (en sautant les quelques passages techniques), qui tente de présenter plus clairement et sans nécessiter

4. L'impact des transactions sur les prix, si singulier soit-il, a trouvé une explication dans au moins quatre courants de modélisation fondamentalement différents ! C'est dire à quel point l'esprit humain est capable d'identifier des *patterns* et d'assembler astucieusement les briques qu'il connaît pour expliquer toutes sortes de choses.

beaucoup de pré-requis la théorie dynamique générale de l'offre et la demande qui se déduit des deux articles précédents, et discute de ses implications sur plusieurs questions fondamentales concernant la formation des prix et la modélisation multi-agents en micro-économie et en finance. Ces dernières seront ensuite détaillées plus en avant (et en français) dans le Chapitre 8, qui tentera ensuite de les illustrer par divers exemples. Le dernier article (Chapitre 9), assez court et également relativement accessible pour le lecteur non initié, tâche de démontrer à partir de l'étude du Bitcoin l'intérêt d'une compréhension microstructurelle des marchés pour expliquer l'origine – et anticiper – les événements macroscopique extrêmes comme les bulles et des crashs, « bouclant ainsi la boucle » sur la compréhension de la liquidité à toutes les échelles. Le Chapitre 10 conclura ce travail.

Plusieurs travaux non publiés seront également présentés en appendice. L'Appendice A introduira une nouvelle sorte de processus mathématique, représentant l'évolution du prix sous une pression extérieure. Il s'agit d'une généralisation du processus trouvé en Chapitre 6, pour lequel une dualité entre dynamique du carnet d'ordres et processus de prix est mise en évidence. Ce nouveau processus étant plus solidement microfondé et moins « effectif » que de nombreux modèles récents (modèles à noyaux, modèles basés sur des processus de Hawkes), son étude mathématique plus approfondie me semble être un sujet important. L'Appendice B traitera ensuite de l'exécution optimale dans les modèles du Chapitre 6, en présentant des résultats analytiques asymptotiques de liquidation sans « alpha » – c'est-à-dire sans information sur les prix futurs – ainsi que des résultats numériques à la fois pour des problèmes de liquidation avec ou sans « alpha » et pour des problèmes d'arbitrage. Pour finir, l'Appendice C validera empiriquement une nouvelle prédiction du modèle du Chapitre 7 sur la dynamique du volume échangé sur un marché après l'enchère d'ouverture.

Chapitre 3

Microstructure et écologie des marchés financiers modernes

3.1 L'organisation moderne des marchés et le *carnet d'ordres*

Beaucoup ont probablement déjà vu cette image du trading à la corbeille où des dizaines voire des centaines de *traders* échangent « à la criée » des titres dans une cacophonie infernale (voir le film *Rogue Trader* (1999) par exemple). Aujourd'hui en 2016, les choses ont bien changé. La corbeille de la bourse de Paris au Palais Brongniart n'est plus que le vestige d'un passé tourbillonnant – une pièce de musée au-dessus de laquelle les cotations du dernier jour de bourse encore affichées (le 10 juillet 1987 !) donnent l'illusion d'un temps figé. Celle de *Wall Street* n'est guère plus fonctionnelle. Aujourd'hui, la plupart des valeurs sont cotées sur des places de marché (NASDAQ, BATS, NYSE EURONEXT,...) grâce à un *carnet d'ordres* électronique. Plus récemment, la libéralisation des places de marché (en conséquence des directives MiFID en Europe et Reg NMS aux Etats-Unis) a entraîné leur rapide prolifération si bien qu'une même valeur est souvent cotée simultanément sur plusieurs places de marché [1]. Celles-ci se différencient par leur niveau de technologie, leurs tarifs, les types d'ordres qu'elles permettent de placer (cf ci-dessous) ou encore leurs conditions d'entrée (certaines places de marché interdisent par exemple le trading à haute fréquence). Le point commun à toutes les places de marché est le pas de cotation (*tick* en anglais) pour chaque valeur cotée. Celui-ci est la plupart du temps égal à un centime (notamment pour les actions américaines), mais est parfois fixé à des valeurs différentes en fonction du prix du titre et de sa liquidité pour faciliter les échanges, définissant une grille de prix possibles commune à toutes les places de marché. Ainsi, pour chaque valeur il est possible de reconstruire un carnet d'ordres consolidé qui rassemble les carnets d'ordres

1. Pour une compréhension plus exhaustive du fonctionnement des marchés et une analyse de leur évolution récente, voir Lehalle and Laruelle (2013).

de chaque place de marché pour cette valeur : aux Etats-Unis par exemple, le *National Best Bid and Offer* (NBBO) représente le meilleur prix à l'offre et à la demande sur l'ensemble des places de marché pour une valeur donnée. Dans la suite de cet ouvrage, nous éluderons cette question de la fragmentation en nous intéressant toujours au carnet d'ordres *consolidé* (1 valeur cotée = 1 carnet d'ordres).

Le principe du carnet d'ordres est simple pour qui est familier avec les systèmes d'enchères (pensez à Ebay par exemple) : un individu (un *trader*) souhaitant acheter un titre (ou une certaine quantité de titres, qu'on appellera le *volume* de l'ordre ou sa *taille*) pourra soit l'acheter immédiatement au meilleur prix disponible (il s'agit alors d'un ordre *marché*, que l'on décrit souvent comme étant du trading *aggressif*), soit placer un ordre d'achat à un prix plus bas, en attendant qu'un hypothétique vendeur consente à lui vendre à ce prix (ce qu'on appelle un ordre *limite*, que l'on décrit comme étant du trading *passif*). A l'inverse, un individu souhaitant vendre un titre pourra soit le vendre immédiatement au plus offrant avec un ordre marché, soit en demander un prix plus élevé en plaçant un ordre limite et en attendant que se manifeste un acheteur hypothétique. En raison de caractère symétrique entre acheteurs et vendeurs, ce mécanisme de marché est souvent appelé *double-enchère continue*. Les ordres limites d'achat constituent ce qu'on appelle le *bid*, et les ordres limites de vente ce qu'on appelle l'*ask*. On appellera ainsi le NBBO, de manière équivalente, le *best bid* et le *best ask*. Notons qu'une transaction a toujours lieu entre un ordre marché et un ordre limite, puisqu'il faut bien qu'un des deux protagonistes ait été en train d'attendre passivement pour que l'autre puisse venir le saisir agressivement. Notons également que le prix d'un ordre marché placé à l'achat sera toujours supérieur à celui d'un ordre limite (sous-entendu, à l'achat également), puisque par définition ce dernier est utilisé dans le but d'obtenir dans le futur un prix plus avantageux. En contrepartie, un ordre marché assure une exécution immédiate alors qu'un ordre limite pourrait ne jamais l'être si le prix demandé est irréaliste ou si le prix du marché s'éloigne en sens contraire. Terminons enfin cette rapide explication du carnet d'ordres par la notion (indispensable) de priorité pour les ordres limite : si plusieurs ordres limite sont présents lorsqu'un ordre de marché arrive, il faut en effet choisir dans quel ordre les exécuter (notamment si la taille l'ordre marché ne permet pas de les exécuter tous). La première convention universelle est la priorité au prix : les ordres limites dont les prix sont les plus avantageux sont exécutés en premier. Lorsque plusieurs ordres limite proposent le même prix, la convention souvent utilisée est la priorité temporelle : il y a donc une file d'attente par prix possible. Plus un ordre limite a été posté tôt, meilleure est sa position dans la file d'attente à un prix donné, si bien que l'ordre marché exécutera d'abord les ordres limite les plus anciens avant d'exécuter les plus récents si sa taille le permet. A noter que les ordres limite peuvent n'être exécutés que partiellement par un ordre marché : dans ce cas, la partie non exécutée restera en ordre limite sur le carnet d'ordre. Un système où les ordres limites de même prix sont exécutés au pro-rata de leur volume existe également sur les marchés de contrats futurs. Un schéma du carnet d'ordres illustrant son fonctionnement est présenté en Fig.3.1.

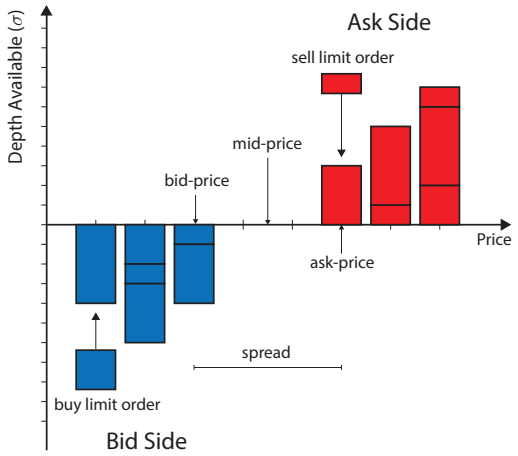

FIGURE 3.1 – Schéma illustrant le fonctionnement d'un carnet d'ordres, tiré de la revue Gould et al. (2013). L'axe vertical représente le volume disponible à chaque prix, chaque ordre étant représenté par un rectangle de hauteur proportionnelle à son volume. Sur un marché à priorité temporelle, un ordre arrivant vient se placer après les ordres déjà existants au même prix, et ne sera exécuté qu'après ces derniers. La différence de prix entre l'ordre limite d'achat au prix le plus élevé et l'ordre limite de vente au prix le plus bas est appelée le *spread*.

De nombreux types d'ordres sont en réalité possibles de nos jours (et pas seulement les ordres limite ou les ordres marché), jouant par exemple sur l'affichage ou non au public de l'ordre (e.g. les ordres cachés ou les ordres iceberg, qui se mélangent au carnet d'ordres mais ne sont révélés qu'*ex post* une fois exécutés, ou encore les ordres placés dans des *dark pools*, indexés sur le NBBO, où l'intégralité de la liquidité [2] est cachée tant qu'elle n'est pas exécutée), ou mixant les caractéristiques (e.g. un ordre limite qui se transforme en ordre marché au bout d'un certain temps s'il n'est pas encore exécuté [3]). On voit tout de même qu'on est loin du prix Black-Scholes à ces échelles : la notion de prix de marché est elle-même mal définie dans le cadre du carnet d'ordres, et plusieurs définitions peuvent être adoptées en fonction du problème :

1. *Les prix auquel les transactions ont lieu* : Cette définition présente deux inconvénients : le prix n'est défini qu'au moment des transactions, à moins de définir le prix en temps continu comme le prix de la dernière transaction qui a eu lieu (ou de la suivante, si l'on regarde les données *a posteriori*). Le second inconvénient, plus pratique, est que les transactions ont lieu alternativement de part et d'autre du spread, ce qui génère une *mean-reversion* importante aux échelles de temps courtes.
2. *Une moyenne glissante des prix de transactions :* Aussi appelés VWAP (*Volume-Weighted Average Price*) quand les prix sont pondérée par les volumes des transactions correspondantes,

2. La liquidité est un concept à multiples facettes. Ici, comprendre : l'ensemble des ordres (passifs) présents dans le carnet d'ordres.

3. Notons que dans le cas où l'ordre limite n'est pas exécuté et un ordre de marché agressif est envoyé, le prix obtenu peut se révéler extrêmement désavantageux ! Il s'agit là d'un premier *hola* à l'idée reçue qu'un ordre limite permet d'obtenir un meilleur prix qu'un ordre marché : cela n'est vrai que *si* il est exécuté, dans le cas contraire il faut s'attendre à payer le prix fort.

et TWAP (*Time-Weighted Average Price*) quand la moyenne est simplement temporelle. Ces prix servent souvent de benchmark aux stratégies d'exécution des brokers, qui s'y comparent pour savoir s'ils ont fait « mieux » ou « moins bien » que le marché sur leur exécution.

3. *Le centre du spread, ou* mid-price *:* Cette définition permet d'éviter en grande partie le bruit de microstructure, et est donc souvent utilisée en pratique. Cependant, en fonction des applications, il peut être plus pertinent de considérer les prix « réels » de transactions, plutôt que ce prix virtuel.

4. *Un autre* microprice *calculé à partir du carnet d'ordres :* Par exemple, la moyenne des prix du bid et de l'ask pondérée par l'inverse des volumes. Cette définition est plus souvent utilisée à des fins de *prédiction*, car elle contient – en principe – plus d'information sur l'évolution future des prix de transactions que les définitions précédentes : s'il y a beaucoup plus d'acheteurs au bid qu'il n'y a de vendeurs à l'ask, on peut s'attendre à ce que les prix de transactions augmentent dans un futur proche.

3.2 L'écosystème, les acteurs et leurs objectifs

Du fonctionnement en ordres limites/ordres marché apparaît la nécessité de la présence des deux types d'ordres pour qu'un tel marché fonctionne, et notamment celle des ordres limite. Sans ceux-ci en effet, pas de carnet d'ordres, pas de NBBO, et pas de prix, dans le sens où personne ne pourrait jamais savoir à quel prix il obtiendrait une transaction s'il la demandait. Historiquement, ce rôle d'assurer en permanence la présence d'ordres limites dans le carnet d'ordres – de *fournisseur de liquidité* – était rempli par des teneurs de marchés désignés (en anglais, *designated market makers*), qui en échange de ce service maintenaient un *spread*, c'est-à-dire qu'ils se proposaient d'acheter en permanence à des prix plus bas qu'ils ne se proposaient de vendre, réalisant un gain sur chaque transaction [4]. Tous les autres acteurs étaient alors des preneurs de liquidité, étant obligés de réaliser chacune de leurs transactions contre un market maker, et en payant le prix. De nos jours, la distinction entre fournisseurs de preneurs de liquidité est moins nette, puisque chacun peut utiliser les ordres limites et les ordres marché à sa convenance. L'activité de market making a par conséquent une tendance naturelle à se délocaliser et à se fragmenter : aujourd'hui, la plus grande partie de la liquidité est fournie par des traders haute fréquence, que la compétition a amené à réduire drastiquement la taille des spreads (qui étaient de l'ordre de 70 bps [5] entre 1900 et 1980, et ne valent plus que quelques bps de nos jours !), de telle sorte que leur profit moyen par transaction soit proche de zéro (au grand bénéfice des preneurs de liquidité).

4. Ce fonctionnement devrait être familier à tout voyageur ayant déjà utilisé un bureau de change : les taux de change réciproques (e.g. EUR→USD et USD→EUR) sont toujours différents l'un de l'autre, et toujours en la défaveur du particulier.

5. 1bp (basis point) = 0.01%, donc 70bps = 0.7% !

Comment l'activité de market making pourrait-elle ne plus être profitable, sachant que par définition l'ordre limite d'achat le plus haut est inférieur à l'ordre limite de vente le plus bas [6], et donc qu'à tout moment le prix d'achat proposé d'un market maker est inférieur d'au moins un tick à son prix de vente proposé (ce qui paraît générer au minimum un profit incompressible d'un demi-tick par transaction) ? La réponse tient en le fait – apparemment simple – que le prix n'est pas fixe mais évolue en permanence, soumettant le market maker à ce qu'on appelle la *sélection adverse*. Pour expliquer ce phénomène, intéressons-nous d'abord à un deuxième type de comportement : la *spéculation*. Spéculer, c'est acheter parce qu'on pense que le prix va monter, dans l'espoir d'effectuer une plus-value à la revente ; ou bien vendre à découvert (*short-selling* en anglais) en prévision d'une chute des prix pour racheter ensuite à un prix plus bas et réaliser un gain. Les marchés financiers, presque par définition, contiennent une grande part de spéculation : c'est en effet ce qui donne un sens aux prix, et permet aux adeptes de l'efficience d'écrire « prix $= \mathbb{E}[V \mid \mathcal{F}]$ » comme évoqué dans l'introduction ! Certains marchés, comme le Bitcoin ou l'action Twitter à l'instant présent, sont même presque entièrement spéculatifs, la possession de l'actif n'occasionnant (quasiment) aucune jouissance présente [7] : leur valeur ne reflète donc que la promesse d'une jouissance future (ou de celle, supposée, de celui qui nous rachètera l'actif). Mais, si tous ces spéculateurs achètent en moyenne lorsque le prix est sur le point de monter, ou vendent en moyenne que lorsque le prix est sur le point de descendre, c'est que le market maker fait le contraire, et donc doit forcément s'attendre à une perte future sur chacune de ses transactions ! L'avantage informationnel de ceux que de nombreux modèles appellent traders *informés* sur le market maker (moins bien informé par définition), est à l'origine de cette *sélection adverse* pour le market maker, qui réalise la plupart du temps ses transactions dans les situations qui lui sont les moins favorables. C'est entre autres pour compenser celle-ci, que ce dernier maintient un *spread* entre son prix d'achat et son prix de vente (et donc, un demi-spread sépare le prix qu'il pense être juste du best bid qu'il propose ainsi que de son best ask). Dans une situation de compétition parfaite, celui-ci pourrait donc réaliser un profit nul en moyenne malgré le spread, si ce dernier compense exactement son déficit d'information sur les prix futurs [8]. Il apparaît donc que le profit d'un market maker pour chaque transaction s'écrit comme :

$$\text{profit} = \text{demi-spread - sélection adverse}$$

Etant obligé pour être exécuté de proposer un meilleur prix que ses concurrents ou du moins un prix égal, peu de latitude est laissée au market maker concernant le demi-spread – à moins d'accepter de réaliser des transactions de manière beaucoup moins fréquentes. En revanche, il peut tenter de minimiser la sélection adverse à laquelle il fait face, en essayant d'obtenir lui aussi de l'information

6. Car sinon ils réaliseraient une transaction l'un avec l'autre et disparaîtraient du carnet d'ordres.

7. Que vaut un actif lorsqu'aucun dividende n'est prévu à l'horizon ? Il vaut... sa valeur future. En un sens, on peut dire que les marchés voient trop loin, contrairement à l'idée répandue du marché court-termiste : si je pense qu'un actif aura de la valeur dans 200 ans, j'ai tout intérêt à y investir, car je trouverai toujours quelqu'un pour me le racheter. Si mon intuition se précise, j'aurai réalisé un bénéfice.

8. Glosten and Milgrom (1985) furent les premiers à formaliser cette notion de sélection adverse.

sur les prix futurs. On voit donc bien pourquoi les rôles se mélangent aujourd'hui, le spéculateur ayant tout intérêt à utiliser des ordres limites en plus des ordres marché, et le market maker ayant tout autant intérêt à se tenir autant informé que possible. [9]

Impossible enfin de conclure cette section sans évoquer le trading à haute fréquence (en anglais *High-Frequency Trading* ou *HFT*) qui a fait couler beaucoup d'encre dans la presse ces derniers temps. C'est tout naturellement que nous y arrivons : une des manières de se tenir informé, dans un contexte où de nombreux produits corrélés [10] sont échangés sur de nombreuses plateformes en des lieux physiquement différents, est en effet de traiter et de transporter l'information d'une plateforme à une autre plus rapidement que les autres. Si le prix d'une valeur monte fortement sur une plateforme, on s'attend à ce qu'il en soit de même sur les autres plateformes, et le premier à profiter de cette prédiction triviale sera celui qui gagnera le plus d'argent – ou en perdra le moins, pour un market maker passif. Ce constat a été à l'origine de l'explosion du trading à haute fréquence ces dernières années, qui en réduisant les temps de trajet entre plateformes au temps de parcours de la lumière [11], a amené les différentes plateformes à une synchronisation sur une échelle de temps de l'ordre de quelques millisecondes [12]. Cette rapidité est toutefois souvent assortie de comportements plus complexes, certains étant activement chassés par les régulateurs car considérés comme parasites. Nous ne rentrerons pas dans ce débat ici, qui nécessite bien plus d'un paragraphe [13].

Je terminerai cette zoologie rapide en présentant un type d'acteur qui me sera utile dans ce qui suit, car il servira souvent à personnifier mon discours lorsqu'il sera question d'*impact* : il s'agit du « gros investisseur » (par exemple l'investisseur institutionnel), ou le cas échéant de son broker. Il s'agit d'un acteur souhaitant exécuter de grosses quantités sur le marché, pour une raison qui lui est propre (souvent reliée à l'information tout de même, ce qui en fait une sorte de spéculateur), mais se trouvant obligé de l'exécuter incrémentalement, la liquidité du marché instantanée étant trop faible pour obtenir un prix d'ensemble convenable – ou simplement inférieure au volume qu'il souhaite exécuter. Cet acteur, ou son broker s'il délègue cette mission, se voit donc confronté à un

9. Certains market makers font également de l'exécution pour le compte de clients, ce qui leur donne une information précieuse sur les tendances futures.

10. Par exemple le même produit échangé sur deux plateformes de marché différentes, mais aussi des produits partiellement corrélés, comme un contrat futur et son sous-jacent ou bien encore deux actions d'entreprises dont les activités sont d'une manière ou d'une autre reliées (Boeing et Airbus, Samsung et Apple...).

11. Le jeu étant de trouver un chemin toujours plus court pour relier deux plateformes : fibre optique plus directe, micro-ondes...

12. On ne serait pas complet sans parler de la *colocation* : pour optimiser le temps de transfert de l'information entre les produits d'une même plateforme, celles-ci proposent aux firmes financières des serveurs dans leurs locaux mêmes. Pour que le service soit standard et équitable, la longueur des fibres optiques qui les relient aux serveurs de l'exchange sont mesurées au millimètre (pour une précision temporelle de l'ordre de la microseconde). Pour plus de détails sur le monde fascinant du trading à haute fréquence, le lecteur peut se reporter aux travaux d'Alexandre Laumonier et lire son livre 6|5.

13. Je retiens principalement un argument en leur faveur : la réduction drastique des coûts de spread/d'impact évoquée plus haut, et un argument contre : la capacité de stockage que le grand nombre d'ordres (pas de transactions !) qu'ils soumettent forcent les acteurs de marché à déployer, et le brouillage de l'information « réelle » que cela provoque. Pour d'autres arguments lire Menkveld (2013); Brogaard et al. (2014); Boehmer et al. (2014); Jones (2013).

problème d'*exécution* : sachant qu'il veut acheter ou vendre un quantité donnée d'un actif avant telle date (souvent, avant la fin de la journée), il doit répartir son exécution dans le temps pour obtenir le meilleur prix possible – ou du moins, un prix correct. Comme le présentera le chapitre suivant, cette exécution de *méta-ordres* constitue une expérience de choix pour tester la réponse du prix, et donc de l'offre et de la demande, à une pression acheteuse ou vendeuse. Elle sera donc le point de départ de l'intégralité de mes travaux.

Chapitre 4

Offre, demande et impact

Il peut paraître étonnant qu'il y ait encore des choses à comprendre sur l'offre et la demande, sujets phares s'il en est de la science économique depuis Walras au XIX$^{\text{ème}}$ siècle. C'est qu'à l'époque il n'y avait pas autant de *données*, et les théories ne pouvaient qu'y être faiblement confrontées. Par conséquent, les théories les plus simples (linéaires) suffisaient, et s'imposaient naturellement face aux lames exigeantes du rasoir d'Ockham. Ce chapitre est destiné à introduire les notions d'offre et de demande ainsi que leurs interprétations Walrasiennes historiques, et expliquera en quelques mots pourquoi celles-ci ne sont pas adaptées aux marchés financiers malgré de nombreuses idées reçues[1]. Il évoquera différentes manières communes de déduire un prix d'une offre et d'une demande, et soulignera leur manque de pertinence malgré leur utilisation dans de nombreux modèles. Il définira ensuite les concepts de *méta-ordre* et d'impact, et présentera certaines pistes classiques de modélisation de ce dernier qui introduiront le coeur des présents travaux. Il se terminera enfin en discutant quelques questions simples mais centrales sur la formation des prix.

4.1 Quelques questions fondamentales sur les prix

Maintenant que le lecteur s'est familiarisé avec le côté *pratique* des marchés financiers, nous pouvons plonger dans le coeur de la discussion. Reprenons les questions évoquées en introduction, auxquelles nous allons tenter de répondre au cours des chapitres suivants :

Question 1 (expérience individuelle) : Si je souhaite acheter une certaine quantité d'un bien ou d'un actif sur un marché à un moment donné, quel prix obtiendrai-je ?

Cette question est en réalité beaucoup plus profonde que son aspect pratique ne le laisse penser. Comme évoqué dans l'introduction, elle défie l'existence d'un prix *universel*, et suggère que le prix est en réalité une quantité mal définie : quel sens donner à un prix si je ne peux pas l'obtenir lors

1. Elles n'étaient d'ailleurs historiquement pas destinées à l'être.

d'un échange ? Et en effet, nous avons vu que le prix pour acheter immédiatement une unité d'un actif est toujours strictement supérieur (d'au moins un tick) au prix disponible pour la vendre : cet écart est ce qu'on appelle le *spread*. On pourrait alors penser qu'il existe non pas *un* prix mais *deux* : un prix d'achat et un prix de vente. La réalité est en fait plus complexe, et le prix que l'on peut espérer obtenir dépend de la quantité d'actif que l'on souhaite acheter ou vendre : plus celle-ci est grande, plus le prix nous sera défavorable. Il suffit alors de réaliser que chaque action effectuée sur le marché est une mesure (directe ou indirecte) de l'offre et de la demande locale, pour se rendre compte du caractère expérimental – et donc de l'intérêt scientifique – de chaque transaction. De nombreux articles, intéressés le plus souvent par des questions de trading optimal, de réduire cet impact à un simple *coût* de transaction : cela serait éluder la question de la formation des prix, et oublier leur caractère *dynamique* et *influençable*. Cela nous amène directement à la question suivante :

Question 2 (expérience individuelle) : Si je souhaite acheter une certaine quantité d'un bien ou d'un actif sur un marché à un moment donné, comment affecterai-je le marché (immédiatement, et dans le futur) ?

Bien qu'étroitement reliée à la précédente, cette question soulève un problème des plus fondamentaux : mon action sur le marché n'influence pas uniquement le prix que j'obtiens, comme la vision de *coût d'impact* pourrait le laisser penser, mais aussi les prix qu'obtiendront les autres après moi – et potentiellement toutes leurs actions futures et tous les prix futurs. Cette observation donne à la notion d'impact toute sa profondeur : il ne s'agit pas seulement d'une sonde de la structure *instantanée* de l'offre et de la demande, mais de toute sa *dynamique*. Elle pose également la question de l'impact de marché (en anglais, *market impact*), qui comme son nom l'indique représente l'impact d'une action sur l'ensemble du marché et pas seulement sur le prix (ce que l'on désigne en anglais par *price impact*). Toutefois, les prix étant les quantités observables les plus évidentes et les moins bruitées, nous emploierons par abus de langage les deux expression de manière équivalente pour désigner l'impact sur les prix. On se rend mieux compte à ce stade que l'impact fait partie intégrante de la formation des prix, et que ce que l'on connaît sous le nom de « prix de marché » est la résultante de l'agrégation des actions individuelles – et, implicitement, de leurs impacts. Cela mène à la troisième et dernière question :

Question 3 (expérience collective) : Comment les prix se forment-ils avec l'offre et la demande ?

Cette question, plus macroscopique que les deux précédentes, a fait l'objet de nombreuses études théoriques et empiriques. Cependant, il me semble extrêmement difficile d'y répondre directement, si l'on ne répond pas d'abord aux questions sur l'impact individuel. De telles tentatives, qui foisonnent dans la littérature, amènent en général à décrire les prix par des processus effectifs *ad-hoc*, comme les

modèles de propagateur, les modèles auto-régressifs, les modèles de processus gaussiens... qui sont ensuite fittés à l'aide des quantités observées moyennes agrégées sur tout le marché. De tels procédés ne permettent pas d'aboutir à une compréhension profonde des mécanismes de marché, et ignorent par construction toute propriété fine non incluse *a priori* dans le modèle [2]. L'approche alternative, qui consiste à partir de l'action élémentaire pour en déduire l'effet de l'agrégation des actions, tient tout autant de la physique que de la micro-économie : c'est celle-ci que nous adopterons.

4.2 Généralités sur l'offre et la demande

Considérons un bien échangeable entre au moins deux participants. Supposons qu'un prix p soit fixé par une autorité extérieure. L'*offre* agrégée pour ce bien au prix p, notée $S(p)$, désigne la quantité de bien que l'ensemble des participants est prêt à vendre à ce prix. Symétriquement, la *demande* agrégée de bien $D(p)$ est la quantité de bien que l'ensemble des participants souhaite acheter à ce prix. En général, l'offre est une fonction croissante du prix : plus le prix augmente, plus les participants sont prêts à vendre. Inversement, la demande est en général une fonction décroissante du prix : plus le prix est bas, plus les participants souhaitent acheter le bien – cf. Figure 4.1 [3]. Mettons-nous donc à la place d'un commissaire-priseur qui connait les courbes d'offre et de demande et qui souhaite maximiser la quantité de bien échangée entre acheteurs et vendeurs en fixant un prix d'enchères bien choisi. S'il fixe le prix trop haut, de nombreux vendeurs se manifesteront mais trop peu d'acheteurs ; s'il fixe le prix trop bas, peu de vendeurs seront au contraire attirés, et de nombreux acheteurs resteront sur leur faim. Dans les deux cas, la quantité échangée sera faible, et une partie des participants sera frustrée car ils étaient prêts à échanger au prix d'enchère mais n'ont pas pu, faute de contrepartie. Il s'avère, sous certaines conditions, qu'une unique valeur du prix permet à la fois de maximiser la quantité échangée, et de ne frustrer personne, dans le sens où tout acheteur et tout vendeur prêts à échanger au prix fixé pourront effectivement le faire : il s'agit de p^* tel que $D(p^*) = S(p^*)$ ($\equiv \mathcal{Q}^*$). En effet, la quantité échangée si le prix est fixé à p est $\min(S(p), D(p))$, et il est facile de voir que cette fonction admet p^* pour unique maximum dès lors que S et D sont strictement monotones. Dans cette situation, il est facile de déterminer la sensibilité du prix à une variation de l'offre ou de la demande : supposons qu'une quantité Q vienne s'ajouter à la demande, alors le nouveau prix optimal $p^*(Q)$ sera tel que $D(p^*(Q)) + Q = S(p^*(Q))$ ce qui

2. D'aucuns décriraient ce genre de modèles – sans propriétés émergentes – comme des modèles financiers de deuxième génération. Nous défendrons ici une troisième générations de modèles, microfondés.

3. Des cas pathologiques peuvent exister, par exemple dans le luxe, où des prix hauts peuvent être au contraire des facteurs attractifs. Nous ne nous intéresserons pas dans ce qui suit à ce genre de situations, en partie car ces phénomènes se manifestent sur des grandes échelles de prix, alors que nous nous intéresserons d'abord à leur microstructure.

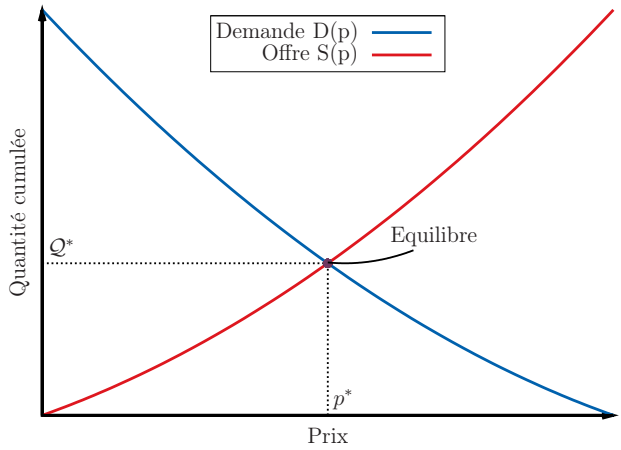

FIGURE 4.1 – Schéma des courbes d'offre et de demande cumulées telles qu'observées par notre commissaire-priseur virtuel. Si les deux courbes sont strictement monotones, elles se coupent en un unique prix p^*, qui sera choisi comme prix d'équilibre, et permettra l'échange d'une quantité $\mathcal{Q}$.

donne, pour Q petit, en développant à l'ordre 1 :

$$p^*(Q) \simeq p^* + \lambda Q, \quad \text{où} \quad \lambda \equiv \frac{1}{S'(p) - D'(p)} > 0. \tag{4.1}$$

Dans cette petite expérience de pensée, un changement dans l'offre ou la demande (plus précisément, l'arrivée d'un acheteur au-dessus du prix ou d'un vendeur en-dessous du prix) affecte donc le prix d'équilibre de manière linéaire, avec un coefficient de proportionnalité qui dépend de la pente des courbes d'offre et de demande autour du prix (et donc, comme nous le verrons plus loin avec l'introduction de l'offre et de la demande *marginales*, de la densité d'acheteurs et de vendeurs autour du prix).

Cependant, utiliser la représentation de la figure 4.1 sur les marchés financiers serait commettre un première erreur évidente : les échanges y ayant lieu en continu, la courbe d'offre en-dessous du prix est nulle, car si un vendeur s'y trouvait sa transaction serait immédiatement exécutée, le faisant instantanément disparaître de la courbe d'offre (et de même pour un acheteur au-dessus du prix) ! L'exemple Walrasien du commissaire priseur est en fait très différent : il suppose que toute l'offre et la demande a eu au préalable le temps nécessaire pour s'accumuler sans effectuer de transaction ! Le caractère continu des marchés financiers semble donc crucial dans la structuration de l'offre et la demande. Cela va en réalité beaucoup plus loin que cette simple remarque, comme l'expliquera le Chapitre 7.

Les notions de courbes d'offre et de demande ne sont pas pour autant caduques : elles semblent juste devoir être adaptées à la situation particulière des marchés. L'introduction du *carnet d'ordres latent* par Tóth et al. (2011) a été un premier pas dans cette direction. Il s'agissait alors de s'affranchir du carnet d'ordres *réel* tel qu'affiché par les exchanges, qui ne révèle que peu d'information

sur l'état global de l'offre et de la demande, et ce pour deux raisons. Premièrement, la soumission d'un ordre limite étant peu engageante, une partie de l'activité du carnet d'ordres est découplée des intentions réelles des agents : il s'agit souvent plus d'un jeu consistant à glaner un maximum d'information sur les intentions des autres participants, et à brouiller les siennes. Deuxièmement, en ce qui concerne les intentions « réelles », seuls les agents immédiatement intéressés par une transaction aux alentours du prix actuel envoient en général leurs ordres limite au marché, dans l'espoir de se faire exécuter le plus rapidement possible. L'offre et la demande plus loin des prix actuels va au contraire rester latente (c'est-à-dire, sous forme d'intentions) et ne se matérialisera sur le carnet d'ordres que quand le prix se rapprochera suffisamment. L'explication est simple : il s'agit de donner le moins d'information possible sur ses intentions de trading, car les autres participants de marché pourraient l'exploiter à nos dépens. Et en effet, le carnet d'ordres n'est habituellement rempli que sur quelques bps autour du prix sur les marchés financiers modernes (1 bp=0.01%). A y regarder de plus près, il apparaît vite que les courbes d'offre et de demande peuvent tout simplement être interprétés comme les carnets d'ordres latents cumulés à partir du prix – et inversement, les carnet d'ordres latent représentent l'offre et la demande marginale au voisinage d'un prix, c'est-à-dire formellement leur dérivée, d'où leur appellation dans le Chapitre 7 : *offre/demande marginales*. Les concepts sont donc cohérents, et l'on pourra interpréter les résultats sur le carnet d'ordres latent comme des résultats sur les courbes d'offre et de demande le cas échéant.

4.3 Quelques (mauvaises) manières d'écrire les prix

Le désir d'écrire un modèle à agents pour décrire le marché nécessite en premier lieu de stipuler comment leurs actions impactent les prix (ce qui, joint à des *comportements* qui stipulent comment les agents réagissent ensuite aux variations de prix, permet d'aboutir à un système autonome). Encouragé par la description Walrasienne des prix présentée ci-dessus, de nombreuses hypothèse simples – mais pas toutes pertinentes – peuvent venir à l'esprit, qui pourtant conditionnent l'ensemble des résultats prédits par le modèle. J'ai choisi d'en mentionner quelques-unes ici, pour expliquer où elles pèchent, et mettre en avant les propriétés fondamentales des marchés à côté desquelles elles passent.

La solution la plus simple, si l'on reprend l'interprétation walrasienne des prix, est de partir à chaque instant t des courbes d'offre et de demande, notées $S_t(p)$ et $D_t(p)$, déterminées à partir des comportements des agents, et d'en déduire le prix de marché implicite p_t qui maximiserait la quantité échangée. Déterminer une dynamique du prix revient dans ce cas à déterminer dans un premier temps des dynamiques pour l'offre et la demande, puis à résoudre à tout instant l'équation implicite $S_t(p_t) = D_t(p_t)$. Cette représentation « top-down » *offre/demande* $\rightarrow$ *prix* n'est justifiée que si ce qui se passe sur le marché ne rétroagit pas sur l'offre et la demande, par exemple si ces dernières se reconstruisent totalement d'une période à l'autre, « oubliant » à chaque fois les événements des périodes précédentes (t est alors forcément discret) – ce qui peut être le cas pour

des biens non stockables. Sur un marché financier coté en continu et pour lequel les biens échangés sont non-périssables, l'équilibre est plus subtil, et prendre en compte la rétroaction des transactions sur l'offre et la demande devient nécessaire – nous y viendrons pus tard.

Dans ce même contexte, une méthode de fixation des prix qui peut venir à l'esprit (Patzelt and Pawelzik, 2013), et qui semble naturelle pour des questions d'unités de mesure, est de fixer à tout moment le prix comme le ratio entre une demande et une offre, $p_t = \tilde{D}(t)/S(t)$ où $\tilde{D}(t)$ est une demande en unités de *devise* (et non pas en unités de titres comme ci-dessus) et $S(t)$ est une offre en unités de titres. Cette méthode de fixation du prix résulte implicitement d'une hypothèse simple : elle suppose que chaque agent décide d'une quantité du bien qu'il possède (des devises s'il achète, des titres s'il vend) pour lequel il demande un échange à n'importe quel prix : le prix est donc fixé à partir des volumes, sans qu'aucun agent ne fixe lui-même de prix. La fonction de demande est donc $D_t(p) = \tilde{D}(t)/p$ (c'est le nombre de titres que les acheteurs peuvent acquérir avec une quantité de devises $\tilde{D}(t)$ si le prix est p) et la fonction d'offre est simplement $S_t(p) = S(t)$ – elle ne dépend pas du prix. Si l'on fait ensuite la même hypothèse que précédemment, où à chaque période $S_t(p_t) = D_t(p_t)$, on obtient $\tilde{D}(t)/p_t = S(t)$ et donc $p_t = \tilde{D}(t)/S(t)$. Cette méthode est par conséquent sujette aux mêmes commentaires que la précédente – en plus de l'hypothèse de l'indifférence au prix, qui semble peu pertinente en réalité.

Si utiliser des modèles aux hypothèses simplistes peut avoir un sens lorsque l'on souhaite pouvoir se concentrer sur les calculs intermédiaires et obtenir des formules analytiques, il faut bien garder en mémoire que des ingrédients non réalistes produisent des modèles non réalistes, et n'accorder qu'une foi limitée à leurs conclusions. Au mieux, peuvent-ils nous faire prendre conscience de l'existence de subtilités potentielles – sans pour autant rien dire sur leur existence ou non dans le monde réel.

4.4 *Price impact* : définitions et énigmes

L'impact des actions individuelles sur les prix est le sujet central de cette thèse, ou du moins son point de départ. Il convient donc de l'introduire proprement, de manière la plus pédagogique possible. L'expérience idéale d'impact serait la suivante : (i) choisir aléatoirement une action (acheter/vendre) et un instant d'exécution t_0 (ii) choisir une quantité Q et (iii) exécuter la transaction et enregistrer le(s) prix obtenu(s) (et éventuellement, les prix futurs). En réalité, cette expérience ne permet pas de tirer de réels enseignements : parce que les participants de marché cachent souvent leurs intentions jusqu'au dernier moment, les ordres limites présents sur le carnet d'ordres ne sont qu'une fraction des intentions réelles « latentes » [4], et cette expérience ne fera que sonder l'offre ou la demande *affichée* [5]. Pour laisser le temps aux participants de matérialiser leurs intentions, il convient donc – et c'est ce

4. Cf discussion de la section 4.2 sur la distinction entre carnet d'ordres réel et carnet d'ordres latent.

5. Ce qui ne sert en fait à rien car il suffirait de lire le carnet d'ordres – si l'on oublie l'existence d'ordres cachés ou icebergs, cf Section 3.1.

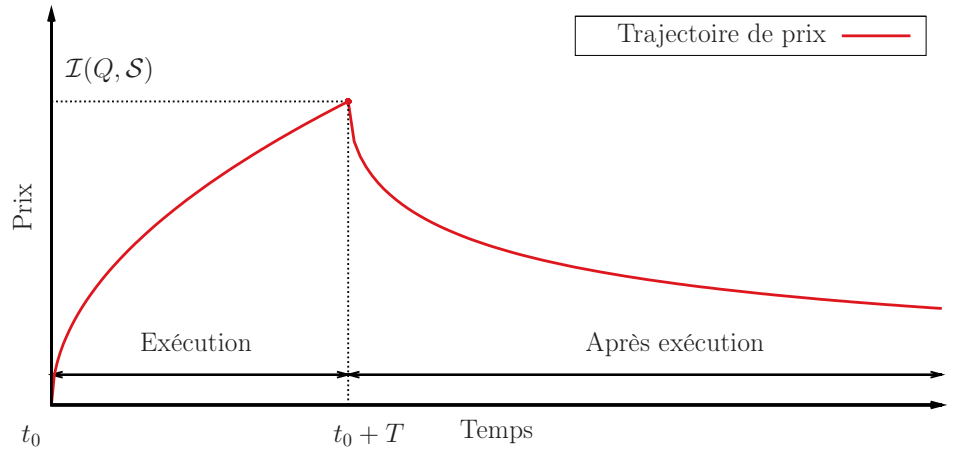

FIGURE 4.2 – Schéma d'une trajectoire d'impact typique pendant un méta-ordre à l'achat d'une durée T et d'une quantité Q. La pression acheteuse pousse le prix à la hausse, puis ce dernier tend à revenir vers sa valeur initiale lorsqu'elle s'arrête (ce qu'on appelle : *résilience*).

que font les gros investisseurs qui souhaitent échanger de grosses quantités – d'exécuter la quantité Q incrémentalement [6]. Cela change quelque peu l'expérience, et l'on se voit obligé d'introduire un *horizon* d'exécution T, ainsi qu'une stratégie d'exécution sur cet intervalle $\mathcal{S}$ – la stratégie la plus raisonnable pour une expérience étant d'adopter un taux d'exécution constant, ce que nous considèrerons souvent implicitement par souci de simplicité [7]. Reprenons donc l'expérience :

i Choisir aléatoirement une action (acheter/vendre) et un instant de départ t_0,

ii Choisir une quantité Q et une stratégie $\mathcal{S}$ (que l'on suppose contenir l'horizon d'exécution $T(\mathcal{S})$, toujours noté T par souci de simplicité),

iii Exécuter l'ordre suivant la stratégie $\mathcal{S}$, en enregistrant la trajectoire de prix obtenue, et éventuellement les prix des transactions qui auront lieu après T (ou un autre prix comme le *mid-price*).

Cet ordre exécuté incrémentalement est appelé *méta-ordre* (en anglais *meta-order*) ou ordre *parent*. Les ordres élémentaires qui le constituent sont appelés ordres *enfants* (en anglais, *child orders*). L'impact d'un tel méta-ordre sur le prix (par exemple, sur le *mid-price*) présente une anatomie universelle, représentée sur la figure 4.2 (dans le cas d'un méta-ordre à l'achat, le cas d'un méta-ordre à la vente étant symétrique) : la pression exercée a tendance à pousser le prix vers le haut, puis ce dernier revient lorsque la pression s'arrête. Bien entendu, d'autres agents peuvent affecter les prix pendant l'exécution du méta-ordres, et à cette trajectoire d'impact se superpose en général un bruit de marché : nous parlerons donc ci-dessous de trajectoires moyennes, pour éliminer ce bruit. Nous pouvons donc définir les quantités suivantes :

6. Une limite tacite est de ne pas dépasser environ 20% du flux du marché lorsque l'on exécute un tel ordre.
7. Par exemple, acheter Q/T titres toutes les secondes pendant T secondes, à conditions de marché constantes.

Definition 4.4.1. *L'impact final $\mathcal{I}(\mathcal{S}(Q))$ est la différence entre le prix à la fin de l'exécution du volume Q par la stratégie $\mathcal{S}$, et le prix au début de l'exécution.*

La trajectoire d'impact étant en moyenne monotone pour une exécution à vitesse constante (croissante pour un méta-ordre à l'achat, décroissante pour un méta-ordre à la vente), on emploie souvent le terme *pic d'impact* (en anglais *peak impact*) pour désigner l'impact final lorsque l'on parle de trajectoires moyennes. Pour d'autres styles d'exécution, le maximum de l'impact pourrait très bien être atteint avant T en moyenne, et cette terminologie n'aurait pas vraiment de sens [8]. Notons que Q intervient comme argument de la stratégie $\mathcal{S}$, et que l'horizon T est absent de la définition car il est inclus dans la stratégie. En pratique, nous verrons que l'impact dépend essentiellement du volume exécuté Q, de manière modérée de l'horizon d'exécution T et relativement peu de la stratégie utilisée. Pour ces raisons, nous écrirons souvent $\mathcal{I}(\mathcal{S}(Q))$ comme $\mathcal{I}(Q,T)$, ou même $\mathcal{I}(Q)$.

Definition 4.4.2. *Le* chemin d'impact $\{I(\mathcal{S}(q_t))\}_{t\in[0,T]}$ *(en anglais* impact path*), où q_t est le volume exécuté au temps t, décrit le chemin par lequel le prix passe durant l'exécution pour atteindre l'impact final.*

Par abus de notation, nous avons utilisé $\mathcal{S}(q_t)$ pour désigner la sous-partie de stratégie $\mathcal{S}(Q)$ employée jusqu'au temps t, c'est-à-dire la stratégie $\tilde{\mathcal{S}}$ définie comme la stratégie $\mathcal{S}$ arrêtée au temps d'arrêt $t = \inf\{s \mid q(s) \geq q_t\}$. Cet abus est justifié par l'intuition que l'impact de l'exécution de la quantité q_t sur le prix ne dépend pas de ce qui se passe ensuite (il est « $\mathcal{F}_t$-mesurable »), et est donc égal à l'impact *final* d'une exécution hypothétique $\tilde{\mathcal{S}}(q_t)$ s'arrêtant à t. Ces définitions étant posées, nous pouvons définir le slippage :

Definition 4.4.3. *Le* slippage $\Delta S(\mathcal{S}(Q))$ *est la différence entre le prix moyen réalisé et le prix de décision pour une stratégie d'achat (et l'opposé pour une stratégie de vente, de manière à être positif en moyenne dans les deux cas).*

Le slippage est directement relié au coût d'impact (en principe positif lui aussi), qui s'écrit comme $C(\mathcal{S}(Q)) = |Q| * \Delta S(\mathcal{S}(Q))$. Cette quantité est très utilisée en pratique pour mesurer la performance d'une stratégie d'exécution. Terminons enfin par une notion des plus cruciales :

Definition 4.4.4. *L'*impact permanent $\mathcal{I}^{\infty}(\mathcal{S}(Q))$ *(en anglais* permanent impact*) désigne l'impact sur le prix qui reste de manière permanente longtemps après l'exécution ($t \to \infty$) – par opposition à l'impact transitoire qui désigne la partie qui décroît intégralement après l'exécution.*

La mesure de l'impact permanent est cruciale pour la théorie de la formation des prix : une action, même aléatoire – et donc sans aucun fondement économique – peut-elle affecter les prix à tout jamais, ou est-elle nécessairement absorbée en totalité si l'on attend suffisamment longtemps ?

8. C'est ce qui se passe dans Zarinelli et al. (2015).

La question de savoir si l'impact est en partie *permanent* ou s'il est uniquement *transitoire* est donc cruciale dans la question de l'efficience des marchés [9].

4.5 Pistes de modélisation de l'impact

Les principales définitions étant maintenant posées, nous pouvons aller plus en avant dans la compréhension du phénomène d'impact.

4.5.1 Vision Walrasienne

Reprenons l'écriture Walrasienne du prix avec l'offre et la demande de la section 4.2, avec $S(p)$ et $D(p)$ les courbes d'offre et de demande, et p^* le prix tel que $S(p^*) = D(p^*)$. Détaillons un peu plus l'obtention de l'équation 4.1 : supposons qu'un individu additionnel prenne part à l'enchère pour un volume Q (positif s'il achète, négatif s'il vend), changeant alors la courbe de demande $D(p)$ en $D(p)+Q$ s'il est acheteur ou la courbe d'offre $S(p)$ en $S(p)-Q$ s'il est vendeur. Le nouveau prix d'équilibre $\hat{p}^*$ doit alors être tel que $S(\hat{p}^*) = D(\hat{p}^*) + Q$ s'il est acheteur ou $S(\hat{p}^*) - Q = D(\hat{p}^*)$ s'il est vendeur. En soustrayant les équations d'équilibre avec et sans Q, respectivement associées à p^* et $\hat{p}^*$, on obtient dans les deux cas :

$$S(\hat{p}^*) - S(p^*) = D(\hat{p}^*) - D(p^*) + Q \tag{4.2}$$

Un simple développement de Taylor de $S(p)$ et $D(p)$ autour de p^* donne :

$$(\hat{p}^* - p^*)\, S'(p) \simeq (\hat{p}^* - p^*)\, D'(p) + Q \tag{4.3}$$

ce qui peut se réécrire sous la forme :

$$\hat{p}^* \simeq p^* + \lambda Q, \quad \text{où} \quad \lambda = \frac{1}{S'(p) - D'(p)} > 0. \tag{4.4}$$

Dans l'expérience de pensée de l'enchère Walrasienne, l'impact sur le prix d'une action est proportionnelle à son volume, et le coefficient de proportionnalité est donné par l'inverse de l'offre et de la demande marginales $S'(p)$ et $D'(p)$. [10] Cette linéarité est également retrouvée par le célèbre modèle de Kyle, que nous présentons maintenant.

9. Pour des études empiriques et des discussions sur l'impact permanent, se reporter à Gomes and Waelbroeck (2015); Brokmann et al. (2015) ou au Chapitre 5.

10. L'offre et la demande marginales correspondent à une intuition simple : si le prix se déplace de p à $p + \mathrm{d}p$, le nombre de nouveaux vendeurs qui va se manifester est $S'(p)\mathrm{d}p$, et inversement $-D'(p)\mathrm{d}p$ nouveaux acheteurs se manifesteront si le prix de déplace de p à $p - \mathrm{d}p$. L'offre et la demande marginale sont donc assimilables à la *liquidité* autour d'un prix donné, et le prix est d'autant plus robuste qu'elles sont importantes.

4.5.2 Le modèle de Kyle

La vision Walrasienne souffre d'un problème majeur lorsque l'on parle de marchés financiers : elle représente un monde statique, alors que les marchés fonctionnent en continu, absorbant un flux d'information lui même dynamique. Cet aspect dynamique introduit une toute nouvelle dimension : l'arbitrage. Celui-ci se manifeste de deux manières : la spéculation sur la valeur future du bien – basée sur de l'information – et l'arbitrage d'inefficiences comportementales. Partant de ce constant, Kyle propose en 1985 un modèle à une période[11] où trois types d'agents coexistent : un trader *informé* qui tente d'exploiter une information qu'il possède sur le prix futur V en soumettant un volume $x = X(V)$ bien choisi, un *noise trader* (trader non informé) qui soumet de manière aléatoire un volume u, et un *market maker* qui reçoit le volume agrégé (net) $x+u$ de ces deux acteurs et doit proposer un prix $p = P(x+u)$ auquel il accepte de réaliser les transactions demandées (la fonction P(.) sera appelée *règle de pricing*). Au début du jeu, le prix est supposé égal à p_0[12].

L'équilibre doit être tel que : (i) le market maker, supposé compétitif, réalise un profit moyen nul et (ii) la stratégie du trader informé soit optimale vis-à-vis de la règle de pricing du market maker. Kyle montre alors l'existence d'un unique équilibre si la règle de pricing du market maker est linéaire et si le volume soumis par le *noise trader* suit une loi gaussienne :

$$\begin{cases} P(x+u) = p_0 + \lambda(x+u) \\ X(V) = \dfrac{V-p}{2\lambda}, \end{cases} \tag{4.5}$$

générant un profit espéré nul pour le market maker et un profit espéré optimal de $\frac{(V-p)^2}{4\lambda}$ pour le trader informé. Le paramètre λ représente la sensibilité du prix du market maker aux volumes, c'est-à-dire, le coefficient d'*impact*. L'impact apparaît alors sous un angle nouveau : il s'agit d'une *prime* que le market maker fait payer aux agents pour compenser son déficit d'information face aux traders informés, et qui représente *exactement* son anticipation des prix futurs, sachant le flux d'ordres qui lui parvient (c'est ce qu'impose la condition de profit nul). L'expression du paramètre λ est importante : il représente simplement le ratio entre l'avantage informationnel moyen du trader informé et le volume moyen soumis par le noise trader. Le modèle permet donc de prédire les déterminants d'un marché *liquide* : plus le volume du noise trader est important (et donc plus les acteurs de marché sont stupides), moins market maker est sensible et plus le coefficient d'impact est faible, car une unité de volume ne contient alors en moyenne que peu d'information sur les prix futurs. Le trader informé doit donc exécuter un volume important s'il veut incorporer intégralement son information dans les prix (ce qui lui permet au passage de réaliser un profit important). Cette

11. Il introduit surtout un modèle en temps continu. Nous nous intéressons ici seulement au jeu à une période car il est plus simple et fournit déjà de bonnes intuitions.

12. Le market maker et le trader informé sont donc initialement d'accord sur la valeur du prix, avant que ce dernier n'acquière une information privée.

robustesse permet entre autres d'assurer une certaine résistance du prix face aux comportements malveillants ou stupides. En revanche, si tout le flux d'ordres est informé, le market maker sait qu'il sera perdant à chaque coup et doit donc faire payer très cher l'impact – dans ce cas, de très petits volumes suffisent à bouger les prix de manière importante, et les prix sont fragiles. Ces conclusions qualitatives sont tout à fait en accord avec les comportements réels, puisque lorsqu'une nouvelle importante est attendue les market makers tendent à diminuer la liquidité qu'ils offrent au marché, craignant de subir la sélection adverse occasionnée par un flux d'ordres trop informé. Lors de l'arrivée d'une telle nouvelle, il n'est pas rare de voir les prix subir des sauts importants sans qu'aucune transaction n'ait lieu (ou très peu).

Si l'on prend un peu de recul, on se rend compte que dans ce modèle *personne* n'impacte réellement les prix. Souvenons-nons de l'expression du prix :

$$\text{prix} = \mathbb{E}\left[V \mid \mathcal{F}\right]. \tag{4.6}$$

L'*impact* dont le modèle de Kyle fait mention, c'est en fait la mise à jour du set d'information $\mathcal{F}$ du market maker : le prix impacté n'est qu'une avancée dans la direction du *vrai* prix. D'où l'interprétation de Hasbrouck que les ordres *prévoient* les prix – plus qu'ils ne les *impactent*. Dit simplement, les ordres ne changent rien au prix du marché : si on leur observe un impact, c'est simplement qu'ils avaient bien prédit les changements de prix futurs. Cette croyance en un prix fondamental transcendant est très caractéristique de nombreux économistes. La principale reproche que l'on peut formuler à son égard est son caractère religieux : dès que l'hypothèse est faite, elle peut en effet tout expliquer. Nous revenons là aux critiques formulées dans l'introduction au sujet de l'efficience et de la valeur fondamentale – concepts absolument nécessaires pour ce genre de modèles.

Ce modèle a tout de même de nombreuses vertus pédagogiques, et permet de voir l'impact sous un angle différent : l'impact apparaît comme un coût pour les traders que le market maker fait payer pour compenser le désavantage informationnel qu'il s'attend à avoir. Si le flux d'ordres qu'il reçoit est en moyenne non informé, le risque qu'il encourt est faible et il peut charger une prime faible par transaction. Dans le cas contraire, il doit faire payer très cher chaque transaction, car il sait qu'il est le *dindon de la farce*. La spéculation devient donc l'art de posséder plus d'information que la moyenne des autres pour un volume donné, de manière à surpasser le coût d'impact – ce qui peut se faire entre autres en diminuant le volume, d'où la question cruciale de *dimensionnement* et de *capacité* d'une stratégie.

4.5.3 L'énigme de l'impact concave

La théorie walrasienne de l'offre et de la demande, tout comme le modèle original de trading informé de Kyle (1985) – qui reste un modèle central de l'économie financière – prédisent donc un impact linéaire en la quantité exécutée. Toutefois, depuis plus de 20 ans de nombreuses études

empiriques ont démontré que la fonction d'impact $\mathcal{I}(Q)$ était fortement concave, mettant en doute leurs conclusions – et pire, leurs hypothèses. En fait, il s'avère que l'impact moyen d'un méta-ordre de volume Q sur le prix d'un actif est remarquablement bien prédit par la formule suivante, quelles que soient les caractéristiques du marché considéré (voir par exemple Tóth et al. (2011)) :

$$\mathcal{I}(Q) = Y\sigma_D\sqrt{\frac{Q}{V_D}}, \tag{4.7}$$

où Q est le volume du méta-ordre, σ_D est la volatilité journalière, V_D le volume journalier échangé sur le marché et Y un pré-facteur sans unité proche de 1, souvent appelé Y-ratio. Malheureusement, les modèles évoqués ci-dessus ne fournissent aucune piste d'explication des observations empiriques d'un tel impact concave. Pour expliquer ce phénomène, de nombreuses pistes d'explications ont été proposées ces dernières années. Cette section a pour but d'en réaliser un rapide tour d'horizon, pour introduire la subtilité de la question – et aussi pour montrer à quel point la question est controversée et l'esprit humain plein de ressources.

1. Raison de risque : BARRA (Torre and Ferrari, 1997) propose une idée simple, où l'impact n'est pas une compensation de la sélection adverse mais du risque d'inventaire que le market maker subit. Supposons que le market maker commence avec une position nulle et que le prix suive un mouvement brownien σW_t auquel le market maker peut trader une quantité μV par unité de temps, où V est le volume moyen échangé par unité de temps sur le marché et μ son taux de participation. Pour se débarrasser d'un inventaire Q, il aura donc besoin d'un temps $T = Q/\mu V$ et le prix obtenu suivra une loi $\mathcal{N}(p_0, \sigma^2\frac{Q}{\mu V})$. Sachant cela, il paraît naturel pour le market maker de faire payer une prime proportionnelle à l'écart-type du prix obtenu, car cela permet d'assurer un gain dans pour un pourcentage des cas contrôlé et défini à l'avance – un contrôle de risque comme un autre. Cette prime s'exprime alors comme $I(Q) = C\sigma\sqrt{\frac{Q}{\mu V}} = \frac{C}{\sqrt{\mu}}\sigma_D\sqrt{\frac{Q}{V_D}}$ en introduisant les volatilités et volumes journaliers, et où $C \simeq 1$. Bien que produisant un résultat intéressant, cette idée se base sur de mauvaises hypothèses : elle suppose entre autres (i) que l'impact provient d'une condition de risque, alors qu'il a été montré que la relation *impact = sélection adverse* est remarquablement vérifiée [13], (ii) que le market maker est la seule contrepartie du métaordre, qu'il prend de plus dans son intégralité, alors que l'échelle des market makers est en principe bien plus courte que celle des métaordres, et (iii) que le market maker commence avec un inventaire nul – dans le cas contraire la singularité en $\sqrt{Q}$ disparaît au profit d'une relation linéaire. Il faudrait donc une synchronisation exceptionnelle du métaordre avec tous les market makers pour observer cette singularité. De plus, les hypothèses initiales sont irréalistes sur un marché comme le Bitcoin,

13. Il est d'ailleurs intéressant de noter que cette explication voit uniquement l'impact comme un *coût* et aucunement comme un ingrédient de la formation des prix avec l'offre et la demande : l'impact des ordres sur les prix n'affecte pas son mouvement, imperturbable, qui suit la marche aléatoire σW_t indépendamment des actions des participants de marché.

où très peu de market making algorithmique a lieu (voir Chapitre 5). Enfin, elle prédit un Y-ratio plusieurs fois supérieur à sa valeur empirique.

2. Raison de corrélations : l'idée est de reprendre l'argument original de Kyle, où un teneur de marché compétitif donne un prix à un trader informé de manière à réaliser un profit nul, mais avec l'hypothèse que le trader informé exécute ses méta-ordres incrémentalement sur plusieurs enchères successives. Cela produit donc une corrélation entre les volumes soumis lors d'enchères successives, de telle sorte que pour contrer celle-ci le teneur de marché doive imposer un impact concave – un impact linéaire produirait en effet des variations de prix autocorrélées, donc prévisibles [14]. Cette idée a été développée sous plusieurs formes par Donier (2012); Farmer et al. (2013); Jaisson (2015). Toutefois elle ne prédit qu'une concavité asymptotique, ce qui nous le verrons n'est pas satisfaisant. De plus, un certain nombre des hypothèses effectuées ne sont pas validées empiriquement – ce que l'on verra au Chapitre 5.

3. Raison d'analyse dimensionnelle : si l'on suppose qu'exécuter de manière directionnelle une fraction du volume total échangé sur le marché dans un intervalle de temps donné produit une volatilité due à l'impact égale à une fraction de la volatilité habituelle dans cet intervalle de temps (par exemple qu'exécuter 100% du volume journalier de manière directionnelle produit un impact égal à 100% de la volatilité journalière), alors la seule loi d'impact possible pour des raisons de dimensions est $I(Q) = \sigma_D \sqrt{\frac{Q}{V_D}}$. [15] Cependant, cette hypothèse n'est ni explicative ni intuitive !

4. Raison mécanique : la loi d'impact tient *malgré* les individus, sans être *produite* intentionnellement par eux (comme cela est le cas dans les interprétations précédentes et dans celle de Kyle), et émergeant mécaniquement de l'interaction entre de nombreux agents hétérogènes. Cette piste sera fortement appuyée par les résultats empiriques du Chapitre 5, et fera l'objet de plusieurs chapitres de cet ouvrage

4.6 Manipulation de prix et arbitrage

Si les mathématiciens sont en général plus ou moins agnostiques sur la formation des prix, ils imposent en général aux modèles de marché une condition de cohérence bien précise : l'absence d'*arbitrage*. De nombreuses définitions de l'arbitrage ont été données, qui correspondent à des notions et à des critères mathématiques à chaque fois différents – voir par exemple Huberman and Stanzl

14. Il s'agit également de l'intuition des modèles à propagateurs, dans lesquels l'autocorrélation du flux d'ordres doit être compensé par une décroissance de l'impact (qui produit une concavité dans le cas de métaordres) pour que les prix restent efficients. Une manière équivalente de le présenter, est que l'information d'ordres successifs dans la même direction devient redondante : il faut donc qu'ils impactent de moins en moins les prix au fur et à mesure qu'ils arrivent.

15. En effet, trader xQ doit alors produire un impact $I(xQ) = \frac{\sigma_{xT}}{\sigma_T} I(Q) = \sqrt{x} I(Q)$, ce qui en remplaçant x par $1/Q$ produit $I(Q) = I(1)\sqrt{Q}$. Pour $Q = V_D$, cela donne $\sigma_D = I(V_D) = I(1)\sqrt{V_D}$ d'où $I(1) = \sigma_D/\sqrt{V_D}$, CQFD.

(2004); Gatheral (2010). La notion d'arbitrage que nous utiliserons est l'arbitrage *en moyenne* : il sera nécessaire, pour qu'un modèle d'impact soit considéré comme valide, qu'il ne permette pas de réaliser des profits en espérance pour toute stratégie de trading à inventaires initial et final nuls qui ne se base sur aucune information sur le prix futur. La condition de non-arbitrage en espérance s'apparente en fait à une condition de bonne définition mathématique du problème de minimisation de l'impact : sans elle, la question de minimisation de l'impact en moyenne s'apparenterait à une question d'arbitrage d'un modèle mal posé. Elle est également cohérente avec l'intuition que la construction d'une machine de pompage pour extraire mécaniquement des profits du marché est impossible – il faut au moins utiliser de l'information.

4.7 FAQ

Avant de conclure ce chapitre, j'aimerais répondre à une question qui revient de manière particulièrement fréquente sur l'évolution des prix :

Puisque chaque transaction comporte autant *de volume acheteur que de volume vendeur, pourquoi les prix bougent-ils ?*

En effet, l'idée naïve que le prix baisse quand « tout le monde vend » et monte quand « tout le monde achète » s'effondre face au constat simpliste de la symétrie des transactions ! Pourtant, il est souvent considéré qu'un ordre marché à l'achat pousse le prix à la hausse (et inversement à la vente), et il faut bien de toute manière que les prix évoluent. La subtilité se trouve dans l'asymétrie entre ordre agressif (ordre marché) et ordre passif (ordre limite). Si l'ordre marché affecte le prix au moment de la transaction, c'est en fait bien avant cela que l'ordre limite a eu un impact : cela a eu lieu dès le moment où il s'est retrouvé au best bid/ask (ou même dans le carnet d'ordres). Même si *in fine* tout se passe comme si aucun des deux ordres n'avait été soumis, le prix étant alors fixé par les ordres limite restants, c'est ce caractère asynchrone qui permet au prix d'avoir une dynamique. En fait, si chaque ordre limite est considéré comme temporaire – car il sera probablement annulé ou exécuté un jour ou l'autre – on s'aperçoit que le prix de marché n'est défini *que* par des ordres temporaires en attente d'être exécutés, et pas par les transactions passées ! Une telle vision donne un peu le vertige : le prix y semble tellement fragile...

On rétorquera que la distinction ordre limite/ordre marché est artificielle, et qu'on pourrait très bien imaginer des mécanismes de marché où cette distinction n'est pas nécessaire (en mettant par exemple au point un système d'enchères successives). Cependant, on se rend rapidement compte que le prix ne peut être cohérent dans le temps que si une partie des intentions des participants *persiste* entre les enchères pour empêcher un comportement de prix trop erratique... [16] ce qui nous ramène

16. En effet, si les intentions des participants de marché étaient entièrement effacées après une enchère et tirées selon une nouvelle distribution aléatoire pour l'enchère suivante, des fluctuations de type bruit blanc viendraient s'ajouter à l'évolution du prix, ce qui ne serait pas très satisfaisant, ni rassurant.

aux ordres limites ! Ce sont donc bien les ordres *non exécutés* qui définissent le prix, même dans ce mécanisme (ou plus généralement, les intentions pour les transactions futures) – ce qui montre la nécessité de modéliser la dynamique du carnet d'ordres sous une forme ou une autre.

Cet exemple nous montre cependant que la formation des prix a un caractère plus général que l'interaction entre ordres limites et ordres marchés sur un carnet d'ordres en temps continu. Il semble donc important de développer une théorie plus générale de l'offre et de la demande qui ne dépende pas du choix pratique d'implémentation du marché [17] – à laquelle nous arriverons au Chapitre 7.

17. D'autant plus que certains en ce moment souhaiteraient remplacer la cotation en continu par des enchères successives fréquentes pour contrer le trading à haute fréquence (Budish et al., 2013; Fricke and Gerig, 2014).

Deuxième partie

Etude des données et pistes de modélisation

Chapitre 5

Analyse de l'impact sur le Bitcoin, ou comment utiliser des données pour répondre à des questions profondes sur l'origine de l'impact

From
A Million Metaorder Analysis of Market Impact on the Bitcoin
with Julius Bonart
(Donier and Bonart, 2014)

5.1 Préface (français)

Nous voyons donc que les questions sur le rôle de l'impact des actions des agents dans la formation des prix et sur ses déterminants ne manquent pas. Face à autant d'interrogations, il fallait donc trouver un moyen de discriminer – pour autant que ce soit possible. C'est ce que l'article qui suit s'est donné pour but, en étudiant en détail l'impact sur un marché très particulier : le marché d'échange du Bitcoin[1]. De par son caractère amateur, son évolution rapide[2] et à cause des frais de transactions importants qui y étaient en vigueur à l'époque de l'étude, on pouvait avant même de commencer y exclure toute explication de l'impact basée sur l'arbitrage et l'efficience, du moins pour toute variation de prix inférieure à 1% – ce qui s'est avéré être le cas pour l'écrasante majorité des méta-ordres. Si la formule d'impact de l'Equation 4.7 y était vérifiée, on devrait alors y trouver

1. Et essentiellement le marché Bitcoin/USD qui à l'époque de l'étude était de loin le plus important. Il est aujourd'hui surpassé par les marchés chinois Bitcoin/Yuan.

2. Pour se faire une idée, il suffit de regarder l'évolution du prix du Bitcoin entre 2008 et 2014, ainsi que de son nombre d'adeptes (cf https ://blockchain.info/charts) !

une autre explication. Ce fut en effet le cas : cet article, court mais assez dense, a donc joué un rôle crucial dans ces travaux et lancé l'ensemble de l'effort de modélisation qui va suivre.

En tant qu'adepte habituel de l'efficience, je ne puis toutefois m'empêcher de poser la question malgré tout : Quel rôle joue-t-elle dans tout cela ? Elle n'est en effet pas tout à fait absente : mais nous discuterons de ce point plus loin.

Abstract : We present a thorough empirical analysis of market impact on the Bitcoin/USD exchange market using a complete dataset that allows us to reconstruct more than one million metaorders. We empirically confirm the "square-root law" for market impact, which holds on four decades in spite of the quasi-absence of statistical arbitrage and market marking strategies. We show that the square-root impact holds during the whole trajectory of a metaorder and not only for the final execution price. We also attempt to decompose the order flow into an "informed" and "uninformed" component, the latter leading to an almost complete long-term decay of impact. This study sheds light on the hypotheses and predictions of several market impact models recently proposed in the literature and promotes heterogeneous agent models as promising candidates to explain price impact on the Bitcoin market – and, we believe, on other markets as well.

5.2 Introduction

Most financial markets have undergone rapid changes in the past 10 years. The implementation of limit order books, the electronization of trade exchanges, the rise of high-frequency trading, and the introduction of algorithmic and automated executions are the most emblematic features of the new financial economy. Both supporters Hendershott et al. (2011); Brogaard et al. (2014) and critics Budish et al. (2013); Kirilenko et al. (2015) of this evolution hold the view that the recent speed revolution has had a profound impact on the functioning of financial markets. While this is obviously true from a technological and microstructural point of view, it is less clear from an economists' perspective : In the end, a stock exchange is a market place where buyers and sellers meet and agree on a price ; and it is not straightforward to assess the relevance of technological changes to this process.

It is however possible to directly compare financial markets which strongly differ in their respective degrees of technological sophistication and latency. The recent development of cryto-currencies such as the Bitcoin, traded against classical currencies on automated exchange platforms, provides us with a unique benchmark for a comparison with high-speed markets. Indeed, the Bitcoin exchange is an example of a proto-typical financial market, maintained – at the time of the study – in a rudimentary competitive state by high trading fees [3] which inhibit the development of substantial market making or arbitrage.

3. 0.6% fees per transaction for the vast majority of the main exchange MtGox's users

An important aspect of a financial market is how it absorbs the new information conveyed by a trade or a sequence of trades into the market price Kyle (1985). Such incorporation of new information may not be instant due to market frictions Beja and Goldman (1980) ; nor is the submission of the trader's order, since order splitting and inventory considerations create a serial dependence of trades Lillo and Farmer (2004); Toth et al. (2015). During the execution of a large sequential order (meta-order), spread over a certain time period, the difference between the impacted price and the initial price can be quantified by measuring the market impact of the meta-order.

In this paper we undertake a thorough empirical analysis of market impact on the Bitcoin/USD exchange market. Previous studies Almgren et al. (2005); Moro et al. (2009); Tóth et al. (2011); Bershova and Rakhlin (2013); Gomes and Waelbroeck (2015); Mastromatteo et al. (2014a) found that average market impact of a meta-order of total volume Q approximately follows the "square-root law",

$$I(Q) \approx \pm Y \sigma \left(\frac{Q}{V_D} \right)^{\delta} , \tag{5.1}$$

where $I(Q)$ quantifies the average difference between the impacted price and the initial price (with the positive sign corresponding to buy orders and vice versa), V_D and σ the daily traded volume and daily volatility of the stock. The exponent δ has been consistently found to be approximately $1/2$. When the meta-order terminates so the pressure it was exerting on the price stops, the price is usually observed to revert (partially or totally) towards the initial (unimpacted) price Brokmann et al. (2015); Gomes and Waelbroeck (2015). With our high-quality dataset we shall confirm that above impact law holds for the Bitcoin/USD exchange market. Therefore, the rise of algorithmic trading may not have had as a profound effect on the functioning of markets as is often advocated. On the Bitcoin, both the response of the market to trades and the serial dependence of orders are similar to what is observed on mature liquid financial markets, in such a way that market impact can be specified by the same empirical law.

This observation is central : The precise mechanism which is behind the peculiar square-root law appears to be universal across markets which significantly differ with respect to their trade characteristics (latency, daily traded volume, volatility), microstructural parameters (tick or lot sizes) and fee structure (i.e. the high fees on the Bitcoin). This latter point is especially important. Statistical arbitrage is not profitable on the Bitcoin on scales below the fee level of 60 bps (120 bps for a round-trip !) and price efficiency is therefore not ensured in this region. The resulting large bid-ask spread frequently exceeds the peak impact of typical meta-orders ; concepts such as price efficiency and arbitrage are hence not suitable to explain market impact on the Bitcoin exchange.

Therefore, we conclude that competitive equilibria between different agents, for instance amongst market makers or between informed liquidity takers and uninformed liquidity providers Glosten and Milgrom (1985), should not be used as the fundamental starting point for the theoretical modelling of market impact on the Bitcoin. Nor should one use the martingale conditions for the market price,

as the notion of "price" is not precisely defined under the scale of 60 bps (∼spread/fees). In fact, as we shall see in the following, impact describes *how trades dig into the opposite (supply/demand) side* – before some post-trade mean-reversion occurs on this side – rather than how the *market price* itself is affected. Although these two definitions are equivalent when the spread is tight, they do not coincide in the case of the Bitcoin. The fact that very similar impact laws are found on mature financial markets suggests that this observation holds for these markets, as well.

5.2.1 Comparison with related literature

The square-root impact formula quantifies how market prices are affected by trades. This has implications for *market stability* : Price impact may lead to unexpected price swings Lehalle and Lasnier (2012) and stock crashes Kyle and Obizhaeva (2012) as well as to well-known phenomena such as stock pinning Avellaneda et al. (2012). Second, the control of *execution costs*, i.e. market impact, is of great practical interest to financial institutions. Almgren and Chriss (2001), for instance, determined optimal execution strategies based on the assumption that impact is *linear* (i.e. that $\delta = 1$). This is the simplest theoretical setting that excludes round-trip execution strategies (zero terminal inventory) with negative execution costs (profit making). Since then, a huge effort has been made to model non-linear market impact, motivated by a flurry of empirical studies. These approaches can be broadly split into three categories :

— Concave market impact of metaorders can be reproduced by *propagator models*, in which each trade is assumed to have a transient impact which decreases according to some time-dependent kernel. Summing up the impacts of individual trades leads to the impact function of the metaorder which is then found to be concave. The possible absence of dynamical arbitrage in these settings Gatheral (2010) allows one to solve convex optimization problems for optimal execution and find optimal liquidation strategies Alfonsi and Schied (2013) within this framework. While these models yield fairly realistic results and are analytically tractable, they are however purely phenomenological and do not provide a mechanism to explain impact.

— *Equilibrium models* assume a competitive equilibrium between liquidity providers and takers. They can be regarded as an extension of the original argument in Glosten and Milgrom (1985) from a single market order to a sequence of trades. Equilibrium models typically use two constraints to fix the bid and the ask during the metaorder execution, usually using the martingale condition and a subtle fair pricing argument Farmer et al. (2013); Donier (2012), which states that the transient impact of a metaorder anticipates its permanent impact on the price so that neither the informed trader nor the market maker should expect profits for any metaorder (on average). In the same vein, Jaisson (2015) shows how non-linearities can emerge from anticipations when the order flow is correlated by a Hawkes process, even though impact is linear *ex post*. These arguments, together with strong correlations in the

order flow finally leads to an – asymptotically – concave metaorder impact. However, as such mechanisms are rather unlikely to be in force on the Bitcoin market, they seem not appropriate to explain the seemingly universal shape of impact.

— The third class of models, initiated in a different context in Bak et al. (1997), are *statistical models* of supply and demand Tóth et al. (2011); Mastromatteo et al. (2014a,b); Donier et al. (2015), that may also be seen as *heterogeneous agents models* Donier and Bouchaud (2015b). The dynamics they give to the supply and demand that underlie the order book is such that both generically increase quadratically with distance from the mid-price, in turn leading to an exact and universal square-root impact at all scales.

Empirical studies have been mostly concerned with the *peak impact* of meta-orders, i.e. the impact measured between the extremal points of the execution path. Further studies have shown that impact is to a large part transient : after execution the price falls from its peak to some intermediate level, that some studies Moro et al. (2009); Bershova and Rakhlin (2013); Gomes and Waelbroeck (2015) argue to be close to 2/3 of the peak impact, in agreement with equilibrium models Farmer et al. (2013). In Brokmann et al. (2015) the authors find that this high permanent level may be due to the correlation between the trader's execution decision and the residual order flow [4] and to the price signal that triggered the decision to trade : By taking into account these effects they argue that the "bare" permanent impact (or mechanical permanent impact) is much lower and possibly even zero. Finally, Gomes and Waelbroeck (2015) conduct a separate impact study of informed trades and cash-flow (uninformed) trades, to find that the latter have no permanent effect on the price (although the transient impact are similar for both types of trades). Hitherto, these results are not well accounted for nor understood.

Amongst the empirical studies of market impact in the past literature, the vast majority has relied on partial datasets. Usually, market participants have only access to their own proprietary data Tóth et al. (2011); Bershova and Rakhlin (2013); Brokmann et al. (2015) which leads to an unavoidable conditioning to their trading strategies (even though in some cases many different strategies are collated together so part of the conditioning may average out). Two notable exceptions do however exist : In Moro et al. (2009) hidden metaorders are directly inferred from brokerage codes, while Zarinelli et al. (2015) have unprecedented access to the start times, end times and volumes of a huge amount of metaorders stemming from Ancerno's clients [5].

In this paper, we use a dataset which allows on the contrary to identify *each* trade with a *unique* trader, thereby leading to a *complete picture of the market*. Our dataset is large enough to study market impact as a function of volume, participation rate and even as a function of the behaviour

4. After a buy (resp. sell) metaorder, the order flow is on average biased towards buy orders (resp. sell orders). This correlation with future order flow artificially inflates its measured impact.

5. This study reports a log impact as a better overall fit to metaorder impact, although they actually mention some caveats due to the low level of details of their dataset.

of the residual market. This allows us to retrieve pseudo-random metaorders, i.e. metaorders that are uncorrelated from the residual order flow, either because they do not convey any information or because the information that triggered them is not shared by the residual market.

5.2.2 Brief outline

After a presentation of the dataset (Sec. 5.3), we introduce the main definitions and methodology (Sec. 5.4) with a focus on the methods we use to identify distinct meta-orders. The price impact of metaorders is introduced in Sec. 5.5. The first major result of our study is that, despite its prototypical micro-structure, the square root law holds on the Bitcoin. Moreover, in Sec. 5.6 we show that not only the peak impact of metaorders, but rather the *whole impact trajectory* follows a square-root. We show unprecedented pictures of the mid-price, bid and ask trajectories during and after impact and discuss permanent and transient market impact as a function of the execution speed and its dependence on the residual order flow.

We are able to differentiate the mata-orders with respect to their correlation with the residual order flow. The permanent impact is identified as the information content of meta-orders : Pseudo-random metaorders (i.e. metaorders uncorrelated with the residual order flow) have little permanent impact, in agreement with previous studies Brokmann et al. (2015); Gomes and Waelbroeck (2015).

Finally, in Sec. 5.7 we summarize our main findings and in the last section we conclude and discuss the implications of our empirical findings for the most common impact models so far proposed in the literature.

5.3 Bitcoin market at a glance and data

5.3.1 Bitcoin : a prototypical market

Bitcoin is a crypto-currency introduced by an anonymous programmer in 2008 Nakamoto (2008), designed to allow exchanges of money without the need of a central authority (e.g. state or bank) to enforce trust. The money is issued progressively through a process called *mining* – by analogy to gold mining – to the people who give their computing power to help build the transactions ledger (the *blockchain*). As a currency, it has some intrinsic value (even though still quite imprecise) and can be exchanged against usual currencies on organized markets. The Bitcoin market started in 2008 and literally exploded in 2013 with a peak market capitalization above 10B dollars, a daily number of transactions over 100k and a daily traded volume above 100M dollars at the end of 2013.

Similarly to financial markets, trading takes place on an order book, where Bitcoins can be exchanged against other currencies. Notably, the role of the order book is quite important since a large fraction of the liquidity is not hidden, but actually posted in the order book. A more quantitative analysis indeed shows that typically $30 - 40\%$ of the volume traded during the day is

already present in the order book in the morning. This is to be compared with a ratio below 1% on more traditional financial markets, say stocks Wyart et al. (2008). The tick size, i.e. the minimal price increment between two consecutive prices at which it can be exchanged, is of USD 10^{-5} on MtGox so that the order book can be considered as a continuous price grid.

Bitcoin market microstructure is quite unique for several more fundamental reasons. First, because of the very high level of fees (compared to other markets) of 0.6% per transaction on MtGox, resulting in an average spread of $\approx$ 0.6%. This, to a large extent, hobbles high frequency arbitrage/market making strategies on such a market, who only take part in a few percent of the transactions : the Bitcoin market is essentially a market between end users. Second, all relevant information is concentrated on one asset and one exchange (MtGox) on the time span considered, with very little notion of a "fundamental price". This is a very unique example of a single-asset economy, with little correlations with any other asset on the planet (for the time being).

5.3.2 Data

This study was realized using a database of all 13M trades trades that happened on MtGox Bitcoin-USD exchange between August 2011 and November 2013, in which traders are uniquely identified [6]. Thus, we have a complete description of the market as a whole, in contrast to most hedge funds' proprietary data (reflecting their trading decision) which does not allow for a global analysis of trading decisions and market impact. As a consequence, we do not face the problem of the potentially strong conditioning of impact paths on the particular strategy of the agent. We rather have an overall insight on how agent's decisions entangle. Since there is very little brokerage intermediation in this market, we have direct knowledge of the actual initiators of each trade (only as an anonymized numerical ID code). One can thus assume that with very good approximation *all metaorders can be fully identified*. This is a most valuable property for studying impact, since it allows for the explanatory variable – the metaorder – to be fully characterized. This data quality is probably very difficult to obtain on other financial markets, because of brokerage intermediation, multiplicity of venues, and multiplicity of correlated instruments on which a trade can be executed. In addition, our dataset is the largest so far with such precision, and with full knowledge of the trajectories for the 1M identified metaorders.

6. The data we use was provided to us by an anonymous source, and is available upon request under certain conditions. The public part of this data file has been checked to match otherwise known price and volume data at www.bitcoincharts.com, and the private part to match all private trading data that are to the knowledge of the authors. Apart from these consistency checks, no insurance is given that all data is accurate. However the high precision of our findings suggests that they are.

5.4 Definitions and methodology

5.4.1 Definitions

When a trader (she) wishes to bring an excess of supply or demand Q on the market, she is confronted to the question of how to achieve it in a reasonable way. If the volume she wants to sell or buy is small, then she will probably do it all at once if enough liquidity is available on the order book. However, if she wants to invest or sell back a quantity that exceeds the available offer or demand (think of large speculators or professional service providers), an instantaneous execution might destabilize the market and incur her larger costs than planned : she therefore has to split her large order into chunks, so that the imbalance can be slowly digested by the market Bouchaud et al. (2009). We refer to this *total* quantity Q as a *metaorder*, denoting T its duration and $\mu := Q/T$ its execution speed, and study the quantities

— $\mathcal{I}_{\text{path}}(r, Q, \mu)$, defined as the impact on the price of the first $r\%$ of a metaorder of size Q and execution speed μ. By definition $r \in [0, 1]$ is the part of the volume already executed, and $\mathcal{I}_{\text{path}}(0, Q, \mu) = 0$ is the initial price (gauged to zero) ;

— $\mathcal{I}(Q, \mu) := \mathcal{I}_{\text{path}}(1, Q, \mu)$ is the price at the end of the metaorder, referred to as the *peak impact* of the metaorder ;

— $\mathcal{I}_{\text{exec}}(Q, \mu) := \int_0^1 \mathcal{I}_{\text{path}}(r, Q, \mu)\mathrm{d}r$ is the average execution price ;

— $\mathcal{I}^{\infty}(Q, \mu)$ is the average price long after the end of the metaorder and can be decomposed into a predictable part $\mathcal{I}^{\infty}_{\text{info}}(Q, \mu)$ and the response to the metaorder in question $\mathcal{I}^{\infty}_{\text{mec}}(Q, \mu)$. The latter is the most relevant quantity regarding price formation, since the former represents *alpha* biases that has *a priori* no link with the market's mechanical reaction to the order.

Note that all these quantities are implicitly averaged over all residual noise or variables. The effect of such metaorders on the price is qualitatively well-known. While the metaorder is executed, the pressure exerted on the price tends to make it rise so $\mathcal{I}_{\text{path}}(r, Q, \mu)$ is an increasing function of r. When the metaorder is completed the price reverts to the *permanent* level $\mathcal{I}^{\infty}(Q, \mu) \leq \mathcal{I}(Q, \mu)$. Previous empirical studies have found the *peak* impact to be approximately square root of the volume and the amplitude of the *post-impact decay* to be about 1/3 on average (so that $\mathcal{I}^{\infty}(Q, \mu) \approx \mathcal{I}_{\text{exec}}(Q, \mu) \approx \frac{2}{3}\mathcal{I}(Q, \mu)$ for square root impact, as required by equilibrium models since this ensures that the price paid for the execution is fair *ex-post*). More recent studies Brokmann et al. (2015); Gomes and Waelbroeck (2015) have shown that when subtracting the *predictable part* $\mathcal{I}^{\infty}_{\text{info}}$ from the price, the reversion goes all the way back to zero – so after waiting a long enough time the *mechanical impact* $\mathcal{I}^{\infty}_{\text{mec}}(Q, \mu)$ of trades on the price is zero.

One can also define the *execution rate* μ_V, defined as the ratio between the volume of the metaorder Q and the total volume traded by the market during the same period, V_M,

$$|Q| = \mu_V V_M \ . \tag{5.2}$$

We will thus study the peak impact $\mathcal{I}(Q,\mu)$, the average executed price $\mathcal{I}_{\text{exec}}(Q,\mu)$ and the permanent impact $\mathcal{I}^{\infty}(Q,\mu)$, in order to identify any dependence on T (or equivalently μ) that goes beyond the usual rule of thumb (9.4) which predicts impact to be independent of the execution speed. Throughout this study, price is taken to be the *traded price*. Since we only consider aggressive metaorders (cf. Sec. 5.4.2), this amounts to studying the ask for buy metaorders and the bid for sell metaorders.

5.4.2 Metaorder decomposition and properties

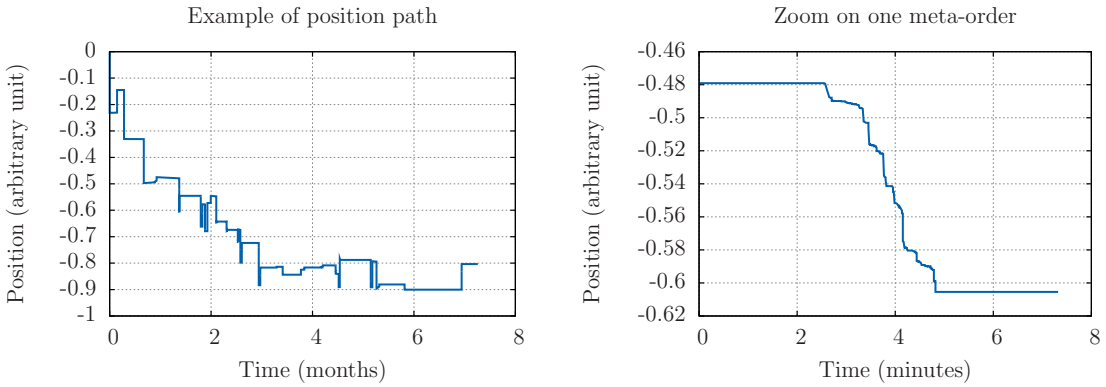

FIGURE 5.1 – Typical position vs time for one of the 1% most active traders in terms of volume. *(left)* position path during approximatively 12 months. One can see that metaorders are clearly identifiable and alternate with long periods of inactivity. *(right)* Zoom on a 2 minute sell metaorder composed by a dozen trades (zoom 1 : 50000).

