Principal Trading Arrangements: When Are Common Contracts Optimal?*

Markus Baldauf† Christoph Frei‡ Joshua Mollner§

August 8, 2020

Abstract

Many financial arrangements reference market prices that are yet to be realized at the time of contracting and consequently susceptible to manipulation. Two of the most common such arrangements are: (*i*) market-on-close contracts, which reference the price prevailing at the end of an execution window, and (*ii*) guaranteed VWAP contracts, which reference the volume-weighted average price (VWAP) prevailing over the execution window. To study such situations, we introduce a stylized model of financial contracting between a client, who wishes to trade a large position, and her dealer. Market-on-close contracts are generally not optimal in this principal-agent problem. In contrast, we provide conditions under which guaranteed VWAP contracts are optimal. These results question the usage of market-on-close contracts in practice, explain the usage of guaranteed VWAP contracts, and also suggest considerations for the design of financial benchmarks.

Keywords: agency conflict, benchmark manipulation, dealer-client relationship, foreign exchange fix, front-running, market-on-close, pre-trade hedging, principal trading, volume-weighted average price, VWAP

JEL Codes: G11, G14, G18, G23, D82, D86

*We thank Robert Almgren, Hendrik Bessembinder, Songzi Du (discussant), Piotr Dworczak, Lorenzo Garlappi, Will Gornall, Lawrence Harris (discussant), Burton Hollifield (discussant), Lukas Kremens, Pete Kyle, Kasper Larsen, Esen Onur, Andreas Park (discussant), Duane Seppi, Yang Song (discussant), Chester Spatt, Alberto Mokak Teguia, Nicholas Westray, and Tina Zhang (discussant) for helpful comments. We thank Colis Cheng and Kateryna Melnykova for excellent research assistance. Markus Baldauf and Christoph Frei gratefully acknowledge support by the Social Sciences and Humanities Research Council of Canada. We also acknowledge support from the 2019 SFS Cavalcade Best Paper in Asset Pricing Award. This paper evolved out of an earlier project entitled "Contracting for Financial Execution."

†University of British Columbia, Sauder School of Business, 2053 Main Mall, Vancouver, BC, Canada V6T 1Z2: baldauf@mail.ubc.ca, https://sites.google.com/site/mbaldauf/.

‡University of Alberta, 621 Central Academic Building, Edmonton, AB, Canada T6G 2G1: cfrei@ualberta.ca, https://www.math.ualberta.ca/∼cfrei/.

§Kellogg School of Management, Northwestern University, 2211 Campus Drive, Evanston, IL 60208: joshua.mollner@kellogg.northwestern.edu, https://sites.google.com/site/joshuamollner/.

1 Introduction

Many trading arrangements between a client and a dealer reference benchmark prices whose values are yet to be determined at the time of contracting. For example, a client might agree to conduct an off-market (i.e., over-the-counter) trade with her dealer, and the dealer would typically pursue offsetting on-market trades in the interim. Two of the most common benchmarks used to price the off-market trade between the client and the dealer are: (*i*) the market price prevailing at the end of the execution window, as in a *market-on-closed contract* (henceforth, simply 'MOC contract'), and (*ii*) the volume-weighted-average price prevailing over the window, as in a *guaranteed VWAP contract* (henceforth, simply 'VWAP contract'). Importantly, both of these pricing benchmarks are endogenous. It follows that the dealer may trade on the market in a way that influences the chosen benchmark to his advantage, and to the detriment of the client. Financial markets abound with examples of such trade-based benchmark manipulation.[1] Although this manipulative behavior is often ruled to be illegal, it continues to occur, likely because monitoring is difficult and because of legal gray areas between acceptable and prohibited forms of trading, as between pre-trade hedging and front-running (FINRA, 2013).

Of course, a predetermined, fixed benchmark price could not be manipulated in this way. But alternative arrangements involving such prices (e.g., 'arrival price' contracts, which refer to the price prevailing at the time of contracting) have their own disadvantage: they may leave the dealer with a great deal of exposure to price fluctuations, which could be inefficient if he is more risk averse or capital constrained than the client. Likewise, regulatory requirements that dealers hold capital in amounts corresponding to their exposure might produce the effects of risk aversion and induce the same considerations.

In this paper, we formulate a principal-agent problem so as to study how a client might contract in the presence of the two frictions alluded to above: (*i*) the possibility of hard-to-detect benchmark manipulation, and (*ii*) risk aversion on the part of the dealer. We begin by

[1] To illustrate this client-dealer conflict, one recent example involved a client agreeing, in advance, to trade $3.5 billion US dollars with her dealer in exchange for British pounds at the 'fix', a popular benchmark for foreign currency transactions. The client alleged that the dealer, HSBC, manipulated the fix by conducting a large volume of its offsetting trades during the fixing interval (Bloomberg, 2016; DOJ, 2018a,b). Several other examples that we discuss in the text—including some involving other asset classes—can be thought of as manifestations of the same underlying conflict (e.g., Bloomberg, 2013; WSJ, 2018; CFTC, 2008, 2013; WSJ, 2011; SEC, 2014).

considering the MOC contract. Although such arrangements are common and might appear innocuous at first glance, we show that they generally are not optimal. We then turn to the question of which contracts are. In general, the optimal contract depends in potentially complex ways on the various parameters. However, we show that under some conditions, the VWAP contract is uniquely optimal.

In the model, the client offers a contract to the dealer. The contract specifies that they will conduct an off-market trade at a future time T, in which the client will purchase a quantity normalized to one share from the dealer at a price that may be an arbitrary function of the history of market volumes and prices. If the dealer accepts the contract, then he obtains the share by purchasing it on the market, dividing his trades across the intervening periods. Trading creates price impact, and as a result, the total expected cost of a trade depends on how it is divided across time. Moreover, the severity of this price impact is influenced by exogenous market conditions (representing, e.g., the activity of other traders). The dealer's trading expertise is modeled as superior information about these market conditions, hence superior information about the optimal division of trades. There is also hidden action, in that the client cannot verifiably observe the dealer's on-market trades.

While the client cannot observe either the dealer's trades or his knowledge of market conditions, she can observe two outcomes that are influenced by them: market prices and volumes. The client's problem is to propose a contract that uses these market outcomes to formulate a benchmark that will serve to price her off-market trade with the dealer. In doing so, it is important to account for how the contract will shape the dealer's incentives about how to trade on the market: because the market outcomes are endogenous, the dealer may be incentivized to trade so as to manipulate the benchmark. Formally, the client chooses a contract and recommends a trading policy so as to minimize her expected payment to the dealer, subject to incentive compatibility and individual rationality constraints.

Our first contribution is to propose a general model of this contracting problem. Our second contribution is to show that the MOC contract fails to be optimal across nearly all parameterizations of this general model. (The result relies on only a mild condition about the nature of price impact.) MOC contracts (and analogous arrangements) are widely used in practice. But our result questions whether this is wise. Given the agency problem that we formulate and analyze, MOC contracts quite generally do not accomplish the goal of minimizing trading costs.

Our third contribution is to formulate conditions under which the VWAP, a widely-used benchmark, is in fact the price referenced by the optimal contract. Those conditions include the assumptions that price impact is purely temporary and that the dealer's knowledge of market conditions is sufficient to perfectly forecast volumes. Although these conditions might not be descriptive of all markets, they hold in approximation for many markets, and for such markets our results apply also in approximation. These conditions define what we call 'the specialized model.' Our initial description of it (in Section 4.1) portrays the dependence of prices and volumes on the dealer's trades in a reduced-form way, allowing for a wide class of functional forms. However, we also provide (in Section 4.2) a micro-foundation for one particular member of this class.

Solving for the first-best benchmark (i.e., the version of the problem in which the client can observe the dealer's trades and has his knowledge of market conditions) resembles the type of problem considered in the optimal execution literature. In that literature, a trader chooses how to work an order across time (or across venues) in order to minimize execution costs, taking as given an exogenously-specified market model that specifies how his trades create price impact. In the specialized model, the first-best trading strategy is equivalent to what is known in practice as a *volume participation strategy,* wherein the order is split over time so as to be proportional to the volume profile of the market. Whether the dealer's trading decisions actually accord with the first-best benchmark will, however, depend on the form of the contract.

In the specialized model, the optimal contract references the VWAP benchmark. The proof sketch involves two steps. First, the VWAP contract incentivizes the dealer to pursue a volume participation strategy, just as in the first-best benchmark. To build intuition, suppose that the dealer deviates from this first-best policy, by shifting some volume from one period to another. His expected trading costs increase (trivially, by the definition of first best), but the question is whether, in expectation, the VWAP (i.e., his payment from the client) will increase by more. And the answer is no, because the change in the VWAP is muted by the volume accounted for by outside traders. Second, note that when the dealer does pursue this volume participation strategy, his realized on-market trading costs equal the VWAP. Thus, the VWAP contract prescribes a payment to the dealer that covers his trading costs exactly, which perfectly insures him against price variation and also leaves him with zero surplus. The outcome therefore features efficient trading, efficient risk sharing,

and all surplus going to the client. Clearly, no contract can do better than that. The forces underlying this result may help explain the wide use of VWAP as a benchmark in practice, and the use of VWAP contracts in particular.

To illustrate uniqueness of the VWAP contract's optimality in the specialized model, it is instructive to consider a few familiar alternatives in more detail. First, consider the MOC contract, where the client pays the dealer at the price prevailing in period T. As discussed above, this contract is not optimal—not even in the general model. In the specialized model, we can provide an additional intuition: it incentivizes the dealer to deviate from the first-best trading policy by tilting his trades toward the last period, a behavior that is known in practice as 'banging the close' and that has been known to occur in conjunction with MOC contracts and analogous arrangements.[2] Intuitively, price impact from this deviation moves the price in period T, which in expectation creates a gap between this price and the average price that the dealer obtains for his on-market trades. Thus, unlike the VWAP contract, the MOC contract fails to induce efficient trading. Second, consider a contract in which the client pays the dealer at an exogenously predetermined and fixed price. Because it does not depend on future market outcomes, such a price cannot be manipulated by the dealer. But the dealer then assumes the price risk of the position, and if he is risk averse (or subject to minimum capital requirements) then he would require compensation for doing so. Thus, unlike the VWAP contract, fixed price contracts fail to induce efficient risk sharing when the dealer is risk averse.[3] Indeed, in the large class of contracts that we allow for, only the VWAP contract induces both efficient trading and efficient risk sharing.

We also apply our results to a number of settings that involve questions of benchmark design. Benchmarks are particularly relevant to markets in which it is difficult or impossible for clients to observe prices or volumes directly. Nevertheless, a third party, such as a regulator or platform, may publish a benchmark that summarizes these quantities—that benchmark might then be observed and contracted on. Section 6.1 considers how to design such a benchmark. Our results suggest that it may be desirable to compute this benchmark

[2]For example, consider the fixing trade mentioned in footnote 1 (involving a foreign exchange transaction by HSBC). This arrangement resembles a MOC contract. And indeed, at the heart of the ensuing litigation was this type of aggressive trading toward the end of the execution window (Bloomberg, 2016; DOJ, 2018a,b).

[3]Note that such arrangements do induce efficient risk sharing in the special case where the dealer is risk neutral. Moreover, by standard arguments, an appropriately-specified fixed price contract would be optimal (analogous to the so-called 'sell-the-firm' contract in canonical principal-agent models).

as the VWAP, since, in that case, it becomes possible for the client to replicate the contract that is optimal in the specialized model by referring to the benchmark.[4] Section 6.2 considers how to compute the settlement price for certain futures contracts. We show that our model explains a certain type of manipulative trading that occurs in practice. Moreover, the results of our specialized model suggest that use of a VWAP-based settlement price could mitigate this manipulation. Then in Section 6.3, we show that similar considerations argue for using VWAP when computing the net asset values of various funds.

Finally, we turn to a discussion of the extent to which our main results on the VWAP contract can be generalized. The contracts that the client optimizes over in the baseline fall in the class 'principal trading' arrangements, in which the dealer is the counterparty of the agent. Such arrangements are our primary focus, and the main result of our specialized model is that the VWAP contract is uniquely optimal in this class. However, in Section 7.1, we also consider 'agency trading' arrangements, in which the dealer merely acts as a matchmaker in locating a counterparty. We show that such arrangements can also be optimal in the model, and we discuss how a variety of unmodeled forces might make agency trading either more or less attractive than the VWAP contract. Section 7.2 considers an extension of the specialized model in which the dealer possesses more flexibility in making his trading decisions. Under certain additional assumptions on other aspects of the model, the VWAP contract remains optimal. Next, Section 7.3 investigates the implications of relaxing the assumptions of the specialized model that price impact is purely temporary and that the dealer's knowledge of market conditions is sufficient to perfectly forecast volumes. We establish a continuity result, showing that if these assumptions hold in approximation, then the main result extends in approximation: a VWAP-based contract will be nearly optimal.