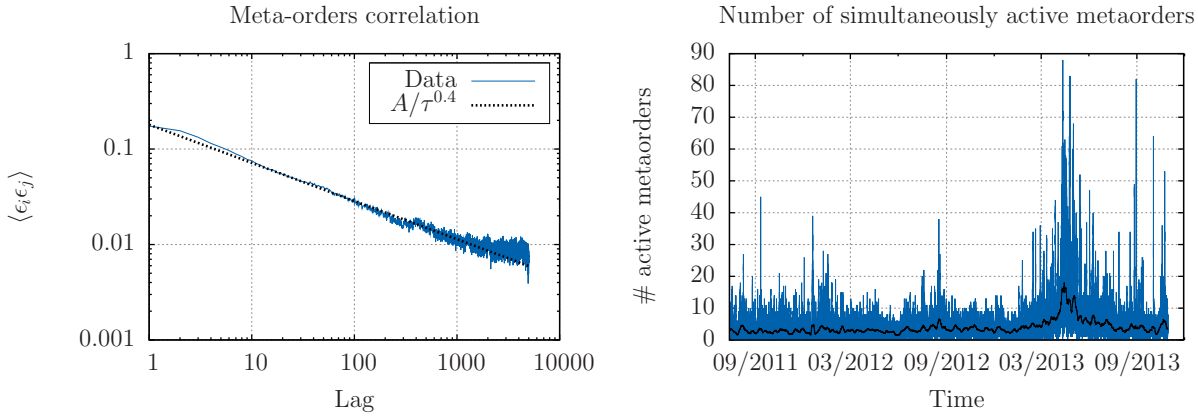

FIGURE 5.2 – *(left)* Autocorrelation function of the sign of *metaorders*, ordered by starting time. Like financial markets, not only trades but metaorders are strongly auto-correlated. *(right)* Number of simultaneously active metaorders vs time. The typical value is around 5, and we can clearly observe clustering.

The first operation needed in order to study market impact is to spot the metaorders : due to the extreme irregularity and heterogeneity between the traders' typical position paths, usual time series decomposition methods Toth et al. (2010) are not relevant here. In order to identify

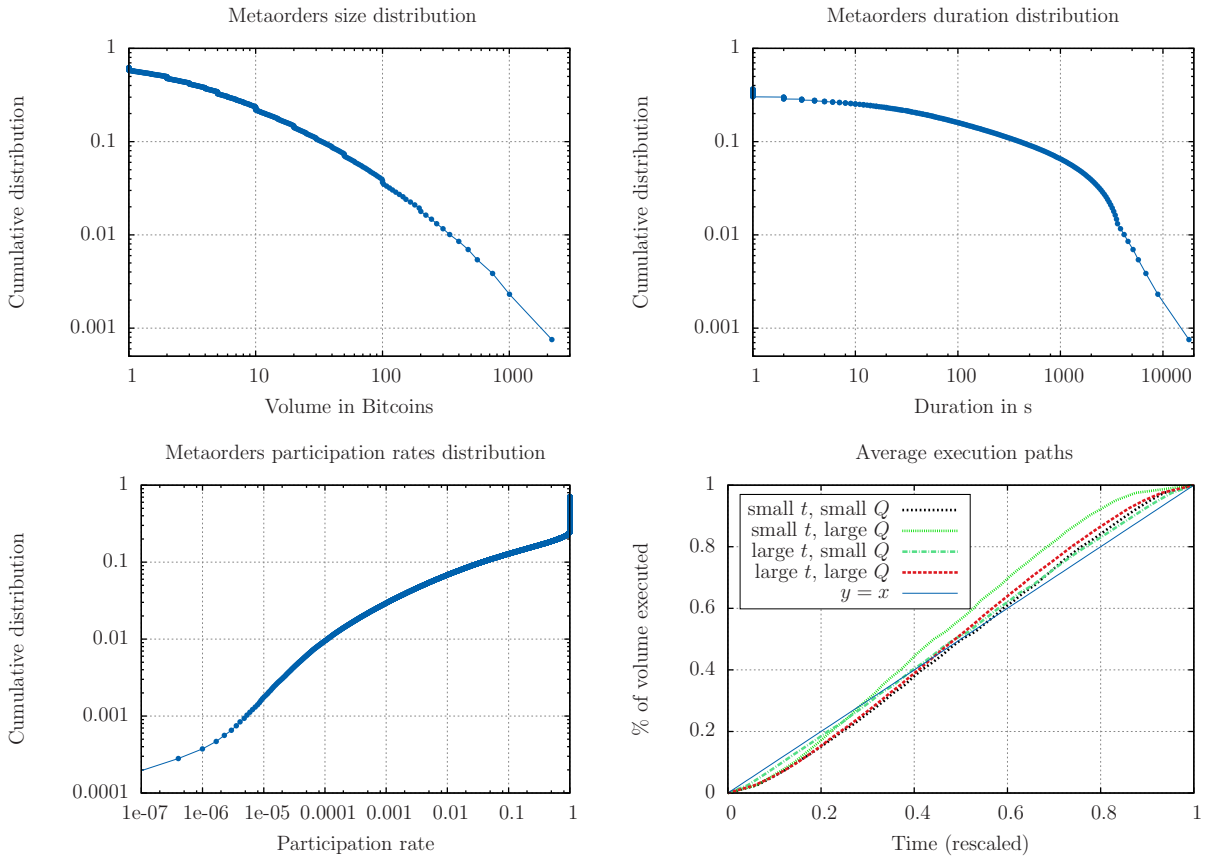

FIGURE 5.3 – *(top and bottom left)* Metaorder size, duration and participation rate distributions for the whole market, with no clear power-law fit for any. Note that the durations are much shorter than usual metaorder durations on financial markets. *(bottom right)* Percentage of volume executed vs time elapsed since the start of the metaorder, which appears to grow roughly linearly for all ranges of volume and duration (here the volume threshold between small and large has been fixed at 200BTC and the time threshold at 100s).

these large buy and sell metaorders for this particular data, in such a way that no conditioning in the start/end sequences of the metaorders is introduced (most intuitive techniques may create mean-reversion biases), we used the following : For each trader, we defined the start of a metaorder to coincide with the first aggressive [7] order placed after a given period of inactivity [8]. We define the end point of the metaorder either as the last order before a (new) inactivity period, or as the point where the trader trades in the opposite direction. While the introduction of a time scale to define inactivity periods could seem to be arbitrary, it is the case in practice that the metaorders are so clearly distinguishable that the sensitivity to this time scale is minimal (Fig. 5.1) : for one particular trader, metaorders are very clearly separated from each other. This way, from over 14

7. We only consider aggressive orders since they are a much better proxy for system perturbations. Limit orders in the contrary may have been posted long in advance without a specific intention nor view on the price : By ignoring them we therefore limit any adverse selection that might bias our study. Besides, the fact that the execution schedule is roughly linear strongly suggests that we do not bias the results by making this choice.

8. The relevant scale for this inactivity period has been empirically determined to be about one hour.

# of child trades	1	$2 \leq n \leq 4$	$5 \leq n \leq 9$	$10 \leq n$
% of metaorders	61%	29%	6.5%	3.5%

TABLE 5.1 – Number of child orders per metaorder.

million trades we recover over 1 million metaorders of variable sizes/durations. A few percent of the trades are not assigned to a metaorder, corresponding to mean-reverting trades which by definition have a conditioning bias. Table 5.4.2 presents the repartition of metaorders in terms of number of child orders. As expected, more than half are one-shot trades whereas 10% percent are composed of more than 5 trades. However, as a result of the tick being small, more than 20% of the one-shot metaorders cross several price levels.

Fig. 5.2 shows the autocorrelation function of metaorder signs which is slowly decaying with a power-law exponent around 0.4 – which happens to be similar to the autocorrelation exponent of trades themselves – as well as the number of metaorders that are simultaneously active in the market at any point in time, which present clear clustering with a typical value around 5. The salient statistical characteristics of metaorders are presented in Fig. 5.3. It presents the metaorder sizes and durations distributions, which are crucial ingredients in many market impact and price models Farmer et al. (2013); Donier (2012); Gabaix et al. (2003). Contrary to what is usually required in such models, none of these distributions are clear power-laws – and particularly not on *all* time scales [9]. This challenges such impact models, since in these models the impact function is closely related to metaorders distributions – and a power-law with exponent 3/2 is required to reproduce the square-root impact. This is a strong hint that the distribution of order sizes is not a fundamental input to explain the shape of market impact, since it is neither clearly a power-law nor universal. Most importantly, above figure shows that the average execution profile is linear in time. This property ensures that μ is well defined in the sense that it is on average constant during execution (one also needs to check that the executed volume in the market is constant, which is the case as shown e.g. by Fig. 5.4). This will allow us to properly compare points within impact trajectories, $\mathcal{I}_{\text{path}}(r, Q, \mu)$, with peak impacts $\mathcal{I}(rQ, \mu)$.

5.4.3 Conditioning checks

It has been shown in the previous section that average execution style is linear in time, so that we can really study the effect on the price of metaorders with constant execution speed on average [10]. In this section, we proceed to some other checks concerning the behaviour of the rest of the market while metaorders are being executed. Indeed, the usual argument against the square-root law assumes them to be conditioned to the rest of the market in such a way, that the apparent

9. If however one estimates a tail exponent using the Hill estimator Hill (1975) for volumes greater than 10BTC, one finds a tail exponent of -2 for the probability density function.

10. as opposed e.g. to Zarinelli et al. (2015) in which they find strongly front-loaded executions.

concavity of the marginal impact is only an artefact. In particular, anticipating Sec. 5.6, we look into the buy/sell market order flow while a metaorder is executed, to check whether dynamical effects exist in order flow that could result in concavity : for example, a higher correlation with market imbalance at the beginning of the metaorder would result in sharper price changes at the start of the trajectories[11]. Since we have the start and end times of every metaorders as well as their execution rates, we can compute at any point in time the number/volume of active buy/sell metaorders in the market. In Fig. 5.4 we compute these quantities during the execution of chosen metaorders (here all metaorders of duration about 2 minutes) and plot the average paths over all these metaorders to look for such dynamical effects. The data shows that such a synchronisation of

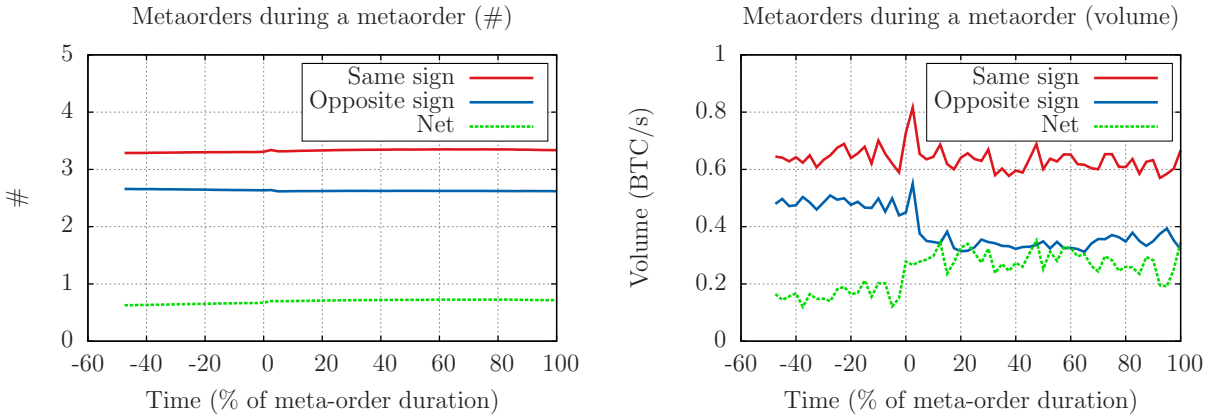

FIGURE 5.4 – Here we selected all metaorders with duration of approx. 2 minutes. *(left)* We plotted the number of active metaorders in the same direction than the metaorder considered after subtracting it (green), in the opposite direction (red) and the difference (blue). *(right)* The same plot, in terms of volume. On both plots, time is normalized so that the metaorders start at 0 and end at 100.

metaorders is almost non-existent, either in terms of number of metaorders or in terms of volume, so that concavity of impact is not generated by some kind of synchronization between the agents[12]. Since the study of impact paths in Sec. 5.6 will show that the square-root law holds for the whole *trajectories* and not only for the peak impact, one can assert that concavity is *true* in the sense that it does not stem from conditioning but is really the way the market absorbs an excess of supply or demand.

11. This would imply some sort of synchronization between the agents, either exogenous (e.g. the agents react to the same news thus starting their metaorders at the same time, but some stopping before the others) or endogenous (e.g. arbitrageurs are able to detect metaorders and push the price up at its beginning to sell back later when the price has risen).

12. One can note the interesting (unrelated) fact that market order volume in the opposite direction slightly decreases while a metaorder is being executed. Also, during a buy (resp. sell) metaorder the other metaorders present on the market tend to go in the same direction. This confirms the fact that metaorders are "informed" on average. A more in-depth discussion on this notion of information can be found in Donier and Bouchaud (2015b).

5.5 The square root impact law on the Bitcoin/USD market

5.5.1 The square root law of peak impact for individual metaorders

For each of the 1M metaorders we identified above, we considered the impact defined as

$$\mathcal{I}(Q,\mu) = \mathcal{I}_{\text{path}}(r=1,Q,\mu), \tag{5.3}$$

i.e. the difference between the first and the last executed price, quantifying the reaction of the market to the trader's order. Note that here impact is measured as the peak price (with the initial price gauged to zero). The result is shown in Fig. 5.5 : In spite of the very special features of the Bitcoin market, a concave impact law (depicted as a straight line)

$$\langle \mathcal{I}(Q,\mu)\rangle_\mu \approx \widetilde{Y} Q^\delta \ , \tag{5.4}$$

fits the data points very well from the smallest scales and over 4 decades with $\delta \approx 0.5$ and $\widetilde{Y} \approx 4.5 \cdot 10^{-2}$. Normalizing by Bitcoin average volatility and daily volume gives a *Y-ratio* (as defined in Eq. 9.4) of $Y \approx 0.9$, close to the value reported on "mature" financial markets, e.g. futures or stocks Tóth et al. (2011); Brokmann et al. (2015). For a more in-depth study of the *Y-ratio*, see Sec. 5.5.3 below. Thus, peak impact for the Bitcoin is consistent with the square-root law despite the

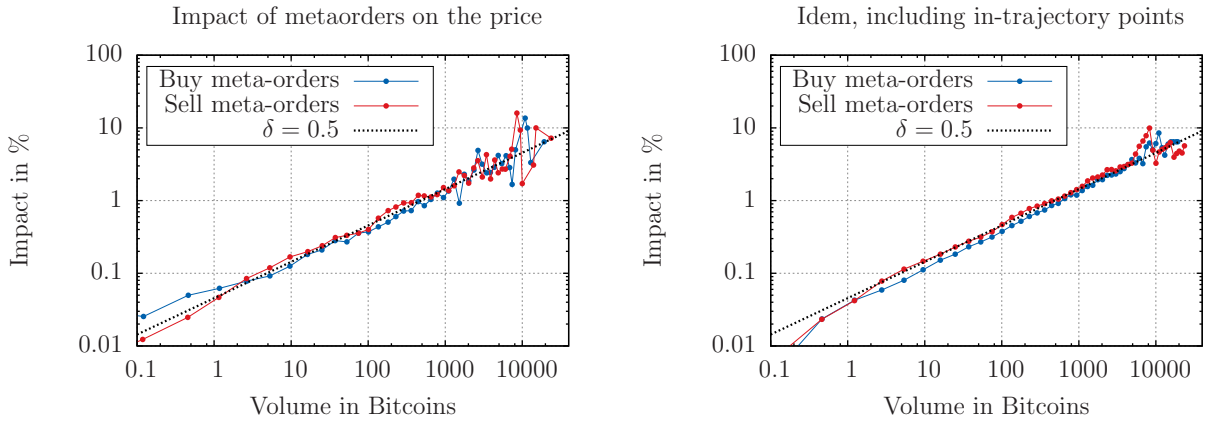

FIGURE 5.5 – Market impact $\langle \mathcal{I}(Q,\mu)\rangle_\mu$ (averaged over all execution rates μ), follows the same square root law as is observed by banks and hedge funds on financial markets (plots in log-log scale). Each point represents the average impact of all metaorders in a given range of volume. The impact exponent δ is found to be very close to 0.5, and the Y-ratio is around 0.9. One should emphasize that this power-law behaviour appears at the smallest scales and holds over 4 decades. *(left)* Only end points of metaorders (∼1M data). *(right)* 41 point per metaorder (every 2.5% quantile of volume), giving 27M data points. Part of these points being degenerate, one can assess the number of effective points around a few millions.

prototypical nature of the Bitcoin market. This confirms the view expressed in Tóth et al. (2011) and Mastromatteo et al. (2014a) that market impact depends neither on microstructure (with e.g. arbitrage and high-frequency trading), nor on a clear metaorder distribution. As such, it directly

challenges the explanation of concave market impact in terms of rational equilibrium theories.

5.5.2 Square root impact trajectories

We now turn to the study of impact trajectories, i.e. the quantity $\mathcal{I}_{\text{path}}(r, Q, \mu)$ for given Q, μ and r varying from 0 to 1. Unless mentioned otherwise, we average over all other quantities in the following (like e.g. daily volatility, daily traded volume etc.). Fig. 5.6 shows the results, putting into light two facts of particular interest. The first is the answer to question whether the impact trajectories follows the same law as peak impacts. On the Bitcoin, where execution paths are roughly executed linearly in time, the agreement is remarkable, meaning that while the metaorder is not finished the market makes no difference between a metaorder that will stop soon and a metaorder that will continue [13]. We find empirically that

$$\mathcal{I}_{\text{path}}(r, Q, \mu) = \mathcal{I}(rQ, \mu) \ . \tag{5.5}$$

Hence, if t is the time elapsed since the start of the execution, and since μ is relatively constant for the metaorders in our dataset, one can write impact as

$$\mathcal{I}_{\text{path}}(r, Q, \mu) = \mathcal{I}(t, \mu) = f(\mu)t^{\delta} (:= \widetilde{f}(\mu)(\mu t)^{\delta}) \tag{5.6}$$

where $\delta \approx 1/2$. This in particular allows one to include in-trajectory points in the measurement of peak impact, cf. Fig. 5.5 (right) as they would have been the peak impact of the metaorder, had it stopped before. Note that in the above equation we implicitly average over any other variable so that f only depends on μ (with an assumption of independence).

Second, one observes that the value of the $\widetilde{Y}$-factor appears to be noisy as impact lines do not superimpose, but are all perfectly parallel, across different Q's. A possible cause is the time variations of $\widetilde{Y}$ during the period considered (see Fig. 5.6), leading to a conditioning effect as soon as the distributions of Q is not similar during high- and low-$\widetilde{Y}$ periods. This suggests that the price during a metaorder execution can be written as

$$I(t, \mu) = \widetilde{Y}(t)\sqrt{\mu t} + \sigma W_t \ , \tag{5.7}$$

where $\widetilde{Y}(t)$ accounts for a (slowly) time-varying liquidity [14] and σ represent some additional market noise.

13. This is at variance with Zarinelli et al. (2015) where average executions are not linear in time but rather front-loaded execution, resulting in very particular price trajectories.

14. By writing this integrated form one implicitly assumes that $\widetilde{Y}(t)$ is constant throughout the execution and varies at scales that are slower.

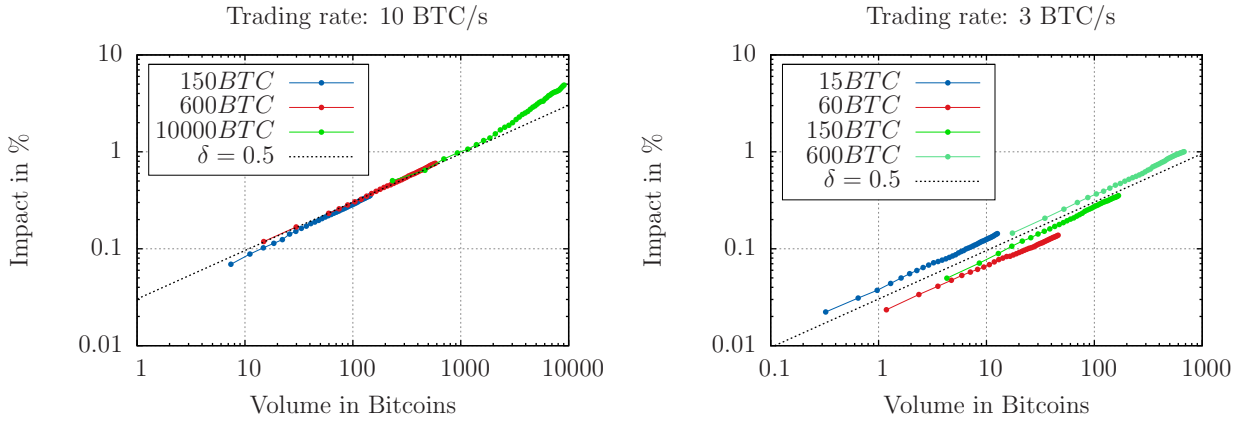

FIGURE 5.6 – Impact paths $\mathcal{I}^{\text{path}}(p, Q, \mu)$ in decimal loglog plot, for different metaorder volumes (cf. legends), for *(left)* $\mu = 10BTC/s$ and *(right)* $\mu = 3BTC/s$, and for $r \in [0, 1]$ for each couple (Q, μ). The first value has intentionally been chosen high, so that it survives the criticism raised in Section 5.6.2 that on average other metaorders in the same direction are observed, which tends to artificially increase impact measures. One can observe a liquidity breakdown leading to an asymptotically linear impact when important pressures are maintained for too long on the same side of the order book.

5.5.3 The "Y-ratio"

Until now, much emphasis has been put in the literature on the study of the dependence in Q but very few studies have been realized on the pre-factor $\widetilde{Y}$ – or equivalently the *Y-ratio* defined in Eq. 9.4. We devote this section to a temporal analysis of both these pre-factors. For each metaorder i, we compute its individual $\widetilde{Y}_i$ as the ratio $\mathcal{I}(Q_i)/\sqrt{|\ Q_i\ |} \times \text{sign}(Q_i)$. For each day, we compute the daily average $\widetilde{Y}$ as a volume-weighted average of all individual ratios. In parallel, we compute the daily realized volatility σ_D and traded volume V_D, and we compare the daily $\widetilde{Y}$'s to the corresponding $\sigma_D/\sqrt{V_D}$ ratios for every day in our dataset. For each day, we also plot the actual *Y-ratio*. Results are presented on Fig. 5.7 and show that after such rescaling, the *Y-ratio* becomes nearly time independent. Its distribution is plotted on Fig. 5.8 and is well approximated by a Gaussian distribution $\mathcal{N}(Y_0, \Sigma_Y)$ with mean $Y_0 = 0.9$ and standard deviation $\Sigma_Y = 0.35$.

From these results, we can draw two conclusions of particular interest. First, it validates the scaling form of Eq. 9.4 proposed in Torre and Ferrari (1997); Grinold and Kahn (2000); Tóth et al. (2011). Indeed, the non-stationariness of the impact pre-factor $\widetilde{Y}$ is well encoded in the ratio $\sigma_D\sqrt{V_D}$. Besides, the residual a-dimensional *Y-ratio* is shown to be of order unity with a standard deviation of the same order of magnitude so that it essentially lies in the interval $[0, 2]$. In this light, Eqs. 9.4 and 5.7 can be merged together so that impact reads

$$I(t, \mu) = \sigma \left[(Y_0 + \sigma_Y \eta) \sqrt{\frac{\mu t}{V_D}} + W_t \right] , \tag{5.8}$$

where σ_Y accounts for some possible noise on the value of Y_0. The relationship between the *liquidity*

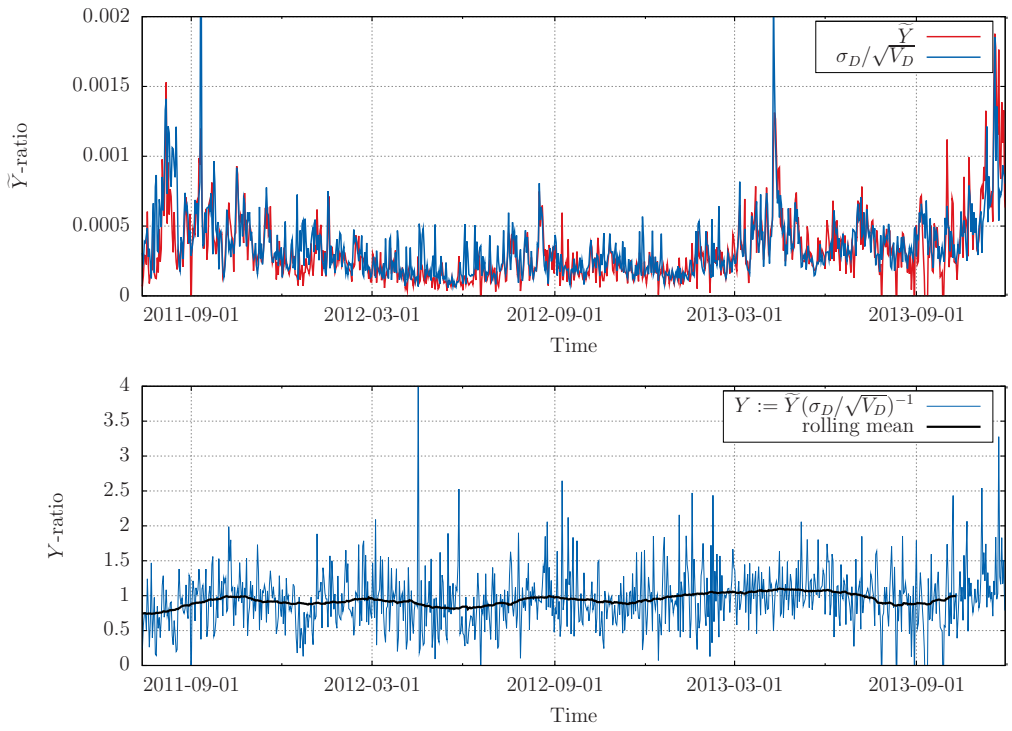

FIGURE 5.7 – *(top)* Raw impact pre-factor $\widetilde{Y}$ vs time. We also plot the usual normalization $\sigma_D\sqrt{V_D}$ to show that it accounts for the major part of the non-stationariness, particularly during extreme market events (e.g. April 10, 2013 major crash). *(bottom) Y-ratio* as defined in Eq. 9.4, which oscillates around its mean value $Y_0 \simeq 0.9$.

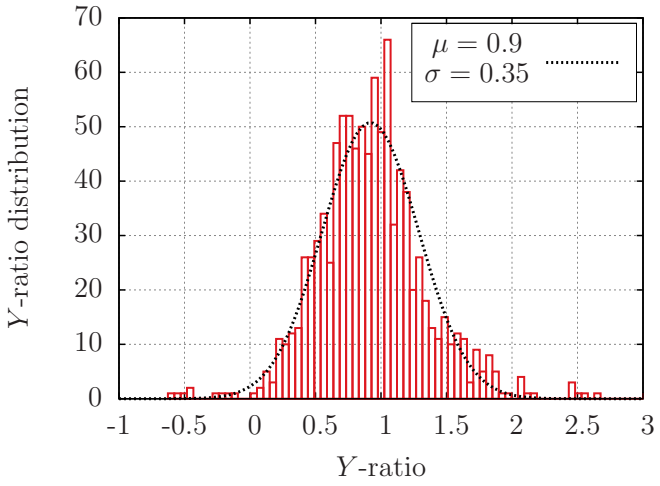

FIGURE 5.8 – *Y-ratio* distribution and its Gaussian approximation $\mathcal{N}(Y_0, \Sigma_Y)$ with mean $Y_0 = 0.9$ and $\Sigma_Y = 0.35$.

noise σ_Y and the *market noise* σ [15] – and their dynamics – have not been investigated here, although it is a topic of interest for further studies.

15. Both of which contribute to the Σ_Y evoked above.

The other conclusion is more macroscopic and relates to the study of market stability. Indeed, the scaling form for $\widetilde{Y}$ holds particularly well during extreme market events, such as the major crash that occurred on April 10, 2013, as evidenced in Donier and Bouchaud (2015a) who show that the impact pre-factor $\widetilde{Y}$ is a relevant proxy for market liquidity, even at scales much larger than that of impact. This definitely relates the microscopic aspect of price formation to its macroscopic characteristics such as its propensity to crash – see Kyle and Obizhaeva (2012) for similar discussions on financial markets. In this light, the understanding of how trades impact prices appears more crucial than ever to understand, detect – and perhaps even control ? – market instabilities.

5.6 Impact, execution speed and correlations with the order flow

5.6.1 Impact trajectories of the bid and the ask

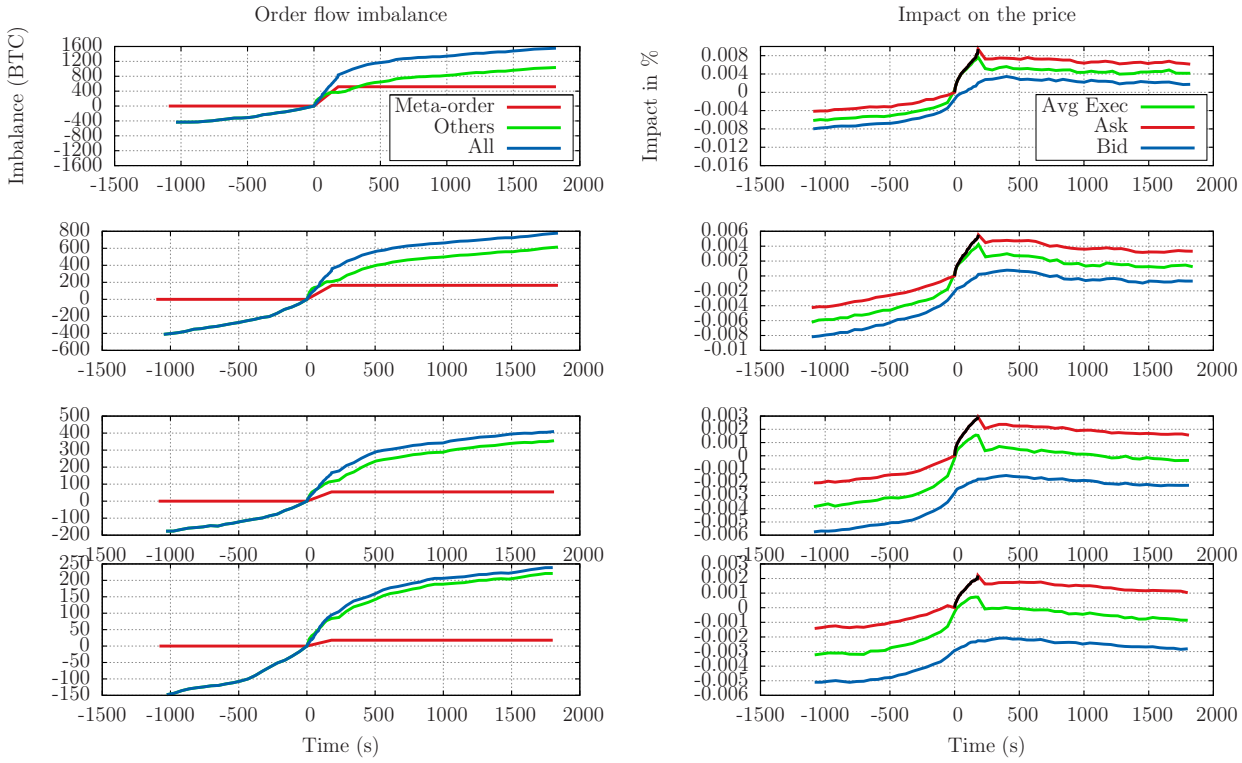

FIGURE 5.9 – (left) Order flow imbalances measured as the cumulated volumes of signed market orders during the execution of metaorders, for different metaorder volumes. For each panel, the metaorder starts at time 0 as shown by its executed volume in red. The blue line is the signed cumulated order flow of the whole market, whereas the green curve subtracts the metaorder from it to keep only the residual order flow. (right) Corresponding impact paths. In black, the actual execution path. In red, the ask path, and in light blue, the bid path. In green, we also plot the average execution price of the market. One can see that even for large volumes, impact is at most 1 or 2 spreads, not mentioning the fees that are of same order of magnitude than the spread : impact costs are dominated by friction costs, challenging martingale and fair pricing conditions.

The large fees on Bitcoin allow for a separate study of the bid and the ask during the execution and prevents the data to be too noisy due to high-frequency arbitrage. For the sake of simplicity, we assume that metaorders are buy orders, so that the ask denotes the opposite side of the order book, on which the trader executes his/her metaorder. The facts of interest are the following :

- Before the execution, the spread is roughly constant, and the execution direction is on average positively correlated with both the bid and the ask evolution before the execution (Fig. 5.9).
- During the execution, the ask rises sharply – following the same square root as the execution price – whereas the bid follows more linearly. This is very reminiscent of what is found theoretically in Donier et al. (2015).
- After this quick reversion, the bid and the ask remain roughly constant at a non-zero permanent level, in spite of the order flow pressure from the rest of the market that continues some time after the metaorder.
- These observations hold when we condition the metaorder to be trend-following or mean-reverting, as shown on Fig. 5.10. In particular, the fact that impact is still square root at small scales for trend-following metaorders is non-trivial and reveals that *trading speed matters*[16].
- In any case, impact does not exceed very much the order of magnitude of the spread and the fees, which questions any interpretation based on "fair-price" arguments, all the more so that impact is square root even at the smallest scales.

Consistent with the observations of Fig. 5.4, one observes that on average metaorders are positively correlated with the remaining order flow. We will see below that this plays a crucial role in the fact that the permanent component of impact is non-zero in these plots.

Finally, one notices that the market VWAP in green lies around the mid-price in the absence of the metaorder, and is naturally biased towards the execution side during the execution – all the more so that the execution is aggressive.

5.6.2 Impact and execution speed

The question of dependence of impact on the execution speed μ has been addressed very recently by Zarinelli et al. (2015) even though practitioners have been looking into it for at least a few years. In this section we study the *impact surface* described by the bivariate function $\mathcal{I}_{\text{exec}}(Q, \mu_V)$ [17] For

16. It contradicts in particular equilibrium models that only consider the aggregate order flow as a relevant quantity, since in these models *trading speed* is transparent.

17. We prefer here the average quantity $\mathcal{I}_{\text{exec}}(Q, \mu_V)$ to the peak impact $\mathcal{I}(Q, \mu_V)$ since latter is much noisier. The mean execution price is anyway far more relevant in practice since it is directly related to *execution costs*. We also preferred the execution rate μ_V to the execution speed μ as a fundamental variable since for this particular study this quantity is more relevant and yields cleaner pictures.

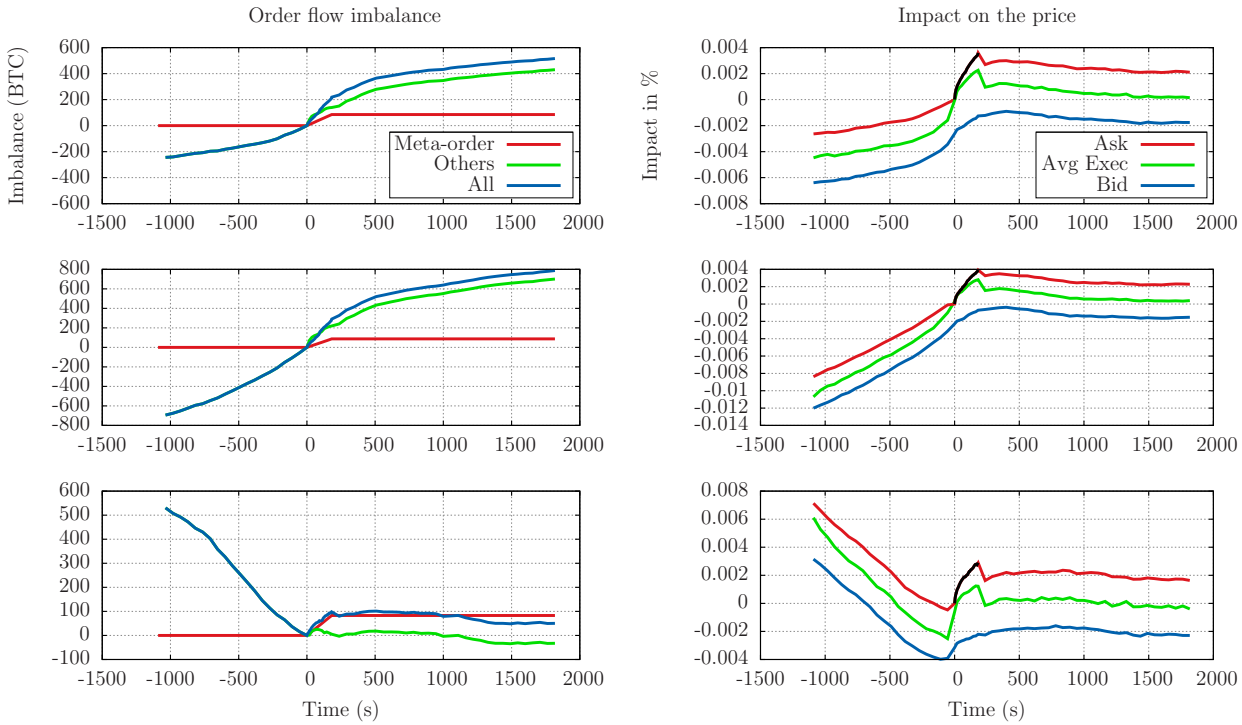

FIGURE 5.10 – Same pictures as in Fig. 5.9 for various types of conditioning *(top)* Unconditioned metaorders. *(middle)* Trending metaorders. *(bottom)* Mean-reverting metaorders.

each fixed μ_V the dependence on Q is found to be square root :

$$\langle \mathcal{I}^{\text{exec}}(Q, \mu_V) | \mu_V \rangle \sim \sqrt{Q} \,. \tag{5.9}$$

However, the dependence on μ is somewhat surprising. First, for very high participation rates (close to 100% of the volume during the period), the impact becomes unusually high. This is not surprising since the market breaks down in this regime [18]. More importantly however, while one could expect that impact monotonically decreases as the execution rate decreases, the observed impact actually *increases* again (cf. Fig. 5.11), at variance with intuition and previous findings on financial markets Zarinelli et al. (2015). The best fit to the empirical data reads $I^{\text{exec}}(Q, \mu_V) \sim Q^{\delta}/\mu_V^{\delta'}$, with $\delta \approx 0.5$ and $\delta' \approx 0.4$ (cf. Fig. 5.11). The slower the execution, the larger the measured impact : this strange dependence on μ_V stresses the difference between mechanical and informational impact. Clearly, a slow execution gives other market participants the opportunity to detect the same signal and the information content of the metaorder realizes during its execution. On the other hand, fast execution leads mostly to mechanical impact as the *alpha* realizes itself afterwards. Following this first non-intuitive discovery – which is very specific to the Bitcoin –, for each of the data points of

18. this fact is of importance since such an observation would probably be impossible on financial markets, where participation rates rarely exceed $20 - 30\%$ – for this very reason.

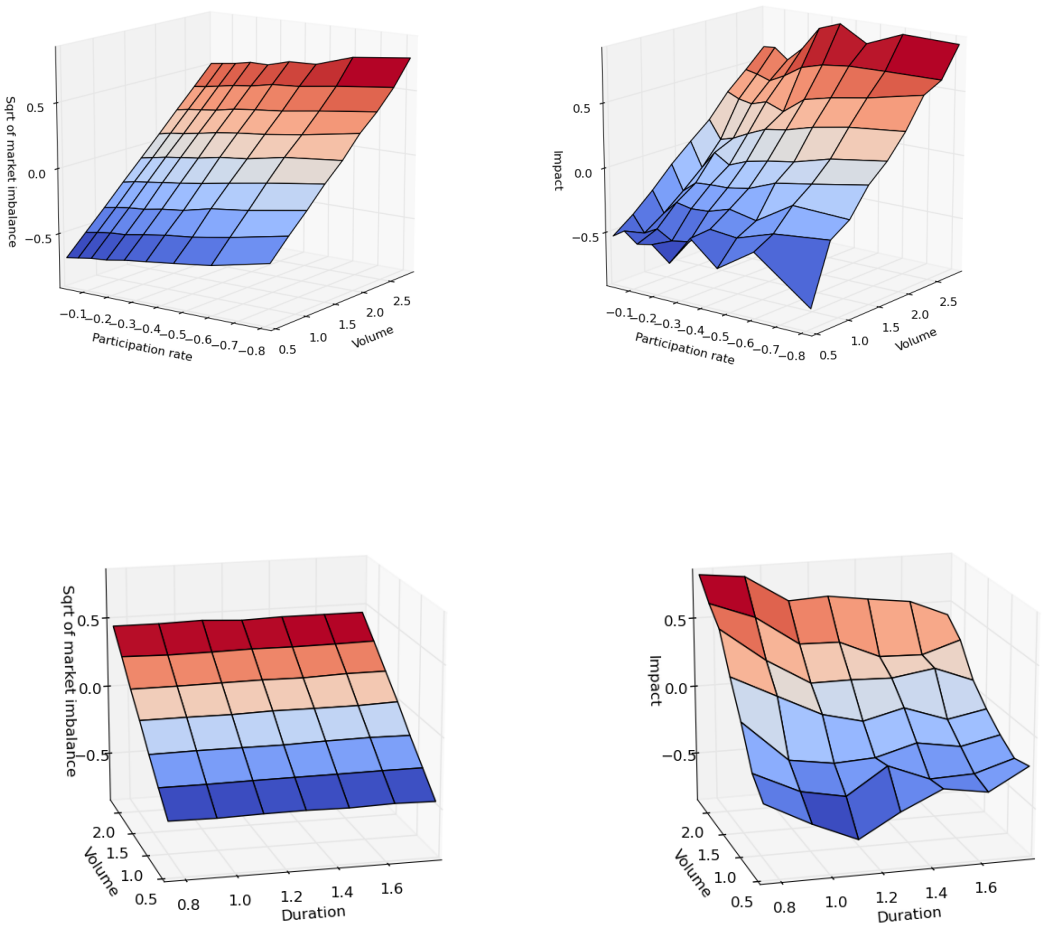

FIGURE 5.11 – The impact surface (in log-log-log scale) for typical metaorders (top) which are correlated with the total order flow. During the execution span of the metaorder, the market reacts rather to total market imbalance as to the individual market order : For all participation rates market impact is almost perfectly proportional to the square root of the total order flow as defined in the main text. Bottom images : Impact of an isolated metaorder (for which the residual market order flow remains neutral). The total market imbalance corresponds to the volume of the metaorder ; hence, the participation rate is not a suitable measure. Rather, the execution time should be regarded. The figure clearly displays an impact that decreases when the execution time increases.

Fig. 5.11 we plotted the whole market order imbalance $\text{sign}(Q) \cdot V_M^{\text{signed}}$ [19]. The similarity of both pictures (see Fig. 5.11 top images) is striking : during metaorders, impact is a very nice square root of *global market imbalance*. This leads to the following conclusion : Market impact is not a reaction to *individual* metaorders, but to the *whole* order flow. This seems rather natural since orders are anonymous, hence the aggregated order flow should be the only relevant quantity.

19. $V_M^{\text{signed}} = \sum_i v_i \epsilon_i$ where $\epsilon_i = \pm 1$ according to whether the trade is triggered by the buyer (resp. seller) and v_i is its volume.

5.6.3 Permanent impact and correlation

One question remains however : How to study the mechanical impact of *one isolated* metaorder ? We can answer this question by searching the data for metaorders that are not correlated with the rest of the market, i.e. in the course of which the residual order flow does not trend nor anti-trend : we selected all metaorders that account for more than 75% of the market net imbalance on $[t_0, t_0+10T]$ where t_0 is the starting time of the metaorder and T is its duration, so that the rest of the market can be considered neutral during the measurements – hence the terminology of *isolated* metaorder. Fig. 5.12 compares the impact of randomly chosen metaorders (which are positively correlated to the rest of the market), that we will refer to as "informed", and isolated metaorders (that we will refer to as "uninformed") and clearly shows that impact of isolated metaorders decreases far below the 2/3 threshold. Note that the above selection labels about 3% of the metaorders as isolated. This illustrates the fact that impact is built up by an "informational" component, that reveals itself

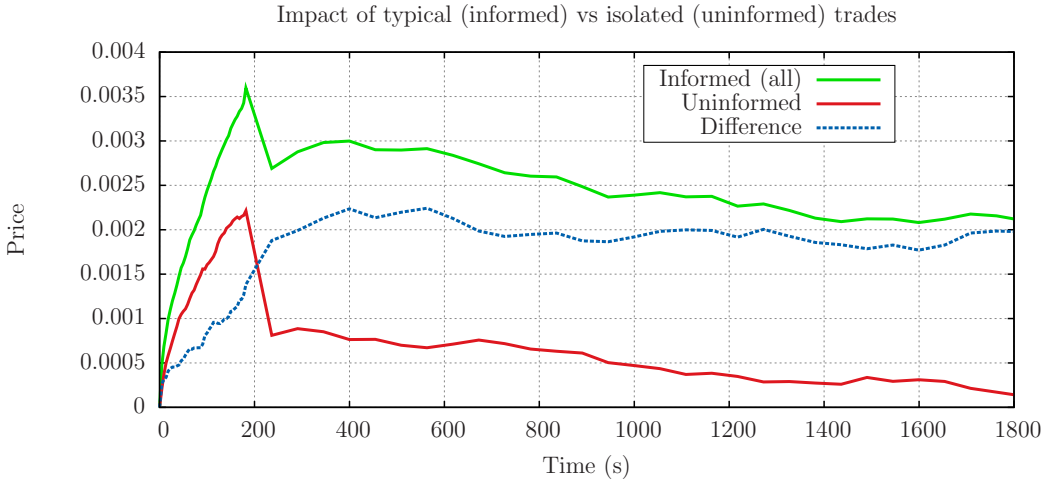

FIGURE 5.12 – Impact of "informed" metaorders vs. "uninformed" metaorders (i.e. isolated metaorders) on the opposite best price. While the former have a permanent impact due to their correlation with the residual order flow, the latter do not affect the price in the long run – or very few.

during and after the execution and results in an apparent permanent impact [20] (in the case of isolated orders this component is by definition zero or very small) and a "mechanical" component whose shape after execution is consistent with a decay all the way to zero, in agreement with the findings of Gomes and Waelbroeck (2015). This suggests that the mechanical peak impact is found by removing the permanent part of the impact :

$$\mathcal{I}^{\infty}_{\text{mec}}(Q,\mu) \approx \mathcal{I}(Q,\mu) - \mathcal{I}^{\infty}(Q,\mu)\ . \tag{5.10}$$

20. This is Hasbrouck's Hasbrouck (1991) definition of "information".

One might ask whether the "informed/uninformed" terminology is appropriate here, since we never wonder about external information. Actually, the fact that isolated metaorders have no permanent impact, and therefore no "informational content" on average, indicates that the permanent component of the impact – that one may refer to as the *informational content* of the metaorder – should be interpreted as a *correlation* between the metaorder and future traders' behaviours and not as some information about any "fundamental price" (see Donier and Bouchaud (2015b)). Concerning the dependence of impact on the execution speed, the same kind of impact surface can be plotted after conditioning on isolated metaorders only, showing that after accounting for the bias the impact actually *decreases*, as expected intuitively, when the execution speed decreases :

$$\mu_1 < \mu_2 \Rightarrow \mathcal{I}^{\infty}_{\text{mec}}(Q, \mu_1) < \mathcal{I}^{\infty}_{\text{mec}}(Q, \mu_2) \ .$$

However, the measure is too noisy to quantify precisely this dependence. Such a behaviour is strikingly reminiscent of what is obtained within the reaction-diffusion framework – and less consistent with equilibrium models which so far do not deal with the issue of execution speed.

5.7 Summary of main results

We have presented a very detailed analysis of market impact on the Bitcoin exchange market. For the sake of a clear understanding let us summarize here our main results :

1. Large orders are split into small trades which confirms the wide-spread belief that large orders have to be executed incrementally. On our MtGox dataset over 1 million metaorders are clearly identifiable, corresponding to 14M trades.

2. Metaorders size and duration distributions are not power-law. However, on Bitcoin they have the very nice property of being executed at a constant average rate.

3. The market impact of a metaorder executed incrementally and linearly in time can be reduced in a dependence on the execution speed and on the time elapsed since the start of the execution and written $\mathcal{I}_{\text{path}}(r, Q, \mu) = \mathcal{I}(t, \mu)$. The square-root law is clearly confirmed as it even holds *trajectory-wise* and is exact even at the smallest scales, which once again invalidates a broad class of equilibrium theories. Besides it has been shown that this cannot be explained by biases : the square-root law describes how market digests local excesses in supply /demand.

4. The impact pre-factor $\widetilde{Y}$ is subject to large fluctuations, that are well explained by the ratio $\sigma_D/\sqrt{V_D}$ where σ_D is the daily volatility and V_D the daily volume. The residual *Y-ratio*, as defined in Eq. 9.4, fluctuates around a mean-value of order unity : to be more precise, its distribution is well approximated by a Gaussian distribution of mean $Y_0 = 0.9$ and standard deviation $\Sigma_Y = 0.35$. The full stochastic impact formula we propose is thus given by Eq. 5.8.

5. We have presented strong evidence that the market reacts to the total order flow and not to distinct metaorders. This implies that metaorders are not detected by other market participants, as assumed in some equilibrium models Farmer et al. (2013). More importantly, the impacts of different metaorders are in fact *dependent* and *do not add up linearly.*
6. To subtract the effect of the order flow, and to measure the marginal effect of *one isolated* metaorder, we have selected metaorders that were not correlated to the market's local direction. This allowed us to show that for slow execution speeds the average execution cost $\mathcal{I}_{\text{exec}}(Q, \mu)$ decreases – even though measures are too noisy to make a proper fit – and that the mechanical component of the permanent impact $\mathcal{I}^{\infty}_{\text{mec}}(Q, \mu)$ is close to zero. Such a behaviour is consistent with kernel models and statistical order book models that naturally generalize them.
7. The marginal impact of these isolated metaorders drops far below the 2/3 level predicted by equilibrium theories. The permanent impact of isolated trades is probably zero, as concluded in Brokmann et al. (2015) and Gomes and Waelbroeck (2015).

5.8 Conclusion

Using a remarkable dataset which covers all transactions on the MtGox Bitcoin/USD exchange, we have conducted a comprehensive analysis of the market impact of over one million metaorders on Bitcoin. The Bitcoin market has two important features which motivate such a study : First, it corresponds to a single-asset economy, a quite unique example. Second, since each transaction is charged fees of 0.6% of its amount (60 bps), the existence of market makers and arbitrageurs is the exception rather than the rule.

The fact that the square-root law for impact holds in these conditions allows us to assess some underlying hypotheses of current impact models. Due to the large fees, neither statistical arbitrage nor high-frequency market making can be at the origin of the square-root law (on the Bitcoin) and theoretical explanations of this stylized fact must be found elsewhere. Accordingly, the notion of fair pricing is irrelevant in this market, as it seems not reasonable to assume that investors behave in such a way that they break even on their impact (of order a fraction of a percent) when they have to pay fees that are much larger. Nor should one rely on any martingale conditions for the price, as the notion of "price" is not precisely defined under the scale of 60 bps due to the large spread. The fact that the square-root impact law holds on the Bitcoin market so precisely at all scales clearly demonstrates that equilibrium and arbitrage mechanisms are not the underlying mechanism of market impact. Furthermore, our study suggests that impact should be regarded as how trades dig into the opposite side rather than how they affect the "price" itself, which is not equivalent when the spread is not tight.

To conclude, this study incites us to think that impact is driven by a generic fundamental and local mechanism that emerges together with the simple phenomenon of supply and demand and pre-

exist any notion of arbitrage (although it might remain compatible with it in the end as behaviours adapt to it). A promising avenue in our view, is the use of heterogeneous agent models Donier and Bouchaud (2015b), in which supply and demand are the output of interactions in a complex system that cannot be reduced to a few representative agents, and which seem to reproduce impact pictures remarkably well Donier et al. (2015).

Acknowledgements

We warmly thank Jean-Philippe Bouchaud for countless discussions and for carefully reading the manuscript. We also thank N. Kornman and A. Tilloy for their precious insights on the Bitcoin and finally C.A. Lehalle and H.Waelbroeck for very useful discussions.

5.9 Postface (français)

Nous avons donc retrouvé sur le Bitcoin la loi d'impact habituellement observée sur les marché matures. De deux choses l'une : ou bien il s'agit d'une coïncidence, et deux mécanismes différents mènent à la même loi, ou bien le mécanisme est le même et doit être cherché ailleurs que dans des hypothèses d'efficience. Je trouve pour ma part la première possibilité difficile à concevoir : je n'aime pas les coïncidences. Il va donc falloir chercher une explication à l'impact compatible avec l'amateurisme [21].

21. Ne serait-ce que pour expliquer les résultats sur le Bitcoin ! Car si les deux lois n'était similaires que par coïncidence, la loi d'impact en racine trouvée sur le Bitcoin en deviendrait une découverte des plus importantes !

Troisième partie

Théorie

Chapitre 6

Un modèle minimal et cohérent pour l'impact non linéaire

From
A fully consistent, minimal model for non-linear market impact
with Julius Bonart, Iacopo Mastromatteo and Jean-Philippe Bouchaud
(Donier et al., 2015)

6.1 Préface (français)

Si l'on devait caractériser le marché du Bitcoin et les *Bitcoiners*, le premier mot qui viendrait à l'esprit serait sans doute : *nombreux*, ou peut-être *hétérogènes* (à condition bien sûr d'exclure *geeks* ou *fous* de la liste). Difficile de savoir pourquoi exactement ils contribuent au marché, ou d'imaginer que leur action soit le résultat de choix rationnels et optimaux qui leur permettraient de cerner le prix avec une précision inférieure au dixième de pourcent [1]. Cela ouvre grand la voie de la modélisation *statistique* : et si l'on avouait ne pas comprendre grand chose aux motivations des *Bitcoiners* et ne prenait en compte que leur hétérogénéité – jusqu'où pourrions-nous aller [2] ?

Il ne s'agirait pas de la première fois que ce genre de modèles serait évoqué en finance : Bak et al. (1997); Maslov (2000); Smith et al. (2003), ou plus récemment Lasry and Lions (2007) avaient déjà eu l'intuition d'une modélisation statistique des carnets d'ordres pour expliquer la dynamique des prix. Il s'est toutefois avéré que ces modèles ne permettaient pas de produire des prix approximativement diffusifs – condition *sine qua non* d'un modèle admissible – et contrastaient avec les mesures réalisées

1. Rappelons encore une fois que les frais de transaction s'élèvent à 0.6% du prix sur le marché que nous avons étudié, et que la volatilité journalière moyenne y est de 7% !

2. Les physiciens ont bien réussi à établir la loi des gaz parfaits en renonçant à décrire le comportement individuel de chaque particule. Toutefois, je préfère préciser que *l'analogie s'arrête là* : il ne s'agit pas de calquer un problème économique sur un problème physique connu, seulement de s'inspirer de l'idée statistique.

sur les carnet d'ordres affichés par les places de marchés – ils furent donc laissés de côté. C'est que personne ne s'y était encore posé la question de l'impact, ni n'avait pu voir à quel point leurs prédictions étaient bonnes le concernant.

L'analyse empirique du Bitcoin présentée dans la section précédente a pourtant fortement encouragé la vision statistique de l'offre et la demande pour rendre compte de l'impact, signe qu'aux modèles statistiques évoqués ci-dessus il ne manque sans doute qu'un peu de réflexion. Le premier des concepts manquants a été introduit par Tóth et al. (2011) : il s'agit du carnet d'ordres *latent* (qui en réalité n'est rien d'autre qu'une manière nouvelle de représenter l'offre et la demande, cf Chapitre 7). L'idée de l'existence d'ordres latents (ou *intentions*), non obligatoirement affichés publiquement sur un carnet d'ordres observable bien qu'ancrés dans les esprits des agents, aura en effet permis de s'abstraire des contraintes imposées par l'observation des carnets d'ordres réels, sur lesquels les agents n'affichent en réalité leurs intentions que lorsqu'elles sont sur le point d'être exécutées, dans un grand jeu de cache-cache de plus en plus brouillé par l'essor du trading algorithmique et l'arrivée des traders à haute fréquence. L'aspect remarquable du carnet d'ordres latent, est qu'il semble donner une piste d'explication à la « loi d'impact en racine » sur des échelles de prix potentiellement bien plus grandes que la profondeur du carnet d'ordre affiché : il s'agit là du premier soutien empirique fort à ce genre de « modèles à particules » [3].

Enthousiasmés par la piste d'explication de l'impact en racine, Mastromatteo et al. (2014b) ont été les premiers à dériver une formule analytique pour l'impact d'une force constante exercée sur l'offre ou la demande dans le cadre du carnet d'ordres latent. Le résultat, en accord avec les résultats numériques de Tóth et al. (2011), fut que dans un modèle de réaction-diffusion le prix se déplace comme la racine carrée exacte du volume exécuté [4] lorsque l'exécution se fait à taux constant. Il s'agit en réalité du même calcul que celui du déplacement de l'interface dans un problème physique connu sous le nom de « Two-phases Stefan Problem ».

Trois choses manquent cependant : (i) la diffusivité des prix qui n'est toujours pas expliquée, celui-ci étant fortement confiné par le carnet d'ordres, (ii) une formule générique pour l'impact qui donne l'effect sur le prix d'une action dans le cas *général* – et pas seulement dans le cas d'une pression constante et (iii) l'assurance que le modèle ainsi généré est cohérent en interne, c'est-à-dire, qu'il est *non-arbitrable* (cf Chapitre 4.6). Ces trois éléments sont fournis par l'article qui suit [5].

Abstract : We propose a minimal theory of non-linear price impact based on the fact that the (latent) order book is locally linear, as suggested by diffusion-reaction models and general arguments. Our framework allows one to compute the average price trajectory in the presence of a meta-order,

3. Connu en physique sous le nom de *réaction-diffusion*.

4. Ou plus précisément, du temps : la trajectoire d'impact est en racine mais son préfacteur qui dépend du taux d'exécution. Nous comprendrons plus loin ce qui se passe exactement.

5. Et tout cela, dans un modèle minimal, comme l'évoque le titre ci-dessus : il ne nécessite en effet que deux paramètres scalaires !

that consistently generalizes previously proposed propagator models. We account for the universally observed square-root impact law, and predict non-trivial trajectories when trading is interrupted or reversed. We prove that our framework is free of price manipulation, and that prices can be made diffusive (albeit with a generic short-term mean-reverting contribution). Our model suggests that prices can be decomposed into a transient "mechanical" impact component and a permanent "informational" component.

6.2 Introduction

The study of market impact (i.e. the way trading influences prices in financial markets) is arguably among the most exciting current themes in theoretical finance, with many immediate applications ranging from trading cost modelling to important regulatory issues. What is the meaning of the market price if the very fact of buying (or selling) can substantially affect that price ? The questions above would on their own justify a strong research activity that dates back to the classic Kyle (1985) paper. But as often in science, it is the empirical discovery of a genuinely surprising result that explains the recent spree of activity on the subject (see e.g. Almgren et al., 2005; Tóth et al., 2011; Farmer et al., 2013; Mastromatteo et al., 2014a; Skachkov, 2014 and refs. therein). In strong contrast with the predictions of the Kyle model, market impact appears to be neither *linear* (in the traded quantity Q) nor *permanent*, i.e. time independent (Bouchaud et al., 2009). As now firmly established by many independent empirical studies, the average price change induced by the sequential execution of a total volume Q (which we call *meta-order*) appears to follow a sub-linear, approximate $\sqrt{Q}$ law (Torre and Ferrari, 1997; Grinold and Kahn, 2000; Almgren et al., 2005; Moro et al., 2009; Tóth et al., 2011; Mastromatteo et al., 2014a; Gomes and Waelbroeck, 2015; Bershova and Rakhlin, 2013; Brokmann et al., 2015). At the end of the meta-order, impact is furthermore observed to decay (partially or completely) towards the unimpacted price (Gomes and Waelbroeck, 2015; Bershova and Rakhlin, 2013; Brokmann et al., 2015; Donier and Bonart, 2014).

Quite strikingly, the square-root law appears to be *universal*, as it is to a large degree independent of details such as the type of contract traded (futures, stocks, options or even Bitcoin, see Donier and Bonart, 2014...), the geographical position of the market venue (US, Europe, Asia), the time period (1995 $\rightarrow$ 2014), the maturity of the market (e.g. Bitcoin vs. S&P500), etc. While the impact of *single orders* is non universal and highly sensitive to market micro-structure, the impact of *meta-orders* appears to be extremely robust against micro-structural changes. For example the rise of high-frequency trading (HFT) in the last ten years seems to have had no effect on its validity (compare Torre and Ferrari, 1997; Almgren et al., 2005 that uses pre-2004 data with Tóth et al., 2011; Mastromatteo et al., 2014a; Gomes and Waelbroeck, 2015 that use post-2007 data). This universality strongly suggests that simple, "coarse-grained" models should be able to reproduce the square-root impact law and other slow market phenomena, while abstracting away from many

microscopic details that govern order flow and price formation at high frequencies. This line of reasoning is very similar to many situations in physics, where universal large scale/low frequency laws appear for systems with very different microscopic behaviour. A well known example is the behaviour of weakly interacting molecules which on large length scales can be accurately described by the Navier-Stokes equation, with a single "emergent" parameter (the viscosity) that encodes the microscopic specificities of the system. The Navier-Stokes equation can in fact be derived either from the statistical description of the dynamics of molecules, through an appropriate coarse-graining procedure, or from general considerations based on symmetries, conservation laws and dimensional arguments. Along this path, two pivotal ideas have recently emerged. One is the concept of a *latent* order book (Tóth et al., 2011) that contain the intentions of low-frequency actors at any instant of time, which may or may not materialize in the observable order book. Indeed, since the square-root impact is an aggregate, low-frequency phenomenon, the relevant object to consider cannot be the "revealed" order book, which chiefly reflects the activity of high frequency market-makers. Simple orders of magnitude confirm that the latent liquidity is much higher than the revealed liquidity : whereas the total daily volume exchange on a typical stock is around 1/200th of its market capitalisation, the volume present in the order book at any instant in time is 1000 times smaller than this. Market-makers only act as small intermediaries between much larger volume imbalances present in the latent order book, that can only get resolved on large time scales.

The second idea is that the dynamics of the latent order book can be faithfully modelled by a so-called "reaction-diffusion" model, at least in a region close to the current price where this dynamics becomes universal, i.e. independent of the detailed setting of the model – and hence, as emphasized above, of the detailed micro-structure of the market and of its high-frequency activity. The reaction-diffusion model in one dimension posits that two types of particles (called B and A), representing in a financial context the intended orders to buy (*bids*) and to sell (*ask*) diffuse on a line and disappear whenever they meet $A + B \to \emptyset$ – corresponding to a transaction. The boundary between the B-rich region and the A-rich region therefore corresponds to the price p_t. This highly stylized order book model was proposed in the late 90's by Bak et al. (1997) (see also Tang and Tian, 1999) but never made it to the limelight because the resulting price dynamics was found to be strongly mean-reverting on all time scales, at odds with market prices which, after a short transient, behave very much like random walks. However, some of us (Mastromatteo et al., 2014b) recently realized that the analogue of market impact can be defined and computed within this framework, and was found to obey the square-root law exactly.

This opens the door to a fully consistent theoretical model of non-linear impact in financial markets, which we propose in the present paper. We show how all the previously discussed ingredients can be accommodated in a unifying coarse-grained model for the dynamics of the latent order book that is consistent with price diffusion, with a single emergent parameter – the market *liquidity* $\mathcal{L}$, defined below. When fluctuations are neglected (in a sense that will be specified below), the impact

of a meta-order can be computed exactly, and is found to exhibit two regimes : when the execution rate is sufficiently slow, the model becomes identical to the linear propagator framework proposed in Bouchaud et al. (2004, 2009), with a bare propagator decaying as the inverse square-root of time. When execution is faster, impact becomes fully non-linear and obeys a non-trivial, closed form integral equation. In the two regimes, the impact of a meta-order grows as the square-root of the volume, but with a pre-factor that depends on the execution rate in the slow regime, but becomes independent of it in the fast regime – as indeed suggested by empirical data. The model predicts interesting price trajectories when trading is interrupted or reversed, leading to effects that are observed empirically but impossible to account for within a linear propagator model. We demonstrate that prices in our model cannot be manipulated, in the sense that any sequence of buy and sell orders that starts and ends with a zero position on markets leads to a non-positive average profit. This is a non trivial property of our modelling strategy, which makes it eligible for practical applications. Finally, we discuss how our framework suggests a clear separation between "mechanical" price moves (i.e. induced by the impact of random trades) and "informational" price moves (i.e. the impact of any public information that changes the latent supply/demand). This is a key point that allows us to treat consistently, within the same model, diffusive prices and memory of the order book – which otherwise leads to strongly mean-reverting prices (see the discussion in Tóth et al., 2011; Mastromatteo et al., 2014a; Taranto et al., 2014). We discuss in the conclusion some of the many interesting problems that our modelling strategy leaves open – perhaps most importantly, how to consistently account for order book fluctuations that are presumably at the heart of liquidity crises and market crashes.

6.3 Dynamics of the latent order book

Our starting point is the zero-intelligence model of Smith et al. (2003), reformulated in the context of the latent order book in Tóth et al. (2011) and independently in Lehalle et al. (2011). We assume that each trading (buy/sell) intention of market participant is characterized by a reservation price and a volume.[6] In the course of time, the dynamics of intentions can be essentially of four types : a) reassessment of the reservation price, either up or down ; b) partial or complete cancellation of the intention of buying/selling ; c) appearance of new intentions, not previously expressed and finally d) matching of an equal volume of buy/sell intentions, resulting in a transaction at a price that delimits the buys/sells regions, and removal of these intentions from the latent order book. It is clear that provided very weak assumptions are met : i) the changes in reservation prices are well behaved (i.e. have a finite first and second moment) and short-ranged correlated in time ; ii) the volumes have a finite first moment, one can establish – in the large scale, low frequency,

6. In fact, each participant may have a full time dependent supply/demand curve with different prices and volumes, with little change in the effective model derived below.

"hydrodynamic" limit – the following set of partial differential equations for the dynamics of the *average* buy (resp. sell) volume density $\rho_B(x,t)$ (resp. $\rho_A(x,t)$) at price level x :

$$\frac{\partial \rho_B(x,t)}{\partial t} = -V_t \frac{\partial \rho_B(x,t)}{\partial x} + D\frac{\partial^2 \rho_B(x,t)}{\partial x^2} - \nu \rho_B(x,t) + \lambda \Theta(p_t - x) - \kappa R_{AB}(x,t); \quad \text{(1-a)}$$

$$\frac{\partial \rho_A(x,t)}{\partial t} = \underbrace{-V_t \frac{\partial \rho_A(x,t)}{\partial x} + D\frac{\partial^2 \rho_A(x,t)}{\partial x^2}}_{\text{a- Drift-Diffusion}} \underbrace{- \nu \rho_A(x,t)}_{\text{b- Cancel.}} + \underbrace{\lambda \Theta(x - p_t)}_{\text{c- Deposition}} \underbrace{- \kappa R_{AB}(x,t)}_{\text{d- Reaction}}; \quad \text{(1-b)}$$

where the different terms in the right hand sides correspond to the four mechanisms a-d, on which we elaborate below, and p_t is the *coarse-grained* position of the price (i.e. averaged over high frequency noise), defined from the condition

$$\rho_A(p_t,t) - \rho_B(p_t,t) = 0. \quad (6.1)$$

— a- *Drift-Diffusion :* The first two terms model the fact that each agent reassess his/her reservation price x due to many external influences (news, order flow and price changes themselves, other technical signals, etc.). One can therefore expect that price reassessments contain both a (random) agent specific part that contributes to the diffusion coefficient [7] D and a *common* component V_t that shifts the entire latent order book. This shift is due to a collective price reassessment due for example to some publicly available information (that could well be the past transactions themselves). The drift component V_t is at this stage very general ; one possibility that we will adopt below is to think of V_t as a white noise, such that the price p_t is a diffusive random walk. Since the derivation of these first two terms and the assumptions made are somewhat subtle, we devote Appendix A to a more detailed discussion and alternative models ; see in particular Eq. (6.30).

— b- *Cancellations :* The third term corresponds to partial or complete cancellation of the latent order, with a decay time ν^{-1} independent of the price level x (but see previous footnote[2]). Consistent with the idea of a common information, cancellation could be correlated between different agents. However, this does not affect the evolution of the average densities $\rho_{B,A}(x,t)$, while it might play a crucial role for the fluctuations of the order book, in particular to explain liquidity crises.

— c- *Deposition :* The fourth term corresponds to the appearance of new buy/sell intentions, modelled by a "rain intensity" λ modulated by an arbitrary increasing function $\Theta(u)$, expressing that buy orders mostly appear below the current price p_t and sell orders mostly appear above p_t. The detailed shape of $\Theta(u)$ actually turns out to be, to a large extent, irrelevant for

7. In full generality, the diffusion constant D could depend on the distance $|x - p_t|$ to the transaction price. We neglect this possibility in the present version of the model, for reasons that will become clear later : see Appendix B. A similar remark applies to the cancellation rate ν as well.

the purpose of the present paper (see Appendix B for details) ; for simplicity we will choose below a step function, $\Theta(u > 0) = 1$ and $\Theta(u < 0) = 0$.

— d- *Reaction :* The last term corresponds to transactions when two orders meet with "reaction rate" κ ; the quantity $R_{AB}(x,t)$ is formally the average of the product of the density of A particles and the density of B particles, i.e. $R_{AB}(x,t) \approx \rho_A(x,t)\rho_B(x,t)$ + fluctuations. However, the detailed knowledge of $R_{AB}(x,t)$ will not affect the following discussion. We will consider in the following the limit $\kappa \to \infty$, which corresponds to the case where latent limit orders close to the transaction price *all become instantaneously visible* limit orders that are duly executed against incoming market orders.

Let us insist that Eqs. (6.1-a,b) only describe the *average* shape of the latent order book, i.e. fluctuations coming from the discrete nature of orders are neglected at this stage : see Fig. 6.1 for an illustration. In particular, the instantaneous position of the price $p_t^{\text{inst.}}$ – where the density of buy/sell orders vanishes – has an intrinsic non-zero width even in the limit $\kappa \to \infty$ (Barkema et al., 1996), corresponding to the average distance $a - b$ between the highest buy order $x = b$ and the lowest sell order $x = a$.[8] The instantaneous price can then be conventionally be defined as $p_t^{\text{inst.}} = (a + b)/2$, but will in general not coincide with the coarse-grained price p_t defined by the average shape of the latent order book through Eq. (6.1). Indeed, as shown in Barkema et al. (1996), the diffusion width (i.e. the typical distance between $p_t^{\text{inst.}}$ and p_t) is also non zero and actually larger than the intrinsic width, but only by a logarithmic factor.

In the following, we will neglect both the intrinsic width and the diffusion width, which is justified if we focus on price changes much larger than these widths. This is the large scale, low frequency regime where our coarse-grained equations Eqs. (6.1-a,b) are warranted. Formally, Eqs. (6.1-a,b) become valid when the market *latent liquidity* $\mathcal{L}$ (defined below) tends to infinity, since both the intrinsic width and the diffusion width vanish as $\mathcal{L}^{-1/2}$ (Barkema et al., 1996).

6.4 Stationary shape of the latent order book

A remarkable feature of Eqs. (6.1-a,b) is that although the dynamics of ρ_A and ρ_B is non-trivial because of the reaction term (that requires a control of fluctuations, see Barkema et al., 1996) the combination $\varphi(x,t) := \rho_B(x,t) - \rho_A(x,t)$ evolves according to a linear equation independent of κ :[9]

$$\frac{\partial \varphi(x,t)}{\partial t} = -V_t \frac{\partial \varphi(x,t)}{\partial x} + D\frac{\partial^2 \varphi(x,t)}{\partial x^2} - \nu\varphi(x,t) + \lambda \,\text{sign}\,(p_t - x)\,, \tag{6.2}$$

8. We indeed assume that latent orders become instantaneously visible when close to $p_t^{\text{inst.}}$, in such a way that the latent order book and the observable order book become identical at the best limits. This is of course needed to identify $p_t^{\text{inst.}}$ with the "real" mid-price. It is very interesting to ask what happens if the conversion speed between latent orders and real orders is not infinitely fast, or when market orders become out-sized compared to the prevailing liquidity. As we discuss in the conclusion, this is a potential mechanism for crashes, and the simple coarse-grained framework discussed here has to be adapted to deal with these situations.

9. The disappearance of κ can be traced to the conservation of $\#A - \#B$ for each reaction $A + B \to \emptyset$.

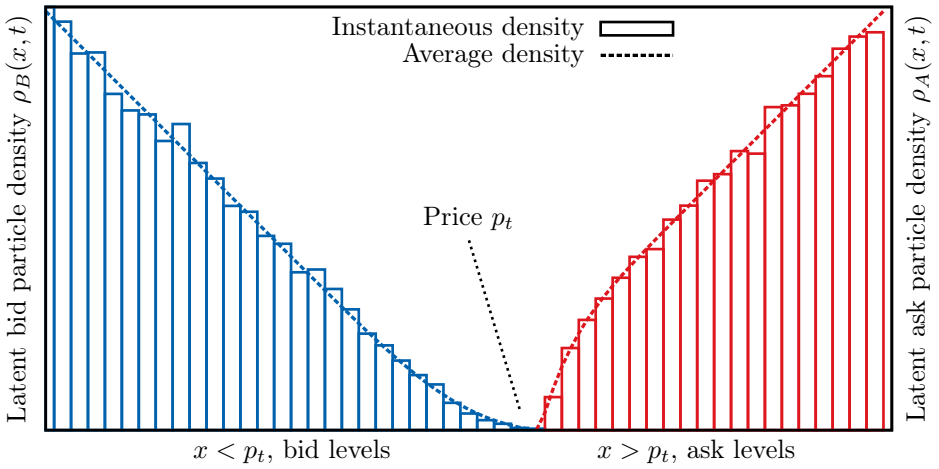

FIGURE 6.1 – Snapshot of a latent order book in the presence of a meta-order, with bid orders (blue boxes) and ask orders (red boxes) sitting on opposite sides of the price line and subject to a stochastic evolution. The dashed lines show the mean values the order densities $\rho_{A,B}(x,t)$, which are controlled by Eqs. (6.1).

where p_t is the solution of $\varphi(p_t,t)=0$. This solution is expected to be unique for all $t>0$ if it is unique at $t=0$ (see also Lehalle et al., 2011). Note that Eq. (6.2), without the drift-diffusion terms, has recently been obtained as the hydrodynamic limit of a Poisson order book dynamics in Gao et al. (2014).

Introducing $\widehat{p}_t = \int_0^t \mathrm{d}s\, V_s$, the above equation can be rewritten in the reference frame of the latent order book $y = x - \widehat{p}_t$ as :[10]

$$\frac{\partial\varphi(y,t)}{\partial t} = D\frac{\partial^2\varphi(y,t)}{\partial y^2} - \nu\varphi(y,t) + \lambda\,\mathrm{sign}\,(p_t - \widehat{p}_t - y)\,. \tag{6.3}$$

Starting from a symmetric initial condition $\varphi(y,t=0) = -\varphi(-y,t=0)$ such that $p_{t=0} = \widehat{p}_{t=0} = 0$, it is clear by symmetry that the equality $p_t = \widehat{p}_t$ is a solution at all times, since all terms in the above equation are odd when $y \to -y$. For more general initial conditions, p_t converges to $\widehat{p}_t$ when $t\to\infty$ and the stationary solution of Eq. (6.3) reads, in the limit $\mu\to\infty$:

$$\varphi_{\mathrm{st.}}(y\le 0) = \frac{\lambda}{\nu}\left[1 - e^{\gamma y}\right]; \qquad \varphi_{\mathrm{st.}}(y\ge 0) = -\varphi_{\mathrm{st.}}(-y), \tag{6.4}$$

10. One should be careful with the "Ito" term when V_s is a Wiener noise, which adds a contribution to D, see Appendix A and Eq. (6.31)

with $\gamma^2 = \nu/D$. This is precisely the solution obtained in Tóth et al. (2011) which behaves linearly close to the transaction price. But as emphasized in Tóth et al. (2011) and in Appendix B, this linear behaviour in fact *holds for a very wide range of models* – for example if the appearance of new orders only takes place at some arbitrary boundary $y = \pm L$, as in Mastromatteo et al. (2014b), or else if the coefficients D, ν are non-trivial (but sufficiently regular) functions of the distance to the price $|y|$, etc.

6.5 Price dynamics within a locally linear order book (LLOB)

We will therefore, in the following, "zoom" into the universal linear region by taking the formal limit $\gamma \to 0$ with a fixed current

$$J = D|\partial_y \varphi_{\text{st.}}|_{y=0} \equiv \lambda/\gamma. \tag{6.5}$$

This current can be interpreted as the volume transacted per unit time in the stationary regime, i.e. the total quantity of buy (or sell) orders that get executed per unit time. As a side remark, it is important to realize that if the drift V_t contains a Wiener noise component, or jumps, this drift does in fact contribute to J and does not merely shift the latent order book around without any transactions (see Appendix A).

In the limit $\nu, \lambda \to 0$ with $\lambda/\gamma = J$ fixed, the stationary solution $\varphi_{\text{st.}}(y)$ becomes exactly linear :

$$\varphi_{\text{st.}}(y) = -Jy/D. \tag{6.6}$$

This is the regime we will explore in the present paper, although we will comment below on the expected modifications induced by non-zero values of ν, λ. Note that $\mathcal{L} = J/D \equiv \lambda\sqrt{D/\nu}$ can be interpreted as the *latent liquidity* of the market, which is large when deposition of latent orders is intense (λ large) and/or when latent orders have a long lifetime (ν small). The quantity $\mathcal{L}^{-1}$ is the analogue, within a LLOB, of Kyle's "lambda" for a flat order book.

In terms of order of magnitudes, it is reasonable to expect that the latent order book has a memory time ν^{-1} of several hours to several days (Tóth et al., 2011) – remember that we are speaking here of slow actors, not of market makers contributing to the high-frequency dynamics of the revealed order book. Taking D to be of the order of the price volatility, the width of the linear region γ^{-1} is found to be of the order of 1% of the price (see Eq. (6.4)). Therefore, we expect that restricting the analysis to the *linear* region of the order book will be justified for meta-orders lasting up to several hours, and impacting the price by less than a fraction of a percent. For larger impacts and/or longer execution times, a more elaborate (and probably less universal) description may be needed.

We now introduce a "meta-order" within our framework and work out in detail its impact on the price. Working in the reference frame of the *unimpacted* price $\widehat{p}_t$ defined above, we model a

meta-order as an extra current of buy (or sell) orders that fall exactly on the transaction price p_t. Introducing $y_t \equiv p_t - \widehat{p}_t$, the corresponding equation for the latent order book reads, within a LLOB that precisely holds when $\nu, \lambda \to 0$:

$$\begin{cases} \dfrac{\partial \varphi(y,t)}{\partial t} = D \dfrac{\partial^2 \varphi(y,t)}{\partial y^2} + m_t \delta(y - y_t) \\ \dfrac{\partial \varphi(y \to \pm\infty, t)}{\partial y} = -\mathcal{L}, \end{cases} \tag{6.7}$$

where m_t is the (signed) trading intensity at time t ; $m_t > 0$ corresponding to a buy meta-order. Note that the meta-order will be assumed to be small enough not to change the behaviour of the rest of the market (i.e. the parameters D, ν and λ), so that $\mathcal{L}$ is a fixed parameter in the above equation. Of course, this assumption might break down when the meta-order is out-sized, leading to a sudden increase of the cancellation rate ν and a corresponding drop of the liquidity $\mathcal{L}$, which might in turn result in a crash (see the discussion in the conclusion).

We will now consider a meta-order that starts at a random time that we choose as $t = 0$, with no information on the state of the latent order book. This means that at $t = 0$, there is no conditioning on the state of the order book that can be described by its stationary shape, $\varphi_{\text{st.}}(y) = -Jy/D$. For $t > 0$, the latent order book is then given by the following exact formula :

$$\varphi(y,t) = -\mathcal{L}y + \int_0^t \frac{\mathrm{d}s\, m_s}{\sqrt{4\pi D(t-s)}} e^{-\frac{(y-y_s)^2}{4D(t-s)}}, \tag{6.8}$$

where y_s is the transaction price (in the reference frame of the book) at time s, defined as $\varphi(y_s, s) \equiv 0$. This leads to a self-consistent integral equation for the price at time $t > 0$:

$$y_t = \frac{1}{\mathcal{L}} \int_0^t \frac{\mathrm{d}s\, m_s}{\sqrt{4\pi D(t-s)}} e^{-\frac{(y_t-y_s)^2}{4D(t-s)}}. \tag{6.9}$$

This is the central equation of the present paper, which we investigate in more detail in the next sections. [11]

As a first general remark, let us note that provided impact is small, in the sense that $\forall t, s$, $|y_s - y_t|^2 \ll D(t-s)$, then the above formula exactly boils down to the *linear propagator model* proposed in Bouchaud et al. (2004, 2009) (see also Gatheral, 2010), with a square-root decay of impact :

$$y_t = \frac{1}{\mathcal{L}} \int_0^t \frac{\mathrm{d}s\, m_s}{\sqrt{4\pi D(t-s)}}. \tag{6.10}$$

This linear approximation is therefore valid for very small trading rates m_s, but breaks down for more aggressive executions, for which a more precise analysis is needed. An ad-hoc non-linear

11. When m_s has a non-trivial time dependence, the above equation may not be easy to deal with numerically. It can be more convenient to iterate numerically the Eq. (6.7) and find the solution of $\varphi(y_t, t) = 0$.

generalisation of the propagator model was suggested by Gatheral (2010), but is difficult to justify theoretically (and leads to highly singular optimal trading schedules in the continuous time limit, see Curato et al., 2014). We believe that Eq. (A.2) above is the correct way to generalize the propagator model, such that all known empirical results can be qualitatively accounted for.

Note that one can in fact define a volume dependent "bid" (or "ask") price $y_t^{\pm}(q)$ for a given volume q as the solution of :

$$\int_{y_t^-(q)}^{y_t} \mathrm{d}y\, \varphi(y,t) = -\int_{y_t}^{y_t^+(q)} \mathrm{d}y\, \varphi(y,t) = q. \tag{6.11}$$

Clearly, in the equilibrium state, and for q small enough, $y_t^{\pm}(q) = y_t \pm \sqrt{2q/\mathcal{L}}$. After a buy meta-order, however, we will find that strong asymmetries can appear.

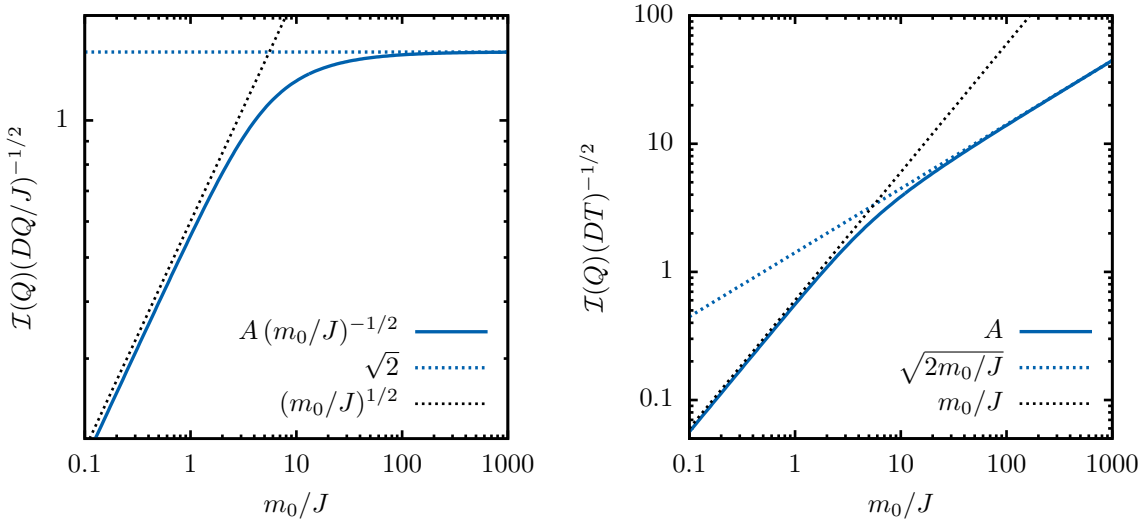

FIGURE 6.2 – Left : Dependence of the ratio $A/\sqrt{m_0/J}$ upon the trading rate parameter m_0/J. (This ratio coincides with the empirically used Y ratio if σ^2 is identified with D and V with J). The curve interpolates between a $\sqrt{m_0/J}$ dependence observed at small trading trading rates and an asymptotically constant regime $\approx \sqrt{2}$ for large m_0/J. This is consistent with the weak dependence of Y upon the trading rate observed in CFM empirical data. Right : Dependence of the impact $\mathcal{I}(Q)$ on Q for a *fixed* execution time T – i.e. a variable $m_0 = Q/T$. Note the crossover between a linear behaviour at small Q and a square-root behaviour for large Q.

6.6 The square-root impact of meta-orders

The simplest case where a fully non-linear analysis is possible is that of a meta-order of size Q executed at a constant rate $m_0 = Q/T$ for $t \in [0, T]$. In this case, it is straightforward to check that $y_s = A\sqrt{Ds}$ is an *exact* solution of Eq. (A.2), where the constant A is the solution of the following equation :

$$A = \frac{m_0}{J} \int_0^1 \frac{\mathrm{d}u}{\sqrt{4\pi(1-u)}} e^{-\frac{A^2(1-\sqrt{u})}{4(1+\sqrt{u})}}. \tag{6.12}$$

It is easy to work out the asymptotic behaviour of A in the two limits $m_0 \ll J$ and $m_0 \gg J$. In the first case, one finds $A \approx m_0/J\sqrt{\pi}$, while in the second case $A \approx \sqrt{2m_0/J}$. The impact $\mathcal{I}$ of a meta-order of size Q, is defined as : [12]

$$\mathcal{I}(Q) = \langle \varepsilon \cdot (p_{t+T} - p_t) | Q \rangle, \tag{6.13}$$

where $\langle \dots | Q \rangle$ denotes an average over all meta-orders of sign ε and volume Q, executed over the time interval $[t, t+T]$.

We assuming for now that the meta-order is uninformed, in the following sense :

$$\langle \varepsilon \cdot (\hat{p}_{t+T} - \hat{p}_t) | Q \rangle = 0, \tag{6.14}$$

such that the only contribution is the "mechanical" impact on the dynamics of y_t. The case of informed meta-orders will be treated in Sect. IX. The mechanical impact at the end of the meta-order is then given by $y_T = A\sqrt{DT}$, i.e. : [13]

$$\mathcal{I}(Q) = \frac{A}{\sqrt{m_0}}\sqrt{DQ} \approx \sqrt{\frac{m_0}{J\pi}} \times \sqrt{\frac{Q}{\mathcal{L}}} \qquad (m_0 \ll J); \qquad \mathcal{I}(Q) \approx \sqrt{2\frac{Q}{\mathcal{L}}} \qquad (m_0 \gg J), \tag{6.15}$$

i.e. precisely a square-root impact law.