2 Related literature

Empirics. At the heart of our model is a conflict of interest stemming from trade-based benchmark manipulation on the part of the dealer. An important benchmark is the closing price, and a number of papers have found empirical evidence of closing-price manipulation (Harris, 1989; Felixson and Pelli, 1999; Carhart, Kaniel, Musto and Reed, 2002; Hillion and

[4]This constitutes an interesting parallel to Duffie and Dworczak (2018) whose solution—albeit to a different problem—in some cases resembles a VWAP benchmark.

Suominen, 2004; Ben-David, Franzoni, Landier and Moussawi, 2013; Comerton-Forde and Putniņš, 2011, 2014; Henderson, Pearson and Wang, 2019). In some cases, those papers have also sought to identify the party behind the manipulation and to explain the motivation for it. Another important benchmark is the VIX, a measure of implied volatility calculated from prices of S&P 500 index options. Griffin and Shams (2017) have found evidence of trading patterns in out-of-the-money options that are consistent with VIX manipulation.

Our analysis focuses on just one aspect of a client's relationship with her broker or dealer in which a conflict of interest may exist. But other points of conflict have also been identified in the literature. For example, evidence suggests that brokers sometimes route client orders suboptimally in order to collect rebates (Battalio, Corwin and Jennings, 2016), to collect payments from high-frequency liquidity providers (Battalio, Hatch and Sağlam, 2019), or to use their own alternative trading systems (Anand, Samadi, Sokobin and Venkataraman, 2019). In addition, Barbon, Di Maggio, Franzoni and Landier (2019) provide evidence that brokers leak information about client trades to their other clients, where the latter then engage in predatory trading that reduces execution quality for the former.

Theory. Previous literature has also studied conflicts of interest between broker-dealers and clients. One type of conflict is created by so-called dual trading (e.g., Röell, 1990; Fishman and Longstaff, 1992; Bernhardt and Taub, 2008) in which a dealer may engage in proprietary trading alongside the trades that he makes on behalf of his clients. In contrast, we analyze a setting in which conflict may arise even in the absence of any proprietary net positions of the dealer.

The choice of a benchmark price is important in our model because of its effect on the trading incentives of the dealer. Similarly, benchmark choices shape incentives in many other aspects of financial markets. Interest rate benchmarks are one example. Banks may have incentives to move these rates in a particular direction, and a benchmark administrator may wish to select a benchmark that is less prone to manipulation of this sort (Duffie and Dworczak, 2018; Coulter, Shapiro and Zimmerman, 2018). Benchmarks for assessing the performance of fund managers are another example. Fund managers may have incentives to distort their trading decisions so as to perform well under the chosen metric, so that wide use of a manipulation-proof performance measure could be beneficial (Ingersoll, Goetzmann,

Spiegel and Welch, 2007).[5] Finally, Duffie, Dworczak and Zhu (2017) analyze how benchmarks affect the incentives of traders in search markets, finding that the publication of a benchmark can raise social surplus.

In the specialized model, the contract that emerges as optimal, the VWAP contract, is fairly simple. In that sense, this paper relates to a literature on foundations for contracts possessing simple features such as linearity (e.g., Holmström and Milgrom, 1987; Carroll, 2015).

While our primary focus is on the contracting friction between a client and her dealer, solving for the first-best benchmark is equivalent to the type of problem considered in the optimal execution literature (e.g., Bertsimas and Lo, 1998; Almgren and Chriss, 2001). In the specialized model, the first-best solution amounts to a volume participation strategy. Similarly, Kato (2015) provides conditions under which such a participation strategy achieves optimal execution. Note that if a participation strategy is used, then the price paid for the order necessarily equates to the VWAP over the trading period. Consequently, the participation strategy can be equivalently thought of as a strategy designed to target VWAP. Given the importance of VWAP in practice, a considerable literature studies how to devise such strategies (e.g., Humphery-Jenner, 2011; Frei and Westray, 2015; Cartea and Jaimungal, 2016). We depart from this literature in the sense that our primary focus is not on the first-best benchmark, but rather on the second-best—that is, the version of the problem in which there is a client-dealer relationship and an agency problem between them stemming from the client's inability to observe the dealer's trades and the dealer's superior knowledge of market conditions. In the specialized model, our main finding is that the VWAP contract is the unique contract that incentivizes the dealer to pursue a volume participation strategy, just as he does in the first-best benchmark. In consequence, the client actually obtains her first-best payoff despite these additional frictions.

[5] Also related are papers that analyze how the composition of a benchmark portfolio may affect trading behavior, asset pricing, and other outcomes (e.g., Brennan, 1993; Cuoco and Kaniel, 2011; Basak and Pavlova, 2013; Buffa, Vayanos and Woolley, 2019; Kashyap, Kovrijnykh, Li and Pavlova, 2019). However, these papers largely take the benchmark portfolio as given and do not investigate implications for optimal benchmark design.

3 General model

A client (the principal) needs to purchase a fixed quantity of a particular security, which we normalize to one share, and she offers her dealer (the agent) a contract regarding the intended trade.[6] If the dealer accepts the offer contract, then he purchases from the market the share that he will subsequently sell to the client. Importantly, the payment that he obtains from the client might, depending on the terms of the contract, be influenced by his trading activity. The main friction involves hidden action: the client cannot observe the dealer's precise sequence of on-market trades.

3.1 Model details

Trading. There are finitely many discrete trading periods $t \in \{1, 2, \ldots, T\}$.[7,8] Before time 1, the client contracts to purchase the share from the dealer after time T.[9] In advance of this transaction, the dealer must purchase the required share in the market. Letting x_t denote the number of shares purchased by the dealer in period t, we therefore require $\sum_{t=1}^{T} x_t = 1$. We also require all x_t to be nonnegative so that the dealer does not sell in any period.[10] We refer to a vector $\boldsymbol{x} = (x_t)_{t=1}^{T}$ that satisfies these conditions as a *trading schedule*.

Remark. Note that the requirement $\sum_{t=1}^{T} x_t = 1$ implies that the dealer ends with a net inventory of zero. Thus, the dealer merely intermediates between the client and the market, neither trading with the client out of his own inventory nor taking on a proprietary position of his own. Thus, our model shuts down sources of conflict that are already well-understood, thanks to the literature on dual trading. Nevertheless, there remains potential for conflict in terms of how the dealer chooses the timing of his intermediating trades.

[6] Symmetric analysis applies to the case in which the client's need is to sell.

[7] In some cases, the model could be alternatively interpreted to capture settings in which trading is done by splitting orders not across time periods but across venues (or even across both time and venues). For these interpretations, t may be taken to index venues (or time-venue pairs).

[8] Alternatively, we could consider a continuous-time model with trading during an interval $[0, T]$; see the remark at the end of Section 7.2.1.

[9] The horizon T is taken to be an exogenous parameter for the purposes of our model. However, one might think of it as being derived from the degree of the client's desire for timely execution.

[10] This restriction is only to simplify the presentation; with suitable adjustments to the model (e.g., to distinguish between signed and unsigned quantities), the subsequent results would hold even without the restriction.

Information structure. The dynamics of prices and market volumes will depend not only on the dealer's trading schedule $\boldsymbol{x}$ but also on an exogenous random variable. To model the possibility of the dealer having superior knowledge about these dynamics—which might, after all, be why the client engages the dealer in the first place—we suppose that this is a multidimensional random variable with two components $(\boldsymbol{\eta}, \boldsymbol{\varepsilon})$.[11] The first component $\boldsymbol{\eta}$ is observed by the dealer (but not by the client) before trading begins at time 1, but after contracting takes place.[12] Neither the dealer nor the client observe the second component $\boldsymbol{\varepsilon}$.

Prices and volumes. Let $\boldsymbol{p} = (p_t)_{t=1}^T$ denote the sequence of prices, and let $\boldsymbol{v} = (v_t)_{t=1}^T$ denote the sequence of market volumes (including but not limited to the dealer's trades). In general, both of these will be functions of $\boldsymbol{x}$ and $(\boldsymbol{\eta}, \boldsymbol{\varepsilon})$, although that dependence will often be suppressed in the notation.

Remark. Note that we take a reduced-form approach in specifying how the market outcomes $\boldsymbol{p}$ and $\boldsymbol{v}$ depend on the trades of the dealer. This aspect of our model resembles the approach of the optimal execution literature. Nevertheless, in Section 4.2, we provide an example of how these dependencies might be micro-founded.

Contracts. In specifying the set of feasible contracts, we imagine that the client can verifiably observe the sequence of prices $\boldsymbol{p}$ and the sequence of volumes $\boldsymbol{v}$ at time T. We allow the client to offer contracts that are arbitrary functions of these market outcomes.[13] Formally, the set of contracts that we allow for consists of measurable functions $\tau : \mathbb{R}^T \times \mathbb{R}_{++}^T \to \mathbb{R}$, specifying that the client will pay $\tau(\boldsymbol{p}, \boldsymbol{v})$ to the dealer at time T in exchange for one share of the security when the market outcomes are $\boldsymbol{p}$ and $\boldsymbol{v}$.

[11] When we specialize the model in Section 4.1, we further assume that each component is a vector consisting of one real-valued random variable per period: $\boldsymbol{\eta} = (\eta_t)_{t=1}^T$ and $\boldsymbol{\varepsilon} = (\varepsilon_t)_{t=1}^T$. But that additional structure is not necessary here.

[12] Under this formulation, the dealer does not possess hidden information at the time of contracting. Nevertheless, the subsequent results about optimality of the VWAP contract in the specialized model would remain unchanged if the dealer were to observe $\boldsymbol{\eta}$ before contracting takes place. This is because (*i*) when accepting the optimal contract τ^{VWAP}, the dealer receives a payoff of zero for every realization of $\boldsymbol{\eta}$ and so would not wish to condition his decision to accept or reject on the basis of $\boldsymbol{\eta}$; and (*ii*) from offering the optimal contract τ^{VWAP}, the client already receives her first-best payoff, and so she could not benefit from screening on the basis of $\boldsymbol{\eta}$.

[13] In contrast, we imagine that other quantities are not verifiably observable, and so they cannot be contracted upon (except through their influence on prices and volumes). For example, the realization of $\boldsymbol{\eta}$ constitutes information of the dealer that is hidden from the client. Similarly, the trading schedule $\boldsymbol{x}$ constitutes actions of the dealer that are hidden from the client.

This formulation permits the client to propose a wide variety of trading arrangements including several familiar ones. Many contracts observed in practice specify that the client pay the dealer according to a particular yet-to-be-determined benchmark price. Common benchmarks include (*i*) the closing price, which corresponds to the MOC contract $\tau^{MOC} \equiv p_T$, (*ii*) the time-weighted average price (TWAP), which corresponds to $\tau^{TWAP} \equiv \frac{1}{T}\sum_{t=1}^{T} p_t$, and (*iii*) the VWAP, which corresponds to $\tau^{VWAP} \equiv \frac{\sum_{t=1}^{T} p_t v_t}{\sum_{s=1}^{T} v_s}$. Another set of possibilities are contracts that specify a predetermined fixed price, which corresponds to $\tau = \tau_0$ for some constant τ_0.[14]

Timing. The timing of events is as follows. First, the client offers a contract τ to the dealer. The dealer either accepts or rejects the contract. If he rejects the contract, then he receives an outside option of 0. If he accepts, then the dealer chooses a *trading policy* $\boldsymbol{x}(\cdot)$, which is a rule for mapping information that he will obtain into a trading schedule. Our analysis will consider two classes of trading policies:

- The *dynamic trading policies* are the measurable vectors $\boldsymbol{x}(\cdot) = (x_t(\cdot))_{t=1}^{T}$ satisfying (*i*) for all t, x_t depends on $\boldsymbol{\eta}$ and $(p_s)_{s=1}^{t-1}$, (*ii*) for all t, $x_t\big(\boldsymbol{\eta}, (p_s)_{s=1}^{t-1}\big) \geq 0$ almost surely, and (*iii*) $\sum_{t=1}^{T} x_t\big(\boldsymbol{\eta}, (p_s)_{s=1}^{t-1}\big) = 1$ almost surely.
- The *static trading policies* are given by the subset of dynamic trading policies in which each x_t depends only on $\boldsymbol{\eta}$. In other words, a static trading policy is a function that maps each $\boldsymbol{\eta}$ into a trading schedule.

Next, $\boldsymbol{\eta}$ is realized. Given $\boldsymbol{\eta}$ and $\boldsymbol{x}(\cdot)$, prices $\boldsymbol{p}$, volumes $\boldsymbol{v}$, and the trading schedule $\boldsymbol{x}$ are realized. After trading, the client pays the dealer as specified by τ.