In fact, the empirical result is often written as $\mathcal{I}(Q) = Y\sigma\sqrt{Q/V}$ where σ is the daily volatility and $V \equiv JT_d = D\mathcal{L}T_d$ the daily traded volume ($T_d \equiv 1$ day), and Y a constant of order unity. Assuming that $\sigma^2 \propto DT_d$ (which is the case if $D_0 = 0$, see Appendix A), we see that Eq. (6.15) exactly reproduces the empirical result, with Y proportional to $\sqrt{m_0/J}$ for small trading intensity m_0 and becoming independent of m_0 for larger trading intensity – see Fig. 6.2. [14] CFM's empirical data indeed suggests that Y only very weakly depends on the trading intensity, which is nicely

12. Note that $\mathcal{I}(Q)$ is a slight abuse of notations since the impact in fact depends in general in the whole trajectory m_s.

13. The results in the two limits are (up to prefactors) those obtained in Mastromatteo et al. (2014b) within an explicit reaction-diffusion setting.

14. Note that in agreement with our interpretation of the latent order book, the quantity JT must be interpreted as the volume of "slow" orders executed in a time T, removing all fast intra-day activity that averages out and therefore cannot withstand (other than temporarily) the incoming meta-order.

explained by the present framework.

6.7 Impact decay : beyond the propagator model

The next interesting question is impact relaxation : how does the price behave after the meta-order has been executed, i.e. when $t > T$. Mathematically, the impact decay is given by the solution of :

$$y_t = \frac{Dm_0}{J}\int_0^T \frac{\mathrm{d}s}{\sqrt{4\pi D(t-s)}} e^{-\frac{(y_t - A\sqrt{Ds})^2}{4D(t-s)}}, \qquad (t > T) \tag{6.16}$$

In the small m_0/J limit, the linear propagation model is appropriate and predicts the following impact relaxation :

$$\frac{\mathcal{I}(Q, t > T)}{\mathcal{I}(Q)} = \frac{\sqrt{t} - \sqrt{t-T}}{\sqrt{T}}, \tag{6.17}$$

that behaves as $1 - \sqrt{(t-T)/T}$ very shortly after the end of the meta-order and as $\sqrt{T/t}/2$ at long times.

The analysis of Eq. (6.16) at large m_0/J is more subtle, in particular at short times. The full analysis is given in Appendix C and reveals that the rescaled initial decay of impact is, quite unexpectedly, still exactly given by Eq. (6.17), independently of m_0/J. For large times, $y_t \to 0$, which implies that asymptotically $|y_t - A\sqrt{Ds}| \ll \sqrt{Dt}$, i.e. the exponential term in Eq. (6.16) is approximately equal to one, leading to an asymptotic rescaled impact decay as $\sqrt{m_0 T/2\pi J t}/4$. We plot in Fig. 6.3 the normalized free decay of impact for different values of m_0/J for the "mid-price" p_t, and in Fig. 6.4 the corresponding evolution of the effective "bid-ask" $p_t^{\pm}(q)$ for a given volume q, illustrating how the latent order book becomes more and more asymmetric as m_0/J increases.

The above analysis can be extended to the case where trading is reverted after time T, i.e. $m_t = m_0$ for $t \in [0, T]$ and $m_t = -m_0$ for $t \in [T, 2T]$. This case is particularly interesting since it puts the emphasis on the lack of liquidity behind the price for large execution rates. Within the linear propagator approximation, it is easy to show that the time needed for the price to come back to its initial value (before continuing to be pushed down by the sell meta-order) is given by $T/4$. In the non linear regime $m_0 \gg J$, the price goes down much faster, and reaches its initial value after a time given by $JT/2m_0 \ll T/4$ – see Figs. 6.5, 6.6. Such an asymmetry is indeed seen empirically, and means that such a simple round-trip is necessary costly, since the average sell price is below the average buy price. We shall see below that this property (called absence of price manipulation in Huberman and Stanzl (2004); Alfonsi and Schied (2010)) holds in full generality within our framework.

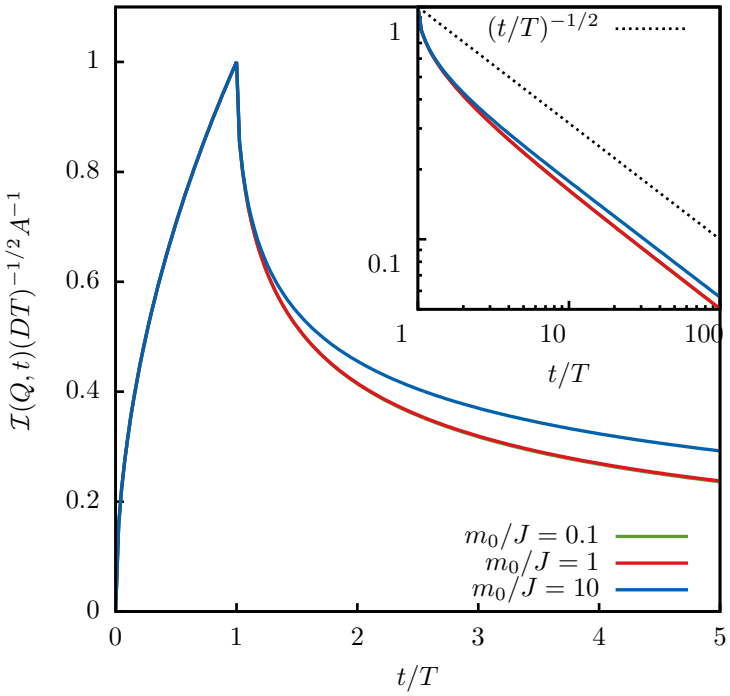

FIGURE 6.3 – Impact $\mathcal{I}(Q,t)$ as a function of rescaled time for various trading rate parameters m_0/J. The initial growth of the impact follows exactly a square-root law, and is followed by a regime shift suddenly after end of the meta-order. While for $t = T^+$ the slope of the impact function becomes infinite, at large times one observes an inverse square relaxation $\sim \sqrt{T/t}$ with an m_0/J dependent pre-factor. Note that the curves for $m_0/J = 0.1$ and $m_0/J = 1$ are nearly indistinguishable.

6.8 Price trajectory at large trading intensities

Our general price equation Eq. (A.2) is amenable to an exact treatment in the large trading intensity limit $m_t \gg J$, provided m_t does not change sign and is a sufficiently regular function of time. In such a case, the change of price is large and therefore justifies a saddle-point estimate of the integral appearing in Eq. (A.2). This leads to the following asymptotic equation of motion :

$$\mathcal{L} y_t |\dot{y}_t| \approx m_t \left[1 + D \left(3 \frac{\ddot{y}_t}{\dot{y}_t^3} - 2 \frac{\dot{m}_t}{m_t \dot{y}_t^2} \right) + O\left(\frac{J^2}{m^2} \right) \right] ; \tag{6.18}$$

see Appendix D for details of the derivation and for the next order term, of order J^2/m^2.

When m_t keeps a constant sign (say positive), the leading term of the above expansion therefore

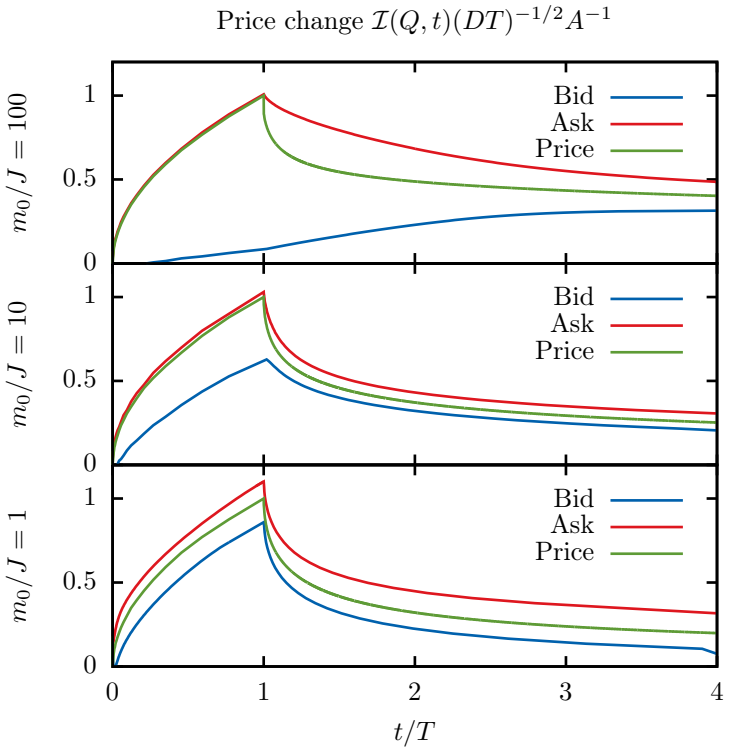

FIGURE 6.4 – Evolution in time of the bid $p_t^-(q)$ (blue line) and the ask $p_t^+(q)$ (red line) while executing a meta-order at a rate $m_0/J \in \{1, 10, 100\}$. The price p_t (green line) is also shown for comparison. The three curves correspond to the execution of a constant volume $Q = m_0 T$, while the threshold q has been set by $q = 10^{-3}\,Q$. The plot illustrates how a large execution rate m_0/J induces a locally asymmetric liquidity profile around the price, see also Fig. 6.5.

yields the following average impact trajectory :

$$y_t \approx \sqrt{\frac{2}{\mathcal{L}} \int_0^t \mathrm{d}s\, m_s}, \tag{6.19}$$

i.e. a price impact that only depends on the total traded volume, but not on the execution schedule. This is a stronger result than the one obtained above, where impact was found to be independent of the trading intensity for a uniform execution scheme. This path independence is in qualitative agreement with empirical results obtained at CFM. Using Eq. (6.18), systematic corrections to the above trajectory can be computed (see Appendix D). Perhaps surprisingly, the execution cost of a given quantity Q is found to be *independent* of the trading schedule even to first-order in J/m – see Appendix D for a proof. Exploring the optimal execution schedule within the full non-linear price equation Eq. (A.2), and comparing the results with those obtained in Curato et al. (2014), is left

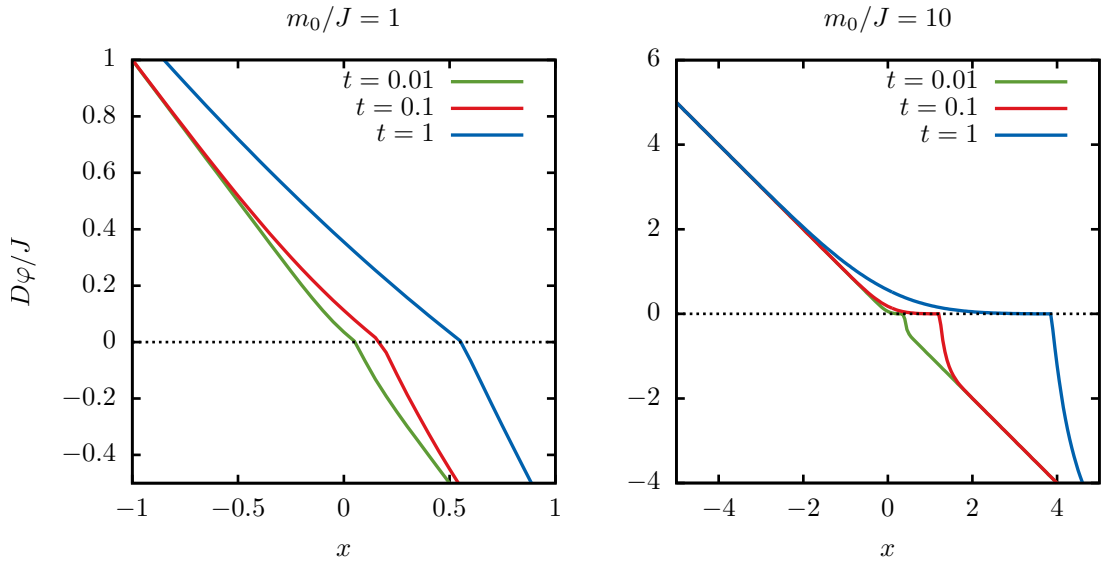

FIGURE 6.5 – Evolution of the order book shape $\varphi(x,t)$ during the execution of a meta-order at small trading rate $m_0/J = 1$ (left plot) and large trading rate $m_0/J = 10$ (right plot). The solid lines indicate the profile of the book at $t = 0.01$ (green line), $t = 0.1$ (red line) and $t = 1$ (blue line). While the displacement of the mid-price follows a square root law, the function $D\varphi(x,t)/J + x$ satisfies a scaling relation determined by the parameter m_0/J – see also Appendix C and Fig. 6.8.

for a future study.

6.9 Absence of price manipulation

We now turn to a very important issue, that of price manipulation. Although not proven to be impossible in reality, it looks highly implausible that one will ever be able to build a money machine that "mechanically" pumps money out of markets. Any viable model of price impact should therefore be such that mechanical price manipulation, leading to a positive profit after a closed trading loop, is impossible in the absence of information about future prices (Huberman and Stanzl, 2004). [15] Here, we show that the non-linear price impact model defined by Eq. (A.2) is free of price manipulation, generalizing the result of Alfonsi and Schied (2010) for the linear propagator model (see also Alfonsi and Blanc, 2016). We start by noticing that the average cost of a closed trajectory is given by :

$$\mathcal{C} = \int_0^T \mathrm{d}s\, m_s y_s, \qquad \text{with} \qquad \int_0^T \mathrm{d}s\, m_s = 0 \tag{6.20}$$

15. Note that property is highly important for practical purposes as well, since using an impact model with profitable closed trading trajectories in – say – dynamical portfolio algorithms would lead to instabilities.

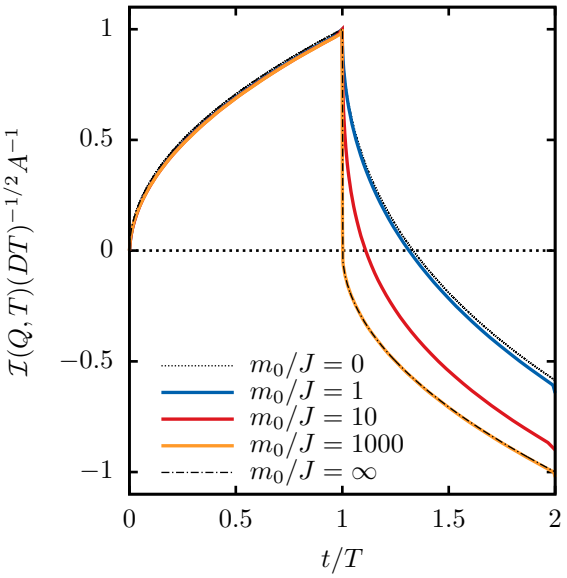

FIGURE 6.6 – Trajectory of the average price before and after a sudden switch of the sign of a meta-order. We have considered $m_t = m_0$ for $t < T$ and $m_t = -m_0$ for $t > T$, and plotted the expected price change as a function of time for different values of m_0. The curves for finite m_0 (solid lines) are also compared with the theoretical benchmark $m_0 = 0$, corresponding to the propagator model (dotted line), and to the $m_0 = \infty$ limit (dot-dashed line). We find that non-linear effects in the large m_0 regime makes to propagator approximation invalid, and increase considerably the impact of the reversal trade.

and y_s given by Eq. (A.2). The above formula simply means that the executed quantity $m_s \mathrm{d}s$ between time s and $s + \mathrm{d}s$ is at price y_s.[16] Because the initial and final positions are assumed to be zero, there is no additional marked-to-market boundary term. Using Eq. (A.2), it is not difficult to show that $\mathcal{C}$ can be identically rewritten as a quadratic form :

$$\mathcal{C} = \frac{1}{2}\int_0^T \int_0^T \mathrm{d}s\mathrm{d}s'\, m_s\, M(s,s')\, m_{s'}, \tag{6.21}$$

where $M(s,s')$ is a non-negative operator, since it can be written as a sum of "squares" $KK^\dagger$, or more precisely :

$$M(s,s') = \frac{D}{\mathcal{L}}\int_{-\infty}^{\infty} \mathrm{d}z\, z^2 \int_{-\infty}^{+\infty} \mathrm{d}u\, K_z(s,u)K_z^*(s',u), \qquad K_z(s,u) \equiv \Theta(s-u)e^{-Dz^2(s-u)+izy_s}. \tag{6.22}$$

16. The alert reader might wonder whether m_s is really the *executed* quantity, rather than the *submitted* quantity, as the definition of m_s as a flux of buy/sell orders suggest. However, within the present framework where m_s is deposited precisely at the mid-price p_t, one can check that in the limit $\kappa \to \infty$, and provided latent and real liquidity are the same close to p_t, the opposite flow of limit orders immediately adapts to absorb exactly the incoming meta-order.

This therefore proves that $\mathcal{C} \geq 0$ for *any* execution schedule, i.e. price manipulation is impossible within a LLOB (see Skachkov, 2014 for loosely related ideas). We note, *en passant*, that this proof extends to a much larger class of Markovian order book dynamics, where the reservation price of latent orders evolves, for example, according to a Lévy process (and not necessarily a diffusion, as assumed heretofore – see Appendix A).

6.10 Mechanical vs. informational impact

We now imagine that the agent executing his/her meta-order has some information about the future price, i.e. that the execution flow m_t is correlated with the future motion of the latent order book $V_{t'}$ for $t' > t$. The apparent impact of the meta-order will now contain two contributions that are, within our framework, *additive*. Assuming again, for simplicity, that $m_t = m_0$, one finds that the average price difference can be written as :

$$\langle \varepsilon \cdot (p_t - p_0) | Q \rangle = \langle \varepsilon \cdot (\widehat{p}_t - \widehat{p}_0) | Q \rangle + \langle \varepsilon \cdot (y_t - y_0) | Q \rangle, \tag{6.23}$$

where now the first term is non-zero. More explicitly, this leads to :

$$\langle p_t - p_0 \rangle = m_0 \int_0^t \mathrm{d}s \int_0^s \mathrm{d}s' C(s - s') + A(m_0)\sqrt{Dt}, \qquad (t \leq T) \tag{6.24}$$

where $C(s - s') \propto \langle V_s m_{s'} \rangle$ is a measure of the temporal correlation between meta-orders and future collective latent order moves. Let us insist that we do not assume any causality here : $C(s - s')$ can be interpreted either as the information content of the order that *predicts* future price moves (i.e. the so-called "alpha"), or as the collective reaction of the market to the order flow, i.e. the fact that agents may change their valuation as a result of the trading itself (see Bouchaud et al., 2009; Bouchaud, 2010 for a discussion of this duality).

The second term in the right hand side of Eq. (6.24) corresponds to the "mechanical" component of the impact discussed above, corresponding to the square-root impact. The first term, on the other hand, may behave very differently as a function of T. For example, if $C(s - s')$ has a range much smaller than T, the first term is expected to grow like Q and not $\sqrt{Q}$.

When $t > T$, i.e. after the end of the meta-order, the informational contribution adds to the impact decay computed above and can substantially change the apparent evolution of $\langle p_t - p_0 \rangle$. In order to fix ideas, let us assume that $C(s - s') = \Gamma\zeta e^{-\zeta(s-s')}$ (other functional forms would not change the qualitative conclusions below). The behaviour of the "total" impact for $t > T$ is then given by :

$$\mathcal{I}_{\text{tot.}}(Q, t > T) = \mathcal{I}(Q, t > T) + \Gamma Q - \frac{m_0 \Gamma}{\zeta}(1 - e^{-\zeta T})e^{-\zeta(t-T)} \underset{t\to\infty}{\longrightarrow} \Gamma Q, \tag{6.25}$$

which shows that on top of the relaxing mechanical impact (the first term), there is a growing contribution coming from the informational content of the trade (or alternatively from the collective reaction of the market to that trade) that saturates at large time to a finite value proportional to Q – see Fig. 6.7. This corresponds to a "permanent" component of impact. That the permanent component of impact should be linear in Q conforms well with the assumptions of Kyle (1985); Almgren et al. (2005). However, our calculation shows that the empirical determination of the mechanical component of impact should carefully take into account any possible information content of the analyzed trades, as well as the possible auto-correlation of the trades. This parallels the discussion offered in Gomes and Waelbroeck (2015); Brokmann et al. (2015), where attempts are made to measure the decay of mechanical impact $\mathcal{I}(Q, t > T)$ in equity markets, with the conclusion that the mechanical component of impact seems indeed to relax to zero at large times.

The possibility of generating a permanent impact by correlating the collective drift V_t with the flow of meta-orders m_s is in fact important to make our model internally consistent. Absent the permanent impact component, a random flow of meta-orders would give rise to a strongly mean-reverting contribution to the price (on top of the random walk contribution $\widehat{p}_t = \int_0^t \mathrm{d}s\, V_s$), and therefore potentially profitable mean-reversion/market making strategies. This profit can however be reduced to zero by increasing the permanent impact component (i.e. the Γ factor above), that acts as an adverse selection bias for market makers. On this point, see the discussion in Wyart et al. (2008).

6.11 Possible extensions and open problems

The LLOB framework presented above is surprisingly rich and accounts for many empirical observations, but it can only be a first approximation of a more complex reality. First, we have neglected effects that are in principle contained in Eqs. (6.1-a,b) but that disappear in the limit of slow latent order books $\nu T \ll 1$ (such that the memory time $1/\nu$ is much longer than the meta-order) and large liquidity $\mathcal{L}$, such that the meta-order only probes the linear region of the book. Re-integrating these effects perturbatively is not difficult ; for example, one finds that the impact $\mathcal{I}(Q)$ of a meta-order of size Q, executed at a constant rate, is lowered by a quantity proportional to νT when $\nu T \ll 1$. In the opposite limit $\nu T \gg 1$, one expects that $\mathcal{I}(Q)$ becomes linear in Q, since impact must become additive in that limit (see Tóth et al., 2011). Any other large scale regularisation of the model will lead to the same conclusion. One also expects that the deposition, with rate λ, of new orders behind the moving price should reduce the asymmetry of impact when the trade is reversed. We leave a more detailed calculation of these effects for later investigations.

Another important line of research is to understand the corrections to the LLOB induced by fluctuations, that are of two types : first, as discussed in section 6.3, the theory presented here only deals with the *average* order book $\varphi(x, t)$, from which the price p_t is deduced using the definition

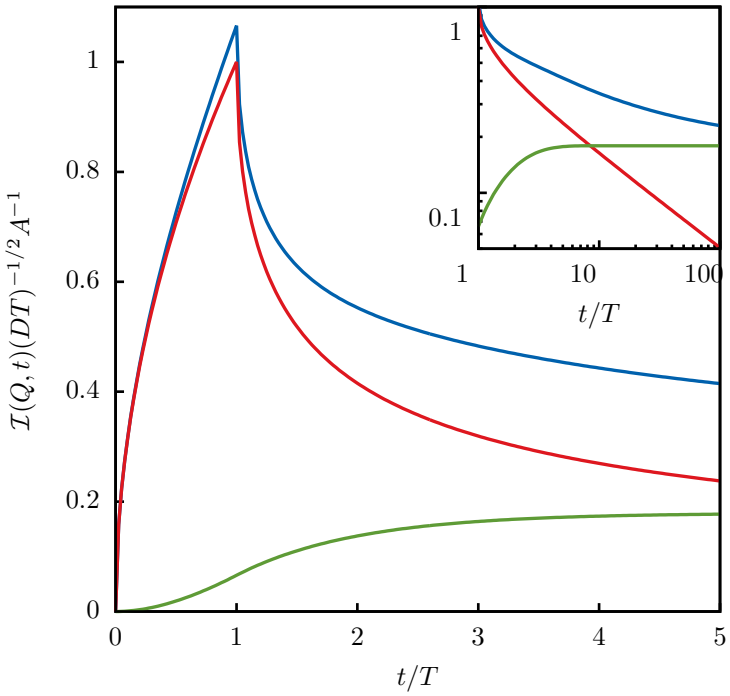

FIGURE 6.7 – The figure illustrates the relative rôles of mechanical and informational impact in determining the price trajectory during and after the execution of a meta-order. We have chosen in particular the set of parameters $D = J = \zeta = m = T = 1$ and $\Gamma = 0.1$. The figure indicates that the mechanical part of the impact is the dominating effect at small times. The permanent, informational component of the impact becomes relevant only after the slow decay of the mechanical component, as shown the inset. From a theoretical point of view, the permanent component is important since it counterbalances potential market-making/mean-reversion profits coming from the confining effect of the latent order book on the price.

$\varphi(p_t, t) = 0$, that allowed us to compute the average impact of a meta-order. However, one should rather compute the impact from the instantaneous definition of the price $p_t^{\text{inst.}}$ (that takes into account the fluctuations of the order book) and then take an average that would lead to $\mathcal{I}(Q)$. The numerical simulations shown in Mastromatteo et al. (2014b) suggest however that the approximation used here is quite accurate for long meta-orders, which is indeed expected as the difference $|p_t^{\text{inst.}} - p_t|$ becomes small compared to $\mathcal{I}(Q)$.

Second, we have assumed that the rest of the market is in its stationary state and does not contribute to the source term modelling the meta-order. One should rather posit that the flow of meta-order m_s has a random component that adds to the particular meta-order that one is particularly interested in. There again, a calculation based on the average order book is not sufficient, since

the interaction with other uncorrelated meta-orders then trivially disappears. Following Barkema et al. (1996), one finds that random fluctuations in m_s do contribute to a strongly mean-reverting term in the variogram of the price trajectory, that should be taken into account in a consistent way. Interestingly, this generic mean-reverting component leads to an excess short-term volatility that is commonly observed in financial markets ; more quantitative work on that front would therefore be worthwhile.

Finally, other extensions/modifications may be important in practice : as noted above, the cancellation rate ν is expected to increase with the intensity of meta-orders. Furthermore, the incoming flow of latent orders λ and/or the lifetime of orders $1/\nu$ can be expected to be increasing functions of the distance to the price $|x-p_t|$, i.e. better prices should attract more, and more patient buyers (or sellers), in such a way that the latent order book becomes *convex* at large distances. This would naturally explain why all impact data known to us appear to grow even slower than $\sqrt{Q}$ at large Q (Zarinelli et al., 2015; Bacry et al., 2014, and CFM, unpublished data). Another interesting path would be to allow the "drift" term V_t in Eq. (6.1-a,b) to become non-Gaussian and thereby study a cumulant expansion of the square-root law.

6.12 Conclusion

In this paper, we have proposed a minimal theory of non-linear price impact based on a linear (latent) order book approximation, inspired by diffusion-reaction models and general arguments. As emphasized in Tóth et al. (2011); Mastromatteo et al. (2014a,b), our modelling strategy does not rely on any equilibrium or fair-pricing conditions, but rather relies on purely statistical considerations. Our approach is strongly bolstered by the universality of the square-root impact law, in particular on the Bitcoin market – as recently documented in Donier and Bonart (2014) – where fair-pricing arguments are clearly unwarranted because impact is much smaller than trading fees.

Our framework allows us to compute the average price trajectory in the presence of a meta-order, that consistently generalizes previously proposed propagator models. Our central result is the dynamical Eq. (A.2), which not only reproduces the universally observed square-root impact law, but also predicts non-trivial trajectories when trading is interrupted or reversed. Quite surprisingly, we find that the short time behaviour of the free decay of impact is identical to that predicted by a propagator model, whereas the impact of a reversed trade is found to be much stronger. The latter result is in qualitative agreement with empirical observations. We have shown that our model is free of price manipulation, which makes it the first consistent, non-linear and time dependent theory of impact. Our setting also suggests how prices can be naturally decomposed into a transient "mechanical impact" component and a permanent "informational" component, as initially proposed by Almgren et al. (2005), and recently exploited in Gomes and Waelbroeck (2015); Brokmann et al. (2015) – see Section 6.10. Let us insist once again that this decomposition allowed us to construct

diffusive prices (albeit with a generic short-term mean-reverting contribution). [17]

Although our calculations are based on several approximations (restricting to a locally linear order book and neglecting fluctuations), we believe that it provides a sound starting point for further extensions where the neglected effects can be progressively reinstalled. Of particular importance is the potential feedback loop between price moves, order flow and the shape of the latent and of the revealed order books. In particular, we have assumed that latent orders instantaneously materialize in the real order book as the distance to the price gets small : any finite conversion time might however contribute to liquidity droughts, in particular when prices accelerate, leading to an unstable feedback loop. As emphasized in Lillo and Farmer (2005); Bouchaud (2011); Tóth et al. (2011) this might be triggered by the anomalous liquidity fluctuations induced by the vanishingly small liquidity in the vicinity of the price. This mechanism could explain the universal power-law distribution of returns that appear to be unrelated to exogenous news but rather to unavoidable, self-induced liquidity crises.

Acknowledgements

We warmly thank M. Abeille, R. Benichou, X. Brokmann, J. de Lataillade, C. Deremble, J. D. Farmer, J. Gatheral, J. Kockelkoren, C. A. Lehalle, Y. Lempérière, F. Lillo, E. Sérié and in particular M. Potters and B. Tóth for many discussions and collaborations on these issues. We also thank P. Blanc, N. Kornman, T. Jaisson, M. Rosenbaum and A. Tilloy for useful remarks on the manuscript. One of us (IM) benefited from the support of the "Chair Markets in Transition", under the aegis of "Louis Bachelier Finance and Sustainable Growth" laboratory, a joint initiative of École Polytechnique, Université d'Évry Val d'Essonne and Fédération Bancaire Française.

6.13 Postface (français)

Faisons abstraction de la technique [18] et concentrons-nous sur les principales leçons de cet article. Pour éviter le confinement du prix et obtenir sa diffusivité, il a été nécessaire d'introduire une synchronisation des agents à l'échelle du système [19] : nous avons donc atteint la limite de l'analo-

17. A clear theoretical justification of the square-root impact law is also important if one wants to promote the idea of impact discounted mark-to-market accounting rules, as advocated in Caccioli et al. (2012).

18. Il est vrai que celle-ci peut sembler effrayante : malheureusement, il est fréquent de devoir utiliser des techniques compliquées pour montrer des résultats intuitifs. Lorsque l'on s'intéresse à ce genre de modélisation, le mieux est souvent de faire abstraction des mathématiques pour se concentrer sur les idées, même si l'on y passe en réalité le plus clair de notre temps.

19. Sans supposer quoi que ce soit sur sa nature. Il peut par exemple s'agir de nouvelles médiatiques, ou bien d'une synchronisation « irrationnelle » due à des effets de foule. Il pourrait aussi tout simplement s'agir d'agents non infinitésimaux en taille qui influenceraient le système de manière macroscopique. La modélisation présentée ne permet toutefois pas d'inclure de tels agents, essentiellement pour une raison de difficulté technique et pour conserver des formules analytiques. Cette idée est toutefois à garder en tête, et à inclure en priorité dans un modèle de marché plus complet.

gie physique avec un système à particules de type réaction-diffusion. Cela nous rappelle que nous considérons là un système humain, qui doit être compris en tant que tel avant d'être modélisé : ici, la notion manquante était l'information, ou du moins un monde extérieur changeant – ce dont tout économiste aurait pu avoir l'intuition.

Commentons d'abord ses résultats en eux-mêmes, qui ne sont pas dénués d'intérêt : ils promeuvent en effet une toute nouvelle classe de processus mathématiques en apparence extrêmement riches et qui me semblent digne du plus grand intérêt. L'Appendice A leur sera consacrée, et mettra en évidence la dualité *processus de prix/dynamique du carnet d'ordres* qui leur fournit une interprétation élégante ainsi qu'une vraie micro-fondation. Quant aux problèmes de trading optimal qu'ils engendrent, ils seront évoqués dans l'Appendice B.

Les résultats principaux cependant sont plus indirects : en validant pour la première fois l'ensemble des critères théoriques et empiriques attendus sur l'impact (formule d'impact en racine, dépendance en le taux de participation, relaxation de l'impact, coût d'un aller-retour, absence de manipulation...), cet article valide surtout une dynamique de l'offre et de la demande sous-jacentes – souvenons-nous d'une des raisons principales de l'étude de l'impact : il s'agit d'une sonde dynamique de l'état de l'offre et de la demande ! En particulier, il met en lumière le rôle crucial de la statistique et de l'hétérogénéité dans la modélisation écono-financière. Et même si cette hétérogénéité n'a pu être traitée ici que dans le cas d'agents infinitésimaux, rien n'empêche les modèles futurs d'inclure des agents macroscopiques pour compléter le tableau, comme précisé dans la note de bas de page précédente 19 : ce que cet article valide, plus qu'un modèle, c'est une approche, une hypothèse, un comportement sous-jacent. Ne perdons donc pas de vue notre objectif : l'article qui suit va tenter de prendre un peu de hauteur, et se concentrer sur cette dynamique de l'offre et la demande mise en évidence, dans un cadre plus général que celui du « carnet d'ordres limites en temps continu ».

Appendix A : Derivation of the drift/diffusion term

In order to give more flesh to the microscopic assumptions underlying the drift/diffusion equation written in Eqs. (6.1-a,b), let us assume first that each agent contributes to a negligible fraction of the latent order book, which is probably a good approximation for deep liquid markets. A model for thin markets, where some participants contribute to a substantial fraction of the liquidity, is discussed below, but leads to a very similar final result.

Between t and $t + \delta t$, each agent i revises its reservation price p_i to $p_i + \beta_i \xi_t + \eta_{i,t}$, where ξ_t is common to all i representing some public information (news, but also the price change itself or the order flow, etc.) and $\beta_i > 0$ is the sensitivity of agent i to the news, which we imagine to be a random variable from agent to agent, with a pdf $\Pi(\beta)$ mean normalized to $[\beta_i]_i = 1$. $[[...]_i$ represents a cross-sectional average over agents.] Some agents may over-react, others under-react ; β_i might in fact be itself time dependent, but we assume that the *distribution* of β's is stationary. The

completely idiosyncratic contribution $\eta_{i,t}$ is an independent random variable both across different agents and in time, with distribution $R(\eta)$ of mean zero and rms Σ. We assume that within each price interval $x, x + \mathrm{d}x$ lie latent orders from a large number of agents. The density of latent orders $\rho(x,t)$ therefore evolves according to the following Master equation :

$$\rho(x, t+\delta t) = \int_0^\infty \mathrm{d}\beta \Pi(\beta) \int_{-\infty}^\infty \mathrm{d}\eta R(\eta) \int \mathrm{d}y \rho(y,t) \delta(x - y - \beta\xi_t - \eta), \tag{6.26}$$

or :

$$\rho(x, t+\delta t) = \int_0^\infty \mathrm{d}\beta \Pi(\beta) \int_{-\infty}^\infty \mathrm{d}\eta R(\eta) \rho(x - \beta\xi_t - \eta, t). \tag{6.27}$$

Assuming that the price revisions $\beta\xi_t + \eta$ over a small time interval δt are small enough, a second-order expansion Kramers-Moyal of the above equation leads to (see Gardiner, 2009 for an in-depth discussion of this procedure) :

$$\rho(x, t+\delta t) - \rho(x,t) = -\xi_t \rho'(x,t) + \frac{1}{2}\left([\beta^2]\xi_t^2 + \Sigma^2\right)\rho''(x,t) + \dots \tag{6.28}$$

At this stage, one can either assume that formally $\xi_t = V_t \delta t$ and $\Sigma^2 = 2D_0\delta t$ in which case the continuous time limit reads :

$$\frac{\partial \rho(x,t)}{\partial t} = -V_t \frac{\partial \rho(x,t)}{\partial x} + D_0 \frac{\partial^2 \rho(x,t)}{\partial x^2} \tag{6.29}$$

or that $\xi_t = V_t\sqrt{\delta t}$, where V_t is now a Gaussian white noise of variance σ^2, and again $\Sigma^2 = 2D\delta t$, in which case the continuous time limit should be written as :

$$\mathrm{d}\rho(x,t) = -\mathrm{d}W_t \frac{\partial \rho(x,t)}{\partial x} + D_1 \mathrm{d}t \frac{\partial^2 \rho(x,t)}{\partial x^2} \tag{6.30}$$

with $\mathrm{d}W_t$ a Wiener noise and $D_1 \equiv D_0 + [\beta^2]\sigma^2/2$, to wit, the diffusion constant involves both the idiosyncratic component and the dispersion of reaction to random information. This is the interpretation we will mostly follow in the present paper. A careful derivation of the corresponding equation in the reference frame of the price $\widehat{p}_t = \int_0^t \mathrm{d}W_s$ finally gives the diffusion part of Eq. (6.3) in the main text :

$$\frac{\partial \rho(y,t)}{\partial t} = D\frac{\partial^2 \rho(y,t)}{\partial y^2}; \qquad D \equiv D_0 + \frac{\sigma^2}{2}\int \mathrm{d}\beta \Pi(\beta)(\beta - 1)^2; \tag{6.31}$$

i.e. only the dispersion of reaction $\beta - 1$ can contribute to the diffusion term, as expected.

Another interpretation of this last equation is to imagine that between t and $t + \mathrm{d}t$, a fraction $\phi \in [0,1]$ (possibly time dependent) of agents change their price estimate by an amount $\mathrm{d}W_t$, with

no other idiosyncratic component. This leads to :

$$\mathrm{d}\rho(x,t) = \phi\left[\rho(x-\mathrm{d}W_t,t)-\rho(x,t)\right] = -\phi\mathrm{d}W_t\rho'(x,t) + \frac{1}{2}\phi\sigma^2\mathrm{d}t\rho''(x,t), \tag{6.32}$$

that essentially corresponds to the case above with $\Pi(\beta) = (1-\phi)\delta(\beta) + \phi\delta(\beta-1)$. In the price reference frame $\widehat{p}_t = \int_0^t \phi\mathrm{d}W_s$, one finds Eq. (6.31) with $D \equiv \phi(1-\phi)\sigma^2/2$. Note that, clearly, these price revisions must by themselves induce transactions whenever $0 < \phi < 1$.

Interestingly, this last derivation may also be interpreted as the order book dynamics with macroscopic agents : The fraction ϕ introduced above would then correspond to the relative size of the agent with respect to the market. We notice that in this case, there is no need for synchronization between agents via the common component ξ_t, since single agents may have non-negligible effects on the order book and on the price. More generally, a drift component will be obtained as soon as a non-negligible fraction of the latent volume moves in the same direction, and a diffusion-style component will be obtained as soon as these moves are heterogeneous among agents.

If price revisions cannot be considered as small, the resulting evolution of $\rho(x,t)$ should include jumps in the continuous time limit, i.e. one would find an integro-differential equation rather than a partial differential equation for $\rho(x,t)$. However, if the jump process is homogeneous in space, one can diagonalize the evolution operator in Fourier space. This allows one to show that price manipulation is impossible in that case as well.

Appendix B : A generically linear latent order book

Let us consider the case where the deposition flow is not constant. This leads to the following equation for the stationary state of the latent order book :

$$D\frac{\partial^2\varphi_{\mathrm{st.}}(y)}{\partial y^2} - \nu\varphi_{\mathrm{st.}}(y) + \lambda\left(\Theta(y)-\Theta(-y)\right) = 0, \tag{6.33}$$

with $\varphi_{\mathrm{st.}}(y) = -\varphi_{\mathrm{st.}}(-y)$ (and in particular the market clearing condition $\varphi_{\mathrm{st.}}(y=0)=0$).

Let us assume that $\Theta(y)-\Theta(-y)$ behaves, for $y\to\infty$, as a constant that we can set to unity. The solution $\varphi_{\mathrm{st.}}(y)$ then converges to λ/ν for large y, so we set :

$$\varphi_{\mathrm{st.}}(y) = \frac{\lambda}{\nu} + \Psi(y), \tag{6.34}$$

where $\Psi(y) = 1-\Theta(y)+\Theta(-y)$ with $\Psi(y\to\infty)\to 0$, and :

$$D\frac{\partial^2\Psi(y)}{\partial y^2} - \nu\Psi(y) = \lambda\Xi(y), \tag{6.35}$$

where $\Xi(y\to\infty)\to 0$. The boundary condition on $\Psi(y)$ at large y means that we can look at a

solution of the form :

$$\Psi(y) = \psi(y) e^{-\sqrt{\nu/D}\, y}, \tag{6.36}$$

so that :

$$D\psi''(y) - 2\sqrt{\nu D}\psi'(y) = \lambda \Xi(y) e^{\sqrt{\nu/D}\, y}. \tag{6.37}$$

Finally, one finds :

$$\varphi_{\text{st.}}(y) = \frac{\lambda}{\nu}\left[1 - e^{-\sqrt{\nu/D}\, y}\right] + \frac{\lambda}{D} e^{-\sqrt{\nu/D}\, y} \int_0^y \mathrm{d}y' e^{2\sqrt{\nu/D}\, y'} \int_{y'}^{\infty} \mathrm{d}y'' e^{-\sqrt{\nu/D}\, y''} \Xi(y''). \tag{6.38}$$

The solution given in the main text corresponds to $\Xi \equiv 0$, so that only the first term survives. From the above explicit form, one sees that provided the integral $\int_0^\infty \mathrm{d}y'' e^{-\sqrt{\nu/D} y''} \Xi(y'')$ is finite, the behaviour of $\varphi_{\text{st.}}(y)$ is *always* linear in the vicinity of $y = 0$. Only a highly singular the deposition rate, diverging faster than $1/y$ when $y \to 0$, would jeopardize the local linearity of the latent order book (see also Bouchaud et al., 2002, where this property was first discussed).

Appendix C : Shape of the order book during constant rate execution and initial relaxation of impact

When the trading rate is a constant (equal to m_0), one can exhibit an *exact* scaling solution of the time dependent order book of the form $\varphi(x,t) = m_0\sqrt{t/D}\, F(\frac{x}{\sqrt{Dt}})$, where F is the solution of :

$$2F''(u) + uF'(u) - F(u) = -2\delta(u - A). \tag{6.39}$$

As we show below, this equation can be solved and gives the exact shape of the book at $t = T$, from which the initial relaxation (after trading has stopped) can be deduced.

Writing $F = uG$ in the above equation, one finds a first order linear equation for $H = G'$:

$$H'(u) + \left(\frac{u}{2} + \frac{2}{u}\right) H(u) = \frac{1}{u}\delta(u - A) \tag{6.40}$$

which is easily solved as :

$$H(u) = \frac{H_0}{u^2} e^{-u^2/4}, \quad (u < A); \qquad H(u) = \frac{H_0 - A e^{A^2/4}}{u^2} e^{-u^2/4}, \quad (u > A). \tag{6.41}$$

There are two boundary conditions that are useful. One is the very definition of the price position, $x = A\sqrt{Dt}$ or $A = u$, for which $\phi(x,t) = x$ or $F(A) = AJ/m_0$. The second remark is that when

$u = 0$, the integral defining F can be computed, leading to

$$F(0) = \frac{A}{4\sqrt{\pi}} e^{A^2/4} \int_{A^2/4}^{\infty} \frac{\mathrm{d}v}{v^{3/2}} e^{-v}. \tag{6.42}$$

This allows one to fix H_0 since $G'(u) = F'(0)/u - F(0)/u^2 \approx -F(0)/u^2$ when $u \to 0$, to be compared with $H(u) \approx H_0/u^2$ in the same limit. Hence $H_0 = -F(0)$.

The final solution for $F(u)$ is easily obtained from integrating H and multiplying by u (see Fig. 6.8). Using $G(A) = J/m_0$, one finds :

$$G(u) = F(0) \int_u^A \frac{\mathrm{d}v}{v^2} e^{-v^2/4} + J/m_0, \qquad (u \leq A) \tag{6.43}$$

and

$$G(u) = -(F(0) + Ae^{A^2/4}) \int_A^u \frac{\mathrm{d}v}{v^2} e^{-v^2/4} + J/m_0, \qquad (u \geq A) \tag{6.44}$$

Of special interest is the slope of F for $u = A^{\pm}$; with $F' = uH(u) + G(u)$ one finds :

$$F'(A^-) = J/m_0 - \frac{F(0)}{A} e^{-A^2/4}; \qquad F'(A^+) = J/m_0 - \frac{F(0)}{A} e^{-A^2/4} - 1. \tag{6.45}$$

Now, shortly after the meta-order has stopped, one can look at the solution of the diffusion equation in the vicinity of the final price $p_T = A\sqrt{DT}$, using a piece-wise linear function for the initial condition, with slopes given by $F'(A^{\pm})$. The solution then reads, with $t - T = \Delta$ small :

$$\varphi(x,t) = m_0\sqrt{T\Delta} \left[F'(A^-)z + (F'(A^+) - F'(A^-)) \int_z^{\infty} \mathrm{d}u \frac{(u-z)}{\sqrt{4\pi}} e^{-u^2/4} \right], \tag{6.46}$$

with $z = (p_T - x)/\sqrt{D\Delta}$. The position of the price is given by $p_t = p_T - \sqrt{D\Delta} z^*$, with z^* such that :

$$F'(A^-)z^* + (F'(A^+) - F'(A^-)) \int_{z^*}^{\infty} \mathrm{d}u \, \frac{(u-z^*)}{\sqrt{4\pi}} e^{-u^2/4} = 0. \tag{6.47}$$

Using the expression for $F(0)$ and the result above for $F'(A^{\pm})$, this equation simplifies to :

$$z^* \int_{A^2/4}^{\infty} \frac{\mathrm{d}v}{v^{3/2}} e^{-v} = 2 \int_{z^*}^{\infty} \mathrm{d}u \, (u - z^*) e^{-u^2/4}. \tag{6.48}$$

Changing variables in the RHS from u to $v = u^2/4$, and integrating by parts, one finds :

$$z^* \int_{A^2/4}^{\infty} \frac{\mathrm{d}v}{v^{3/2}} e^{-v} = z^* \int_{z^{*2}/4}^{\infty} \frac{\mathrm{d}v}{v^{3/2}} e^{-v}, \tag{6.49}$$

which leads to $z^* = A$ for all m_0/J. [The other solution, $z^* = 0$, is spurious].

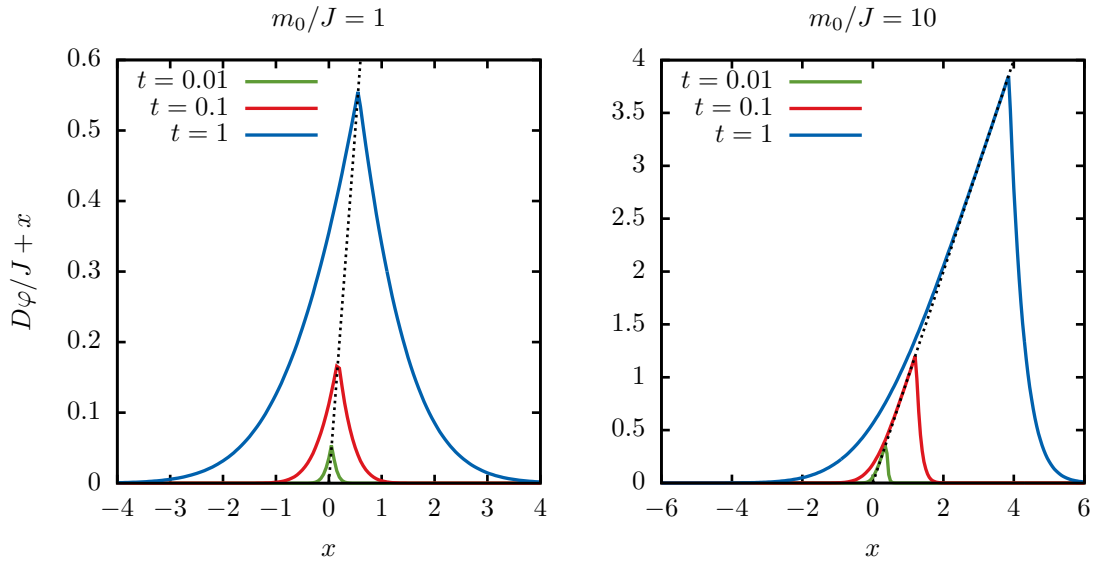

FIGURE 6.8 – Evolution of the average order book $\varphi(x,t)$, represented for two different values of the perturbation parameter m_0/J. The curves in different colors are snapshots taken at different times of the difference between the perturbed and the unperturbed average of the book. The scaling form of the book $\varphi(x,t) = m_0\sqrt{t/D}F(\frac{x}{\sqrt{Dt}})$ is clear from the plots, which illustrate how a flat region of the book is formed at large values of m_0/J.

Hence, the initial stage of the impact relaxation can be written in a super-universal way :

$$p_t \underset{t\to T^+}{\approx} p_T\left[1-\sqrt{\frac{t-T}{T}}\right], \tag{6.50}$$

i.e. exactly the result from the propagator model, even in the non-linear regime !

Appendix D : A saddle point approximation for large trading rates

In this Appendix we develop a systematic procedure in order to solve perturbatively Eq. (A.2). The first step in order to find a solution is to introduce an expansion parameter $\epsilon \ll 1$, which we use in order to control the amplitude of the trading rate through $m_t \to m_t\epsilon^{-1}$. Such substitution implies a scaling of the solution of the form $y_t \to y_t\epsilon^{-1/2}$, leading to an equation for the price of the form :

$$\mathcal{L}y_t = \int_0^t \frac{\mathrm{d}s\, m_s}{\sqrt{4\pi D\epsilon(t-s)}} e^{-\frac{(y_t-y_s)^2}{4D\epsilon(t-s)}}. \tag{6.51}$$

which is equivalent to the one which one would have by leaving invariant m_t and by performing the substitutions $D \to D\epsilon$ and $J \to J\epsilon$. Hence, the large trading regime is equivalent to the one of slow

diffusion.

In this case, it is evident that the integral is dominated by times s close to t, which suggests to Taylor expand both the trading rate m_s and the price y_s around $s = t$, so to insert the resulting series in the integral appearing in Eq. (A.2). The dominating term results

$$m_t \int_0^\infty \mathrm{d}u \frac{1}{\sqrt{4\pi D u \epsilon}} e^{-\dot{y}_t^2 \frac{u}{4D\epsilon}} = m_t |\dot{y}_t|^{-1} \,. \tag{6.52}$$

The successive corrections to above result can be computed systematically, as they involve developing the square and the exponential function in the Gaussian term in Eq. (A.2). In particular, by exploiting the identity

$$\int_0^\infty \mathrm{d}u \, e^{-z^2 u} u^\alpha = \Gamma(1+\alpha)|z|^{-2(1+\alpha)} \tag{6.53}$$

it is possible to derive the expansion

$$\begin{aligned}
\mathcal{L} y_t |\dot{y}_t| &= m_t \Bigg[1 + (D\epsilon) \left(\frac{3\ddot{y}_t}{\dot{y}_t^3} - \frac{2\dot{m}_t}{m_t \dot{y}_t^2} \right) \\
&+ (D\epsilon)^2 \left(\frac{6\ddot{m}_t \dot{y}_t^2 - 30 \dot{m}_t \ddot{y}_t \dot{y}_t - 10 m_t \dddot{y}_t \dot{y}_t + 45 m_t \ddot{y}_t^2}{m_t \dot{y}_t^6} \right) \\
&+ 5(D\epsilon)^3 \left(\frac{-4\dddot{m}_t \dot{y}_t^3 + 42 \ddot{m}_t \ddot{y}_t \dot{y}_t^2 + 28 \dot{m}_t \dddot{y}_t \dot{y}_t^2 + 7 m_t \ddddot{y}_t \dot{y}_t^2}{m_t \dot{y}_t^9} \right) \\
&+ 5(D\epsilon)^3 \left(\frac{-168 \dot{m}_t \ddot{y}_t^2 \dot{y}_t - 112 m_t \ddot{y}_t \dddot{y}_t \dot{y}_t + 252 m_t \ddot{y}_t^3}{m_t \dot{y}_t^9} \right) \\
&+ O(\epsilon^4) \Bigg] \,,
\end{aligned} \tag{6.54}$$

whose first-order terms match the form reported in Eq. (6.18). Each of the contributions of order ϵ^n can be seen equivalently either as suppressed by the small value of the diffusion constant diffusion (through a D^n factor) or by the large value trading rate (through a factor of the order of m_t^{-n}).

Finally, note that the implicit equation above needs to be inverted in order to obtain a relation yielding y_t as a function of m_t. This is possible by using Eq. (6.54) as an iterative scheme for y_t, which allows to calculate

$$\begin{aligned}
\mathcal{L} y_t |\dot{y}_t| &= m_t \\
&+ (J\epsilon) \left(-3 + \frac{2 Q_t \dot{m}_t}{m_t^2} \right) \\
&+ (J\epsilon)^2 \left(-\frac{12 Q_t \dot{m}_t}{m_t^3} - \frac{6\, s\, \dot{m}_t}{m_t^2} + \frac{4\dot{m}_t}{m_t^2} \int \mathrm{d}s \frac{Q_s \dot{m}_s}{m_s^2} + \frac{16 Q_t^2 \dot{m}_t{}^2}{m_t^5} - \frac{4 Q_t^2 \ddot{m}_t}{m_t^4} \right) \\
&+ O(\epsilon^3) \,,
\end{aligned} \tag{6.55}$$

where $Q_t = \int_0^t \mathrm{d}s \, m_s$.

As a simple application of the above formula, consider the case where $m_t \geq 0, \forall t \in [0, T]$. In

this case, $\dot{y}_t$ is also non negative and we can remove the absolute value in the above equation. To order ϵ, the solution of the above equation is (assuming $y_0 = 0$) :

$$\frac{1}{2}\mathcal{L}y_t^2 = Q_t + J\epsilon\,(-t - \frac{2Q_t}{m_t}), \tag{6.56}$$

where we have used integration by parts. Now, the cost $\mathcal{C}(Q)$ associated to buying a total quantity Q in time T is given by :

$$\mathcal{C}(Q) = \int_0^T \mathrm{d}s\, m_s y_s, \qquad Q = \int_0^T \mathrm{d}s\, m_s. \tag{6.57}$$

Therefore, to order ϵ :

$$\mathcal{C}(Q) = \sqrt{\frac{2}{\mathcal{L}}}\int_0^T \mathrm{d}s\, m_s\sqrt{Q_s} - \frac{J\epsilon}{\sqrt{2\mathcal{L}}}\int_0^T \mathrm{d}s\,\left[2\sqrt{Q_s} + \frac{sm_s}{\sqrt{Q_s}}\right]. \tag{6.58}$$

After integrating by parts the last term, one finally finds that the impact cost is, to order ϵ, *independent* of the trading schedule m_s, and given by :

$$\mathcal{C}(Q) = \frac{2}{3}\sqrt{\frac{2}{\mathcal{L}}}Q^{3/2}\left[1 - \frac{3J\epsilon T}{2Q}\right]. \tag{6.59}$$

The correction term is negative, as expected since a slower trading speed leaves time for the opposing liquidity to diffuse towards the traded price.

Chapitre 7

Théorie dynamique de l'offre et de la demande

From
From Walras' auctioneer to continuous time double auctions :
A general dynamic theory of supply and demand
with Jean-Philippe Bouchaud
(Donier and Bouchaud, 2015b)

"The paper posits a theory of continuous trading with marginal supply and demand evolving over time. Unfortunately, the marginal supply and demand curves are simply assumed without any underlying economics."
[Anonymous referee]

7.1 Préface (français)

L'article qui suit est sans doute l'article central de cette thèse, non pas par son innovation empirique ou technique mais par la mise en perspective qu'il propose des résultats obtenus dans les deux articles précédents (ainsi que de leurs prédécesseurs). L'impact nous a permis de sonder l'offre et la demande et a mis en évidence ses propriétés dynamiques : faisons-en maintenant abstraction, et essayons plutôt de comprendre les implications de nos travaux sur la modélisation de l'offre et la demande en économie – en la comparant par exemple à la vision Walrasienne classique mentionnée en introduction.

Il m'a semblé jusqu'ici que les différentes disciplines qui s'intéressent à la modélisation de l'offre et de la demande (micro-économie, économie financière, mathématiques financières, éconophysique [1]...)

1. Certains préfèrent l'appellation de *data-driven modelling*.

interagissaient de manière plutôt limitée, à la fois physiquement et dans la littérature : chacun a pourtant un point de vue valable, mais qui gagnerait beaucoup à emprunter quelques concepts à ses voisins – selon moi. Un de objectifs principaux de l'article qui suit a justement été de formuler les conclusions de ces cent premières pages d'une manière compréhensible par chacun [2], et de revisiter à leur lumière les hypothèses, visions et modèles classiques de chacune de ces disciplines. Il vise également à poser un cadre à l'intérieur duquel chaque discipline pourra travailler, des économistes adeptes de la rationalité aux physiciens qui préfèrent donner à leurs agents des règles de décision heuristiques, en mettant en lumière un certain nombre de propriétés universelles – parfois importantes, et non triviales ! – que les deux approches retrouveront [3].

La citation en tête de ce chapitre, rapport de *referee* anecdotique reçu dans le cadre d'une grande conférence d'économie, est toutefois lourde de sens, et preuve de l'attachement des communautés à leurs outils de prédilection – nonobstant les conclusions de l'article qui démontre justement que dans ce cas précis on peut passer outre cet attachement. Le chemin à parcourir pour unifier les forces semble long : mais l'objectif me semble valoir l'effort.

Abstract : In standard Walrasian auctions, the price of a good is defined as the point where the supply and demand curves intersect. Since both curves are generically regular, the response to small perturbations is linearly small. However, a crucial ingredient is absent of the theory, namely transactions themselves. What happens after they occur ? To answer the question, we develop a dynamic theory for supply and demand based on agents with heterogeneous beliefs. When the inter-auction time is infinitely long, the Walrasian mechanism is recovered. When transactions are allowed to happen in continuous time, a peculiar property emerges : close to the price, supply and demand vanish quadratically, which we empirically confirm on the Bitcoin. This explains why price impact in financial markets is universally observed to behave as the square root of the excess volume. The consequences are important, as they imply that the very fact of clearing the market makes prices hypersensitive to small fluctuations.

7.2 Introduction

One of the most time-worn statement of economic science is that "prices are such that supply matches demand". In order to explain how this really comes about, one usually invokes a Walras auctioneer, who attempts to measure the supply and demand curves $S(p)$ and $D(p)$, that give the total amount of supply/demand for a given good (or asset), would the price be set to p. The equilibrium price p^* is then such that $D(p^*) = S(p^*)$, which maximizes the amount of good exchanged

2. Le résultat n'est qu'imparfait : mais il est difficile de s'arracher totalement à son environnement. La tâche est d'ailleurs peut-être vaine, et la solution d'écrire plusieurs article, chacun dans le langage d'une discipline particulière.

3. Ce qu'ils trouveront de différent, en revanche, sera la valeur exacte des paramètres, qui dépendra des ingrédients qu'ils choisissent en entrée.

among agents, given the set of preferences corresponding to the current supply and demand curves Walras (1954). In reality, the full knowledge of $S(p)$ and $D(p)$ is problematic, and Walras envisioned his famous *tâtonnement* process as a mean to observe the supply/demand curves. However, there is a whole aspect of the dynamics of markets that is totally absent in Walras' framework. While it describes how a pre-existing supply and demand would result in a clearing price, it does not tell us anything about what happens *after* the transaction has taken place. In this sense, the Walrasian price is of very limited scope, since the theory ceases to apply as soon as the price is discovered.

A practical solution to match supply and demand is the so-called "order book" Harris et al. (1990); Glosten (1994), where each agent posts the quantities s/he is willing to buy or sell as a function of the price p. $S(p)$ (resp. $D(p)$) is then the sum of all sell (buy) quantities posted at or above (below) price p. At each time step, the auctioneer can then clear the market by finding the (unique) price such that $D(p^*) = S(p^*)$. This is in fact how most financial markets worked before the advent of electronic matching engines, when market makers played the role of "active" Walrasian auctioneers, in the sense that they would themselves contribute to the order book as to insure orderly trading and stable prices Glosten and Milgrom (1985); Madhavan (2000).

Although close to Walras' idealization, order book based auctions are still confronted with a fundamental problem : agents do not necessarily reveal their intentions by placing visible orders, for fear of giving away information to the rest of the market – among other reasons Handa and Schwartz (1996); Bongiovanni et al. (2006). It is plausible that only agents with the most urgent need to buy or to sell reveal their intentions. Only close to the transaction price is the order book expected to reveal the true underlying supply and demand curves $S(p)$ and $D(p)$, where they however get intertwined with the orders of market makers/high frequency traders who play strategic "hide and seek" games Handa et al. (2003); Roşu (2006); Foucault et al. (2013); Bouchaud et al. (2009). The visible order book is a sort of *Potemkin village* that reveals only very little about the true underlying supply and demand [4] and whose features strongly depend to the precise design of the market (time priority, pro-rata matching, small or large tick, presence of hidden orders, etc. – see e.g. Kockelkoren (2010)). A direct empirical observations of the dynamics of the full supply and demand curves $S(p)$ and $D(p)$ is therefore difficult (except in particular markets such as Bitcoin, see below and Donier and Bouchaud (2015a)). But since the dynamics of prices is essentially governed by that of supply and demand, we need a plausible theoretical framework to model the (unobservable) evolution of the time dependent curves $S(p,t)$ and $D(p,t)$, where t is time, to account for the (observable) evolution of prices. This would allow one to construct a "Walrasian" description of market dynamics, offering a much deeper level of understanding than simply postulating ad-hoc stochastic models for prices, such as the standard (geometric) Brownian motion Bachelier (1900); Black and Scholes (1973).

4. The overwhelming activity market makers/HFT in the visible order book is a distraction from our story here : since their position quickly mean-reverts around zero, they chiefly act as intermediaries (as they should) and only weakly contribute to the supply and demand curves. See the detailed discussion in Donier et al. (2015) and below.

There are indeed many questions, some of utmost fundamental and practical importance, which cannot be addressed within these stochastic models and require the knowledge of the underlying supply and demand structure and dynamics. One of them is *price impact* Bouchaud (2010), i.e. how much does an additional unconditional buy/sell quantity Q move the price up/down ? This is important both for practitioners who want to estimate the costs associated to the impact of their trading strategies Almgren and Chriss (2001), and for regulators who want to understand the stability of markets and the price sensitivity to large "freak" orders (see e.g. Donier and Bouchaud (2015a)). It is also of interest for the general understanding of price discovery and market efficiency : how much noise do "noise traders" Kyle (1985) introduce in markets through their impact on prices ? How relevant is marked to market accounting ? Amihud and Mendelson (1986); Caccioli et al. (2012), etc.

In a Walrasian context, the impact $\mathcal{I}$ of a small buy quantity Q is easily shown to be linear in Q, simply because the slopes of the supply and demand curves around the price p^* (that would prevail for $Q = 0$) are generically non zero. More precisely, writing that $S(p_Q) = D(p_Q) + Q$ and Taylor expanding $S(p)$ and $D(p)$ around p^* to first order in Q, one readily obtains : [5]

$$S(p^*) + (p_Q - p^*)\partial_p S(p^*) = D(p^*) + (p_Q - p^*)\partial_p D(p^*) + Q \qquad \Rightarrow \qquad \mathcal{I}(Q) \equiv p_Q - p^* = \lambda Q \quad (7.1)$$

with $\lambda^{-1} = \partial_p S(p^*) - \partial_p D(p^*) > 0$, since one expects $S(p)$ to be a strictly increasing function of p and $D(p)$ a strictly decreasing function of p (see Figure 1). Whenever the derivatives of the supply and demand curves do not simultaneously vanish at p^*, the price response to a perturbation must be *linear*. This intuitive result can also be justified using much more elaborate arguments, such as provided by the Kyle model Kyle (1985), where noise traders, market makers and an informed trader interact in the market place. As shown by Kyle, the optimal strategy of market makers is to shift the price linearly in the market imbalance, with the coefficient λ ("Kyle's lambda") proportional to the volatility of the asset σ and inversely proportional to the typical volume V traded by the whole market.

Until very recently, the above linear relation between order imbalance and price changes was taken for granted in most academic papers. Remarkably, a series of independent empirical studies of the impact of proprietary orders in financial markets published since the mid-nineties suggests otherwise Torre and Ferrari (1997); Grinold and Kahn (2000); Almgren et al. (2005); Moro et al. (2009); Tóth et al. (2011); Mastromatteo et al. (2014a); Gomes and Waelbroeck (2015); Bershova and Rakhlin (2013); Brokmann et al. (2015); Bacry et al. (2014); Zarinelli et al. (2015); Donier and Bonart (2014). All these studies report a *strongly concave* price impact, even in the regime where

5. The notation ∂_p means that one takes the derivative with respect to the price p. More generally, ∂_x means the derivative with respect to any variable x.

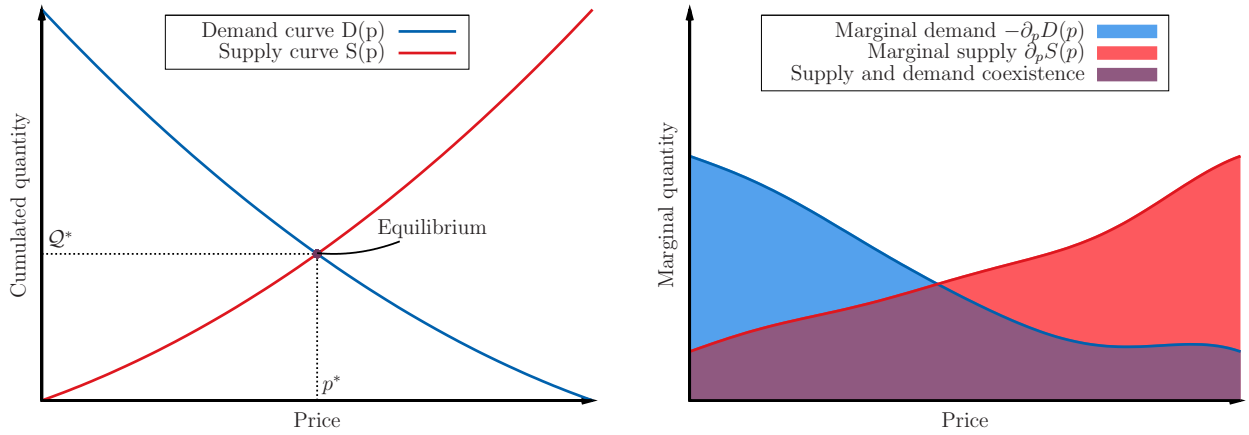

FIGURE 7.1 – Left : Illustration of the supply and demand curve (SD), and the resulting price according to Walras' law. Right : Marginal supply and demand curves (MSD) corresponding to the figure on the left.

Q/V is very small (say between 10^{-4} and 0.1). In fact, a simple square-root law

$$\mathcal{I}(Q) = Y\sigma\sqrt{\frac{Q}{V}}$$

where Y is a constant of order unity, accounts surprisingly well for a number of observations, quite independently of the type of markets (stocks, futures, FX, options,..), geographical zones, epochs (pre-2005, before the advent of massive HFT, or post-2005), trading style, etc. The empirical evidence is now so compelling – see in particular the Bitcoin data in Donier and Bonart (2014) – that it is difficult to avoid looking for a consistent theoretical explanation for such a universal non-linear impact law.

There are actually two reasons to believe that usual equilibrium models are not relevant to explain the observed square-root impact. The first one is about orders of magnitude in the Bitcoin case, where the square-root impact is perfectly obeyed for price changes 30 times smaller than transaction fees themselves, and 300 times smaller than the daily volatility – see Donier and Bonart (2014). Imagining that the typical amateur trader on Bitcoin is able to optimize anything with this level of precision seems a total utopia. The second reason is that a square-root impact formally corresponds to $\partial_Q \mathcal{I}(0) = \lambda \to \infty$, i.e. a situation where $\partial_p S(p^*)$ and $\partial_p D(p^*)$ are both zero. This is a fact that equilibrium models (as in Kyle (1985)) cannot reproduce, as they would always predict a linear impact for small quantities.

A more likely possibility is that this square-root law is not willingly enforced by any market participant but is rather an emergent property. Stepping away from classical ideas, a detailed scenario for the divergence of Kyle's λ was proposed in Tóth et al. (2011). The main assumption in that paper is that the shape of the supply and demand curves is primarily the result of the inter-

action between order flow and past transactions themselves. This is in line with the idea that large population of interacting agents, each using heuristic decision rules, can lead to universal emergent behaviour (see e.g. Hommes (2006) for a review of heterogeneous agent models in economics and finance Gualdi et al. (2015b) for a recent didactic discussion). This scenario in fact emphasizes the *transient* aspects of market dynamics, which are usually discarded in equilibrium models. The resulting detailed theory, elaborated in Mastromatteo et al. (2014b); Donier et al. (2015), accounts well for the above square-root impact law and for many other empirical observations. In fact, the model presented in Tóth et al. (2011); Donier et al. (2015) and in the present paper is not incompatible with game theoretic approaches as it can be rephrased in the language of the mean field games introduced in Lasry and Lions (2007), see below, Section 7.4.3.

The aim of the present paper is to revisit and extend the Walrasian theory by proposing a dynamic theory of supply and demand in the light of these recent results. After a brief review of the literature in Section 7.3, we propose in Section 7.4 a dynamical theory for the evolution of the supply and demand curves $S(p,t)$ and $D(p,t)$, including transactions. Our theory only relies on weak, general hypotheses about the behaviour of agents, in particular the assumption of heterogeneous beliefs in a very large population. The Walrasian auctions setting can be seen as a limiting case of our theory, corresponding to an infinitely long time between auctions. We then show (Section 7.5) that as soon as the inter-auction time τ is finite, the impact of small volumes is linear, but with a coefficient λ that diverges as $1/\sqrt{\tau}$ in the $\tau \to 0$ limit (corresponding to continuous time double auctions) where we recover an exact square-root impact. This reflects the fact that the supply and demand curves both vanish quadratically close to the current price p^*, a property that we validate empirically using Bitcoin data. For small, but finite τ, impact is linear for very small Qs and becomes square-root beyond a crossover value $Q^*(\tau) \propto \tau$ when $\tau \to 0$. Finally, in Sections 10,E.7 we discuss some of the conceptual aspects of our framework, as well as some insights concerning market design and market stability, in particular in view of the recent proposals to curb the HFT activity by reintroducing periodic batch auctions Budish et al. (2013); Fricke and Gerig (2014).

7.3 Review of the literature

The literature on price formation is obviously old and vast. It is divided into two distinct branches : microeconomics and financial economics, with quite different perspectives on the problem. The microeconomic community often attempts to determine how an economy with a pre-defined set of agents and preferences produces equilibrium prices, and study their properties (uniqueness, stability, computability, convergence etc.). This is mostly a static view, whereas financial economists are mostly interested in the dynamics of these prices. However, the assumption that markets are instantaneously arbitraged and efficient imposes (semi-)martingale properties for the price, the dynamics of which is entirely driven by *news*, subsuming all knowledge of the actual dynamics of

supply and demand itself. The quantities of interest are then the volatility of prices, the distribution of returns, etc., as well as the micro-structural properties of the immediate supply and demand visible in the order book, which however corresponds to an infinitesimal fraction of the total supply and demand (see below). The aim of this section is to give a (rough) overview of these two different approaches and position the present work with respect to both of them.

7.3.1 Theory of supply and demand in economics

As stated above, the question of how prices emerge from supply and demand has fuelled more than a century of economic research, based on the assertion that prices are such that supply equals demand for every asset in the economy. The immediate questions that arise are whether this equilibrium exists, is unique and is stable (in some sense) Mas-Colell et al. (1995). Whereas these questions are rather subtle in a multi-asset economy Hicks (1946); Samuelson (1983), they become trivial in the case of a single asset economy as soon as the supply and demand curves are strictly monotonous. We will only consider a single asset economy in the present paper, since it fully suffices for our purpose and allows one to better focus on the essential part of our message, as we skip some of the usual problems that arise in multi-dimensional settings (is the equilibrium unique, stable, computable, etc.).

As many have noted, such a static description of prices is not fully satisfying, and a description of the dynamics of prices would be highly valuable. However, several interpretations of what "dynamics" actually means can be found in the literature.

1. Dynamics might refer to the way prices converge towards equilibrium. To address this point and the unrealistic fact that the Walrasian mechanism does not allow agents to trade until the equilibrium is reached, some economists have introduced the concept of *non-tâtonnement* in which agents are allowed to trade before the equilibrium has been reached Fisher (1989). Whether such convergence dynamics, even in the presence of trading, should be identified to the dynamics of market price itself, is far from obvious. In fact, as mentioned in Mas-Colell et al. (1995), such a model should be thought of *not as modelling the evolution of a supply-and-demand driven economy, but rather as a tentative trial-and-error process taking place in fictional time [...] [to find] the equilibrium level of prices*. We are thus speaking here of a *transient* dynamics in an otherwise stable world.

2. A second way of introducing dynamics is to consider a multiple period economy in which the supply and demand may evolve at each period, resulting in a new price (see for example Mankiw (2014), Ch. 14). This collection of static equilibria is a rather weak notion of dynamics, that is closer to a *quasi-static* evolution without transactions, in which the price is always the outcome of an equilibrium supply and an equilibrium demand. Hicks (1946) noted that *we shall find [...] that there is a way of reducing the dynamic problem into terms where it becomes*

formally identical with that of statics, showing that something important is somehow missing. As we shall indeed argue below, this figment misses the essential point that when transactions occur, supply and demand curves are both immediately depleted, thus affecting subsequent transactions. Only when the time between market clearing auctions is large enough can the supply and demand curve again be considered in an equilibrium state prior to any further transaction.

3. Following up on the last remark, our view is that a complete dynamic description must account for the evolution of supply and demand in an ever changing world, both in-between auctions/transactions and right when the auction takes place. This appears as a necessity if one wants to understand the formation of real prices, in particular in financial markets where supply and demand permanently interact, and where transactions prevent the supply and demand curves from being in an equilibrium at any point in time. To our knowledge, whereas many papers have worked in the direction of understanding price formation under continuous double auction Gjerstad and Dickhaut (1998); Biais (1993); Cason and Friedman (1996); Easley and Ledyard (1993), such a general dynamical theory of supply and demand is not available at this stage. This is what the present paper aims at achieving.

As we shall see below, our proposal is at odds with classical approaches, which usually consider "rigid" supply and demand curves, that shift uniformly with respect to each other (see e.g. Mankiw (2014), Ch. 14). In contrast, we propose a partial derivative equation that describes the evolution and deformation of the full price dependent supply and demand curves as the core ingredient of our model.

7.3.2 Financial economics : The Kyle model

One strategy to model financial markets that has been popular since Kyle's seminal paper Kyle (1985) is to consider markets as one- or several-period(s) equilibrium(a) between two or more (representative) agents, each representing a well-identified trading behaviour. In Kyle's original paper, an informed trader competes with a market maker who provides liquidity for every trade, in the presence of a noise trader who trades at random. This modelling strategy has been followed in many subsequent papers, where a small number of rational agents optimize a given utility function and maintain an ecological equilibrium.

Whereas the results of such models often yield useful qualitative intuition, they miss – in our opinion – two essential features of markets. First, trading occurs in continuous time and reasoning in terms of periods (e.g. one period per trading day) is not appropriate in that respect. Second, the number of (representative) agents is usually pre-defined and small ; typically, three strategies in Kyle's model. This has to be contrasted with real markets where agents are strongly heterogeneous, and the very idea of representative agents dubious. As a result, these equilibrium models do not

reproduce some essential market features such as the square-root impact law and therefore probably miss some fundamental aspects of price formation.

At variance with these usual models, the present paper suggests that relevant agent-based modelling should incorporate three essential features : (i) a one-dimensional definition of the price dynamics *via* the order book, (ii) continuous time and (iii) heterogeneous agents. Based on these three ingredients, we define and solve below a particular tractable class of models that appears to capture faithfully some essential feature of price formation in continuous double-auction markets. But we believe that our results would hold for a much wider class of models based on the same ingredients.

7.3.3 Financial economics : Models of the limit order book

Instead of the above “macroscopic" considerations on the (unobservable) supply and demand, a whole branch of financial mathematics (concerned with “market microstructure”) has recently emerged. The focus is on the actual evolution of the limit order book and of price formation on financial markets Gould et al. (2013). The limit order book is described as a queueing system, in which buyers (resp. sellers) post quantities on a discrete price grid (the elementary price change is called the “tick”) and wait for being executed. The buyers (resp. sellers) that offer the best price are then executed by aggressive sellers (resp. buyers) according to a *first in, first out* policy. When the whole volume present on the best ask/bid queue is executed, then either some sell/buy volume replenishes the queue and the price is unchanged, or the queue is replenished by opposite side traders and the price moves by one tick. An obvious motivation for such research is to give practical answers to many questions from the financial industry (concerning optimal market making, optimal execution, optimal trading, etc. Cont and Kukanov (2013); Cartea and Jaimungal (2015)) as well as from regulators (tick size, market ecology, market design and stability etc. Robert and Rosenbaum (2011); Laruelle et al. (2011)). Much effort has been devoted to understand and model the mechanics of the limit order books, how it is affected by market design and the ecology of traders (in particular High Frequency Trading), and how it relates to macroscopic variables such as price volatility, etc Foucault et al. (2013); Handa et al. (2003); Hasbrouck (2006).

The availability of detailed data where all market events are recorded (i.e. trades, quotes, cancellations, etc.) has generated a flurry of empirical papers, describing many aspects of price formation at the microstructural level (for a review see Biais et al. (1995); Bouchaud et al. (2009)). Correspondingly, a host of stylized models of the order book have appeared, with different starting points and objectives. For example, “zero-intelligence” models Smith et al. (2003); Bouchaud et al. (2002); Farmer et al. (2005); Cont et al. (2010); Gareche et al. (2013) form an important class of models of the order book, where one assumed that agents act mechanically (rather than strategically) leading to simple Poissonian statistics for the order flow. Although obviously too simple to account for

what goes on in financial markets, such models reveal some interesting relationships between observables (spreads, volatility, activity, etc.) Farmer et al. (2005); Cont and De Larrard (2013). Much more elaborate models have also been developed, taking into account the heterogeneity, strategies and preferences of market participants Foucault (1999); Roşu (2009, 2014); Maglaras and Moallemi (2011); Lachapelle et al. (2013), some including the queues behind the best buy/sell prices Huang et al. (2014).

The present paper is clearly partly inspired by the above strand of papers on real limit order books, in particular Smith et al. (2003); Bouchaud et al. (2002). However, we depart from these models on one very fundamental issue. Instead of trying to describe the evolution of the *visible* order book (where only a tiny fraction of the outstanding liquidity is revealed, and whose dynamics is dominated by highly strategic market-makers/HFT), we want to describe the much deeper and much slower "latent" order book, introduced in Tóth et al. (2011), that contains all buy/sell *intentions*, whether displayed or not by market participants. In other words, we model the true underlying supply and demand curves that would materialize if the transaction price was to move closer to the reservation prices. The distinction between the visible limit order book (which, as stated above, gives a very poor indication on liquidity at larger scales) and the true supply and demand curves is absolutely crucial for all that follows. The model described below is a generalisation of the ideas introduced in Tóth et al. (2011); Mastromatteo et al. (2014a) and in Donier et al. (2015). It builds upon the intuition that agents can revise their reservation prices in an heterogeneous manner, introduced long ago in Bak et al. (1997) and recently revisited in completely different contexts in Lasry and Lions (2007); Lehalle et al. (2011) and in Tóth et al. (2011). The motivation of the latter paper was to explain the universal concave ("square-root") impact of directional trade sequences mentioned in the introduction, that deeply challenges standard equilibrium models.

In summary, the aim of the present paper is to reconcile the insights gained by the financial literature on price formation with a more Walrasian view of supply and demand that provides us with a macroscopic theory of price formation. We believe that this reconciliation has important conceptual consequences from an economic perspective (in particular in emphasizing the *dynamical aspects* of price formation and liquidity), as well as practical implications for market design and regulation (in particular concerning the crucial issue of market stability).

7.4 A dynamic theory of the supply & demand curves

7.4.1 Definitions

The classical supply and demand curves $S(p,t)$ and $D(p,t)$ (SD) represent respectively the amount of supply and demand that would reveal themselves if the price were to be set to p at time t. In classical Walrasian auctions, the equilibrium price p_t^* is then set to the value that matches

both quantities so that $D(p_t^*, t) = S(p_t^*, t)$. This equilibrium is unique provided the curves are strictly monotonous[6]. The supply and demand curves, as well as the resulting equilibrium price, are represented on Figure 7.1 (left).

In order to define the dynamics of the supply and demand curves, we also introduce the *marginal supply and demand curves* (MSD), on which we will focus in the rest of this paper. They are defined as the derivative of the SD curves

$$\rho_S(p,t) = \partial_p S(p,t) \geq 0;$$
$$\rho_D(p,t) = -\partial_p D(p,t) \geq 0,$$

with the following interpretation : For any price p, $\rho_S(p,t)\mathrm{d}p$ (resp. $\rho_D(p,t)\mathrm{d}p$) is, at time t, the quantity of supply (resp. demand) that would materialize if the price changed from p to $p + \mathrm{d}p$ (resp. $p - \mathrm{d}p$). The MSD curves can thus be seen as the density of supply and demand intentions in the vicinity of a given price. Figure 7.1 (right) shows MSD curves corresponding to the SD curves : Higher MSD levels correspond to larger slopes for the SD curves.

In the Walrasian story, supply and demand pre-exist and the Walrasian auctioneer gropes (*tâtonne*) to find the price p_t^* that maximizes the amount of possible transactions. The auction then takes place at time t and removes instantly all matched orders. Assuming that all the supply and demand intentions close to the transaction price were revealed before the auction and were matched, the state of the MSD just after the auction is simple to describe, see Figures 7.1 & 7.7 :

$$\begin{cases} \rho_S(p,t^+) &= \rho_S(p,t^-) \quad (p > p_t^*) \\ &= 0 \quad (p \leq p_t^*) \\ \rho_D(p,t^+) &= \rho_D(p,t^-) \quad (p < p_t^*) \\ &= 0 \quad (p \geq p_t^*). \end{cases} \tag{7.2}$$

But what happens next, once the auction has been settled ? So far the story does not tell (to the best of our knowledge). The aim of the following is to set up a general framework for the dynamics of the supply and demand curves. This will allow us to describe, among other questions, how the supply and demand curves evolve from the truncated shape given by Equation (7.2) up to the next auction at time $t + \tau$ (where τ is the inter-auction time).

7.4.2 General hypotheses about the behaviour of agents

The theory that we present here relies on weak and general assumptions on agents behaviours that translate into a simple and universal evolution of the MSD curves, with only very few para-

6. By definition, or simply by common sense, the demand curve is a decreasing function of the price whereas the supply curve is increasing.

meters [7]. The MSD curves aggregate the intentions of all agents, which would materialize in the "real" order book if it was not for fear of being picked off by more informed traders, or of revealing some information to the market. This is why the MSD curves were called the "latent" order book (LOB) in Refs. Mastromatteo et al. (2014a,b); Donier et al. (2015), as initially proposed in Tóth et al. (2011).

We will assume that there is a so-called "fundamental" price process $\widehat{p}_t$ which is only partially known to agents, in a sense clarified below (see Section 10). For simplicity, we will also posit that $\widehat{p}_t$ is an additive Brownian motion. In the absence of transactions, the MSD curves evolve according to three distinct mechanisms, that we model as follows :

— New intentions, not present in the supply and demand before time t, can appear. The probability for new buy/sell intentions to appear between t and $t+\mathrm{d}t$ and between prices p and $p+\mathrm{d}p$ is chosen to be $\omega_{\pm}(p-\widehat{p}_t)$, where $\omega_{+}(x)$ is a decreasing function of x and $\omega_{-}(x)$ is an increasing function of x.

— Already existing intentions to buy/sell at price p can simply be cancelled and disappear from the supply and demand curves. The probability for an existing buy/sell intention around price p to disappear between t and $t+\mathrm{d}t$ is chosen to be $\nu_{\pm}(p-\widehat{p}_t)$.

— Already existing intentions to buy/sell at price p can be revised. Between t and $t+\mathrm{d}t$, each agent i revises his/her reservation price p^i to $p^i+\beta^i\mathrm{d}\xi_t+\mathrm{d}W_{i,t}$, where $\mathrm{d}\xi_t$ is common to all i, representing some public information. β^i is the sensitivity of agent i to the news, which we imagine to be a random variable from agent to agent, with a mean normalized to 1. Some agents may over-react ($\beta^i>1$), others under-react ($\beta^i<1$). The idiosyncratic contribution $\mathrm{d}W_{i,t}$ is an independent Wiener noise both across different agents and in time, with distribution of mean zero [8] and variance $\Sigma_i^2\mathrm{d}t$, that may depend on the agent (some agents might be more "noisy" than others).

We will furthermore assume that the "news" term $\mathrm{d}\xi_t$ is a Wiener noise of variance $\sigma^2\mathrm{d}t$, corresponding to a Brownian motion for the fundamental price $\widehat{p}_t=\int^t\mathrm{d}\xi_{t'}$ with volatility σ. Normalising the mean of the β^i's to unity thus corresponds to the assumption that agents are on average unbiased in their interpretation of the news – i.e. their intentions remain centred around the fundamental price $\widehat{p}_t$ in the course of time – but see the expanded discussion of this point in Section 10.

Our central assumptions are *heterogeneity*, together with the hypothesis that idiosyncratic behaviours "average out" in the limit of a very large number of participants, i.e., no single agent accounts for a finite fraction of the total supply or demand. While not strictly necessary, this assumption

7. In fact, as shown in Donier et al. (2015) and below, only two parameters suffice to describe the problem in the vicinity of the price : one is the price volatility, and the other one is related to market activity (traded volume per unit time).

8. One could generalize the calculations below to the case where the mean is non zero (modelling for example the tendency of agents to revise their reservation price in the direction of the traded price). This would affect none of the conclusions below, at least in the limit where the inter-auction time τ becomes very small.

leads to a deterministic aggregate behaviour and allows one to gloss over some rather involved mathematics.

7.4.3 The model in terms of optimizing agents

The above assumptions might appear obscure to those used to think in terms of rational optimizing agents and equilibria. Here we rephrase these assumptions in a language closer to standard economic intuition.

We consider an *open* economic system, in which many heterogeneous, infinitesimal agents operate. Each agent i has a certain utility $\mathcal{U}_i(p, \theta | \widehat{p}_t^i, \mathcal{F}_t)$ for buying ($\theta = +1$) or selling ($\theta = -1$) a unit (small) quantity at price p, given his/her estimate of the fundamental price $\widehat{p}_t^i$ and all the information about the rest of the world, available at time t, encoded in $\mathcal{F}_t$. The third option available to agent i is to be inactive ($\theta = 0$), in which case the number of goods s/he owns remains constant. Agents are heterogeneous in the sense that both their utility function and their estimates of the fundamental price are different ; one can think of them as random members of some adequate statistical ensembles. For the sake of simplicity, we consider no interest rate and no risk of any kind.