The dealer's payoffs. The dealer's utility function over money is denoted u, which is assumed to be strictly increasing and weakly concave. From accepting a contract τ and choosing a trading policy $\boldsymbol{x}(\cdot)$, the dealer receives expected utility

$$\mathbb{E}\big[u\big(\tau(\boldsymbol{p}, \boldsymbol{v}) - \boldsymbol{p} \cdot \boldsymbol{x}\big)\big].$$

[14] Often, the fixed price τ_0 would be specified in relation to the price prevailing at the time of contracting (the 'arrival price'). For example, it might be that price itself, or that price plus a commission.

The client's payoffs. We assume that if the dealer rejects the offered contract, then the client receives a payoff of negative infinity.[15] On the other hand, if the contract is accepted, then the client is risk neutral over monetary outcomes. In particular, if the dealer accepts a contract τ and chooses a trading policy $\boldsymbol{x}(\cdot)$, then the client receives expected utility

$$-\mathbb{E}\big[\tau(\boldsymbol{p},\boldsymbol{v})\big].$$

Remark. The client of our model most closely resembles a large institution, but one that is unsophisticated as regards financial trading. 'Large' because the client's order may affect prices and because of our assumption that the client is risk neutral. And 'unsophisticated' because the client has inferior information about price and volume dynamics and for that reason is accessing the market through an external dealer rather than one in house. As a concrete example of a large and unsophisticated client, consider the episode involving HSBC mentioned in footnote 1 in the introduction. In that case, the client was Cairn Energy Plc, a British oil explorer. In connection with selling part of its ownership interest in a subsidiary company, Cairn Energy realized a one-time need to convert $3.5 billion USD into British pounds.

3.2 The client's problem

Given this framework, the client's problem is to choose a contract τ, as well as a 'recommended trading policy' $\boldsymbol{x}(\cdot)$ to maximize her expected utility (equivalently, to minimize her expected payment) subject to individual rationality and incentive compatibility constraints of the dealer:

$$\begin{aligned}
&\min_{\tau,\boldsymbol{x}(\cdot)} \mathbb{E}[\tau(\boldsymbol{p},\boldsymbol{v})] \quad \text{subject to} \\
&\mathbb{E}\big[u\big(\tau(\boldsymbol{p},\boldsymbol{v})-\boldsymbol{p}\cdot\boldsymbol{x}\big)\big] \geq u(0) && \text{(IR)} \\
&\forall \hat{\boldsymbol{x}}(\cdot) : \mathbb{E}\big[u\big(\tau(\boldsymbol{p},\boldsymbol{v})-\boldsymbol{p}\cdot\boldsymbol{x}\big)\big] \geq \mathbb{E}\big[u\big(\tau(\boldsymbol{p},\boldsymbol{v})-\boldsymbol{p}\cdot\hat{\boldsymbol{x}}\big)\big] && \text{(IC)}
\end{aligned}$$

[15] Assuming that the client receives infinite disutility from rejection simply ensures that acceptance occurs on path, which is the interesting case. But the assumption can be significantly relaxed. Acceptance would similarly occur on path provided the client's disutility from rejection is any exogenous value that exceeds her expected payment to the dealer under the optimal contract.

This is the *second-best problem* (or simply, 'the client's problem'). A contract τ is *optimal* if it is part of a solution to this problem.

Remark. Note that this formulation closely follows standard approaches in contract theory (e.g., Mirrlees, 1976; Hölmstrom, 1979). For instance, we consider just a single principal (the client) and a single agent (the dealer). The client has all the bargaining power.[16] The client is risk-neutral. The client can choose the dealer's trading policy so long as (IR) and (IC) are satisfied—thus, she can resolve indifference on the part of the dealer however she prefers. Some of those items might not perfectly accord with aspects of the markets that we seek to model. Nevertheless, the standard approach seems a natural starting point.

As a benchmark, we also consider the *first-best problem:* the version in which the client can observe the dealer's trades and has his knowledge of price dynamics—formally, the version in which τ can also condition directly on $\boldsymbol{x}$ and $\boldsymbol{\eta}$. A trading policy is *first-best* if it is part of a solution to this version of the problem.

Remark. For the first-best problem, note that the client can use a forcing contract to implement any desired $\boldsymbol{x}(\cdot)$ while satisfying (IC). Then given (IR), it is optimal to set $\tau(\boldsymbol{p}, \boldsymbol{v}) = \boldsymbol{p} \cdot \boldsymbol{x}$ when the dealer acts according to the desired $\boldsymbol{x}(\cdot)$. Plugging that in to her objective, the client's problem reduces to $\min_{\boldsymbol{x}(\cdot)} \mathbb{E}\big[\boldsymbol{p} \cdot \boldsymbol{x}\big]$. In other words, solving for the first-best trading policy reduces to a problem of optimal execution.

3.3 Suboptimality of the MOC contract

Given this fairly general setup, we are ready to state our first result. The aforementioned MOC contract $\tau^{MOC} = p_T$ is generally not optimal, in the sense that it does not solve the client's problem.

Condition 1. Prices and the dealer's utility function satisfy the following:

(i) There exists some time period $s \in \{1, 2, \ldots, T-1\}$ such that $\mathbb{E}[\bar{p}_T|\boldsymbol{\eta}] > \mathbb{E}[\bar{p}_s|\boldsymbol{\eta}]$ on a set of scenarios with strictly positive probability, where $\bar{p}_t$ denotes the price in period t when the dealer uses the trading schedule $\boldsymbol{x} = (0, \ldots, 0, 1)$.

(ii) Prices $\boldsymbol{p}$ are continuously differentiable in x_s at $\boldsymbol{x} = (0, \ldots, 0, 1)$.

(iii) The dealer's utility function $u(w)$ is continuously differentiable at $w = 0$.

[16] This might be interpreted as an 'as if' representation of perfect competition among multiple dealers.

Parts (ii) and (iii) of the condition are simple continuity requirements. The main substance of Condition 1 lies in its part (i), which rules out the case in which, under the first-best trading policy, the dealer almost always buys only in the last period. This is a natural and mild condition. Indeed, it is implied, for example, when expected prices without price impact are nondecreasing and there is a strictly positive price impact of buy orders.

Proposition 1. *If Condition 1 holds, then τ^{MOC} is not optimal.*

The proofs of Proposition 1 and of all other results are relegated to the Appendix. To see the intuition for this result, first note that under τ^{MOC}, if the dealer were to always buy only in the last period, then his payment from the client would always equal his trading costs, and his payoff would equal his outside option. But if Condition 1 holds, then we show that he can do better by sometimes shifting some volume to an earlier period s, obtaining an expected payoff that strictly exceeds his outside option.[17] Thus, τ^{MOC} leaves some money on the table and therefore cannot be optimal.[18]

MOC contracts are widely used in practice. However, Proposition 1 demonstrates that they generally do not accomplish the goal of minimizing trading costs given the agency problem that we model. As such, the proposition might be interpreted as calling into question the wisdom of these trading arrangements. And because the model is fairly general, the result is a strong one.

4 Specialized model

Having established that τ^{MOC} generally does not achieve optimality, we now turn to the question of which contracts do. The general model laid out in the previous section does not lend itself to a solution unless additional structure is imposed. But with additional structure, we can get a sharp result: the VWAP contract is uniquely optimal in a certain class of settings. In this sense, our model can explain usage of the VWAP contract in practice. Section 4.1 defines the class of settings, and Section 4.2 provides a micro-foundation for a particular element of that class.

[17] He can do this whether he has access to only the static trading policies or to any superset thereof (e.g., the dynamic trading policies).

[18] Indeed, the proposition could be generalized to $\tau = p_t$ not being optimal for any fixed $t \in \{1, 2, \ldots, T\}$ (with an appropriate modification to Condition 1).

4.1 Model details

This section states the additional conditions under which the VWAP contract attains optimality. The primary substance of these conditions is to require (*i*) that the dealer is limited to static trading policies, (*ii*) that price impact is temporary, and (*iii*) that volumes can be perfectly forecast given the dealer's knowledge of market conditions. These assumptions are strong, but with them, we obtain this sharp result despite making relatively weak assumptions about other aspects of the model.

Trading policies. We assume that the dealer is limited to static trading policies. Recall that a static trading policy is one in which each x_t can depend only on $\boldsymbol{\eta}$. In particular, x_t cannot depend on $(p_s)_{s=1}^{t-1}$.

Remark. In one interpretation of the model, t indexes venues (*cf.* footnote 7). For that interpretation, the restriction to static trading policies would be particularly natural—indeed, if the dealer accesses different venues simultaneously, then his trading at one venue cannot depend on outcomes observed at other venues. But for the interpretation in which t indexes time, requiring the dealer to commit to an entire trading schedule at the outset might be seen as an artificial constraint on his actions. However, we revisit this assumption in Section 7.2, showing that our results survive even if the dealer has access to the broader set of dynamic trading policies, provided that certain conditions are imposed on other aspects of the model.

Note also that not even dynamic trading policies permit the choice of x_t to depend on the contemporaneous price p_t. Thus, the dealer should be thought of as conducting his trades using market orders, in both the static and dynamic cases.

Information structure. We now impose additional structure on the random variables that affect the dynamics of prices and market volumes: $\boldsymbol{\eta}$ (which the dealer observes) and $\boldsymbol{\varepsilon}$ (which the dealer does not observe). Each of these will be a vector consisting of one real-valued random variable per period: $\boldsymbol{\eta} = (\eta_t)_{t=1}^T$ and $\boldsymbol{\varepsilon} = (\varepsilon_t)_{t=1}^T$. For reasons that will become clear below, when we parametrize how these variables influence prices and volumes, we use the following terminology. We refer to $\boldsymbol{\eta}$ as the vector of *market conditions* (or, more precisely, the dealer's knowledge about market conditions).[19] And we refer to $\boldsymbol{\varepsilon}$ as the vector of *price*

[19]For parsimony, we assume that all relevant aspects of the dealer's knowledge of the market conditions in period t can be captured by the single scalar η_t. Our approach is to treat η_t as an abstract object that affects

shocks.

We require each η_t to take strictly positive values, but we do not impose any assumptions on the joint distribution of $\boldsymbol{\eta}$. In particular, its elements do not need to be independent, nor do they need to be identically distributed. We require each ε_t to have the same expectation conditional on $\boldsymbol{\eta}$, namely, $\mathbb{E}[\varepsilon_t|\boldsymbol{\eta}] = \mu$ almost surely for all t and some constant μ. However, we impose no further assumptions on the distribution of $\boldsymbol{\varepsilon}$. In particular, its elements do not need to be independent of each other or of $\boldsymbol{\eta}$, and neither do they need to be identically distributed.

Prices. We assume the price that prevails in period t is

$$p_t = h\left(\frac{x_t}{\eta_t}\right) + \varepsilon_t,$$

where h is a strictly increasing function such that $yh(y)$ is strictly convex. Note that $\boldsymbol{\varepsilon}$ can be thought of as the price dynamics that would prevail in the absence of the dealer's trading, and our assumption of constant conditional expectations nests the case in which this evolves as a random walk.

Note also that for a given number of shares x_t, the impact on price is influenced by the market conditions prevailing in the period. That is, the price impact depends on x_t measured relative to η_t, and not on x_t itself. In light of this, the market condition η_t can be thought of as parametrizing how steep price impact will be in period t (i.e., the liquidity of period t).[20] A class of price impact functions nested by our approach is $h(x_t/\eta_t) = (x_t/\eta_t)^a$ for $a > 0$. Such specifications are supported by both theoretical and empirical results in the literature, where the predominant configurations are between $a = 0.5$ (square root price impact) and $a = 1$ (linear price impact).[21]

Remark. Note also that the price depends only on contemporaneous values of x_t and η_t. In

price and volume in ways specified below. Nevertheless, one concrete example (to which we will return in Section 7.2) is for η_t to represent the unsigned volume traded on the market by outside traders in period t. Another example will be pursued in Section 4.2, where η_t represents the number of outside traders.

[20] In the case where η_t represents outside volume, using precisely the ratio x_t/η_t makes the price impact dimensionless (Almgren, Thum, Hauptmann and Li, 2005).

[21] Using a large data set on US equity, Almgren, Thum, Hauptmann and Li (2005) estimate an exponent a of 0.6 while Mastromatteo, Tóth and Bouchaud (2014) report exponents a in the range of 0.4–0.7 across different markets (equities, futures, and foreign exchange).

that sense, the specialized model should be interpreted as one of temporary price impact. As an example of a market in which this assumption of purely temporary price impact may hold in approximation, consider stock index futures. Berkman, Brailsford and Frino (2005) quantify permanent price impact in that setting, finding it to be small. As they point out, this is consistent with the theory of Subrahmanyam (1991), who develops a model in which markets for basket products feature relatively little informed trading, which therefore suggests that permanent price impact is likely to be similarly small for other types of basket products (e.g., ETFs, index funds).

Volumes. We assume the market volume (including the dealer's trades) at time t is given by $v(x_t, \eta_t)$, where v is a positive function with domain $\text{dom}(v) \subseteq \mathbb{R}_+ \times \mathbb{R}_{++}$ that for all $x \neq 0$ takes the form $v(x, \eta) = xV(x/\eta)$ for some function $V(y)$ that is strictly decreasing for $y \neq 0$. For example, in the case where η_t represents the unsigned outside volume traded in period t, $v(x_t, \eta_t) = x_t + \eta_t$, which is indeed of the desired form, with $V(y) = 1 + 1/y$ for $y > 0$.

This assumption implies that volume $v(x, \eta)$ is increasing in η. This is natural: in light of the fact that η_t can be thought of as parametrizing the steepness of price impact in period t, we obtain the intuitive relationship that, holding fixed the dealer's volume, price impact is smaller when (total market) volume is larger. The assumption also implies that volume $v(x, \eta)$ is homogenous of degree one. This is natural as well, since it implies that price impact is invariant to the scale of the market. Indeed, in the case where $v(x, \eta)$ is homogenous of degree one, proportionate increases in x and η correspond to an increase in scale of the market: both the dealer's volume and (total market) volume increase at the same rate. And at the same time, price impact as measured by $h(x/\eta)$ remains constant. The following proposition establishes that the above assumption on $v(x, \eta)$ not only implies these two realistic properties but also is implied by them.

Proposition 2. *Let v be a positive function with domain $\text{dom}(v) \subseteq \mathbb{R}_+ \times \mathbb{R}_{++}$. The following are equivalent:*

(i) $v(x, \eta) = xV(x/\eta)$ for $x \neq 0$ and a function $V(y)$ that is strictly decreasing for $y \neq 0$.

(ii) $v(x, \eta)$ is homogeneous of degree one and strictly increasing in η for $(x, \eta) \in \text{dom}(v)$ with $x \neq 0$.

Remark. Whereas we have modeled prices as stochastic, this formulation assumes that volumes are a deterministic function of the dealer's trading schedule and his knowledge of market conditions.[22,23] This is, of course, not completely realistic: the path of volumes—or, following footnote 22, the path of relative volumes—is not perfectly predictable in practice. Yet, there do exist many known empirical regularities in trading volume (e.g., the so-called 'liquidity smile'). This stands in stark contrast to the unpredictability of the price path—under the weak-form efficient market hypothesis, there are *no* such regularities in prices—which is what underlies our motivation for modeling prices as stochastic and volumes as deterministic, conditional on the dealer's information set. In many markets, a sophisticated trader would be able to forecast volumes with a fair degree of accuracy (e.g., Exhibit 9 of Satish, Saxena and Palmer, 2014), so that this assumption of predicable volumes may hold in approximation.

For these last two assumptions of temporary price impact and forecastable volume, Section 7.3.1 establishes an intuitive continuity result: if the assumptions hold in approximation, then our main result also holds in approximation. Finally, we point out that, although this section imposes some strong assumptions, it nevertheless allows for a great deal of generality in the price impact function h, the volume function v, the dealer utility function u, the distribution of market conditions $\boldsymbol{\eta}$, and the distribution of price shocks $\boldsymbol{\varepsilon}$.

4.2 Micro-foundation

In the model as it is laid out in Sections 3.1 and 4.1, prices and volumes in a period t are jointly influenced by both the dealer's trades x_t and the random variables (η_t, ε_t). Our approach encompasses a broad class of reduced-form dependencies. In this section, we complement that previous analysis by presenting a micro-foundation for one of these functional

[22] Our later arguments go through if price impact and volume are of the forms $h\big(\frac{x_t}{E\eta_t}\big)$ and $v(x_t, \eta_t) = x_t V\big(\frac{x_t}{E\eta_t}\big)$, respectively, where the unobserved random vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_T, E)$ contains a random variable E that is independent from $(\varepsilon_t)_{t=1}^T$. As before, the dealer observes $(\eta_t)_{t=1}^T$, which provides only a noisy signal of the 'true' market conditions $(E\eta_t)_{t=1}^T$. Nevertheless, this signal does fully reveal the relative market conditions $\Big(\frac{E\eta_t}{\sum_{s=1}^T E\eta_s}\Big)_{t=1}^T$.

[23] One interpretation of this assumption is that the dealer is the only trader who is large enough to appreciably influence market prices and volumes. Traders who constitute the rest of the market might be thought of as atomistic, meaning that, as by a law of large numbers, they create no aggregate randomness in the volume traded. In that sense, our framework is in the spirit of models that feature a monopolist and a competitive fringe, a connection that we make more explicit in the micro-foundation of Section 4.2.

forms. (This section could be skipped by a reader who wishes to proceed directly to the main results.)

Suppose the client and the dealer are as before. What is different is that we more precisely specify the other traders in the market and the nature of trading. In addition to the dealer, there exists, in every period, a continuum of outside traders who receive liquidity shocks and trade with demand schedules. The quantity η_t, previously referred to as 'market conditions,' now denotes the measure of outside traders present in period t. The main appeal of this micro-foundation is that prices and volumes will be derived endogenously from a market-clearing condition. Moreover, they take a form that is nested by our previous analysis.

Outside traders. In each period t, a positive measure η_t of outside traders arrive, who then depart the market after the period has ended. In all other respects, $\boldsymbol{\eta} = (\eta_t)_{t=1}^T$ retains the properties that were assumed of $\boldsymbol{\eta}$ previously.

Trading. In each period t, the dealer submits a market order x_t, as before. In addition, each of the outside traders submits a demand schedule. We assume that an outside trader i arriving in period t submits the schedule

$$y_i(p_t) = \theta_i - p_t,$$

where $\theta_i = \psi_i + \varepsilon_t$.[24] The idiosyncratic component ψ_i is an independent draw from a standard normal distribution, and the common components $\boldsymbol{\varepsilon} = (\varepsilon_t)_{t=1}^T$ are random variables, which retain the properties that were assumed of $\boldsymbol{\varepsilon}$ previously. The security is in zero net supply, and the price p_t is chosen to clear the market.

[24]There might be many ways to micro-found this form of demand schedule, but one is as follows. Suppose the security has a liquidation value of $V \sim N(0,1)$. Suppose that trader i arrives to the market having received an endowment shock of $-\theta_i$ units of the security and that his utility over final wealth is given by $u_i(w) = -\exp(-w)$. From acquiring y shares at the per-share price p_t, he receives expected utility $v_i(y, p_t) = \mathbb{E}\left[-\exp(-[V(y-\theta_i) - p_t y])\right] = -\exp(p_t y + (y-\theta_i)^2/2)$. Given a continuum of outside traders, trader i acts as a price taker. Taking the first-order condition with respect to y, we obtain that trader i optimally submits the demand schedule stated in the text.

Solution. In period t, η_t outside traders are active. Substituting for $\theta_i = \psi_i + \varepsilon_t$ in the demand schedule submitted by outside trader i, the market-clearing condition becomes

$$x_t + \eta_t \int_{-\infty}^{\infty} (z + \varepsilon_t - p_t)\, \phi(z)\, dz = 0,$$

where $\phi(\cdot)$ denotes the standard normal probability density function. Likewise, $\Phi(\cdot)$ will denote the standard normal cumulative distribution function in what follows. Solving for price, the market-clearing condition becomes

$$p_t = \frac{x_t}{\eta_t} + \varepsilon_t.$$

We therefore obtain linear price impact for the dealer's trades, which is indeed nested by our previous analysis (with $h(y) = y$).

Having submitted the demand schedule $y_i(p_t)$, the number of shares purchased by trader i at this market-clearing price will be

$$y_i\left(\frac{x_t}{\eta_t} + \varepsilon_t\right) = \theta_i - \frac{x_t}{\eta_t} - \varepsilon_t = \psi_i - \frac{x_t}{\eta_t}.$$

Thus, trader i will be a buyer only if $\psi_i > \frac{x_t}{\eta_t}$. In consequence, the number of shares bought in period t (and therefore also the volume traded) will be

$$\begin{aligned} v(x_t, \eta_t) &= x_t + \eta_t \int_{\frac{x_t}{\eta_t}}^{\infty} \left(z - \frac{x_t}{\eta_t}\right) \phi(z)\, dz \\ &= x_t \Phi\left(\frac{x_t}{\eta_t}\right) + \eta_t \phi\left(\frac{x_t}{\eta_t}\right). \end{aligned}$$

This function is positive since $\eta_t > 0$ and of the form $v(x_t, \eta_t) = x_t V\left(\frac{x_t}{\eta_t}\right)$ for a strictly decreasing function V, so that the expression for volume is also nested by our previous analysis. Indeed, the function $V(y) = \Phi(y) + \frac{1}{y}\phi(y)$ is strictly decreasing because

$$V'(y) = \phi(y) + \frac{1}{y} \underbrace{\phi'(y)}_{=-y\phi(y)} - \frac{1}{y^2}\phi(y) = -\frac{1}{y^2}\phi(y) < 0.$$

Moreover, note that volume in period t is a deterministic function of (x_t, η_t). It is because this micro-foundation assumes a competitive 'fringe' of atomistic outside traders that there is no aggregate randomness.

5 Optimal contracts

Throughout this section, we restrict ourselves to the conditions of the specialized model. Our main results concern the optimality of the VWAP contract under those conditions. Before coming to them, we begin by characterizing the first-best trading policy.

5.1 The first-best trading policy

As observed in Section 3.2, a first-best trading policy minimizes the expected cost of acquiring one share: it is how the client would trade herself if she were to possess the dealer's knowledge of market conditions. Formally, a trading policy $\boldsymbol{x}(\cdot)$ is first best if it satisfies the following condition almost surely

$$\boldsymbol{x}(\boldsymbol{\eta}) \in \arg\min_{\boldsymbol{x}} \mathbb{E}[\boldsymbol{p} \cdot \boldsymbol{x} | \boldsymbol{\eta}].^{25} \tag{FB}$$

In other words, solving for the first-best trading policy reduces to a problem of optimal execution.

To simplify the exposition, we focus on trading policies that satisfy (FB) for all realizations of the market conditions $\boldsymbol{\eta}$. Given the structure of the specialized model, this determines a unique first-best trading policy, which Lemma 3 characterizes as that in which the dealer's trades x_t are proportional to market conditions η_t. Intuitively, when η_t is large, price impact in that period is small, and so it is optimal to trade a larger volume x_t. To prove the result, we show that this policy equates the marginal cost of trading an extra unit across time periods. The final part of the lemma states that this policy leads the dealer to use a volume participation strategy: he trades in proportion to the volume profile of the market.

[25] We remind the reader that the dependence of $\boldsymbol{p}$ on $\boldsymbol{x}$, $\boldsymbol{\eta}$, and $\boldsymbol{\varepsilon}$ is suppressed in the notation. Similarly, $\boldsymbol{v}$ depends on $\boldsymbol{x}$ and $\boldsymbol{\eta}$, and we will make that dependence explicit with the notation $\boldsymbol{v}(\boldsymbol{x}, \boldsymbol{\eta}) = (v(x_t, \eta_t))_{t=1}^T$ where it helps to clarify the dependence structure.

Lemma 3. *The first-best trading policy is*

$$\boldsymbol{x}^{FB}(\boldsymbol{\eta}) = \left(\frac{\eta_t}{\sum_{s=1}^{T} \eta_s} \right)_{t=1}^{T}.$$

The expected trading cost incurred by this policy is $\mathbb{E}[\boldsymbol{p} \cdot \boldsymbol{x}^{FB}(\boldsymbol{\eta})] = \mu + \mathbb{E}\Big[h\Big(\frac{1}{\sum_{t=1}^{T} \eta_t}\Big)\Big]$. *Moreover, the following equality holds*

$$\boldsymbol{x}^{FB}(\boldsymbol{\eta}) = \left(\frac{v(x_t^{FB}(\boldsymbol{\eta}), \eta_t)}{\sum_{s=1}^{T} v(x_s^{FB}(\boldsymbol{\eta}), \eta_s)} \right)_{t=1}^{T}.$$

The optimality of volume participation strategies in the specialized model is consistent with and might help to explain their extensive usage in practice (e.g., TheTrade, 2019). Note, moreover, that if the dealer obeys such a strategy—so that the relative volume curve of his trades corresponds to that of the market—then the realized trading cost incurred by this policy, $\boldsymbol{p} \cdot \boldsymbol{x}^{FB}(\boldsymbol{\eta})$, will always equal the realized market VWAP.