At time t, each agent computes his optimal action p_t^i, θ_t^i as the result of the following optimisation program :

$$(p_t^i, \theta_t^i) = \underset{p,\theta}{\operatorname{argmin}} \ \mathcal{U}_i(p, \theta | \widehat{p}_t^i, \mathcal{F}_t). \tag{7.3}$$

Because of the random evolution of the outside world summarized by $\widehat{p}_t^i, \mathcal{F}_t$, the value of $\theta_t^i \in \{-1, 0, +1\}$ can change between t and $t + \mathrm{d}t$. For the sake of simplicity, we assume that the change of the state of the world in time $\mathrm{d}t$ is never so large as to induce direct transitions from $\mp 1 \to \pm 1$ without pausing at 0. Hence, between t and $t + \mathrm{d}t$, the following transitions (or absence thereof) are possible :

- $0 \to 0$: this clearly induces no change in the MSD curves ;
- $0 \to \pm 1$: in this case, agent i previously absent from the market becomes either a buyer or a seller, with reservation price p_t^i given by Equation (7.3). The assumption that agents are heterogeneous translates in a model where this event is a Poisson process with some arrival rate $\omega_\pm(p)$;
- $\pm 1 \to 0$: in this case, agent i previously present in the market as a buyer or a seller, decides to become neutral, which is modelled as a Poisson process with some cancellation rate $\nu_\pm(p)$;
- $\pm 1 \to \pm 1$: in this case, a buyer/seller remains a buyer/seller, but may change his/her reservation price because the solution of Equation (7.3) has changed. Writing $p_t^i = f_i(\widehat{p}_t^i, t)$,

where f is a regular function if $\mathcal{U}_i$ is regular enough, and applying Itô's lemma, one finds :

$$\begin{aligned} \mathrm{d}p_t^i &= \frac{\partial f_i}{\partial t}\mathrm{d}t + \frac{\partial f_i}{\partial p}\mathrm{d}\widehat{p}_t^i + \frac{\sigma_i^2}{2}\frac{\partial^2 f_i}{\partial p^2}\mathrm{d}t \\ &= \alpha_t^i \mathrm{d}\widehat{p}_t^i + \gamma_t^i \mathrm{d}t. \end{aligned}$$

The drift term γ_t^i will play little role in the following (see previous footnote), and we neglect it henceforth. In order to recover the specification of the above section, we further decompose the price revision $\mathrm{d}p_t^i = \alpha_t^i \mathrm{d}\widehat{p}_t^i$ into a *common* component $\beta^i \mathrm{d}\xi_t$ and an *idiosyncratic* component $\mathrm{d}W_{i,t}$ as above.

Therefore, the mechanism proposed in the above section indeed describes the behaviour of an open system of *infinitesimal* and *heterogeneous* market participants. Note that we do not need to distinguish between fundamental investors, noise traders and market makers, as for example in the Kyle model Kyle (1985). This is due to our assumption that the market contains a very large number of participants, in which case the MSD curves are continuous. Discretization effects (in price and in quantity) would open gaps in the MSD curves, and specific market makers would then be needed to ensure continuous, orderly trading. Finally, and quite importantly, the price dynamics in the above setting is arbitrage free (see Donier et al. (2015)). There is therefore no optimal strategic component that is missing from the above utility maximisation program.

7.4.4 The "free evolution" equation for the MSD curves

Endowed with the above hypothesis, one can derive stochastic partial differential equations for the evolution of the marginal supply ($\rho_S(p,t) = \partial_p S(p,t)$) and the marginal demand ($\rho_D(p,t) = -\partial_p D(p,t)$) in the absence of transactions Donier et al. (2015). It turns out that, as expected, these equations take a simpler form in the reference frame of the (moving) fundamental price $\widehat{p}_t$. Introducing the shifted price $y = p - \widehat{p}_t$, one finds Donier et al. (2015) : [9]

$$\left\{ \begin{array}{rcl} \partial_t \rho_D(y,t) &=& \mathcal{D}\partial_{yy}^2 \rho_D(y,t) - \nu_+(y)\rho_D(y,t) + \omega_+(y); \\ \partial_t \rho_S(y,t) &=& \underbrace{\mathcal{D}\partial_{yy}^2 \rho_S(y,t)}_{\text{Updates}} - \underbrace{\nu_-(y)\rho_S(y,t)}_{\text{Cancellations}} + \underbrace{\omega_-(y)}_{\text{New orders}}, \end{array} \right. \tag{7.4}$$

where $\mathcal{D} = \frac{1}{2}[\mathbb{E}_i(\Sigma_i^2) + \sigma^2 \mathrm{Var}(\beta^i)]$, i.e. part of the diffusion term comes from the purely idiosyncratic "noisy" updates of agents ($\mathbb{E}_i(\Sigma_i^2)$), and another part comes from the inhomogeneity of their reaction to news ($\sigma^2 \mathrm{Var}(\beta^i)$), which indeed vanishes if all β^i's are equal to unity. [10]

9. See also Lasry and Lions (2007); Lehalle et al. (2011) for similar ideas in the context of mean-field games. Note that Equation (6.1) is strictly valid when $\rho_S(p,t)$ and $\rho_D(p,t)$ are be interpreted as the marginal supply and demand curves averaged over the noise processes. Otherwise some noisy component remains, see e.g. Dean (1996).

10. Here we neglect the possibility that buyers and sellers update their price differently, but one could make a distinction between a $\mathcal{D}_+$ and a $\mathcal{D}_-$, or even make $\mathcal{D}$ price/time dependent.

These equations, that are at the core of the present paper, describe the structural evolution of supply and demand around the fundamental price $\widehat{p}_t$. Notice however that $\widehat{p}_t$ has disappeared from the above equations. The dynamics of the MSD curves can be treated independently from the dynamics of the price itself, provided one describes the MSD in the reference frame of the price. There is however a direct relationship between the price volatility σ and the diffusion coefficient $\mathcal{D}$, as expressed above and noted in Donier et al. (2015).

Interestingly, whereas the price is random and follows a rough path (typically a Brownian motion), the structural part is deterministic and smooth, thanks to the assumption of "infinitesimal" orders (that can be made rigourous by considering an appropriate scaling for system parameters that corresponds to a hydrodynamic limit, see Gao et al. (2014)).

The above equations for $\rho_D(y,t)$ and $\rho_S(y,t)$ are linear and can be formally solved in the general case, starting from an arbitrary initial condition such as Equation (7.2), using a spectral decomposition of the evolution operator. This general solution is however not very illuminating, and we rather focus here in the special case where $\nu_\pm(y) \equiv \nu$ does not depend on y nor on the side of the latent order book. The general solution can then be written in a fairly transparent way, as :

$$\rho_{S,D}(y,t) = \int_{-\infty}^{+\infty} \frac{\mathrm{d}y'}{\sqrt{4\pi\mathcal{D}t}}\, \rho_{S,D}(y', t=0^+) e^{-\frac{(y'-y)^2}{4\mathcal{D}t} - \nu t} + \int_0^t \mathrm{d}t' \int_{-\infty}^{+\infty} \frac{\mathrm{d}y'}{\sqrt{4\pi\mathcal{D}(t-t')}}\, \omega_\pm(y') e^{-\frac{(y'-y)^2}{4\mathcal{D}(t-t')} - \nu(t-t')}, \tag{7.5}$$

where $\rho_{S,D}(y, t=0^+)$ is the initial condition, i.e. just after the last auction.

We will now explore the properties of the above solution at time $t = \tau^-$, i.e., just before the next auction, in the two asymptotic limits $\tau \to \infty$, corresponding to very infrequent auctions, and $\tau \to 0$, corresponding to continuous time auctions.

7.5 Discrete Auctions and Price Impact

The aim of this section is to show that the shape of the marginal supply and demand curves can be fully characterized in the limit of very infrequent auctions (corresponding to Walras' auctions) and in the opposite limit of nearly continuous time auctions (corresponding to financial markets), and describe the transition between the two limits. The upshot is that while the liquidity around the auction price is in general finite and leads to a linear impact using the standard argument in Equation (7.1) above, this liquidity vanishes as $\sqrt{\tau}$ when the inter-auction time $\tau \to 0$. This signals the breakdown of linear impact and, as shown at the end of the section, its replacement by the square-root law mentioned in the introduction.

7.5.1 Walras, or the limit of infrequent auctions

Letting $t = \tau \to \infty$ in the above Equation (7.5), one immediately sees that the first term disappears, meaning that one reaches a *stationary solution* $\rho^{\text{st.}}_{S,D}(y)$ that is independent of the initial condition. The second term can be simplified further to give the following general solution :

$$\rho^{\text{st.}}_{S,D}(y) = \frac{1}{2\sqrt{\nu\mathcal{D}}} \int_{-\infty}^{+\infty} \mathrm{d}y' \omega_{\pm}(y')\, e^{-\sqrt{\frac{\nu}{\mathcal{D}}}|y'-y|}. \tag{7.6}$$

A particularly simple case is when $\omega_{\pm}(y) = \Omega_{\pm} e^{\mp \mu y}$, meaning that buyers(/sellers) have an exponentially small probability to be interested in a transaction at high/low prices. In this toy-example, one readily finds that a stationary state only exists when $\nu > \mathcal{D}\mu^2$ and reads :

$$\rho^{\text{st.}}_{S,D}(y) = \frac{\Omega_{\pm}}{\nu - \mathcal{D}\mu^2} e^{\mp \mu y}.$$

Other forms for $\omega_{\pm}(y)$ can be investigated as well, for example $\omega_{\pm}(y) = \omega^0_{\pm} \mathbb{1}_{\{y<>0\}}$ which yields :

$$\rho^{\text{st.}}_{S,D}(y) = \frac{\omega^0_{\pm}}{2\nu} \left[1 \pm \text{sign}(y)(1 - e^{-\sqrt{\nu/D}|y|})\right],$$

that we will use in Figures 7.3 and 7.4. The shape of $\rho^{\text{st.}}_{S,D}(y)$ is generically the one shown in Figure 7.1 with an overlapping region where buy/sell orders coexist. The auction price $p^*_\tau = \widehat{p}_\tau + y^*$ is determined by the condition $D(p^*_\tau, \tau^-) = S(p^*_\tau, \tau^-)$, or else :

$$\int_{y^*}^{\infty} \mathrm{d}y \rho^{\text{st.}}_D(y) = \int_{-\infty}^{y^*} \mathrm{d}y \rho^{\text{st.}}_S(y) \equiv v^*,$$

where v^* is the volume exchanged during the auction. For the simple exponential case above, this equation can be readily solved as :

$$y^* = \frac{1}{2\mu} \ln \frac{\Omega_+}{\Omega_-},$$

with a clear interpretation : if the new buy order intentions accumulated since the last auction happen to outsize the new sell intentions during the same period, the auction price will exceed the fundamental price, and vice-versa. This pricing error is expected to be small if the order book is observable during the inter-auction period, since in that case Ω_+ and Ω_- will track each other and remain close. Otherwise, one expects the imbalance to invert in the next period, leading to a kind of "bid-ask bounce" well known in the context of market microstructure. One can also compute the volume exchanged during the auction v^*. One finds :

$$v^* = \frac{\sqrt{\Omega_+ \Omega_-}}{\mu(\nu - \mathcal{D}\mu^2)}.$$

Just after the auction, the MSD curves start again from $\rho^{\text{st}\cdot}_{S,D}(y)$, truncated below (resp. above) y^*, as in Equation (7.2).

Let us now turn to price impact in this model. From Equation (7.6), it is immediate that for any clearing price y^*, both $\rho^{\text{st}\cdot}_S(y^*)$ and $\rho^{\text{st}\cdot}_D(y^*)$ are strictly positive. This would remain true even if the dependence on y of cancellation rate $\nu_\pm(y)$ was reinstalled. The general argument given in the introduction therefore predicts a *linear impact* for an extra buy/sell quantity given by :

$$\mathcal{I}(Q) = \pm\lambda Q; \qquad \lambda = \frac{1}{\rho^{\text{st}\cdot}_S(y^*) + \rho^{\text{st}\cdot}_D(y^*)}.$$

For the exponential case, this again takes a simple form :

$$\lambda = \frac{\nu - \mathcal{D}\mu^2}{2\sqrt{\Omega_+\Omega_-}},$$

whereas for a general symmetric order flow $\omega_+(y) = \omega_-(-y)$, y^* is obviously equal to zero, leading to :

$$\lambda = \frac{\sqrt{\nu\mathcal{D}}}{\int_{-\infty}^{+\infty} \mathrm{d}y'\omega(y')e^{-\sqrt{\frac{\nu}{\mathcal{D}}}|y'|}}.$$

For $\omega_\pm(y) = \omega^0 \mathbb{1}_{\{y<>0\}}$, one obtains the simple and intuitive result :

$$\lambda = \frac{\nu}{\omega_0},$$

i.e. that the market liquidity, measured by λ^{-1}, grows linearly with the rate of incoming orders and inversely proportionally to the cancellation rate.

The main point of the present section is that when the inter-auction time is large enough, each auction clears an equilibrium supply with an equilibrium demand, with very simple and predictable outcomes. This corresponds to the quasi-static dynamics discussed in item 2., Section 7.3.1, and to the standard representation of market dynamics in the Walrasian context, since in this case only the long-term properties of supply and demand matter and the whole transients are discarded. The next section will depart from this limiting case, by introducing a finite inter-auction time such that the transient dynamics of supply and demand becomes a central feature in the theory.

7.5.2 High frequency auctions

We will now investigate the alternative limit where the inter-auction time τ tends to zero. Since all the supply (resp. demand) curve left (resp. right) of the auction price is wiped out by the auction process, one expects intuitively that after a very small time τ, the density of buy/sell orders in the immediate vicinity of the transaction price will remain small. We will show that this is indeed the case, and specify exactly the shape of the stationary MSD after many auctions have taken place.

Consider again Equation (7.5) just before the $n+1$th auction at time $(n+1)\tau^-$, in the case where the flow of new orders is symmetric, i.e. $\omega_+(y) = \omega_-(y)$, such that the transaction price is always at the fundamental price ($y^* = 0$). We will focus on the supply side and postulate that $\rho_S(y, t = n\tau^-)$ can be written, in the vicinity of $y = 0$, as [11] :

$$\rho_S(y, t = n\tau^-) = \sqrt{\tau}\,\phi_n\left(\frac{y}{\sqrt{\mathcal{D}\tau}}\right) + O(\tau) \tag{7.7}$$

when $\tau \to 0$ (and symmetrically for the demand side). Plugging this ansatz into Equation (7.5), making the change of variable $y' \to \sqrt{\mathcal{D}\tau}w$ and taking the limit $\tau \to 0$ leads to the following iteration equation, exact up to order $\sqrt{\tau}$: [12]

$$\phi_{n+1}(u) = \int_0^{+\infty} \frac{\mathrm{d}w}{\sqrt{4\pi}}\phi_n(w)e^{-(u-w)^2/4} + \sqrt{\tau}\,\omega(0) + O(\tau).$$

Note that ν has entirely disappeared from the equation (but will appear in the boundary condition, see below), and only the value of ω close to the transaction price is relevant at this order.

After a very large number of auctions, one therefore finds that the stationary shape of the demand curve close to the price and in the limit $\tau \to 0$ is given by the non-trivial solution of the following fixed point equation :

$$\phi_\infty(u) = \int_0^{+\infty} \frac{\mathrm{d}w}{\sqrt{4\pi}}\phi_\infty(w)e^{-(u-w)^2/4}, \tag{7.8}$$

supplemented by the boundary condition $\phi_\infty(u \gg 1) \approx \mathcal{L}\sqrt{\mathcal{D}}u$, where $\mathcal{L}$ is a constant to be determined below. [Note that the solution of Equation (7.8) is determined up to a multiplicative factor that must be fixed by some external condition].

Equation (7.8) is of the Wiener-Hopf type and its analytical solution can be found in Atkinson (1974); Boersma (1974). We plot numerically this solution in Figure 7.2 ; it is seen to be numerically very close to an affine function for $u > 0$: $\phi_\infty(u) \approx \mathcal{L}\sqrt{\mathcal{D}}(u + u_0)$ with $u_0 \approx 0.824$. In summary, the stationary shape $\rho_{S,\text{st.}}(y)$ of the marginal supply curve in the frequent auction limit $\tau \to 0$ and close to the transaction price ($y = O(\sqrt{\mathcal{D}\tau})$, has a *universal shape*, independent of the detailed specification of the model (i.e., the functions $\nu_\pm(y)$ and $\omega_\pm(y)$). This supply curve is given by $\sqrt{\tau}\phi_\infty(y/\sqrt{\mathcal{D}\tau})$, which can itself be approximated by a simple affine function that will fully suffice for the purpose of the present paper :

$$\rho_{S,\text{st.}}(y \geq 0) \approx \mathcal{L}(y + y_0); \qquad y_0 = u_0\sqrt{\mathcal{D}\tau}; \qquad (\tau \to 0), \tag{7.9}$$

11. This approximation happens to be exact in the particular setting considered in Donier et al. (2015).

12. An extra correction of order $\sqrt{\tau}$ would appear if a drift term was added to Equation (7.4).

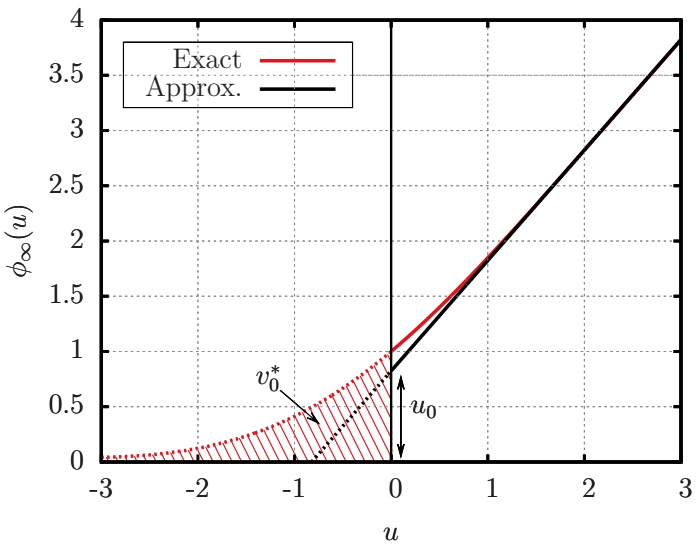

FIGURE 7.2 – Graph of the normalized exact solution $\phi_\infty(u)$, and its affine approximation. The whole picture must be rescaled by a factor $\sqrt{\tau}$ to recover the order book when the inter-auction time is τ. The hatched region corresponds to the volume to be executed, and therefore scales with τ.

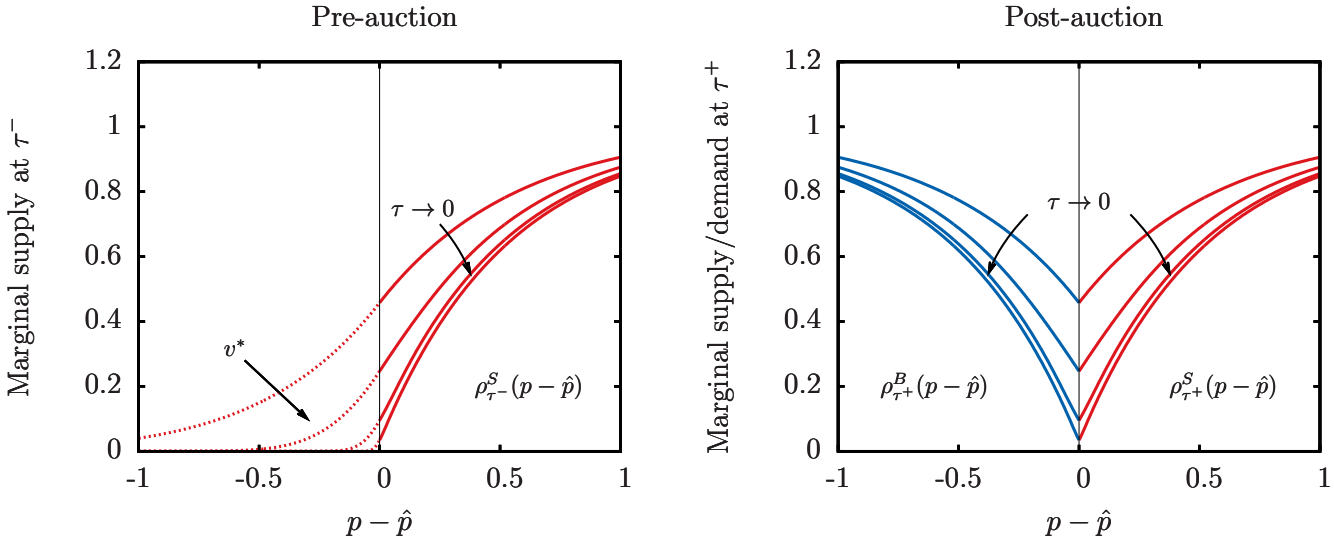

FIGURE 7.3 – Left : Shape of the marginal supply curve immediately before the auctions, for different inter-auction times τ, in the case $\omega_\pm(y) = \omega_\pm^0 \mathbb{1}_{\{y<>0\}}$. Right : Shape of the MSD immediately after the auctions, again for different inter-auction times τ. Note that as $\tau \to 0$, the MSD acquires a characteristic V-shape.

and similarly for $\rho_{D,\text{st.}}(y)$, see Figure 7.3. The detailed interpretation of this result – in terms of market liquidity and price impact – will be given below.

We however still need to find the value of $\mathcal{L}$. This is done by comparing with the stationary

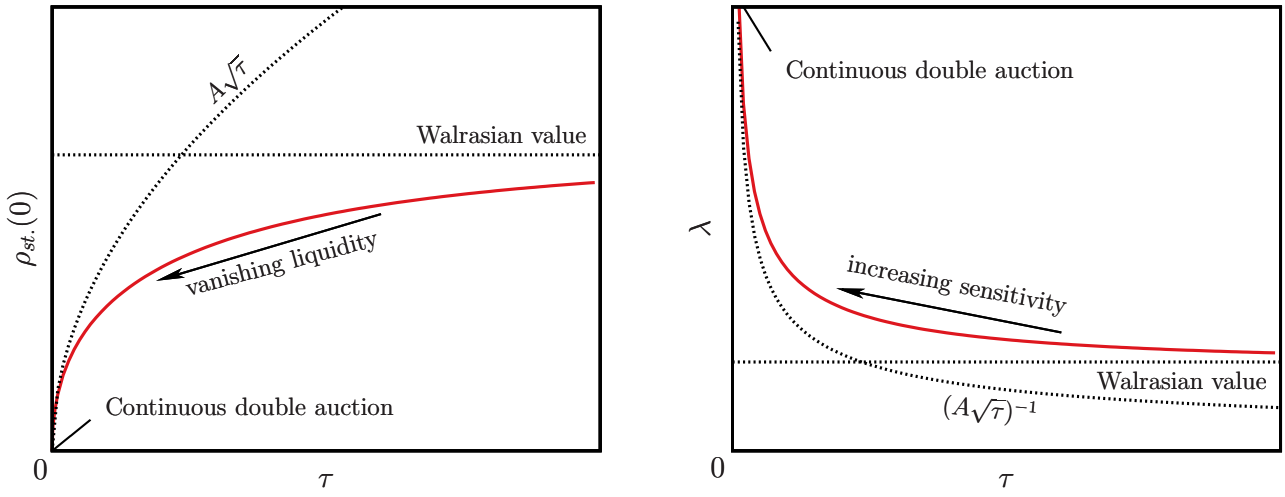

FIGURE 7.4 – Left : As the inter-auction time τ decreases, the liquidity close to the price, $\rho_{\text{st.}}(y=0)$ decreases from its finite Walrasian value to zero in the case of continuous double auctions. Right : Conversely, the relative impact of (small) additional volumes λ diverges for continuous double auctions ($\lambda \to \infty$ when $\tau \to 0$), eventually leading to the square-root impact law.

solution $\varphi_{\text{st.}}(y)$ of Equation (7.4) that satisfies the boundary solution $\varphi_{\text{st.}}(0) = 0$ (valid for $\tau = 0$). For $\nu_{\pm}(y) = \nu$, $\varphi_{\text{st.}}(y)$ can be computed explicitly and is given by :

$$\varphi_{\text{st.}}(y) = \frac{1}{\mathcal{D}} e^{-\sqrt{\nu/\mathcal{D}}\, y} \int_0^y \mathrm{d}y' e^{2\sqrt{\nu/\mathcal{D}}\, y'} \int_{y'}^{\infty} \mathrm{d}y'' e^{-\sqrt{\nu/\mathcal{D}}\, y''} \omega(y'').$$

Expanding $\varphi_{\text{st.}}(y)$ for small y (but still much larger than $\sqrt{\mathcal{D}\tau}$) finally leads to :

$$\varphi_{\text{st.}}(y) \approx \mathcal{L} y; \qquad \mathcal{L} = \frac{1}{\mathcal{D}} \int_0^{\infty} \mathrm{d}y' e^{-\sqrt{\nu/\mathcal{D}}\, y'} \omega(y'),$$

where $\mathcal{L}$ can be seen as a measure of the market liquidity (see Donier et al. (2015) and below). Again in the simple case $\omega_{\pm}(y) = \omega^0 \mathbb{1}_{\{y<>0\}}$, one finds :

$$\mathcal{L} = \frac{\omega_0}{\sqrt{\nu \mathcal{D}}}.$$

Therefore, liquidity increases with the order arrival rate and decreases with their cancellation rate, as above, but also decreases with the diffusion constant $\mathcal{D}$ that can be loosely identified with market volatility (see the discussion above).

Coming back to Equation (7.9), one notes that Kyle's λ behaves as $\lambda^{-1} \equiv 2\rho_{S,\text{st.}}(y=0) \propto \sqrt{\tau}$, which is the pivotal result of the present paper. It means that the marginal supply and demand at the transaction price becomes very small around the transaction price as the auction frequency increases. Intuitively, this is due to the fact that close to the transaction price, liquidity has no time to

rebuild between two auctions. From the point of view of impact, the divergence of Kyle's λ as $1/\sqrt{\tau}$ means that the auction price becomes more and more susceptible to any imbalance between supply and demand. We show in Figure 7.4 (left) $\rho_{\text{st.}}(y=0)$ in the special case where $\omega_{\pm}(y) = \omega^0 \mathbb{1}_{\{y<>0\}}$, illustrating how liquidity vanishes as $\tau \to 0$ as well as (right) the corresponding impact parameter $\lambda(\tau)$ that diverges in this limit. One can thus see how, by increasing the auction frequency, one smoothly departs from the Walrasian equilibrium scenario to reach the limit of continuous double auction corresponding to modern financial markets.

The last item we need is the shape of the supply curve *below* the transaction price just before the next auction, that gives the amount of supply/demand on the "wrong" side of the price, i.e. precisely the volume exchanged at the auction. Using the simple affine approximation of Equation (7.9), one finds :

$$\rho_{S,\text{st.}}(y<0) \approx \mathcal{L}\int_0^{+\infty} \frac{\mathrm{d}y'}{\sqrt{4\pi\mathcal{D}\tau}}\,(y'+y_0)e^{-\frac{(y'-y)^2}{4\mathcal{D}\tau}},$$

or, again setting $y = -u\sqrt{\mathcal{D}\tau}$ and $y' = w\sqrt{\mathcal{D}\tau}$,

$$\rho_{S,\text{st.}}(y<0) \approx \mathcal{L}\sqrt{\mathcal{D}\tau}\int_0^{+\infty} \frac{\mathrm{d}w}{\sqrt{4\pi}}\,(w+u_0)e^{-\frac{(w+u)^2}{4}} = \mathcal{L}\sqrt{\mathcal{D}\tau}\left[\frac{e^{-u^2/4}}{\sqrt{\pi}} + \frac{1}{2}(u_0-u)(1-\text{Erf}(u/2))\right]. \tag{7.10}$$

From this expression, the total volume v^* exchanged during each auction is found to be :

$$v^* = \int_{-\infty}^{0} \mathrm{d}y\rho_{S,\text{st.}}(y) = \mathcal{L}\mathcal{D}\tau\left[\frac{1}{2} + \frac{u_0}{\sqrt{\pi}}\right] \approx 0.965\mathcal{L}\mathcal{D}\tau,$$

whereas the exact result (that can be obtained directly from the diffusion equation in the $\tau \to 0$ limit) is $v^* = \mathcal{L}\mathcal{D}\tau$. The error induced by our simple affine approximation is thus only a few percents. Interestingly, one sees that the total transacted volume V in a finite time interval T, given by $V = v^*T/\tau$, remains finite when $\tau \to 0$, and equal to $V = \mathcal{L}\mathcal{D}T$. This observation should be put in perspective with the recent evolution of financial markets, where the time between transactions τ has become very small, while the volume of each transaction has simultaneously decreased, in such a way that the daily volume has remained roughly constant.

7.5.3 The vanishing liquidity limit : From linear to square-root impact

From the shape of the MSD close to transaction price given by Equation (7.9), it is immediate to compute the supply and demand curves just before an auction when the inter-auction time τ tends to 0. Denoting again as y the difference between the price level p and the fundamental price $\widehat{p}_t$, one finds :

$$\begin{aligned} S(p \geq \widehat{p}_t) &= \mathcal{L}(y_0 y + \frac{1}{2}y^2) + v^* \\ D(p \leq \widehat{p}_t) &= \mathcal{L}(-y_0 y + \frac{1}{2}y^2) + v^* \end{aligned} \tag{7.11}$$

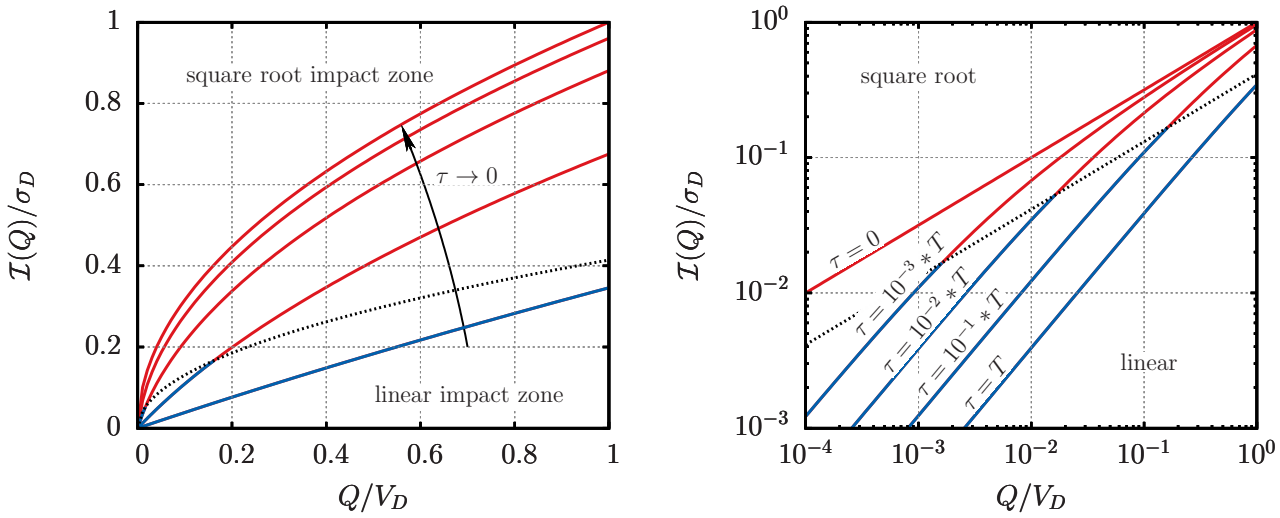

FIGURE 7.5 – The impact of traded volumes Q for a given inter-auction time τ is linear for $Q \ll v^* = \mathcal{L}\mathcal{D}\tau$ and then square-root for $Q \gg v^* = \mathcal{L}\mathcal{D}\tau$. The linear impact zone shrinks to zero when $\tau \to 0$, when one recovers a pure square-root impact, i.e. a diverging Kyle's λ.

where, as found in the previous section, $y_0 \equiv u_0\sqrt{\mathcal{D}\tau} \approx 0.824\sqrt{\mathcal{D}\tau}$. From Equation (7.10) above, it is readily seen that the supply (resp. demand) curve below (resp. above) $\widehat{p}_t$ can be written as $v^* F(y/\sqrt{\mathcal{D}\tau})$, where $F(u)$ is a certain function that goes from $F(0) = 1$ to $F(\infty) = 0$. [13]

The above equation Equation (7.11) immediately allows us to compute the impact $\mathcal{I}(Q) \equiv y^*$ of an extra buy quantity Q, as the solution of $\mathcal{L}(y_0 y^* + \frac{1}{2}y^{*2}) + v^* = Q + v^* F(y^*/\sqrt{\mathcal{D}\tau})$. It is clear that the solution can be written as $y^* = \sqrt{\mathcal{D}\tau}\, Y(Q/\mathcal{L}\mathcal{D}\tau)$, where $Y(q)$ obeys $u_0 Y + \frac{1}{2}Y^2 + (1 - F(Y)) = q$. The limiting behaviours of Y in the limits $q \ll 1$ and $q \gg 1$ are easy to compute, and read :

$$Y(q) \approx_{q\ll 1} 0.555q; \qquad Y(q) \approx_{q\gg 1} \sqrt{2q}.$$

One therefore deduces that the impact $\mathcal{I}(Q)$ is linear in a region where the volume Q is much smaller than $v^* \sim \mathcal{L}\mathcal{D}\tau$, i.e. when the extra volume is small compared to the typical volume exchanged during auctions, as expected. In the other limit, however, one recovers the *square-root impact* observed empirically (as found in Donier et al. (2015) [14]) :

$$\mathcal{I}(Q \gg v^*) \approx \sqrt{\frac{2Q}{\mathcal{L}}},$$

13. This function reads, explicitly :

$$\left[\frac{1}{2} + \frac{u_0}{\sqrt{\pi}}\right] F(u) = \frac{1}{2}(1 - \mathrm{Erf}(u/2))(\frac{u^2}{2} - u_0 u + 1) - \frac{e^{-u^2/4}}{\sqrt{\pi}}(\frac{u}{2} - u_0).$$

14. In that paper, the study of market impact in the $\tau \to 0$ limit has been investigated much more in depth, in particular in the case of a progressive execution over some finite time window.

The impact in the universal small Q region [15], with a linear regime for $Q < v^*$ and a crossover to a square root regime when Q becomes greater that v^*, is shown in Figure 7.5. Clearly, for $\tau = 0^+$, the auction volume $v^* = \mathcal{L}\mathcal{D}\tau$ also tends to zero, so that the region where impact is linear in volume shrinks to zero. In other words, when the interaction time becomes infinitely small, the impact of small trades is *never* linear. This comes from the fact that the MSD curves tend to zero exactly at the trading price when $\tau \to 0$.

7.5.4 Empirical confirmation : the shape of supply and demand on the Bitcoin

Remarkably, we can check directly the above prediction on the shape of the MSD curves using Bitcoin data, where traders are much less strategic than in more mature financial markets and display their orders in the visible order book even quite far from the current price. The graph of the average shape of the Bitcoin limit order book (LOB) and of its integral, that are assumed to be good proxies for the MSD/SD curves at least not too far from the price, is shown on Figure 7.6.

Quite strikingly, the MSD curves are indeed *linear* in the vicinity of the price that corresponds to about 5% range, in perfect agreement with our dynamical theory of supply and demand in the limit of frequent auctions (note in particular that $\partial_y S(p^*) = \partial_y D(p^*) \approx 0$!). Correspondingly, we do expect that impact of meta-orders should be well accounted by a square-root law in this region, which is indeed also found empirically (see Donier and Bonart (2014) for the special case of Bitcoin case, and Torre and Ferrari (1997); Grinold and Kahn (2000); Almgren et al. (2005); Moro et al. (2009); Tóth et al. (2011); Mastromatteo et al. (2014a); Gomes and Waelbroeck (2015); Bershova and Rakhlin (2013); Brokmann et al. (2015); Bacry et al. (2014); Zarinelli et al. (2015) for more mainstream markets). Further away from the price, the non-universal region clearly appears, where the shape of the MSD (here, approximately saturating to some constant value) depends on the detailed characteristics of the order flow, modelled in this paper by the functions $\nu(y)$ and $\omega(y)$.

7.5.5 Summary

The above section presented our story in mathematical terms. The punchline is however quite simple, and well summarized by the graphs plotted in Figure 7.7, where we show (a) the standard Walrasian supply and demand curves just before the auction, from which the equilibrium price p^* can be deduced ; (b) the supply and demand curves just after an auction, when the inter-auction time τ is large enough – in which case the marginal supply and demand are both finite at p^* ; and (c) the supply and demand curves in the continuous time limit $\tau \to 0$, for which the marginal supply and demand curves vanish linearly around the current price, as found in the Bitcoin market.

15. Corresponding to the region where ϕ_∞ is a good approximation of the order book, see Equation 7.7. For very large Q's, the linear approximation describing the MSD curves breaks down, and one enters a presumably non-universal regime that is beyond the scope of the present discussion (cf. e.g. the shape of the MSD on the Bitcoin, Figure 7.6).

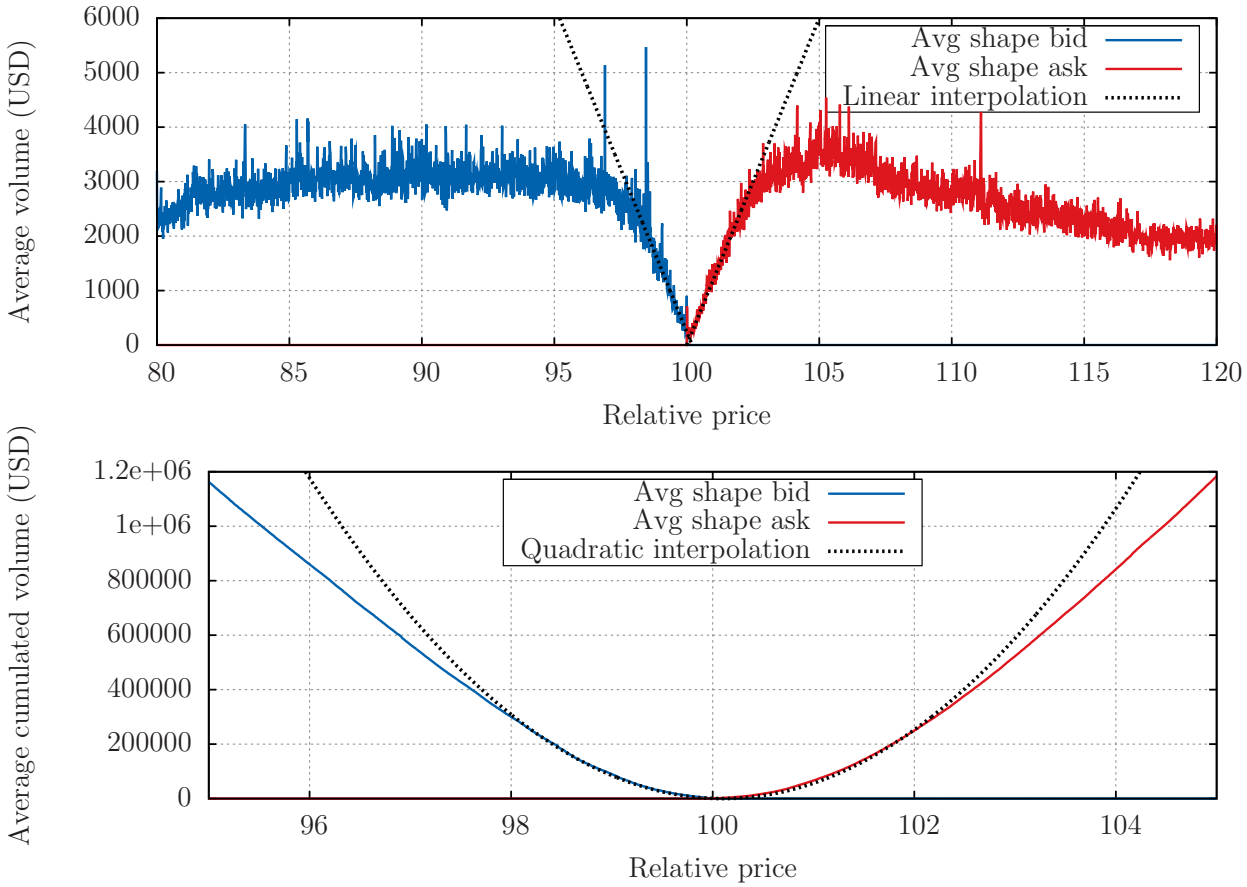

FIGURE 7.6 – Top : The average shape of the limit order book on the Bitcoin market, which we believe to be a representative sample of the MSD curve, in particular close to the price. The data comes from successive snapshots of the full order book of the Bitcoin market, every 15 minutes from May 2013 to September 2013, centred auround the current mid-price. Bottom : Integrated shape of the visible order book, as a proxy for the supply and demand curves on the Bitcoin. The LOB/MSD curves grow linearly with respect to the distance to the price, resulting in a quadratic shape for supply and demand.

7.6 Discussion

Up to now, our presentation has been fairly technical, with the aim of establishing our main results. Still, many points of general interest have glossed over for the sake of readability. We feel that these deserve a more detailed discussion that we provide now.

7.6.1 Price discovery vs. price formation

The interpretation of price moves in financial markets has generated endless theoretical debates, culminating in the split 2013 Nobel prize between the hero of efficient market theory (Fama (1970)), and the beacon of behavioural economics (Shiller (1980)). For the former school of thought, the *fun-*

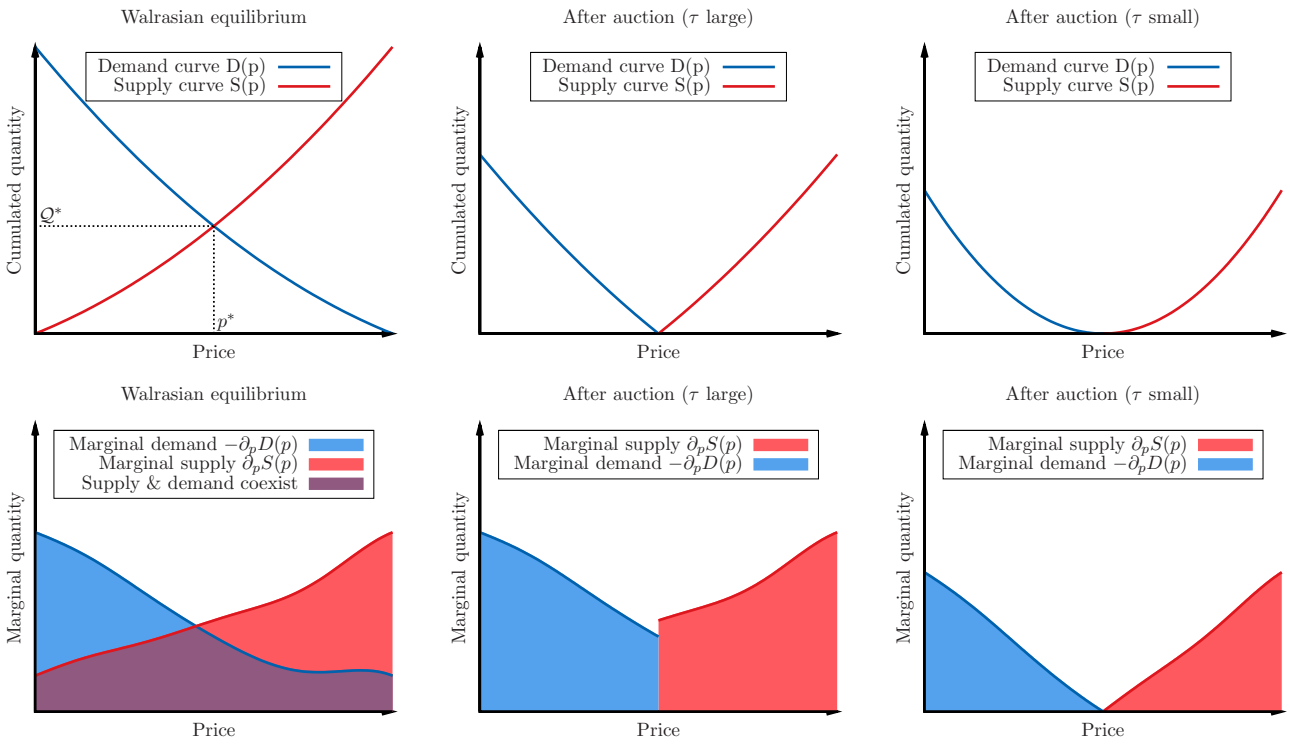

FIGURE 7.7 – Top : Supply and demand curves in (left) Walrasian auctions, (center) immediately after infrequent auctions and (right) immediately after frequent auctions. Bottom : Corresponding MSD curves. When transactions occur, supply and demand cannot cross (center and right). When the market is cleared frequently, supply and demand are depleted close to the price and exhibit a characteristic V-shape (right).

damental price of an asset pre-exists and only waits to be "discovered" by aggregating the unbiased opinions of rational market participants. Up to small self-correcting errors, markets do clear at the right price. This assumption is a pre-requisite for many economic models, e.g. the classical Kyle model Kyle (1985) within which some agents are *informed* : they are supposed to know the fundamental price for the next period and invest accordingly so as to make a maximum profit from of this information.

This perfect Platonian view of markets is however hard to swallow for many followers of Shiller. After all, financial markets are driven by humans who have a very imperfect knowledge of the fundamental price, prone to many behavioural biases. The market is a device that merely aggregates all participants' intentions, regardless on whether they are justified or not, and spits out the "market price". In this view, markets allow *price formation* rather than *price discovery*.

Our modelling strategy provides a very intuitive framework to think about the difference between the two viewpoints. In Section 7.4.2, we postulated that on any time interval $(t, t + \mathrm{d}t)$ an information item is released, leading to a change in the (unobservable) fundamental price $\mathrm{d}\xi_t$, of

variance $\sigma^2 \mathrm{d}t$. This piece of news is however interpreted by agent i as predicting a price change $\beta_t^i \mathrm{d}\xi_t$ plus idiosyncratic noise. As noted above, this means that the news is over-interpreted by some ($\beta^i > 1$) and under-interpreted by others ($\beta^i < 1$). If for any news $\mathrm{d}\xi_t$, the β_t^i's average *exactly* to 1, the market on aggregate perfectly digests the news and the (permanent) increment in perceived price is precisely $\mathrm{d}\xi_t$. If however this is not the case, $\mathbb{E}_i[\beta_t^i]\mathrm{d}\xi_t \neq \mathrm{d}\xi_t$, and this leads to two possibly different definitions of a reference price : the *fundamental* price p_t^{F} (unknown to agents) and the *market* price $\widehat{p}_t$ (encoding agents perceptions) : [16]

$$\begin{aligned} p_t^{\mathrm{F}} &\equiv p_0 + \int_0^t \mathrm{d}\xi_s \\ \widehat{p}_t &\equiv p_0 + \int_0^t \mathbb{E}_i[\beta_s^i]\mathrm{d}\xi_s, \end{aligned} \tag{7.12}$$

where p_0 is an arbitrary reference price, assuming that the market price is equal to the fundamental price initially. Only if the average market reaction $\beta_t \equiv \mathbb{E}_i[\beta_t^i]$ is unbiased at any time can one speak of efficient markets and *price discovery*. Otherwise, *price formation* prevails, and the market price errs away from the fundamental price, as envisaged by Black (1986). More precisely, one can compute the pricing error as :

$$\mathrm{Var}(\widehat{p}_t - p_t^{\mathrm{F}}) = \mathbb{E}\left[\sigma^2 \int_0^t (\beta_s - 1)^2 \mathrm{d}s\right] = \mathrm{Var}(\beta_t)\mathrm{Var}(p_t^{\mathrm{F}}),$$

as well as the variance of the market price as :

$$\mathrm{Var}(\widehat{p}_t) = \mathbb{E}\left[\sigma^2 \int_0^t \beta_s^2 \mathrm{d}s\right] = \left[1 + \mathrm{Var}(\beta_t)\right] \mathrm{Var}(p_t^{\mathrm{F}})$$

showing that even if there is no bias on average (i.e. $\mathbb{E}_s[\beta_s] \equiv 1$), $\mathrm{Var}(\widehat{p}_t) \geq \mathrm{Var}(p_t^{\mathrm{F}})$ with a strict inequality as soon as the market reaction is not perfectly unbiased *at all times* – a highly plausible situation. [17] This is in fact a natural formalization of the conclusions of Gennaioli et al. (2015), that conclude after an empirical investigation of agents' expectations that *plausible models would consider common errors among many economic agents, which therefore would have potential aggregate effects*. This embeds Shiller's famous excess volatility puzzle Shiller (1980) in an interesting formal framework.

We are here at the core of a crucial question in financial economics : what is "information" ? The discussion above naturally leads to two definitions of information :

16. Here and below, we assume that agents perceptions are symmetrically distributed around $\widehat{p}_t$, that we identify with the auction clearing price p^*. In other words, we neglect here any "bid-ask" bounce that may affect the short-time price dynamics.

17. Note however that on long time scales, some weak mean reversion towards the fundamental price should take place, as discussed in e.g. Black (1986); De Bondt and Thaler (1990); Bouchaud and Cont (1998).

(a) information on the *fundamental price* (often called fundamental information), corresponds to some (perhaps noisy) knowledge about the value of p_t^{F}, while

(b) information on the *market price*, corresponds to some (perhaps noisy) knowledge about the future value of the market price $\widehat{p}_t$.

In the latter case, information is only about *correctly anticipating the behaviour of others*, exactly as Keynes envisioned Keynes (1936). The notion of *information* should then rather be replaced by the notion of *correlation* – if all the market participants' β were negative, the correct information for an arbitrageur would correspond to also interpreting the news with a negative β, even if it did not make sense. In this context, the difference between a "noise" trader and an "informed" trader in Kyle's model is merely the level of correlation with the crowd of other market participants : informed traders are positively correlated with it, whereas noise traders are simply uncorrelated with the crowd [18].

Finally, let us give an alternative interpretation of the market price equation Equation (7.12), in terms of the individual estimate of the fundamental price $\widehat{p}_t^i$ and the fraction of market shares F^i of agent $i \in \{1, \dots N\}$, such that $\sum_i F^i = 1$. The individual estimate of the fundamental price evolves as $\mathrm{d}\widehat{p}_t^i = \mathrm{d}\xi_t^i$. Assuming that expectations are symmetrically distributed around $\widehat{p}_t$, the market clearing price $\widehat{p}_t$ is given by the weighted average of individual prices (whether they are truly "informed" or not) :

$$\widehat{p}_t = \sum_i F^i \widehat{p}_t^i,$$

which of course coincides with Equation (7.12). However, one now clearly sees that the *permanent* impact of an agent with market share F^i on the price is $F^i \mathrm{d}\xi^i$ when his view on the price changes by $\mathrm{d}\xi^i$. When the market share of an individual investor is small, the permanent impact of his/her isolated orders on market prices is itself small (in particular much smaller than the transient square-root impact that depends on his/her fraction of the daily volume), except if his/her trade is correlated with the rest of the market, in which case it is often said to have an *alpha* – i.e. a predictive signal on the price (see previous footnote).

7.6.2 Market stability and marked-to-market valuation

The bottom line of the model and analysis presented in the above sections is the possible effect of market design itself on *price stability*. Indeed, the highly singular square root impact consistently measured on financial markets implies that small perturbation (e.g. noise trading) may result in abnormally high returns, questioning the robustness of the price formation and the very stability

18. See the detailed discussion in Donier et al. (2015), and Gomes and Waelbroeck (2015); Donier and Bonart (2014) for empirical studies on how "informed" and "noise" trades impact the price differently on the long run. While both have similar instantaneous impacts, on the long run the price reverts to its initial value for noise trades whereas a permanent component remains for "informed" trades. Interestingly, data suggests that the permanent impact is of the same order of magnitude as the transient impact.

of markets. Indeed, if one freak order, whose volume remains small in comparison to market daily volume, can move the price by a large fraction of the daily volatility, then it is clear that one should not put too much faith in the reliability and resiliency of market prices. As we argued above, this market fragility is the result of continuous market clearing, that leads to the following property :

The price is the point at which a *vanishing* supply meets a *vanishing* demand.

An implication of particular importance is the relevance of mark-to-market accounting rules for the valuation of large portfolios (so as to leave aside the problem of liquidation costs that usually enters in the discussion, one can think of the assets of an insurance firm that are kept until expiry). From the previous discussion, it is meaningless – or even dangerous – to mark too closely portfolios to the market prices, as some noise traders, fat-fingers or even ill-intentioned manipulators can trigger large re-balancing with not-so-large volumes, resulting in inappropriate profits and in unstable prices.

It would be more significant to assess market prices on the basis of the local supply and demand (around the market price) so as to smooth out fluctuations. A practical way to do so is to monitor the market *liquidity* $\mathcal{L}$, that is well proxied by the ratio $\sigma_D/\sqrt{V_D}$ (where σ_D is the daily volatility and V_D is the daily traded volume), as proposed and tested on the Bitcoin market in Donier and Bouchaud (2015a). Several other measures of liquidity might also be relevant, see Amihud and Mendelson (1986); Foucault et al. (2013); Corradi et al. (2015). In the case when the asset is not assumed to be kept until expiry but might incur liquidation costs, the monitoring of liquidity would be a good indicator of the *ex-post* value of the portfolio, once liquidated. This is the idea of impact-discounted mark-to-market value proposed and discussed in Caccioli et al. (2012).

7.6.3 Would batch auctions be beneficial ?

In order to curb potentially nefarious and socially wasteful HFT activities, a possibility that is currently hotly debated Budish et al. (2013) is to change the continuous trading system to frequent batch auctions, that would occur in discrete time with a time interval between auctions of order 100ms-1s. While we are not necessarily convinced that high frequency trading is such an evil [19] and the investment in speed technologies is such a waste on the long run, we believe that the issue of market stability should be of primary concern. Our theory provides a natural framework for studying the effect of such market design changes on the supply and demand curves, and on the stability of the resulting price.

From the above analysis, we concluded that the singularity at the immediate vicinity of the price is regularized when auctions occur in discrete time τ, leading to a reduced price impact

19. The overall profit of HFT firms was estimated to be around 5B$/year at its peak in 2010. This corresponds to an estimated cost of 1 basis point (10^{-4}) per transaction in the US equity markets, compatible with a fraction of the average bid-ask spread in the same period Hendershott et al. (2011). This is probably at least 10 times smaller than the profits of "old school" market makers : the average bid-ask spread on US markets was fluctuating around 60 basis points (!) from 1900 to 1980, before declining sharply Jones (2002).

(linear instead of square-root). However, one should not rejoice too fast, since we also saw that this regularisation only concerns very small volumes, less than the average volume traded in the market during time τ. In order to have a substantial effect and reduce the impact of trade size Q of – say – 1−10% of the average daily volume, the inter-auction time should be, unsurprisingly, of the order of five minutes to one hour and not milliseconds or even seconds. Trying to improve market stability through frequent batch auctions does not seem very useful in the light of our theory, except if one accepts to clear the market every hour or so, which would potentially imply other problems, such as and new source of liquidity risk and the corresponding development of secondary markets where transactions would take place in-between auctions.

7.7 Conclusion

In this paper, we have developed a fully dynamic theory of liquidity, based on weak and general assumptions on investors' behaviours : in a nutshell, heterogeneous reactions to incoming news of a large number of "infinitesimal" investors. Addressing the inability of the Walrasian theory to take transactions into consideration, we allow for auctions to clear the market periodically, and show how the market clearing mechanism itself affects the structural properties of supply and demand. In the case when the time between auctions is very large, we recover classical Walrasian auctions, in which market prices and liquidity are determined by the long-term (im-)balance between the incoming supply and the incoming demand. When auctions are allowed to happen at high frequency, the liquidity around the price mechanically vanishes, which leads to an anomalous, square-root impact of small orders that increases with market *volatility*. This accounts for the universally observed square root impact of small orders on modern financial markets and on the Bitcoin. In order to obtain a direct confirmation of the theory, we measured the shape of the Bitcoin order book (that appears to be a faithful reflection of low frequency supply and demand, at least close enough to the price), which indeed displays a striking "V-shape" for the marginal supply and demand (see Figure 7.6).

Our results highlight an apparent paradox : the more frequent the transactions are allowed to occur (thereby increasing, in theory, market efficiency), the more *fragile* the resulting price is ! In continuous double auctions markets, the price can be seen as the point at which a *vanishing* supply meets a *vanishing* demand, challenging the Platonian view of financial markets that prices are well-defined and stable. Our framework allows us to draw two further conclusions of general interest. First, the local estimates of supply and demand needs to be taken into account when one watches market prices, with important implications on portfolio valuation and stability monitoring Kyle and Obizhaeva (2012); Donier and Bouchaud (2015a). Second, as soon as the reaction to incoming news is not unbiased at all times, the volatility of the market price exceeds the fundamental volatility, embedding Shiller's famous excess volatility puzzle Shiller (1980) in an interesting formal framework.

Although we only considered market design and price stability within a particular angle, we believe that our framework can be extended to address many other practical problems, such as the market maker's problem, cross-impact between markets, optimal execution issues or the effect of taxation and other changes in market design. We also restricted to stationary market conditions, but our formalism can be readily adapted to include fluctuations in liquidity and/or market volatility (that would translate in some stochastic evolution of $\mathcal{D}$). Finally, the surprisingly good agreement between theory and empirical data suggests to extend our set of hypotheses to agent-based models of markets (cf Section 7.3.2), with the aim of producing realistic emergent properties from microscopic agent behaviour. More generally, our results vindicate an approach to dynamics in economic sciences as resulting from complex interactions between many heterogeneous agents, that may – or may not – be rational, in the general vein of, e.g. Hommes (2006); Gualdi et al. (2015b) and refs. therein. The resulting partial differential equation that governs the evolution of agents' preferences offers a much deeper level of understanding than simply postulating ad-hoc stochastic models for prices. We believe that this modelling strategy can be extended to many other situations of economic relevance.

Acknowledgements

We warmly thank M. Abeille, J. Bonart, R. Cont, J. De Lataillade, D. Delli Gatti, M. Gould, T. Hendershott, J. Kockelkoren, C. A. Lehalle, Y. Lempérière, I. Mastromatteo, M. Potters, I. Rosu, D.Thesmar and B. Tóth for many crucial discussions and collaborations on these issues – and in particular G. Zerah for his help on the Wiener-Hopf method. We also thank A. Tilloy for useful remarks on the manuscript.

7.8 Postface (français)

Tout comme l'article précédent, cet article présente deux niveaux de lecture. Le premier niveau concerne ses résultats directs. Basé sur les résultats des Chapitres 5 et 6, il développe une théorie générale dynamique de l'offre et de la demande qui dépasse la dichotomie ordres limites/ordres marché ainsi que le concept même de carnet d'ordres. Il met en évidence les conséquences non triviales du processus de trading (ou d'échanges) lui-même sur la structure de l'offre et la demande, et effectue des prédiction sur la liquidité pour tout mécanisme de marché (notamment les *batch auctions* si populaires en ce moment). Il jette au passage un nouveau regard sur l'excès de volatilité des prix observé par Shiller, et donne une compréhension nouvelle de la dynamique des prix en mettant en lumière une règle de *pricing* à la fois simple et intuitive – du moins dans le cas d'agents infinitésimalement petits. Toutefois, tout n'est pas fini. Nous avons vu dans le chapitre précédent la nécessité – et la difficulté – d'inclure des agents non infinitésimaux pour rendre compte de l'impact :

en un sens, en les ignorant, ce chapitre n'obtient qu'un résultat limite [20]. La tâche n'est toutefois pas simple : il faudrait alors à la fois réintégrer les méta-ordres du chapitre précédent, et y ajouter des ordres limite de taille macroscopique. Cela ferait apparaître des inefficiences, et nombre de comportements particuliers et de propriétés subtiles de l'équilibre des marchés à cette échelle de granularité devraient alors être pris en compte [21], dans un mélange écono-financiaro-physicien des plus délicats – mais peut-être un tel modèle global serait-il trop complexe pour être utile.

Le second de niveau de lecture s'insère quant à lui dans le message global de cette thèse, et concerne plus généralement la manière d'aborder la modélisation des systèmes économique. Le modèle de ce chapitre a été présenté de deux façons : la première résumant la dynamique des intentions d'achat et de vente à son aspect purement statistique (sans structure particulière sous-jacente ni prise en compte des comportements), la seconde se basant sur des utilités – qui au vu des hypothèses ne se révèlent que des coquilles vides qui disparaissent dans le calcul, comme le souhaitait Poincaré ! Les deux approchent produisent les mêmes équations et les même dynamiques : l'agent hétérogène a donc permis, dans un certain périmètre, d'unifier rationalité et irrationalité, en produisant des résultats indifférents à l'hypothèse choisie. Cela ne devrait pas nous étonner : ce qui paraît rationnel à l'un, semble souvent irrationnel à l'autre, et l'*idiosyncrasie* semble pouvoir transcender les approches. Le choix de l'approche n'est cependant pas transparent pour autant : il aura son influence sur les valeurs des paramètres, en produisant par exemple des niveaux d'idiosyncrasie et des synchronisations globales différents. Le physicien se réjouira donc d'avoir capturé un phénomène essentiel des systèmes humains, et l'économiste de pouvoir mesurer l'impact de ses paramètres d'entrée (par exemple les hypothèses formulées sur les utilités des agents) sur le système (essentiellement, la liquidité et la volatilité) – à des fins de régulation, par exemple.

Une des conclusions les plus importantes, est que l'agent hétérogène a permis de reproduire des faits stylisés hautement non-triviaux que les modèles à agents représentatifs semblent incapables de reproduire [22]. L'élément crucial qui donne toute sa chair à la présente théorie, est l'articulation *entre* les agents – plus précisément, leur agencement dans les courbes d'offre et de demande – que l'agent représentatif, par définition, ne permet pas de prendre en compte. L'agent hétérogène apparaît donc comme une base nécessaire dans la modélisation des systèmes économiques de grande taille. Les adeptes de la rationalité se dirigerons alors vers les *Mean-Field Games*, et les autres vers des modèles à agents hétérogènes heuristiques. Entre les deux, la présente étude ne permet pas de trancher, et il semblerait que le choix doive pour quelques temps encore rester idéologique. Il me semble cependant intéressant de voir les différences que les deux approches peuvent produire dans

20. Il me semble juste de dire que nous avons inclus la partie *permanente* de l'impact mais il manque dans ce modèle toute la partie *transitoire* qui se trouve absorbée dans l'infinitésimalité : au Chapitre 6, nous avions dû intégrer un agent macroscopique pour la faire émerger.

21. Quelques mots-clés : *selective liquidity taking*, *market making*, *optimal execution*, *fair pricing* etc.

22. A partir du moment où l'on se concentre sur *un* acteur particulier et que l'on cherche à assimiler le comportement du marché à son comportement individuel, on semble destiné à obtenir un impact localement linéaire.

ce cadre commun : le chapitre suivant est une ébauche de recherche dans cette direction – qui je l'espère profondément continuera à être suivie.

Quatrième partie

Mise en perspective

Chapitre 8

Un cadre d'études pour les modèles multi-agents en économie et en finance

8.1 Préface

Si l'on résume et que l'on prend un peu de recul par rapport à l'article précédent (voire même toute la partie précédente), on se rend compte qu'elle délivre un message simple :

Les EDP[1] *sont des outils pertinents pour décrire les systèmes à agents.*[2]

Cela donne assez envie d'effectuer un petit bond logique, et d'affirmer également :

Les EDP sont des outils pertinents pour décrire les systèmes économiques *à grand nombre d'agents.*

Cette affirmation, qui du reste ne fonctionne que pour les systèmes continus (à agents infinitésimaux par exemple) résume en réalité plusieurs affirmations sous-jacentes (qui, elles, sont supposées rester valables pour des systèmes discrets) :

1. L'ensemble de l'offre et de demande agrégée sur la gamme de prix $(0, +\infty)$ doit être considérée pour comprendre la dynamique du prix et sa liquidité. En particulier, il ne suffit pas de modéliser le prix tout seul comme un processus réel[3].
2. Le temps doit *avancer*[4], en permettant simultanément (i) l'arrivée de nouvelles informations

1. Equations aux dérivées partielles.

2. Vous me direz, il y a bien des EDP dans Black-Scholes !... Mais il s'agit conceptuellement de deux choses différentes : dans un cas (Black-Scholes) la fonction régie par l'EDP est une densité de probabilité sans existence physique décrivant un processus stochastique, dans l'autre (modèles multi-agents) il s'agit d'une densité réelle d'agents de laquelle on peut faire disparaître tout aspect stochastique.

3. Même si dans le cadre du modèle du chapitre 6, l'Appendice A montre qu'une dualité existe entre courbes d'offre/demande et prix. Il ne s'agit là que d'un cas exceptionnel, en principe, qui a d'ailleurs été trouvé en partant de l'offre et de la demande.

4. De manière continue de préférence, mais des modèles à temps discrets peuvent être imaginés.

(donc des changements d'avis des agents) et (ii) le processus de *clearing* à l'oeuvre, le tout dans un cadre dynamique.

3. Les agents doivent être *hétérogènes*[5], notamment dans leur appréciation des prix (et potentiellement, dans tout le reste : fonctions d'utilité, etc).

4. Les intentions exécutées doivent *disparaître* des courbes d'offre/demande[6].

Le dernier point est relativement subtil. On voit bien que moyenner (et faire diffuser) des fonctions linéaires ne peut donner que des fonctions linéaires (comme dans Kyle et al. (2014) par exemple). Dès que les intentions exécutées disparaissent en revanche (ou réapparaissent plus loin, comme nous le verrons ci-dessous), les courbes d'offre/demande individuelles deviennent fortement non-linéaires, ce qui se traduit de manière agrégée par une offre et une demande quadratique autour du prix !

Les ingrédients ci-dessus devraient suffire à construire un modèle à agents réaliste au niveau de la dynamique (et donc produisant la bonne structure de l'offre et de la demande agrégées), même dans le cas où eux-ci ne sont pas infinitésimaux ou lorsque ceux-ci ont des *stratégies*. Comme évoqué ci-dessus, il s'agit d'une question complexe techniquement, et nous serons heureux de nous restreindre au cas des agents infinitésimaux où des EDP décrivent le système lorsque nous chercherons des résultats analytiques. Notons au passage que, comme le montrera l'Appendice A, il peut s'agir d'EDP différentes de celle de diffusion évoquée jusqu'ici : par exemple, des diffusions fractionnaires. Notez également que ces quatre critères ne supposent rien de la rationalité (ou non) des agents. Il ne s'agit donc pas d'un modèle mais simplement d'un cadre dans lequel de nombreux modèles peuvent être (ré-)étudiés. La suite de ce chapitre a pour but d'introduire quelques-uns de ces modèles, en supposant des agents rationnels – ou non – et dans des cadres différents de celui des marchés financiers. Nous tâcherons en particulier d'étudier l'influence des paramètres d'entrée – que ce soit pour les agents rationnels ou les autres – sur les paramètres globaux du système, et éventuellement, sur son comportement macroscopique.

Notez enfin que ce chapitre est plus une ouverture sur des pistes de recherche à venir qu'une fin en soi. Les intuitions y sont préférées aux résultats – même si la plupart d'entre elles s'inspirent des résultats « rigoureux » des chapitres précédents. Certaines sont peut-être même fausses, mais toutes à mon avis sont suffisamment intéressantes pour valoir la peine d'être considérées.

---

5. Certains diraient : « *they agree to disagree* ».

6. Cela est plus simple à comprendre lorsque les agents sont *localisés*, c'est-à-dire qu'ils doivent matérialiser leur courbe d'offre/demande (potentiellement continue) en un (ou plusieurs) ordres élémentaires. Cela prend tout sons sens en-dehors de la finance, où on n'achète en général qu'un seul lave-linge à la fois – et on ne le revend pas immédiatement.

8.2 Schématique du cadre d'études

8.2.1 Schéma général

Continuons à prendre du recul par rapport au modèle du Chapitre 6, et commençons par dessiner le schéma le plus basique de modèle à agents qui permette de modéliser un système financier.

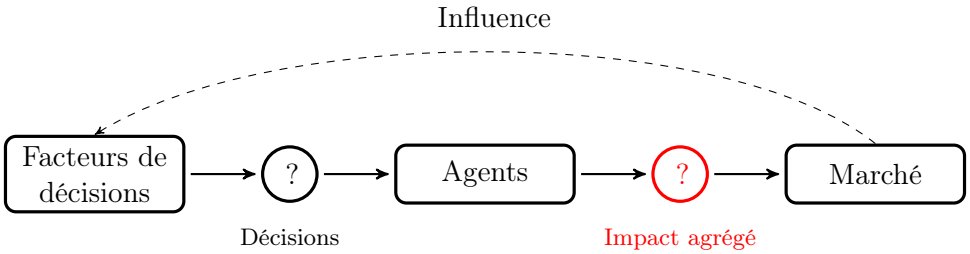

FIGURE 8.1 – Schéma agrégé d'un modèle à agents.

Etendons-nous brièvement sur le schéma 8.1, et détaillons-le en partant de la fin remontant les flèches. Notre objet d'études principal, tout à droite, est le marché. Comme argumenté tout au long de cet ouvrage, celui-ci n'est rien d'autre qu'une synthèse des actions des agents qui y agissent – toute la question étant bien entendu *comment*, nous y reviendrons. Ces actions sont enfin décidées en fonctions d'éléments que nous appellerons des *facteurs*, qui peuvent être exogènes – des *informations* provenant du monde réel – ou endogènes – c'est-à-dire, provenant du marché lui-même.

Toutefois, en présence de plus d'un agent, ce schéma n'est pas très informatif : l'ensemble de la complexité est cachée dans les fonctions de transfert. Il nous faut donc, pour commencer, le détailler au niveau individuel, toujours de la manière la plus générale possible :

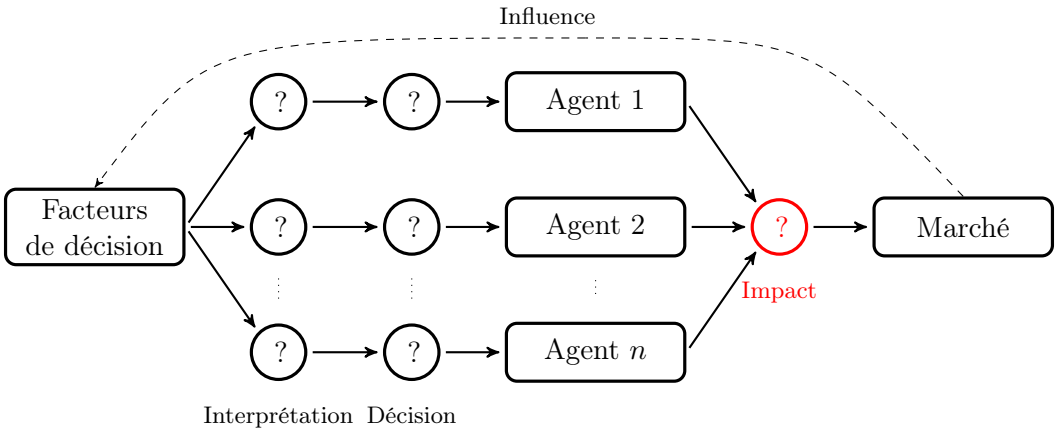

FIGURE 8.2 – Schéma détaillé d'un modèle à agents.

Ce changement de niveau de lecture nous a permis de développer le processus de décision pour chaque agent en deux parties :

i. L'interprétation des facteurs de décision, qui encode à la fois quels facteurs rentrent en compte

dans les décisions de l'agent i et la manière dont il les perçoit,

ii. La décision elle-même, que l'on pourrait appeler *stratégie*.

Notez que dans un modèle à agents représentatifs, les fonctions d'interprétation et de décision sont considérées communes à l'intérieur de chaque classe d'agents, ce qui permet de simplifier grandement le schéma et – quand le setup est suffisamment simple – d'obtenir des formules fermées. Toutefois, si les vertus pédagogiques d'une telle approche sont indéniables, il n'en est pas de même de leur réalisme, et de leur pertinence pour décrire le monde réel.

8.2.2 Schéma du modèle du Chapitre 6

Essayons maintenant de schématiser le modèle présenté aux Chapitres 6 et 7 sous une forme similaire :

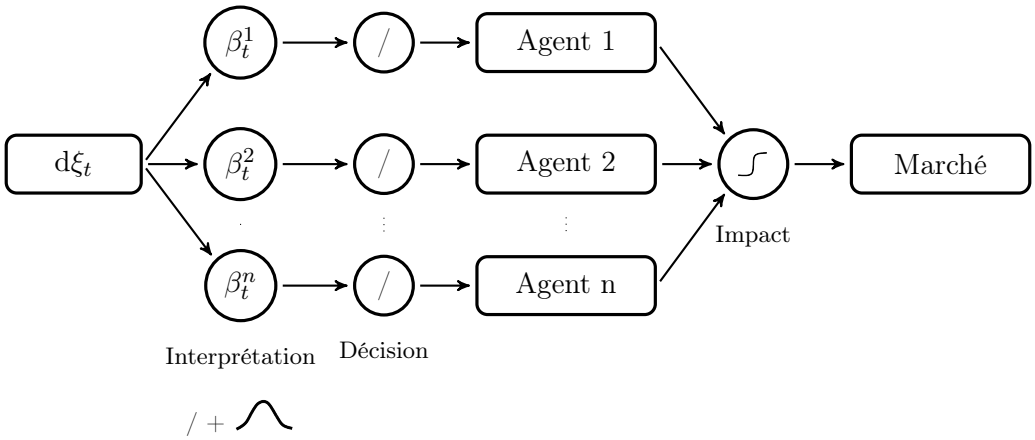

FIGURE 8.3 – Schéma du modèle des Chapitres 6 et 7.

Plusieurs choses ont changé par rapport au schéma général, ce qui met en avant les hypothèses effectuées :

i La rétroaction possible du marché sur les facteurs de décision est ignorée,

ii Les facteurs sont interprétés de manière *hétérogène* par les agents, ce qui est résumé par un coefficient β_t^i pour chaque agent qui n'est pas forcément égal à 1 (1 correspondant à une interprétation parfaite des signaux/informations), auquel s'ajoute une erreur idiosyncratique gaussienne,

iii Les décisions sont tout simplement linéaires, i.e. il n'y a aucune notion de *stratégie* : si un agent pense que le prix d'un bien augmente, il révise d'autant ses intentions d'achat ou de vente (même si nous avons vu au Chapitre 7 qu'introduire des fonctions d'utilité ne change rien au phénomène décrit),

iv Une hypothèse qui n'apparaît pas sur ce schéma est également $n \to \infty$ (n étant le nombre d'agents) ainsi que l'hypothèse que chaque agent est *petit* (au sens de la fraction du capital total dont il dispose).

Si ces hypothèses nous ont permis d'obtenir des résultats analytiques et de développer des intuitions, elles ne sont pas satisfaisantes pour autant. Il est donc temps de prendre un peu de hauteur, en laissant de côté la contrainte de la tractabilité mathématique, pour réfléchir à ce que serait un modèle de marché réaliste.

8.2.3 Simuler un marché

Cette section se donne un objectif précis : en se basant sur tout ce qui a été fait jusqu'à présent, elle se veut de proposer une manière de réaliser un simulateur de marché réaliste du point de vue de l'impact mais également au niveau de l'écosystème et de la microstructure réels des marchés financiers. Les enjeux sous-jacents sont triples :

1. **Pour le praticien :** Permettre de backtester les stratégies d'investissement en prenant en compte son propre impact sur le marché, ce qui est impossible sur des données historiques – non impactées par définition. En fonction du type de stratégie, prendre l'impact en compte peut s'avérer crucial dans les décisions d'investissement car celui-ci peut être à l'origine de coûts importants.
2. **Pour le régulateur :** Tester à coût faible la réponse du marché à un changement réglementaire. Cela pourrait du moins précéder ou remplacer certaines expérience pilotes, parfois coûteuses et compliquées à mettre en place.
3. **Pour le chercheur :** Développer un cadre d'études générique pour étudier les systèmes composés de nombreux agents qui capture correctement leur structure et leur dynamique, en finance et en-dehors (le paragraphe 8.5 traitera l'exemple du *yield management*).

Le modèle développé jusque là est clairement insuffisant pour chacun de ces points, car trop simpliste. Toutefois, je crois qu'il capture de manière unique un aspect essentiel de la dynamique de l'offre et la demande, et de leur structure à l'échelle locale : il faudrait donc en extraire le coeur – i.e. garder ses propriétés concernant l'impact – et l'intégrer à un modèle plus riche, où des propriétés comme la dynamique des queues, le tick, la distribution des volumes, la selection adverse, l'impact permanent,... seraient également reproduites.

Conjecture 8.2.1. *Le schéma 8.4 permet de répondre au cahier des charges ci-dessus.*

L'intuition derrière cette conjecture est la suivante : le modèles des Chapitres 6 et 7 nous a fait comprendre que l'impact concave « en racine » provenait – ou plutôt, pouvait provenir – de l'*hétérogénéité* entre les agents à cause de laquelle les courbes d'offre et de demande se vident au voisinage du prix – les agents s'y trouvant réalisant des transactions les uns avec les autres et en disparaissant.[7] Cette hétérogénéité n'est pas incompatible avec la notion de stratégie, exclue à

7. Mathématiquement, qui permettait d'écrire une équation d'évolution de type EDP avec condition absorbante au prix, résultant immédiatement en un carnet d'ordres linéairement croissant au voisinage du prix – sous certaines conditions de régularité.

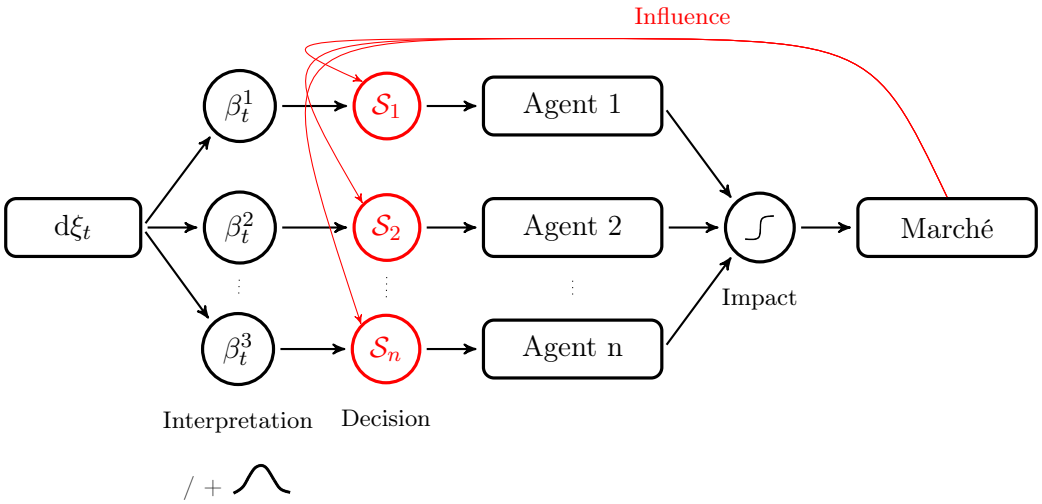

FIGURE 8.4 – Simulateur de marché. $\mathcal{S}_i$ est la stratégie de l'agent i (qui peut être une stratégie précise ou une meta-stratégie, par exemple : broker, market maker, arbitrageur, trader à haute fréquence...).

dessein jusqu'ici. Cette approche a d'ailleurs dû sembler étrange à certains habitués des modèles d'économie financière où la stratégie est reine, et le reste simplifié à l'extrême : mais nous pouvons maintenant les réconcilier, et à l'intérieur du cadre développé dans cette thèse (i.e. du schéma 8.4) qui apporte l'impact, les stratégies ont beaucoup à apporter – i.e. tout le reste.

On voit maintenant l'intérêt d'avoir découpé le problème en deux parties, et de s'être concentré sur le problème « ouvert » de l'impact d'une action sur le marché : car maintenant que le cadre est établi et que l'impact n'est plus une question – il sera toujours plus ou moins réaliste maintenant, quoi que l'on fasse – il ne nous reste plus qu'à nous concentrer sur les stratégies – et par là, « fermer » le modèle [8]. Les stratégies deviennent donc la question cruciale – et plusieurs décennies de littérature économique et financière peuvent à présent nous inspirer. De l'agent rationnel à l'agent à rationalité limitée, de l'agent déterministe à l'agent stochastique, de l'agent averse au risque à l'agent en excès de confiance, de l'agent précurseur à l'agent suiveur, de l'agent fondamental à l'agent chartiste – les possibilités sont infinies, et chacun pourra les exploiter comme il l'entend. A partir du moment où il n'oublie pas l'ingrédient fondamental : *les agents sont hétérogènes.*

Nous sommes arrivés au plus haut dans la prise de hauteur, et approchons par la même occasion la fin de cette thèse. Avant de changer de chapitre et de jeter un oeil plus macroscopique sur la notion de liquidité, il nous reste juste à illustrer la rhétorique ci-dessus par quelques exemples.

8. Ou tout simplement, le créer, car nous avons jusqu'ici un cadre d'études sans réel modèle implémenté à l'intérieur, ou juste un modèle trivial.

8.3 Un exemple de modèle à agents rationnels

Imaginons un marché dans lequel les agents suivent une politique d'investissement optimal en fonction de leurs prédicteurs de prix – mais supposons que ceux-si sont hétérogènes. Les investisseurs sont rationnels dans le sens où, sachant leurs prédicteurs, ils suivent la stratégie d'investissement qui maximise leurs profits. En revanche, supposons que les investisseurs sont en *excès de confiance* par rapport à leurs prédicteurs respectifs, et pensent chacun mieux prédire les prix futurs que la moyenne. [9]

Supposons que chaque investisseur ait un coût associé à une décisions d'investissement, et que ces coûts soient linéaires en la quantité échangée – i.e. des coûts par unité de capital investi, qui peuvent également être hétérogènes parmi la population. Ces derniers peuvent représenter des coûts de *spread*, des frais de transactions sur les plateformes de marché, ou résulter plus indirectement de seuils de profitabilité exigés par les investisseurs. [10] Supposons enfin que les prédicteurs des investisseurs soient tels qu'ils oscillent autour de zéro, c'est-à-dire, que leur évaluation du prix oscille autour du prix du marché lui-même. Cela peut provenir d'un mécanisme de régulation exogène – e.g. les investisseurs trop loin de la « vérité » se rendent compte de leur erreur et finissent par s'en rapprocher – ou tout simplement d'un mécanisme endogène où les investisseurs tendent à se rapprocher du prix du marché par peur de se tromper eux-mêmes. [11] Les trois hypothèses ci-dessus se résument en le modèle suivant :

1. Les agents sont infinitésimaux et leur nombre $n \to \infty$, comme dans le chapitre précédent (ils n'ont donc pas d'impact),
2. Les prédicteurs p_t^i des agents suivent des processus mean-reverting dans le référentiel du prix, par exemple Ornstein-Uhlenbeck centrés en zéro et de force de rappel α^i et de volatilité σ (choisie égale pour tous les investisseurs, pour des raisons de tractabilité),
3. Les agents subissent des coûts (réels ou effectifs) proportionnels associés à leur investissements, notés c^i,

9. De manière évidente, il est impossible que chacun prédise mieux l'avenir que la moyenne. Comment les investisseurs peuvent-ils rester en excès de confiance s'ils perdent de l'argent ? La réponse est double : une est comportementale – c'est la tendance à penser que les gains sont obtenus par le talent alors que les pertes ne sont qu'une mauvaise chance – et l'autre est qu'ils ne perdent pas forcément de l'argent directement : puisque le marché monte en moyenne, tout le monde gagne – seulement, certains gagnent moins que le marché et s'en satisfont.

10. Supposons qu'un investisseur ne souhaite rebalancer son portefeuille que mensuellement, et qu'il souhaite réaliser un *ratio de Sharpe* annualisé supérieur ou égal à 1 sachant que la volatilité annuelle de l'investissement en question est de 24%. Il devra donc réaliser un gain minimum de $\frac{24*1}{12} = 2\%$ sur chaque investissement mensuel, faute de quoi il aura bloqué du capital pendant 1 mois pour un trop faible arbitrage. Cette contrainte peut être incorporée simplement en ajoutant un coût effectif d'investissement de 2% – ce qui est bien supérieur aux coûts de spread et aux frais de transaction. En des mots simple, cette contrainte permet d'allonger la durée d'investissement à rentabilité fixée.

11. Ces deux situations sont très intéressantes : dans la première, le prix sera toujours proche de la « valeur fondamentale », alors que dans la seconde il peut s'en éloigner tout en restant dans une situation auto-cohérente – cela nous rappelle le concours de beauté *à la* Keynes.

4. Les agents ont des limites d'inventaire strictes X^i, et suivent chacun une stratégie optimale *à la* De Lataillade et al. (2012).

La loi jointe des variables α, c et x est notée $q(\alpha, c, x)$ – les variables n'étant pas nécessairement indépendantes. Nous nous placerons pour tout ce qui suit dans le référentiel du prix. La stratégie d'investissement optimal *à la* De Lataillade et al. (2012) est simple : lorsqu'un investisseur dont le prédicteur suit un processus d'Ornstein-Uhlenbeck subit des coûts proportionnels aux volumes qu'il engage, sa stratégie optimale est de ne rien faire jusqu'à ce que son prédicteur dépasse un certain seuil $\theta(\alpha, c)$, auquel cas il investit l'intégralité de son capital disponible. De la même manière, il garde ensuite son investissement jusqu'à ce que son prédicteur dépasse un seuil négatif – auquel cas il revend tout immédiatement. A tout instant, chaque agent a donc en tête un prix auquel il achèterait instantanément son volume maximal autorisé (ou revendrait s'il a déjà acheté). L'agrégation de ces intentions d'achat ou de vente est un des exemples les plus simples de carnet d'ordres latent. Un échange a alors lieu dès que deux ordres latents de signes opposés se rencontrent au même prix.

Lorsqu'une quantité δ est échangée, les acheteurs concernés se retrouvent donc collectivement avec une quantité supplémentaire δ à vendre, et les vendeurs avec une quantité supplémentaire δ à acheter. Pour comprendre cela, et comment ces nouvelles intentions intègrent le carnet d'ordres, mettons-nous à la place d'un de ces acheteurs : s'il vient d'acheter, c'est parce que son prédicteur a atteint son seuil θ_i – il estime donc, dans le référentiel du prix, que l'actif vaut θ_i. Son seuil de revente est alors $2\theta_i$: son prix estimé θ_i, *plus* son seuil de revente qui vaut également θ_i. Par conséquent, un acheteur de seuil θ_i replace immédiatement un ordre de vente à un prix $2\theta_i$. Symétriquement, un vendeur de seuil θ_i replace immédiatement un ordre d'achat à un prix $-2\theta_i$. En appelant $\lambda_{A/B}(\theta, t)$ la distribution des seuils parmi les investisseurs qui réalisent des transactions au temps t, pondérées par leurs inventaires[12], on obtient immédiatement le système d'équations suivant pour décrire la dynamique du carnet d'ordres, comme pour l'Equation 6.1 :

$$\frac{\partial \rho_B(x,t)}{\partial t} = \frac{\partial}{\partial x}\alpha(x)\rho_B(x,t) + \frac{\sigma^2}{2}\frac{\partial^2 \rho_B(x,t)}{\partial x^2} + \frac{1}{2}\lambda_B(x/2,t) - \kappa R_{AB}(x,t); \qquad \text{(1-a)}$$

$$\frac{\partial \rho_A(x,t)}{\partial t} = \underbrace{\frac{\partial}{\partial x}\alpha(x)\rho_A(x,t) + \frac{\sigma^2}{2}\frac{\partial^2 \rho_A(x,t)}{\partial x^2}}_{\text{a- Drift - Diffusion}} + \underbrace{\frac{1}{2}\lambda_A(x/2,t)}_{\text{b- Deposition}} - \underbrace{\kappa R_{AB}(x,t)}_{\text{c- Reaction}}; \qquad \text{(1-b)}$$

où les conventions sont les même que dans le Chapitre 6, et $\alpha(x)$ dépend entre autres du processus mean-reverting choisi (ici Ornstein-Uhlenbeck) et de la distribution des α : si ceux-ci sont identiques pour tous les agents, elle vaut simplement $\alpha(x) = \alpha x$. Comme précédemment, l'équation sur $\varphi = \rho_A - \rho_B$ fait disparaître le terme de réaction, et donne simplement :

$$\frac{\partial \varphi(x,t)}{\partial t} = \frac{\partial}{\partial x}\alpha(x)\varphi(x,t) + \frac{\sigma^2}{2}\frac{\partial^2 \varphi(x,t)}{\partial x^2} + \frac{1}{2}\lambda(x/2,t). \qquad (8.1)$$

12. Qui dépend « uniquement » des paramètres et de la condition initiale...

Nous retrouvons donc une équation très similaire à l'Equation 6.2, mais dans laquelle le taux d'annulation qui permettait de stabiliser le système a été remplacé par un drift de rappel vers le prix par le biais de $\alpha(x)$. Dans le cas où la condition initiale est symétrique autour du prix, la solution stationnaire se calcule comme une combinaison linéaire des solutions stationnaires obtenues pour chaque sous-ensemble d'investisseurs de même seuil θ (car dans ce cas, on peut faire en sorte de réaliser les transactions de manière à ce que chaque investisseur ne réalise ses transactions qu'avec des investisseurs de même seuil !), pondérées par leur taille.

Jetons un coup d'oeil aux résultats du problème similaire – mais plus facilement soluble – où $\alpha(x) = \alpha_0 \left(\mathbb{1}_{x\geq 0} - \mathbb{1}_{x<0}\right)$, dont les résultats sont extrêmement intéressants et méritent d'être développés plus en détail. [13] Pour une valeur de seuil θ fixée, l'équation stationnaire est tout simplement :

$$\begin{cases} \dfrac{\partial \varphi_\theta(x,t)}{\partial t} = \alpha_0 \dfrac{\partial \varphi_\theta(x,t)}{\partial x} + \dfrac{\sigma^2}{2}\dfrac{\partial^2 \varphi_\theta(x,t)}{\partial x^2} + \dfrac{1}{2}\delta(x/2-\theta), \\ \dfrac{\partial \varphi_\theta(x,t)}{\partial t} = -\alpha_0 \dfrac{\partial \varphi_\theta(x,t)}{\partial x} + \dfrac{\sigma^2}{2}\dfrac{\partial^2 \varphi_\theta(x,t)}{\partial x^2} - \dfrac{1}{2}\delta(x/2+\theta), \\ \varphi_\theta(0^+,t) = \varphi_\theta(0^-,t) \end{cases}$$

Si les conditions initiales sont symétriques, alors par symétrie $\forall t, \varphi_\theta(0^+,t) = \varphi_\theta(0^-,t) = 0$ et l'on peut résoudre séparément les équations pour $x \geq 0$ et $x \leq 0$. Concentrons-nous donc sur $\mathbb{R}_+$. La solution stationnaire obéit à l'équation simple

$$\begin{cases} \alpha_0 \dfrac{\partial \varphi^{st}_\theta(x)}{\partial x} + \dfrac{\sigma^2}{2}\dfrac{\partial^2 \varphi^{st}_\theta(x)}{\partial x^2} + \dfrac{1}{2}\delta(x/2-\theta) = 0 \\ \varphi^{st}_\theta(0) = 0, \end{cases} \tag{8.2}$$

et s'écrit donc simplement comme :

$$\varphi^{st}_\theta(x) = \begin{cases} \dfrac{\sigma^2}{2\alpha_0}\left[1 - e^{-\frac{2\alpha_0}{\sigma^2}x}\right], & 0 \leq x < 2\theta, \\ \dfrac{\sigma^2}{2\alpha_0}\left[e^{\frac{4\alpha_0\theta}{\sigma^2}} - 1\right] e^{-\frac{2\alpha_0}{\sigma^2}x}, & 2\theta \leq x. \end{cases} \tag{8.3}$$

On vérifiera que $\frac{\partial}{\partial x}\varphi^{st}_\theta(2\theta^+) - \frac{\partial}{\partial x}\varphi^{st}_\theta(2\theta^-) = \frac{\partial}{\partial x}\varphi^{st}_\theta(0)$, et donc que l'injection d'ordres à $x = 2\theta$ compense bien exactement la disparition d'ordres au niveau du prix. On peut facilement calculer l'intégrale de φ^{st}_θ sur $\mathbb{R}_+$, qui vaut $\int_{x\in\mathbb{R}_+} \varphi^{st}_\theta \mathrm{d}x = \frac{\sigma^2\theta}{\alpha_0}$.

Considérons maintenant le problème initial d'une population suivant une distribution de seuils (pondérés par les inventaires) $\lambda(\theta)$. La solution stationnaire du système global s'écrit comme une

13. En supposant que la stratégie optimale reste une stratégie à seuil *à la* De Lataillade et al. (2012) dans ce cas, ce qui semble vrai – mais que je ne prouverai pas.

combinaison linéaire des solutions pour chaque θ :

$$\varphi^{st}(x) = \int_{\mathbb{R}_+} C(\theta)\varphi_\theta^{st,0}(x)\mathrm{d}\theta, \tag{8.4}$$

où pour chaque θ, la valeur de $C(\theta)$ doit être telle que la quantité d'investisseurs dans $[\theta, \theta + \mathrm{d}\theta]$ est $\lambda(\theta)\mathrm{d}\theta$, c'est-à-dire telle que $C(\theta)\int_{x\in\mathbb{R}_+}\varphi_\theta^{st}\mathrm{d}x = \lambda(\theta)$. Cela donne directement $C(\theta) = \frac{\alpha_0}{\sigma^2}\frac{\lambda(\theta)}{\theta}$, et finalement :

$$\varphi^{st}(x) = e^{-\frac{2\alpha_0}{\sigma^2}x}\int_0^{\frac{x}{2}}\frac{\lambda(\theta)}{2\theta}\left[e^{-\frac{4\alpha_0\theta}{\sigma^2}} - 1\right]\mathrm{d}\theta + \left[1 - e^{-\frac{2\alpha_0}{\sigma^2}x}\right]\int_{\frac{x}{2}}^{\infty}\frac{\lambda(\theta)}{2\theta}\mathrm{d}\theta. \tag{8.5}$$

Si λ est continu et si $\lambda(0) = \lambda_0 > 0$, on peut calculer la forme de $\varphi^{st}(x)$ pour $x \to 0$, ce qui donne

$$\varphi^{st}(x) \underset{x\to 0}{=} \frac{\alpha_0\lambda_0}{\sigma^2}x\ln\left(\frac{1}{x}\right) + O(x), \tag{8.6}$$

et en particulier $\varphi^{st}(x) \underset{x\to 0}{\to} 0$. Notez que φ^{st} croît dans ce cas presque linéairement – mais pas tout à fait. L'impact instantané n'est donc plus exactement en racine, mais en est très proche [14]. [15] Si $\lambda(\theta)$ tend polynomialement vers zéro en revanche, ou s'annule sur un voisinage de zéro, alors on retrouve le carnet d'ordres linéaire des chapitres précédents :

$$\varphi^{st}(x) \underset{x\to 0}{=} f(\lambda)\frac{\alpha_0}{\sigma^2}x + o(x), \tag{8.7}$$

où $f(\lambda)$ est une constante qui dépend de la forme de la fonction λ. Si pour finir cette dernière a un comportement singulier $\lambda(\theta) \underset{x\to 0}{\sim} \theta^{-\gamma}$ où $0 < \gamma < 1$, alors le carnet d'ordres se comporte au voisinage de zéro comme :

$$\varphi^{st}(x) \underset{x\to 0}{=} f(\lambda)\frac{\alpha_0}{\sigma^2}x^{1-\gamma} + o(x^{1-\gamma}), \tag{8.8}$$

résultant en un impact instantané $I(Q) \sim Q^{\frac{1}{2-\gamma}}$ où $\frac{1}{2} < \frac{1}{2-\gamma} < 1$. La concavité de l'impact diminue donc avec l'augmentation massive de l'activité haute fréquence – ce qui pourrait expliquer que sur certains marchés financiers l'impact mesuré soit plus proche d'une loi puissance d'exposant $0.5 < \delta < 1$ (Mastromatteo et al., 2014a), que d'une racine carrée « pure » comme sur le Bitcoin (voir aussi l'Appendice A pour une explication possible basée sur le *grey brownian motion*).

Comme promis, nous pouvons également tirer des conclusions concernant l'influence des paramètres d'entrée sur le comportement global du système. En particulier, nous voyons que lorsque la

14. Pour ceux qui veulent vraiment savoir : $\mathcal{I}(Q) = Ce^{\frac{1}{2}W(-\frac{y}{e})}$ où W est la fonction W de Lambert qui a un comportement similaire à celui du logarithme à grand argument : $\mathcal{I}(Q)$ se comporte donc comme une racine lorsque Q n'est pas trop petit. Etant donné la précision des mesures d'impact sur les données, il est de toute manière difficile de trancher entre ces formes fonctionnelles.

15. Cette singularité logarithmique subsisterait-elle pour un drift Ornstein-Uhlenbeck (non singulier en 0) ? Il est sans doute possible de répondre à cette question, mais les calculs sont plus compliqués.

volatilité σ augmente, la liquidité au voisinage du prix diminue, résultant en un marché moins robuste à des petites perturbations. Le paramètre de drift α quant à lui a l'effet inverse, et a tendance à localiser la liquidité autour du prix, tout en augmentant de manière proportionnelle le nombre de transactions (et donc le taux de *turnover* du marché). Mais les conclusions les plus intéressantes concernent sans doute la distribution de seuils des investisseurs $\lambda(\theta)$. Si celle-ci n'est pas trop singulière, le carnet d'ordres conserve donc sa forme en V au voisinage du prix, garantissant un impact toujours approximativement en racine. En revanche, si la part en capital des investisseurs « haute fréquence » devient importante [16], ils renforcent le carnet d'ordre autour du prix, et l'impact devient moins singulier – i.e. il se rapproche d'une forme linéaire. Cette conclusion est intéressante du point de vue de la modélisation elle-même : on voit en effet qu'une description plus fine au niveau de l'agent peut permettre d'aboutir à des résultats légèrement différents et plus modulables que l'approche *mean-field* utilisée jusqu'à maintenant.

Pour finir cette section, penchons-nous sur un phénomène extrêmement intéressant qui apparaît lorsque l'équation 8.2 est modifiée de la manière suivante :

$$\begin{cases} \dfrac{\partial \varphi_\theta(x,t)}{\partial t} = \alpha(t)\dfrac{\partial \varphi_\theta(x,t)}{\partial x} + \dfrac{\sigma^2}{2}\dfrac{\partial^2 \varphi_\theta(x,t)}{\partial x^2} + 2\delta(x/2-\theta), \\ \dfrac{\partial \varphi_\theta(x,t)}{\partial t} = -\alpha(t)\dfrac{\partial \varphi_\theta(x,t)}{\partial x} + \dfrac{\sigma^2}{2}\dfrac{\partial^2 \varphi_\theta(x,t)}{\partial x^2} - 2\delta(x/2+\theta), \\ \varphi_\theta(0^+,t) = \varphi_\theta(0^-,t) \end{cases}$$

avec $\alpha(t)$ dépendant de l'intensité du trading $\frac{\sigma^2}{2}\frac{\partial \varphi_\theta(0,t)}{\partial x}$ (i.e. le nombre d'échanges par unité de temps) *via* la relation :

$$\alpha(t) = \alpha_0 + C\frac{\sigma^2}{2}\frac{\partial \varphi_\theta(0,t)}{\partial x} \tag{8.9}$$

Dans ce cas en effet, il est remarquable que l'équation obtenue est identique à celle considérée par Gualdi et al. (2015a) dans un contexte différent, et pour laquelle ils ont montré l'apparitions d'oscillations notamment de la quantité $L(t) \equiv \int_0^\infty \varphi_\theta(x,t)\mathrm{d}x$, qui représente ici la liquidité totale présente sur le carnet d'ordres. Des oscillations entre période de liquidité normale et périodes de liquidité faible apparaissent alors de manière endogène – c'est la première fois dans cette thèse que nous parlons de fluctuations de liquidité.

8.4 Un exemple de modèle à agents heuristiques

Oublions maintenant l'aspect d'optimisation, et ajoutons des comportements macroscopiques à nos agents. Nous avons vu dans la partie précédente qu'un comportement de retour à la moyenne (e.g. investisseurs influençables) rajoutait un terme en $-\frac{\partial}{\partial x}\alpha(x)\varphi(x,t)$, avec par exemple $\alpha(x) = \alpha x$.

16. En capital total et non pas en nombre de transactions : ce dernier est naturellement d'autant plus grand que leur taux élevé de turnover est élevé, mais cela n'a aucune influence ici !

Ce terme de retour au prix avait uniquement des effets sur la liquidité, et en particulier ne créait aucun effet de long terme sur le prix, ni ne produisait d'impact permanent pour les métaordres non informés. Dans cette section, nous allons étudier l'effet d'un comportement de *trend-following/ mean-reversion* – comportement que l'on pourrait qualifier d'irrationnel. Nous verrons que ceux-ci affectent de manière directe les prix – plus que la liquidité.

Reprenons le modèle du Chapitre 6, avec déposition uniforme selon une fonction indicatrice Θ et une annulation des ordres à taux ν. Supposons maintenant que tous les agents sont *trend-followers*, et que leur estimée de prix suit la dynamique suivante :

$$\begin{aligned} p^i_{t+\mathrm{d}t} \to & p^i_t + \beta^i \mathrm{d}\xi_t + \mathrm{d}W_{i,t} + \mu \mathrm{d}t \int_{-\infty}^{t} \frac{e^{-\frac{t-s}{\tau}}}{\tau} \mathrm{d}p_s \\ := & p^i_t + \beta^i \mathrm{d}\xi_t + \mathrm{d}W_{i,t} + \frac{\mu}{\tau}\mathrm{d}t\left(p_t - p^\tau_t\right), \quad p^\tau_t := \int_{-\infty}^{t} \frac{e^{-\frac{t-s}{\tau}}}{\tau} p_s \mathrm{d}s \end{aligned} \tag{8.10}$$

où nous avons décrit le trend-following de deux manières équivalentes [17] : en regardant les variations de prix passées et en les pondérant par un noyau décroissant de portée τ, ou en comparant le prix actuel aux prix passés sur une fenêtre d'ordre τ. Dans ce setup où tous les agents ont le même paramètre de trend-following, on ne fait que décaler le carnet d'ordres par rapport au processus sans trend-following. On peut donc écrire la dynamique de prix de manière autonome, en prenant la dynamique de prix que produit le modèle sans trend-following (par exemple celle du Chapitre 6) et en lui appliquant un certain filtre.

Plus précisément, si p^0_t est le processus sans trend-following (supposé connu) et p_t le processus avec trend-following, on peut écrire le système suivant :

$$\begin{cases} \mathrm{d}p_t = \mathrm{d}p^0_t + \frac{\mu}{\tau}\mathrm{d}t\left(p_t - p^\tau_t\right) \\ \mathrm{d}p^\tau_t = \frac{1}{\tau}\mathrm{d}t\left(p_t - p^\tau_t\right) \end{cases} \tag{8.11}$$

La résolution de ce système donne :

$$\mathrm{d}p_t = \mathrm{d}p^0_t + \frac{\mu}{1-\mu}\mathrm{d}t \int_{-\infty}^{t} \frac{1-\mu}{\tau} e^{-\frac{1-\mu}{\tau}(t-s)} \mathrm{d}p^0_s. \tag{8.12}$$

Lorsque $\mu = 0$ on retrouve bien le prix sans trend-following $p_t = p^0_t$. On voit bien sur cette équation que le trend-following amplifie l'effet des variations passées en le répercutant sur le présent. On peut même étudier la réponse d'une perturbation ponctuelle sur les prix futurs : en posant $\mathrm{d}p^0_t = \Delta p \delta_{t=0}$, on obtient :

$$\begin{cases} p^0_t = \Delta p, \quad t \geq 0 \\ p_t = \Delta p + \frac{\mu}{1-\mu}\left[1 - e^{-\frac{(1-\mu)t}{\tau}}\right]\Delta p \underset{t\to\infty}{\to} \frac{\Delta p}{1-\mu}. \end{cases} \tag{8.13}$$

17. On peut passer de l'une à l'autre par intégration par parties.

A $t = 0^+$, on a bien $p_t^0 = p_t = \Delta p$, mais le trend-following amplifie graduellement la perturbation initiale pour finalement la multiplier par un facteur $\frac{1}{1-\mu}$. Lorsque le trend-following est trop fort (i.e. $\mu > 1$), cette quantité diverge, et le marché est instable (on voit aussi que l'équation 8.12 n'est plus définie en général). Les équations ci-dessus décrivent également l'amplification de l'impact d'un métaordre, et montrent que si celui-ci décroît à zéro sans trend-following, il décroitra également à zéro en sa présence. En revanche, si un impact permanent existe, il sera amplifié par un facteur $\frac{1}{1-\mu}$.

Tout cela est finalement assez peu intéressant car cela aurait très bien pu être trouvé en-dehors du cadre des agents hétérogènes : les aspects prix et liquidité y sont quasiment découplés. La seuls chose qu'on retiendra, c'est que la liquidité garde toutes les propriétés qu'elle avait en l'absence de trend-following – en particulier, le carnet d'ordres en V. Les variantes plus intéressantes où l'aspect hétérogène pourra se manifester et affecter le processus de prix de manière non triviale, seront par exemple des situations où les paramètres μ et τ varieront d'agent en agent, avec possiblement des μ négatifs (comportements de mean-reversion). L'apport d'arbitrageurs pourrait également être un élément clé, car en leur absence le prix est facilement prévisible. Au-delà de ces embryons d'idées, il y en a certainement de nombreuses auxquelles je ne pense pas, et auxquelles le cadre des agents hétérogènes pourrait donner du sens. Ce sont des questions pour des temps futurs.

8.5 L'exemple du *Yield Management*

Le cadre exposé dans cette thèse, bien qu'initialement motivé par la compréhension de la microstructure des marchés financiers et par la résolution de problèmes pratiques comme le trading optimal, se trouve avoir une portée bien plus large. En effet, rien ne l'oblige à se restreindre au cas des double-enchères continues (comprendre : des marchés financiers modernes) qui a initialement justifié son étude. En fait, bien plus que dans les prédictions sur la *structure* de l'offre et la demande dans ce cas particulier, la réelle avancée se trouve dans la compréhension de leur *dynamique*. Nous pouvons donc nous attaquer à une classe bien plus large de problèmes, quel que soit le mécanisme de marché ou les propriétés micro- et macroscopiques du système. Une première étape dans cette direction a été effectuée an Chapitre 7, en tentant de comprendre comment la fréquence du *market clearing* affecte le marché, et comment l'on passe de l'équilibre Walrasien classique aux marchés financiers modernes en permettant des transactions en continu. La seconde étape est maintenant de sortir du cadre habituel des marchés financiers : c'est ce que l'on va tenter d'ébaucher dans cette section.

Nous allons pour cela nous intéresser au cas où un côté de l'enchère est contrôlé par un monopole et où l'autre côté (la demande) est constituée de nombreux clients. La firme en monopole doit alors déterminer dynamiquement des prix optimaux de manière à maximiser une fonction d'utilité donnée (qui peut dépendre par exemple de ses revenus attendus, de leur variance, de contraintes sur l'expérience utilisateur, etc.). Ce problème est bien connu dans l'industrie sous le nom de *reve-*

nue management (ou *yield management*) et est adressé par de nombreuses entreprises (companies aériennes, hôtels, etc.).

L'idée de cette courte section est d'utiliser la compréhension des systèmes à agents que nous avons obtenue dans le cadre des marchés financiers pour donner une dynamique aux clients, que nous allons modéliser par un ensemble d'agents hétérogènes. Nous proposons ainsi une sorte de couche additionnelle qui pourra se superposer aux modèles macroscopiques habituellement utilisés (données de tendances sur la demande, inventaire restant, temps avant l'échéance, etc.) pour produire des courbes de demande dynamiques – et optimiser les prix en fonction de celles-ci.

8.5.1 Formulation générale du problème

Soit $\varphi(x,t)$ la densité d'acheteurs potentiels au prix x et au temps t. Cette densité est supposée suivre une dynamique d'agents hétérogènes, représentée par l'équation suivante :

$$\begin{cases} \dfrac{\partial \varphi}{\partial t} = \mathcal{L}_x \varphi \\ \forall x \geq p(t), \quad \varphi(x,t) = 0 \end{cases} \tag{8.14}$$

où $\mathcal{L}_x$ est un opérateur différentiel approprié qui capture la dynamique des clients, et $p(t)$ est le prix de vente proposé par la firme au temps t. La signification exacte de φ est laissée vague intentionnellement, une possibilité est d'y inclure tous les clients en recherche active du bien ou du service, mais on peut imaginer y inclure des intentions latentes comme pour le cas des marchés financiers. En toute généralité, le problème à résoudre pour la firme est donc le suivant :

$$\max_p \int_t^\infty e^{-r(s-t)} \mathrm{d}\mathcal{U}_s \tag{8.15}$$

où p est la politique de pricing, $\mathrm{d}\mathcal{U}_s$ est son utilité incrémentale et r est un taux d'actualisation.

8.5.2 Un exemple

Supposons pour faire simple que la firme souhaite maximiser ses revenu futurs sans aucune autre contrainte, et sans taux d'actualisation. Le problème devient alors [18] :

$$\max_p \int_t^\infty -p(s) \frac{\partial \varphi^p}{\partial x}(0,s) \mathrm{d}s \tag{8.16}$$

Supposons également que la demande suit la dynamique évoquée au Chapitre 6, avec (i) des nouveaux entrants par unité de temps représentés par une densité $\lambda(x,t)$, (ii) des intentions annulées représentés par un taux d'annulation ν_t, et (iii) des changements d'avis globaux (V_t) et idiosyncra-

18. Sous cette forme l'intégrale diverge, il faut donc voir l'intégrale comme la limite de l'intégrale jusqu'à T pondérée par $1/T$ pour que le problème de maximisation ait un sens. Cela revient à résoudre le problème 8.15 pour $r \to 0$.

tiques (D_t). Ces hypothèses sont représentées par l'opérateur suivant :

$$\mathcal{L}_x\varphi = D_t\frac{\partial^2\varphi}{\partial x^2} + V_t\frac{\partial\varphi}{\partial x} + \lambda(x,t) - \nu_t\varphi. \tag{8.17}$$

Les évolutions temporelles des paramètres D_t, V_t, $\lambda(x,t)$ et ν_t pourront typiquement être données en amont par un modèle macroscopique plus « classique ». Pour illustrer ces propos par un exemple, considérons la dynamique suivante[19] :

$$D_t = D; \quad V_t = V\sin(\mu t); \quad \lambda(x,t) = \lambda\mathbb{1}_{x<\bar{p}}; \quad \nu_t = \nu, \tag{8.18}$$

c'est-à-dire une situation où l'agressivité des clients (ou leur impatience) est périodique et où tous les autres paramètres sont constants. Cette périodicité peut par exemple provenir d'effets journaliers, hebdomadaires, mensuels, annuels, etc. La fonction de pricing la plus simple pour la firme est simplement de garder un prix constant p^{const} : cela constituera notre benchmark (le profit obtenu pour la meilleure valeur possible de p^{const} est Π^{const} et est normalisé à $\Pi^{\text{const}} = 100$). Une estimation numérique de la stratégie optimale, obtenue par recuit simulé, est présentée sur la figure 8.5. Naturellement, celle-ci a également un aspect périodique, et permet de réaliser un revenu de $\Pi^{\text{rs}} = 116$, soit 16% de plus que la meilleurs stratégie de prix constant. La meilleure stratégie sinusoïdale, en comparaison, réalise un profit de $\Pi^{\text{sin}} = 108.7$, soit une amélioration de 8.7% par rapport au benchmark. Il est intéressant de noter que les politiques de pricing optimales sont décalées par rapport à la courbe d'agressivité des clients. Il semble en effet qu'il soit optimal de monter les prix *avant* que les clients ne deviennent plus agressifs, probablement pour permettre à un plus grand réservoir de se constituer en prévision du moment où les prix seront hauts. Cet effet est d'ailleurs quasi-parfaitement synchronisé entre la politique en sinus et celle déterminée par recuit simulé. Le pattern du volume associé à la politique optimale est également intéressant : lors de chaque période, il augmente progressivement jusqu'à un pic, puis redescend immédiatement aux alentours de zéro : une fois le pic d'agressivité passé, il est plus intéressant de permettre au réservoir de clients de se reconstituer en vue du pic suivant, plutôt que de réaliser les transactions. Cela se manifeste sur les courbes de profits, moins lisses que dans le cas constant.

Il ne s'agit bien sûr là que d'un exemple, dont rien n'assure la pertinence dans le monde réel. Les questions de savoir si la modélisation choisie ici est réaliste, ou si son apport par rapport à des modèles plus simples à optimiser est suffisamment conséquent pour justifier de leur utilisation en pratique, ne relèvent pas de cette thèse. Quoi qu'il en soit, ce type de modèles peut *a minima* permettre de comprendre certains aspects qualitatifs dont il faut être conscient – comme ceux décrits à la fin du paragraphe précédent – ou encore de benchmark pour tester la pertinence des modèles

19. ...qui n'a pas été choisie pour son réalisme, mais qui permet de développer certaines intuitions. Les valeurs des paramètres choisis sont : $D = 10$; $V = 10$; $\mu = \frac{10*\pi}{T}$; $T = 10$; $\lambda = 1$; $\nu = 1$ (unités arbitraires).

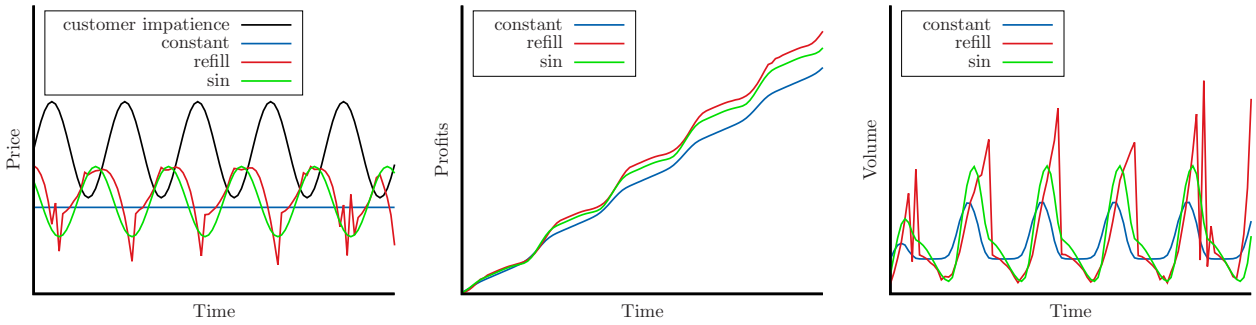

FIGURE 8.5 – (gauche) En noir, l'agressivité des clients. En couleur, les politiques de pricing optimales (à erreurs numériques près) dans le cas d'un pricing constant (bleu), sinus (vert) et une solution numérique obtenue par recuit simulé (rouge). (milieu) Profits cumulés associés à chacune de ces politiques. La politique en sinus améliore le benchmark constant de 8.7%, et le recuit simulé de 16%. (droite) Volumes de transactions correspondants.

plus simples que l'efficacité de la production exige.

8.6 Conclusion et postface

Nous arrivons à la fin de ce chapitre un peu plus conceptuel – et un peu moins technique. Cela clôt la tentative de réponse aux questions soulevées dans l'introduction de cette thèse, qui a tout de même nécessité quelques étapes. Nous avons plongé dans des données uniques et complémentaires à celles des marchés financiers classiques, pour y trouver les hypothèses qui allaient ensuite structurer la théorie. Nous avons développé cette théorie dans une limite où elle était soluble – i.e. dans la limite où les agents sont tous suffisamment petits – obtenant entre autres une formule fermée pour décrire l'impact d'une action agressive sur le prix, et démontrant une absence structurelle d'arbitrage. Puis nous avons commencé à prendre du recul, et nous avons extrait la dynamique des agents du contexte limité des marchés financiers continus, pour tenter de comprendre l'effet du mécanisme de marché sur le marché lui-même, faisant apparaître le lien – et les différences – entre les théories classiques de l'offre et la demande et les marchés financiers modernes qui cotent en continu. Après cette abstraction nécessaire tant du point de vue technique que du point de vue conceptuel, mais qui a exigé de postuler certaines hypothèses forcément limitantes, nous avons finalement discuté des différentes manières d'utiliser et de mettre à profit le cadre des agents hétérogènes proposé ici, dans toute sa généralité – que ce soit en finance avec la réalisation d'un simulateur de marché, ou en-dehors comme nous venons de le voir dans la section précédente. Nous avons donc l'impression d'avoir compris un mécanisme fondamental de la microstructure des marchés, d'avoir extrait les équations qui le décrivent et d'avoir enfin développé un cadre général qui parvienne à le capturer pour créer des modèles pertinents et réalistes. Mais à quoi bon si cela reste strictement confiné à cette micro-échelle ? C'est qu'en réalité cela ne le reste pas, comme va le montrer le chapitre suivant.

Chapitre 9

Bulles, crashs, liquidité et impact : comprendre les événements extrêmes par la microstructure

From
Why do markets crash ? Bitcoin data offers unprecedented insights
with Jean-Philippe Bouchaud
(Donier and Bouchaud, 2015a)

9.1 Préface (français)

Nous avons maintenant à notre disposition une belle théorie micro- et mésoscopique de la formation des prix avec l'offre et la demande, mais pour « boucler la boucle » quelque chose manque encore. De même que l'on cherche des lois physiques à l'échelle de l'atome pour expliquer des phénomènes à plus grande échelle, de même que la micro-économie cherche dans les comportements individuels des clés pour comprendre le système dans son ensemble, pour gagner toute sa pertinence la présente théorie doit encore prouver ses implications macroscopiques. L'article qui suit, probablement le plus « tape à l'oeil » des quatre, se charge donc de faire le lien entre les échelles à travers une quantité : la *liquidité*. Son but, en se basant – encore ! – sur le Bitcoin et en enquêtant sur quatorze de ses principaux crashs de l'année 2013, est d'établir la pertinence des mesures de liquidité micro-et mésoscopiques mises en avant par la théorie, et de montrer grâce à des données inédites sur la profondeur de l'offre et de la demande en temps réel leur aptitude à prédire les crises de liquidité macroscopiques. En particulier, la détermination en temps réel d'une *bande de liquidité* que le prix sera capable de traverser en période extrême nous renseigne à tout moment sur la propension d'un marché au crash : l'élargissement de cette bande est un signe avant-coureur de périodes troubles,

qui dans le cas du Bitcoin aurait pu nous alerter sur l'état de bulle dans lequel il s'est trouvé – au moins – à deux reprises, en Avril et en Décembre 2013.

En un sens, il ne s'agit là que d'un travail préparatoire, servant essentiellement à valider la pertinence macroscopique de la présente théorie. L'étude de la liquidité et de sa dynamique sur un large univers d'actifs financiers serait maintenant de la plus grande valeur.

Abstract : Crashes have fascinated and baffled many canny observers of financial markets. In the strict orthodoxy of the efficient market theory, crashes must be due to sudden changes of the fundamental valuation of assets. However, detailed empirical studies suggest that large price jumps cannot be explained by news and are the result of endogenous feedback loops. Although plausible, a clear-cut empirical evidence for such a scenario is still lacking. Here we show how crashes are conditioned by the market liquidity, for which we propose a new measure inspired by recent theories of market impact and based on readily available, public information. Our results open the possibility of a dynamical evaluation of liquidity risk and early warning signs of market instabilities, and could lead to a quantitative description of the mechanisms leading to market crashes.

9.2 Introduction

Why do market prices move ? This simple question has fuelled fifty years of academic debate, reaching a climax with the 2013 Nobel prize in economics, split between Fama and Shiller who promote radically different views on the question Shiller (2013). Whereas Fama argues that markets are efficient and prices faithfully reflect fundamental values, Shiller has shown that prices fluctuate much more than what efficient market theory would suggest, and has insisted on the role of behavioural biases as a source of excess volatility and price anomalies. Of particular importance is the origin of the largest changes in prices, aka market crashes, that may have dire consequences not only for market participants but also for the society as a whole Taleb (2010). It is fair to say that after centuries of market folly Mackay (2012); Kindleberger and Aliber (2011); Sornette (2009); Reinhart and Rogoff (2009), there is no consensus on this issue. Many studies Fair (2002); Joulin et al. (2008); Cornell (2013) have confirmed the insight of Cutler, Poterba & Summers Cutler et al. (1989) who concluded that *[t]he evidence that large market moves occur on days without identifiable major news casts doubts on the view that price movements are fully explicable by news....* The fact that markets appear to crash in the absence of any remarkable event suggests that destabilising feedback loops of behavioural origin may be at play Smith et al. (1988); Lillo and Farmer (2005); Hommes et al. (2005); Bouchaud (2013). Although plausible, a clear-cut empirical evidence for such an endogenous scenario is still lacking. After all, crashes are not that frequent and a convincing statistical analysis is difficult, in particular because of the lack of relevant data about the dynamics of supply and demand during these episodes.

In this respect, the Bitcoin Nakamoto (2008); Ali et al. (2014); Böhme et al. (2014) market is quite unique on many counts. In particular, the absence of any compelling way to assess the fundamental price of Bitcoins makes the behavioral hypothesis highly plausible. For our purpose, the availability of the full order book [1] at all times provides precious insights, in particular before and during extreme events. Indeed, at variance with most financial markets where participants hide their intentions, the orders are placed long in advance by Bitcoin traders over large price ranges. Using two highly informative data-sets – the trade-by-trade MtGox data between December 2011 and January 2014, and the full order book data over the same period – we analyse in depth the liquidity of the Bitcoin market. We find that what caused the crash was not the selling pressure per se, but rather the dearth of buyers that stoked the panic. Following up on this observation, we show that three different liquidity measures that aim at quantifying the presence of buyers (or sellers) are highly correlated and correctly predict the amplitude of potential crashes. Whereas two of them are direct probes of the prevailing liquidity but difficult to access on financial markets, the third one – which is also firmly anchored theoretically Donier et al. (2015) – only uses readily available, public information on traded volumes and volatility, and is therefore a promising candidate for monitoring the propensity of a market to crash.

9.3 Anatomy of April 10, 2013 crash

Amongst all crashes that happened on the Bitcoin and for which we found some data, the April 10, 2013 crash is probably the most interesting one since on that day the price dropped by more than 50% of its value in a few hours. At that time, MtGox was by far the leading exchange (its market share was over 80% on the BTC/USD spot market) so our data-set captures a large fraction of the investors' behaviour. Intuitively, the main driver of market crashes is the mismatch between the aggregate market order flow imbalance ($\mathcal{O}$, defined below) that becomes strongly negative and the prevailing liquidity on the buy side, i.e. the density of potential buyers below the current price. Whereas the former quantity can be easily reconstructed from the series of trades, the notion of "prevailing liquidity" is only at best ambiguous. It is only when the price starts heading down, that one expects most of the interested buyers to declare themselves and post orders in the order book. Therefore, the liquidity cannot in principle be directly inferred from the information on the publicly available order book. The dynamic nature of liquidity has been clearly evidenced Weber and Rosenow (2005); Bouchaud et al. (2006), and has led to the notion of "latent" liquidity that underpins recent theories of impact in financial markets Tóth et al. (2011); Mastromatteo et al. (2014a,b); Donier et al. (2015).

However, Bitcoin is quite an exceptional market in this respect, since a large fraction of the

1. The order book is the record of all intentions to buy or sell at a given point in time, each volume coming with an offering price.

liquidity is not latent, but actually posted in the order book – possibly resulting from less strategic participants on a still exotic market – and thus directly observable (see Fig. 9.1). A more quantitative analysis indeed shows that typically $30-40\%$ of the volume traded during the day is already present in the order book in the morning. This is to be compared with a ratio below 1% on more traditional financial markets, say stocks [2]. This allows us to test in detail the respective roles of aggregate imbalance and liquidity in the triggering of market crashes. We first study the "aggressive" order flow defined as the aggregated imbalance of market orders for every 4 hours window between January 2013 and August 2013. In fact, two definitions are possible. One is defined as the average of the signed number of Bitcoin contracts sent as market orders [3] Because of this need for immediacy, one often refers to them as aggressive orders ; $\mathcal{O}_B = \sum_i \epsilon_i q_i$, where each i is a different market order of sign ϵ_i ($\epsilon_i = +1$ for buyer-initiated trades and -1 for seller-initiated trades) and number of contracts q_i, and the sum runs over consecutive trades in a 4 hour window. The second is the volume imbalance expressed in USD : $\mathcal{O}_\$ = \sum_i \epsilon_i q_i p_i$, where p_i is the i-th transaction price. These two quantities are shown in Fig. 9.2 and reveal that (a) large sell episodes are more intense than large buy episodes ; (b) when expressed in Bitcoin, the sell-off that occurred on April, 10 (of order of 30,000 BTC on a 4h window) is not more spectacular than several other sell-offs that happened before or after that day ; (c) however, when expressed in USD, the April 10 sell-off indeed appears as an outlier.

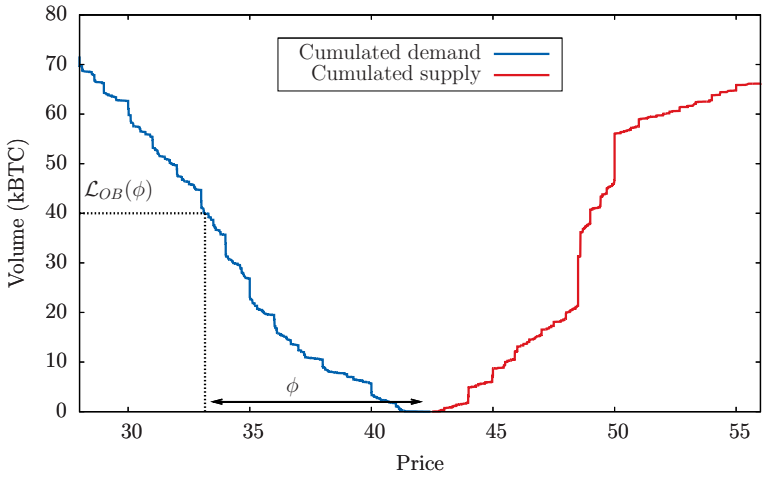

FIGURE 9.1 – **Instantaneous cumulated order book.** Snapshot of the cumulated supply and demand displayed on the order book taken on March 8, 2013, with a graphical representation of the order book liquidity $\mathcal{L}_{OB}(\phi)$ defined in Def. 9.4.1.

The difference between $\mathcal{O}_B$ and $\mathcal{O}_\$$ originates from the fact that a large fraction of this selling activity occurred at the peak of the "bubble" that preceded the crash, see Fig. 9.3, top. The BTC

2. The total volume in the order book of major stocks is 5-10 times the volume at the best quotes, which is itself $\sim 10^{-3}$ of the daily turnover, see e.g. Wyart et al. (2008).

3. A market order is an order to trade immediately at the best available price.

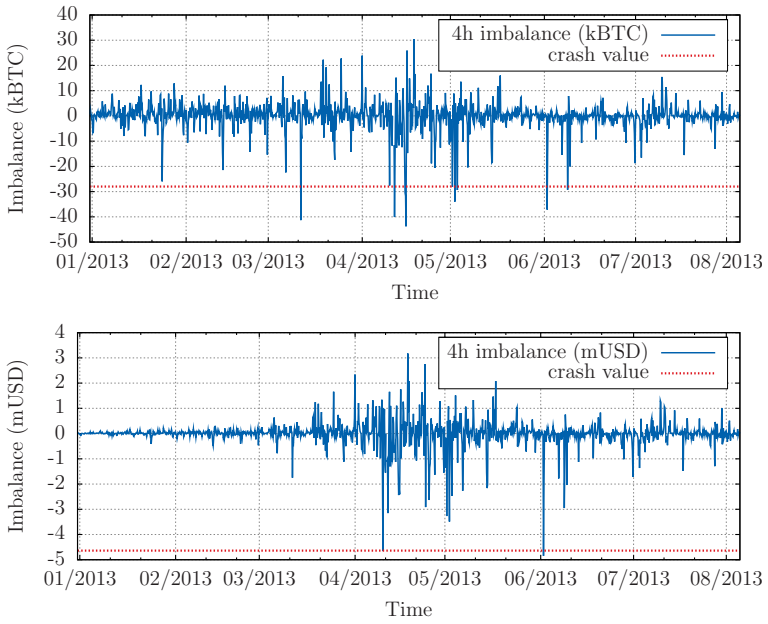

FIGURE 9.2 – **Order flow imbalances in USD and BTC.** Top : Aggressive imbalance in order flow $\sum_i \epsilon_i v_i$ (where $\epsilon_i = \pm 1$ is the sign of the transaction, and v_i its volume in Bitcoins), aggregated over periods of 4 hours between January 2013 and August 2013, expressed in Bitcoins. April 10, 2013 (for which the realised imbalance is represented as a dashed horizontal line) does not appear as an outlier. Bottom : aggressive imbalance in order flow $\sum_i \epsilon_i v_i p_i$, aggregated over periods of 4 hours between January 2013 and August 2013 and expressed in USD. April 10, 2013 now clearly appears as an outlier.

price rose from \$13 in early January to \$260 just before the crash. In Fig. 9.3, we represent a "support" level p_S^{40k} such that the total quantity of buy orders between p_S^{40k} and the current price p_t is 40,000 BTC, see Fig. 9.1. One notices that the price dramatically departed from the support price during the pre-crash period, which is a clear sign that Bitcoin price was engaged in a bubble. Although the liquidity expressed in USD was actually *increasing* during that period (see Fig. 9.3, middle), the BTC price increased even faster, resulting in a thinner and thinner liquidity on the buy side of the order book *expressed in BTC*, see Fig. 9.3, bottom. This scenario is precisely realised in some Agent Based Models of markets Giardina and Bouchaud (2003).

The conclusion of the above analysis, that may appear trivial, is that the crash occurred because the price was too high, and buyers too scarce to resist the pressure of a sell-off. More interesting is the fact that the knowledge of the volume present in the order book allows one to estimate an expected price drop of $\approx 50\%$ in the event of a large – albeit not extreme – sell-off. Of course, the possibility to observe the full demand curve (or a good approximation thereof) is special to the Bitcoin market, and not available in more mainstream markets where the publicly displayed liquidity is only of order 1% of the total daily traded volume. Still, as we show now, one can built accurate proxies of the latent liquidity using observable quantities only, opening the path to early warning signs of an impeding crash.

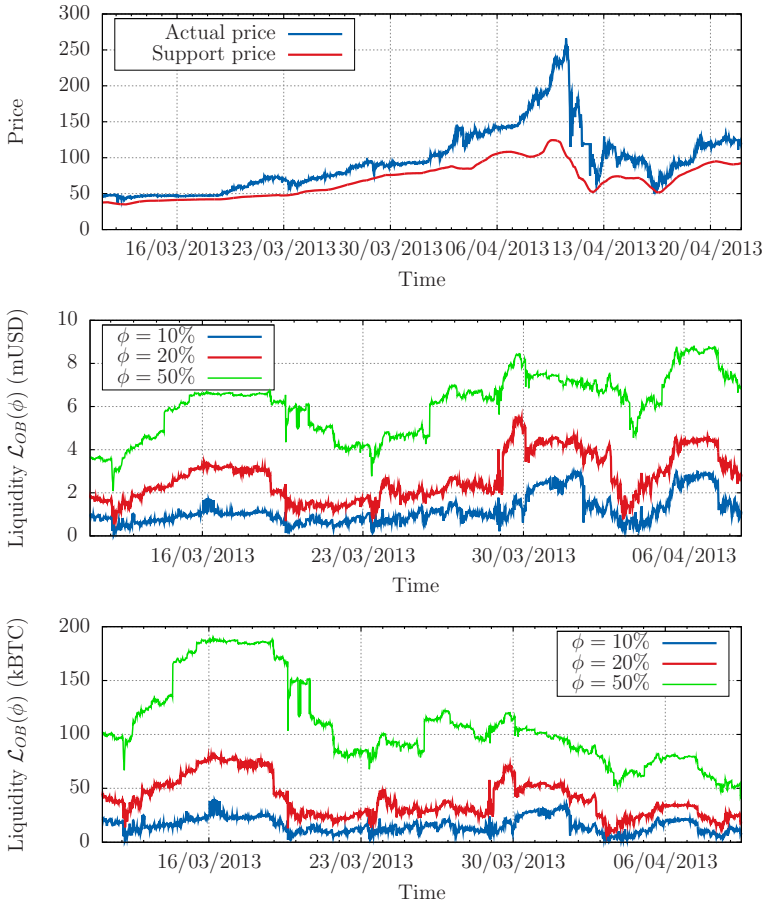

FIGURE 9.3 – **Liquidity and support price.** Top : Actual price p_t (blue) vs. support price $p_S^{40\text{k}}$ (red) defined as the price that would be reached if a typical sell-off of 40,000 BTC was to occur instantaneously. Note that $p_S^{40\text{k}}$ is $\approx 50\%$ below the price p_t just before the crash, explaining the order of magnitude of the move that happened that day. Middle (resp. Bottom) : Buy volume $\mathcal{L}_{\text{OB}}(\phi)$ in USD (resp. BTC) in the order book, during the months preceding the crash of April 10, 2013, measured as the volume between the current price p_t and $p_t(1-\phi)$ where $\phi = 10\%$, 20% and 50%. One can see that for any quantile the liquidity in USD tended to increase by an overall factor $\simeq 2$ during the period, while the liquidity in BTC was decreased by a factor $\simeq 2-3$ as an immediate consequence of the bubble.

9.4 Three definitions of "liquidity"

More formally, the market liquidity measure discussed above is defined as :

Definition 9.4.1. *The* order-book liquidity $\mathcal{L}_{OB}(\phi)$ *(on the buy side) is such that (cf. Fig. 9.1 above) :*

$$\int_{p_t(1-\phi)}^{p_t} \mathrm{d}p\rho(p,t) := \mathcal{L}_{OB}(\phi) \ , \tag{9.1}$$

(and similarly for the sell-side). In the above equation, p_t is the price at time t and $\rho(p,t)$ is the density of demand that is materialised on the order book at price p and at time t.

Conversely, the price drop $-\phi^* p_t$ expected if a large instantaneous sell-off of size Q^* occurs is

such that :

$$\phi^* = \mathcal{L}_{OB}^{-1}(Q^*), \tag{9.2}$$

where $\mathcal{L}_{OB}^{-1}$ *is a measure of illiquidity.*

An *a posteriori* comparison between realised returns and the liquidity-adjusted imbalance for the 14 most extreme negative returns that have occurred between Jan 1, 2013 and Apr 10, 2013 is shown in Fig. 9.4. These events, which corresponds to dramatic jumps in the cumulated order flow process, are found to have a characteristic scale of about 4h with a standard deviation of 2.5h, justifying the choice made in Fig. 9.2 to plot imbalances at a 4h time scale. The analysis shows that the quantity $\mathcal{L}_{\mathrm{OB}}^{-1}(\mathcal{O}_B)$ nearly perfectly matches crashes amplitudes, vindicating the hypothesis that most of the liquidity is indeed present in the visible order book for the Bitcoin.

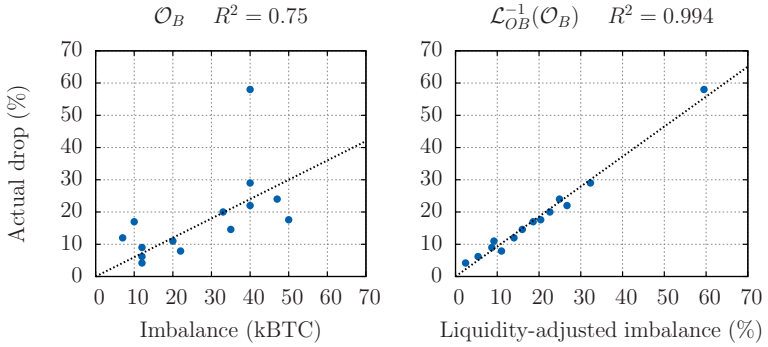

FIGURE 9.4 – **Forecast of crashes amplitudes using order book volumes.** For the 14 most extreme negative returns that have occurred between Jan 1, 2013 and Apr 10, 2013, we compare the realised return with : (Left) the net imbalance $\mathcal{O}_B$ during the period (usually a few hours) and (Right) the *liquidity-adjusted* imbalance $\mathcal{L}_{\mathrm{OB}}^{-1}(\mathcal{O}_B)$. This illustrates the relevance of the $\mathcal{L}_{\mathrm{OB}}$ liquidity measure to predict the amplitudes of crashes – even in the most extreme cases.

However, as recalled above, the visible order book on standard financial markets usually contains a minute fraction of the real intentions of the agents. Therefore the use of $\mathcal{L}_{\mathrm{OB}}(\phi)$ deduced from the observable order book would lead to a tremendous underestimation of the liquidity in these markets Sandås (2001); Weber and Rosenow (2005). Liquidity is in fact a dynamic notion, that reveals itself progressively as a reaction (possibly with some lag) to the incoming order flow Weber and Rosenow (2005); Bouchaud et al. (2006). Another definition of liquidity, that accounts for the progressive appearance of the latent liquidity as orders are executed, is based on a measure of *market impact*. With enough statistics, the average (relative) price move $I(Q) = \langle \Delta p/p \rangle$ induced by the execution of a meta-order [4] can be measured as a function of their total volume Q. Since these meta-orders are executed on rather long time scales (compared to the transaction frequency), it is reasonable to think that their impact reveals the "true" latent liquidity of markets Tóth et al.

4. A meta-order is a sequence of individual trades generated by the same trading decision but spread out in time, so as to get a better price and/or not to be detected Tóth et al. (2011).

(2011); Mastromatteo et al. (2014a,b); Donier et al. (2015). This leads us to a second definition of liquidity, based on market impact :

Definition 9.4.2. *The* impact liquidity $\mathcal{L}_I(\phi)$ *is defined as the volume of a meta-order that moves, on average, the price* p_t *by* $\pm\phi p_t$*, or, more precisely,* $\mathcal{L}_I(\phi)$ *is fixed by the condition :*

$$I(\mathcal{L}_I(\phi)) = \phi, \tag{9.3}$$

since the impact $I(Q)$ *is usually measured in relative terms. As above, the price drop expected if a large sell-off of volume imbalance* Q^* *occurs is simply given by* $\mathcal{L}_I^{-1}(Q^*) = I(Q^*)$*.*

The problem with this second definition is that it requires proprietary data with sufficient statistics, available only to brokerage firms or to active asset managers/hedge funds. It turns out to be also available for Bitcoin Donier and Bonart (2014) – see below. However, a very large number of empirical studies in the last 15 years have established that the impact of meta-orders follows an extremely robust "square-root law" Torre and Ferrari (1997); Almgren et al. (2005); Moro et al. (2009); Donier and Bonart (2014); Bladon et al. (2012); Tóth et al. (2011); Kyle and Obizhaeva (2012); Bershova and Rakhlin (2013); Gomes and Waelbroeck (2015); Mastromatteo et al. (2014a); Brokmann et al. (2015). Namely, *independently* of the asset class, time period, style of trading and micro-structure peculiarities, one has :

$$I_{\rm TH}(Q) \;\approx\; Y\sigma_{\rm d}\sqrt{\frac{Q}{V_{\rm d}}}\,, \tag{9.4}$$

where Y is an a-dimensional constant of order unity, $V_{\rm d}$ is the daily traded volume and $\sigma_{\rm d}$ is the daily volatility. This square-root law has now been justified theoretically by several authors, building upon the notion of latent liquidity Tóth et al. (2011); Mastromatteo et al. (2014a,b); Donier et al. (2015) (see Ref. Farmer et al. (2013) for an alternative story). Assuming that the above functional shape of market impact is correct leads to a third definition of liquidity :

Definition 9.4.3. *The* theoretical liquidity $\mathcal{L}_{TH}(\phi)$ *is the theoretical volume of a meta-order required to move the price* p_t *by* $\pm\phi p_t$ *according to formula Eq. (9.4) above, i.e. :*

$$I_{TH}(\mathcal{L}_{TH}(\phi)) = \phi. \tag{9.5}$$

Together with Eq. (9.4), this amounts to consider $\sigma_{\rm d}/\sqrt{V_d}$ as a measure of market illiquidity. Clearly, since both $\sigma_{\rm d}$ and V_d can be estimated from public market data, this last definition of liquidity is quite congenial. It was proposed in Ref. Caccioli et al. (2012) as a proxy to obtain impact-adjusted marked-to-market valuation of large portfolios, and tested in Ref. Kyle and Obizhaeva (2012) on five stock market crashes, with very promising results. However, there is quite a leap of faith in assuming that our above three definitions are – at least approximately – equivalent. This is

why the Bitcoin data is quite unique since it allows one to measure all three liquidities $\mathcal{L}_{\rm OB}, \mathcal{L}_{\rm I}, \mathcal{L}_{\rm TH}$ and test quantitatively that they do indeed reveal the very same information.

9.5 Comparing the liquidity measures

We measured the order book liquidity $\mathcal{L}_{\rm OB}$ at the daily scale by averaging the volume present at all prices in the buy side of the order book for each day. The empirical impact is obtained following Ref. Donier and Bonart (2014) by measuring the full $I(Q)$, obtained as an average over all meta-orders of a given volume Q on a given day. Finally, the theoretical impact Eq. (9.4) is obtained by measuring both the traded volume of the day $V_{\rm d}$ and the corresponding volatility $\sigma_{\rm d}$ [5]. The daily scale has been chosen so as to average out market noise and intraday patterns in the measure of $\mathcal{L}_{\rm I}^{-1}$ and $\mathcal{L}_{\rm TH}^{-1}$, while remaining reactive to liquidity fluctuations : Fig. 9.3 indeed shows how much liquidity can fluctuate in a few days.

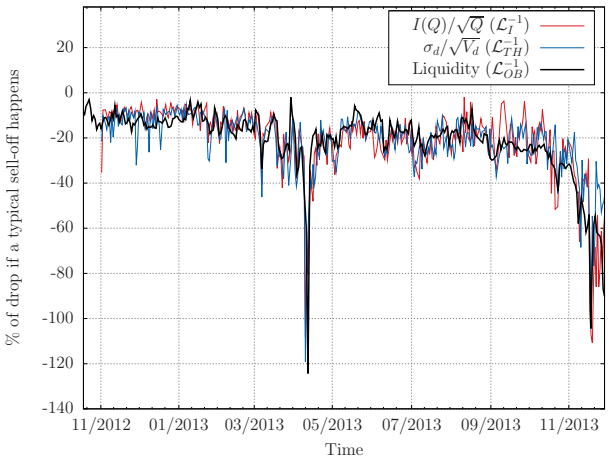

FIGURE 9.5 – **Comparison between the three (il-)iquidity measures.** Parallel evolution of the three price drops ϕ^* deduced from our three estimates of illiquidity $\mathcal{L}_{\rm OB}^{-1}, \mathcal{L}_{\rm I}^{-1}, \mathcal{L}_{\rm TH}^{-1}$ defined above. The estimates based on $\mathcal{L}_{\rm I}^{-1}, \mathcal{L}_{\rm TH}^{-1}$ have been rescaled by a factor 6.10^4 to match the average order book data prediction.

These three estimates allow us to compare, as a function of time (between November 2012 and November 2013) the expected price drop for a large sell meta-order of size – say – $Q^* = 40,000$ BTC, see Fig. 9.5. We have rescaled by a constant factor the predictions based on $\mathcal{L}_{\rm I}$ and $\mathcal{L}_{\rm TH}$, so as to match the average levels. The agreement is quite striking, and shown in a different way in Fig. 9.6 as a scatter plot of $\mathcal{L}_{\rm OB}^{-1}$ vs $\mathcal{L}_{\rm I}^{-1}$ or $\mathcal{L}_{\rm TH}^{-1}$, either on the same day, or with a one day lag. As coinciding times, the R^2 of the regressions are ≈ 0.86 and only fall to ≈ 0.83 with a day lag, meaning that one can use past data to predict the liquidity of tomorrow. As a comparison, when

5. defined as $\sigma_{\rm d}^2 = \frac{1}{T}\sum_{t=1}^{T}\left(0.5\ln\left(H_t/L_t\right)^2 - (2\ln(2)-1)\ln\left(C_t/O_t\right)^2\right)$ where $O_t/H_t/L_t/C_t$ are the open/high/low/close prices of the sub-periods Garman and Klass (1980).

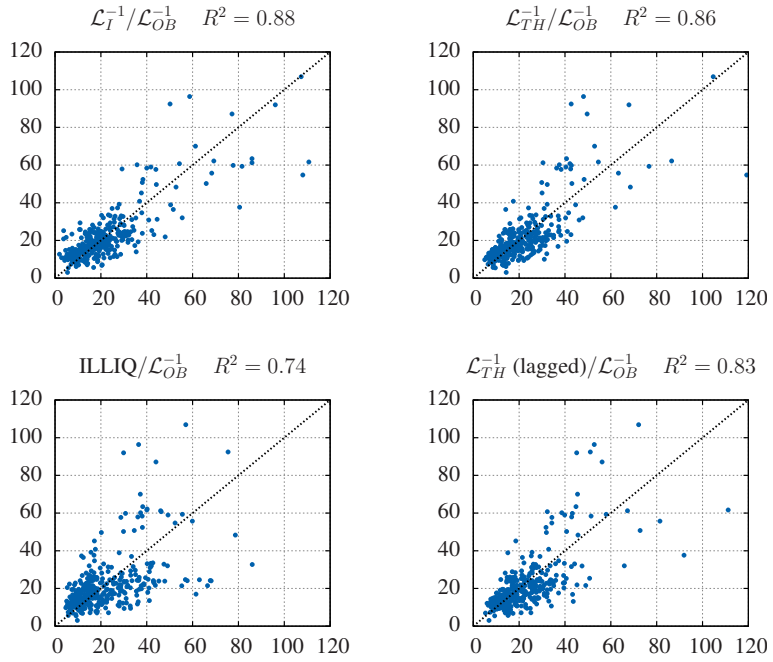

FIGURE 9.6 – **Regression of the actual (il-)liquidity against the different (il-)liquidity measures.** Regression of the actual illiquidity $\mathcal{L}_{OB}^{-1}$ on three same-day illiquidity measures (after rescaling so that the samples means coincide) : The direct measure of orders market impact $\mathcal{L}_{\mathrm{I}}^{-1}$, the publicly available measure $\mathcal{L}_{\mathrm{TH}}^{-1}$ that corresponds to the theoretical and empirical impact, and the well-known Amihud ILLIQ measure Amihud (2002). Both $\mathcal{L}_{\mathrm{I}}^{-1}$ and $\mathcal{L}_{\mathrm{TH}}^{-1}$ outperform ILLIQ ($R^2 \approx 0.86$ vs. 0.74). Note that a high predictability remains when lagging $\mathcal{L}_{\mathrm{TH}}^{-1}$ by one day ($R^2 \approx 0.83$ vs. 0.71). The regression slopes for the four graphs are, respectively : $0.9, 0.95, 0.87$ and 0.93.

using instead Amihud's Amihud (2002) measure of illiquidity $\sigma_{\mathrm{d}}/V_{\mathrm{d}}$, one obtains R^2 of resp. 0.74 and 0.71.

That the estimates based on $\mathcal{L}_{\mathrm{I}}$ and $\mathcal{L}_{\mathrm{TH}}$ match is no surprise since the square-root law was already tested with a high degree of precision on the Bitcoin Donier and Bonart (2014). But that the theoretical measure of liquidity $\mathcal{L}_{\mathrm{TH}}$ based on easily accessible market data is able to track so closely the information present in the whole order book is truly remarkable, and suggests that one can indeed faithfully use $\mathcal{L}_{\mathrm{TH}}$ on markets where reliable information on the latent order book is absent (as is the case for most markets).

9.6 Discussion

Thanks to the unique features of the Bitcoin market, we have been able to investigate some of the factors that determine the propensity of a market to crash. Two main features emerge from our study. First, the price level should lie within a range where the underlying demand (resp. supply) is able to support large – but expected – fluctuations in supply (resp. demand). When the price is clearly out of bounds (for example the pre-April 2013 period for Bitcoin) the market

is unambiguously in a precarious state that can be called a *bubble*. Our main result allows one to make the above idea meaningful in practice. We show that three natural liquidity measures (based, respectively, on the knowledge of the full order book, on the average impact of meta-orders, and on the ratio of the volatility to the square-root of the traded volume, $\sigma_{\rm d}/\sqrt{V_{\rm d}}$) are *highly correlated* and do predict the amplitude of a putative crash induced by a given (large) sell order imbalance.

Since the latter measure is entirely based on readily available public information, our result is quite remarkable. It opens the path to a better understanding of crash mechanisms and possibly to early warning signs of market instabilities. However, while we claim that the amplitude of a potential crash can be anticipated, we are of course not able to predict when this crash will happen – if it happens at all. Still, our analysis motivates better dynamical risk evaluations (like value-at-risk), impact adjusted marked-to-market accounting Caccioli et al. (2012) or liquidity-sensitive option valuation models. As a next step, a comprehensive study of the correlation between the realised crash probability and $\sigma_{\rm d}/\sqrt{V_{\rm d}}$ on a wider universe of stocks – expanding the work of Ref. Kyle and Obizhaeva (2012) – would be a highly valuable validation of the ideas discussed here.

Acknowledgments

We thank A. Tilloy for his insights on the Bitcoin and for reading the manuscript ; P. Baqué for reading the manuscript ; and J. Bonart for useful discussions.

9.7 Postface (français)

Le message que contient cet article est relativement simple, et il n'est pas grand chose que je puisse ajouter à sa préface. Peut-être puis-je conclure en soulignant les similitudes et différences structurelles entre le marché du Bitcoin et les marchés financiers *mainstream*. Même si l'écologie et le niveau de sophistication des agents ainsi que les fondamentaux économiques sous-jacents sont très différentes dans les deux cas, l'universalité de la loi d'impact montre que ceux-ci partagent une *hétérogénéité*, une *micro-structure*[6] et une dynamique similaires. La principale différence entre ce marché amateur et les marchés financiers modernes, qui semble structurelle et incontournable, concerne le degré d'affichage des intentions sur le carnet d'ordres (autrement dit, la proportion du carnet d'ordres latent affiché publiquement). Si les *Bitcoiners* semblent peu se soucier de divulguer à l'avance leurs intentions, du moins pour l'instant, cela n'est pas le cas des acteurs professionnels des marchés modernes : impossible donc de mesurer la liquidité macroscopique directement sur le carnet d'ordres pour ces derniers. Il fallait donc le Bitcoin pour confirmer la théorie : cela n'empêche pas bien sûr, maintenant que le lien est établi, de mener à bien une étude similaire sur les actifs

6. Comprendre : une même structure locale de l'offre et la demande.

classiques en se basant sur d'autres mesures publiques de liquidité [7].

7. Des mesures locales, ou plus globales, comme la mesure de $\sigma/\sqrt{V}$ à une échelle journalière, comme proposé par l'article. Cf Goyenko et al. (2009) pour d'autres mesures de liquidité.

Chapitre 10

Discussion de fin

"If you can identify a delusional popular belief,
you can find what lies hidden behind it :
the contrarian truth."
(Thiel and Masters, 2014)

10.1 Discussion des résultats

Ici se termine cette thèse, et il est de bon ton de rappeler les plus importants de ses résultats. A la question de la formation des prix sur les marchés financiers posée en introduction, nous avons répondu en plusieurs étapes. Avant d'entreprendre toute analyse empirique, la première des questions, cruciale s'il en est, était de déterminer une question empirique pertinente à laquelle cette analyse pourrait répondre. Avant même le commencement de cette thèse, les preuves supportant la théorie de l'impact concave sur les marchés financiers ne manquaient pas dans la littérature financière : confirmer les résultats de ces analyses n'apportait pas grand chose en soi, et il fallait quelque chose d'autre – de différent. Ce *quelque-chose* fut à la fois une *écologie* de marché différente et une extrême qualité des données, qui nous permettait pour la première fois d'effectuer des mesures en masse sur un marché d'amateurs. En retrouvant au Chapitre 5 les résultats d'impact classiques sur le marché si particulier du Bitcoin, nous avons compris que l'écologie de marché n'était pas leur principale origine, et que la réponse se trouvait probablement plus dans la statistique des agents que dans leurs comportements précis. Nous avons donc consacré le Chapitre 6 à une étude approfondie de l'impact dans un modèle à agents hétérogènes « statistiques », pour y obtenir des résultats théoriques en étonnant accord avec l'ensemble des études empiriques réalisées à ce jour. Au-delà de leur importance concernant les questions de trading optimal, ceux-ci ont surtout permis de valider la pertinence d'un concept, celui de l'agent hétérogène en finance, que nous avons ensuite développé sous plusieurs angles dans les Chapitres 7 et 8, tentant tour à tour de revisiter la théorie dynamique de l'offre et de la demande, d'étudier l'impact du mécanisme de marché sur sa liquidité, de poser les

bases d'un simulateur de marché réaliste ou encore de nous attaquer à la modélisation de systèmes non financiers comme le *revenue management*. Nous avons enfin terminé en montrant au Chapitre 9 l'importance de cette compréhension microstructurelle de l'offre et de la demande pour comprendre et anticiper les phénomènes de bulles et de crashs, preuve que la liquidité est un concept important et cohérent à travers les échelles. Mais au-delà des enseignements « concrets » sur la dynamique des systèmes financiers que ces résultats nous ont permis de tirer, ceux-ci donnent en outre un relief tout particulier aux discussions générales de modélisation ébauchées au Chapitre 1 : c'est ce que nous allons maintenant discuter une dernière fois.

10.2 Sur le statut des hypothèses d'efficience et de martingale

Nous avons évoqué en introduction le mot d'ordre : *Find the essential in the non-obvious*. Cela faisait essentiellement référence à des faits empiriques, mais j'aimerais étendre légèrement sa signification aux hypothèses fondamentales des modèles [1]. Quoi de moins évident pour tout bon élève que l'affirmation suivante :

Les prix ne sont pas martingales.

Imprévisibles, ils le sont pourtant bien, du moins aux échelles qui permettraient de gagner de l'argent sans supporter trop de risque. Là n'est pas exactement la question. Ce que conteste cette affirmation, c'est plutôt la confiance totale en l'hypothèse de martingale (on pourra remplacer de manière indifférente « martingale » par « efficience » dans ce qui suit), souvent évoquée comme un fait établi au commencement des modèles et que personne ne saurait contester – comme s'il s'agissait d'une hypothèse triviale, au même titre que « 1+1=2 », ou « la terre est ronde ». Pourtant, il ne faut pas oublier qu'une martingale ou encore un point fixe [2] sont des concepts *mathématiques*, et qu'en temps que tels ils ne peuvent pas représenter la réalité, et ne le pourront jamais. A partir de là, pourquoi ne pas la relâcher, et chercher d'autres bases pour expliquer le monde ? Rien n'empêche bien sûr celui-ci d'être compatible *in fine* avec une hypothèse de martingale ou d'efficience, ou de s'en révéler arbitrairement proche – le fait que le slippage ait été trouvé relativement proche de l'impact permanent au Chapitre 5 en est un exemple. Mais les *conditions* dans lesquelles cet état est atteint seront alors non triviales et potentiellement riche en enseignements – alors qu'un raisonnement basé sur la martingale s'assortira forcément des conditions les plus simples [3] et ratera sans doute

1. Ce qui n'est pas le cas de tout le monde : le papier du Chapitre 7 s'est fait refuser la publication dans une revue d'économie pour la raison que « *the standard approach is to study a rational expectations equilibrium in which agents' price expectations are correct (or at least not contradicted by the equilibrium)* ». Cet attachement aux standards me semble toutefois être contraire à la définition même de la science.

2. Rappelons que l'hypothèse d'anticipations rationnelles requiert une coïncidence entre anticipations et réalisations.

3. L'article de Zhang (1999), développant un exemple de convergence impossible vers un marché parfaitement efficient, est une illustration intéressante de ces propos. Voir aussi Bonart et al. (2014) pour un exemple d'économie dynamique qui ne converge pas vers son équilibre.

la structure sous-jacente, car il se suffit à lui-même et ne nécessite pas de structure. Pour cette raison, il ne permettra pas de tirer des conclusions qui dépassent le cadre dans lequel il se trouve : quelles leçons permet-il de tirer en effet sur les systèmes non martingales ? Que nous apprend-il sur les événements rares ? Quelles renseignements nous donne-t-il sur la dynamique de l'offre et de la demande fondamentale ? [4] N'oublions pas que les marchés financiers leur sont subordonnés, et que les market makers et arbitrageurs ne sont qu'une sur-couche supposée rendre le marché plus fluide et les prix efficients. Comprendre le système, ce n'est pas seulement comprendre son état *asservi*, mais aussi la dynamique *libre* de ses composants : pour modéliser la dynamique de la liquidité, les mécanismes sous-jacents des bulles et des crashs, ou tout simplement pour répondre à des questions hors marché financier [5], la structure de l'offre et de la demande fondamentales et leur dynamique *doivent* être comprises. Si le normatif tente de comprendre le système dans son état *asservi*, le descriptif souhaite le comprendre aussi dans son état *libre* – et cela me semble nécessaire.

10.3 Sur la modélisation en finance

Pourquoi modélise-t-on ? D'un point de vue de physicien, la réponse est assez claire. Modéliser, c'est tenter de comprendre les lois qui régissent le monde qui nous entoure. Les intuitions se transforment en modèles à l'aide d'outils mathématiques qui permettent de les formaliser, ces modèles effectuent des prédictions que l'on peut ensuite confronter à la réalité. Si celles-ci sont conformes aux prédictions du modèle, on juge le modèle bon. Sinon, on tente de comprendre ce qui ne marche pas – et l'on adapte son modèle, ou l'on en change. Notons qu'un modèle peut être bon pour certains aspects mais pas pour d'autres : c'est d'ailleurs souvent comme cela que la science progresse, les modèles les plus simples étant finalement mis en défaut et remplacés par des modèles plus complets. [6] Ce type de modélisation est une manière d'assouvir une curiosité naturelle sur le monde qui nous entoure, et une tentative de le comprendre tel qu'il est dans le moindre de ses détails.

Lorsque l'on passe au monde des humains en revanche, les choses deviennent plus compliquées. Contrairement à la nature, les comportements n'ont *a priori* aucune raison d'obéir à des lois universelles et immuables. La première conséquence, qui affecte directement l'approche physicienne de la modélisation, est qu'un modèle n'est jamais ni tout à fait bon ni tout à fait mauvais : la réalité qu'il tente d'approcher n'est tout simplement pas unique. [7] Ainsi, de nouveaux modèles sont créés et détruits au gré des expériences, et leurs paramètres peuvent et doivent toujours être réajustés aux

4. Ainsi, l'absurdité de critiques telles que : « *the concept of absence of arbitrage is so strong that it supersedes any empirical evidence* » (le concept d'absence d'arbitrage est tellement fort qu'il supplante toute preuve empirique), qui permettent malgré tout d'empêcher d'autorité la publication de certains papiers.

5. Je pense au yield management, à la demande des ménages, etc.

6. Ainsi, la physique quantique ou la théorie de la relativité sont nées des déviations du monde réel à la physique classique.

7. Le physicien répond souvent à ce problème en scindant la modélisation en deux parties : une partie interactionnelle « mécanique » et une partie comportementale, cette dernière pouvant changer au cours du temps.

temps présents – ce qui produit une littérature changeante, moins établie et hiérarchisée que dans les sciences plus « dures ». Ce n'est pas là le seul problème : le plus souvent, les expériences sont difficiles à reproduire et présentent un nombre de données limitées, si bien que l'on se contente d'extraire des tendances (moyennes, corrélations, ratio d'endogénéité, etc.) plutôt que de comprendre en profondeur des résultats trop bruités et qui plus est datés. Tout cela promeut une approche alternative de la modélisation des systèmes sociaux, où le modèle sert plus à véhiculer des intuitions qu'à produire des résultats quantitatifs. Les vertus pédagogiques, notamment auprès des publics à orientations scientifiques, n'en sont que meilleures – ou en tout cas, sont meilleures que de longues proses.[8] La littérature économique a trouvé une approche encore plus radicale pour faire face à l'absence de cohérence temporelle du monde réel. Elle a créé ce qu'elle appelle la modélisation *normative* – par opposition à la modélisation *descriptive* – qui ne cherche plus à représenter le monde tel qu'il est, mais tel qu'il devrait être : ainsi le monde a tort, pas le modèle. On est cependant en droit de rester dubitatif sur la pertinence d'une telle approche lorsqu'il s'agit de prendre des actions quantitatives concrètes sur le monde réel – et notamment, de le réguler.

Le raisonnement normatif me semble tenir de moins en moins en finance de nos jours. Il s'agit sans doute du domaine qui a généré la plus grande quantité de données comportementales durant les dernières décennies (peut-être aujourd'hui supplantée par les réseaux sociaux et la publicité) – sans forcément le savoir. Un marché étant par définition standardisé, ces données sont très structurées et le nombre d'observations est colossal, ce qui rend la tâche de les comprendre *a priori* accessible. Avec l'arrivée ces dernières années d'énormes quantités de données propriétaires auparavant inaccessibles[9], une fenêtre sans précédent s'ouvre sur l'esprit humain qui va sans doute profondément transformer notre compréhension de ses actions et de ses interactions. A leur lumière, il devient

8. Il m'a paru tellement plus simple de comprendre la *Théorie générale de l'emploi, de l'intérêt et de la monnaie* de Keynes après avoir lu quelques équations (qu'il aurait très bien pu écrire lui-même d'ailleurs, étant mathématicien à l'origine)... ces proses interminables me rappellent toujours l'intérêt de la formalisation mathématique, qui a permis de remplacer de énoncés auparavant assez indigestes par des expressions simples et claires. Un exemple qui me fait toujours sourire est la solution originale de l'équation $x^3 + px = q$, énoncée par Cardan (x étant la « chose », p le « nombre de la chose » et q le « nombre de l'équation ») :

> *Le tiers du nombre de la chose au cube étant obtenu,*
> *on y ajoute le carré de la moitié du nombre de l'équation et du tout,*
> *on extrait la racine carrée que l'on met de côté.*
> *Le demi-nombre que l'on a élevé au carré, tu ajoutes ou tu enlèves à l'autre :*
> *tu as le binôme avec son apotome.*
> *En extrayant la racine cubique de l'apotome et celle de son binôme,*
> *le résidu de leurs différences est la valeur de la racine.*

que le formalisme mathématique moderne a remplacé par une expression aussi simple que :

$$\sqrt[3]{\frac{q}{2}+\sqrt{\left(\frac{p}{3}\right)^3+\left(\frac{q}{2}\right)^2}}-\sqrt[3]{-\frac{q}{2}+\sqrt{\left(\frac{p}{3}\right)^3+\left(\frac{q}{2}\right)^2}} \tag{10.1}$$

qui, il faut le dire, est quand même plus simple à comprendre et à manipuler (Cardano and Witmer, 1993).

9. Essentiellement, des données dans lesquelles les agents possèdent des identifiants qui permettent de suivre anonymement leurs actions dans le temps.

donc possible de viser une compréhension plus structurelle et quantitative des systèmes sociaux ou financiers, à condition de ne faire l'impasse ni sur la *physique* des systèmes sociaux, ni sur leur *économie*. J'espère pour ma part que cette thèse aura pu apporter une petite pierre à ce grand édifice.

Appendices

Annexe A

Définition d'une classe de *Processus de Prix Impacté* (IPP) et introduction aux *Path-Dependent Kernels*

Depuis Bachelier au début du siècle dernier (Bachelier, 1900), la recherche de processus de prix toujours plus réalistes a occupé une place importante dans la littérature financière. Partant du constant que les prix sont au premier order martingale, c'est la modélisation de la *volatilité* qui a été l'objet de tous les intérêts, du mouvement brownien géométrique aux processus de Lévy, en passant pas les modèles autorégressifs (ARCH (Bollerslev et al., 1994) et maintenant Hawkes (Bacry and Muzy, 2014)), les processus à volatilité stochastique (Heston, 1993), les processus multifractaux (Muzy et al., 2000), les mouvements browniens fractionnaires (Gatheral et al., 2014), etc. Ces dernières années, avec la popularisation du concept de liquidité, la recherche de processus permettant de décrire de manière effective l'impact d'une action sur les prix s'est peu à peu développée, avec pour modèles phares les modèles de propagateur (Bouchaud and Potters, 2003; Gatheral, 2010; Almgren and Chriss, 2001) et plus récemment les modèles de Hawkes (Alfonsi and Blanc, 2016; Bacry and Muzy, 2014). Le Chapitre 6 a déjà défini un processus de prix grâce à l'équation A.2, sous la forme :

$$y_t = \frac{1}{\mathcal{L}} \int_0^t \frac{\mathrm{d}s\, m_s}{\sqrt{4\pi D(t-s)}} e^{-\frac{(y_t - y_s)^2}{4D(t-s)}} := \frac{1}{\mathcal{L}} \int_0^t m_s P(y_t - y_s, t-s)\mathrm{d}s, \tag{A.1}$$

où P est la fonction de Green associée à la diffusion classique pour une volatilité $\sqrt{2D}$. Toutefois, ce processus est restrictif pour une raison simple : à temps grands, il se comporte comme un propagateur en $1/\sqrt{t}$. Or, Bouchaud and Potters (2003) a montré que pour obtenir un processus diffusif basé seulement sur un modèle de propagateur en loi puissance, la relation $\alpha = \frac{1-\gamma}{2}$ doit être vérifiée, où α est l'exposant du propagateur (ici $1/2$) et $\gamma > 0$ est l'exposant de l'autocorrélation des trades qui

permet de contrer la mean-reversion du propagateur. Pour satisfaire cette relation avec $\alpha = 1/2$, il faudrait donc $\gamma = 0$ ce qui n'a pas de sens. Ainsi, quel que soit le processus de trades dans le modèle du Chapitre 6, la composante liquidité *mean-revert* toujours à 0, et une synchronisation exogène est nécessaire pour rendre le prix diffusif (due à de l'information publique par exemple). Elargir la classe de modèles à des processus à *mean-reversion* moins forte est donc intéressant : la composante de liquidité pourrait alors jouer un rôle à part entière dans le processus de prix.

Nous développons dans ce premier appendice une classe de processus qui élargit celle du Chapitre 6, en introduisant des *path-dependent kernels* généraux $P(x_t - x_s, t - s)$. Un choix de fonctions particulières permet dans certaines limites de retrouver des modèles de propagateur en $1/t^\alpha$ où $\alpha \neq 1/2$. Lorsque $0 < \alpha < 1/2$, ces modèles peuvent être microfondés et correspondent à des agents qui évoluent selon un *Grey Brownian Motion*. Le système agrégé suit alors une équation de diffusion fractionnaire, et comme précédemment ces processus reproduisent les principales propriétés de l'impact – en particulier un impact en racine universel à temps longs. Le cas $1/2 < \alpha$ peut également être écrit, mais ne correspond *a priori* pas à une dynamique d'agents sous-jacente.

N'étant pas issue d'un article publié mais plutôt un résumé de recherches en cours, l'annexe qui suit n'est pas exempte d'erreurs – et si elle en contient je m'en excuse.

Abstract We generalize the notion of path-dependent kernel that has emerged naturally in the derivation of price impact in supply- and demand-driven financial markets. We write the price as the convolution of past trades m_t with a path-dependent kernel $P(x_t - x_s, t - s)$ where x_t is the price at time t. In this framework, one can define an order book φ that can be expressed in terms of P, so that in some cases properties of P can be interpreted in a dual way as properties on how the orders to buy/sell behave on the order book. This compatibility with an order book representation is novel for that kind of processes, and seems crucial in the understanding and interpretation of price impact in a context where prices are the result of supply and demand. We show that the order book follows a PDE for some markovian kernels – amongst which, α-stable kernels – and that the absence of manipulation holds for a very general shape of kernels. Finally, one of the main results is the universal square root upper bound for price impact, attained for large trading rates.

A.1 Path-dependent kernels

A.1.1 Definition for the price under an external pressure

We introduce here the general notion of *path-dependent kernels* (PDK) $P(x, t)$ where x represents the price and $t \geq 0$ represents the time. We define the following class of price dynamics under a pressure m_t (that can be interpreted as the derivative of position of the investor, i.e. his trading process) :

$$x_t = \frac{1}{\mathcal{L}} \int_{-\infty}^{t} m_s P(x_t - x_s, t-s)\mathrm{d}s, \tag{A.2}$$

where $\mathcal{L}$ is a liquidity parameter, P is such that $P(x,0) = \delta_0$ and x_t represents the expected price at time t under the perturbation $(m_s)_{s\leq t}$. Note that this equation gives a deterministic price path, that should be interpreted as the average price path over some noise.

Definition A.1.1. *A path-dependent kernel P is said to be* admissible *if it defines a unique price process for any (regular) investment process $(m_t)_t$ and* complete *if any (regular) price trajectory is attainable by at least one trading process.*

Conjecture A.1.1. *The Gaussian kernel $P(x,t) = \frac{e^{-\frac{x^2}{2t}}}{\sqrt{2\pi t}}$ (used in DBMB) is admissible and complete.*

Because of the physics underlying the equation, I have little doubt that this conjecture is true, however proving it seems quite technical and I did not find a way so far. I hope that it will be proven in a near future though.

A.2 Dual definition of the order book

Following this definition, we now define an order book $\varphi(x,t)$, such that $\forall t \geq 0$:

$$\varphi(x,t) = \varphi^0(x) + \int_0^t m_s P(x - x_s, t-s)\mathrm{d}s, \qquad \varphi^0(x) \equiv -\mathcal{L}x \tag{A.3}$$

Such definition of an order book might be a little abrupt without a little background. The first convention in the interpretation of φ, is that it represents for each price level the difference between buy and sell orders densities, so that it is positive left to the price (where by definition only buy orders are awaiting) and negative right to the price (where one finds only awaiting sell orders). In this way, all order book information is concentrated into one only function, instead of dealing with one function for buy orders and one for sell orders – exactly like in Chapter 6. Note that some constraints on the motion of orders are necessary to take this step – which we don't develop here. Second, this function stands for *latent* orders, i.e. any trading intention that will reveal itself when it crosses the price [1]. The above definition calls for a few early comments :

— Because P is admissible, the price x_t is the only zero of φ. $\varphi(x < x_t) > 0$ and $\varphi(x > x_t) < 0$ represent respectively the density of buy and sell orders around price x (called Marginal Supply and Demand in Chapter 7).

1. In particular, it does not only represent the orders that are publicly displayed on the official order book, as these are only a tiny part of the "true" supply and demand at any time.

— The shape of the order book in the absence of perturbation m is $\varphi^0(x)$. Unless mentioned otherwise we will always consider that the order book is in the linear state $\varphi^0(x) = -\mathcal{L}x$ at $t = 0$, with $m_{t<0} = 0$ (see below for a justification of this choice).

— A perturbation of the order book (i.e. a volume δm that is added or removed at some price x and at time t) propagates on other price levels x' at subsequent times t' according to $P(x' - x, t' - t)$. In particular, perturbations add up linearly to the order book.

— The condition $P(x, 0) = \delta_0$ means that new orders are injected *exactly* at the price. If one steps away from the financial interpretation, more generally the process given by Eq. A.2 stands for the evolution of a reaction-diffusion system on which an external action is operated at the interface between the two phases.

From the last two remarks, it appears that the choice of a kernel P amounts to giving particular behaviours to the orders. A Gaussian kernel for example implies an independent, markovian, short range evolution of orders, as seen in Chapter 6. Other behaviours will be investigated later in the article. Before studying particular kernels, one can come with the following, general definitions :

Definition A.2.1. *A kernel P is* (strictly) positive *iff* $\forall t\ P(x,t) \geq (>)0$.

The positivity of a kernel can be intuitively interpreted by saying that the effect of an extra buy (resp. sell) volume at any point x in the order book at time t can only increase the volume of buy orders/decrease the number of sell orders (resp. decrease the volume of buy orders/increase the number of sell orders) at any price x' and at any subsequent time $t' > t$. While one could think of negative effects in very general situations, they would not make much sense in our framework so we will restrict ourselves to positive kernels from now on.

Definition A.2.2. *A kernel is* conservative *iff* $\forall t\ V(t) \equiv \int_{-\infty}^{\infty} P(x,t)\mathrm{d}x = 1$. *If $V(t)$ is a decreasing function of t, the order book is* absorbent. *If $V(t)$ is increasing, it is* magnifying.

Remark A.2.1. *Note that since $P(x, 0) = \delta_0$, one always has $V(0) = 1$.*

Remark A.2.2. *We will also assume that $\exists t$, s.t. $P(x,t) \neq \delta_0$ – otherwise there is no dynamics.*

This definition also calls for an interpretation. With conservative kernels, new orders cannot appear and orders cannot disappear otherwise than in a transaction. In physical terms, this amounts to saying that conservative kernels preserve the mass of the system. With absorbent kernels, an extra volume triggers an opposite market reaction, with the appearance of new sell orders (or the disappearance of existing buy orders), therefore reducing its impact. Conversely, with an magnifying kernel an extra buy volume triggers a same-side market reaction, with the appearance of new buy orders (or the disappearance of existing sell orders), therefore exacerbating its impact (and leading to possible price manipulations, see below). Unless mentioned otherwise, we will restrict ourselves to conservative kernels in the following. Let us now end this section by giving a last definition :

Definition A.2.3. *A kernel is* centred *iff* $\int_{-\infty}^{\infty} xP(x,t)\mathrm{d}x = 0\ \forall t$.

This condition imposes that the effect of a perturbation is centred on 0, affecting both buy orders below the price and sell orders above the price in a balanced way. From the above three definitions, one can derive some relations between the state of the order book at a time $t > 0$, and the trading path $(m_s, x_s)_{0\leq s\leq t}$.

Proposition A.2.1. *Starting from the stationary shape at $t = 0$, for any execution schedule $(m_s)_{0\leq s\leq t}$ one has the following equality :*

$$m \star V_t \equiv \int_0^t m_s V(t-s)\mathrm{d}s = \int_{-\infty}^{\infty} \left[\varphi(x,t) - \varphi(x,0)\right] \mathrm{d}x \tag{A.4}$$

In particular, for conservative kernels, one has

$$Q_t \equiv \int_0^t m_s \mathrm{d}s = \int_{-\infty}^{\infty} \left[\varphi(x,t) - \varphi(x,0)\right] \mathrm{d}x \tag{A.5}$$

where Q_t is the total executed quantity.

Proof A.2.1. *By definition of φ, $\varphi(x,t) = \varphi(x,0) + \int_0^t m_s P(x_t - x_s, t-s)\mathrm{d}s$ hence by integrating on x and switching integrals on the right-hand side,*

$$\begin{aligned} \int_{-\infty}^{\infty} \left[\varphi(x,t) - \varphi(x,0)\right] \mathrm{d}x &= \int_0^t \int_{-\infty}^{\infty} m_s P(x - x_s, t-s)\mathrm{d}x\mathrm{d}s \\ &= \int_0^t m_s V(t-s)\mathrm{d}s \equiv m \star V_t \end{aligned} \tag{A.6}$$

The result for conservative kernels follows by noticing that in this case $V = 1$ by definition.

This result for conservative kernels means that the volume traded at the price at past times remains as an excess volume spread out in the order book at later times. For non-conservative kernels, this excess volume can be magnified or absorbed.

Proposition A.2.2. *Iff the order book is centred and conservative, then $\forall (m_s)_{s\leq t}$,*

$$C_t \equiv \int_0^t m_s x_s \mathrm{d}s = \int_{-\infty}^{+\infty} x \left[\varphi(x,t) - \varphi(x,0)\right] \mathrm{d}x \tag{A.7}$$

where C_t can be interpreted by the ex-ante cost of a trading trajectory. In particular, if $\int_0^T m_s \mathrm{d}s = 0$, C_t is the cost of the round-trip.

Proof A.2.2. *Similarly to the previous proof, one has that*

$$\begin{aligned}\int_{-\infty}^{\infty} x\left(\varphi(x,t)-\varphi(x,0)\right)\mathrm{d}x &= \int_0^t\int_{-\infty}^{\infty} m_s x P(x-x_s,t-s)\mathrm{d}x\mathrm{d}s \\ &= \int_0^t\int_{-\infty}^{\infty} m_s (x+x_s) P(x,t-s)\mathrm{d}x\mathrm{d}s \\ &= \int_0^t x_s m_s V(t-s)\mathrm{d}s + \int_0^t m_s \int_{-\infty}^{\infty} x P(x,t-s)\mathrm{d}x\mathrm{d}s \\ &= \int_0^t x_s m_s V(t-s)\mathrm{d}s\end{aligned} \tag{A.8}$$

where the last equality comes from the fact that P is centred. When the order book is also conservative, $V=1$ and $\int_{-\infty}^{\infty} x\left[\varphi(x,t)-\varphi(x,0)\right]\mathrm{d}x = \int_0^t x_s m_s \mathrm{d}s = C_t$.