5.2 Optimality of VWAP

Having derived the first-best trading policy, we now turn our attention to the second best. Recall that the client's problem is to choose τ and $\boldsymbol{x}(\cdot)$ to minimize her expected payment subject to individual rationality and incentive compatibility constraints:

$$\min_{\tau, \boldsymbol{x}(\cdot)} \mathbb{E}\big[\tau\big(\boldsymbol{p}, \boldsymbol{v}(\boldsymbol{x}(\boldsymbol{\eta}), \boldsymbol{\eta})\big)\big] \quad \text{subject to}$$

$$\mathbb{E}\big[u\big(\tau\big(\boldsymbol{p}, \boldsymbol{v}(\boldsymbol{x}(\boldsymbol{\eta}), \boldsymbol{\eta})\big) - \boldsymbol{p} \cdot \boldsymbol{x}(\boldsymbol{\eta})\big)\big] \geq u(0) \tag{IR}$$

$$\forall \hat{\boldsymbol{x}}(\cdot) : \mathbb{E}\big[u\big(\tau\big(\boldsymbol{p}, \boldsymbol{v}(\boldsymbol{x}(\boldsymbol{\eta}), \boldsymbol{\eta})\big) - \boldsymbol{p} \cdot \boldsymbol{x}(\boldsymbol{\eta})\big)\big] \geq \mathbb{E}\big[u\big(\tau\big(\boldsymbol{p}, \boldsymbol{v}(\hat{\boldsymbol{x}}(\boldsymbol{\eta}), \boldsymbol{\eta})\big) - \boldsymbol{p} \cdot \hat{\boldsymbol{x}}(\boldsymbol{\eta})\big)\big] \tag{IC}$$

Letting $\boldsymbol{x}^{FB}$ be as defined in Lemma 3, we can derive useful intermediate results about properties of optimal contracts (i.e., τ that solve the client's problem).

Lemma 4. *The following conditions are together sufficient for a contract* τ *to be optimal:*

(i) for all $\hat{\boldsymbol{x}}(\cdot)$*:*

$$\mathbb{E}\big[u\big(\tau\big(\boldsymbol{p}, \boldsymbol{v}(\boldsymbol{x}^{FB}(\boldsymbol{\eta}), \boldsymbol{\eta})\big) - \boldsymbol{p} \cdot \boldsymbol{x}^{FB}(\boldsymbol{\eta})\big)\big] \geq \mathbb{E}\big[u\big(\tau\big(\boldsymbol{p}, \boldsymbol{v}(\hat{\boldsymbol{x}}(\boldsymbol{\eta}), \boldsymbol{\eta})\big) - \boldsymbol{p} \cdot \hat{\boldsymbol{x}}(\boldsymbol{\eta})\big)\big]$$

(ii) $\tau\big(\boldsymbol{p}, \boldsymbol{v}(\boldsymbol{x}^{FB}(\boldsymbol{\eta}), \boldsymbol{\eta})\big) = \boldsymbol{p} \cdot \boldsymbol{x}^{FB}(\boldsymbol{\eta})$ *almost surely.*

Moreover, if some such contract exists, and if u *is strictly concave, then the conditions are also necessary.*

For the first part of the lemma, notice that condition (i) is equivalent to $(\tau, \boldsymbol{x}^{FB}(\cdot))$ satisfying (IC). And condition (ii) implies that under $(\tau, \boldsymbol{x}^{FB}(\cdot))$, the dealer is fully insured and (IR) is satisfied with equality. Thus, $(\tau, \boldsymbol{x}^{FB}(\cdot))$ implements the efficient outcome—both the efficient trading policy and efficient risk sharing—and leaves the dealer with zero surplus. This then provides the client with her first-best payoff, and clearly, no contract can do better than that. The second part of the result observes that if some contract satisfies those conditions, then *all* optimal contracts must implement the efficient outcome and leave the dealer with zero surplus. Indeed, to implement the efficient trading policy, condition (i) must hold. And, if the dealer is risk averse, then to implement efficient risk sharing and leave the dealer with zero surplus, condition (ii) must hold.

Building on Lemma 4, we now state our main results, both of which pertain to the VWAP contract

$$\tau^{VWAP} \equiv \frac{\sum_{t=1}^{T} p_t v_t}{\sum_{s=1}^{T} v_s}.$$

Proposition 5 states that the VWAP contract is optimal in the specialized model. Proposition 6 says that if the dealer is risk averse and a certain full support condition is satisfied, then the VWAP contract is also the *unique* contract that is optimal. Because *any* function of market prices and volumes constitutes a feasible contract, this uniqueness result is a strong one.

Proposition 5. *The contract* τ^{VWAP} *is optimal.*

Proposition 6. *If* u *is strictly concave and the distributions of* $\boldsymbol{\varepsilon}$ *and* $\boldsymbol{v}(\boldsymbol{x}^{FB}(\boldsymbol{\eta}), \boldsymbol{\eta})$ *have full support over* $\mathbb{R}^T$ *and* $\mathbb{R}^T_{++}$*, respectively, then a contract* τ *is optimal only if* $\tau = \tau^{VWAP}$ *almost everywhere on its domain.*

To prove Proposition 5, we establish that τ^{VWAP} satisfies the conditions of Lemma 4. The lemma then immediately implies that τ^{VWAP} is optimal and in fact establishes an even stronger property: that τ^{VWAP} provides the client with her first-best payoff. To see that condition (ii) of Lemma 4 is satisfied, recall that, by the last part of Lemma 3, the first-best trading policy corresponds to a volume participation strategy. As a result, the realized costs incurred by the policy, $\boldsymbol{p} \cdot \boldsymbol{x}^{FB}(\boldsymbol{\eta})$, always equal the realized market VWAP, which also equals the payment $\tau^{VWAP}(\boldsymbol{p}, \boldsymbol{v}(\boldsymbol{x}^{FB}(\boldsymbol{\eta}), \boldsymbol{\eta}))$. The meat of the argument lies in establishing that condition (i) of Lemma 4 is satisfied.

To see why condition (i) of the lemma is satisfied, suppose that a dealer who is compensated according to τ^{VWAP} considers deviating from $\boldsymbol{x}^{FB}(\cdot)$ to trade $\delta > 0$ fewer shares at time t and δ more shares at time t'. By the definition of $\boldsymbol{x}^{FB}(\cdot)$, this will raise the dealer's expected costs for acquiring the position from the market, $\mathbb{E}[\boldsymbol{p} \cdot \boldsymbol{x}(\boldsymbol{\eta})]$. Indeed, this deviation will lower the price at t and raise the price at t' so that more shares are being acquired at a higher price and fewer shares at a lower price. But for a similar reason, this will also raise the dealer's expected revenue received as payment from the client, $\mathbb{E}\left[\frac{1}{\sum_{s=1}^{T} v(x_s(\boldsymbol{\eta}), \eta_s)} \boldsymbol{p} \cdot \boldsymbol{v}(\boldsymbol{x}(\boldsymbol{\eta}), \boldsymbol{\eta})\right]$. The key observation is that costs increase by more than revenue. Indeed, what matters for the magnitude of the change in costs is the fraction of the dealer's volume that is shifted from t to t', which in this case is δ. But what matters for the change in revenue is the fraction of (total market) volume that is shifted. This is a smaller fraction than δ—the economic intuition is that the effect is muted by the volume accounted for by outside traders.

In fact, as the previous argument suggests, something stronger than Condition (i) of Lemma 4 holds. That condition requires a contract to provide weak incentives for the dealer to pursue the first-best policy. But τ^{VWAP} in fact provides *strict* incentives for doing so. Mathematically, we show in the proof of Proposition 5 that for all trading policies $\hat{\boldsymbol{x}}(\cdot)$ not equal to $\boldsymbol{x}^{FB}(\cdot)$ almost surely, (IC) holds with strict inequality:

$$\mathbb{E}\big[u\big(\tau^{VWAP}\big(\boldsymbol{p}, \boldsymbol{v}(\boldsymbol{x}^{FB}(\boldsymbol{\eta}), \boldsymbol{\eta})\big) - \boldsymbol{p} \cdot \boldsymbol{x}^{FB}(\boldsymbol{\eta})\big)\big] > \mathbb{E}\big[u\big(\tau^{VWAP}\left(\boldsymbol{p}, \boldsymbol{v}(\hat{\boldsymbol{x}}(\boldsymbol{\eta}), \boldsymbol{\eta})\right) - \boldsymbol{p} \cdot \hat{\boldsymbol{x}}(\boldsymbol{\eta})\big)\big].$$

Although risk aversion may intensify this effect, the inequality is strict even when the dealer is risk neutral.

For proving the uniqueness result of Proposition 6, we build on the conclusion that τ^{VWAP} satisfies the conditions of Lemma 4 to deduce that all optimal contracts τ must satisfy those

same conditions. These conditions are demanding and severely restrict the possibilities for τ. With the full support assumptions, they in fact pin down τ to equal τ^{VWAP} almost everywhere.[26]

Some intuition for the uniqueness result in Proposition 6 can be gleaned by studying why contracts tied to other common benchmark prices fail to be optimal in the specialized model:

- First, consider the MOC contract $\tau^{MOC} = p_T$. That this contract is not optimal was already shown in Proposition 1, using an argument that applies even outside the parameterization of the model that we have since specialized to. Within this parameterization, yet another intuition can be given:

  The contract τ^{MOC} incentivizes the dealer to deviate from the first-best trading policy by tilting his trades toward the last period, a behavior known in practice as 'banging the close.' Price impact from this will move the price in period T, and a risk-neutral dealer would trade so as to maximize the expected gap between this price and the average price that he pays to acquire the position on the market.[27] At the other extreme, an infinitely risk-averse dealer would concentrate his trading entirely in the last period so as to insure himself against price shocks. In these boundary cases as well as the intermediate ones, the ensuing failure to induce the first-best trading policy imply that τ^{MOC} does not achieve optimality.

  Consider once again the episode mentioned in footnote 1 in the introduction, involving a client that contracted to conduct a large trade with HSBC at the fix. Within our model, this arrangement might be thought of as the MOC contract τ^{MOC}. Based on reports, HSBC did trade in just the way that the above analysis predicts (Bloomberg, 2016).

- Next, consider a contract benchmarked to the TWAP: $\tau^{TWAP} = \frac{1}{T}\sum_{t=1}^{T} p_t$. This contract also incentivizes the dealer to deviate from the first-best trading policy, and so it similarly fails to be optimal. In this case, the dealer would deviate by smoothing trading across time periods more than the first-best policy prescribes—in a sense, being insufficiently responsive to his information about market conditions.

- Finally, consider fixed price contracts. By paying the dealer a predetermined amount,

[26] The full-support assumption could be abandoned if the uniqueness statement were weakened to $\tau(\boldsymbol{p}, \boldsymbol{v}(\boldsymbol{x}^{FB}(\boldsymbol{\eta}), \boldsymbol{\eta})) = \tau^{VWAP}(\boldsymbol{p}, \boldsymbol{v}(\boldsymbol{x}^{FB}(\boldsymbol{\eta}), \boldsymbol{\eta}))$ almost surely, where $\boldsymbol{p}$ are the prices induced by $\boldsymbol{x}^{FB}(\boldsymbol{\eta})$.

[27] This type of suboptimal execution also appears in Saakvitne (2016) who develops a model of dealers who are incentivized on the basis of such contracts and therefore trade in this way.

regardless of the prices or volumes that are realized, these contracts require the dealer to bear some price risk. And if the dealer is risk averse, then this constitutes inefficient risk sharing, which destroys optimality.

The last observation also highlights why the uniqueness result requires risk aversion: an appropriately-specified fixed price (i.e., 'sell-the-firm') contract would also attain optimality under risk neutrality. Similarly, uniqueness also requires the full support assumptions, for otherwise there would be certain 'irrelevant' regions of the domain of τ in which the contract could be altered without affecting optimality.

We would stress that, beyond its optimality in the specialized model, an additional virtue of the VWAP contract is its simplicity. It is easy to calculate, and in many asset classes, the daily VWAP is even available as a pre-computed market statistic. But perhaps even more striking, the contract is detail-free: it is optimal across a wide set of assumptions about the distributions of $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon}$, as well as about the functional forms of u, h, and v. Thus, the contract satisfies a certain robustness property in the sense that it does not require the client to possess detailed knowledge of those quantities.

Finally, note that to the extent our model is only an approximation of reality, the aforementioned results on the VWAP contract might be expected to hold only in approximation. In particular, it might be necessary to add a small commission in order to ensure that the (IR) constraint remains satisfied. Consistent with this, such 'VWAP plus commission' contracts are commonly observed in practice.

6 Applications to benchmark design

To this point, we have focused on bilateral contracting, investigating which benchmark price a client should reference when contracting with her dealer. We concluded that it is generally suboptimal to reference the closing price, but that it might be desirable to reference the VWAP. Yet benchmarks also play a role in many other settings, suggesting further applications for our results. This section highlights three. Although these applications are somewhat more speculative than what we have analyzed so far, many of the same economic arguments can be made in favor of using VWAP benchmarks in these settings as well.