This result implies that the profit of a trading strategy can be computed ex-post at any time by looking at the shape of the perturbed order book.

A.3 Dynamics of markovian (latent) order books

The above approach of defining directly a price process and from it deriving an order book is rather unusual. A more natural and common construction is to first define an order book, for instance by defining its dynamics, and only then to regard the price as the point at which supply meets demand (so in our words, the zero of the φ function) : this is actually how the Chapter 6 came up with the Gaussian PDK, starting from a diffusive behaviour of orders. In this section, we will get back to this more concrete approach and work out the link between the "abstract" price process defined by Eq. A.2 and the dynamics of its dual order book for particular shapes of kernels.

A.3.1 Consistency of the dynamics and stationary state

When one interprets Eq. A.3 as an evolution equation for the orders present on the book φ, one may ask the particles that are already present at $t=0$ to follow the same dynamics that those which respond to the perturbation m. Therefore, one would like the following relation instead of Eq. A.3, where the initial condition is subject to the same evolution than the perturbation :

$$\varphi(x,t) = \int_{-\infty}^{\infty} \varphi(y,0)P(x-y,t)\mathrm{d}y + \int_{-\infty}^{t} m_s P(x-x_s,t-s)\mathrm{d}s \tag{A.9}$$

Let us denote $\varphi_{st}(y)$ the stationary shape in the absence of perturbation ($m=0$), if it exists, so that $\varphi_{st}(x) = \int_{-\infty}^{\infty} \varphi_{st}(y)P(x-y,t)\mathrm{d}y$.

Proposition A.3.1. *If the system admits a stationary state φ_{st} normalized so that the price (if*

unique) is zero, such that $\exists n \in \mathbb{N}$ *s.t.* $\varphi_{st}^{(n)} \in L^2(\mathbb{R})$*, then for any* conservative *kernel, one has :*

$$\varphi_{st}(x) = -\mathcal{L}x \tag{A.10}$$

where $\mathcal{L}$ *is called the* liquidity *of the system. Furthermore, the corresponding kernel must be* centred.

Proof A.3.1. *Let* t *be fixed.* $\varphi^{(n)} \in L^2(\mathbb{R})$ *so one can take the Fourier transform on the equation* $\varphi_{st}^{(n)}(x) = \int_{-\infty}^{\infty} \varphi_{st}^{(n)}(y) P(x-y,t)\mathrm{d}y$ *to find* $\hat{\varphi}_{st}^{(n)} = \hat{\varphi}_{st}^{(n)} \hat{P}_t$ *where* $P_t \equiv P(.,t)$*. Since* P *is conservative, then* $\forall \lambda \neq 0$ $\hat{P}_t(\lambda) \neq 1$ *and therefore* $\hat{\varphi}_{st}^{(n)}(\lambda) = 0$*. Therefore* $\varphi_{st}^{(n)} = Cte$ *and* φ_{st} *is polynomial of degree* $d \leq n$*. Let us assume that* $d \geq 2$ *(so* $n \geq 2$*). By writing* $\varphi_{st}^{(n-2)}(x) = ax^2 + bx + c$ *with* $a \neq 0$*, one can find that* $a = 0$ *by solving for the coefficients in* $ax^2+bx+c = \int_{-\infty}^{\infty}(ay^2+by+c)P(x-y,t) = \int_{-\infty}^{\infty}(a(y+x)^2 + b(y+x) + c)P(y,t)\mathrm{d}y$*, since one has*

$$a = C_0 a, b = C_0 b + 2C_1 a, c = C_0 c + C_1 b + C_2 a \tag{A.11}$$

where $C_i \equiv \int_{-\infty}^{\infty} y^i P(y,t)\mathrm{d}y$*, so if one assumes that* $a \neq 0$ *then* $C_0 = 1$ *from the first equality which forces* $C_1 = 0$ *from the second and therefore* $C_2 = 0$*, contradiction with the fact that* $C_2 > 0$ *(because of Remark A.2.2), so* $a = 0$*, again contradiction. Therefore* $d < 2$ *and* $\varphi_{st}(x) = bx + c$*. By writing the same condition on the coefficients as above, one finds that the second equations becomes* $b = C_0 b$ *which forces either* $C_0 = 1$ *or* $b = 0$ *which in turn forces* $C_0 = 1$ *from the third equation. So,* $C_0 = 1$*, and then* $C_1 = 0$ *from the third equation (which is a good news, since the kernel is conservative so that had to be satisfied). The normalization chosen so that the price is zero finally forces* $c = 0$*. This demonstrates that* φ_{st} *is necessarily of the form* $\varphi_{st}(x) = -\mathcal{L}x$*, and that* $\int_{-\infty}^{\infty} yP(y,t)\mathrm{d}y = 0$ *so the kernel must be centred.*

The above proposition is important for two reasons. First, implies that the density of orders grows linearly when one goes further from the price, as observed on the Bitcoin, and as expected to generate the square root impact law in the regime of fast trading (see below). Second, it justifies the choice of conservative, centred kernels that allow the order book to follow a consistent dynamics. The next section will be devoted to describing further the dynamics of φ in the particular cases where its evolution follows a PDE.

A.3.2 Markovian kernels and PDE for the order book

A useful class of processes is that for which the order book follows a partial differential equation (PDE) that drives its evolution. In such cases indeed, one can use some PDE tools to compute price paths, instead of computing the solution of an integral equation. In such cases, φ is markovian, which we can use to characterize the corresponding kernels. In particular, one has the following proposition :

Proposition A.3.2. *If the dynamics is markovian, then P must be of the form*

$$P(x,t) = (C)\int_{-\infty}^{+\infty} e^{f(\lambda)t+i\lambda x}d\lambda \tag{A.12}$$

Proof A.3.2. *If the dynamics is markovian, then for any $\varphi \in C_c^\infty$ one has that*

$$\forall s<t, \int_{-\infty}^{+\infty}\varphi(y,0)P(x-y,t) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\varphi(y,0)P(z-y,s)P(x-z,t-s) \tag{A.13}$$

and therefore

$$P(y,t) = \int_{-\infty}^{+\infty} P(y-z,t-s)P(z,s) \quad \forall s<t \tag{A.14}$$

Taking the Fourier transform in y, $\widehat{P}(\lambda,t) = \widehat{P}(\lambda,t-s)\widehat{P}(\lambda,s) \quad \forall s<t$ implying that $\widehat{P}(\lambda,t) = e^{f(\lambda)t}$ so the kernel P reads $P(x,t) = \int_{-\infty}^{+\infty} e^{f(\lambda)t+i\lambda x}d\lambda$.

Proposition A.3.3. *$f(0)$ is real, and the following equivalences are true :*

- *$f(0) = 0 \Leftrightarrow P$ is conservative,*
- *$f(0) < 0 \Leftrightarrow P$ is absorbent,*
- *$f(0) > 0 \Leftrightarrow P$ is magnifying.*

Proof A.3.3. *Simply integrate Eq. A.12 over x to find that $V(t) = e^{f(0)t}$.*

Proposition A.3.4. *Is f real then P is centred.*

Proof A.3.4. *Replace x by $-x$ in Eq. A.12, and use the fact that P is real.*

Question A.3.1. *Is the reverse true ? What are the conditions for f to be real ?*

Below we will only consider functions f that are real, since it is likely that the behaviour of P is strange otherwise. One can find additional properties on the function f. In particular, f is an odd function since P is real. One case that deserves to be investigated, is when $f(\lambda) = \sum_{k\geq 1} a_{2k}\lambda^{2k}$. In such cases, φ follows the PDE

$$\partial_t\varphi(x,t) = \sum_{k\geq 1} a_{2k}(-1)^k\partial_{2k}\varphi(x,t) \tag{A.15}$$

The Gaussian case is a particular of such a markovian kernel since it corresponds to $f(\lambda) = -a\lambda^2$. Another particular case is that of PDK P that have a scaling form $P = \frac{1}{t^{\alpha^{-1}}}F(\frac{x}{t^{\alpha^{-1}}})$ for $\alpha \in (0,2]$, which are given by

$$P(x,t) = \int_{-\infty}^{+\infty} e^{-a|\lambda|^\alpha t+i\lambda x}\mathrm{d}\lambda \tag{A.16}$$

On long time scales, one recovers propagators with exponent $\alpha^{-1} > \frac{1}{2}$. It is tempting to say that these kernels correspond to alpha-stable random walks for φ's particles. However, in this case, particles can jump over the price and hence the reaction term would no longer be localized : the interpretation is thus wobbly.

A.3.3 Impact and relaxation for α-stable kernels

α-stable kernels have the property that for small perturbations, the price reads

$$\begin{aligned} x_t &= x_0 + \frac{1}{\mathcal{L}} \int_0^t m_s \int_{-\infty}^{+\infty} e^{-a|\lambda|^\alpha (t-s)} \mathrm{d}\lambda \mathrm{d}s \\ &= x_0 + \frac{1}{\mathcal{L}} \int_0^t \frac{m_s I(\alpha) \mathrm{d}s}{(a(t-s))^{\alpha^{-1}}} \\ &\equiv x_0 + \frac{1}{\mathcal{L}} \int_0^t m_s G^{a,\alpha}(t-s) \mathrm{d}s \end{aligned} \tag{A.17}$$

where $I(\alpha) = \int_{-\infty}^{+\infty} e^{-|\lambda|^\alpha} d\lambda$. The diffusion exponent of the particles on the order book thus determine how impact behaves asymptotically for small perturbations, and to which classical propagator model the model converges in this regime. However, as soon as $\alpha < 2$, for a constant pressure $m_t = m$ the propagator predicts a price trajectory $x_t \sim t^{1-\frac{1}{\alpha}}$ with $1 - \frac{1}{\alpha} < \frac{1}{2}$ so for small t this always breaches the square root bound derived below : The "propagator" regime can only appear for large enough t, after a "square-root impact" regime for small t. This would be compatible with some empirical observations, where the impact becomes more concave than square root for large volumes.

A.3.4 Non-arbitrage

Definition A.3.1. *The price is said to be* non-arbitrable *if* $\mathcal{C} \equiv \int_{-\infty}^{+\infty} m_s x_s \mathrm{d}s \geq 0 \ \forall (m_s)_s$. *It is said to be* strictly non-arbitrable *if* $\mathcal{C} = 0 \iff m_s = 0 \ \forall s$.

Proposition A.3.5. *All markovian kernels such that f is real and negative are non-arbitrable.*

Proof A.3.5. *A markovian kernels has the form* $P(x,t) = (C) \int_{-\infty}^{+\infty} e^{f(\lambda)t + i\lambda x} d\lambda$. *One can easily check that* $\mathcal{C}$ *can be identically rewritten as a quadratic form :*

$$\mathcal{C} = \frac{1}{2} \int_0^T \int_0^T \mathrm{d}s\mathrm{d}s' \, m_s \, M(s,s') \, m_{s'}, \tag{A.18}$$

where $M(s,s')$ is a non-negative operator, since it can be written as a "square", $M = KK^\dagger$, or more

precisely :

$$M(s,s') = \frac{D}{\mathcal{L}} \int_{-\infty}^{\infty} \mathrm{d}\lambda\, \lambda^2 \int_{-\infty}^{+\infty} \mathrm{d}u\, K_\lambda(s,u) K_\lambda^*(s',u), \qquad K_\lambda(s,u) \equiv \Theta(s-u) e^{f(\lambda)(s-u)+i\lambda x_s}. \tag{A.19}$$

A.4 Grey Brownian Motion, fractional diffusion and propagators with $0 < \alpha < 1/2$

One important case is when the orders follow a Grey Brownian Motion (Schneider, 1990). This case is currently under study, but we present some first ideas below (although some might be wrong). In this case, the order book function φ can be shown to follow a fractional diffusion equation :

$$\varphi(x,t) = \varphi(x,0) + \frac{1}{\Gamma(2\alpha)} \int_0^t \mathrm{d}s (t-s)^{2\alpha-1} \Delta\varphi(x,s), \quad 0 < \alpha \leq \frac{1}{2}. \tag{A.20}$$

The Green's function $G_\alpha(x,t)$ is given in Schneider (1990), and can be written as :

$$G_\alpha(x,t) = \frac{1}{t^\alpha} g_\alpha(\frac{x}{t^\alpha}), \tag{A.21}$$

where g_α is some function described in Schneider (1990). At large times, this process is therefore equivalent to a propagator with exponent $\alpha \in (0, \frac{1}{2})$. These results are important for two reasons :

i The price process obtained is micro-founded,

ii The price process that corresponds to this limiting propagator can be made diffusive if the trades have a power-law autocorrelation with exponent $\gamma = 1 - 2\alpha$ (Bouchaud et al., 2009). The question of whether this would remain true in this case is still open however.

iii The square root impact hold not only for high participation rates, but also for all long enough execution.

Let us develop this last point further. Let us look for an auto-consistent power-law solution $y_t = At^\delta$ of the equation :

$$y_t = \frac{1}{\mathcal{L}} \int_0^t \frac{\mathrm{d}s m_0}{(t-s)^\alpha} g_\alpha \left(\frac{y_t - y_s}{(t-s)^\alpha} \right). \tag{A.22}$$

By doing the change of variable $s \to t - s$, one has :

$$At^\delta = \frac{m_0}{\mathcal{L}} \int_0^t \frac{\mathrm{d}s}{s^\alpha} g_\alpha \left(\frac{At^\delta - A(t-s)^\delta}{s^\alpha} \right). \tag{A.23}$$

Since the function $g_\alpha(x)$ decays very quickly to 0 for $x \gg 1$, the integral bound can be restricted to s's such that $At^\delta - A(t-s)^\delta < Cs^\alpha$, where C is some large enough generic constant whose value can vary from equation to equation. This inequality can be rewritten under the form $s < Ct^{\frac{1-\delta}{1-\alpha}}$.

One therefore has that :

$$\begin{aligned} t^\delta &\underset{t\to\infty}{\sim} \int_0^{Ct^{\frac{1-\delta}{1-\alpha}}} \frac{\mathrm{d}s}{s^\alpha} g_\alpha\left(\frac{At^\delta - A(t-s)^\delta}{s^\alpha}\right) \\ &\underset{t\to\infty}{\sim} \int_0^{Ct^{\frac{1-\delta}{1-\alpha}}} \frac{\mathrm{d}s}{s^\alpha} \\ &\underset{t\to\infty}{\sim} t^{1-\delta}. \end{aligned} \tag{A.24}$$

The only auto-consistent solution is therefore $\delta = 1/2$, i.e. exactly a square-root impact ! The prefactor A *a priori* depends on the system parameters m_0 and $\mathcal{L}$. For small t however, the impact of a metaorder with constant rate m_0 is behaves as $t^{1-\alpha}$, as it would be the case for the usual propagator of exponent α.

A.5 Properties of price impact

Theorem A.5.1. *For any (strictly) positive, conservative PDK, the price impact of signed quantities Q_t (starting from the stationary order book shape) has the upper bound*

$$x_t < \sqrt{\frac{2Q_t}{\mathcal{L}}} \tag{A.25}$$

Proof A.5.1. *Let $Q_t \equiv \int_0^t m_s \mathrm{d}s$ be the (buy) executed quantity between 0 and t, with $\varphi(x,0) = -\mathcal{L}x$. Then by (strict) positivity of P, $\varphi(x,t) = \varphi(x,0) + \int_0^t m_s P(x_t - x_s, t-s)\mathrm{d}s > \varphi(x,0)$. Together with lemma A.2.1, this gives us that $Q_t = \int_{-\infty}^{\infty} (\varphi(x,t) - \varphi(x,0))\,\mathrm{d}x \geq (>) \int_0^{x_t} (\varphi(x,t) - \varphi(x,0))\,\mathrm{d}x$ where x_t is the price at T. Since x_t is uniquely defined, one must have that $\varphi(x < x_t) > 0$ so by replacing $-\varphi(x,0)$ by $\mathcal{L}x$, one finds that $Q_t > \int_0^{x_t} \mathcal{L}x\mathrm{d}x = \frac{1}{2}\mathcal{L}x_t^2$ or equivalently $x_t < \sqrt{\frac{2Q_t}{\mathcal{L}}}$.*

It seems obvious that the upper bound is reached for infinitely fast executions, as in this limit the orders on the order book do not move and therefore the results should be the same as in the gaussian case – but this has still to be proven.

A.6 Conclusion

We have introduced a new class of integral processes, defined as the convolution of a path-dependent kernel with some source process. In the case of financial markets, such processes would give the impact on the price of an incoming trading flow. In general, this would give the response to a reaction-diffusion system to a perturbation that occurs at the interface between the two phases. In some cases, our process is compatible with an order-book representation that gives some micro-foundations to price impact. Prices are found to be non-manipulable for a large family of kernels, and the price/order book duality is evidenced. We show how the choice of the kernel amounts to

giving particular dynamics to the underlying supply and demand. We show that propagators with exponent $\alpha \in (0, \frac{1}{2})$ are found when orders follow a grey brownian motion, i.e. a brownian motion where particules can be frozen for some amount of time with a well-chosen distribution, which may represent agents with different time horizons. We finally show that the price impact of directional trading is upper bounded by the square root of the volume. A next step would be to consider a time-varying liquidity $\mathcal{L}_t$, which would be crucial to account for the reality of financial markets and thus in order to derive optimal execution strategies.

Annexe B

Exécution optimale : résultats numériques et asymptotiques

Ce second appendice est consacré au problème de l'exécution optimale dans le modèle du Chapitre 6. Celui-ci étant analytiquement compliqué, nous nous limiterons à une étude asymptotique dans les régimes d'impact extrêmes (exécution très rapide ou exécution très lente), ainsi qu'à la simulation numérique des politiques de trading optimales dans en situation de liquidation et en situation d'arbitrage. Pour finir, nous évaluerons la performance, à l'intérieur de notre modèle, des politiques de trading optimal préconisées par les modèle d'impact classiques.

B.1 Résultats théoriques asymptotiques

Avant de nous lancer dans les simulations, nous présentons d'abord quelques résultats asymptotiques sur l'impact de métaordres qui nous seront très utile pour interpréter les résultats numériques.

B.1.1 Exécution lente

Comme nous l'avons vu à plusieurs reprises dans le Chapitre 6, lorsque le taux d'exécution est suffisamment faible pour que les variations de prix soient telles que $\Delta y^2 \ll D\Delta t$ (ce qui rappelons-le est le cas s'il vérifie $m \ll J$), alors le prix est tel que

$$y_t = \frac{1}{\mathcal{L}} \int_0^t \frac{\mathrm{d}s\, m_s}{\sqrt{4\pi D(t-s)}} e^{-\frac{(y_t - y_s)^2}{4D(t-s)}} \simeq \frac{1}{\mathcal{L}} \int_0^t \frac{\mathrm{d}s\, m_s}{\sqrt{4\pi D(t-s)}} := y_t^{\text{slow}}. \tag{B.1}$$

Plus précisément, si $\Delta y^2 < \epsilon D\Delta t$ avec $\epsilon \ll 1$,

$$
\begin{aligned}
\mid y_t - y_t^{\text{slow}} \mid &\leq \frac{1}{\mathcal{L}} \int_0^t \mid \frac{\mathrm{d}s\, m_s}{\sqrt{4\pi D(t-s)}} \left(1 - e^{-\frac{(y_t - y_s)^2}{4D(t-s)}}\right) \mid \\
&\leq \frac{1}{\mathcal{L}} \int_0^t \frac{\mathrm{d}s \mid m_s \mid}{\sqrt{4\pi D(t-s)}} \left(1 - e^{-\epsilon}\right) \\
&\leq \epsilon \frac{1}{\mathcal{L}} \int_0^t \frac{\mathrm{d}s \mid m_s \mid}{\sqrt{4\pi D(t-s)}} \\
&\propto \epsilon \mid y_t^{\text{slow}} \mid,
\end{aligned}
\tag{B.2}
$$

où le symbole $\propto$ doit être compris comme « est du même ordre de grandeur que ». Ainsi, le prix pour $\epsilon \to 0$ est « aussi proche que l'on veut » dans un certain sens du prix y_t^{slow} donné par le propagateur simple. [1]

B.1.2 Exécution rapide

Reprenons l'équation 6.18 du Chapitre 6 :

$$
\mathcal{L} y_t |\dot{y}_t| \approx m_t \left[1 + D\left(3\frac{\ddot{y}_t}{\dot{y}_t^3} - 2\frac{\dot{m}_t}{m_t \dot{y}_t^2}\right) + O\left(\frac{J^2}{m^2}\right)\right]. \tag{B.3}
$$

Celle-ci peut être réécrite en utilisant l'approximation $\mathcal{L} y_t |\dot{y}_t| \approx m_t$ dans le membre de droite, ce qui donne

$$
\mathcal{L} y_t |\dot{y}_t| \approx m_t + 2J\left(\frac{\dot{m}_t Q_t}{m_t^2} - \frac{3}{2}\right). \tag{B.4}
$$

Après intégration par parties, on obtient donc

$$
\begin{aligned}
y_t &\approx \sqrt{\frac{2}{\mathcal{L}} \int_0^t m_s + 2J\left(\frac{\dot{m}_s Q_s}{m_s^2} - \frac{3}{2}\right) \mathrm{d}s} \\
&\approx \sqrt{\frac{2}{\mathcal{L}}} \sqrt{Q_t \left(1 - 2\frac{J}{m_t}\right) - Jt},
\end{aligned}
\tag{B.5}
$$

qui nous amène à l'expression suivante du coût d'exécution :

$$
\begin{aligned}
\mathcal{C} &:= \int_0^T m_t y_t \mathrm{d}t \\
&\approx \sqrt{\frac{2}{\mathcal{L}}} \int_0^T m_t \sqrt{Q_t \left(1 - 2\frac{J}{m_t}\right) - Jt} \mathrm{d}t
\end{aligned}
\tag{B.6}
$$

Remarquez que la correction en $\frac{Q_t J}{m_t}$ est d'ordre Jt si m_t est suffisamment lisse, puisque $Q_t := \int_0^t m_s \mathrm{d}s$. Le cas le plus simple est lorsqu'un volume Q est exécuté à taux constant dans une fenêtre

1. Pour obtenir une preuve mathématiquement rigoureuse, il faudrait commencer par définir une notion de proximité pertinente pour ce problème : nous ne nous en occuperons pas ici.

d'exécution T :

$$\begin{aligned} m_t &= \frac{Q}{T} \\ Q_t &= \frac{Qt}{T} \end{aligned} \tag{B.7}$$

Dans ce cas, l'expression B.6 donne immédiatement $\mathcal{C} \approx \left(1 - \frac{3}{2}\frac{JT}{Q}\right)\mathcal{C}_{inst}$, où $\mathcal{C}_{inst} := \frac{2\sqrt{2}}{3}\frac{Q^{\frac{3}{2}}}{\mathcal{L}^{\frac{1}{2}}}$ est le coût associé à une exécution instantanée ($T \to 0$).

Plus généralement, on peut aller plus loin dans l'approximation de l'équation B.6 en développant la racine carrée et écrire :

$$\begin{aligned} \mathcal{C} &\approx \sqrt{\frac{2}{\mathcal{L}}}\int_0^T \left(m_t\sqrt{Q_t} - J\sqrt{Q_t} - \frac{Jtm_t}{2\sqrt{Q_t}}\right)\mathrm{d}t \\ &\approx \sqrt{\frac{2}{\mathcal{L}}}\left(\frac{2}{3}Q^{\frac{3}{2}} - \int_0^T J\sqrt{Q_t}\mathrm{d}t - \int_0^T \frac{Jtm_t}{2\sqrt{Q_t}}\mathrm{d}t\right) \\ &\approx \sqrt{\frac{2}{\mathcal{L}}}\left(\frac{2}{3}Q^{\frac{3}{2}} - \int_0^T J\sqrt{Q_t}\mathrm{d}t - JT\sqrt{Q_t} + \int_0^T J\sqrt{Q_t}\mathrm{d}t\right) \\ &\approx \sqrt{\frac{2}{\mathcal{L}}}\left(\frac{2}{3}Q^{\frac{3}{2}} - JT\sqrt{Q}\right) \\ &\approx \left(1 - \frac{3}{2}\frac{JT}{Q}\right)\mathcal{C}_{inst}. \end{aligned} \tag{B.8}$$

Toute notion de stratégie d'exécution a complètement disparue : le coût ne dépend que du volume total exécuté Q et de l'horizon T (et l'expression est donc la même que celle obtenue pour un taux d'exécution constant) ! Cela suggère que toutes les stratégies ont des coûts équivalents au premier ordre, dès qu'elles vérifient $m_t \gg J \ \forall t$. Autrement dit, lorsque l'on souhaite liquider rapidement une position, la manière exacte de le faire importe peu : seul l'horizon importe.

B.2 Résultats numériques : liquidation optimale et exploitation d'*alpha*

Procédons maintenant à des simulations numériques pour confirmer les résultats ci-dessus et atteindre les régimes intermédiaires, ainsi que pour acquérir une intuition sur le modèle lui-même. Deux situations très différentes nous intéresseront : (i) la situation de *liquidation*, où un gérant de portefeuille doit liquider une quantité Q sur un horizon T fixé en minimisant ses coûts d'impact, et (ii) la situation d'exploitation d'*alpha* où un gérant de portefeuille essaie de tirer le meilleur parti d'une prédiction qu'il a réalisée sur les prix futurs.

B.2.1 Liquidation optimale

Intéressons-nous donc à la liquidation d'un volume donné Q sur un intervalle fixé T. Pour cela, nous simulons l'Equation 6.7 du Chapitre 6 grâce à un schéma numérique de Crank-Nicolson, et

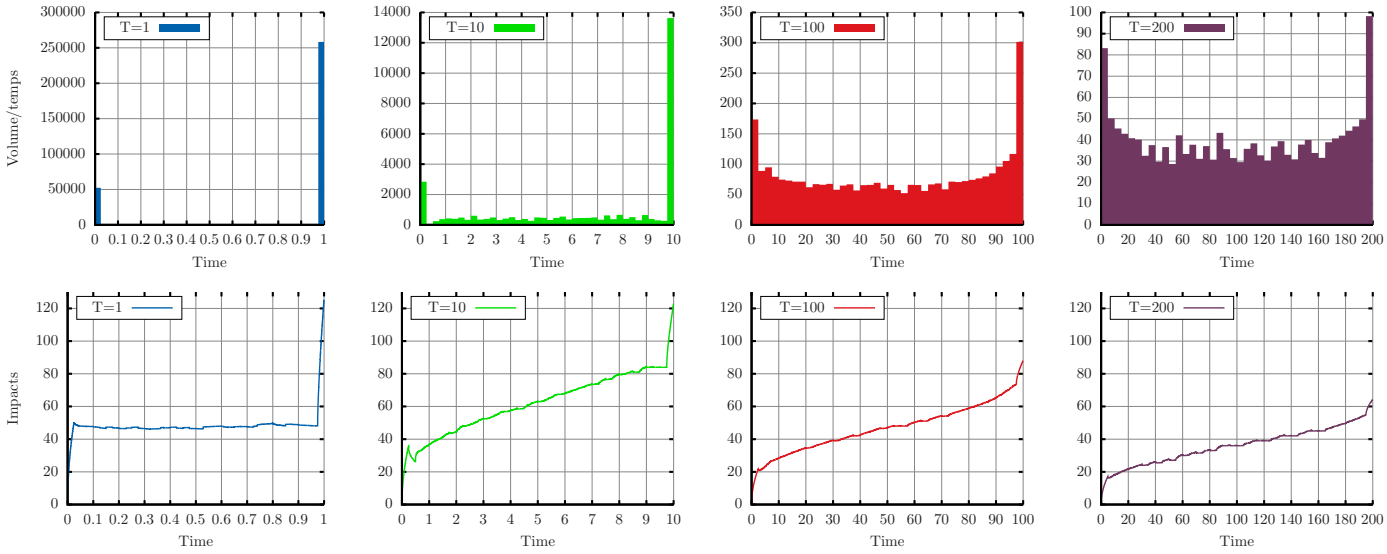

FIGURE B.1 – Profils de liquidation et trajectoires d'impact pour quatre horizons différents. (haut) Profils de liquidation définis comme volume exécuté par unité de temps. (bas) Profils d'impact correspondant. Plus l'horizon T est court, plus l'exécution se concentre aux bords de la fenêtre. A horizons longs, on retrouve la solution symétrique du propagateur. Le prix correspondant suit une trajectoire en S d'autant plus marquée que T est petit, et l'impact maximal diminue avec T.

utilisons une méthode de descente de gradient pour obtenir une approximation numérique de la solution optimale. Prenons quatre exemples : un exemple où l'horizon T est court (l'exécution est plus rapide que la vitesse d'adaptation du marché), un example à horizon intermédiaire et deux exemples à horizon long. Les paramètres choisis pour les 4 simulations sont les suivants :

a $J = 16$, $D = 32$, $Q = 4000$, $T = 1$,

b $J = 16$, $D = 32$, $Q = 4000$, $T = 10$,

c $J = 16$, $D = 32$, $Q = 4000$, $T = 100$,

d $J = 16$, $D = 32$, $Q = 4000$, $T = 200$.

Les profils d'exécution optimaux pour chacune de ces stratégies sont présentés en Figure B.1, ainsi que les trajectoires impact correspondantes. On remarques les propriétés suivantes :

1. Lorsque l'horizon est suffisamment long (taux d'exécution faible), le profil optimal de volume se rapproche bien du profil optimal théorique obtenu pour le modèle de propagateur associé, avec une exécution plus agressive au début et à la fin de la période d'exécution (confirmant les équations 6.10 et B.2). Cela provient naturellement du phénomène de décroissance de l'impact : en concentrant l'exécution au début et à la fin, la stratégie permet de maximiser (sous contrainte) le temps entre les trades et ainsi profite au maximum de la décroissance de l'impact.

2. Notez également que le profil de volume est symétrique par renversement du temps lorsque

T est grand (ce qui est attendu pour un propagateur, le coût s'écrivant $\frac{1}{2}\int_0^T\int_0^T G(\mid t-s\mid)\mathrm{d}m_s\mathrm{d}m_s$ qui est invariant par $t \to T-t$).

3. Lorsque l'horizon se raccourcit, la stratégie optimale se concentre plus fortement sur les bords, et tend vers une stratégie « bucket-shaped » qui rappelle les stratégies optimales obtenues dans Predoiu et al. (2011). Toutefois, à la différence de ces modèles, le profil de volume optimal est *back-loaded*, c'est-à-dire plus agressif à la fin de l'exécution. Il s'agit là d'une propriété du modèle plutôt inattendue – et sans doute d'un signe que la contrainte T fixé n'est pas des plus pertinentes pour l'exécution : elle force une singularité à la fin de l'horizon, dont on se passerait volontiers.

4. Les profil de prix sont également intéressants : ils semblent quant à eux plus symétriques que les profils de volume, et exhibent une forme en S, d'autant plus prononcée que l'horizon est court – le prix devient alors constant entre le bucket de début et le bucket de fin.

Analyse des gains

Quel crédit accorder aux stratégies ainsi obtenues – c'est-à-dire, au-delà de leur optimalité, quelles économies permettent-elles réellement de réaliser par rapport à une exécution simple ? La Figure B.2 présente les coûts associés aux stratégies optimales pour une gamme d'horizons d'exécution, et les compare aux coûts associés aux stratégies à taux constants. Si à T grands suivre une politique optimale peut permettre d'économiser jusqu'à 2%, lorsque T devient court les gains ainsi obtenus deviennent dérisoires, en accord avec les prédictions de l'équation B.8.

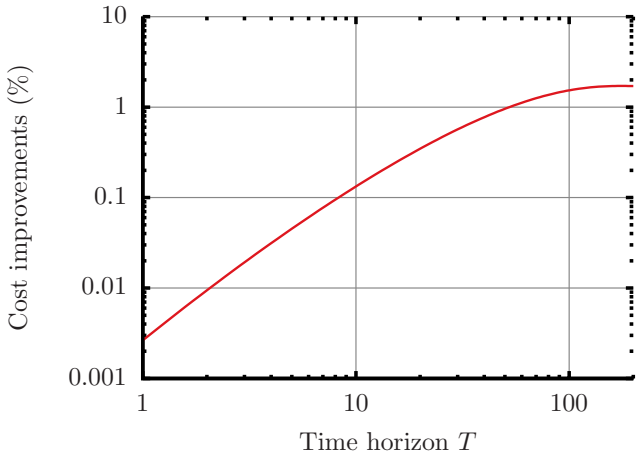

FIGURE B.2 – Gains réalisés par rapport à une exécution à taux constant (%). Comme prédit dans la section B.1.2, les gains sont négligeables pour les exécutions rapides. Ils atteignent un maximum d'environ 2% à T grand, ce qui correspond aux gains attendus dans le cas du propagateur linéaire en $1/\sqrt{t}$.

Effet du tick

Puisque que nous savons simuler le problème et en extraire des solutions « optimales », nous pouvons nous intéresser à des problèmes analytiquement plus délicats mais numériquement tout autant intéressants – si ce n'est plus. Cette section se consacrera à la liquidation optimale dans le cas d'un asset à gros *tick*.[2] Cela se simule en fait très naturellement : il suffit de choisir un pas de discrétisation spatial suffisamment gros dans la simulation de l'EDP. La Figure B.3 présente le profil d'exécution pour une valeur de $T = 100$.

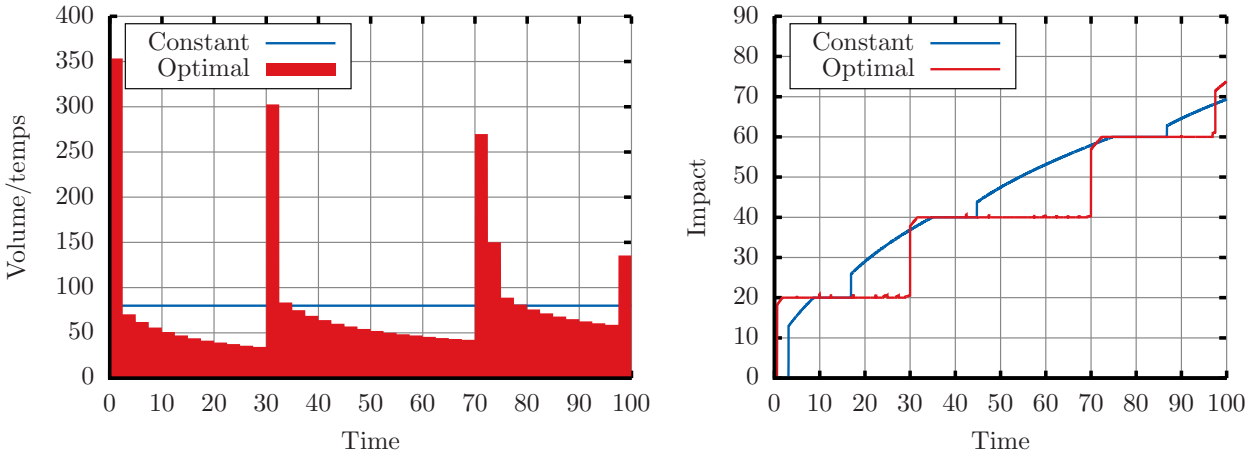

FIGURE B.3 – Profil d'exécution optimal dans le cas d'un gros *tick* et trajectoire de prix associée. L'algorithme fait en sorte d'exécuter les volumes niveau de prix après niveau de prix, ce qui génère un pattern dans le profil de volume.

Les résultats sont très surprenants lorsque l'on regarde le profil de volume : on voit apparaître des pics d'agressivité quasi-périodiques qui relaxent ensuite avant l'arrivée du pic suivant. Le profil d'impact (i.e. de prix) est en revanche plus explicite : on voit que l'exécution se réalise par paliers de prix. L'algorithme fait donc en sorte de ne jamais se trouver entre deux niveaux de prix : lorsqu'il arrive sur un niveau de prix, il le consomme en totalité, reste à ce niveau pendant encore quelques pas de temps en exécutant les volumes qui y réapparaissent, et lorsque plus suffisamment de volume de revient passe au niveau suivant. Si l'on compare cette fois les coûts obtenus à ceux associés à une exécution à taux constante, on obtient une amélioration plus conséquente de 5.8% pour l'exemple $T = 100$ (contre 1.8% précédemment).

La Figure B.4 représente les gains associés à une exécution optimale, en fonction de la taille du tick. On voit nettement que plus le tick augmente, plus les gains sont potentiellement élevés. Cela suggère la conclusion suivante : en liquidation optimale, l'essentiel de l'optimisation se trouve dans la bonne exécution au niveau du tick/spread et dans le choix de l'horizon d'exécution plus que dans le *scheduling* précis : il est alors au moins aussi important d'avoir un bon modèle de

2. Rappelons que le tick est l'incrément de prix minimum entre deux prix de transactions autorisés.

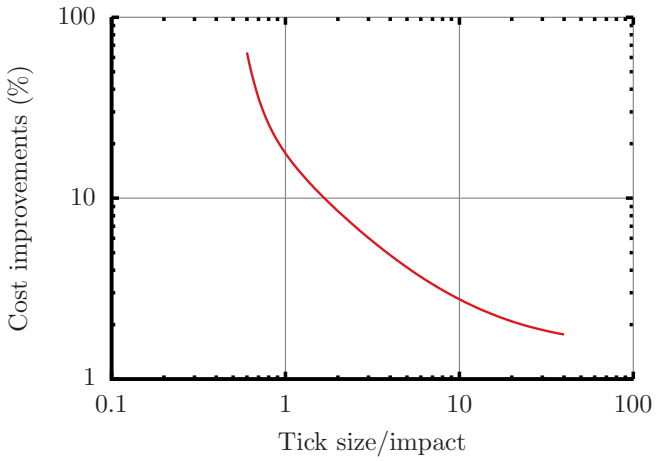

FIGURE B.4 – Comparaison entre les coûts associés à la stratégie optimale et ceux associés à une exécution à taux constant, pour plusieurs tailles de tick, en pourcentage d'économies réalisées. Celui-ci est d'autant plus grand que le tick est grand par rapport à l'échelle de l'impact.

carnet d'ordres et de risque à ces échelles, qu'un bon modèle d'impact ! Nous verrons dans la section suivante l'intérêt des modèles d'impact lorsque le volume à exécuter n'est pas pré-déterminé (ni l'horizon d'exécution).

B.2.2 Exploitation d'*alpha*

Tournons-nous maintenant vers le problème d'exploitation optimale d'une prédiction sur l'évolution sur les prix futurs. Souvenons-nous que dans le modèle la composante « prix fondamental » et la composante « liquidité » s'additionnent. Le problème que nous posons est donc le suivant : sachant que l'on possède une information sur l'évolution future de la composante fondamental, quel stratégie mettre en place pour réaliser les meilleurs profits (cette stratégie affectant la composante liquidité uniquement) ? [3] La Figure B.5 représente une estimation numérique de la stratégie d'exploitation optimale du signal :

$$\alpha(t) = 30(1 - e^{-0.1t}), \tag{B.9}$$

dans le cas où l'inventaire final est évalué au *mark-to-market* final, c'est-à-dire $\alpha(t \to \infty)$ (donc sans contrainte de revente). Naturellement, celle-ci consiste essentiellement à acheter avant que l'*alpha* ne se réalise. Mais en plus de nous renseigner sur la capacité de la stratégie en terme de volume, l'expérience permet de tirer quelques leçons moins évidentes :

1. L'exploitation n'est pas instantanée, signe d'une compétition entre l'alpha qui se réalise (incitation à accélérer) et la décroissance de l'impact (incitation à ralentir).
2. L'algorithme exploite son propre decay d'impact et revend lorsque le prix dépasse la valeur

3. Il y a de nombreuses sortes de signaux sur les prix, et je ne prétends pas que cette forme précise réaliste. Elle permet seulement de poser un problème simple.

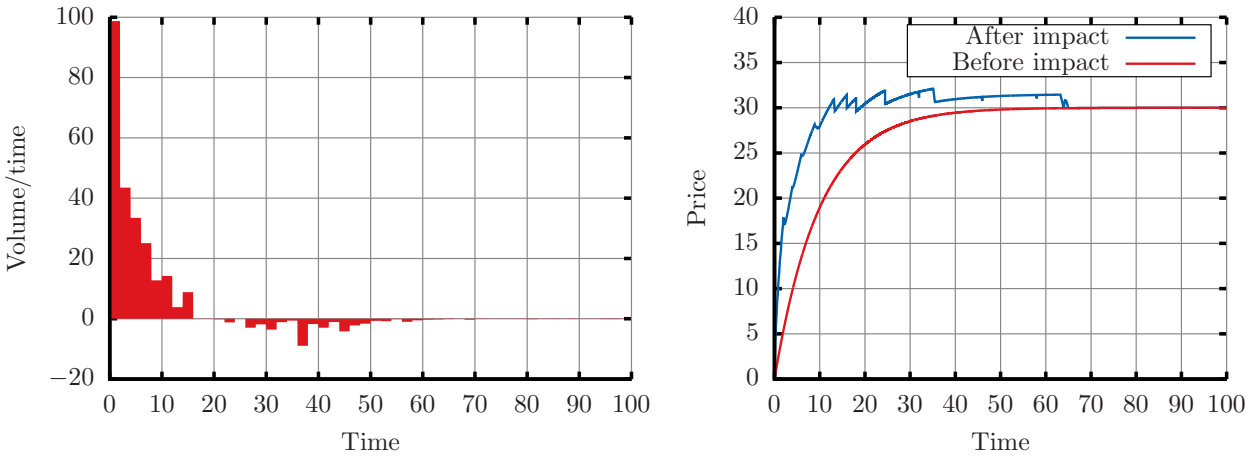

FIGURE B.5 – Stratégie optimale d'exploitation de l'*alpha* exponentiel $\alpha(t) = 30(1 - e^{-0.1t})$, sans contrainte d'inventaire final nul. (gauche) Profil de volume. Les volumes positifs sont à l'achat et les volumes négatifs sont à la vente. (droite) Prix associé. Les dents de scie qui apparaissent sont dues à des effets de discrétisation.

maximale de l'*alpha* ! Cela est bien entendu lié à la structure d'*alpha* choisie au départ – cela n'est reste pas moins intéressant.

3. Le prix « overshoot » la prédiction maximale à cause de l'impact – avant d'y revenir grâce au decay. Cet étrange phénomène est étroitement lié avec le point précédent.

4. L'arbitrage d'un signal ne le fait donc pas disparaître immédiatement, mais accélère sa convergence vers la valeur prédite.

Pricer l'inventaire au au *mark-to-market* final peut toutefois sembler étrange, dons un contexte où l'impact est justement primordial ! La Figure B.6 présente la stratégie optimale d'exploitation de l'*alpha* de l'équation B.9, sous contrainte d'inventaire final nul : l'algorithme doit maintenant prendre en compte l'impact des trades qui lui permettent de déboucler sa position. La position prise est donc naturellement plus faible que dans le cas précédent, et les gains sont diminués de 46%. On remarque pour finir que :

1. La revente de la position se fait essentiellement à prix constant, avec une revente intense dès que l'alpha commence à disparaître suivie par une revente plus lente.

2. L'effet de bord à la fin de la période du à la décroissance de l'impact (« bucket de fin »), que l'on avait déjà observé dans le cas de la liquidation, apparaît à nouveau.

Terminons par un dernier exemple, identique à celui de la Figure B.6 mais dans lequel la vitesse de relaxation du carnet d'ordres est diminuée d'un facteur 100. La stratégie optimale obtenue dans ce cas est représentée Figure B.7. Dans ce cas, la capacité se trouve grandement réduite à cause de la difficulté à retourner la direction de trading, et les profits sont diminués drastiquement de 77%. Cet

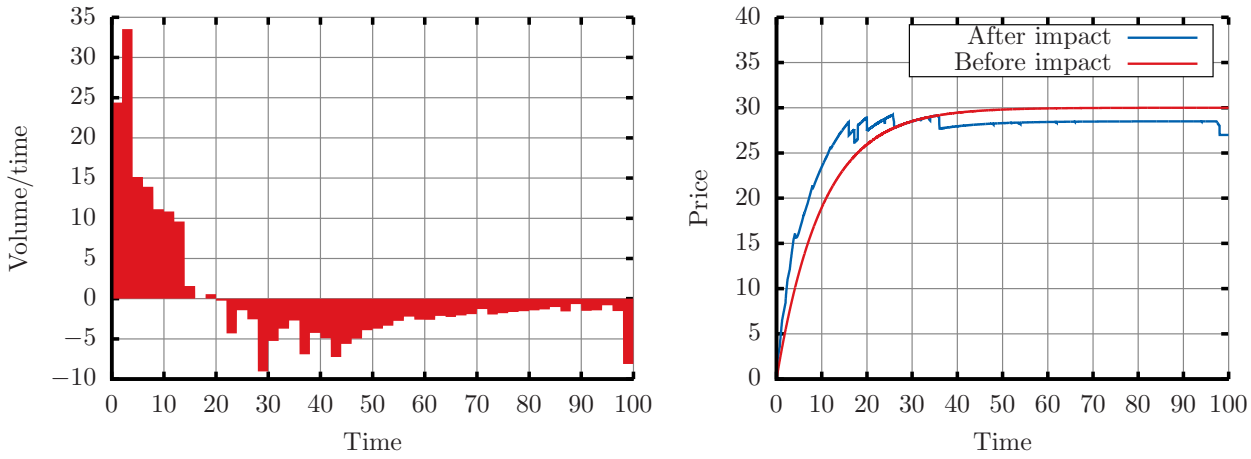

FIGURE B.6 – Stratégie optimale d'exploitation de l'*alpha* exponentiel $\alpha(t) = 30(1 - e^{-0.1t})$, avec contrainte d'inventaire final nul. (gauche) Profil de volume. Les volumes positifs sont à l'achat et les volumes négatifs sont à la vente. (droite) Prix associé.

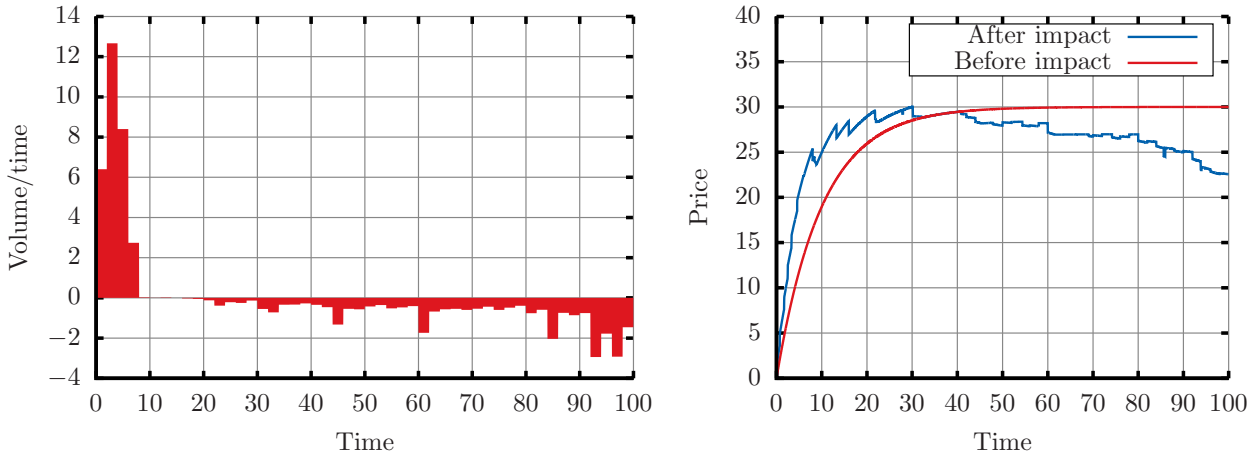

FIGURE B.7 – Stratégie optimale d'exploitation de l'*alpha* exponentiel $\alpha(t) = 30(1 - e^{-0.1t})$, avec contrainte d'inventaire final nul, pour une relaxation du carnet d'ordres 100 fois plus lente que pour la Figure B.6. Les profits sont réduits de 77%.

aspect est entièrement absent des modèles à propagateur, et souligne l'importance de comprendre, non seulement la liquidité des marchés, mais aussi sa vélocité.

Il est probable que tous les effets qualitativement mis en avant dans cette section puissent l'être également dans un modèle simple de propagateur. Mais les intuitions développées, tant sur la réalité que sur le modèle, me semblent suffisamment importantes pour mériter leur place ici.

B.3 Conclusion

Si je devait conclure cette section, ce serait essentiellement pour discuter de l'intérêt des modèles d'impact en trading optimal. La Section B.2.1, et en particulier la Figure B.2, nous montre bien que pour le problème de liquidation à volume et horizon de temps fixés, le coût obtenu dépend peu du modèle : l'économie maximale réalisée par un bon *scheduling* est minime (au maximum $1-2\%$ pour les exécutions lentes). Le choix de l'horizon quant à lui est plus déterminant, comme suggéré par les résultats asymptotiques.

Dans le cadre de la liquidation, la modélisation du carnet d'ordre à l'échelle du tick/spread devient alors primordiale : lorsque les effets de discrétisation se manifestent à l'échelle de l'impact, les économies réalisables deviennent potentiellement conséquentes. Des tâches similaires et non décrites ici, comme le routage à travers les différentes plateformes d'échanges, seront sans doute toutes aussi cruciales pour une exécution performante. [4]

En revanche, dès lors que la tâche est de déterminer les volumes, avoir un bon modèle d'impact est crucial, à la fois pour déterminer la capacité de la stratégie en terme de volume, et pour adopter un profil de volume optimal en fonction du profil d'*alpha* en prenant en compte entre autres l'impact des trades passés sur le carnet d'ordres. En particulier, bien comprendre la relaxation du carnet d'ordres et du prix est crucial dans le dimensionnement et le *scheduling* de la stratégie. Ici, le modèle d'impact peut facilement faire la différence entre une stratégie non profitable et une stratégie profitable.

4. Cela semble dire que les modèles d'impact importent peu pour tout ce qui est liquidation. Mais il me semble que l'intérêt d'un modèle est aussi de déterminer dans quels cas celui-ci a un intérêt réel – et dans quels cas il n'en a pas.

Annexe C

Etude empirique : prédiction sur le pattern de volume après l'enchère du matin

En donnant une dynamique continue à l'offre et la demande, le modèle des chapitres 6 et 7 permet d'effectuer des prédictions dans des situations moins standard que le régime stationnaire en trading continu. Cet appendice s'intéresse au pattern de volume après l'enchère du matin sur 10 des principaux stocks du NASDAQ, où le trading a lieu essentiellement entre 9h30 et 16h – et est arrêté la nuit. Les données semblent confirmer les prédictions du modèle d'un pattern de volume en $1/\sqrt{t-t_0}$ où t_0 est l'heure d'ouverture du marché.

C.1 Model predictions

Based on the above theory for the dynamics of supply and demand, we can study the realistic case where continuous trading takes place during some opening hours, with rather long periods of interruption (typically overnights, week-ends, lunch pauses, etc.). In order to avoid large instabilities at the start of the trading day, market venues usually open the day with a morning auction. This is in fact naturally explained within our framework : the free evolution of the supply and demand leads to overlapping MSD curves, thus the need for an auction. This is an interesting test case for our theory, since the latent order book should evolve from the stationary state reached at the close of the previous day, and characterized by the MSD curves $\rho_{S,\mathrm{st.}}(y)$ and $\rho_{D,\mathrm{st.}}(y)$ computed above, until the next morning. We assume that this evolution is characterized by Eqs. (6.1) during an effective time T, that is not necessarily the physical duration of the overnight. It is known that the volatility corresponding to overnights amounts to roughly $T = 2$ hours of intraday trading. At the end of this period (i.e., just before the morning auction), the MSD will have evolved to a new shape

$\rho_{S,D}(y, t = T)$ that has the typical form shown in Figures 7.1, 7.7, left. Assuming for simplicity that the MSD is symmetrical, the auction will clear the market at $y^* = 0$ with a certain volume v_T^*, and again abruptly truncates to zero $\rho_D(y > 0, t = T)$ and $\rho_S(y < 0, t = T)$. Then, continuous time trading resumes. Clearly, the shape of the MSD will not be immediately given by Eq. (7.9), because the excess volume present in $\rho_{S,D}(y, t = T^+)$ needs to be absorbed by the market. The aim of this section is to compute this excess executed volume (per unit time) as a function of time and compare this prediction with available data on stock markets.

Since we now directly place ourselves in the context of continuous time auction, any order that reaches $y = 0$ is immediately executed and disappears from the latent order book. Therefore, the MSD curves are given by :

$$\begin{aligned}\rho_S(y, t > T) &= \int_0^{+\infty} \frac{\mathrm{d}y'}{\sqrt{4\pi\mathcal{D}(t-T)}}\, \rho_S(y', T^+) e^{-\nu(t-T)} \left[e^{-\frac{(y'-y)^2}{4\mathcal{D}(t-T)}} - e^{-\frac{(y'+y)^2}{4\mathcal{D}(t-T)}} \right] \\ &+ \int_T^t \mathrm{d}t' \int_{-\infty}^{+\infty} \frac{\mathrm{d}y'}{\sqrt{4\pi\mathcal{D}(t-t')}}\, \omega(y') e^{-\nu(t-t')} \left[e^{-\frac{(y'-y)^2}{4\mathcal{D}(t-t')}} - e^{-\frac{(y'+y)^2}{4\mathcal{D}(t-t')}} \right], \quad \text{(C.1)}\end{aligned}$$

and symmetrically for $\rho_D(y, t > T)$; we have assumed that the flux of new orders is symmetrical to keep the auction price centred around the fundamental price (i.e. $y_t^* \equiv 0$). To make the mathematical analysis tractable, we assume that $\mathcal{D}(t - T)$ is small enough so that we can forget the price dependence of $\rho_S(y', T^+)$ and $\omega(y')$ that we set, respectively, to $\rho_S(0, T^+)$ and $\omega(0)$. Eqs. (6.1) also allows us to compute the number $J(t)\mathrm{d}t$ of sell/buy orders that "cross" $y = 0$ and get executed between t and $t + \mathrm{d}t$, from :

$$J(t) = \mathcal{D}\, \partial_y \rho_S(y, t > T)|_{y=0} = -\mathcal{D}\, \partial_y \rho_D(y, t > T)|_{y=0}\,. \quad \text{(C.2)}$$

Using Eq. (C.1) and for an initial affine MSD given by Eq.(7.9), one then obtains :

$$J(t) = \sqrt{\mathcal{D}}\rho_S(0, T^+) \frac{e^{-\nu(t-T)}}{\sqrt{\pi(t-T)}} + J_{\text{st.}}\mathrm{Erf}(\nu(t-T)), \quad \text{(C.3)}$$

where $J_{\text{st.}} = \omega(0)\sqrt{\mathcal{D}/\nu} = J(t \to \infty)$ is the stationary volume executed per unit time (see Donier et al. (2015)), long after the morning auction. In the short time limit $(t-T) \to 0$, the above equation leads to a very simple prediction :

$$J(t \to T^+) \approx \frac{\sqrt{\mathcal{D}}\rho_S(0, T^+)}{\sqrt{\pi(t-T)}} + O(\sqrt{t-T}), \quad \text{(C.4)}$$

i.e. a square-root decay of the excess volume traded per unit time just during the morning session, after the opening auction.

C.2 Empirical results

The average intraday volume pattern is shown in Figure C.1, and compared to the $1/\sqrt{t-T}$ prediction, which is surprisingly well vindicated. Note that the afternoon session is characterized by an increased volume as the close gets nearer, that one may attribute to agents willing to finish their trades or close their inventories to avoid the overnight volatility, but this regime is clearly beyond the scope of the present framework. [1]

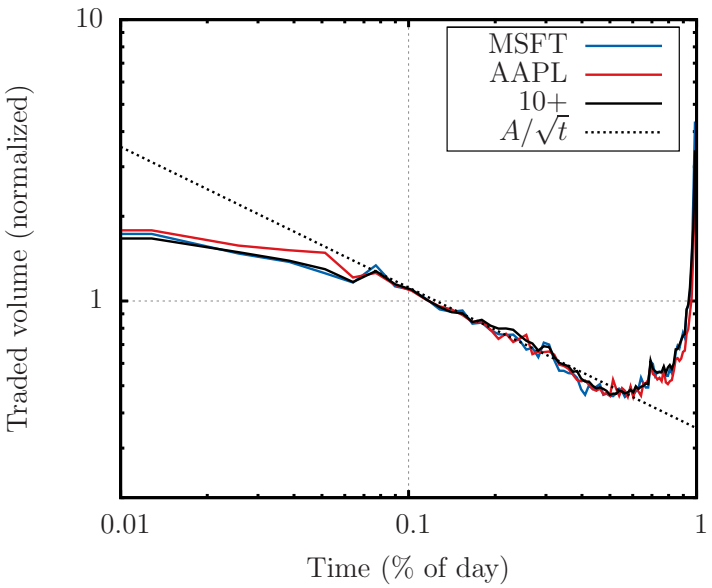

FIGURE C.1 – The average intraday volume pattern for Apple, Google, Amazon, Facebook, Visa, IBM, Bank of America, McDonald's, Citigroup, Microsoft in the period Jan 2014 → April 2015 shows good agreement with the theory, that predicts an inverse square root decay of volume after the morning auction before reaching its stationary value. The overshoot of our prediction very shortly after the opening of the market comes from discretization effects (the volume is measured here as an average over 5 minutes bin) and possibly from non-diffusive effects in early trading. The increase of volume when one gets closer to the evening auction is due to a different phenomenon that is outside of the scope of the present theory. We show the individual data for Apple and Microsoft, and the average over the 10 stocks.

C.3 Conclusion

Although we focus only on the volume pattern here (and for instance not on the volatility pattern which also exists), we find that the model predictions of an inverse square root decay is rather well vindicated. We find a deviation right after the auction, maybe sign that orders are not

1. In the terms of the model, it might be interpreted as an increase of the deposition rate ω and/or of the diffusion speed $\mathcal{D}$ that in turn lead to a larger $J_{\text{st.}}$.

perfectly diffusive – but maybe subject to different motions, like the grey brownian motion evoked in Appendix A. None of this of course is a proof that the model is correct : this only suggests that it might correctly capture some part of market dynamics – although it has to be worked out with non constant parameters now.

Annexe D

Dissection de l'hypothèse de *trading invariance*

From
Unravelling the trading invariance hypothesis
with Michael Benzaquen and Jean-Philippe Bouchaud
(Benzaquen et al., 2016)

Les deux articles qui suivent ne font pas directement partie de l'histoire que j'essaye de raconter dans cette thèse, mais méritent tout de même une place dans ce manuscrit, si ce n'est pour leurs résultats, au moins pour l'effort qu'ils ont représenté pour chacun des auteurs. J'en profite ici pour remercier Michael Benzaquen (l'auteur principal de l'article qui suit), qui bien que novice sur ces sujets, a su mener à bien avec persévérance et brio cette étude à la recherche d'une « loi des marchés parfaits ».

Abstract : We confirm and substantially extend the recent empirical result of Andersen et al. (2015), where it is shown that the amount of risk W exchanged in the E-mini S&P futures market (i.e. price times volume times volatility) scales like the $3/2$ power of the number of trades N. We show that this $3/2$-law holds very precisely across 12 futures contracts and 300 single US stocks, and across a wide range of times scales. However, we find that the "trading invariant" $I = W/N^{3/2}$ proposed by Kyle and Obizhaeva (2010) is in fact quite different for different contracts, in particular between futures and single stocks. Our analysis suggests I/S as a more natural candidate, where S is the bid-ask spread. We also establish two more complex scaling laws for the volatility σ and the traded volume V as a function of N, that reveal the existence of a characteristic number of trades N_0 above which the expected behaviour $\sigma \sim \sqrt{N}$ and $V \sim N$ hold, but below which strong deviations appear, induced by the size of the tick.

D.1 Introduction

Understanding the dynamics of financial markets is of obvious importance for the financial industry, but also for decision makers, central bankers and regulators. It is also a formidable intellectual challenge that has attracted the interest of many academic luminaries, with perhaps Benoit Mandelbrot as a legendary figure. He was the first to propose the idea of *scaling* in this context Mandelbrot (1997), a concept that in fact blossomed in statistical physics before getting acceptance in economics and finance (for a review, see Gabaix (2009)). In the last twenty years, many interesting scaling laws have been reported, concerning different aspects of price and volatility dynamics. One particular question that has been the focus of many studies is the relation between volatility and trading activity, measured as the number of trades and/or the volume traded (see e.g. Clark (1973); Tauchen and Pitts (1983); Jones et al. (1994); Bollerslev and Jubinski (1999); Ané and Geman (2000); Liesenfeld (2001); Tauchen and Pitts (1983); Engle (2000) and more recently Zumbach (2004); Eisler and Kertész (2006); Wyart et al. (2008)). Revisiting these results, Kyle and Obizhaeva (2010) (KO) recently proposed a bold but inspiring hypothesis, coined as the *trading invariance principle*.

Their original idea primarily relies on *dimensional analysis*, which is very common in physics and states that any "law" relating different observables must express one particular dimensionless (or unit-less) combination of these observables as a function of one or several other such dimensionless combinations. The simplest example might be the ideal gas law, that amounts to realizing that pressure p times volume v has the dimension of an energy. Hence pv must be divided by the thermal energy RT of a mole of gas to yield a dimensionless combination. The right-hand side of the equation must be a function of other dimensionless variables, but in the case of non-interacting point-like particles, there is none – hence the only possibility is $pv/RT = \text{cst}$. Deviations from the ideal gas law are only possible because of the finite radius of the molecules, or the strength of their interaction energy, that allows one to create other dimensionless combinations (and correspondingly new interesting phenomena such as the liquid-gas transition !).

In the search of an "ideal market law" for stocks, several possible observable quantities that characterize the trading activity come to mind : the total market capitalisation M (in dollars), the share price P (in dollars per share), the square volatility σ^2 (in $\%^2$ per day), the amount traded V (shares per day), and the volume of individual "bets" Q (in shares) [1]. Other, more microstructural quantities might come into play, such as the difference between the best bid and best offer price, called the spread S (in dollars), the tick size s (in dollars) that fixes the smallest possible price change, the lot size ℓ (in shares) that fixes the smallest amount of exchanged shares, the average volume available at the best quotes, and perhaps other quantities as well.

Kyle and Obizhaeva further postulate the existence of a universal invariant I in dollars, that they

1. A bet is the ensemble of trades that originate from a single trading decision. It is alternatively called a "metaorder" in the literature.

interpret as the average "cost" of a single bet, and keep only P, σ^2, V and Q as relevant variables. Dimensional analysis then immediately leads to the following relation :

$$\frac{PQ}{I} = f\left(\frac{Q\sigma^2}{V}\right) \tag{D.1}$$

where f is a certain function that cannot be determined on the basis of dimensional analysis only. At this point, Kyle invokes the Modigliani-Miller theorem and argues that capital restructuring between debt and equity should keep $P \times \sigma$ constant, while not affecting the other variables. This suggests that $f(x) \sim x^{-1/2}$, finally leading to the KO *trading invariance principle* :[2]

$$I = \frac{P\sigma Q^{3/2}}{V^{1/2}} := \frac{W}{N^{3/2}} \ , \tag{D.2}$$

where $W := PV\sigma$ is a measure of exchanged risk (precisely the dollar amount of risk traded per day), also referred to as *trading activity* by Andersen et al. (2015), and $N := V/Q$ represents the number of bets per day.

This simple scaling relation was empirically confirmed by KO using portfolio transition data Kyle and Obizhaeva (2010). Portfolio transitions correspond to rebalancing decisions by institutional investors, that are then executed by brokers who collated the corresponding data. However, these trades only reflects part of the market activity, and it is furthermore not obvious that these portfolio transitions can be associated with elementary bets. Andersen et al. (2015) reformulated KO's invariance principle in a way that can be tested on public trade-by-trade data. The idea is simply to interpret single bets as single *trades* and therefore define Q as the average volume of trades and N as the total number of trades within some time interval τ. Using trade-by-trade data on the E-mini S&P 500 futures contract, Andersen et al. (2015) showed that Eq. (D.2) holds remarkably well for $\tau = 1$ minute time intervals. Because the activity of the market has significant intraday variability, notably marked by the switching from Asian to European and American trading hours, N typically varies over almost two decades, indeed allowing one to test the scaling relation $W \sim N^{3/2}$ quite convincingly (see Fig. 1 below).

Such a remarkable empirical result, and its purported universal status, clearly cries for further scrutiny and interpretation. The goal of this paper is to dissect the trading invariance hypothesis on a wide range of futures contracts and individual stocks. Eq. (D.2) can actually be interpreted in different ways, depending on the degree of universality attached to its validity :

1. *No universality* : The scaling relation $W \sim N^{3/2}$ (the "3/2-law" henceforth) holds for some contracts and some time intervals τ (over which W and N are computed). In the cases where the scaling law holds, the prefactor I has a non-universal value (that depends on the contract and/or on τ).

2. Up to a redefinition of I, one can always set $f(x) = x^{-1/2}$ without any numerical prefactor.

2. *Weak universality* : The 3/2-law holds for all contracts and some (possibly all) time intervals τ, but with a non-universal value of I.
3. *Strong universality* : The 3/2-law holds for all contracts and all time intervals τ, with a universal value of I, independent of τ and of the contract type.

The last case might in fact be too strong : it would already be a remarkable result that I only depends on the contract type (say stocks) and on the geographical zone (say the U.S.). In fact, from general considerations it would be very strange that I (in dollars) is completely universal, for one thing because the value of the dollar itself is time dependent. As we will show in detail below, our results favor the second interpretation of "weak universality" where the 3/2-law holds for all contracts, and all time intervals τ. However, the value of I itself varies significantly, both within the universe of US stocks and among the different futures contracts. Furthermore, the separate analysis of the scaling of σ vs. N on the one hand and V vs. N on the other (the product of the two essentially leading to the 3/2-law) reveals a surprisingly rich and universal behaviour, and suggests that $W \sim N^{3/2}$ might only be an approximation.

The outline of the paper is as follows. In section 1, we replicate and confirm Andersen *et al.*'s results on E-mini S&P 500 futures contract and extend them to eleven other futures contracts. We show that the 3/2-law does hold both across time and across contracts, but that the average value of I (and the whole distribution of I, for that matter) clearly depends on the considered contract. In section 2, we confirm the 3/2-law across a pool of 300 US stocks and show that microstructure effects play a much more important role than in the case of futures contracts. In section 3 we propose a unifying picture that decomposes the 3/2-law into two more fundamental scaling laws, that allow us to rescale all futures contracts and all time scales onto two universal master curves. Similar to the deviations away from the ideal gas law example alluded to above, our results suggest that additional microstructural variables must be involved in the search of a relation generalizing Eq. (D.1), where the bid-ask spread and the tick size, among other things, should play an important role – like the molecular size in the ideal gas analogy. In section 4, we suggest an alternative and more natural definition for trading invariant that accounts some of microstructural details mentioned above.

D.2 Futures contracts

We have analysed tick by tick data for the best bid and offer of twelve different futures contracts spanning over three years, from January 2012 to December 2014 (see Tab. D.1). We consider front month contracts only, among which three index futures, four energy futures, two agriculture futures, one bond future, one FX future and one metals future. All contracts are traded basically twenty-four hours a day, five days a week, on the CME, NYBOT, NYMEX, ECBOT, COMEX, ICUS and IPE electronic platforms. Three trading regimes can be distinguished corresponding respectively to Asian, European and American regular trading hours. At variance with the analysis of Andersen

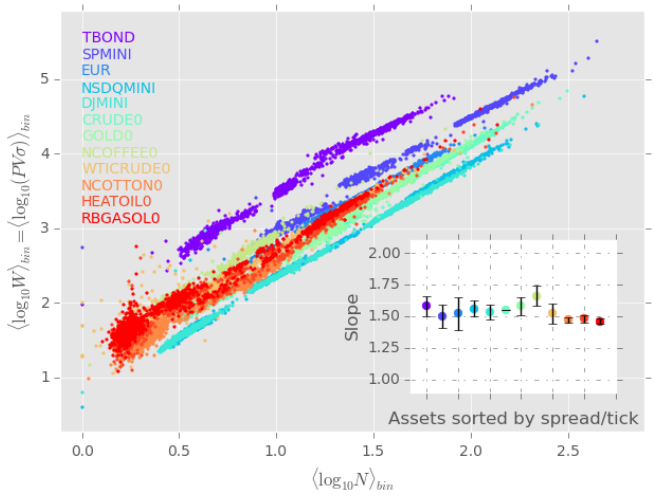

FIGURE D.1 – Scatter plot of $\langle \log_{10} W \rangle_{\text{bin}}$ *vs.* $\langle \log_{10} N \rangle_{\text{bin}}$ for twelve different futures contract sorted by spread over tick values from cold (large ticks) to warm colours (small ticks). The inset shows the slopes α obtained from linear regression of the data, which are all clustered around 3/2. Spread over tick values as well as the slopes α are provided in Tab. D.1.

et al. (2015), we do not discard any time intervals from our study since we found that doing so did not significantly change the results.

For each contract, we group the trades by market time stamp, under the assumption that simultaneous trades correspond to a market order originating from a single participant. We then compute trading volume V, number of trades N, average trade size $Q = V/N$ and average price P within each one minute bin ($\tau = 1$ min). We also compute the volatility σ, from the average of ten second squared-returns. At variance with Andersen et al. (2015), we do not annualize our volatilities. Average values of these quantities for $\tau = 1$ min, as well as average volume at the bid and the ask and average spread, are provided in Tab. D.1. Note that throughout the paper we will elicit power-laws by considering linear regression of log quantities (for example $\log W$ vs. $\log N$). Consistent with this procedure, averages shall be defined with respect to the log-transform, and we will write $\langle X \rangle := \exp\left[\mathbb{E}(\log X)\right]$.

Following the method of Andersen to test the intraday trading invariance hypothesis, we first average over all days of the logarithm of the aforementioned quantities, for each fixed one minute bin. The latter averaging operator shall be noted $\langle\, . \,\rangle_{\text{bin}}$. Note that taking the logarithm prior to averaging dampens the influence of outliers and leads to a robust estimate of the "typical value" of these quantities. The linear regression of $\langle \log W \rangle_{\text{bin}}$ *vs.* $\langle \log N \rangle_{\text{bin}}$ is displayed in Fig. D.1 and Tab. D.1, and indeed confirms the 3/2-law for all twelve contracts independently. However, the conjecture that the quantity $I = WN^{-3/2}$ – which visually corresponds to the y-intercept of the linear regressions shown in Fig. D.1 – is invariant across different contracts is clearly rejected (see Tab. D.1). The top right inset of Fig. D.2 displays $\langle I \rangle$ for the twelve futures contracts sorted by spread over tick.

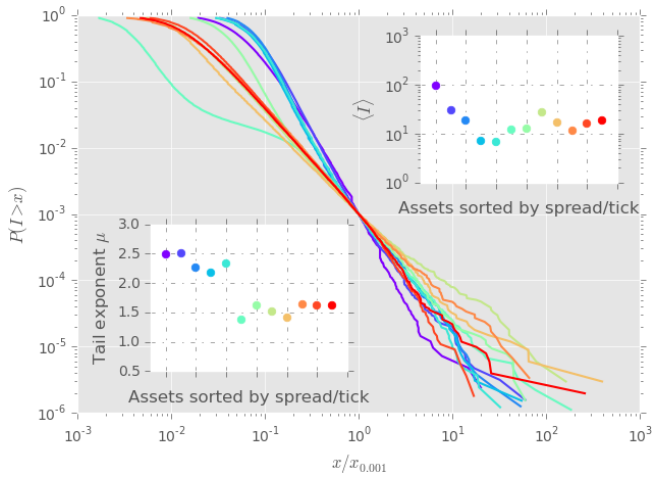

FIGURE D.2 – Rescaled complementary cumulative distribution function of $I = WN^{-3/2}$ for twelve different futures contracts sorted by spread over tick values from cold (large ticks) to warm colours (small ticks). The insets show the average values of I (in dollars) and the tail exponents computed according to the Hill estimator with a cutoff at $P(I > x) = 10^{-2}$. Spread over tick values as well as the average values and tail exponents are provided in Tab. D.1.

For robustness, we checked that the above results also stand on sub-intervals of one year of the full period 2012-2014. In particular we observe that the variations of $\langle I \rangle$ across contracts (more than a factor 10) are much larger than the variations from one year to the next for a given contract (around $\sim 20\%$). The role of the bin size τ is also very interesting. Averaging over one, five and ten minute bins across days shows consistent results. The analysis on longer time scales (thirty minute, one and two hour bins) however shows a slight but systematic underestimation of the predicted 3/2

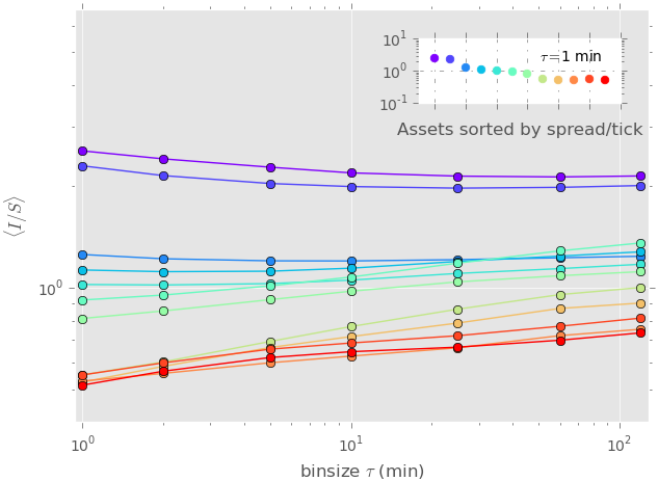

FIGURE D.3 – Plot of $\langle I/S \rangle$ where S denotes the spread (in dollars) as a function of bin size τ (in minutes) for twelve different futures contracts. Note that this ratio is nearly τ-independent. The inset, showing $\langle I/S \rangle$ at $\tau = 1$ min sorted by spread over tick values, should be compared to the top right inset of Fig. D.2.

slope of $\langle \log W \rangle_{\text{bin}}$ versus $\langle \log N \rangle_{\text{bin}}$ which disappears when the volatility estimator based on ten-second squared returns is replaced by the Rogers-Satchell volatility estimator Rogers and Satchell (1991), known to be more adequate when the underlying follows a geometric Brownian motion with an unknown drift. Note that the Rogers-Satchell estimator measures zero volatility whenever the open price matches the high/low *and* the close price matches the low/high, which are not rare events for small bin sizes. Enforcing that the volatility must be non-vanishing leads to discarding a substantial fraction of the data at high frequencies. However, we checked that removing the zero volatility intervals has no material impact on the results. In the following we shall thus consistently use the Rogers-Satchell estimator to compute the volatility. [3] The conclusion of our analysis is that the 3/2-law holds across all futures contracts and across all time intervals τ. Figure D.3 displays a plot of I rescaled by the spread S (in dollars), a choice that will be motivated in section 4. At this point one should note that a) I/S is now a dimensionless quantity of order unity, and b) I/S appears to be significantly more stable across assets than I itself, especially when $\tau = 2$ hours (see Fig. D.3).

Finally, the trading invariance hypothesis – at least in the initial KO formulation – states that the full probability distribution of $I = W/N^{3/2}$ (and not only its average value) should be invariant across time and across contracts. To test this point, we have computed the complementary cumulative distribution function $P(I > x)$ for the twelve futures contracts (see Fig. D.2). For the sake of readability, the main plot of Fig. D.2 displays these distributions with the x-axis rescaled by $x_{0.001}$ defined by $P(I > x_{0.001}) = 10^{-3}$. As one can see, the tail of the distributions are all close to power laws. The tail exponent μ – defined as $P(I > x) \sim x^{-\mu}$ – however varies significantly from $\mu \approx 2.5$ for the larger tick futures to $\mu \approx 1.5$ for the smaller tick futures. Tail exponents were computed using the Hill estimator with a cutoff at $P(I > x) = 10^{-2}$ [4]. The values of the tail exponents are provided in Tab. D.1.

The conclusions so far are thus :

1. We fully confirm the 3/2-law found by Andersen et al. (2015) on the E-mini S&P futures on one-minute intervals ;
2. The 3/2-law holds surprisingly accurately for all contracts and all time intervals ;
3. The "invariant" I is in fact not universal : both its average value and the shape of its distribution function depends quite significantly on the chosen contract. However, I for a given contract is to a good approximation τ-independent.

We now extend our analysis to a much wider sample of single stocks, and find that the above conclusions are indeed vindicated.

3. We have in fact checked that other estimators based on the open, high, low, close prices lead to very similar results.

4. The Hill estimator (1975) allows to compute the tail behaviour of a distribution. Defining the tail exponent μ as $P(X > x) \sim x^{-\mu}$, one has $\mu = \left[\frac{1}{k}\Sigma_{i=0}^{k-1} \log(X_i/X_k)\right]^{-1}$ where k denotes the rank of the cutoff.

D.3 US Stocks

Our analysis is conducted on a pool of three hundred US stocks, chosen to be as representative as possible in terms of market capitalisation and tick size. Note that the large number of assets – and their diversity – allows for great statistical significance. We consider five-minute bins using trades and quotes data from January 2012 to December 2012, extracted from the primary market of each stock (NYSE/NASDAQ). We remove auction time intervals, as well as thirty minutes after the opening and before the closing of the market, so as to avoid any artefact due to these specific trading periods. To compute the volatility, we again use the Rogers-Satchell estimator for which only the high, low, open and close prices are needed Rogers and Satchell (1991). The average values of N, Q, V, and σ are provided in Tab. D.2 for a random selection of twelve stocks within the pool.

As in the previous section, we perform a linear regression of $\log W$ versus $\log N$ for each stock. Fig. D.4 is analogous to Fig. D.1, only here we do not compute the average of the bins across days as was done in the previous section. This is due to the fact that, unlike futures, the stocks we consider are exclusively traded during American hours and thus lack the "three-continent" seasonality of the futures. Proceeding as suggested by Andersen *et al.* for futures would thus significantly reduce the range of possible values of V, Q, N and σ, thereby degrading the determination of the slopes of the fits. For the sake of readability, Fig. D.4 shows a centred rolling average along $\log N$, with window size of one hundred data points. However, all regressions were performed before the rolling average. For the three hundred stocks, a cross sectional determination of the slope yields $\alpha = 1.54 \pm 0.11$, where the uncertainty here is the root mean square cross-sectional dispersion. This is again in very

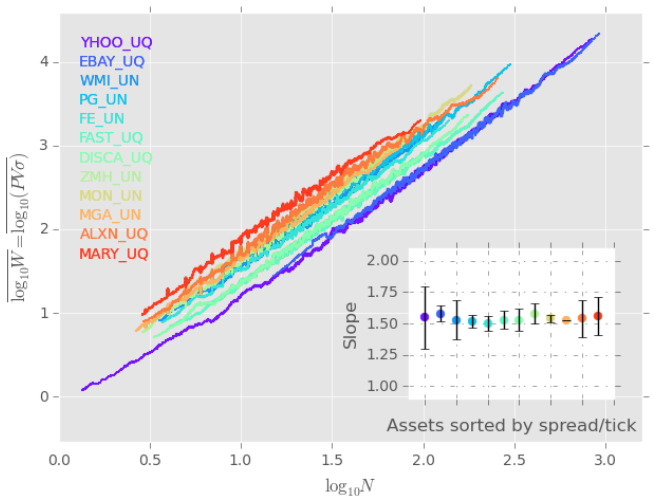

FIGURE D.4 – Centred rolling average (window size = 100) of the scatter plot of $\log_{10} W$ *vs.* $\log_{10} N$ for a random subset of twelve different stocks chosen from a pool of three hundred US stocks sorted by spread over tick values from cold (large ticks) to warm colours (small ticks). The inset shows the slopes obtained from linear regression of the data (before performing the rolling average). Spread over tick values as well as the slopes α obtained from the linear regressions are provided in Tab. D.2.

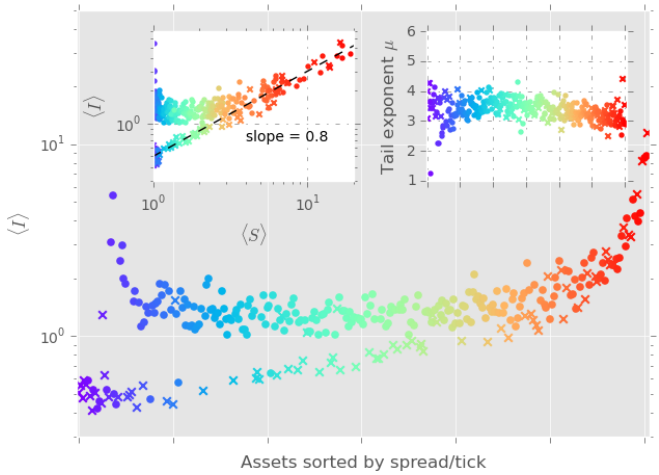

FIGURE D.5 – Main plot : average value of I (in dollars) for three hundred US stocks sorted by spread over tick values from cold (large ticks) to warm colours (small ticks). NASDAQ/NYSE stocks are marked with crosses/filled circles respectively. Top left inset : average value $\langle I \rangle$ as a function of average spread $\langle S \rangle$ (in ticks). Top right inset : tail exponents μ of the complementary cumulative probability distributions $P(I > x) \sim x^{-\mu}$. Numerical values of $\langle I \rangle$ as well as tail exponents are provided in Tab. D.2 for a random subset of twelve stocks.

good agreement with the prediction $\alpha = 3/2$, thereby considerably bolstering the results of the previous section. We also checked that these results hold unchanged for lower frequencies, and in particular at daily time scales $\tau = 6$ hours.

Figure D.5 displays the average values of I (computed using $\tau = 5$ min) as well as the tail exponents μ of the complementary cumulative probability distributions for three hundred stocks, sorted by spread over tick from left to right. After accounting for a factor $\sqrt{5}$ between futures ($\tau = 1$ min) and stocks ($\tau = 5$ min), the typical value of I for the stocks is on average one order of magnitude smaller than for futures – i.e. the "bet sizes" are smaller in dollars on individual stocks than on futures, which is not very surprising. The main plot reveals a striking feature : the appearance of two distinct branches in the larger ticks region. The higher branch presents an intriguing U-shape, similar to what was observed for futures contracts, while the lower branch is consistent with a nearly linear dependence on the average spread : $\langle I \rangle \sim \langle S \rangle^{0.8}$ (see top left inset, and the last section below for a quantitative interpretation). Remarkably, the two branches correspond chiefly to stocks traded on the NYSE (upper branch) and NASDAQ (lower branch) platforms. For better readability, NASDAQ stocks are represented by crosses while NYSE stocks appear as filled circles. Several points could actually explain this difference – although our understanding of this effect is only partial. For example a non-negligible fraction of the trades on NASDAQ happen within the spread (hidden trades), a particularity that would naturally affect the dynamics of large tick stocks and leave unaltered the small tick stocks. It is also know that fees/rebates are slightly higher on NASDAQ than on NYSE. We noticed that the main difference actually lies in the trade size, which appear to be on average smaller on NASDAQ than on NYSE for the large tick stocks. In

some sense, one could say that the large ticks on NASDAQ have a small tick behaviour, consistent with the possibility of having trades within the spread. As was the case for futures contracts, the distributions of I have power law tails, with tail exponents fluctuating around $\mu \approx 3.5$, but with no particular tick dependence. The values of μ for twelve randomly chosen stocks can be found in Tab. D.2.

D.4 Theoretical Analysis

We now turn to a theoretical analysis of the above results, with the aim of gaining a better understanding of the 3/2 scaling law, observed both on futures contracts and on single stocks. In most of this section, we redefine the trading activity as $\widetilde{W} = V\sigma$, without the price P which is irrelevant for the points we want to make. The role of the price will be discussed in the next section.

We first propose a very simple argument that suggests to decompose the N-dependence of the trading activity $\widetilde{W}$ into two parts, one coming from the N-dependence of σ and the other coming from the N-dependence of V. This decomposition reveals a much more subtle picture, where neither σ nor V behave as naively expected, but the product of the two indeed scales approximately as $N^{3/2}$. Most of this section is about futures, for which the story is surprisingly complex, whereas stocks behave more trivially and are discussed at the end.

D.4.1 A naive argument

If one assumes that there is a well-defined average trading frequency ϕ (defined as the number of trades per unit time) and a well-defined average trade size Q_0, then after time τ one expects the following two relations :

$$N = \phi \times \tau; \qquad V = Q_0 \times N. \tag{D.3}$$

Since the (log)-price is close to a random walk, one should also have :

$$\sigma = \varsigma_0 \sqrt{\tau}, \tag{D.4}$$

where ς_0 is a constant. Hence, eliminating τ,

$$\widetilde{W} = V\sigma = \frac{\varsigma_0 Q_0}{\sqrt{\phi}} N^{3/2}. \tag{D.5}$$

This appears to fully explain the 3/2 scaling law, which would then be an almost trivial observation. Although this will indeed turn out to be the correct mechanism for individual stocks, futures contracts reveal a much more intricate story, at least for large ticks and small time intervals τ – more precisely when the volatility on scale τ is small compared to the tick size s.

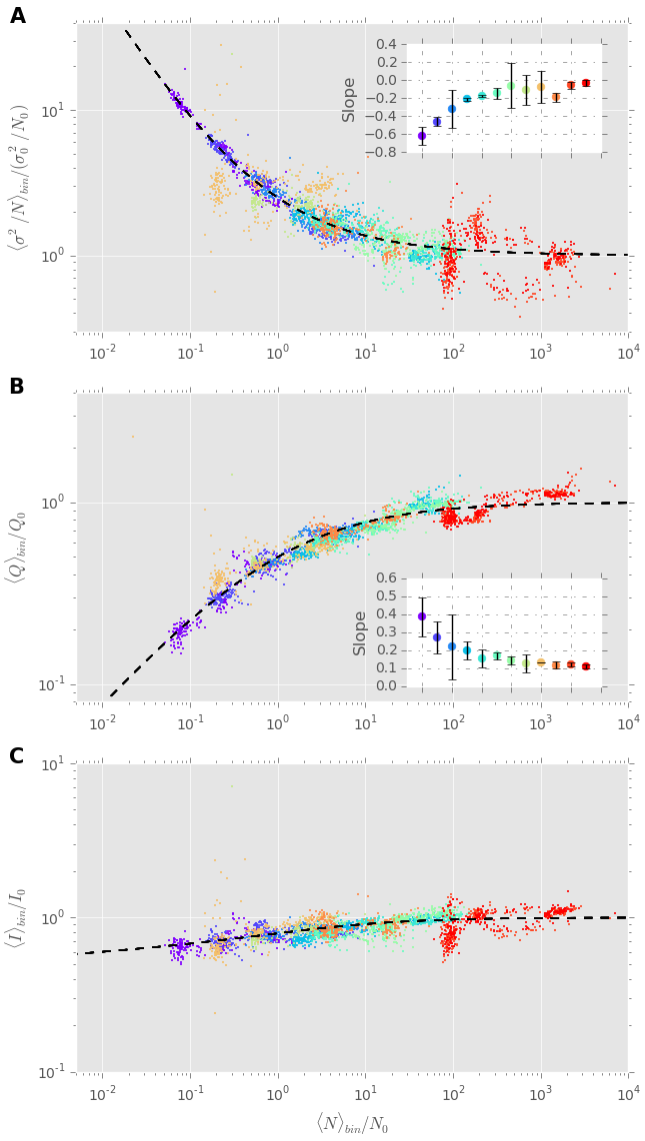

FIGURE D.6 – Data for twelve futures contracts at high frequency (5 minute bins). (A) Rescaled signature plot obtained by fitting $\langle \sigma^2/N \rangle_{\text{bin}}$ against $\langle N \rangle_{\text{bin}}$ as given by Eq. (D.6), with $a = 0.5$. (B) Rescaled average trade size obtained by fitting Eq. (D.7) with $\nu = 0.54$ to the data, as a function of $\langle N \rangle_{\text{bin}}/N_0$. (C) Rescaled plot of $\langle \tilde{I} \rangle_{\text{bin}}/I_0$ against $\langle N \rangle_{\text{bin}}/N_0$, resulting from (A) and (B), and consistent with Eq. (D.8). The values of N_0, σ_0 and Q_0 are reported in Tab. D.3. The insets of plots (A) and (B) display the slopes obtained from linear regression of $\langle \sigma^2/N \rangle_{\text{bin}}$ and $\langle Q \rangle_{\text{bin}}$ against $\langle N \rangle_{\text{bin}}$ respectively for each of the twelve futures contracts at hand.

D.4.2 A more complex picture

We first analyze independently the above two scaling laws ($\sigma \sim \sqrt{N}$, $V \sim N$) on our pool of twelve futures contracts. We focus first on $\tau = 5$-minute bins, a good trade-off between high frequency and noise. The insets of Figs D.6(A) and (B) show the exponents obtained by a power-law fit of $\langle \sigma^2/N \rangle_{\text{bin}}$ *vs.* $\langle N \rangle_{\text{bin}}$ (the so-called signature plot) and $\langle Q \rangle_{\text{bin}}$ *vs.* $\langle N \rangle_{\text{bin}}$ respectively. As can

be seen in these figures, small tick futures are indeed consistent with the expected $\sigma \sim \sqrt{N}$ and $V \sim N$ behaviour. For the large tick futures one rather finds $\sigma \sim N^{\beta}$ and $V \sim N^{\gamma}$, with $\beta < 1/2$ and $\gamma > 1$, suggesting of (i) a sub-diffusive price dynamics and (ii) an effective average trade size that increases with N. The rather puzzling fact, however, is the two exponents appear to conspire to give $\beta + \gamma \approx 3/2$ such that the scaling $\widetilde{W} \sim N^{3/2}$ indeed holds regardless of tick size.

D.4.3 Delayed diffusion for large ticks

A sub-diffusive behaviour for large tick contracts is in fact expected at short times, because a continuous random walk $B(\tau)$ that is constrained to take integer values $[B(\tau)] = n \times s$ (where n is an integer and s the tick size) can easily be shown to fluctuate as $\tau^{1/4}$ when τ is small (instead of the usual $\sqrt{\tau}$ behaviour). Furthermore, one expects a large amount of microstructural high frequency noise on the price when the tick is large. A simple way to account for these two effects is to postulate the following effective diffusion law : [5]

$$\sigma \quad = \quad \sigma_0 \left[a + \left(\frac{N}{N_0} \right)^{\frac{1}{2}} + \frac{N}{N_0} \right]^{\frac{1}{2}} \quad , \tag{D.6}$$

where a accounts for the high-frequency noise and N_0 is a characteristic number of trades such that the usual random walk behaviour is expected for $N \gg N_0$. One expects that $N = N_0$ roughly corresponds to a one tick move, so that σ_0 should be of order s and, correspondingly, a of order unity (since the amount of microstructural noise should be set by the tick size). We will see below that these expectations are indeed confirmed by the data (see Tab. D.3). Note that for large ticks and small trade sizes, one has $N_0 \gg 1$ and a very wide region where the anomalous sub-diffusion law $N^{1/4}$ holds. In the other limit $N_0 \lesssim 1$, the diffusive regime is almost immediately reached.

D.4.4 Master curves for volatility and volumes

Now, as shown in Figure D.6(A), the signature plots of *all* our futures contracts can be quite convincingly rescaled on a unique master curve given by Eq. (D.6), with appropriately chosen values of σ_0 and N_0 that are reported in Tab. D.3. We fixed $a = 0.5$ for all contracts, consistent with an overall goodness of fit when considering the twelve futures together. As expected, σ_0 is indeed found to be of the order of the tick size.

We now turn to the effective trade size $Q = V/N$, which can be similarly rescaled on a unique

5. For an inspiring approach of the price dynamics of large tick assets, see a recent paper by Dayri and Rosenbaum Dayri and Rosenbaum (2015). Their analysis might shed light on the results discussed here.

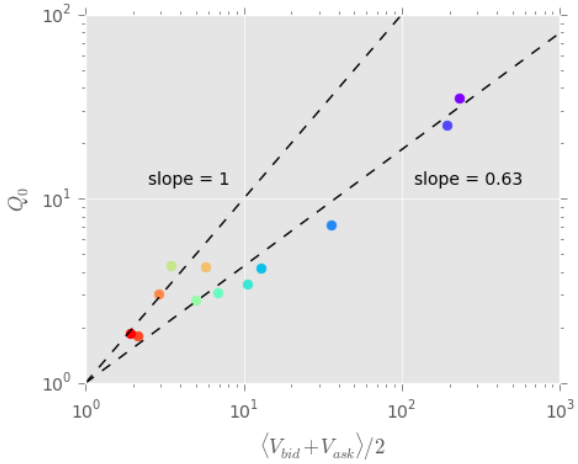

FIGURE D.7 – Plot of Q_0 – as obtained from fitting Eq. (D.7) to the data – as a function of average volume at the bid/ask (see Tab. D.3), for twelve futures contracts.

master curve by the following formula – see Fig. D.6(B) :

$$Q \;=\; Q_0 \left[1 + \left(\frac{N}{N_0}\right)^{-\nu}\right]^{-1} , \tag{D.7}$$

where the value of N_0 is fixed, contract by contract, to the very value favored by the rescaling of the signature plot. The only free parameters are Q_0 (reported in Tab. D.3), and the exponent $\nu \approx 0.54$, determined by the minimization of the overall error function of the aggregated data from all the futures. Note that Q_0 is the asymptotic value (for large N) of the average volume per trade. Figure D.7 displays Q_0 against the average volume at the bid/ask $V_{\text{best}} = (V_{\text{bid}} + V_{\text{ask}})/2$. As expected $Q_0 \sim V_{\text{best}}$ for small tick stocks, but grows sub-linearly for large tick stocks where trades only represent a smaller and smaller fraction of the available volume.

D.4.5 Deviations from the 3/2-law

We have shown that the deviations from simple diffusion and naive additivity of trade sizes can be rationalized by two more sophisticated scaling laws, Eqs. (D.6) and (D.7), leading to two universal master curves. Combining these two laws and letting $n = N/N_0$ and $I_0 = Q_0\sigma_0/\sqrt{N_0}$, allows one to write :

$$\frac{\widetilde{W}}{N^{3/2}} = \tilde{I} = I_0 \, \frac{\left(an^{-1} + n^{-1/2} + 1\right)^{1/2}}{1 + n^{-\nu}} \; . \tag{D.8}$$

Equation (D.8) offers a quantitative unifying picture of the above observation that $\beta + \gamma \approx 3/2$ regardless of tick size. Note that for $N \gg N_0$, $\tilde{I} \to I_0$, while for $N \ll N_0$, $\tilde{I} \to I_0 a\, n^{\nu-1/2}$ which is also nearly constant when $\nu \approx 1/2$. Therefore, our scaling analysis suggests that $\widetilde{W} N^{-3/2}$ has in

fact a residual N dependence. Figure D.6(C) displays a plot of $\langle \tilde{I} \rangle_{\text{bin}} / I_0$ against $\langle N \rangle_{\text{bin}} / N_0$, showing that the data can be rescaled onto a single master curve, as given by Eq. (D.8), and indeed revealing a small, but significant variation with N.

D.4.6 Time rescaling

Eq. (D.6) describes the crossover between a sub-diffusive regime for small N and a purely diffusive regime at large N, and was calibrated on different contracts for the same bin size $\tau = 5$

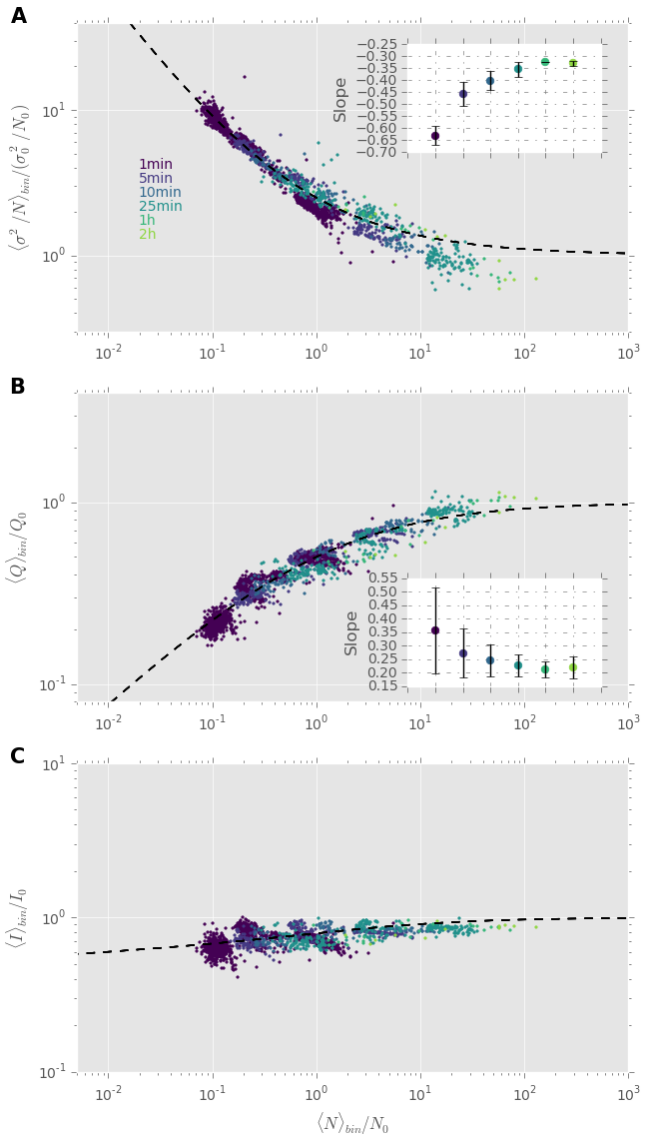

FIGURE D.8 – Figure analogous to Fig. D.6, only for the SPMINI futures contract at different sampling frequencies : $\tau = 1$ min, 5 min, 10 min, 25 min, 1h and 2h bins. The value of N_0 has been set to that measured on 5 minute bins. The values of σ_0 and Q_0 are left free but are found to be roughly constant across sampling frequencies (see Tab. D.4). The mild increase of I with τ (bottom graph) should be compared to the results shown in Fig. D.3 .

min. If our line of reasoning is correct, the *very same rescaling* should hold when focusing on a given contract but letting τ vary, in such a way that N/N_0 itself increases. This assumption is indeed in agreement with the data on all futures contracts. For the sake of clarity, we only show data for the SPMINI contract – but other contracts behave similarly. We considered $\tau = 1$ min, 5 min, 10 min, 25 min, 1h and 2h, and set N_0 to the value obtained in the previous paragraph for 5 minute bins. Figure D.8 displays plots analogous to those of Fig. D.6, but for the SPMINI contract across the different sampling frequencies. As one can see, our extend scaling hypothesis allows one to explain both the variations across contracts for a given τ and across time intervals. The scaling exponents of $\sigma(N)$ and $V(N)$ do converge to their natural values (1/2 and 1) as N becomes much larger than N_0. However, the fact that Eqs. (D.6) and (D.7) hold for all τ explains why $\tilde{I}/I_0$ increases mildly with τ, as Fig. D.3 above also demonstrated.

D.4.7 Naive scaling for single stocks

We now test the naive scalings $\sigma \sim \sqrt{N}$ and $V \sim N$ for $\tau = 5$ min by regressing $\log \sigma$ and $\log Q$ *vs.* $\log N$ for each of the three hundred stocks individually. Keeping with the notations introduced above, we find that the slopes β and γ of these regressions show no significant systematic dependence on the tick size. A cross sectional determination of these two exponents yields $\beta = 0.51 \pm 0.06$ and $\gamma = 1.04 \pm 0.07$, where the uncertainty again reflects the root mean square cross-sectional dispersion. At the daily time-scale one has equivalently $\beta = 0.54 \pm 0.10$ and $\gamma = 1.02 \pm 0.12$. Therefore, for all time scales $\tau \geq 5$ min, one can assume that the natural asymptotic scaling holds for all stocks, which trivially leads to the 3/2-law.

Still, it is surprising that the deviation from $\sigma \sim \sqrt{N}$, clearly observed for futures, does not seem to be present for stocks. In order to understand this difference, we display in Fig. D.9 the scatter plot σ^2/N vs. N for both our largest tick future (TBOND) and our largest tick stock (Applied Materials Inc, see Tab. D.2). The black line represents an average on consecutive log-spaced bins. As one can see, while for the future the average slope of the black line is consistent with the sub-diffusive behaviour discussed above, this is not the case for the stock which is on average rather flat.

This effect can actually be attributed to the fact that futures are traded on three different time zones, so that 5-minute bins where trading is slow (i.e. N small compared to N_0) are much more represented in the data than for stocks. The latter are indeed only active on American regular trading hours for which periods of very low activity are much rarer [6]. Therefore the regression slope β for futures is expected to be more sensitive to sub-diffusive effects. Furthermore, the volatility of single stocks is a factor $2-4$ smaller than the volatility of large tick futures, meaning that for the same relative tick size, discretisation effects are expected to be smaller for stocks. In this sense, large tick futures have a "larger tick" than large tick stocks!

6. Note that we only have 5-minute binned data for stocks, preventing us to zoom on shorter time intervals τ for

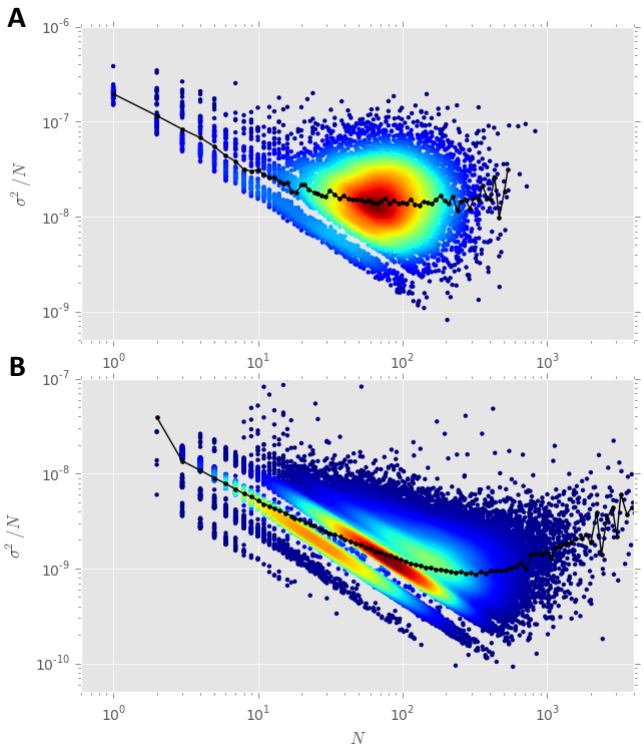

FIGURE D.9 – Signature scatter plot for (A) our largest tick stock (Applied Materials Inc, see Tab. D.2) and (B) our largest tick future (TBOND) computed with all 5-minute bin data. The black line represents an average on consecutive log-spaced bins, and the color code indicates the density of data. This graph shows that discretisation effects are much larger for the TBOND than for Applied Materials Inc.

D.5 Prices, spreads and a new definition of the trading invariant

As noted in Fig. D.3 and the left inset of Fig. D.5, the quantity I/S, where the spread S is given in dollars, seems to be more universal across assets than I itself, both for futures and for stocks. As a matter of fact, the theory presented in Wyart et al. (2008) – based on zero marginal profits for market makers, see also Madhavan et al. (1997) – predicts that for small tick contracts, $\sigma(N) = cS\sqrt{N}$, where S is the spread and c a numerical constant of order unity. Such a prediction was found to be very accurately obeyed by data, see Wyart et al. (2008). This observation naturally leads to a slightly amended definition of the trading invariant, that has the additional virtue of leading to a unit-less quantity, at variance with KO's definition where the invariant has dollar units. Our proposal is thus to consider the quantity $\mathcal{I}$, defined as :

$$\mathcal{I} = \frac{PV\sigma}{SN^{3/2}}, \tag{D.9}$$

which the sub-diffusive behaviour for large ticks should eventually show up.

where both the price and the spread are expressed in dollars. The quantity $\mathcal{I}$ is less scattered than I itself, although for stocks the existence of two branches is still visible in $\mathcal{I}$. Still, the NASDAQ branch is now much more universal, as anticipated by the observation made above that $\langle I \rangle$ is nearly proportional to $\langle S \rangle$ for these stocks.

Therefore, although the data is not convincingly supporting the existence of a universal value for $\mathcal{I}$ ($\langle \mathcal{I} \rangle$ is found to be a mildly decreasing function of the average spread both for the futures and the stocks), it is quite remarkable indeed that this quantity is to a first approximation constant across all contracts. In any case, we find it much more convincing to define a unit-less quantity as a plausible candidate for a genuine market invariant or quasi-invariant, in a sense we discuss now.

D.6 Conclusion

Let us summarize what we have achieved in this work :

— The most important result, to our eyes, is the 3/2-law, stating that the amount of risk W exchanged in markets (i.e. price times volume times volatility) scales like the 3/2 power of the number of trades N. We have shown that this holds very precisely across all 12 futures contracts and 300 single stocks, and across all times scales τ, thereby considerably extending the results obtained by Andersen et al. (2015) on the E-mini S&P futures.

— The second result is that the trading invariant $I = W/N^{3/2}$ proposed by Kyle and Obizhaeva (2010) is in fact quite different for different contracts, in particular between futures and single stocks. Furthermore, this quantity has dollar units, which makes its invariance property dubious. On the basis of a combination of dimensional, theoretical and empirical arguments, we have proposed that a more natural candidate should rather be $\mathcal{I} = \mathcal{I}/\mathcal{S}$, where S is the bid-ask spread. Whether the remaining weak dependence of $\mathcal{I}$ on the tick size is real or comes from some spurious biases is left for future investigations.

— Third, we have unveiled two remarkable master curves for the volatility σ and the traded volume V as a function of N, in the case of large tick futures contracts. We have argued for the existence of a characteristic number of trades N_0 above which the naively expected behaviour $\sigma \sim \sqrt{N}$ and $V \sim N$ hold, but below which strong deviations appear, induced by the size of the tick.

A synthetic way to summarize all our findings is to generalize and amend the dimensional analysis formula, Eq. (D.1), as :

$$\frac{PQ}{S} = f_{\text{asset}}\left(\frac{Q\sigma^2}{V}, \frac{P\sigma\sqrt{\tau}}{s}\right), \tag{D.10}$$

where the left-hand side is dimensionless. The function f_{asset} now depends on the asset class (futures *vs.* stocks) as well as on a second dimensionless argument which involves the time interval τ and the tick size s. For large enough τ, or small enough s, one expects this dependence to disappear,

Name	⟨Spread⟩	⟨N⟩	⟨Q⟩	⟨V⟩	⟨P⟩	⟨σ⟩	⟨$V_{\rm bid}$⟩	⟨$V_{\rm ask}$⟩	α	⟨I⟩	μ
TBOND	1.007	18	10.8	191	140960	1.22	206.1	206.5	1.58	98.0	2.49
SPMINI	1.011	45	11.2	499	81025	1.46	183.8	185.4	1.50	30.5	2.51
EUR	1.021	22	4.1	89	163831	1.70	33.9	33.7	1.52	18.6	2.26
NSDQMINI	1.120	17	2.8	47	58969	2.04	11.9	12.0	1.56	7.0	2.18
DJMINI	1.157	13	2.5	33	73777	1.98	9.8	9.8	1.53	6.7	2.33
CRUDE0	1.227	21	2.2	48	94290	2.50	6.4	6.4	1.55	12.0	1.37
GOLD0	1.323	21	2.2	45	153092	2.61	4.6	4.6	1.58	12.5	1.63
NCOFFEE0	2.066	6	2.0	12	59407	1.57	3.1	3.0	1.66	24.7	1.52
WTICRUDE0	2.258	9	1.8	17	94520	2.71	5.2	5.1	1.52	16.5	1.41
NCOTTON0	3.628	5	2.0	10	40375	2.38	2.5	2.5	1.47	10.9	1.64
HEATOIL0	4.950	9	1.7	15	122626	6.56	2.0	2.0	1.48	15.6	1.62
RBGASOL0	6.052	9	1.7	15	116292	7.49	1.8	1.8	1.46	18.1	1.61

TABLE D.1 – Summary table for twelve futures contracts. Values are computed in one minute bins. The average spread and volatility are given in units of tick, the trade size is given in number of contracts, the volume is given in contracts per unit time. The "invariant" $I = PV\sigma/N^{3/2}$ is given in dollars. All averages are defined as $\langle X\rangle := \exp\left[\mathbb{E}(\log X)\right]$.

i.e. $f(x, y \to \infty) = x^{-1/2}$ leading back to Kyle and Obizhaeva's hypothesis (up to the presence of the spread S, rather than I, in the left hand side). In the other limit, $f(x, y \to 0)$, could in principle behave very differently, but our detailed analysis above has revealed that $f(x, y)$ remains close to $x^{-1/2}$, with only a weak dependence on the second argument – see again Fig. D.6(C). In other words, the 3/2-law holds much beyond the regime $y \gg 1$, where it is expected on the basis of naive scaling. We do not have, at this stage, a detailed understanding of whether this is merely coincidental, or whether there is a deeper principle enforcing this property. We leave this as an open question for future work.

D.6.1 Acknowledgments

We wish to thank F. Patzelt for helping with the data analysis, as well as S. Hardiman, I. Mastromatteo, L. Duchayne, Z. Eisler, J. Kockelkoren, M. Potters, M. Vladkov, A. Darmon and C.-A. Lehalle for very fruitful discussions.

Name	⟨Spread⟩	⟨N⟩	⟨Q⟩	⟨V⟩	⟨P⟩	⟨σ⟩	α	⟨I⟩	μ
YHOO_UQ	1.006	68	325.2	22025	16.0	1.28	1.55	0.51	3.67
EBAY_UQ	1.097	124	153.4	19045	41.9	4.00	1.58	0.55	4.26
WMI_UN	1.099	21	249.3	5146	33.6	2.05	1.53	1.13	3.56
PG_UN	1.172	43	280.9	12057	66.1	3.18	1.52	1.37	3.42
FE_UN	1.335	21	187.8	3888	44.9	2.60	1.50	1.07	3.98
FAST_UQ	1.574	35	117.4	4133	44.9	3.66	1.53	0.72	4.17
DISCA_UQ	1.923	26	111.5	2853	52.5	3.45	1.53	0.76	3.89
ZMH_UN	2.204	17	148.0	2490	62.6	3.72	1.58	1.34	3.70
MON_UN	2.628	30	153.5	4577	82.7	6.06	1.54	1.70	3.95
MGA_UN	3.319	11	132.1	1508	43.1	3.63	1.53	1.42	3.58
ALXN_UQ	5.568	24	110.5	2628	93.5	7.85	1.54	1.78	3.47
MARY_UQ	7.613	13	126.9	1679	58.1	6.82	1.56	2.38	3.57
AMAT_UQ	1.003	55	338.6	18535	11.4	1.10	1.49	0.51	3.20

TABLE D.2 – Summary table for a random subset of twelve stocks and Applied Materials Inc. (used in Fig. D.9(A)). Values are computed in five minute bins. Units are identical to those of Tab. D.1. All averages are defined as $\langle X\rangle := \exp\left[\mathbb{E}(\log X)\right]$.

Name	⟨Spread⟩	N_0	σ_0	Q_0	I_0	⟨V_{best}⟩
TBOND	1.008	177.63	1.23	35.18	3.25	224.6
SPMINI	1.012	156.88	1.34	24.77	2.65	190.3
EUR	1.022	40.06	1.26	7.18	1.43	35.3
NSDQMINI	1.130	8.07	0.98	4.16	1.44	12.7
DJMINI	1.171	4.71	0.84	3.40	1.31	10.4
CRUDE0	1.242	5.85	1.08	3.06	1.36	6.8
GOLD0	1.340	4.46	0.99	2.79	1.31	4.9
NCOFFEE0	2.143	20.16	1.87	4.32	1.80	3.3
WTICRUDE0	2.304	45.17	2.78	4.24	1.75	5.6
NCOTTON0	3.819	2.16	1.25	3.01	2.56	2.8
HEATOIL0	5.092	0.08	0.51	1.80	3.28	2.1
RBGASOL0	6.253	0.07	0.55	1.86	3.78	1.9

TABLE D.3 – Values of N_0, σ_0 (in ticks), Q_0 obtained by fitting the data of twelve futures (with $\tau = 5$ min) to Eqs. (D.6) and (D.7) (see Fig. D.6), $I_0 = Q_0\sigma_0/\sqrt{N_0}$ as well as $V_{\text{best}} = (V_{\text{bid}} + V_{\text{ask}})/2$. Averages are defined as $\langle X\rangle := \exp\left[\mathbb{E}(\log X)\right]$.

Binsize	σ_0	Q_0	I_0
1min	1.21	33.80	3.27
5min	1.34	24.77	2.65
10min	1.45	21.98	2.54
25min	1.64	19.14	2.51
1h	1.83	16.93	2.47
2h	1.94	15.91	2.46

TABLE D.4 – Values of σ_0 (in ticks) and Q_0 obtained by fitting the data of the SPMINI at different sampling frequencies to Eqs. (D.6) and (D.7) (see Fig. D.8), and $I_0 = Q_0\sigma_0/\sqrt{N_0}$.

Annexe E

Processus de Hawkes quadratique pour la modélisation des prix

From
Quadratic Hawkes processes for financial prices
with Pierre Blanc and Jean-Philippe Bouchaud
(Blanc et al., 2015)

Tout comme le précédent, l'article qui suit ne s'insère pas directement dans l'histoire de cette thèse. Au lieu de rechercher une modélisation directement basée sur des comportements, il tente de déterminer un processus qui permette de reproduire les principaux fait stylisés connus à ce jour sur les prix et leur volatilité, à savoir une asymétrie par renversement du temps et des queues épaisses, en prenant pour point de départ les processus de Hawkes très populaires en finance en ce moment. Je remercie Pierre Blanc, son principal auteur, pour tout le travail qu'il a accompli et sans lequel cet article n'aurait jamais vu le jour.

Abstract : We introduce and establish the main properties of QHawkes ("Quadratic" Hawkes) models. QHawkes models generalize the Hawkes price models introduced in E. Bacry et al. (2014), by allowing all feedback effects in the jump intensity that are linear and quadratic in past returns. Our model exhibits two main properties, that we believe are crucial in the modelling and the understanding of the volatility process : first, the model is time-reversal asymmetric, similar to financial markets whose time evolution has a preferred direction. Second, it generates a multiplicative, fat-tailed volatility process, that we characterize in detail in the case of exponentially decaying kernels, and which is linked to Pearson diffusions in the continuous limit. Several other interesting properties of QHawkes processes are discussed, in particular the fact that they can generate long memory without necessarily be at the critical point. A non-parametric fit of the QHawkes model on NYSE stock data shows that the off-diagonal component of the quadratic kernel indeed has a

structure that standard Hawkes models fail to reproduce. We provide numerical simulations of our calibrated QHawkes model which is indeed seen to reproduce, with only a small amount of quadratic non-linearity, the correct magnitude of fat-tails and time reversal asymmetry seen in empirical time series.

E.1 Introduction : fBMs, GARCHs and Hawkes

The hunt for a "perfect" statistical model of financial markets is still going on. Since the primitive Brownian motion model first proposed by Bachelier, droves of more and more sophisticated mathematical frameworks have been devised to describe the salient stylized facts of financial time series, namely : fat (power-law) tails of the return distribution, volatility (or trading activity) clustering with slow decay of correlations, negative return-volatility correlations (the so-called leverage effect), etc. The two most successful family of models to date are : a) GARCH-like models with slowly decaying memory kernels (e.g. FIGARCH models) and b) stochastic volatility models where the log-volatility follows a fractional Brownian motion with a small Hurst exponent (e.g. the Multifractal Random Walk (Bacry et al., 2001) or, more recently, the "rough volatility" model of Gatheral et al. (2014)). Although these models are remarkably parsimonious and convincingly capture many features of financial time series, they are still unsatisfactory on several counts. First, the returns in the simplest version of these models are conditionally Gaussian and therefore never "fat enough", even with a fluctuating volatility. Non-Gaussian residuals (or jumps) must be therefore be introduced "by hand" to match empirical probability distributions. [1] Second, these models are not derived from deeper assumptions on the underlying mechanisms giving rise to fat-tails and volatility clustering. The theorist dream would be to start from, e.g. agents with simple trading rules or behavioral biases, and find that upon aggregation, their collective actions lead to a certain class of stochastic model. Many attempts in this direction have been documented, in particular agent-based models of markets, stylized population dynamics models, or "Minority Games" – for reviews see e.g. Challet et al. (2013); Cristelli et al. (2011). Still, it is fair to say that none of these proposals has yet been widely accepted as a convincing "micro-based" explanation of the stylized facts recalled above.

A further, less discussed, but in our eyes highly relevant stylized fact is related to the time-reversal (a)symmetry (TRS/TRA) of financial time series. As initially emphasized by Zumbach and Lynch (2001) (following earlier ideas, see e.g. Pomeau (1982); Ramsey and Rothman (1988, 1996); Müller et al. (1997); Arneodo et al. (1998)), financial time series are *not* statistically symmetrical when past and future are interchanged ; see Zumbach (2009). There are (at least) two distinct effects that break this symmetry : one is the leverage effect alluded to above : *past* returns r affect (negatively) *future* volatilities σ, but not the other way round. This is an effect that breaks

1. For various generalisations of the Multifractal Random Walk, see the in depth review proposed in Bacry et al. (2008).

both TRS and the up-down symmetry $r \to -r$. There is another effect though, that *is* invariant under $r \to -r$, namely : past large scale realized volatilities are more correlated with future small scale realized volatilities than vice-versa (Zumbach and Lynch, 2001). A more transparent way to explain this rather abstract notion is as follows : take r to be daily returns (say) and σ to be an estimator of volatility based on (say) five minute returns. Then consider, as in Chicheportiche and Bouchaud (2014), the average $\langle r_t^2 \sigma_{t+\tau}^2 \rangle_t$ with $\tau > 0$, which measures the correlation between past daily volatilities with future five minutes volatilities. The Zumbach effect, rephrased and empirically confirmed in Chicheportiche and Bouchaud (2014), is that $\langle r_t^2 \sigma_{t+\tau}^2 \rangle_t > \langle r_{t+\tau}^2 \sigma_t^2 \rangle_t$. It is clear that this criterion is invariant under $r \to -r$, and is thus unrelated to the leverage effect. Where does such an asymmetry come from and what are the models consistent with TRA ?

Interestingly, all continuous time stochastic volatility models, from the famous CIR-Heston model (Cox et al., 1985; Heston, 1993) to the Multifractal Random Walk model alluded to above, obey TRS by construction, and therefore *cannot* account for the empirical TRA of financial time series. GARCH-like models, on the other hand, do lead to strong TRA (Zumbach and Lynch, 2001), in fact stronger than seen in data (Chicheportiche and Bouchaud, 2014). This is expected ; GARCH models do encode a feedback from past to future : large past realized returns lead to large future volatilities. This self-exciting mechanism is actually very similar to the one underlying "Hawkes processes" (invented in the context of earthquake statistics), that have attracted a considerable amount of interest recently (for recent reviews, see Bacry et al. (2015); Laub et al. (2015)). In a financial context, Hawkes processes can be seen as a mid-way between purely stochastic models and agent based models. One postulates that the activity rate at time t, λ_t, depends on the history of the point process itself $N_{s<t}$ via the auto-regressive relation

$$\lambda_t = \lambda_\infty + \int_{-\infty}^{t} \phi(t-s)\, \mathrm{d}N_s, \tag{E.1}$$

where λ_∞ is a baseline intensity and ϕ is a non-negative, measurable function such that $||\phi||_1 = \int_0^\infty \mathrm{d}s \phi(s) \leq 1$. Hawkes processes are called "self-exciting", because every jump $\mathrm{d}N_s \neq 0$ increases the probability of future events for $t > s$ via the kernel ϕ ; this in turn leads to activity clustering with an enticing causal interpretation : each event is a new signal for the rest of the market, triggering more activity. When calibrated on financial data, two remarkable features of the Hawkes process are found (Brémaud and Massoulié, 2001; Hardiman et al., 2013; Hardiman and Bouchaud, 2014; Bacry et al., 2015; Lallouache and Challet, 2014) : its kernel $\phi(s)$ shows long-range (power-law) decay $s^{-1-\epsilon}$, and its L1 norm $||\phi||_1$ is very close to unity, meaning that the process is on the verge of becoming unstable (see however Filimonov and Sornette (2015)). This is quite interesting since this is precisely the regime where the corresponding continuous time limit for the squared volatility (identified here with the activity) is a fractional CIR-Heston process (Jaisson and Rosenbaum, 2015b), with local Hurst exponent $H = \epsilon - 1/2$. This seems to close the loop : since ϵ is empirically found to be

close to 1/2 (Hardiman et al., 2013), one has at hand a "micro-model" (the Hawkes process) that generates on coarse-grained scales a rough volatility process, which generalizes the CIR-Heston model to account for a slow, multi-timescale decay of volatility. Unfortunately, the situation is not as rosy yet : first, the fractional CIR-Heston process has tails that are much too thin (exponentially decaying) to account for the empirical distribution of volatility. Jaisson and Rosenbaum (2015b) therefore suggest to interpret the Hawkes process as a model for the *log*-volatility, but this is not natural. Second, following our discussion on TRS above, the fractional CIR-Heston process (as on fact the normal CIR-Heston one) is strictly TRS, and therefore fails to capture the observed TRA of financial time series !

The long story above sets the stage for our contribution, which is in an attempt to address the above deficiencies of the Hawkes formalism – when applied to financial time series – and take a step closer to the "perfect" model alluded to in our opening sentence. We propose a generalized version of the Hawkes process (called the QHawkes below) that includes features of the QARCH model introduced by Sentana (1995) and revisited in depth in Chicheportiche and Bouchaud (2014). The idea is that the self-exciting mechanism is not only from market activity onto market activity but also from actual price changes onto market activity. To make our motivation clear, consider a sequence of price moves, all with the same amplitude $|r| := \psi$ at the micro-scale. One expects that local trends, i.e. a succession of price moves in the same direction (up or down), triggers more volatility than a succession of compensated price moves, even though the high-frequency activity – the number of price moves – is exactly the same. The extra term we need to include in our generalized Hawkes process, beyond being motivated by empirical data, will encode mathematically this effect. We will see how this modification not only generates the needed fat tails in the return distribution (coming from the fact that the log-activity will indeed appear as a natural variable), but also accounts quantitatively for the TRA of returns, at least on the intraday time scales on which we calibrate the model. We will see that in a particular case, the continuous-time limit of our model boils down to a simple, tractable two-dimensional Pearson diffusion, which can then be used as a low-frequency proxy for the volatility process.

This paper is segmented into two distinct parts : the first part is mostly theoretical and analyzes the mathematical properties of our general QHawkes framework. The second part is concerned with calibration on financial data and simultation results. The detailed outline is as follows : we first introduce our general model in Section E.2, and highlight some of its core properties. Section E.2.4 introduces a particular sub-family of factorized QHawkes processes, which we call ZHawkes after Zumbach (as it captures the mechanism proposed by Zumbach to generate TRA). Section E.3 works out the autocorrelation structure of QHawkes processes in terms of the diagonal (linear) and off-diagonal (quadratic) kernels. We show in Section E.4 that in the case of exponential kernels the process is Markovian, and we write the corresponding stochastic differential equation as well as its continuous counterpart. This allows us to obtain the exact asymptotic power-law tail of the

volatility distribution in these special cases. Turning to the empirical part of the paper, Section E.5 works out the parallel with QARCH models, which we calibrate on intra-day US stock data using the methodology similar to Chicheportiche and Bouchaud (2014), clearly eliciting the off-diagonal structure of the feedback kernel. We then show in Section E.6, using numerical simulations, that the order of magnitude of the TRA generated by our ZHawkes process matches data quite well, and produce a volatility process with the right amount of fat-tails. Section E.7 finally concludes. More technical issues are relegated to Appendices, in particular our QARCH-based calibration procedure.

E.2 The QHawkes model

E.2.1 A general model

Similarly to Hawkes processes (E.1), the QHawkes (Quadratic Hawkes) process $(P_t)_{t\geq 0}$ is a self-exciting point process, whose intensity λ_t is dependent on the past realization of the process itself. As the name suggests, we model the intensity of price changes as the most general self-exciting point process that is Quadratic in $\mathrm{d}P_{s<t}$:

$$\lambda_t = \lambda_\infty \; + \; \frac{1}{\psi}\int_{-\infty}^{t} L(t-s)\; \mathrm{d}P_s \; + \; \frac{1}{\psi^2}\int_{-\infty}^{t}\int_{-\infty}^{t} K(t-s,t-u)\; \mathrm{d}P_s\; \mathrm{d}P_u, \tag{E.2}$$

where P is the high-frequency price, which is a pure jump process with signed increments. More precisely, whenever an event occurs between t and $t+\mathrm{d}t$, with probability $\lambda_t \mathrm{d}t$, the price jumps by an amount ξ, where ξ is a random variable of zero mean and variance ψ^2. A simple case that we will consider below is $\xi = \pm\psi$ with probability $\frac{1}{2}, \frac{1}{2}$, where ψ can be seen as the tick size. Although the only necessary condition for the results below to be valid is that jumps have finite variance, the $\pm\psi$ case nicely illustrates the fact that the properties on returns obtained on longer time scales do *not* rely on fat-tailed innovations. In the above equation, $L : \mathbb{R}^+ \to \mathbb{R}$ is a "leverage" kernel, coupling linearly price changes to market activity and $K : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$ is a quadratic feedback kernel. λ_∞ is again the baseline intensity of the process (in the absence of any feedback). Note that the above equation can be seen as a systematic expansion of the intensity of price changes in powers of past price changes, truncated to second order. One could of course generalize the model further by adding, e. g. a third order term as $\int_{-\infty}^{t}\int_{-\infty}^{t}\int_{-\infty}^{t} K_3(t-s,t-u,t-v)\mathrm{d}P_s\mathrm{d}P_u\mathrm{d}P_v$, etc., but we will not consider this path further in the following.

Although it is necessary to account for the leverage effect on daily time scales, we will find later that on intra-day scales, the kernel L is not significant, so for many applications one can focus on the quadratic kernel only and write

$$\lambda_t = \lambda_\infty \; + \; \frac{1}{\psi^2}\int_{-\infty}^{t}\int_{-\infty}^{t} K(t-s,t-u)\; \mathrm{d}P_s\; \mathrm{d}P_u. \tag{E.3}$$

It is easy to see that model (E.3) is a generalisation of the simple Hawkes process for prices introduced in Bacry et al. (2013) : when choosing unit price jumps $\mathrm{d}P_t = \pm\psi \mathrm{d}N_t$ where ψ can be seen as the tick and discarding any off-diagonal quadratic effects (so that $K(t,s) = \phi(t)\delta_{t-s}$), we recover a Hawkes process of kernel [2] $\phi(s) = K(s,s)$.

It is well known that the linear Hawkes process (E.1) can be seen as a branching process, where each « immigrant » event from the exogenous intensity λ_∞ gives birth to a number of « children » events distributed as a Poisson law of parameter $n_H = ||\phi||_1$, where $||\phi||_1$ is the L^1 norm of the kernel ϕ. Each of these children in turn gives birth to a second generation of children with the same probability law and so on. When $n_H < 1$, each immigrant gives birth on average to $n_H/(1-n_H) < \infty$ descendants. n_H can thus be seen as a measure of endogeneity of the process, since it corresponds to the fraction of events that are triggered internally, reaching zero in the case of simple Poisson process and one in the special case of Hawkes process without ancestors (Brémaud and Massoulié, 2001). The intuition behind the QHawkes in terms of a branching process is very similar, except that now the rate of events also depends on the interaction between the pairs of events. We will consider a positive feedback $K(s,t) \geq 0$ such that two mother events with the same sign (i.e. two prices moves in the same direction) increase the probability of a new event to be triggered in the future (i.e. increase future volatility), whereas compensated events have inhibiting effects, in line with (and directly motivated by) the empirical observations of Chicheportiche and Bouchaud (2014), as emphasized in the introduction. In fact, even when the conditions $L \equiv 0$ and $K \geq 0$ are satisfied, the inhibiting effects may in principle drive the process intensity λ below zero, whereas it should remain non-negative for the process to be well defined. A standard positive definiteness constraint is imposed in the next section to prevent such behaviour – which will trivially hold in the case of factorized ZHawkes.

E.2.2 Mathematical framework

Let us start by specifying the mathematical definition of the objects present in Equation (E.2) :

— $(P_t)_{t\in\mathbb{R}}$ is a pure jump process of stochastic intensity $(\lambda_t)_{t\in\mathbb{R}}$, with unpredictable i.i.d. jump sizes ξ of common law p on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. We assume that $\int_{\mathbb{R}} \xi \; p(\mathrm{d}\xi) = 0$ and $\int_{\mathbb{R}} \xi^2 \; p(\mathrm{d}\xi) = \psi^2 < +\infty$, i.e. that jumps are centered and have a finite variance.

— $\mathcal{F}_t = \sigma(P_s, s \leq t)$ is the natural filtration of P.

— $m(\mathrm{d}t, \mathrm{d}\xi)$ is the Punctual Poisson Measure associated to P, such that for all $t \in \mathbb{R}$ and $A \in \mathcal{B}(\mathbb{R})$,

$$\lim_{h\to 0} \frac{1}{h} \mathbb{E}\left[m([t, t+h[, A)\middle|\mathcal{F}_t\right] = \lambda_t \; p(A).$$

2. In fact, if the kernels K, K_3, etc. to arbitrary order are all diagonal, the model boils down to a Hawkes process with leverage, i.e. $\lambda_t = \lambda_\infty + \int_{-\infty}^t \phi(t-s)\mathrm{d}N_s + \psi^{-1}\int_{-\infty}^t L(t-s)\mathrm{d}P_s$, with adequately redefined kernels ϕ and L such that $\phi(s) - |L(s)| + \lambda_\infty \geq 0$ to ensure positivity of the intensity.

The quadratic kernel $K : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$ is assumed to satisfy

— Symmetry : $\forall s, t \geq 0,\ K(t,s) = K(s,t)$,

— Positive definiteness : $\forall d \geq 1,\ \forall \tau, c \in \mathbb{R}^d$ s.t. $0 \leq \tau_1 < \tau_2 < \cdots < \tau_d$, $\sum_{1 \leq i,j \leq d} K(\tau_i, \tau_j) c_i c_j \ \geq 0$,

— Non-explosion : $\int_0^{+\infty} |K(t,t)| \ \mathrm{d}t \ < +\infty$.

K defines an integral operator $T_K : L^2(\mathbb{R}^+) \to L^2(\mathbb{R}^+)$ which maps $f \in L^2(\mathbb{R}^+)$ to $T_K f : t \mapsto \int_0^{+\infty} K(t,s) f(s) \ \mathrm{d}s$. If K is continuous, this operator is Hilbert-Schmidt and thus compact and one has $K(t,t) \geq 0$ for all $t \geq 0$ (see Buescu (2004)). We define the trace of K

$$\mathrm{Tr}(K) = \int_0^{+\infty} K(t,t) \ \mathrm{d}t \ < +\infty.$$

The leverage kernel $L : \mathbb{R}^+ \to \mathbb{R}$ is assumed to be a measurable function. By analogy with QARCH models (see Chicheportiche and Bouchaud (2014)) it should be dominated by K in some way to ensure the positivity of the intensity λ_t (see footnote 2 above). Since the leverage kernel is found empirically negligible in the sequel, we leave this positivity condition for future research.

E.2.3 Necessary condition for first order time stationarity

In the case of linear Hawkes processes, it has been shown that stationarity is obtained as soon as the norm of the kernel verifies $\|\phi\|_1 < 1$. Intuitively, this means that each event triggers on average less than one child event, so that the clusters generated by each ancestor eventually die out. If this condition is violated, the probability that an ancestor generates an infinite number of events is non-zero, which can result in a stationary process only in the case $\|\phi\|_1 = 1$ and $\lambda_\infty = 0$ studied in Brémaud and Massoulié (2001), see also Hardiman et al. (2013). Because of the quadratic feedback, the QHawkes process cannot be interpreted as a simple branching process, making things somewhat trickier. The goal of this section is to find a necessary condition for (first order) time stationarity. We define the jump process (N_t) that has the same jump times as (P_t), with $\Delta N_\tau = (\Delta P_\tau)^2/\psi^2$ for any jump time τ, where we recall that ψ^2 denotes the variance of the jumps[3], and re-write Equation (E.2) as

$$\lambda_t = \lambda_\infty + \mathcal{L}_t + H_t + 2M_t \tag{E.4}$$

with the notations

$$\begin{cases} \mathcal{L}_t = \frac{1}{\psi} \int_{-\infty}^t L(t-u) \ \mathrm{d}P_u & \text{(leverage)} \\ H_t = \int_{-\infty}^t K(t-u, t-u) \ \mathrm{d}N_u & \text{(Hawkes/diagonal)} \\ M_t = \frac{1}{\psi^2} \int_{-\infty}^t \Theta_{t,u} \ \mathrm{d}P_u & \text{(off-diagonal)} \end{cases}$$

3. Note that $\Delta N_\tau = 1$ iff ΔP_τ has unit jumps $\Delta P_\tau = \pm\psi$.

where $\Theta_{t,u} = \int_{-\infty}^{u-} K(t-u, t-r)\ \mathrm{d}P_r$ is $(\mathcal{F}_u)_{u\leq t}$-adapted for t fixed. Since P is a martingale, one has $\mathbb{E}[M_t] = 0$ and $\mathbb{E}[\mathcal{L}_t] = 0$. Therefore,

$$\begin{aligned}\mathbb{E}[\lambda_t] &= \lambda_\infty + \frac{1}{\psi^2}\mathbb{E}\left[\int_{\mathbb{R}}\int_{-\infty}^{t} K(t-s, t-s)\ \xi^2\ m(\mathrm{d}s, \mathrm{d}\xi)\right] \\ &= \lambda_\infty + \mathbb{E}\left[\int_{-\infty}^{t} K(t-s, t-s)\ \lambda_s\ \mathrm{d}s\right]\end{aligned}$$

by definition of the punctual Poisson measure $m(\mathrm{d}s, \mathrm{d}\xi)$. We obtain

$$\mathbb{E}[\lambda_t] = \lambda_\infty + \int_{-\infty}^{t} K(t-s, t-s)\ \mathbb{E}[\lambda_s]\,\mathrm{d}s.$$

A necessary condition for the process $(\lambda_t)_{t\in\mathbb{R}}$ to be in a stationary state is that its expected value $\overline{\lambda} \equiv \mathbb{E}[\lambda_t]$ is constant, positive and finite. This yields $\overline{\lambda} = \lambda_\infty + \overline{\lambda}\mathrm{Tr}(K)$, thus if $\lambda_\infty > 0$,

$$\overline{\lambda} \;=\; \frac{\lambda_\infty}{1 - \mathrm{Tr}(K)}.$$

This leads to the necessary stationarity condition [4]

$$\lambda_\infty > 0 \text{ and } \mathrm{Tr}(K) < 1 \tag{E.5}$$

$$\text{or } \lambda_\infty = 0 \text{ and } \mathrm{Tr}(K) = 1. \tag{E.6}$$

The existence of a finite average intensity $\overline{\lambda}$ is of course necessary for the process to reach a stationary state. However, the existence of higher moments of the intensity require stronger conditions on $K(t, s)$, similarly to the QARCH case studied in Chicheportiche and Bouchaud (2014). In particular, the decay of the off-diagonal part of K must be fast enough to ensure the existence of two-point and three-point correlations of the process (see below).

E.2.4 A special case : the ZHawkes model

Motivated by the discussion in the introduction and by the empirical (intraday) results presented below, we will specialize the QHawkes model to the case where there is no leverage ($L \equiv 0$) and the quadratic feedback kernel K is of the form

$$K(t, s) = \phi(t)\delta_{t-s} + k(t)k(s),$$

4. In the case of linear Hawkes processes, this condition is also sufficient to obtain stationarity in the case $\mathrm{Tr}(K) < 1$ (whereas the case $\mathrm{Tr}(K) = 1$ is more subtle, see Brémaud and Massoulié (2001)).

i.e. the sum of a diagonal Hawkes component and of a factorisable, rank one kernel. We assume that $\phi, k : \mathbb{R}^+ \to \mathbb{R}^+$ are two positive, measurable functions that satisfy

$$||\phi||_1 \equiv \int_0^{+\infty} \phi(u)\, \mathrm{d}u < +\infty \quad , \quad ||k^2||_1 \equiv \int_0^{+\infty} k(u)^2\, \mathrm{d}u < +\infty.$$

Equation (E.2) becomes in that case

$$\lambda_t = \lambda_\infty + H_t + Z_t^2, \tag{E.7}$$

where

— The « Hawkes term » is given by

$$H_t := \int_{-\infty}^t \phi(t-s)\, \mathrm{d}N_s; \qquad N_t - N_{t^-} := \frac{1}{\psi^2}(P_t - P_{t^-})^2$$

— The « Zumbach term » given by Z_t^2 where

$$Z_t = \frac{1}{\psi} \int_{-\infty}^t k(t-s)\, \mathrm{d}P_s.$$

In the special case of the ZHawkes process, the endogeneity ratio introduced above is given by :

$$\mathrm{Tr}(K) = ||\phi||_1 + ||k^2||_1 \equiv n_H + n_Z,$$

where n_H is the standard « Hawkes norm » while $n_Z \equiv ||k^2||_1$ is the « Zumbach norm ». We call Z_t the Zumbach term since it is directly inspired by the series of empirical observations made by G. Zumbach on the volatility process (Zumbach and Lynch, 2001; Zumbach, 2009). Indeed, Z_t is simply a moving average of the past returns (with positive un-normalized weights $k(\tau)$). Therefore, this term will indeed be such that a sequence of returns in the same direction triggers more future volatility than compensated returns, as empirically observed (Zumbach, 2009). [5]

Besides its empirical motivations, the factorization property of the ZHawkes kernel should significantly reduce the risk of over-fitting, since one is left with two one-dimensional kernels instead of the two-dimensional kernel in Eq. E.2. As we see below, this simplified setup still captures the main phenomenology of price volatility, with in particular time-reversal asymmetry and fat tails, *even for short-ranged kernels.*

5. Although Zumbach describes this effect at the daily time scale, whereas we will here study intra-day time scales.

E.3 The auto-correlation structure of QHawkes processes

It is quite useful for such type of models to investigate the relation between the input kernels and the auto-correlation functions of the generated QHawkes process. Indeed, since the latter is directly observable on the data, the underlying kernels can then obtained by inverting such relations. For linear Hawkes processes, one finds a Wiener-Hopf equation that relates the two-points correlation function to the 1-d kernel (Bacry et al., 2012). In our case, one also needs to consider the three-points correlation function, which will lead to a set of closed relations.

E.3.1 Exact set of equations

We take the model with no leverage, $L \equiv 0$. Equation (E.4) becomes (see notations above) :

$$\lambda_t = \lambda_\infty + H_t + 2M_t.$$

We define for $\tau \neq 0$ and $\tau_1 > 0, \tau_2 > 0, \tau_1 \neq \tau_2$, the correlation functions

$$\begin{aligned} \mathcal{C}(\tau) &\equiv \mathbb{E}\left[\frac{\mathrm{d}N_t}{\mathrm{d}t}\frac{\mathrm{d}N_{t-\tau}}{\mathrm{d}t}\right] - \overline{\lambda}^2 \;=\; \mathbb{E}\left[\lambda_t \frac{\mathrm{d}N_{t-\tau}}{\mathrm{d}t}\right] - \overline{\lambda}^2, \\ \mathcal{D}(\tau_1,\tau_2) &\equiv \frac{1}{\psi^2}\,\mathbb{E}\left[\frac{\mathrm{d}N_t}{\mathrm{d}t}\frac{\mathrm{d}P_{t-\tau_1}}{\mathrm{d}t}\frac{\mathrm{d}P_{t-\tau_2}}{\mathrm{d}t}\right] \;=\; \frac{1}{\psi^2}\,\mathbb{E}\left[\lambda_t \frac{\mathrm{d}P_{t-\tau_1}}{\mathrm{d}t}\frac{\mathrm{d}P_{t-\tau_2}}{\mathrm{d}t}\right]. \end{aligned} \tag{E.8}$$

$\mathcal{C}$ is then extended continuously at zero, as in Hawkes (1971). Let us note that by construction $\mathcal{C}$ is even and $\mathcal{D}$ is symmetric. One finds the following exact equations between the auto-correlation functions ($\mathcal{C}$, $\mathcal{D}$) and the kernel K (cf. the derivation in Appendix E.7) :

$$\mathcal{C}(\tau) = \kappa\overline{\lambda}K(\tau,\tau) + \int_0^\infty \mathrm{d}u\, K(u,u)\mathcal{C}(\tau+u) + 2\int_{0^+}^\infty \mathrm{d}u \int_{u^+}^\infty \mathrm{d}r\, K(\tau+u,\tau+r)\mathcal{D}(u,r), \tag{E.9}$$

$$\begin{aligned} \mathcal{D}(\tau_1,\tau_2) =& 2K(\tau_1,\tau_2)[\mathcal{C}(\tau_2-\tau_1) + \overline{\lambda}^2] + \int_{(\tau_2-\tau_1)^+}^{\tau_2} \mathrm{d}u\, K(\tau_2-u,\tau_2-u)\mathcal{D}(u-\tau_2+\tau_1,u) \\ &+ 2\int_{(\tau_2-\tau_1)^+}^{+\infty} \mathrm{d}u\, K(\tau_1,\tau_2+u)\mathcal{D}(\tau_2-\tau_1,\tau_2-\tau_1+u), \end{aligned} \tag{E.10}$$

where $\kappa = \frac{1}{\psi^4}\int_{\mathbb{R}} \xi^4\; p(\mathrm{d}\xi)$ is the fourth moment of price jumps ($\kappa = 1$ in the particular case of constant price jumps). As $\mathcal{C}(\tau)$ and $\mathcal{D}(\tau_1,\tau_2)$ are directly measurable on the data, one can in principle infer the kernel $K(s,t)$ by inverting the above equations.

E.3.2 Asymptotic behaviour in the case of power-law kernels

Whereas the above equations (E.9) and (E.10) are difficult to solve in general, one can investigate the joint tail behaviour as $\tau \to \infty$ when both the kernel and the auto-correlation functions have

power law decays. Let us assume that :

$$\begin{cases} K(\tau,\tau) & \underset{\tau\to\infty}{\sim} c_0\ \tau^{-1-\epsilon} & \text{(diagonal, } \epsilon>0) \\ K(\tau v_1,\tau v_2) & \underset{\tau\to\infty}{\sim} \tilde{K}(v_1,v_2)\ \tau^{-2\delta} & \text{(off-diagonal, } \delta>1/2) \\ \mathcal{C}(\tau) & \underset{\tau\to\infty}{\sim} c_1\ \tau^{-\beta} & \text{(2-points AC)} \\ \mathcal{D}(\tau,\tau) & \underset{\tau\to\infty}{\sim} c_2\ \tau^{-\beta'} & \text{(3-points AC, diagonal)} \\ \mathcal{D}(\tau v_1,\tau v_2) & \underset{\tau\to\infty}{\sim} \tilde{\mathcal{D}}(v_1,v_2)\ \tau^{-2\rho} & \text{(3-points AC, off-diagonal)} \end{cases} \tag{E.11}$$

where c_0, c_1, c_2 are constants and $\tilde{K}(v_1,v_2), \tilde{\mathcal{D}}(v_1,v_2)$ are bounded functions of (v_1,v_2). The constraint $\epsilon>0$ comes from the fact that $||\phi||_1$ must be finite, whereas $\delta>1/2$ insures that the second and third moments are finite as well. We will furthermore assume (for simplicity) that $\epsilon<1$, which is the interesting case in practice, and that the asymptotic behaviour of $K(\tau_1,\tau_2)\sim\tau_1^{-1-\epsilon}$ is restricted to a narrow channel around the diagonal $|\tau_1-\tau_2|\ll\tau_1,\tau_2$, beyond which the off-diagonal power-law takes over.

The exponents β and ρ can then be related to δ and ϵ by plugging these ansätze into Eqs. (E.9) and (E.10) and carefully matching the asymptotic behaviours. One finds several possible phases for the auto-covariance structure :

1. In the *non critical* case $\mathrm{Tr}(K)<1$, we find :

$$\delta>(3+\epsilon)/4 \Rightarrow \beta=1+\epsilon; \qquad \beta'=1+\epsilon; \qquad \rho=\delta, \tag{E.12}$$

$$(2+\epsilon)/3<\delta<(3+\epsilon)/4 \Rightarrow \beta=4\delta-2; \qquad \beta'=1+\epsilon; \qquad \rho=\delta, \tag{E.13}$$

$$\frac{1}{2}<\delta<(2+\epsilon)/3 \Rightarrow \beta=4\delta-2; \qquad \beta'=3\delta-1; \qquad \rho=\delta, \tag{E.14}$$

   The interpretation of these three phases is straightforward. In the first phase (E.12), the tail of the auto-correlation functions directly comes from the tail of the diagonal part of K : direct effects then dominate quadratic feedback effects. In the last two phases (E.13),(E.14) however, a more sophisticated phenomenon comes into play, as off-diagonal effects feedback in such a way that they generate correlations with slower decay than that of the diagonal part of the kernel itself. In these phases, there is a possibility that $\beta<1$ (corresponds to a long memory process) provided the off-diagonal kernel decays slowly enough $\frac{1}{2}<\delta<\frac{3}{4}$. This result is important as it means that QHawkes processes *need not be critical* (i.e. $\mathrm{Tr}(K)=1$) to generate long memory, unlike standard, linear Hawkes processes (Brémaud and Massoulié, 2001; Saichev and Sornette, 2010; Hardiman et al., 2013; Hardiman and Bouchaud, 2014). If the off-diagonal kernel decays faster than $\tau^{-3/2}$, however, the corresponding process can only have long memory if it is critical and with a slowly decaying diagonal kernel.

2. In the *critical* case $\mathrm{Tr}(K)\to1$, $\lambda_\infty\to0$, the situation is more subtle, as in the standard Hawkes

case where the relation between β and ϵ completely changes, and the condition $0 < \epsilon < 1/2$ must hold for the process to even exist (Brémaud and Massoulié, 2001). In the present case, a similar mechanism operates and leads to :

$$\delta > 3/4 \Rightarrow \beta = 1 - 2\epsilon; \qquad \beta' = 1 - \epsilon; \qquad \rho = \delta, \tag{E.15}$$

$$2/3 < \delta < 3/4 \Rightarrow \beta = 4\delta - 2\epsilon - 2; \qquad \beta' = 1 - \epsilon; \qquad \rho = \delta. \tag{E.16}$$

$$(1+\epsilon)/2 < \delta < 2/3 \Rightarrow \beta = 4\delta - 2\epsilon - 2; \qquad \beta' = 3\delta - \epsilon - 1; \qquad \rho = \delta. \tag{E.17}$$

provided $0 < \epsilon < 1/2$ and $\delta > (1+\epsilon)/2$, otherwise the critical process does not exist or is trivial. So in this critical case, the process is always long-memory (i.e. $\beta < 1$), or ceases to exist, as for the linear Hawkes process.

E.4 Volatility distribution in the ZHawkes model

The volatility distribution for the completely general QHawkes process is difficult to characterize analytically. For the ZHawkes model with exponential kernels $\phi(.)$ and $k(.)$, some progress is however possible because the process becomes Markovian and one can write a stochastic differential equation (SDE) to describe its evolution. Although this assumption is quite restrictive, this case allows one to gain a good intuition on the model, so we investigate this limit in details. It also turns out that the Markovian case is actually extremely interesting mathematically.

E.4.1 SDE in the exponential case

For the sake of simplicity, let us assume that the price jumps are binary $\xi = \pm\psi$, and we set $\psi = 1$ without loss of generality. Besides, we note $k(t) = \sqrt{2 n_Z \omega}\, \exp(-\omega t)$ and $\phi(t) = n_H \beta \exp(-\beta t)$, where n_H is the Hawkes norm and n_Z the Zumbach norm. We require :

$$\mathrm{Tr}(K) = n_H + n_Z < 1.$$

Then the model can be written in this case : $\lambda_t = \lambda_\infty + H_t + Z_t^2$ where

$$\begin{cases} \mathrm{d}H_t &= \beta\ [-H_t\ \mathrm{d}t\ + n_H\ \mathrm{d}N_t]\,, \\ \mathrm{d}Z_t &= -\omega\ Z_t\ \mathrm{d}t\ +\ k_0\ \mathrm{d}P_t \end{cases} \tag{E.18}$$

The processes N and P jump simultaneously with intensity λ_t and amplitudes $\Delta N_\tau = 1$ and $\Delta P_\tau = \pm 1$ with equal probability. Although quite simple, this system of jump SDEs lacks tractability compared to a continuous diffusion. Thus, we turn to the low-frequency asymptotics that one obtains as the number of jumps in a given time window becomes large, while their amplitudes are scaled down accordingly. This is the object of the following section.

E.4.2 Low-frequency asymptotics

The low-frequency asymptotics of nearly critical Hawkes processes with short-ranged kernels have been investigated in details by Jaisson and Rosenbaum (2015a,b). They show that for suitable scaling and convergence to the critical point $n_H = 1$, the short memory Hawkes-based price process of Bacry and Muzy (2014) converges towards a Heston process (since the Hawkes intensity converges towards a CIR volatility process). The same authors (Jaisson and Rosenbaum, 2015b) show that when the kernel exhibits power-law behaviour $\phi(t) \sim t^{-1-\epsilon}$ with $1/2 < \epsilon < 1$, the limiting process for the intensity is a fractional Brownian motion with Hurst exponent $H = \epsilon - \frac{1}{2}$. When ϵ is close to $1/2$, as empirical data suggests (Hardiman et al., 2013), the roughness of the latter process is in agreement with the empirical results of Bacry et al. (2001); Gatheral et al. (2014) who find a Hurst exponent H close to zero the *log*-volatility ($H = 0$ for the multifractal model of Bacry et al. (2001)). However, it is unclear how the Hawkes process intensity can be identified with the log-volatility. A fat-tailed behaviour cannot be reproduced by a simple, linear Hawkes process, as it is absent from Heston-CIR processes (see also below).

Here, we want to investigate the low-frequency asymptotics of the Markovian ZHawkes model, which, as we shall see, reveals very interesting new features, induced by quadratic feedback effects.

Choosing a time scale $T > 0$ that will eventually diverge, we define the processes $\bar{H}_t^T = H_{tT}$, $\bar{Z}_t^T = Z_{tT}$, $\bar{N}_t^T = N_{tT}$ and $\bar{P}_t^T = P_{tT}$, with the parameters β_T and ω_T that may depend on T, but with fixed endogeneity parameters n_H and n_Z : Equation (E.18) gives

$$\begin{cases} \mathrm{d}\bar{H}_t^T &= -\beta_T \left[\bar{H}_t^T \; T\mathrm{d}t \; + \; n_H \; \mathrm{d}\bar{N}_t^T \right], \\ \mathrm{d}\bar{Z}_t^T &= -\omega_T \; \bar{Z}_t^T \; T\mathrm{d}t \; + \; \gamma_T \; \mathrm{d}\bar{P}_t^T, \end{cases} \tag{E.19}$$

where $\gamma_T^2 := 2\omega_T n_Z$ and the common jump intensity of $\bar{N}^T$ and $\bar{P}^T$ is $T \times [\lambda_\infty + \bar{H}_t^T + (\bar{Z}_t^T)^2]$. Since the signs of the jumps of $\bar{P}^T$ are assumed to be unpredictable and equal to ± 1, the infinitesimal generator of the process is given by

$$\begin{aligned} \mathcal{A}^T f(h,z) \; = \; & -\beta_T \; h \; T \; \partial_h f(h,z) \; - \; \omega_T \; z \; T \; \partial_z f(h,z) \\ & + T \left[\lambda_\infty + h + z^2\right] \left\{ \frac{1}{2} f\left(h + n_H \beta_T, \; z + \gamma_T\right) + \frac{1}{2} f\left(h + n_H \beta_T, \; z - \gamma_T\right) - f\left(h, z\right) \right\} \end{aligned} \tag{E.20}$$

for any functions f twice continuously differentiable on $(0, +\infty) \times \mathbb{R}$. We now consider the following scaling

$$\beta_T = \overline{\beta}/T, \qquad , \; \omega_T = \overline{\omega}/T, \tag{E.21}$$

with $\overline{\beta}, \overline{\omega} > 0$. Since we fixed the values of n_H and n_Z, our procedure can be called a « constant endogeneity rescaling », as opposed to the scaling used by Jaisson and Rosenbaum (2015a) and Jaisson and Rosenbaum (2015b), where the endogeneity ratio n_H of the process needs to converge to

unity as T goes to infinity. Our choice is partly motivated by the calibration results of Section E.5.2 for intra-day returns, that yield an endogeneity ratio in the range $0.7 - 0.9$, close to what is obtained at the daily time scale in Chicheportiche and Bouchaud (2014) and Blanc et al. (2014), and significantly away from the critical value $n_H = 1$. Equations (E.20) and (E.21) then combine as

$$\begin{aligned}\mathcal{A}^T f(h,z) \;=\; & -\overline{\beta}\; h\; \partial_h f(h,z) \;-\; \overline{\omega}\; z\; \partial_z f(h,z) \\ & + T\left[\lambda_\infty + h + z^2\right]\left\{\frac{1}{2}f\left(h + n_H\frac{\overline{\beta}}{T},\; z + \frac{\overline{\gamma}}{\sqrt{T}}\right) + \frac{1}{2}f\left(h + n_H\frac{\overline{\beta}}{T},\; z - \frac{\overline{\gamma}}{\sqrt{T}}\right) - f\left(h,z\right)\right\},\end{aligned}$$

where we introduced $\overline{\gamma} = \sqrt{2n_Z\overline{\omega}}$. We turn to the low-frequency asymptotics. As T goes to infinity, one has

$$\frac{1}{2}f\left(h + n_H\frac{\overline{\beta}}{T},\; z + \frac{\overline{\gamma}}{\sqrt{T}}\right) + \frac{1}{2}f\left(h + n_H\frac{\overline{\beta}}{T},\; z - \frac{\overline{\gamma}}{\sqrt{T}}\right) - f\left(h,z\right) = \frac{n_H\overline{\beta}}{T}\partial_h f(h,z) + \frac{\overline{\gamma}^2}{2T}\partial^2_{zz} f(h,z) + \mathrm{o}\left(\frac{1}{T}\right),$$

therefore $\mathcal{A}^T f(h,z)$ converges to

$$\mathcal{A}^\infty f(h,z) \;=\; -\overline{\beta}\left[(1-n_H)h - n_H(\lambda_\infty + z^2)\right]\partial_h f(h,z) - \overline{\omega} z \partial_z f(h,z) + n_Z\overline{\omega}\left[\lambda_\infty + h + z^2\right]\partial^2_{zz} f(h,z).$$

The operator $\mathcal{A}^\infty$ is the infinitesimal generator of the diffusion

$$\left\{\begin{array}{rl} \mathrm{d}\bar{H}^\infty_t & = \left[-(1-n_H)\;\bar{H}^\infty_t + n_H\left(\lambda_\infty + \left(\bar{Z}^\infty_t\right)^2\right)\right]\overline{\beta}\mathrm{d}t, \\ \mathrm{d}\bar{Z}^\infty_t & = -\overline{\omega}\;\bar{Z}^\infty_t\;\mathrm{d}t \;+\; \overline{\gamma}\;\sqrt{\lambda_\infty + \bar{H}^\infty_t + \left(\bar{Z}^\infty_t\right)^2}\mathrm{d}W_t, \end{array}\right. \tag{E.22}$$

where W is a standard Brownian motion. A standard argument of Kallenberg (Kallenberg, 2002) (Theorem 19.25) then gives the convergence of the process $(\bar{H}^T, \bar{Z}^T)$ to $(\bar{H}^\infty, \bar{Z}^\infty)$ as T goes to infinity. Hence, one does *not* need that the norm of the process tends to 1 (i.e. that the process is nearly critical) for a non-degenerate limit process to be obtained. The above limiting process is the major result of this section. Although it was derived for a Markovian ZHawkes process, we believe that this is the limiting process for the whole class of non-critical ZHawkes processes with short memory, and is the analogue of the Heston-CIR limiting process for Hawkes, as in Jaisson and Rosenbaum (2015a). The limiting behaviour corresponding to long-memory/critical ZHawkes processes, in the spirit of Jaisson and Rosenbaum (2015b), is left for future investigations. We now investigate some of the properties of the limiting process, Eq. (E.22), in particular the induced tail of the volatility distribution.

E.4.3 Tail of the volatility distribution

From now on we drop the superscript ∞ on $\bar{H}$ and $\bar{Z}$; the fact that we are studying the limiting process is implied. Let us note first that there is no Brownian part in the SDE for $\bar{H}$ so that it can

be solved explicitly as a deterministic function of $(\bar{Z}_s)_{s\leq t}$:

$$\bar{H}_t \;=\; \bar{H}_\infty \;+\; n_H\overline{\beta}\int_{-\infty}^{t}\exp(-(1-n_H)\overline{\beta}(t-s))\bar{Z}_s^2\mathrm{d}s; \qquad \bar{H}_\infty := \frac{\lambda_\infty}{1-n_H}$$

In the considered limit, $\bar{H}_t$ can thus be written as the sum of a constant term and an exponential moving average of the square of $\bar{Z}_s$. We get the autonomous, but non-Markovian SDE for $\bar{Z}_t$:

$$\mathrm{d}\bar{Z}_t = -\overline{\omega}\;\bar{Z}_t\;\mathrm{d}t \;+\; \overline{\gamma}\;\sqrt{\bar{H}_\infty + \bar{Z}_t^2 + n_H\overline{\beta}\left[\int_{-\infty}^{t}\exp(-(1-n_H)\overline{\beta}(t-s))\bar{Z}_s^2\mathrm{d}s\right]}\;\mathrm{d}W_t. \tag{E.23}$$

ZHawkes without Hawkes

We first consider the simpler case where the Hawkes feedback is zero, i.e. $n_H = 0$. This corresponds to the case where only the Zumbach term is present in the starting model, i.e. $\lambda_t = \lambda_\infty + Z_t^2$ in Equation (E.7). As we see in the sequel, this simpler model is still rich enough to reproduce some interesting empirical properties of the volatility process. One gets :

$$\mathrm{d}\bar{Z}_t = -\overline{\omega}\;\bar{Z}_t\;\mathrm{d}t \;+\; \overline{\gamma}\;\sqrt{\lambda_\infty + \bar{Z}_t^2}\;\mathrm{d}W_t, \tag{E.24}$$

which is a particular case of Pearson diffusions, which are extensively described and classified in Forman and Sørensen (2008). The process $\bar{Z}/\sqrt{\lambda_\infty}$ fits in Case 3 of their classification (see Forman and Sørensen (2008) Section 2.1), with the dictionary $\mu \to 0, \theta \to \overline{\omega}$ and $a \to n_Z$. Therefore, $\bar{Z}_t$ is ergodic and its stationary law is a Student t-distribution with $1 + 1/n_Z$ degrees of freedom and scale parameter $\sqrt{n_Z\lambda_\infty/(1+n_Z)}$. This implies that stationary law of the square of $\bar{Z}^\infty$ is a F-distribution with 1 and $1 + 1/n_Z$ degrees of freedom, and scale parameter $n_Z\lambda_\infty/(1+n_Z)$. We will denote as

$$V_t \;=\; \psi^2\left[\lambda_\infty + \bar{Z}_t^2\right]$$

the low-frequency squared volatility of the price (we reintroduced the jump size ψ for completeness). A straightforward change of variables yields the stationary density $q(v)$ of the process V as :

$$q(v) = \frac{\Gamma\left(1+\frac{1}{2n_Z}\right)}{\Gamma\left(\frac{1}{2}+\frac{1}{2n_Z}\right)\sqrt{\pi v_\infty(v-v_\infty)}}\left(\frac{v_\infty}{v}\right)^{\left(1+\frac{1}{2n_Z}\right)}\mathbf{1}_{\{v>v_\infty\}} \tag{E.25}$$

where $v_\infty = \lambda_\infty\psi^2$ is the baseline level of the squared volatility. For the tail exponent of the distribution of V_t, we get a power-law tail :

$$q(v) \underset{v\to+\infty}{\sim} C\;v^{-\left(\frac{3}{2}+\frac{1}{2n_Z}\right)} \tag{E.26}$$

with C an explicit constant. We find this result interesting for two reasons. First, one obtains a power-law behavior that emerges naturally from the fact that since the volatility behaves as $|\bar{Z}_t|$ for large values of $\bar{Z}_t$, the process describing its dynamics is simply a multiplicative Brownian motion with drift (see E.24). This is at variance with the « diagonal » Hawkes counterpart of Jaisson and Rosenbaum (2015a) where the coefficient in front of the Brownian noise is only the *square-root* of the volatility, which inevitably leads to a process that has a characteristic scale and thin tails. Second, the stationary distribution of V only depends on the Zumbach norm n_Z, that can be seen as the endogeneity of the process. This last result suggests that, similar to Hawkes processes where the asymptotic properties only depend on the norm n_H as soon as the kernel is short-ranged, the distribution (E.25) of the squared volatility should hold for any short-ranged kernel.

Another remark is that as soon as $n_Z \geq 1/3$, the variance σ_V^2 of the activity V explodes while its mean μ_V remains finite up to $n_Z \to 1^-$. Now, when fitting the time series generated by this process using a simple Hawkes process, one finds $n_H \approx 1 - \sqrt{\mu_V(W)/\sigma_V^2(W)}$ for a suitable choice of window size W (see Hardiman and Bouchaud (2014)). Therefore, the vanishing of the mean/variance ratio necessarily imposes that the fitted Hawkes process must be critical, i.e. $n_H = 1$! What we argue here is that this *apparent* criticality may in fact be induced by quadratic feedback effects, but does not necessarily imply that the true underlying process is critical.

Finally, note that in the diffusive limit where the price process satisfies the equation $\mathrm{d}\bar{P}_t^\infty = \sqrt{V_t}\mathrm{d}W_t$, the asymptotic stationary distribution for the returns is given by :

$$p(r) \underset{|r|\to\infty}{\sim} \frac{C'}{|r|^{1+\nu}}; \qquad \nu \equiv 1 + \frac{1}{n_Z}.$$

The fat-tail volatility that is generated by our model naturally produces a fat-tail distribution of instantaneous returns, with exponent ν for the cumulative distribution equal to $1 + 1/n_Z \geq 2$. The more endogenous, the fatter the tails for the returns : this interpretation seems intuitive. For a critical process, $n_Z = 1$, the tail is such that the volatility of the returns diverges. A tail exponent for the cumulative distribution $\nu \approx 3$ (the so-called "inverse cubic law", observed on a large universe of traded products) is obtained for $n_Z = 0.5$. Note however that the value of n_Z obtained above from calibrating the model is much smaller, $n_Z \approx 0.06$, leading to $\nu \approx 18$, far too large to explain the tail of financial returns. We will see now that, quite interestingly, the interaction with a non-critical Hawkes kernel can substantially reduce the value of ν.

ZHawkes with Hawkes

The case when $n_H > 0$ is more complicated but, remarkably, the tail exponent of the activity distribution $q(v)$ can still be analytically computed in some limits. The idea is to realize that when

$\bar{Z} \to \infty$, the distribution of $\bar{H}$ conditional to a certain large value of $\bar{Y} := \bar{Z}^2$ is of the form :

$$\Pi(\bar{H}|\bar{Y}) = \frac{1}{\bar{Y}} F\left(\frac{H}{\bar{Y}}\right) + o(\bar{Y}); \qquad (\bar{Y} \to \infty),$$

where $F(.)$ is a certain scaling function which obeys a differential equation derived in Appendix B. Correspondingly, one can show that the far-tail of the distribution of $V_t = \psi^2 \left[\lambda_\infty + \bar{Z}_t^2\right]$ is still a power-law, given by :

$$q(v) \underset{v \to +\infty}{\sim} C'' \, v^{-\left(\frac{3}{2} + \frac{1}{2n_Z(1+a^*)}\right)}, \tag{E.27}$$

where C'' is another constant and a^* is defined as :

$$a^* = \int_0^\infty \mathrm{d}x \; x \, F(x). \tag{E.28}$$

Introducing $\chi := \frac{2\overline{\omega}}{\overline{\beta}}$ as the ratio of the correlation time scale of the Hawkes process to the one of the ZHawkes process, a full solution for F can be found in the two limits $\chi \to 0$ and $\chi \to \infty$, allowing one to fix the value of a^*. One finds (see Appendix B) :

$$a^* \approx \frac{n_H}{1-n_H}\left[1 - \chi \frac{1-n_H-n_Z}{(1-n_H)^2}\right], \quad (\chi \to 0); \qquad a^* \approx \frac{n_H}{\chi(1-n_Z)}, \quad (\chi, \chi n_Z \to \infty). \tag{E.29}$$

Two other limiting cases can be exactly solved : one is when $n_H \to 0$, one finds that $a^* \approx \frac{n_H}{\chi(1-n_Z)}$ still holds provided $a^* \ll 1$, and the other is $n_Z \to 0$, for which we find an explicit expression for a^* as the solution of a second degree equation (see Appendix B).

The corresponding exponent for the asymptotic tail of the cumulative distribution of returns is now given by :

$$\nu = 1 + \frac{1}{n_Z(1+a^*)}, \tag{E.30}$$

with :

— for $n_H = 0$ (ZHawkes without Hawkes), one recovers the previous case where $a^* = 0$ and $\nu = 1 + 1/n_Z$.

— for $0 < n_H < 1$ and $\chi \to 0$ (Hawkes much "faster" than ZHawkes), the exponent ν is *decreased* to $\nu = 1 + (1-n_H)/n_Z + O(\chi)$.

— for $0 < n_H < 1$ and $\chi \to \infty$ (Hawkes much "slower" than ZHawkes), the exponent ν is again *decreased* from $\nu = 1 + 1/n_Z$ by an amount $\sim 1/\chi$.

— In the case $n_Z \to 0$, one finds $\nu = 1 + \frac{b}{n_Z}$, where b can be computed in terms of n_H and χ, see Appendix B.

The results of this section are, we believe, quite interesting. First, the two-dimensional limit process defined by Eqs. (E.22) leads to power-law tails for the volatility that can be exactly characterized in some limits. From a theoretical point of view, the possibility of computing exactly the tail

exponent in this model is potentially important if our ZHawkes process turned out to be a central ingredient to model the dynamics of financial markets. Second, we have found that although the Hawkes kernel per-se does not lead to power-law tails (i.e., $\nu \to \infty$ when $n_Z \to 0$), the Hawkes kernel actually "cooperates" with the ZHawkes kernel to make the tails of the distribution fatter. The case of empirical interest is $n_Z = 0.06$, $n_H \approx 0.8$ leads to $\nu = 1 + (1 - n_H)/n_Z \approx 4$ for $\chi \to 0$, which indeed remains in the experimental range for a non-Markovian ZHawkes process with parameters calibrated on intraday data, as will be shown by numerical simulations in the next section.

We find this phenomenon quite remarkable : whereas the Hawkes feedback alone is not able to explain fat-tails, only a relatively small amount of quadratic (Zumbach) feedback generates power-law tails in the correct range (remember that $n_Z = 0.06 \ll n_H$). Note however that this ZHawkes family of models leads a continuously varying exponent (as a function of the parameters) rather than a fixed, universal exponent like in many physical situations. This begs the question : is there any mechanism that would explain why the feedback parameters n_Z, n_H, χ lie in a rather restricted interval, such as to explain the apparent universality of the tail exponent of (mature) financial markets (on this particular point see e.g. (Plerou et al., 1999; Gopikrishnan et al., 1999; Di Matteo et al., 2005; Zumbach, 2015))

E.5 Calibration : A QHawkes model for intraday data

We now attempt to calibrate our QHawkes model on intraday data. The idea here is that we want to investigate how fat-tails and time reversal asymmetry emerges at relatively short time scales. Longer time scale effects, including the long memory nature of the volatility fluctuations, require some extra treatment to stitch together different days, i.e. a detailed procedure to treat overnight effects. We leave this crucial question for further work but will comment briefly on this issue in the conclusion. Our calibration strategy is based on the relation between the QHawkes model given by (E.2) and the discrete QARCH model introduced in Sentana (1995), and revisited in depth in Chicheportiche and Bouchaud (2014). This will give us a way to calibrate the QHawkes on discretely sampled price time series.

E.5.1 QHawkes as a limit of QARCH

In this section we investigate the link between the QHawkes model and the discrete QARCH model. For a fixed time step $\Delta > 0$, we define for all $t \in \mathbb{R}$:

— the price (or log-price) increment between time $t - \Delta$ and time t : $r_t^\Delta = P_t - P_{t-\Delta}$,

— the volatility at time t : $\sigma_t^\Delta = \sqrt{\mathbb{E}\left[{r_t^\Delta}^2 \middle| \mathcal{F}_{t-\Delta}\right]}$.

The QHawkes model appears as the limit (in some sense) when $\Delta \to 0^+$ of the QARCH model

$$\sigma_t^{\Delta\,2} = \sigma_\infty^{\Delta\,2} \; + \sum_{\tau \geq 1} L^\Delta(\tau)\; r_{t-\tau\Delta}^\Delta \; + \sum_{\tau,\tau' \geq 1} K^\Delta(\tau,\tau')\; r_{t-\tau\Delta}^\Delta r_{t-\tau'\Delta}^\Delta, \tag{E.31}$$

where $\sigma_\infty^{\Delta\,2} = \psi^2 \lambda_\infty \Delta, L^\Delta(\tau) = L(\tau\Delta)\;\Delta$ and $K^\Delta(\tau,\tau') = K(\tau\Delta,\tau'\Delta)\;\Delta$. Indeed, for $t \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}\left[{r_{t+\Delta}^\Delta}^2 \middle| \mathcal{F}_t\right] &= \psi^2\; \mathbb{P}\left(P_{t+\Delta} - P_t \neq 0 \middle| \mathcal{F}_t\right) \; + \; o(\Delta) \\ &= \psi^2\; \lambda_t\; \Delta \; + \; o(\Delta),\end{aligned}$$

which implies the scaling :

$$\frac{\sigma_t^{\Delta\,2}}{\Delta} \underset{\Delta \to 0^+}{\longrightarrow} \psi^2\; \lambda_t.$$

Thanks to this link between the two models, it is possible to calibrate a QARCH model on intra-day 5 minutes bin returns, as in Blanc (2012); Chicheportiche and Bouchaud (2014), and obtain some qualitative and quantitative insight on the structure of the underlying QHawkes model. Indeed, the direct calibration of the latter would be significantly harder – more noisy and computationally more demanding – and certainly beyond the scope of the present paper. Since fitting a model with another might appear somewhat unjustified in spite of the above theoretical arguments, we validate our procedure numerically in appendix E.7, where we simulate a QHawkes process with realistic parameters.

E.5.2 Intra-day calibration of a QARCH model

QARCH models have mainly been calibrated on daily data so far (Sentana, 1995; Chicheportiche and Bouchaud, 2014). To give a starting point to our study of quadratic effects in high-frequency volatility, we calibrate a discrete QARCH on intra-day five-minute returns.

Dataset and notations

We consider the same dataset as in Allez and Bouchaud (2011), which is composed of stock prices on intra-day five-minute bins. It includes 133 stocks of the New York Stock Exchange, that have been traded without interruption between 1 January 2000 and 31 December 2009. This yields 2499 trading days, with 78 five-minute bins per day. For each bin, the open, close, high and low prices $(O, C, H, L > 0)$ are available. We consider the log-price process and define on each bin :

— The return $r = \ln(C/O)$.

— The Rogers-Satchell volatility $\sigma^{\mathrm{RS}} = \sqrt{\ln(H/O) \times \ln(H/C) + \ln(L/O) \times \ln(L/C)}$.

Normalization procedure

To be able to consider that intra-day prices are (approximately) independent realizations of a stationary stochastic process, we need to normalize the data carefully. As a matter of fact, strong intra-day seasonalities may corrupt the calibration results. This can be avoided to some extent through a cross-sectional and historical normalization. We take advantage of our large dataset to compute a cross-sectional intra-day volatility pattern for each trading day and we normalize the returns by this pattern, which dampens the effect of collective shocks on a given day. On the other hand, we use the intra-day/overnight model volatility model of Blanc et al. (2014) to factor out daily feedback effects and focus on pure intra-day dynamics. To fully explain our normalization protocol, we introduce the following notations :

— The 5-min bin index $1 \leq b \leq 78$, the day index $1 \leq t \leq 2499$ and the stock index $1 \leq u \leq 133$.
— The empirical averages : $\langle x_{u,t,.} \rangle$ means conditional average of x over bins, for stock u and day t fixed ; $\langle x_{u,.,b} \rangle$ and $\langle x_{.,t,b} \rangle$ are defined similarly as the conditional averages over days/stocks ; $\langle x \rangle = \langle x_{.,.,.} \rangle$ means average of x over stocks, days and bins.

We compute the cross-sectional volatility pattern of day t, that we use to normalize the data of stock u, as :

$$b \in \{1, \cdots, 78\} \;\mapsto\; v_{u,t}(b) \equiv \sqrt{\langle r^2_{u' \neq u,t,b} \rangle}.$$

For stock u, the value $r^2_{u,t,b}$ is excluded from the average, so that the normalization protocol does not cap the large returns of stock u artificially. We also consider the open-to-close volatility $\sigma^{\mathrm{D}}_{u,t}$ of day t for stock u, as computed by the intra-day/overnight model of Blanc et al. (2014) with the data of stock u over the days $\{1, \cdots, t-1\}$. For $t = 1$, we fix $\sigma^{\mathrm{D}}_{u,1} = 1$.

The normalization protocol is as follows (here the operator $\leftarrow$ should be understood like an assignment in a programming language) : $\forall u, t, b$,

— $r_{u,t,b} \leftarrow r_{u,t,b}/\sigma^{\mathrm{D}}_{u,t}, \quad \sigma^{\mathrm{RS}}_{u,t,b} \leftarrow \sigma^{\mathrm{RS}}_{u,t,b}/\sigma^{\mathrm{D}}_{u,t}, \qquad$ (normalization by open-to-close volatility)
— $r_{u,t,b} \leftarrow r_{u,t,b}/v_{u,t}(b), \quad \sigma^{\mathrm{RS}}_{u,t,b} \leftarrow \sigma^{\mathrm{RS}}_{u,t,b}/v_{u,t}(b). \qquad$ (cross-sectional intra-day normalization)

We further exclude trading days that involve at least one bin where the absolute return is greater than the average plus six standard deviations. This represents approximately 7% of trading days, i.e. one day every three weeks. Combined with the cross-sectional pattern normalization, this data treatment strongly dampens the impacts of exceptional news events, which would require a special treatment and that we do not aim to model here. Eventually, we set the mean of the squares to one and the average return to zero to make the stock universe more homogeneous : $\forall u, t, b$,

— $r_{u,t,b} \leftarrow r_{u,t,b} - \langle r_{u,.,.} \rangle$ so that $\langle r \rangle = 0$,
— $r_{u,t,b} \leftarrow r_{u,t,b}/\sqrt{\langle r^2_{u,.,.} \rangle}$, so that $\langle r^2 \rangle = 1$,
— $\sigma^{\mathrm{RS}}_{u,t,b} \leftarrow \sigma^{\mathrm{RS}}_{u,t,b}/\sqrt{\langle {\sigma^{\mathrm{RS}}}^2_{u,.,.} \rangle}$, so that $\langle {\sigma^{\mathrm{RS}}}^2 \rangle = 1$.