6.1 Benchmark computation in opaque markets

So far, we have proceeded under the assumption that any measurable function of prices and volumes is a feasible contract. This large set of feasible contracts may be appropriate for modeling asset classes with transparent and publicly available trading data (e.g., equities). But for other asset classes, data is more opaque and difficult to access so that it is not possible to contract on prices and volumes in arbitrary ways. Nevertheless, it is often possible to contract on a benchmark that a third party with access to data—perhaps a platform or regulator—computes and makes available. To the extent that clients and dealers are limited to contracting on the benchmark, the feasible contracts are then effectively the functions of prices and volumes that are measurable with respect to the benchmark. Hence, the set of feasible contracts depends on the benchmark formula, which raises questions about optimal benchmark design from the perspective of client-dealer relationships.

As a specific example of the latter type of market, consider foreign exchange, where a prominent benchmark is the WM/Reuters London 4 pm fix ('the fix'). Prior to 2015, the fix was based on the prices prevailing in a one-minute window. In that sense, it could be thought of, more or less, as a closing price. Hence, clients who were limited to contracting on the fix (e.g., because they lacked more detailed data) were essentially limited to MOC contracts—for example, the client that contracted with HSBC in the example that we mentioned in footnote 1 and have been discussing throughout the paper. As we have shown, MOC contracts can lead to poor outcomes: (*i*) our analysis of the general model showed that they nearly always fail to be optimal, and (*ii*) our analysis of the specialized model showed that they may induce the dealer to pursue a certain form of manipulative trading (i.e., 'banging the close'). But in recent years, industry participants have begun to rethink how the fix ought to be computed. For example, in 2015, WM/Reuters widened the relevant window for collecting prices from one minute to five. What is more, a primary motive for this change seems to have been concern about dealers manipulating the fix to the detriment of their clients, exactly the type of conflict that our framework is geared to analyze.

Analysis. If they are to facilitate the writing of desirable contracts, then how should benchmarks like the fix be computed? A simple way of altering our model so as to address this question is as follows. Add a benchmark administrator whose role is to compute and publish a benchmark $b(\boldsymbol{p}, \boldsymbol{v})$. Restrict the set of feasible contracts to functions of the benchmark

$\tau(b)$. Then ask: how should the benchmark administrator design the benchmark function b?

If the benchmark administrator's objective were to maximize the welfare of the client, then an optimal benchmark is the price that would be referenced in the optimal contract under full data availability for prices and volumes. Indeed, with such a benchmark b, a client could replicate the optimal contract by choosing $\tau(b) = b$, achieving the optimum despite her data limitations.

Moreover, if the setting is as in the specialized model, then our previous results imply that an optimal benchmark is the VWAP

$$b^{VWAP} = \frac{\sum_{t=1}^{T} p_t v_t}{\sum_{s=1}^{T} v_s}.$$

Likewise, if the benchmark administrator's objective were, alternatively, to maximize the sum of dealer and client welfare, then similar arguments would nevertheless continue to imply optimality of the VWAP benchmark.[28]

Discussion. To the extent that the foreign exchange market resembles the specialized model, our analysis suggests that the definition of the fix should be amended to more closely resemble a VWAP: the relevant window for collecting prices should be widened even further, and volumes should be included in the formula.[29] Consistent with this, some members of the Financial Stability Board and Foreign Exchange Joint Standing Committee have advanced some of these same suggestions (FSB, 2014; FXJSC, 2008).

One caveat is that ours is a partial equilibrium model in that we assume market conditions (representing, e.g., the activity of other traders) are exogenous. While this seems reasonably appropriate for our baseline application of bilateral contracting between one client and one dealer—other traders would not seem likely to be influenced by, or even aware of, the agreed-upon contract—it may be less appropriate for studying market-wide changes, such

[28] Another possibility for the objective—particularly plausible in the case where the benchmark administrator is a venue—is to maximize volume. A VWAP benchmark ought also be attractive under this objective. Though not captured directly in our model, client-based volume is presumably maximized when the clients' transaction costs are minimized, which, as argued, occurs when the benchmark is computed as the VWAP.

[29] Of course, benchmarks like the fix may have additional purposes beyond facilitating the writing of desirable contracts (e.g., summarizing the most recent price-relevant information about some economic fundamental), which could be hindered by widening the window for collecting prices. However, the tradeoff between potentially conflicting objectives is beyond the scope of our paper.

as benchmark design. In the language of our model, the choice of benchmark may feed back to affect the distribution over $\boldsymbol{\eta}$ in ways that we do not capture. Thus, the above analysis is most applicable to markets in which a large fraction of trading volume is driven by traders whose incentives are not tied directly to the benchmark, so that the aforementioned feedback effects are more likely to be relatively small.

Another caveat is that, in practice, parties may have financial interests tied to the realization of the benchmark beyond the trades that they will directly conduct at that benchmark price. In the setting of foreign exchange, for example, banks may possess financial obligations that are denominated in a particular currency. Interests like these may create incentives to manipulate the benchmark beyond what is captured by our model.[30] Duffie and Dworczak (2018) also tackle the question of benchmark design; one difference between their approach and ours is that they do allow for these incentives. Another difference lies in the criteria by which benchmarks are judged. Their aim is to design a benchmark that is resistant to manipulation and thus close to and informative about an underlying value. In contrast, our aim is to design a benchmark that facilitates the writing of contracts that lead to desirable outcomes. Despite these differences, Duffie and Dworczak (2018) find that in some cases—namely, when agents are able to split their trades undetected—a benchmark that resembles VWAP emerges as optimal.

6.2 Settlement prices of futures contracts

Another application concerns the settlement prices of certain futures contracts and how best to compute them. TAS (Trading At Settlement) is an order matching procedure available for some commodity futures, which allows market participants to trade futures contracts at prices relative to a particular benchmark: the yet-to-be-determined daily settlement price.

In such cases, our analysis of the specialized model suggests that the daily settlement price is susceptible to a certain type of manipulation when it is computed in a manner other than VWAP. For example, if the settlement price were the closing price, then a trader might achieve a positive expected profit through the following scheme: conduct a TAS trade, then

[30] The LIBOR scandal might be interpreted as a stark illustration of the power of these incentives: it came to light that banks had been manipulating their reports so as to enhance their LIBOR-denominated financial positions. (Note that LIBOR differs from our setting in that it was computed from 'cheap talk' reports rather than trades.)

pursue offsetting trades over the course of the day, concluding with a very large trade at the close so as to create a gap between the price of the TAS trade (i.e., the closing price) and the average price of the offsetting trades.[31] This resembles the setting of our model, with the aforementioned trader in the role of the dealer and that trader's TAS counterparty in the role of the client. This manipulation strategy mirrors the trading behavior that the dealer is induced to pursue by the MOC contract in the specialized model (i.e., 'banging the close'). Viewed through the lens of our model, our results suggest that this type of manipulation could be mitigated if the settlement price were computed as the VWAP.

In fact, the behavior described in the previous paragraph is consistent with trading observed in the market for crude oil futures. There, the settlement price is computed as the VWAP between 2:28 and 2:30 pm, the last two minutes before the close, so that it is approximately the closing price. And in fact, the CFTC has sued some traders for exactly the kind of manipulation described above, bringing cases against Optiver (CFTC, 2008) and SHK Management (CFTC, 2013). Similar considerations apply to TAM (Trading At Marker) orders, which are also executed at prices relative to a yet-to-be-determined benchmark price (CME Group, 2018).

Approximately the same economics applies to BTIC (Basis Trade at Index Close), TACO (Basis Trade At Cash Open) futures trades (CME Group, 2018), as well as on-close orders offered by equities exchanges. And indeed, the SEC has also sued traders for similar types of manipulation in equities (SEC, 2014). But for these applications, one caveat is that equity open and close prices are typically computed using a different trading mechanism—an auction—than the limit order book mechanism that is used throughout the day. This may make the open and close prices special. For example, the market is typically deeper during auctions. To the extent that this difference is encapsulated with relatively high values for η_1 and η_T, then our analysis applies. But if the price impact function h and/or the volume function v themselves differ in periods 1 and T relative to other periods, then our analysis may not carry over directly, and the application becomes more speculative. A similar caveat pertains to the application that we discuss next.

[31] This is distinct from the type of manipulation analyzed by Kumar and Seppi (1992), who demonstrate how, by trading in both the futures and spot markets, an uninformed manipulator can extract positive expected profits.

6.3 Valuation of mutual funds

Another application is to the calculation of net asset value (NAV) for open-end funds. In the United States, standard practice is for the NAV to be calculated once per day based on the closing price of the underlying securities. Investors can purchase and redeem shares of the fund at that price, provided they submitted orders to do so prior to the market close.

A potential concern is that this may provide traders with an incentive to manipulate prices. Indeed, consider a trader who begins with a long position in the fund. That trader might be able to convert his position in the fund for a corresponding position in the underlying while at the same time obtaining a positive expected profit through the following scheme: place an order to liquidate his position in the fund, then trade throughout the day to acquire an offsetting position in the underlying securities on the market, concluding with very large trades at the close. This resembles the setting of our model, with the aforementioned trader in the role of the dealer and the fund (i.e., the remaining shareholders) in the role of the client.[32] This manipulation strategy mirrors the trading behavior that the dealer is induced to pursue by the MOC contract in the specialized model (i.e., 'banging the close'). Viewed through the lens of our model, our results suggest that this type of manipulation could be mitigated by computing the fund's NAV using a different benchmark price for the underlying securities: the daily VWAP instead of the closing price.[33]

6.4 Related issues

Related issues appear in many other settings where existing benchmarks are computed from a small number of prices. For reasons similar to the ones we have been discussing, such benchmarks are prone to manipulation.[34] And a potential way to mitigate manipulation would be by (*i*) broadening the set of prices used to compute the benchmark, and (*ii*) weighting the various prices according to volume.

- For example, cash-settled European and American options have payoffs that depend on the price of an underlying security at the exercise time, which may make such options prone

[32] In the baseline application of bilateral contracting, the client is hurt by higher payments to the dealer. In this application, the remaining shareholders are hurt by dilution of the fund's value.

[33] Correspondingly, the deadline for submitting orders for purchases and redemptions would need to be moved to the market open. This could, of course, create delay costs.

[34] Zhang (2020) proposes a metric for measuring the risk of manipulation.

to manipulation. In contrast, Asian options have payoffs that depend on the average price of an underlying security over an interval of time, mitigating the risk of manipulation.

- Henderson, Pearson and Wang (2019) study structured equity products (SEPs), which are notes typically issued by a bank with payments that are based on the stock price of another company and/or a stock index. Some SEPs are priced based on the underlying's closing price, while others are priced based on the daily VWAP. Those authors find evidence consistent with issuers trading the underlying so to manipulate SEP prices. And as expected, relatively more manipulation seems to occur when the relevant underlying price is the close.
- The Brent Index is a benchmark computed from oil prices. Over time, production in the relevant oil fields—and thus trading volume—has declined, which, holding fixed the benchmark formula, would amplify its susceptibility to manipulation. In response, the benchmark has gradually been amended to include new streams of oil (Forties and Oseberg in 2002, Ekofisk in 2007, and Troll in 2018) and a wider window of days in the calculation.
- In the domain of interest rate swaps, parties often agree to conduct a block trade whose pricing entails components that are linked to market prices prevailing at a particular point in time. As a result, these arrangements are quite vulnerable to the possibility that one of the parties in the block trade might conduct on-market trades designed to manipulate the reference prices. For a recent case involving such manipulation, see CFTC v. Christophe Rivoire (CFTC, 2019).

In all these cases, issues arise that resemble the ones we have discussed in the various applications of our model. There is the same underlying problem: trade-based benchmark manipulation. And there is the same potential solution: change the benchmark formula to more closely resemble a VWAP. (Despite that, these various situations might not be completely isomorphic to our model in terms of the mathematics.)

7 Extensions of the specialized model

We now probe the robustness of our results on VWAP optimality by considering some variations on the specialized model. First, in Section 7.1, we enrich the set of contracts available to the client so as to include agency trading. The VWAP contract remains optimal, but no

longer uniquely so: the agency contract is also optimal within the model when it is available.

We then turn to the key restrictions that we imposed when specializing the model: static trading policies, temporary price impact, and forecastable volumes. Section 7.2 addresses the first of these by allowing also for dynamic trading policies. Under additional assumptions on the volume function $v(x, \eta)$ and the dependence structure of price shocks $\boldsymbol{\varepsilon}$, we show that the VWAP contract again remains optimal. Section 7.3 addresses the other two assumptions (*viz.* temporary price impact and forecastable volumes), beginning with a continuity result: if those assumptions hold in approximation, then VWAP-based contracts remain approximately optimal. We then illustrate and discuss the economic forces underlying why these assumptions are required for our results about VWAP optimality.

7.1 Agency trading

The baseline model studied principal trading, wherein the dealer is the counterparty of the agent. And in our model, the feasible contracts were functions of the prices $\boldsymbol{p}$ and volumes $\boldsymbol{v}$. This would be the appropriate set of contracts to consider if that were all the client could verifiably observe and contract on at time T.

But another class of trading arrangements sometimes used in practice is *agency trading*, wherein the dealer acts as a matchmaker in locating a counterparty.[35] Such an arrangement could be thought of within our model as reimbursement in the amount $\tau^A = \boldsymbol{p} \cdot \boldsymbol{x}$. This contract is not included in the set of feasible contracts considered in the baseline model, for the reason that $\boldsymbol{x}$ cannot typically be verified by the client in practice. However, in many asset classes, there exist regulatory bodies who can verify $\boldsymbol{x}$ and do enforce contracts of the form τ^A (though typically they do not also enforce contracts based on arbitrary functions involving $\boldsymbol{x}$).[36]

In such asset classes, it may therefore be natural to extend the class of feasible contracts to include τ^A. Assume all else remains as in the specialized model. Because the VWAP contract τ^{VWAP} satisfies the conditions of Lemma 4, it remains optimal regardless of which other contracts are feasible. In other words, Proposition 5 extends to the version of the

[35]For principal trading arrangements, which are the main focus of our paper, a broker-dealer acts as a dealer, which is why we have referred to him as a 'dealer' throughout. But for agency trading arrangements, a broker-dealer acts as a broker. In spite of that, we continue to refer to him as a 'dealer' throughout this section to maintain consistent terminology.

[36]For example, in the case of equities, agency trades are heavily regulated by the SEC.

specialized model in which the agency contract is available. However, Proposition 6, which states conditions under which τ^{VWAP} is uniquely optimal, does not extend in this way. Indeed, the agency contract also solves the client's problem described in Section 3.2, with $\boldsymbol{x}^{FB}(\cdot)$ as the corresponding recommended trading policy. Consequently, we have the following result as an immediate corollary of Lemma 4.

Corollary 7. *If the agency contract τ^A is feasible, then it is optimal.*

Thus, two contracts perform particularly well under the conditions of the specialized model: τ^{VWAP} and τ^A. However, the model highlights one potentially important difference between these two optima. On one hand, τ^A makes the dealer indifferent among all trading policies, and its optimality relies on this indifference being broken in favor of the first-best policy. Thus, τ^A fails to be robust to many perturbations of the model.[37] On the other hand, and as observed in Section 5.2, τ^{VWAP} provides the dealer with a strict incentive to pursue the first-best policy, rendering it robust to how the dealer breaks his indifference.

In addition to the distinction above, these two contracts also differ in terms of their robustness to several unmodeled elements. For example, consider the following ingredients as additions to the specialized model:

- If there is a stochastic component to volumes, then τ^{VWAP} would require the dealer to bear risk. And if the dealer is risk averse, then τ^{VWAP} would no longer be optimal; in fact, it would not even be individually rational. (But as previously observed, individual rationality can be restored through the addition of a suitable risk premium.) On the other hand, τ^A completely insulates the dealer from risk, even if volumes are stochastic. And mainly for that reason, τ^A would continue to be optimal.
- Suppose there are many dealers (with the same risk aversion) who are heterogeneous in terms of skill, which we model as knowledge of market conditions. For concreteness, assume two types of dealers: (*i*) high-skill dealers, who have perfect knowledge of $\boldsymbol{\eta}$, as in the baseline model, and (*ii*) low-skill dealers, who have imperfect knowledge. Then τ^{VWAP} has the added advantage that it serves as a screening device: high-skill dealers would

[37] For instance, suppose the model is perturbed by adding vanishingly small effort costs for the dealer, where different trading policies may require different amounts of effort. It is natural to assume that $\boldsymbol{x}^{FB}(\cdot)$ would not minimize these effort costs. In this perturbation, $\boldsymbol{x}^{FB}(\cdot)$ would then no longer be incentive compatible under τ^A. See Edelen and Kadlec (2012) for a model of contracting, focusing purely on agency execution, in which such effort costs play a significant role.

accept it, while low-skill dealers would not. On the other hand, τ^A would be equally acceptable to all types of dealers, which could prove expensive to the client if accepted by a low-skill dealer.[38]

7.2 Dynamic trading policies

The specialized model restricts to static trading policies. But for this section (Section 7.2) only, we permit the broader class of dynamic trading policies. Recall that a dynamic trading policy is a measurable vector $(x_t(\cdot))_{t=1}^T$, such that (i) for all t, x_t depends on $\boldsymbol{\eta}$ and $(p_s)_{s=1}^{t-1}$, (ii) for all t, $x_t\big(\boldsymbol{\eta}, (p_s)_{s=1}^{t-1}\big) \geq 0$ almost surely, and (iii) $\sum_{t=1}^T x_t\big(\boldsymbol{\eta}, (p_s)_{s=1}^{t-1}\big) = 1$ almost surely.

7.2.1 Conditions under which VWAP optimality extends

For our results to be robust to widening the class of trading policies in this way, we must narrow the class of volume functions that we consider. For this section (Section 7.2.1) only, we assume that $v(x, \eta) = x + \eta$. With this volume function, η_t gains another interpretation, as the unsigned volume traded on the market by outside traders in period t. In Section 7.2.2, we discuss why, under other choices for v, the VWAP contract can fail to be optimal if the dealer may adjust his strategy in response to previously observed prices.

All results stated in Section 5 extend to this version of the model and the proofs remain unchanged under the mild assumption that $\boldsymbol{\varepsilon}$ satisfies

$$\mathbb{E}[\varepsilon_{t+1} - \varepsilon_t | \boldsymbol{\eta}, \varepsilon_1, \ldots, \varepsilon_{t-1}] = 0 \tag{1}$$

for all $t = 1, \ldots, T-1$. Examples of such processes include sets of independent random variables with constant mean as well as random walks of the form $\varepsilon_{t+1} = \varepsilon_t + \nu_{t+1}$ for $\nu_1, \ldots, \nu_T$ independent zero-mean random variables. Note that assumption (1) implies the assumption of the specialized model: that all ε_t have the same expectation conditional on $\boldsymbol{\eta}$, which we continue to denote by μ. Note also that, given $\boldsymbol{\eta}$, dependence on the previous prices $(p_s)_{s=1}^{t-1}$ is equivalent to dependence on $(\varepsilon_s)_{s=1}^{t-1}$. Therefore, we can equally well condition in (1) on $\boldsymbol{\eta}, p_1, \ldots, p_{t-1}$, which represents the information available to the dealer when making

[38] A similar argument may help to explain why the volume-weighted average price is a commonly-used benchmark in transactions cost analysis (e.g., Berkowitz, Logue and Noser, 1988).

the time-t trading decision. Thus, (1) means that the price shocks at times t and $t+1$ do not differ in their predicted means given the knowledge acquired by the dealer before time t.

Remark. It is also possible to analyze versions of the model with a continuous time trading period $[0, T]$ rather than the discrete trading periods $\{1, 2, \ldots, T\}$. Our results continue to hold for continuous-time versions of both the specialized model and its extensions, provided that technicalities are suitably dealt with. One subtlety is that for the continuous-time version of the extension considered in this section, assumption (1) becomes the martingale property for $\boldsymbol{\varepsilon}$ (with respect to the σ-algebras generated by itself and the entire process $\boldsymbol{\eta}$). By contrast, assumption (1) in discrete time is slightly more general than the martingale property because the conditioning is only over $\boldsymbol{\eta}, \varepsilon_1, \ldots, \varepsilon_{t-1}$ and not $\boldsymbol{\eta}, \varepsilon_1, \ldots, \varepsilon_{t-1}, \varepsilon_t$. We refrain from spelling out the details for the continuous-time model because it would not offer additional insight compared to our discrete-time model.

7.2.2 Why VWAP optimality may fail without those conditions

We have just said that, if the dealer can use dynamic trading policies, then we can still derive all results of Section 5, provided additional restrictions are imposed. In particular, we have shown that $v(x, \eta) = x + \eta$ is a sufficient restriction on the volume function. Here, we explain why τ^{VWAP} may fail to remain optimal without any additional restrictions on $v(x, \eta)$.[39]

Assume that after trading according to $\boldsymbol{x}^{FB}(\cdot)$ in the first trading period, the dealer observes a p_1 that implies the realization of ε_1 was much greater than the expected value μ. If the dealer will be compensated according to τ^{VWAP}, he may then want to deviate from the first-best trading policy if it is possible that by doing so he can distort the daily market volume $\sum_{t=1}^{T} v(x_t, \eta_t)$ downward. To see this, note that this distortion would increase the weight placed on p_1 in the payment specified by τ^{VWAP} above the weight placed on p_1 in the dealer's trading costs $\boldsymbol{p} \cdot \boldsymbol{x}$. If ε_1 is sufficiently high, this deviation would be profitable.

Such a deviation is not possible if the dealer is limited to static trading policies as in the specialized model. In such cases, he must commit to a complete trading policy prior to observing any information about the realized prices, and so he cannot condition any trading

[39] Our focus in this section is on why Section 7.2.1 imposes the additional assumption $v(x, \eta) = x + \eta$. We do not discuss why Section 7.2.1 also assumes that $\boldsymbol{\varepsilon}$ satisfies (1) for the reason that this additional assumption is mild. Nevertheless, it is fairly easy to see that this assumption is important for the previous results to go through. For example, if the dealer can use dynamic trading policies, then without this assumption, even the first-best such policy might not be as characterized by Lemma 3.

decision on p_1. Such a deviation is also not possible if $v(x,\eta) = x+\eta$, as in Section 7.2.1. Even though trading decisions can depend on p_1 in that version of the model, they cannot create the requisite distortion, since the daily market volume $\sum_{t=1}^{T} v(x_t,\eta_t) = \sum_{t=1}^{T} x_t + \sum_{t=1}^{T} \eta_t = 1 + \sum_{t=1}^{T} \eta_t$ does not depend on the trading policy. However, if neither of these conditions holds, then a profitable deviation of this form might exist, in which case τ^{VWAP} would no longer be optimal.[40]

7.3 Permanent price impact and stochastic volumes

In the specialized model, price impact is purely temporary, and the dealer's knowledge of market conditions allows him to perfectly forecast volumes. But in this section, we analyze the situation when there is also permanent price impact and volumes have additional noise. To make the presentation more accessible, we add these elements to a simple parametrization of the specialized model. The setting is as follows: (*i*) there are $T = 2$ trading periods; (*ii*) volume at time t is given by $x_t + \eta_t + \varepsilon_t^v$ for an independent, nonnegative random variable ε_t^v; and (*iii*) the price impact function allows for temporary and permanent price impact of the form

$$p_t = c\sum_{s\le t} x_s + \frac{x_t}{\eta_t} + \varepsilon_t.$$

In terms of the notation of the general model, the random components are $\boldsymbol{\eta} = (\eta_1,\eta_2)$ and $\boldsymbol{\varepsilon} = (\varepsilon_1,\varepsilon_2,\varepsilon_1^v,\varepsilon_2^v)$. Note that when $c > 0$, there is permanent price impact, since the first period volume x_1 affects the second period price p_2. In contrast, when $c = 0$ and $\varepsilon_1^v = \varepsilon_2^v = 0$, the setting is nested by our previous analysis, for the case in which $h(y) = y$ and $v(x,\eta) = x+\eta$.

As before, we begin by deriving the first-best trading policy. Using the notation $x_1 = x$ and $x_2 = 1-x$, we solve

$$\min_{x\in[0,1]} \mathbb{E}\big[p_1 x + p_2(1-x)\big|\eta_1,\eta_2\big],$$

[40] To elaborate, τ^{VWAP} would not be optimal because it would be dominated by a 'VWAP minus commission' contract $\tau^{VWAP} - c$ for an appropriately chosen commission c. Indeed, if $(\tau^{VWAP},\boldsymbol{x}(\cdot))$ satisfies (IC), then so does $(\tau^{VWAP}-c,\boldsymbol{x}(\cdot))$. Furthermore, the existence of the profitable deviation described in the text means that $(\tau^{VWAP},\boldsymbol{x}(\cdot))$ satisfies (IR) with slack. Thus, for some $c > 0$, $(\tau^{VWAP}-c,\boldsymbol{x}(\cdot))$ also satisfies (IR) while being cheaper for the client. Note that while this establishes that τ^{VWAP} is suboptimal, it does not solve for the optimal contract, which could be an even more substantial deviation from VWAP.

which yields the optimal number of shares in the first period

$$x^{FB} = \frac{\frac{c}{2}\eta_1\eta_2 + \eta_1}{c\eta_1\eta_2 + \eta_1 + \eta_2}. \tag{2}$$

Note that the stochastic components of volume $(\varepsilon_1^v, \varepsilon_2^v)$ have no effect on the first-best policy, as only prices are relevant for that problem. The permanent component of price impact, parametrized by c, does affect the first-best policy, by smoothing it toward a uniform rate of trading regardless of how market conditions may vary across time: $\lim_{c\to\infty} x^{FB} = \frac{1}{2}$.[41]

We next turn to the question of the second best. In general, it is intractably complex to derive a closed-form for the optimal contract even with this simple parametrization. However, a natural question is whether the VWAP contract still performs well. We make two sets of observations. First, we show that if the departures from the specialized model are small, then the VWAP contract remains approximately optimal. Second, we consider what happens if these departures are large, and in so doing, we shed light on some of the economic forces at play. For tractability, we consider both of these issues in the special case where the dealer is risk-neutral (i.e., we assume $u(w) = w$).