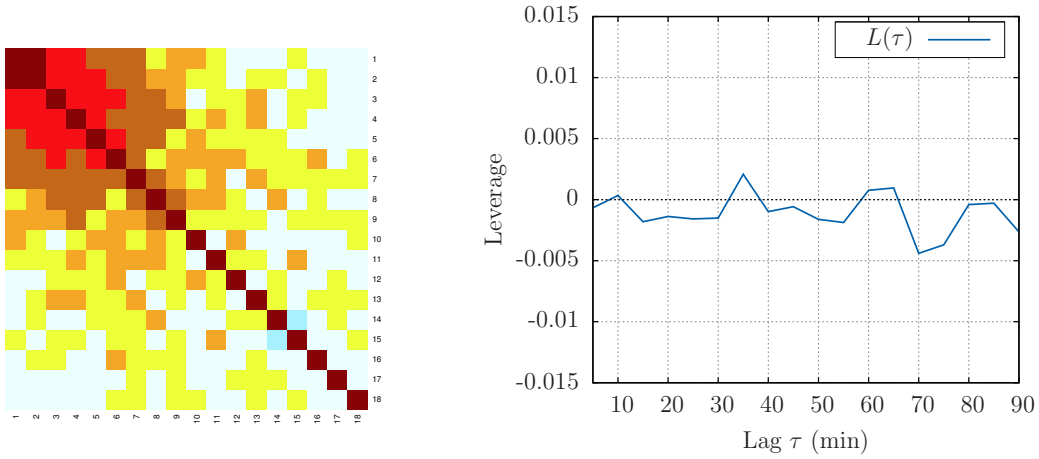

FIGURE E.1 – QARCH kernels calibrated on five-minute intra-day returns for US stocks. The maximum lag is 18 bins, i.e. one hour and a half of trading time. Left : heatmap of the quadratic kernel. White coefficients are close to zero, blue ones are negative and yellow/orange/red ones are positive, with darker shades as they increase in absolute value. We see that all the significant coefficients are positive, with a non-negligible off-diagonal component. Right : leverage kernel. It is hardly distinct from zero and can be considered as pure noise (as opposed to daily models where it is significantly negative).

Calibration results

The calibration process is similar to Chicheportiche and Bouchaud (2014) and Blanc et al. (2014). A first estimate of the kernels is obtained with the Generalized Method of Moments, which uses a set of correlation functions that are empirically observable. Then, using this estimate as a starting point, we use Maximum Likelihood Estimation, assuming that the residuals are t-distributed (which accounts for fat tails that remain in the residuals). This second step significantly improves the precision of the calibration results, compared to a solo GMM estimation.

We find it worth to notice that as opposed to the daily calibration results of Chicheportiche and Bouchaud (2014), a clear off-diagonal structure appears in the feedback matrix in the intra-day case (see Figure E.1). Also, the intra-day leverage kernel is found to be close to zero, justifying the fact that we mainly consider $L \equiv 0$ throughout the paper. The spectral decomposition of quadratic kernel (see Figure E.2) suggests that K is the superposition of a positive rank-one matrix and a diagonal one. Indeed, we obtain to a good approximation (see Figure E.3)

$$K(\tau, \tau') \approx \phi(\tau)\delta_{\tau-\tau'} + k(\tau)k(\tau')$$

where

$$\phi(\tau) = g\tau^{-\alpha} \quad , \quad k(\tau) = k_0 \exp(-\omega\tau),$$

with $g = 0.09,\ \alpha = 0.60,\ k_0 = 0.14,\ \omega = 0.15$. Note that $\omega = 0.15$ corresponds to a characteristic

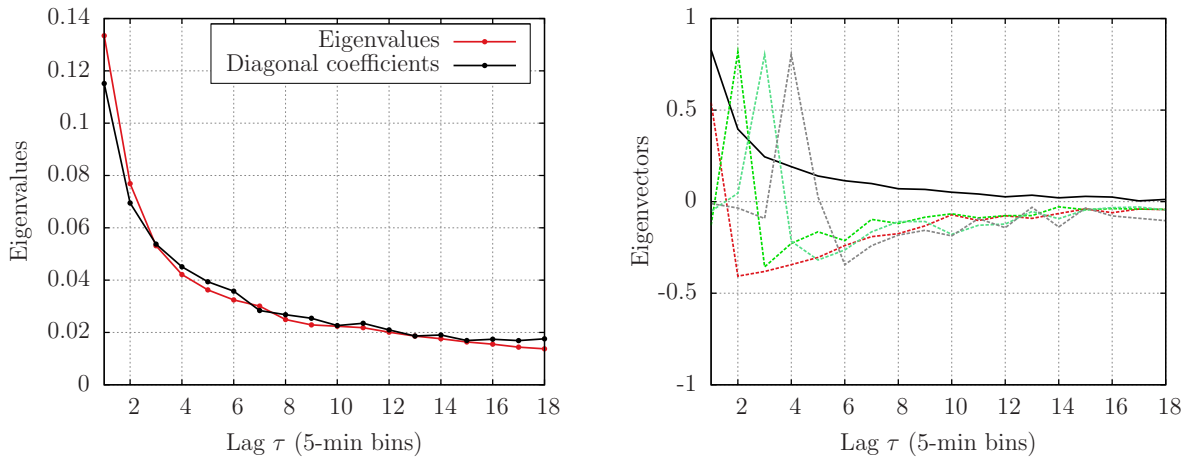

FIGURE E.2 – Spectral decomposition of the quadratic QARCH kernel. Left : ranked eigenvalues (plain dark line) and diagonal coefficients (dashed). One can see that the diagonal coefficients are very close to the eigenvalues, except for the first eigenvalue which is significantly larger than the maximum of the diagonal. Right : eigenvectors corresponding to the five largest eigenvalues. The first eigenvector (plain dark line) is a positive decaying kernel, the others are close to the canonical vectors $e_i(\tau) = \delta_{i-\tau}$.

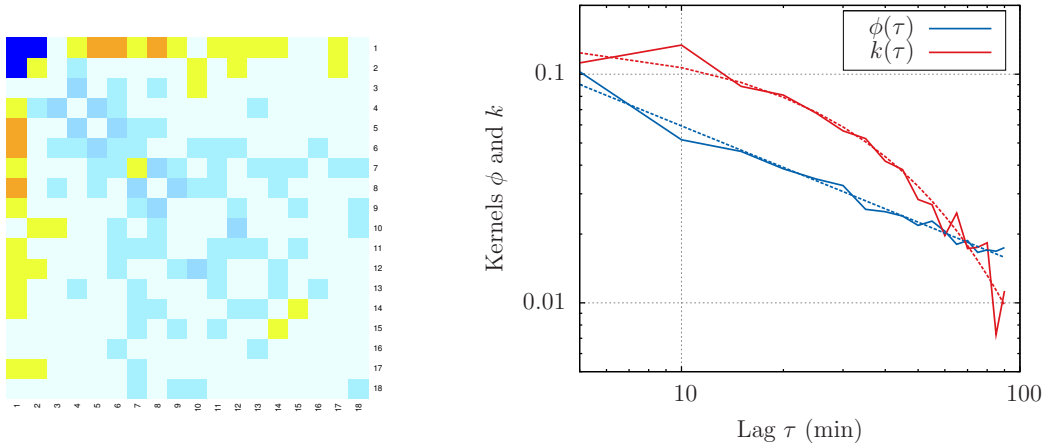

FIGURE E.3 – Fit of the kernel K by the sum of a power-law diagonal matrix and an exponential rank-one matrix. Left : heatmap of the difference between the fitted matrix and the original one. The coefficients are small (white or lightly-colored) except for the upper-left corner : the original matrix features a stronger short-term feedback. Right : kernels $\phi(\tau)$ and $k(\tau)$ that minimize the matrix distance $\sum[K(\tau,\tau') - \phi(\tau)\delta_{\tau-\tau'} - k(\tau)k(\tau')]^2$. The rank-one kernel k is plotted in red (and is larger for small τ's), and the diagonal kernel ϕ is plotted in blue, both in log-log scale. The dashed lines are the power-law fit for $\phi(\tau)$ with exponent $\alpha = 0.6$, and the exponential fit for $k(\tau)$ with characteristic time about 30 min.

time of about thirty minutes (bin size $\times\omega^{-1}$) for the decay of the off-diagonal component. We then fix the off-diagonal part of the kernel K to its fitted value $k(\tau)k(\tau') = k_0^2 \exp(-\omega(\tau+\tau'))$, and we recalibrate the diagonal of K with a longer maximum lag of 60 bins (five hours of trading). We

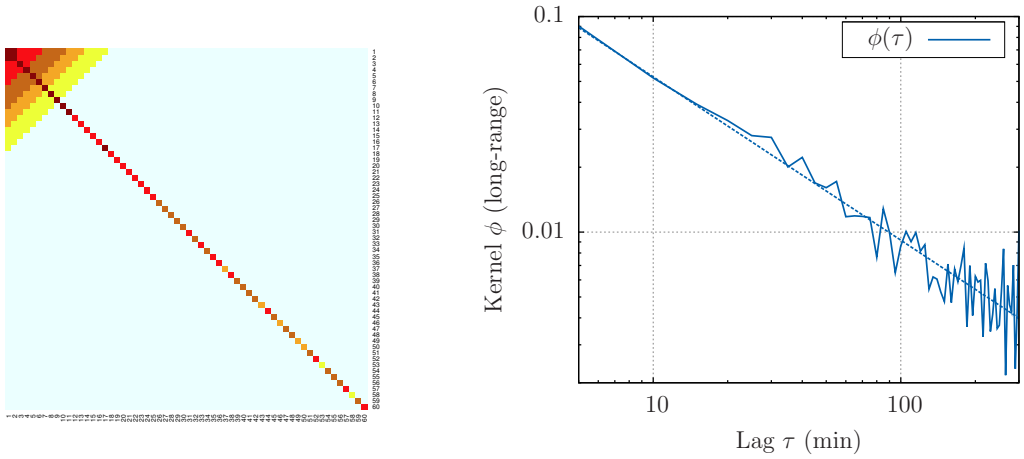

FIGURE E.4 – Long-range kernel K. Left : heatmap of the long-range kernel, with the off-diagonal fixed as its exponential rank-one fit, and with the diagonal calibrated with no constraints. Right : Hawkes kernel $\phi(\tau) = K(\tau, \tau) - k^2(\tau)$ fitted on 60 bins. The kernel $\phi(\tau)$ is plotted in log-log scale, with its power-law fit with exponent $\alpha' = 0.76$ (dashed).

obtain

$$\phi(\tau) = g'\tau^{-\alpha'}$$

with the new coefficients $g' = 0.09$, $\alpha' = 0.76$, not far from those obtained above on a shorter time span. From the numerical experiment conducted in appendix E.7, we however expect α' to underestimate the true power-law decay exponent by $\sim 30\%$.

The residuals ξ_t of the QARCH model, defined by

$$r_t = \sigma_t \xi_t,$$

where r_t is the five-minute return and σ_t is the QARCH volatility, are modeled with Student's t-distribution. The calibration of the model with $K(\tau, \tau') = \phi(\tau)\delta_{\tau-\tau'} + k(\tau)k(\tau')$ yields $\nu \approx 7.9$ degrees of freedom for the residuals, which gives a kurtosis $\kappa \approx 4.5$. This has to be compared with the tail exponent ν_r of r_t itself, which is, as is well known, in the range $3 \to 4$, see also Figure E.5 below. Since ν is more than twice ν_r, the QARCH model with Gaussian residuals and this specific form of K accounts, to a good extent, for the fat tails of five-minute returns, that appear to be nearly entirely induced by the quadratic feedback mechanism. We justify theoretically and numerically why this is the case in Sections E.4 and E.6.

In the QARCH model, the endogeneity ratio of the volatility (i.e. the proportion of the volatility that stems from feedback effects) is given by the trace $\mathrm{Tr}(K)$ of the quadratic kernel. With our

parametrization and a maximum lag of $q \geq 1$, one has

$$\mathrm{Tr}(K) = \sum_{\tau=1}^{q} \phi(\tau) + \sum_{\tau=1}^{q} k^2(\tau).$$

We use the fits $k(\tau) = k_0 \exp(-\omega\tau)$ and $\phi_{\mathrm{lr}}(\tau) = g'\tau^{-\alpha'}$ to compute $\mathrm{Tr}(K)$ for $q = 78$, which is the total number of five-minute bins in a trading day. We obtain

$$\sum_{\tau=1}^{q} \phi(\tau) \simeq 0.74, \quad \sum_{\tau=1}^{q} k^2(\tau) \simeq 0.06 \quad \Rightarrow \quad \mathrm{Tr}(K) \simeq 0.80.$$

This endogeneity ratio implies that 80% of the intra-day volatility is due to short term endogenous feedback effects. Interestingly, it is close to the value obtained for QARCH and ARCH models at a daily time scale, see Chicheportiche and Bouchaud (2014) and Blanc et al. (2014). Note that although high, the endogeneity ratio is significantly below the critical limit $\mathrm{Tr}(K) = 1$, which is the value found by calibrating a standard linear Hawkes process to activity data on much longer time horizons : Hardiman et al. (2013); Hardiman and Bouchaud (2014) report $n_H \approx 0.9$ on a time window of a few hours, and $n_H \approx 0.99$ when the kernel is extended to 40 days. In fact, as emphasized in Hardiman and Bouchaud (2014), for a linear Hawkes process to exhibit long-memory requires criticality, i.e. $n_H \to 1$. As we have seen in Section E.3, this is not necessarily true for QHawkes, but long-memory still requires a power-law decay of the off-diagonal kernel, at variance with the exponential decay suggested by our intraday calibration. Whether the QHawkes also needs to be critical to account for the low frequency properties of financial markets would require a calibration on daily data, an issue we leave for further investigations.

E.6 Numerical simulation results

E.6.1 Empirical tails of the volatility process

In this section, we compare numerically the volatility process generated by the ZHawkes model, with a standard Hawkes-based price model and with the financial data studied in Section E.5.2. We simulate a ZHawkes model with an exponential Zumbach part and a power-law Hawkes part, with parameters inspired by the QARCH calibration of Section E.5.2 : for t expressed in minutes,

$$\phi(t) = 0.0016 \times (1 + 0.01 \times t)^{-1.2}, \quad k(t) = 0.003 \times \exp(-0.03 \times t),$$

so that $n_H = 0.8$, $n_Z = 0.1$ and $\mathrm{Tr}(K) = 0.9$. Note that to simulate a stationary ZHawkes model, we choose a decay exponent α above 1 for ϕ, although the QARCH calibration suggests a slower decay for t corresponding to intraday time scales. However, as noted in appendix E.7, the exponent

$\alpha \approx 0.75$ is probably underestimated by our calibration procedure. In any case, choosing $\alpha > 1$ is the simplest way to enforce stationarity without having to introduce a more complicated functional form for $\phi(t)$ that would model overnight effects and daily time scales. As a benchmark, we also simulate a standard Hawkes-based price process ($n_Z \equiv 0$) with $\phi = (1 + 0.01 \times t)^{-1.3}$, $n_H = 0.99$, which is close to the calibration results of Hardiman et al. (2013).

It is important to note that to simulate the ZHawkes and the Hawkes model, we choose constant price jumps $\Delta P_\tau = \pm\psi$. Therefore, our numerical results for the distribution of the volatility can by no means be attributed to the fat-tails of individual price jumps. For both simulated and real data, we consider the Rogers-Satchell volatility times series for five-minute bins. We use the Hill exponent (Hill, 1975) as an estimator of the empirical tail exponent of the volatility

$$\nu_{\text{hill}} = 1 + \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \log(\sigma_i/\sigma_{\min})}$$

where $\sigma_{\min}$ is some cutoff and $\sigma_i \geq \sigma_{\min}$ are the volatilities in the far tail region of the distribution. One obtains $\nu_{\text{hill}} = 4.50$ for the (normalized) five minutes returns of US stocks (in agreement with many previous determinations of this exponent), $\nu_{\text{hill}} = 5.07$ for the ZHawkes model and $\nu_{\text{hill}} = 12.4$ for the standard Hawkes-based model without ZHawkes feedback. Even with a norm close to one and a slowly-decaying kernel, the standard Hawkes model cannot reproduce the tails observed on US stock data. Instead, the ZHawkes model, with a norm strictly below unity and a short-lived Zumbach effect, naturally produces fat tails very similar to those observed empirically, even with a rather small n_Z. These observations are illustrated by Figures E.5 and E.6.

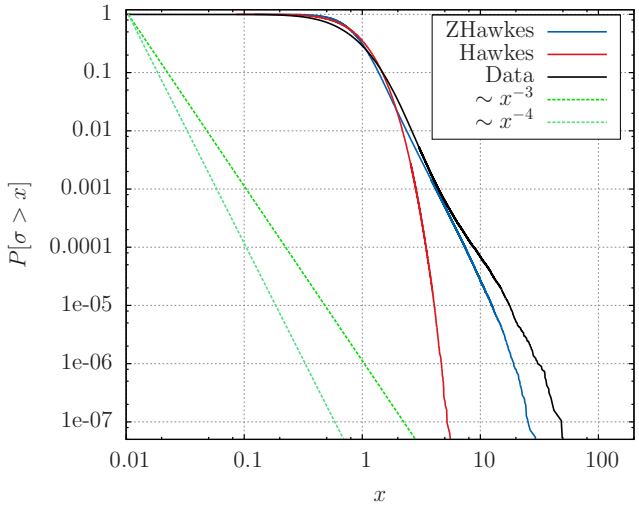

FIGURE E.5 – Cumulative density function of the Rogers-Satchell volatility for US stock data (plain line), simulated Hawkes data (red line), and simulated ZHawkes data (blue line). Notice how well the empirical distribution function is reproduced by the ZHawkes model, calibrated as in Section E.5.2.

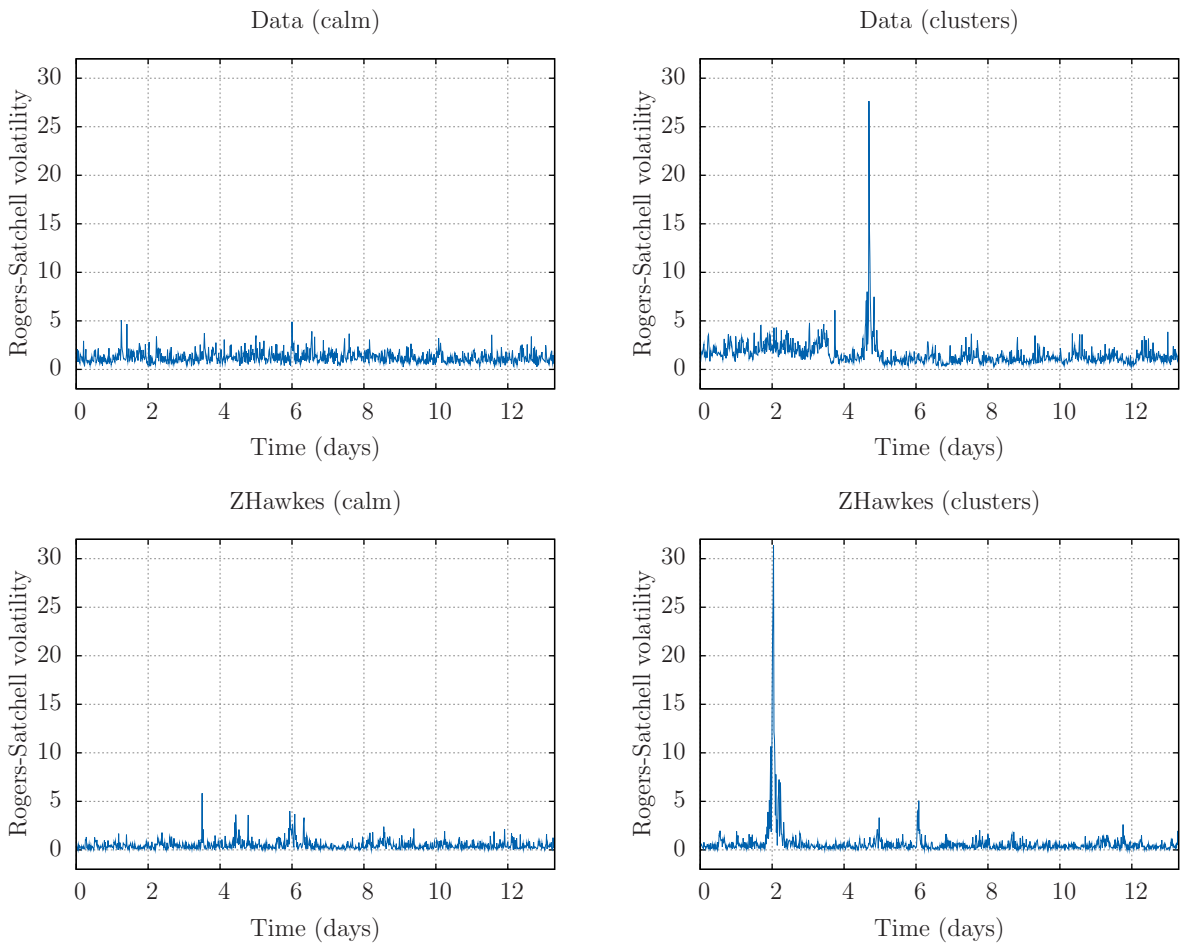

FIGURE E.6 – Time series of Rogers-Satchell volatility. Above : real data ; below : simulated ZHawkes data ; left : period of calm ; right : cluster of intense activity.

E.6.2 Time-reversal asymmetry of ZHawkes processes

Another salient feature of financial markets is, as discussed in the introduction, the time-reversal asymmetry (TRA) of the price time series. The authors of Chicheportiche and Bouchaud (2014) study this feature for stock data on the one hand, and for a simulated FIGARCH volatility process on the other. The chosen observable is the cross-correlation of present Rogers-Satchell volatilities σ_t^2 with past squared returns $r_{t-\tau}^2$, to that of present squared returns with past volatilities, which is found to be such that $\langle r_{t-\tau}^2 \sigma_t^2 \rangle_t > \langle r_t^2 \sigma_{t-\tau}^2 \rangle_t$ for $\tau > 0$, both for real data and FIGARCH processes.

This observation is one of the main motivations for the model introduced in the present paper, since standard models that use Brownian SDEs are TRS by construction and cannot reproduce this asymmetry. In this section, we measure the amount of TRA for the simulated ZHawkes process and for the Hawkes benchmark described in the previous section, and for the financial dataset studied in Section E.5.2.

As in Sections E.5.2, we consider the returns and the Rogers-Satchell volatilities defined for

intra-day five-minute bins. Here, the maximum lag q is fixed to 36 (36 bins of 5 minutes = 3 hours of trading) and the lag index τ varies between 1 and q. We introduce

— The cross-correlation function of the Rogers-Satchell volatility and absolute returns

$$C(\tau) = \frac{\langle \sigma_t^{\mathrm{RS}} \times |r_{t-\tau}| \rangle - \langle \sigma^{\mathrm{RS}} \rangle \langle |r| \rangle}{\sqrt{\langle \sigma^{\mathrm{RS}2} \rangle - \langle \sigma^{\mathrm{RS}} \rangle^2} \sqrt{\langle r^2 \rangle - \langle |r| \rangle^2}}.$$

— The time asymmetry ratio

$$\Delta(\tau) = \frac{\sum_{\tau'=1}^{\tau} [C(\tau') - C(-\tau')]}{2 \sum_{\tau'=1}^{q} \max(|C(\tau')|, |C(-\tau')|)} \in [-1, 1].$$

Note that we choose to compute the cross-correlation function using the *absolute* returns instead of the *squared* returns, since it yields results that are less noisy and more robust to tail events (and thus less sensitive to the normalization method).

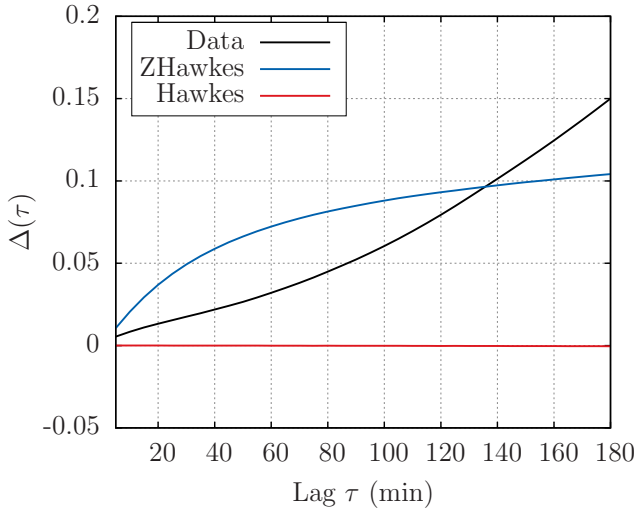

FIGURE E.7 – Time asymmetry ratio $\Delta(\tau)$ for US stock data (plain line), simulated Hawkes model (red line), and simulated ZHawkes model (blue dot-dashed line). Note that the Hawkes process does not generate any detectable TRA.

We compare the time asymmetry ratios $\Delta(\tau)$ for real stock returns, returns simulated with the ZHawkes model and returns simulated with a standard Hawkes-based price model. The results are illustrated by Figure E.7. The standard Hawkes model, perhaps surprisingly, does not generate any detectable TRA : $|\Delta(\tau)| < 10^{-3}$ for all τ. Thus it is clear that the Hawkes model with no off-diagonal quadratic feedback cannot reproduce the time asymmetry observed in intra-day volatility, for which $\Delta(\tau)$ is one hundred times larger. On the other hand, the ZHawkes model with parameters in line with the QARCH calibration of Section E.5 features some time asymmetry, which is not

only of the correct sign but also reproduces the right order of magnitude, without any further parameter adjustment. However, the function $\tau \mapsto \Delta(\tau)$ is found to be concave for the ZHawkes model (as expected on general grounds) and, strangely, convex for stock data. Even with a thorough normalization protocol, intra-day returns are not rigorously stationary, and we believe that the convexity of $\tau \mapsto \Delta(\tau)$ observed on real data is spurious, as it should should saturate to a value less than 1 beyond some time scale. Such convexity would probably be hard to reproduce with a simple model, unless it some non-stationary is added by hand.

E.7 Conclusion

The central message of our study is that the standard Hawkes feedback, where past activity increase the intensity of the current activity, fails at accounting for two essential features of the dynamics of markets : a) the fat-tails in the activity/volatility cannot be reproduced and b) the time-reversal asymmetry between past daily volatilities and future intraday volatilities or vice-versa is completely absent within the Hawkes framework. This was not a priori obvious, since Hawkes processes are constructed on the idea of a feedback from the past. We have thus proposed QHawkes processes as simple, intuitive generalisations of the Hawkes process which posit that the feedback is in fact not only on the past activity, but on past price returns themselves.

A QHawkes model can be seen as a consistent definition of a Quadratic ARCH (QARCH) model as a continuous-time point-processes. This in fact allowed us to calibrate a QHawkes model on the intraday returns of 133 NYSE stocks. We find that the matrix kernel of the QHawkes has a diagonal part (corresponding to the standard Hawkes component) and a off-diagonal, rank-one part that we call "ZHawkes". It corresponds to Zumbach's insight that local trends in the price, both up or down, generate more future activity (Zumbach, 2010). ZHawkes processes have some interesting properties that standard Hawkes processes lack, namely : (i) the quadratic feedback naturally produces a multiplicative dynamics for the volatility, generating power-law tails for the volatility and the returns even even when elementary price changes are strictly bounded ; (ii) it reproduces a level of time-reversal asymmetry (TRA) that is fully compatible with what is measured on actual financial data ; (iii) it can in principle generate long memory without necessarily be at its critical point, although this last point would require calibration on daily data to be further discussed.

The continuous limit SDE corresponding to exponential kernels was found to be a tractable two-dimensional generalization of Pearson diffusions. In particular the tail exponent of the volatility can be exactly computed in several cases and, quite remarkably, fall within the empirical range even when the ZHawkes kernel is of small amplitude. These mathematically tractable diffusions are reminiscent of the log-normal volatility processes considered in Stein and Stein (1991); Bacry et al. (2001); Bergomi (2005) and more recently Gatheral et al. (2014), and provide a natural "microscopic"

mechanism for a multiplicative process for the volatility itself, which up to now has remained quite a mysterious hypothesis (Jaisson and Rosenbaum, 2015b).

We hope our paper motivates more developments on the family of QHawkes models. We have indeed only touched upon the mathematical properties and the empirical relevance of such models but we believe that deeper work on the subject would be valuable, in particular concerning the precise calibration of the model itself. As stated above, a completely open question at this stage is the treatment of overnights and the generalisation of the model to describe longer time scales (our calibration was restricted to intraday data), generalizing the QARCH description proposed by two of us in Blanc et al. (2014). In particular, we know that time-reversal asymmetry can still be detected on time scales of days or weeks (Zumbach and Lynch, 2001; Chicheportiche and Bouchaud, 2014) and this can certainly not be reproduced with a ZHawkes kernel decaying over 30 minutes, as found here. Similarly, multiplicative log-normal models for the volatility have commonly been considered for daily returns. How much is the fat-tailed, long memory of the volatility, recently described within the context of standard Hawkes process, should in fact be traced to the QHawkes mechanism proposed here is, in our opinion, a very interesting question for future research.

To conclude, we believe that a comprehensive understanding of the volatility process, from the scale of the event up to macroscopic scales, would seem very valuable in several respects, in particular that of market design. One would perhaps understand how a change in market microstructural rules (e.g. the tick size) may affect its macroscopic properties (e.g. volatility). Finding a solid, behavioural microscopic foundations to the volatility process seems crucial : when fully understood, simple constraints on the agents might then change the overall, macroscopic market behaviour. We hope that our generalized Hawkes process could provide some clues on this issue.

Acknowledgments

We want to thank R. Chicheportiche, J. Gatheral, S. Hardiman, Th. Jaisson, I. Mastromatteo and M. Rosenbaum for many insightful discussions on these issues. We also thank the referees whose remarks helped improve the quality of our manuscript.

Exact equations relating the kernel and the auto-correlation functions

To simplify notations, we write (in this appendix only) $\varphi(t) = K(t,t)$.

For $s < t$, one has $\mathcal{C}(t-s) = \lambda_\infty \overline{\lambda} - \overline{\lambda}^2 + \mathbb{E}\left[A_t \frac{\mathrm{d}N_s}{\mathrm{d}s}\right] + 2\mathbb{E}\left[M_t \frac{\mathrm{d}N_s}{\mathrm{d}s}\right]$.

$$\mathbb{E}\left[A_t \frac{\mathrm{d}N_s}{\mathrm{d}s}\right] = \int_{-\infty}^{t} \varphi(t-u)\mathbb{E}\left[\frac{\mathrm{d}N_u}{\mathrm{d}u}\frac{\mathrm{d}N_s}{\mathrm{d}s}\right]\mathrm{d}u.$$

For $u \neq s$, $\mathbb{E}\left[\frac{\mathrm{d}N_u}{\mathrm{d}u}\frac{\mathrm{d}N_s}{\mathrm{d}s}\right]\mathrm{d}u = [\mathcal{C}(u-s)+\overline{\lambda}^2]\mathrm{d}u$, and for $u = s$, $\mathbb{E}\left[\left(\frac{\mathrm{d}N_u}{\mathrm{d}u}\right)^2\right]\mathrm{d}u = \kappa\mathbb{E}\left[\frac{\mathrm{d}N_u}{(\mathrm{d}u)^2}\right]\mathrm{d}u = \kappa\overline{\lambda}$, where κ is the kurtosis of the law μ of the jumps of P ($\kappa = 1$ if $\Delta P_\tau = \pm\psi$). Thus,

$$\mathbb{E}\left[A_t\frac{\mathrm{d}N_s}{\mathrm{d}s}\right] = \mathrm{Tr}(K)\overline{\lambda}^2 + \kappa\overline{\lambda}\varphi(t-s) + \int_{-\infty}^{t}\varphi(t-u)\mathcal{C}(u-s)\mathrm{d}u.$$

On the other hand,

$$\begin{aligned}\mathbb{E}\left[M_t\frac{\mathrm{d}N_s}{\mathrm{d}s}\right] &= \frac{1}{\psi^2}\int_{-\infty}^{t}\mathbb{E}\left[\Theta_{t,u}\frac{\mathrm{d}P_u}{\mathrm{d}u}\frac{\mathrm{d}N_s}{\mathrm{d}s}\right]\mathrm{d}u\\ &= \frac{1}{\psi^2}\int_{-\infty}^{t}\int_{-\infty}^{u-}K(t-u,t-r)\mathbb{E}\left[\frac{\mathrm{d}N_s}{\mathrm{d}s}\frac{\mathrm{d}P_u}{\mathrm{d}u}\frac{\mathrm{d}P_r}{\mathrm{d}r}\right]\mathrm{d}r\mathrm{d}u\\ &= \int_{-\infty}^{s-}\int_{-\infty}^{u-}K(t-u,t-r)\mathcal{D}(s-u,s-r)\mathrm{d}r\mathrm{d}u,\end{aligned}$$

since ΔP_τ and $(\Delta P_\tau)^3$ are centered, which implies that $\mathbb{E}\left[\frac{\mathrm{d}N_s}{\mathrm{d}s}\frac{\mathrm{d}P_u}{\mathrm{d}u}\frac{\mathrm{d}P_r}{\mathrm{d}r}\right] = 0$ for $u \geq s$. Taking $t = \tau > 0$ and $s = 0$, we obtain

$$\mathcal{C}(\tau) = \kappa\overline{\lambda}\varphi(\tau) + \int_{-\infty}^{\tau}\varphi(\tau-u)\mathcal{C}(u)\mathrm{d}u + 2\int_{0+}^{\infty}\int_{u+}^{\infty}K(\tau+u,\tau+r)\mathcal{D}(u,r)\mathrm{d}r\mathrm{d}u.$$

For $t > t_1 > t_2$, one has $\mathcal{D}(t-t_1,t-t_2) = \frac{1}{\psi^2}\mathbb{E}\left[A_t\frac{\mathrm{d}P_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\right] + \frac{2}{\psi^2}\mathbb{E}\left[M_t\frac{\mathrm{d}P_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\right]$. The first term gives

$$\begin{aligned}\frac{1}{\psi^2}\mathbb{E}\left[A_t\frac{\mathrm{d}P_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\right] &= \frac{1}{\psi^2}\int_{-\infty}^{t}\varphi(t-u)\mathbb{E}\left[\frac{\mathrm{d}N_u}{\mathrm{d}u}\frac{\mathrm{d}P_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\right]\mathrm{d}u\\ &= \int_{t_1+}^{t}\varphi(t-u)\mathcal{D}(u-t_1,u-t_2)\mathrm{d}u.\end{aligned}$$

The second term is given by

$$\frac{1}{\psi^2}\mathbb{E}\left[M_t\frac{\mathrm{d}P_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\right] = \frac{1}{\psi^4}\int_{-\infty}^{t}\int_{-\infty}^{u-}K(t-u,t-r)\mathbb{E}\left[\frac{\mathrm{d}P_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\frac{\mathrm{d}P_u}{\mathrm{d}u}\frac{\mathrm{d}P_r}{\mathrm{d}r}\right]\mathrm{d}r\mathrm{d}u$$

Since $r < u$ in the integral and $t_2 < t_1$, the expected value is zero if $u \neq t_1$. For $u = t_1$, we have $\mathbb{E}\left[\left(\frac{\mathrm{d}P_u}{\mathrm{d}u}\right)^2\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\frac{\mathrm{d}P_r}{\mathrm{d}r}\right]\mathrm{d}u = \psi^2\mathbb{E}\left[\frac{\mathrm{d}N_u}{(\mathrm{d}u)^2}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\frac{\mathrm{d}P_r}{\mathrm{d}r}\right]\mathrm{d}u = \psi^2\mathbb{E}\left[\frac{\mathrm{d}N_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\frac{\mathrm{d}P_r}{\mathrm{d}r}\right]$. Thus,

$$\mathbb{E}\left[M_t\frac{\mathrm{d}P_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\right] = \frac{1}{\psi^2}\int_{-\infty}^{t_1-}K(t-t_1,t-r)\mathbb{E}\left[\frac{\mathrm{d}N_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\frac{\mathrm{d}P_r}{\mathrm{d}r}\right]\mathrm{d}r.$$

For $r \neq t_2$, one has $\frac{1}{\psi^2}\mathbb{E}\left[\frac{\mathrm{d}N_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\frac{\mathrm{d}P_r}{\mathrm{d}r}\right]\mathrm{d}r = \mathcal{D}(t_1-t_2,t_1-r)\mathrm{d}r$. On the other hand $r = t_2$ yields

$\mathbb{E}\left[\frac{\mathrm{d}N_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}N_r}{(\mathrm{d}r)^2}\right]\mathrm{d}r = \mathbb{E}\left[\frac{\mathrm{d}N_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}N_{t_2}}{\mathrm{d}t_2}\right] = \mathcal{C}(t_1-t_2)+\overline{\lambda}^2$. We obtain

$$\mathbb{E}\left[M_t\frac{\mathrm{d}P_{t_1}}{\mathrm{d}t_1}\frac{\mathrm{d}P_{t_2}}{\mathrm{d}t_2}\right] = K(t-t_1,t-t_2)[\mathcal{C}(t_1-t_2)+\overline{\lambda}^2] + \int_{-\infty}^{t_1-} K(t-t_1,t-r)\mathcal{D}(t_1-t_2,t_1-r)\mathrm{d}r.$$

We eventually obtain by taking $\tau_2 = t > \tau_1 = t-t_1, t_2 = 0$,

$$\mathcal{D}(\tau_1,\tau_2) = 2K(\tau_1,\tau_2)[\mathcal{C}(\tau_2-\tau_1)+\overline{\lambda}^2] + \int_{(\tau_2-\tau_1)+}^{\tau_2}\varphi(\tau_2-u)\mathcal{D}(u-\tau_2+\tau_1,u)\mathrm{d}u + 2\int_{-\infty}^{(\tau_2-\tau_1)-} K(\tau_1,\tau_2-u)\mathcal{D}(\tau_2-\tau_1,\tau_2-\tau_1-u)\mathrm{d}u.$$

Asymptotic analysis of the Hawkes + ZHawkes process

In order to analyze the coupled Hawkes + ZHawkes processes, we first write the Fokker-Planck equation for the joint probability $\Pi(h,y)$ of $h \equiv \bar{H}$ and $y \equiv \bar{Y} = \bar{Z}^2$. Setting $t \leftarrow \beta t$ as the new time, we find :

$$\begin{aligned}\frac{\partial\Pi}{\partial t} &= -\frac{\partial}{\partial h}\left\{[-(1-n_H)h + n_H(\lambda_\infty+y)]\,\Pi\right\} \\ &\quad -\chi\frac{\partial}{\partial y}\left\{[(n_Z-1)y + n_Z(\lambda_\infty+h)]\,\Pi\right\} + 2\chi n_Z\frac{\partial^2}{\partial y^2}\left\{[y(\lambda_\infty+h+y)]\,\Pi\right\}\end{aligned} \tag{32}$$

We will study the stationary distribution of the process, such that the left-hand side of the above equation is zero. We introduce the conditional distribution of h for a given y, $\Pi(h|y)$, and the marginal distribution of y, $\pi(y)$, as :

$$\pi(y) := \int_0^\infty \mathrm{d}h\,\Pi(h,y); \qquad \Pi(h|y) = \frac{\Pi(h,y)}{\pi(y)}, \tag{33}$$

and the generating function of $\Pi(h|y)$, as :

$$Z(z|y) = \int_0^\infty \mathrm{d}h\,e^{-zh}\Pi(h|y), \tag{34}$$

such that $Z(0|y) = 1$ and $Z'(0|y) := -a^*$ is the conditional average of h for a given y.

Now we assume, and self consistently check, that for large y, $\Pi(h|y)$ is of the form $1/yF(h/y)$, which means that h is a random variable of order y. This implies :

$$Z(z|y) = G(x = zy); \qquad G(x) := \int_0^\infty \mathrm{d}u\,e^{-zu}\,F(u). \tag{35}$$

Multiplying Eq.(32) by e^{-zh} and integrating over h then leads, in the stationary state, to :

$$- x\pi(y)\left[(1-n_H)G'(x)+n_H G(x)\right] - \chi\frac{\partial}{\partial y}\left\{\left[(n_Z-1)G(x)-n_Z G'(x)\right]y\pi(y)\right\} \quad (36)$$
$$+ 2\chi n_Z\frac{\partial^2}{\partial y^2}\left\{\left[G(x)-G'(x)\right]y^2\pi(y)\right\} = 0,$$

where we have assumed $y \gg \lambda_\infty$. In the asymptotic limit, $\pi(y)$ behaves as a power law : $\pi(y) \propto A/y^{1+\mu}$. Indeed, injecting this ansatz into the last equation leads to a non-trivial equation for $G(x)$ where y and A have completely disappeared :

$$x\left[(1-n_H)G'(x)+n_H G(x)\right] = \chi\left[\mu n_Z H(x) - n_Z x H'(x) - \mu G(x) + xG'(x)\right] \quad (37)$$
$$+ 2\chi n_Z\left[x^2H''(x) + 2(1-\mu)xH'(x) - \mu(1-\mu)H(x)\right],$$

where we have introduced the shorthand $H(x) = G(x) - G'(x)$. Let us first analyze this equation for $x = 0$; without any further assumptions one has, with $G(0) = 1$ and $G'(0) = -a^*$:

$$\mu n_Z(1+a^*) - \mu - 2n_Z\mu(1-\mu)(1+a^*) = 0 \Rightarrow \mu = \frac{1}{2} + \frac{1}{2n_Z(1+a^*)}, \quad (38)$$

where the unphysical solution $\mu = 0$ was discarded. We thus need to solve Eq. (37) for $G(x)$ and determine a^* from the value of $-G'(0)$. An easy case is $\chi = 0$. One immediately finds :

$$(1-n_H)G'(x) + n_H G(x) = 0 \Rightarrow G_0(x) = e^{-n_H x/(1-n_H)}, \quad (39)$$

leading to $a_0^* = n_H/(1-n_H)$. The small χ expansion is also conveniently performed by setting $G(x) = G_0(x) + \chi g_1(x) + \chi^2 g_2(x) + \ldots$. To first order in χ, the equation for g_1 reads :

$$(1-n_H)g_1'(x) + n_H g_1(x) = G_0(x)\left[\frac{n_H(1-n_H-n_Z)}{(1-n_H)^2} + \frac{n_H^2}{(1-n_H)^2}x\right], \quad (40)$$

and thus, with the right boundary condition for $g_1(x)$,

$$g_1(x) = \left[\frac{n_H(1-n_H-n_Z)}{(1-n_H)^3}x + \frac{n_H^2}{2(1-n_H)^3}x^2\right]e^{-a_0^* x}. \quad (41)$$

To first order, one thus finds :

$$a^* = \frac{n_H}{(1-n_H)}\left[1 - \chi\frac{(1-n_H-n_Z)}{(1-n_H)^2} + O(\chi^2)\right]. \quad (42)$$

In the opposite limit $\chi \to \infty$, one finds that $G(x) = 1$ solves the equation, as expected since in this limit h cannot follow the dynamics of y, and therefore one expects that in the limit $y \to \infty$, $F(u) \approx \delta(u)$ and thus $G(x) = 1$. When χ is large but not infinite, one can expect that $F(u)$ has a

width of order χ^{-1}, and thus that $G(x)$ is a function of x/χ. This means that each derivative of G brings an extra factor χ^{-1}. Setting $a^* = a/\chi$ and matching the terms in Eq. (37), we find :

$$\chi(\mu + x(1-n_Z))G'(x) + (a\mu + n_H x)G(x), \tag{43}$$

or :

$$\ln G(x) = -\frac{1}{\chi(1-n_Z)}\left[n_H x + \mu(a - n_H/(1-n_Z))\ln(\mu + (1-n_Z)x)\right], \tag{44}$$

which shows that our assumption that $G(x)$ is a function of x/χ singles out $a = n_H/(1-n_Z)$ as the only possibility, in which case :

$$G(x) =_{\chi\to\infty} e^{-\frac{n_H x}{\chi(1-n_Z)}}. \tag{45}$$

This means that in this limit, $\Pi(h|y) \approx \delta(h - \frac{n_H}{\chi(1-n_Z)}y)$.

Finally, let us consider the limit $n_Z \to 0$ for a finite χ. The idea now is to postulate that for small n_Z, $\Pi(h|y)$ is stronlgy peaked around a^*y, with a width that goes to zero as $\sqrt{n_Z}$. This translates into the following ansatz for $G(x)$

$$G(x) = e^{-a^*x}\mathcal{G}(\sqrt{n_Z}x). \tag{46}$$

We can now analyze Eq. (37) in the regime $n_Z \to 0$ with fixed $z = \sqrt{n_Z}x$. The leading order terms are of order $1/n_Z$, and lead to an equation that is identically satisfied. The next two orders, $O(1/\sqrt{n_Z})$ and $O(1)$ allow us to fix both the function $\mathcal{G}(z)$ and the value of a^*. We find in particular :

$$\mathcal{G}(z) = \exp\left[\frac{(1+a^*)^2[(1-n_H+\chi)a^* - n_H]}{\chi}z^2\right], \tag{47}$$

which shows that the distribution $\Pi(h|y)$ is in fact gaussian in that limit. We also find that a^* obeys the following equation :

$$(\gamma a^* - n_H)(\gamma + (\gamma + 2\chi)a^*) = a^{*2}\chi^2 \qquad \gamma = 1 - n_H + \chi. \tag{48}$$

The solution takes a simple form in the limits $\chi \to 0$ and $\chi \to \infty$, where we recover the results obtained above.

In the general case, Eq. (37) is a third order, linear ODE for $G(x)$; imposing the correct boundary condition $G(x \to \infty)$ selects special values of a^* for any triplet (n_H, n_Z, χ). The largest admissible value of a^* corresponding to the smallest value of the tail exponent will be the physical solution. Unfortunately, we have not been able to make progress yet on this general case.

QARCH fit of a QHawkes process

The QARCH fit of real data in section E.5 has evidenced a diagonal + rank one structure of the QARCH kernel, which we interpreted as the QHawkes kernel, the naive fit of the latter being too computationally demanding. To justify the relevance of using a QARCH model as a discrete approximation of the QHawkes for fitting, we simulate a QHawkes process with kernel similar to the one obtained on real data and proceed to its QARCH fit. The QHawkes kernel is chosen as in Eq. (E.32) :

$$K(t,t') \;\approx\; \phi(t)\delta_{t-t'} + k(t)k(t')$$

where

$$\phi(t) = \frac{g}{(1+at)^{\alpha}}\, e^{-\frac{t}{t_0}} \quad , \quad k(t) = k_0 \exp(-\omega t),$$

with $g = 0.82,\ a = 60,\ \tau_0 = 400,\ \alpha = 0.7,\ k_0 = 0.08,\ \omega = 0.03$ and where an exponential cutoff has been added at the scale $t_0 = 400\text{min} \approx$ one trading day. These parameters have been chosen to produce respectively the norms $n_H = 0.8$ and $n_Z = 0.1$ while producing an intraday power-law decay with exponent $\alpha < 1$ for the diagonal kernel, as was found in section E.5. Note that t is expressed in minutes here whereas the parameter τ of section E.5 was expressed in number of 5-min bins.

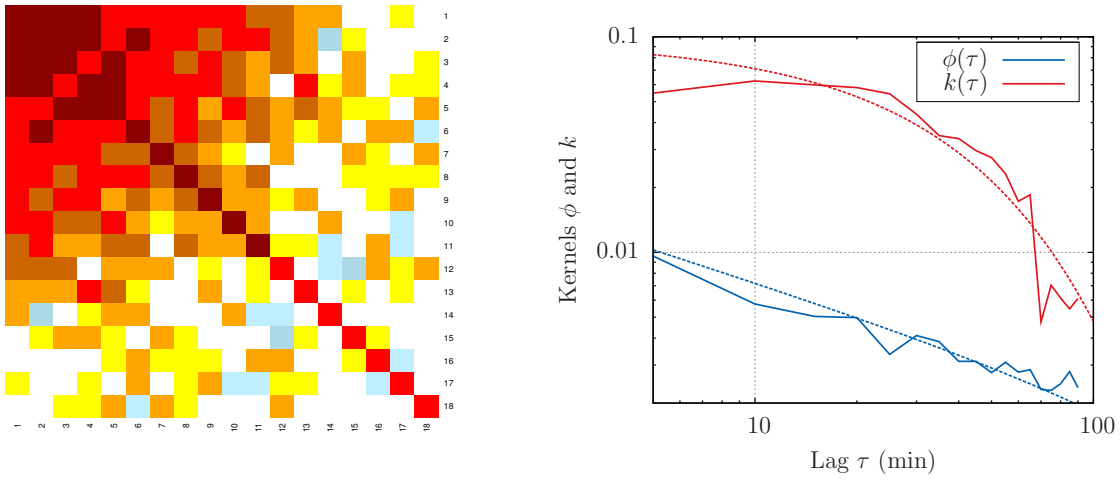

FIGURE 8 – Fit of the simulated QHawkes process. Left : heatmap of the quadratic kernel. White coefficients are close to zero, blue ones are negative and yellow/orange/red ones are positive, with darker shades as they increase in absolute value. The fit evidences a slowly-decaying diagonal component on top of a rank-one structure. Right : kernels $\phi(t)$ and $k(t)$. The rank-one kernel k is plotted in red (and is larger for small t's), and the diagonal kernel ϕ is plotted in blue, both in log-log scale. The dashed lines are the power-law fits for $\phi(t)$ with exponent $\alpha = 0.5$ and the exponential fit for $k(t)$ with characteristic time about 30 min.

The estimation sample consists of 1000 trading days, each composed of 80 five-minute returns

simulated with the ZHawkes model. The heatmap of the fitted kernel is shown on Figure 8, as well as the fits for the diagonal and the rank-one kernels. The results are satisfying as we recover the expected structure, with a rank-one exponential kernel with characteristic time about 30 min and a diagonal power-law kernel with exponent $\alpha \approx 0.5$, $\sim 30\%$ smaller than the expected value $\alpha = 0.7$. The leverage kernel is not shown on the figure, as it is found to be oscillating around zero.

Bibliographie

Aurélien Alfonsi and Pierre Blanc. Dynamic optimal execution in a mixed-market-impact hawkes price model. *Finance and Stochastics*, 20(1) :183–218, 2016.

Aurélien Alfonsi and Alexander Schied. Optimal trade execution and absence of price manipulations in limit order book models. *SIAM Journal on Financial Mathematics*, 1(1) :490–522, 2010.

Aurelien Alfonsi and Alexander Schied. Capacitary measures for completely monotone kernels via singular control. *SIAM Journal on Control and Optimization*, 51(2) :1758–1780, 2013.

Robleh Ali, John Barrdear, Roger Clews, and James Southgate. The economics of digital currencies. *Bank of England Quarterly Bulletin*, page Q3, 2014.

Romain Allez and Jean-Philippe Bouchaud. Individual and collective stock dynamics : intra-day seasonalities. *New Journal of Physics*, 13(2) :025010, 2011.

Robert Almgren and Neil Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3 : 5–40, 2001.

Robert Almgren, Chee Thum, Emmanuel Hauptmann, and Hong Li. Direct estimation of equity market impact. *Risk*, 18 :5752, 2005.

Yakov Amihud. Illiquidity and stock returns : cross-section and time-series effects. *Journal of financial markets*, 5(1) :31–56, 2002.

Yakov Amihud and Haim Mendelson. Asset pricing and the bid-ask spread. *Journal of financial Economics*, 17(2) :223–249, 1986.

Torben G Andersen, Oleg Bondarenko, Albert S Kyle, and Anna A Obizhaeva. Intraday trading invariance in the e-mini S&P 500 futures market. *Available at SSRN*, 2015.

Thierry Ané and Hélyette Geman. Order flow, transaction clock, and normality of asset returns. *The Journal of Finance*, 55(5) :2259–2284, 2000.

Alain Arneodo, Jean-François Muzy, and Didier Sornette. Direct causal cascade in the stock market. *The European Physical Journal B-Condensed Matter and Complex Systems*, 2(2) :277–282, 1998.

Kenneth J Arrow and Gerard Debreu. Existence of an equilibrium for a competitive economy. *Econometrica : Journal of the Econometric Society*, pages 265–290, 1954.

W Brian Arthur, Steven N Durlauf, and David A Lane. *The economy as an evolving complex system II*, volume 28. Addison-Wesley Reading, MA, 1997.

Colin Atkinson. A Wiener-Hopf integral equation arising in some inference and queueing problems. *Biometrika*, 61(2) :277–283, 1974.

Marco Avellaneda, Gennady Kasyan, and Michael D Lipkin. Mathematical models for stock pinning near option expiration dates. *Communications on Pure and Applied Mathematics*, 65(7) :949–974, 2012.

Louis Bachelier. *Théorie de la spéculation*. Gauthier-Villars, 1900.

Emmanuel Bacry and Jean-François Muzy. Hawkes model for price and trades high-frequency dynamics. *Quantitative Finance*, 14(7) :1147–1166, 2014.

Emmanuel Bacry, Jean Delour, and Jean-François Muzy. Modelling financial time series using multifractal random walks. *Physica A : Statistical Mechanics and its Applications*, 299(1) :84–92, 2001.

Emmanuel Bacry, Alexey Kozhemyak, and Jean-François Muzy. Continuous cascade models for asset returns. *Journal of Economic Dynamics and Control*, 32(1) :156–199, 2008.

Emmanuel Bacry, Khalil Dayri, and Jean-François Muzy. Non-parametric kernel estimation for symmetric hawkes processes. application to high frequency financial data. *The European Physical Journal B*, 85(5) :1–12, 2012.

Emmanuel Bacry, Sylvain Delattre, Marc Hoffmann, and Jean-François Muzy. Modelling micro-structure noise with mutually exciting point processes. *Quantitative Finance*, 13(1) :65–77, 2013.

Emmanuel Bacry, Adrian Iuga, Matthieu Lasnier, and Charles-Albert Lehalle. Market impacts and the life cycle of investors orders. *Market Microstructure and Liquidity*, page 1550009, 2014.

Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01) :1550005, 2015.

Per Bak, Maya Paczuski, and Martin Shubik. Price variations in a stock market with many agents. *Physica A : Statistical Mechanics and its Applications*, 246(3) :430–453, 1997.

Gerard T. Barkema, Martin J. Howard, and John L. Cardy. Reaction-diffusion front for $A + B \to 0$ in one dimension. *Physical Review E*, 53(3) :R2017, 1996.

Avraham Beja and M Barry Goldman. On the dynamic behavior of prices in disequilibrium. *The Journal of Finance*, 35(2) :235–248, 1980.

Michael Benzaquen, Jonathan Donier, and Jean-Philippe Bouchaud. Unravelling the trading invariance hypothesis. *Available at SSRN 2730817*, 2016.

Lorenzo Bergomi. Smile dynamics II. *Available at SSRN 1493302*, 2005.

Nataliya Bershova and Dmitry Rakhlin. The non-linear market impact of large trades : Evidence from buy-side order flow. *Quantitative Finance*, 13(11) :1759–1778, 2013.

Bruno Biais. Price formation and equilibrium liquidity in fragmented and centralized markets. *The Journal of Finance*, 48(1) :157–185, 1993.

Bruno Biais, Pierre Hillion, and Chester Spatt. An empirical analysis of the limit order book and the order flow in the paris bourse. *the Journal of Finance*, 50(5) :1655–1689, 1995.

Fischer Black. Noise. *The journal of finance*, 41(3) :529–543, 1986.

Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *The journal of political economy*, pages 637–654, 1973.

Alex J Bladon, Esteban Moro, and Tobias Galla. Agent-specific impact of single trades in financial markets. *Physical Review E*, 85(3) :036103, 2012.

Pierre Blanc. Modélisation de la volatilité des marchés financiers par une structure ARCH multi-fréquence. *Master's thesis, Université de Paris VI Pierre et Marie Curie*, 2012. Available upon request.

Pierre Blanc, Rémy Chicheportiche, and Jean-Philippe Bouchaud. The fine structure of volatility feedback II : overnight and intra-day effects. *Physica A : Statistical Mechanics and its Applications*, 402 :58–75, 2014.

Pierre Blanc, Jonathan Donier, and Jean-Philippe Bouchaud. Quadratic hawkes processes for financial prices. *Available at SSRN*, 2015.

Ekkehart Boehmer, Kingsley YL Fong, and Juan Julie Wu. International evidence on algorithmic trading. In *AFA 2013 San Diego Meetings Paper*, 2014.

Joop Boersma. Note on a wiener-hopf integral equation arising in some inference and queueing problems. 1974.

Rainer Böhme, Nicolas Christin, Benjamin G Edelman, and Tyler Moore. Bitcoin. *Journal of Economic Perspectives, Forthcoming*, pages 15–015, 2014.

Tim Bollerslev and Dan Jubinski. Equity trading volume and volatility : Latent information arrivals and common long-run dependencies. *Journal of Business & Economic Statistics*, 17(1) :9–21, 1999.

Tim Bollerslev, Robert F Engle, and Daniel B Nelson. ARCH models. *Handbook of econometrics*, 4 :2959–3038, 1994.

Julius Bonart, Jean-Philippe Bouchaud, Augustin Landier, and David Thesmar. Instabilities in large economies : aggregate volatility without idiosyncratic shocks. *Journal of Statistical Mechanics : Theory and Experiment*, 2014(10) :P10040, 2014.

Steve Bongiovanni, Milan Borkovec, and Robert D Sinclair. Let's play hide-and-seek : The location and size of undisclosed limit order volume. *The Journal of Trading*, 1(3) :34–46, 2006.

Jean-Philippe Bouchaud. Price impact. *Encyclopedia of quantitative finance*, 2010.

Jean-Philippe Bouchaud. The endogenous dynamics of markets : Price impact, feedback loops and instabilities. *Lessons from the Credit Crisis. Risk Publications*, 2011.

Jean-Philippe Bouchaud. Crises and collective socio-economic phenomena : simple models and challenges. *Journal of Statistical Physics*, 151(3-4) :567–606, 2013.

Jean-Philippe Bouchaud and Rama Cont. A langevin approach to stock market fluctuations and crashes. *The European Physical Journal B-Condensed Matter and Complex Systems*, 6(4) :543–550, 1998.

Jean-Philippe Bouchaud and Marc Potters. *Theory of financial risk and derivative pricing : from statistical physics to risk management*. Cambridge university press, 2003.

Jean-Philippe Bouchaud, Marc Mézard, and Marc Potters. Statistical properties of stock order books : empirical results and models. *Quantitative finance*, 2(4) :251–256, 2002.

Jean-Philippe Bouchaud, Yuval Gefen, Marc Potters, and Matthieu Wyart. Fluctuations and response in financial markets : the subtle nature of 'random' price changes. *Quantitative Finance*, 4 (2) :176–190, 2004.

Jean-Philippe Bouchaud, Julien Kockelkoren, and Marc Potters. Random walks, liquidity molasses and critical response in financial markets. *Quantitative finance*, 6(02) :115–123, 2006.

Jean-Philippe Bouchaud, J Doyne Farmer, and Fabrizio Lillo. How markets slowly digest changes in supply and demand. *Handbook of financial markets : dynamics and evolution*, 1 :57, 2009.

Pierre Brémaud and Laurent Massoulié. Hawkes branching point processes without ancestors. *Journal of applied probability*, 38(1) :122–135, 2001.

Jonathan Brogaard, Terrence Hendershott, and Ryan Riordan. High-frequency trading and price discovery. *Review of Financial Studies*, 27(8) :2267–2306, 2014.

Xavier Brokmann, Emmanuel Serie, Julien Kockelkoren, and J-P Bouchaud. Slow decay of impact in equity markets. *Market Microstructure and Liquidity*, page 1550007, 2015.

Eric B Budish, Peter Cramton, and John J Shim. The high-frequency trading arms race : Frequent batch auctions as a market design response. *Fama-Miller Working Paper*, pages 14–03, 2013.

Jorge Buescu. Positive integral operators in unbounded domains. *Journal of Mathematical Analysis and Applications*, 296(1) :244–255, 2004.

Fabio Caccioli, Jean-Philippe Bouchaud, and J. Doyne Farmer. Impact-adjusted valuation and the criticality of leverage. *Risk*, 2012.

Albert Camus. *L'homme révolté*, volume 2. Gallimard Paris, 1951.

Girolamo Cardano and T Richard Witmer. *Ars magna or the rules of algebra*. 1993.

Álvaro Cartea and Sebastian Jaimungal. Optimal execution with limit and market orders. *Quantitative Finance*, 15(8) :1279–1291, 2015.

Timothy N Cason and Daniel Friedman. Price formation in double auction markets. *Journal of Economic Dynamics and Control*, 20(8) :1307–1337, 1996.

Damien Challet, Matteo Marsili, Yi-Cheng Zhang, et al. Minority games : interacting agents in financial markets. *OUP Catalogue*, 2013.

Rémy Chicheportiche and Jean-Philippe Bouchaud. The fine-structure of volatility feedback I : Multi-scale self-reflexivity. *Physica A : Statistical Mechanics and its Applications*, 410 :174–195, 2014.

Peter K Clark. A subordinated stochastic process model with finite variance for speculative prices. *Econometrica : journal of the Econometric Society*, pages 135–155, 1973.

Rama Cont and Adrien De Larrard. Price dynamics in a markovian limit order market. *SIAM Journal on Financial Mathematics*, 4(1) :1–25, 2013.

Rama Cont and Arseniy Kukanov. Optimal order placement in limit order markets. *Available at SSRN 2155218*, 2013.

Rama Cont, Sasha Stoikov, and Rishi Talreja. A stochastic model for order book dynamics. *Operations research*, 58(3) :549–563, 2010.

Bradford Cornell. What moves stock prices : Another look. *Journal of Portfolio Management*, 39 (3) :32, 2013.

Francesco Corradi, Andrea Zaccaria, and Luciano Pietronero. Liquidity crises on different time scales. *Physical Review E*, 92(6) :062802, 2015.

John C Cox, Jonathan E Ingersoll Jr, and Stephen A Ross. An intertemporal general equilibrium model of asset prices. *Econometrica : Journal of the Econometric Society*, pages 363–384, 1985.

Matthieu Cristelli, Luciano Pietronero, and Andrea Zaccaria. Critical overview of agent-based models for economics. *Proceedings of the International School of Physics "Enrico Fermi" Course CLXXVI, Complex Materials in Physics and Biology, edited by F. Mallamace and H.E. Stanley*, 2011.

Gianbiagio Curato, Jim Gatheral, and Fabrizio Lillo. Optimal execution with nonlinear transient market impact. *Available at SSRN 2539240*, 2014.

David M Cutler, James M Poterba, and Lawrence H Summers. What moves stock prices ? *The Journal of Portfolio Management*, 15(3) :4–12, 1989.

Khalil Dayri and Mathieu Rosenbaum. Large tick assets : implicit spread and optimal tick size. *Market Microstructure and Liquidity*, 1(01) :1550003, 2015.

Werner FM De Bondt and Richard H Thaler. Do security analysts overreact ? *The American Economic Review*, pages 52–57, 1990.

Joachim De Lataillade, Cyril Deremble, Marc Potters, and Jean-Philippe Bouchaud. Optimal trading with linear costs. *arXiv preprint arXiv :1203.5957*, 2012.

David S Dean. Langevin equation for the density of a system of interacting langevin processes. *Journal of Physics A : Mathematical and General*, 29(24) :L613, 1996.

Tiziana Di Matteo, Tomaso Aste, and Michel M Dacorogna. Long-term memories of developed and emerging markets : Using the scaling analysis to characterize their stage of development. *Journal of Banking & Finance*, 29(4) :827–851, 2005.

Jonathan Donier. Market impact with autocorrelated order flow under perfect competition. *Available at SSRN 2191660*, 2012.

Jonathan Donier and Julius Bonart. A million metaorder analysis of market impact on the bitcoin. *Market Microstructure and Liquidity*, page 1550008, 2014.

Jonathan Donier and Jean-Philippe Bouchaud. Why do markets crash ? Bitcoin data offers unprecedented insights. *PloS one*, 10(10) :e0139356, 2015a.

Jonathan Donier and Jean-Philippe Bouchaud. From walras' auctioneer to continuous time double auctions : A general dynamic theory of supply and demand. *arXiv preprint arXiv :1506.03758*, 2015b.

Jonathan Donier, Julius Friedrich Bonart, Iacopo Mastromatteo, and Jean-Philippe Bouchaud. A fully consistent, minimal model for non-linear market impact. *Quantitative finance*, 15(7) :1109–1121, 2015.

David Easley and John Ledyard. Theories of price formation and exchange in double oral auctions. *The double auction market : Institutions, theories, and evidence*, pages 63–97, 1993.

Zoltán Eisler and János Kertész. Size matters, some stylized facts of the market revisited. *Eur. J. Phys.*, B51 :145–154, 2006.

Robert F Engle. The econometrics of ultra-high-frequency data. *Econometrica*, 68(1) :1–22, 2000.

George W Evans and Seppo Honkapohja. *Learning and expectations in macroeconomics*. Princeton University Press, 2001.

Ray C Fair. Events that shook the market. *The Journal of Business*, 75(4) :713–731, 2002.

Eugene F Fama. Efficient capital markets : A review of theory and empirical work. *The journal of Finance*, 25(2) :383–417, 1970.

J Doyne Farmer, Paolo Patelli, and Ilija I Zovko. The predictive power of zero intelligence in financial markets. *Proceedings of the National Academy of Sciences of the United States of America*, 102 (6) :2254–2259, 2005.

J Doyne Farmer, Austin Gerig, Fabrizio Lillo, and Henri Waelbroeck. How efficiency shapes market impact. *Quantitative Finance*, 13(11) :1743–1758, 2013.

Vladimir Filimonov and Didier Sornette. Apparent criticality and calibration issues in the hawkes self-excited point process model : application to high-frequency financial data. *Quantitative Finance*, 15(8) :1293–1314, 2015.

Franklin M Fisher. *Disequilibrium foundations of equilibrium economics*. Number 6. Cambridge University Press, 1989.

Julie Lyng Forman and Michael Sørensen. The pearson diffusions : A class of statistically tractable diffusion processes. *Scandinavian Journal of Statistics*, 35(3) :438–465, 2008.

Thierry Foucault. Order flow composition and trading costs in a dynamic limit order market. *Journal of Financial markets*, 2(2) :99–134, 1999.

Thierry Foucault, Marco Pagano, and Ailsa Röell. *Market liquidity : theory, evidence, and policy*. Oxford University Press, 2013.

Daniel Fricke and Austin Gerig. Liquidity risk, speculative trade, and the optimal latency of financial markets. 2014.

Xavier Gabaix. Power laws in economics and finance. *Annu. Rev. Econom.*, 1(1) :255–294, 2009.

Xavier Gabaix, Parameswaran Gopikrishnan, Vasiliki Plerou, and H Eugene Stanley. A theory of power-law distributions in financial market fluctuations. *Nature*, 423(6937) :267–270, 2003.

Xuefeng Gao, JG Dai, Ton Dieker, and Shijie Deng. Hydrodynamic limit of order book dynamics. *Available at SSRN 2530306*, 2014.

Crispin W. Gardiner. *Stochastic Methods*. Springer, 4 edition, 2009.

Ahcène Gareche, Gaétan Disdier, Julianus Kockelkoren, and Jean-Philippe Bouchaud. Fokker-planck description for the queue dynamics of large tick stocks. *Physical Review E*, 88(3) :032809, 2013.

Mark B Garman and Michael J Klass. On the estimation of security price volatilities from historical data. *Journal of business*, pages 67–78, 1980.

Jim Gatheral. No-dynamic-arbitrage and market impact. *Quantitative Finance*, 10(7) :749–759, 2010.

Jim Gatheral, Thibault Jaisson, and Mathieu Rosenbaum. Volatility is rough. *Available at SSRN 2509457*, 2014.

Nicola Gennaioli, Yueran Ma, and Andrei Shleifer. Expectations and investment. In *NBER Macroeconomics Annual 2015, Volume 30*. University of Chicago Press, 2015.

Irene Giardina and Jean-Philippe Bouchaud. Bubbles, crashes and intermittency in agent based market models. *The European Physical Journal B-Condensed Matter and Complex Systems*, 31 (3) :421–437, 2003.

Steven Gjerstad and John Dickhaut. Price formation in double auctions. *Games and economic behavior*, 22(1) :1–29, 1998.

Lawrence R Glosten. Is the electronic open limit order book inevitable ? *The Journal of Finance*, 49(4) :1127–1161, 1994.

Lawrence R Glosten and Paul R Milgrom. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of financial economics*, 14(1) :71–100, 1985.

Carla Gomes and Henri Waelbroeck. Is market impact a measure of the information value of trades ? market response to liquidity vs. informed metaorders. *Quantitative Finance*, 15(5) :773–793, 2015.

Parameswaran Gopikrishnan, Vasiliki Plerou, Luis A Nunes Amaral, Martin Meyer, and H Eugene Stanley. Scaling of the distribution of fluctuations of financial market indices. *Physical Review E*, 60(5) :5305, 1999.

Martin D Gould, Mason A Porter, Stacy Williams, Mark McDonald, Daniel J Fenn, and Sam D Howison. Limit order books. *Quantitative Finance*, 13(11) :1709–1742, 2013.

Ruslan Y Goyenko, Craig W Holden, and Charles A Trzcinka. Do liquidity measures measure liquidity ? *Journal of financial Economics*, 92(2) :153–181, 2009.

Richard C Grinold and Ronald N Kahn. Active portfolio management. 2000.

Stanislao Gualdi, Jean-Philippe Bouchaud, Giulia Cencetti, Marco Tarzia, and Francesco Zamponi. Endogenous crisis waves : Stochastic model with synchronized collective behavior. *Physical review letters*, 114(8) :088701, 2015a.

Stanislao Gualdi, Marco Tarzia, Francesco Zamponi, and Jean-Philippe Bouchaud. Tipping points in macroeconomic agent-based models. *Journal of Economic Dynamics and Control*, 50 :29–61, 2015b.

Puneet Handa and Robert A Schwartz. Limit order trading. *The Journal of Finance*, 51(5) : 1835–1861, 1996.

Puneet Handa, Robert Schwartz, and Ashish Tiwari. Quote setting and price formation in an order driven market. *Journal of financial markets*, 6(4) :461–489, 2003.

Stephen J Hardiman and Jean-Philippe Bouchaud. Branching-ratio approximation for the self-exciting hawkes process. *Physical Review E*, 90(6) :062807, 2014.

Stephen J Hardiman, Nicolas Bercot, and Jean-Philippe Bouchaud. Critical reflexivity in financial markets : a hawkes process analysis. *The European Physical Journal B*, 86(10) :1–9, 2013.

Lawrence Harris et al. Liquidity, trading rules and electronic trading systems. Technical report, 1990.

Joel Hasbrouck. Measuring the information content of stock trades. *The Journal of Finance*, 46 : 179–206, 1991.

Joel Hasbrouck. *Empirical market microstructure : The institutions, economics, and econometrics of securities trading*. Oxford University Press, 2006.

Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1) :83–90, 1971.

Terrence Hendershott, Charles M Jones, and Albert J Menkveld. Does algorithmic trading improve liquidity ? *The Journal of Finance*, 66(1) :1–33, 2011.

Steven L Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of financial studies*, 6(2) :327–343, 1993.

John R Hicks. *Value and capital*, volume 2. Clarendon press Oxford, 1946.

Bruce M Hill. A simple general approach to inference about the tail of a distribution. *The annals of statistics*, 3(5) :1163–1174, 1975.

Cars Hommes, Joep Sonnemans, Jan Tuinstra, and Henk Van de Velden. Coordination of expectations in asset pricing experiments. *Review of Financial studies*, 18(3) :955–980, 2005.

Cars H Hommes. Heterogeneous agent models in economics and finance. *Handbook of computational economics*, 2 :1109–1186, 2006.

Weibing Huang, Charles-Albert Lehalle, and Mathieu Rosenbaum. Simulating and analyzing order book data : The queue-reactive model. *Journal of the American Statistical Association*, (just-accepted) :00–00, 2014.

Gur Huberman and Werner Stanzl. Price manipulation and quasi-arbitrage. *Econometrica*, 72(4) : 1247–1275, 2004.

Thibault Jaisson. Market impact as anticipation of the order flow imbalance. *Quantitative Finance*, 15(7) :1123–1135, 2015.

Thibault Jaisson and Mathieu Rosenbaum. Limit theorems for nearly unstable hawkes processes. *The Annals of Applied Probability*, 25(2) :600–631, 2015a.

Thibault Jaisson and Mathieu Rosenbaum. Rough fractional diffusions as scaling limits of nearly unstable heavy tailed hawkes processes. *arXiv preprint arXiv :1504.03100*, 2015b.

Charles M Jones. A century of stock market liquidity and trading costs. *Available at SSRN 313681*, 2002.

Charles M Jones. What do we know about high-frequency trading ? *Columbia Business School Research Paper*, (13-11), 2013.

Charles M Jones, Gautam Kaul, and Marc L Lipson. Transactions, volume, and volatility. *Review of Financial Studies*, 7(4) :631–651, 1994.

Armand Joulin, Augustin Lefevre, Daniel Grunberg, and Jean-Philippe Bouchaud. Stock price jumps : News and volume play a minor role. *Wilmott Magazine*, pages 1–7, September/October 2008.

Olav Kallenberg. *Foundations of modern probability*. Springer Science & Business Media, 2002.

John Maynard Keynes. *The general theory of employment, interest and money*. Macmilllan, 1936.

Charles Kindleberger and Robert Aliber. *Manias, Panics and Crashes : a History of Financial Crises*. Palgrave Macmillan, 2011.

Andrei A Kirilenko, Albert S Kyle, Mehrdad Samadi, and Tugkan Tuzun. The flash crash : The impact of high frequency trading on an electronic market. *Available at SSRN 1686004*, 2015.

Alan Kirman. Whom or what does the representative individual represent ? *The Journal of Economic Perspectives*, 6(2) :117–136, 1992.

Alan Kirman. Ants, rationality, and recruitment. *The Quarterly Journal of Economics*, pages 137–156, 1993.

Julianus Kockelkoren. *Order types*. Encyclopedia of Quantitative Finance, 2010.

Albert S Kyle. Continuous auctions and insider trading. *Econometrica : Journal of the Econometric Society*, pages 1315–1335, 1985.

Albert S. Kyle and Anna A. Obizhaeva. Market microstructure invariants. *SSRN 1687965*, 2010.

Albert S Kyle and Anna A Obizhaeva. Large bets and stock market crashes. In *AFA 2013 San Diego Meetings Paper*, 2012.

Albert S Kyle, Anna A Obizhaeva, and Yajun Wang. Smooth trading with overconfidence and market power. *Robert H. Smith School Research Paper No. RHS*, 2423207, 2014.

Aimé Lachapelle, Jean-Michel Lasry, Charles-Albert Lehalle, and Pierre-Louis Lions. Efficiency of the price formation process in presence of high frequency participants : a mean field game analysis. *Mathematics and Financial Economics*, pages 1–40, 2013.

Mehdi Lallouache and Damien Challet. Statistically significant fits of hawkes processes to financial data. *Available at SSRN 2450101*, 2014.

Sophie Laruelle, Charles-Albert Lehalle, and Gilles Pages. Optimal split of orders across liquidity pools : a stochastic algorithm approach. *SIAM Journal on Financial Mathematics*, 2(1) :1042–1076, 2011.

Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese Journal of Mathematics*, 2 (1) :229–260, 2007.

Patrick J Laub, Thomas Taimre, and Philip K Pollett. Hawkes processes. *arXiv preprint arXiv :1507.02822*, 2015.

Charles-Albert Lehalle and Sophie Laruelle. *Market Microstructure in Practice*. World Scientific, 2013.

Charles-Albert Lehalle and Mathieu Lasnier. Mathematical models for stock pinning near option expiration dates. what does the saw-tooth pattern on us markets on 19 july 2012 tell us about the price formation process. *Technical report, CA Chevreux Quantitative Research*, 2012.

Charles-Albert Lehalle, Olivier Guéant, and Julien Razafinimanana. High-frequency simulations of an order book : a two-scale approach. In *Econophysics of Order-driven Markets*, pages 73–92. Springer, 2011.

Roman Liesenfeld. A generalized bivariate mixture model for stock price volatility and trading volume. *Journal of Econometrics*, 104(1) :141–178, 2001.

Fabrizio Lillo and J Doyne Farmer. The long memory of the efficient market. *Studies in nonlinear dynamics & econometrics*, 8(3), 2004.

Fabrizio Lillo and J Doyne Farmer. The key role of liquidity fluctuations in determining large price changes. *Fluctuation and Noise Letters*, 5(02) :L209–L216, 2005.

Charles Mackay. *Extraordinary popular delusions and the madness of crowds*. Start Publishing LLC, 2012.

Ananth Madhavan. Market microstructure : A survey. *Journal of financial markets*, 3(3) :205–258, 2000.

Ananth Madhavan, Matthew Richardson, and Mark Roomans. Why do security prices fluctuate ? a transaction-level analysis of nyse stocks. *Rev. Financ. Stud.*, 10 :1035, 1997.

Costis Maglaras and Ciamac Moallemi. A multiclass model of limit order book dynamics and its application to optimal trade execution. Technical report, Working paper, Columbia Business School, New York, 2011.

Burton G Malkiel and Eugene F Fama. Efficient capital markets : A review of theory and empirical work. *The journal of Finance*, 25(2) :383–417, 1970.

Benoit B. Mandelbrot. *Fractals and scaling in finance : Discontinuity, concentration, risk*. Springer, 1997.

Gregory N Mankiw. *Principles of macroeconomics*. Cengage Learning, 2014.

Andreu Mas-Colell, Michael Dennis Whinston, Jerry R Green, et al. *Microeconomic theory*, volume 1. Oxford university press New York, 1995.

Sergei Maslov. Simple model of a limit order-driven market. *Physica A : Statistical Mechanics and its Applications*, 278(3) :571–578, 2000.

Iacopo Mastromatteo, Bence Toth, and Jean-Philippe Bouchaud. Agent-based models for latent liquidity and concave price impact. *Physical Review E*, 89(4) :042805, 2014a.

Iacopo Mastromatteo, Bence Toth, and Jean-Philippe Bouchaud. Anomalous impact in reaction-diffusion financial models. *Physical review letters*, 113(26) :268701, 2014b.

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

Esteban Moro, Javier Vicente, Luis G Moyano, Austin Gerig, J Doyne Farmer, Gabriella Vaglica, Fabrizio Lillo, and Rosario N Mantegna. Market impact and trading profile of hidden orders in stock markets. *Physical Review E*, 80(6) :066102, 2009.

Ulrich A Müller, Michel M Dacorogna, Rakhal D Davé, Richard B Olsen, Olivier V Pictet, and Jacob E von Weizsäcker. Volatilities of different time resolutions - analyzing the dynamics of market components. *Journal of Empirical Finance*, 4(2) :213–239, 1997.

Jean-François Muzy, Jean Delour, and Emmanuel Bacry. Modelling fluctuations of financial time series : from cascade process to stochastic volatility model. *The European Physical Journal B-Condensed Matter and Complex Systems*, 17(3) :537–548, 2000.

Satoshi Nakamoto. Bitcoin : A peer-to-peer electronic cash system, 2008.

Felix Patzelt and Klaus Pawelzik. An inherent instability of efficient markets. *Scientific reports*, 3, 2013.

Vasiliki Plerou, Parameswaran Gopikrishnan, Luis A Nunes Amaral, Martin Meyer, and H Eugene Stanley. Scaling of the distribution of price fluctuations of individual companies. *Physical Review E*, 60(6) :6519, 1999.

Yves Pomeau. Symétrie des fluctuations dans le renversement du temps. *Journal de Physique*, 43 (6) :859–867, 1982.

Silviu Predoiu, Gennady Shaikhet, and Steven Shreve. Optimal execution in a general one-sided limit-order book. *SIAM Journal on Financial Mathematics*, 2(1) :183–212, 2011.

James B Ramsey and Philip Rothman. Time irreversibility and business cycle asymmetry. *Journal of Money, Credit and Banking*, pages 1–21, 1996.

James Bernard Ramsey and Philip Rothman. *Characterization of the time irreversibility of economic time series : Estimators and test statistics*. CV Starr Center for Applied Economics, New York University, Faculty of Arts and Science, Department of Economics, 1988.

Carmen M Reinhart and Kenneth Rogoff. *This Time is Different : Eight Centuries of Financial Folly*. Princeton University Press, 2009.

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

Ioanid Roşu. Multi-stage game theory in continuous time. Technical report, Technical report, Working Paper, University of Chicago, 2006.

Ioanid Roşu. A dynamic model of the limit order book. *Review of Financial Studies*, page hhp011, 2009.

Ioanid Roşu. Liquidity and information in order driven markets. *Available at SSRN 1286193*, 2014.

L Christopher G Rogers and Stephen E Satchell. Estimating variance from high, low and closing prices. *The Annals of Applied Probability*, pages 504–512, 1991.

Alexander Saichev and Didier Sornette. Generation-by-generation dissection of the response function in long memory epidemic processes. *The European Physical Journal B*, 75(3) :343–355, 2010.

Paul A Samuelson. *Foundations of Economic Analysis, enlarged edition*. Harvard University Press Harvard, 1983.

Patrik Sandås. Adverse selection and competitive market making : Empirical evidence from a limit order market. *Review of Financial Studies*, 14(3) :705–734, 2001.

Thomas J Sargent. Bounded rationality in macroeconomics : The arne ryde memorial lectures. *OUP Catalogue*, 1993.

Thomas C Schelling. Dynamic models of segregation. *Journal of mathematical sociology*, 1(2) : 143–186, 1971.

WR Schneider. Fractional diffusion. In *Dynamics and Stochastic Processes Theory and Applications*, pages 276–286. Springer, 1990.

Enrique Sentana. Quadratic arch models. *The Review of Economic Studies*, 62(4) :639–661, 1995.

Robert J Shiller. Do stock prices move too much to be justified by subsequent changes in dividends ?, 1980.

Robert J Shiller. Sharing nobel honors, and agreeing to disagree. *New York Times*, 26, 2013.

Igor Skachkov. Market impact paradoxes. *Wilmott*, 2014(70) :71–78, 2014.

Eric Smith, J Doyne Farmer, László Gillemot, and Supriya Krishnamurthy. Statistical theory of the continuous double auction. *Quantitative finance*, 3(6) :481–514, 2003.

Vernon L Smith, Gerry L Suchanek, and Arlington W Williams. Bubbles, crashes, and endogenous expectations in experimental spot asset markets. *Econometrica : Journal of the Econometric Society*, pages 1119–1151, 1988.

Didier Sornette. *Why stock markets crash : critical events in complex financial systems*. Princeton University Press, 2009.

Elias M Stein and Jeremy C Stein. Stock price distributions with stochastic volatility : an analytic approach. *Review of financial Studies*, 4(4) :727–752, 1991.

Nassim Nicholas Taleb. *The black swan : The impact of the highly improbable fragility*. Random House, 2010.

Lei-Han Tang and Guang-Shan Tian. Reaction–diffusion–branching models of stock price fluctuations. *Physica A : Statistical Mechanics and its Applications*, 264(3) :543–550, 1999.

Damian E Taranto, Giacomo Bormetti, and Fabrizio Lillo. The adaptive nature of liquidity taking in limit order books. *Journal of Statistical Mechanics : Theory and Experiment*, 2014(6) :P06002, 2014.

George E Tauchen and Mark Pitts. The price variability-volume relationship on speculative markets. *Econometrica : Journal of the Econometric Society*, pages 485–505, 1983.

Peter Thiel and Blake Masters. *Zero to one : notes on startups, or how to build the future*. Crown Business, 2014.

Nicolo Torre and Mark Ferrari. Market impact model handbook, BARRA inc., Berkeley, 1997.

Bence Toth, Fabrizio Lillo, and J Doyne Farmer. Segmentation algorithm for non-stationary compound poisson processes. *The European Physical Journal B-Condensed Matter and Complex Systems*, 78(2) :235–243, 2010.

Bence Tóth, Yves Lempérière, Cyril Deremble, Joachim De Lataillade, Julien Kockelkoren, and Jean-Philippe Bouchaud. Anomalous price impact and the critical nature of liquidity in financial markets. *Physical Review X*, 1(2) :021006, 2011.

Bence Toth, Imon Palit, Fabrizio Lillo, and J Doyne Farmer. Why is equity order flow so persistent ? *Journal of Economic Dynamics and Control*, 51 :218–239, 2015.

Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

Leon Walras. *Elements of Pure Economics, or the Theory of Social Wealth*. 1954. Translated by William Jaff é from the original (1874). Published for American Economic Association and the Royal Economic Society.

Philipp Weber and Bernd Rosenow. Order book approach to price impact. *Quantitative Finance*, 5 (4) :357–364, 2005.

Matthieu Wyart, Jean-Philippe Bouchaud, Julien Kockelkoren, Marc Potters, and Michele Vettorazzo. Relation between bid–ask spread, impact and volatility in order-driven markets. *Quantitative Finance*, 8(1) :41–57, 2008.

Elia Zarinelli, Michele Treccani, J Doyne Farmer, and Fabrizio Lillo. Beyond the square root : Evidence for logarithmic dependence of market impact on size and participation rate. *Market Microstructure and Liquidity*, page 1550004, 2015.

Yi-Cheng Zhang. Toward a theory of marginally efficient markets. *Physica A : Statistical Mechanics and its Applications*, 269(1) :30–44, 1999.

Gilles Zumbach. How the trading activity scales with the company sizes in the FTSE 100. *Quantitative finance*, 4 :441, 2004.

Gilles Zumbach. Time reversal invariance in finance. *Quantitative Finance*, 9(5) :505–515, 2009.

Gilles Zumbach. Volatility conditional on price trends. *Quantitative Finance*, 10(4) :431–442, 2010.

Gilles Zumbach. Cross-sectional universalities in financial time series. *Quantitative Finance*, 15 (12) :1901–1912, 2015.

Gilles Zumbach and Paul Lynch. Heterogeneous volatility cascade in financial markets. *Physica A : Statistical Mechanics and its Applications*, 298(3-4) :521–529, 2001.

Jonathan Donier **10 Octobre 2016**

Agents hétérogènes et formation des prix sur les marchés financiers

Résumé : Cette thèse est consacrée à l'étude de la formation des prix sur les marchés financiers, en particulier lorsque ceux-ci se composent d'un grand nombre d'agents. On commence par l'étude empirique d'un marché émergent – le bitcoin – de manière à mieux comprendre comment les actions individuelles affectent les prix – ce que l'on appelle « l'impact de marché ». On développe ensuite un modèle théorique d'impact basé sur le concept d'agent hétérogène, qui parvient à reproduire les observations empiriques d'un impact concave dans un marché non manipulable. Le cadre de l'agent hétérogène nous permet de revisiter les concepts d'offre et de demande dans un cadre dynamique, de mieux comprendre l'impact du mécanisme de marché sur la liquidité, ou encore de poser les bases d'un simulateur de marché réaliste. On montre enfin, à travers l'étude empirique de plusieurs bulles et crashs sur le marché du bitcoin, le rôle crucial de la micro-structure dans la compréhension des phénomènes extrêmes.

Mots clés : Offre/demande, agents hétérogènes, marchés financiers, micro-structure, impact de marché, bitcoin, efficience, liquidité.

Heterogeneous agents and price formation on financial markets

Abstract : This thesis is devoted to the study of price formation on financial markets, in particular when these are composed of a large number of agents. We start by the empirical study of an emergent market – the bitcoin – in order to better understand how individual actions impact prices – a phenomenon known as « market impact ». We then develop a theoretical model based on the concept of heterogeneous agents, that allows to reproduce the empirical observations of a concave impact in a market that remains non-manipulable. The heterogeneous agents framework allows us to revisit the concepts of supply and demand in a dynamic context, to better understand how the choice of a particular market mechanism can impact liquidity, and to lay some grounds for a realistic market simulator. By studying several bubbles and crashes that happened on the bitcoin market, we finally show how relevant microstructure effects can be, in particular for understanding the occurrence of extreme phenomena.

Keywords : Supply/demand, heterogeneous agents, financial markets, micro-structure, market impact, bitcoin, efficiency, liquidity.