7.3.1 Continuity result

In such a model with permanent price impact and stochastic volumes, the VWAP contract will no longer be optimal. Still, for small permanent price impact and little stochasticity in volumes, such a contract is approximately optimal. More precisely, we give a stochastic continuity result: for any constant γ, some VWAP-plus-fee contract gives the client a payoff that is not worse by γ than her optimal payoff whenever the permanent price impact is small enough and volumes are near enough to deterministic in a suitable sense.

When there is permanent price impact, the VWAP contract τ^{VWAP} might no longer satisfy the individual rationality constraint. We therefore instead consider the possibility that the client offers a contract $\tau^{VWAP} + f$ where f is a constant, interpreted as a fee to make the contract acceptable for the dealer. For a given contract τ, denote by $\mathcal{X}(\tau)$ the set of trading policies $\boldsymbol{x}(\cdot)$ that, together with τ, satisfy both individual rationality and incentive

[41] As an aside, note that if $\eta_1 = \eta_2$, then as in the baseline specification of Bertsimas and Lo (1998), first best corresponds to trading at a uniform rate over time: $x^{FB} = \frac{1}{2}$.

compatibility, which become here

$$\mathbb{E}\big[\tau\big(\boldsymbol{p}, \boldsymbol{x}(\boldsymbol{\eta}) + \boldsymbol{\eta} + \boldsymbol{\varepsilon}^v\big) - \boldsymbol{p} \cdot \boldsymbol{x}(\boldsymbol{\eta})\big] \geq 0 \tag{IR$'$}$$

$$\forall \hat{\boldsymbol{x}}(\cdot) : \mathbb{E}\big[\tau\big(\boldsymbol{p}, \boldsymbol{x}(\boldsymbol{\eta}) + \boldsymbol{\eta} + \boldsymbol{\varepsilon}^v\big) - \boldsymbol{p} \cdot \boldsymbol{x}(\boldsymbol{\eta})\big] \geq \mathbb{E}\big[\tau\big(\boldsymbol{p}, \hat{\boldsymbol{x}}(\boldsymbol{\eta}) + \boldsymbol{\eta} + \boldsymbol{\varepsilon}^v\big) - \boldsymbol{p} \cdot \hat{\boldsymbol{x}}(\boldsymbol{\eta})\big]. \tag{IC$'$}$$

With this notation in place, we can then state the main result of this section. Under suitable assumptions, an analogous result would hold even in substantially more general environments, such as multi-period models with a risk-averse dealer.

Proposition 8. *Assume that* $\mathbb{E}\left[\frac{\eta_1\eta_2+1/\min\{\eta_1,\eta_2\}}{\eta_1+\eta_2}\right] < \infty$. *Then for all* $\gamma > 0$, *there exist* $f > 0$, $\bar{c} > 0$ *and* $\bar{\varepsilon}^v > 0$ *such that, for all* $c \in [0, \bar{c}]$ *and all* $\boldsymbol{\varepsilon}^v \in [0, \bar{\varepsilon}^v]^2$ *almost surely, we have* $\mathcal{X}(\tau^{VWAP} + f) \neq \emptyset$ *and*

$$\sup_{\boldsymbol{x} \in \mathcal{X}(\tau^{VWAP}+f)} \mathbb{E}\big[\tau^{VWAP}\big(\boldsymbol{p}, \boldsymbol{x}(\boldsymbol{\eta}) + \boldsymbol{\eta} + \boldsymbol{\varepsilon}^v\big) + f\big] \leq \inf_{\tau, \boldsymbol{x} \in \mathcal{X}(\tau)} \mathbb{E}\big[\tau\big(\boldsymbol{p}, \boldsymbol{x}(\boldsymbol{\eta}) + \boldsymbol{\eta} + \boldsymbol{\varepsilon}^v\big)\big] + \gamma. \tag{3}$$

The left-hand side of (3) is an upper bound on the client's payment under the VWAP-plus-fee contract $\tau^{VWAP} + f$ and any corresponding incentive compatible trading policy. The right-hand side (excluding γ) is a lower bound on her payment in the optimum. The result says that if c and $\boldsymbol{\varepsilon}^v$ are small, then these two bounds are close, so that $\tau^{VWAP} + f$ is approximately optimal. The argument could also be extended to show that $\tau^{VWAP} + f$ is strictly better for the client than τ^{MOC}, when c and $\boldsymbol{\varepsilon}^v$ are sufficiently small.

The proof involves three steps. First, for small values of c, a small fee f suffices to restore individual rationality, so that $\mathcal{X}(\tau^{VWAP} + f) \neq \emptyset$. Second, for small c and $\boldsymbol{\varepsilon}^v$, the right-hand side (excluding γ) is close to the client's payment under the second-best solution to the specialized model. Third, for small c and $\boldsymbol{\varepsilon}^v$, the elements of $\mathcal{X}(\tau^{VWAP} + f)$ will not differ too much from the first-best trading policy (intuitively, this follows from the fact that, as observed in Section 5.2, such a contract provides the dealer with a strict incentive to pursue the first-best policy in the specialized model), which implies that the left-hand side is also close to the client's payment under the second-best solution to the specialized model.

Two of the key assumptions that we made when specializing the model are (*i*) that price impact is purely temporary, and (*ii*) that the dealer is able to perfectly forecast volumes. As we remarked in Section 4.1: (*i*) in certain markets, the permanent component of price impact is indeed small, and (*ii*) sophisticated traders can indeed forecast volumes with a great deal

of accuracy. Given that the specialized model therefore is a good approximation of certain markets, intuition might have suggested that our main results would be good approximations of what is true in those markets. Proposition 8 establishes this formally.

7.3.2 Non-negligible permanent price impact

The previous analysis indicates that VWAP-based contracts can remain approximately optimal if the permanent component of price impact is small. However, as we show now, the departure from optimality may be more severe if this permanent component is non-negligible. An intuition is that in such settings, the order of events matters: trades in earlier periods influence prices in later periods. In consequence, an optimal contract must account for this, so that early periods and later periods would be handled differently in determining the dealer's compensation. However, τ^{VWAP} does not possess this property, due to the commutative property of the weighted average. Therefore, when offered a VWAP-based contract, it is profitable for the dealer to deviate from the first-best trading policy.

To illustrate with a specific counterexample, we consider the setting described above but where $\varepsilon_1^v = \varepsilon_2^v = 0$ so that volumes are forecastable, as in the specialized model. Recalling equation (2), the number of shares traded by the dealer in the first period under the first-best policy is

$$x^{FB} = \frac{\frac{c}{2}\eta_1\eta_2 + \eta_1}{c\eta_1\eta_2 + \eta_1 + \eta_2}.$$

We next derive the trading policy that is induced by the VWAP contract. Recalling that we have here assumed the dealer to be risk-neutral, he optimizes

$$\max_{x\in[0,1]} \mathbb{E}\big[\tau^{\text{VWAP}} - p_1 x - p_2(1-x)\big|\eta_1, \eta_2\big].$$

Consequently, the number of shares traded by the dealer in the first period is

$$x^{VWAP} = \frac{\frac{c}{2}\eta_1\eta_2(2\eta_1 + \eta_2)/(\eta_1 + \eta_2) + \eta_1}{c\eta_1\eta_2 + \eta_1 + \eta_2}.$$

Our analysis of the specialized model demonstrates that $c = 0$ is a sufficient condition for τ^{VWAP} to be exactly optimal. We now see that it is also necessary. Indeed, if $c > 0$, then $x^{VWAP} \neq x^{FB}$, meaning that τ^{VWAP} induces the dealer to distort his trading decision away

from the first-best policy.[42,43] In consequence, the client obtains something less than her first-best payoff. However—at least in this case of a risk-neutral dealer—an appropriately-specified fixed price contract would achieve the first best and thus dominate the VWAP contract.[44]Although τ^{VWAP} is not itself optimal, we can show that, under the above assumptions, it is unambiguously better for the client than τ^{MOC}.

7.3.3 Non-negligible stochastic volumes

Previous analysis indicated that VWAP-based contracts can remain approximately optimal if the stochastic component of volume is small. However, as we show now, the departure from optimality may be more severe if this stochastic component is non-negligible. An intuition is that the first-best solution is not sensitive to how volume is determined: only price impact matters. But the dealer's incentives under a VWAP contract are shaped by volume, and thus, stochastic volumes distort his trading decisions away from first best.

To illustrate, we consider the setting described at the beginning of Section 7.3, but where $c = 0$ so that price impact is purely temporary (just as in the specialized model). Plugging $c = 0$ into equation (2), the dealer trades

$$x^{FB} = \frac{\eta_1}{\eta_1 + \eta_2}$$

[42]In fact, in this case of a risk-neutral dealer, we have $x^{VWAP} > x^{FB}$. But this need not generalize: under a sufficient amount of risk aversion, the departure from the first-best policy could be in either direction.

[43]From the expression, we can see that $\frac{\mathrm{d}}{\mathrm{d}c}(x^{VWAP}) > 0$, so that permanent price impact leads the dealer to trade more in the first period than $\frac{\eta_1}{\eta_1+\eta_2}$, which is what he selects in the baseline of no permanent price impact. For some intuition into this, note that the dealer's payoff from a choice of x is

$$\tau^{\text{VWAP}} - p_1 x - p_2(1-x) = \underbrace{\frac{x(\eta_1+\eta_2)-\eta_1}{\eta_1+\eta_2+1}}_{\alpha}(p_2 - p_1).$$

Note that $\alpha > 0 \iff x > \frac{\eta_1}{\eta_1+\eta_2}$, so that a choice of $x = \frac{\eta_1}{\eta_1+\eta_2}$ leads to $\alpha = 0$ and zero profits. With permanent price impact (i.e., if $c > 0$), then a choice of $x = \frac{\eta_1}{\eta_1+\eta_2}$ also induces $\mathbb{E}[p_2] > \mathbb{E}[p_1]$. In that case, a risk-neutral dealer could improve his expected payoff by slightly increasing x, so as to raise α. On the other hand, without permanent price impact (i.e., if $c = 0$), then a choice of $x = \frac{\eta_1}{\eta_1+\eta_2}$ makes $\mathbb{E}[p_1] = \mathbb{E}[p_2]$, so the dealer cannot improve his payoff in this way.

[44]In fact, we can show that under risk neutrality and the other above assumptions, where we have permanent price impact $c > 0$, this appropriately-specified fixed price contract is the unique optimal contract in the class of weighted-price contracts of the form $\tau(\boldsymbol{p}, \boldsymbol{v}) = \tau_0 + (\tau_1 v_1 + \tau_2)p_1 + (\tau_1 v_2 + \tau_2)p_2$ for constants (τ_0, τ_1, τ_2). And because fixed price contracts do not insure the dealer, a corollary of this observation is that under the alternative assumption of risk aversion, no contract in that class achieves the first best.

shares in the first period under the first-best policy, just as in the specialized model.

We next derive the trading policy that is induced by the VWAP contract. Recalling that we have assumed the dealer to be risk-neutral, he optimizes

$$\max_{x\in[0,1]} \mathbb{E}\big[\tau^{\text{VWAP}} - p_1 x - p_2(1-x)\big|\eta_1,\eta_2\big].$$

Assuming now that the volume shocks $(\varepsilon_1^v, \varepsilon_2^v)$ are independent of the price shocks $(\varepsilon_1, \varepsilon_2)$ (though not necessarily from $\boldsymbol{\eta}$), the dealer trades

$$x^{VWAP} = \frac{\mathbb{E}\big[\frac{\varepsilon_1^v/\eta_1+\varepsilon_2^v/\eta_2+2\varepsilon_1^v/\eta_2+2\eta_1/\eta_2+2}{\varepsilon_1^v+\varepsilon_2^v+1+\eta_1+\eta_2}\big|\eta_1,\eta_2\big]}{2\mathbb{E}\big[\frac{(1/\eta_1+1/\eta_2)(\varepsilon_1^v+\varepsilon_2^v+\eta_1+\eta_2)}{\varepsilon_1^v+\varepsilon_2^v+1+\eta_1+\eta_2}\big|\eta_1,\eta_2\big]}.$$

shares in the first period. Note that we in general have $x^{VWAP} \neq x^{FB}$, meaning that τ^{VWAP} induces the dealer to distort his trading decision away from the first-best trading policy. Then, as in Section 7.3.2, τ^{VWAP} cannot be optimal in such cases.

Stochastic volumes create several issues. An obvious one is that, under a VWAP-based contract, a dealer who pursues the first-best trading policy would be left bearing risk generated by the volume shocks. If the dealer were risk-averse, then he would need to be compensated for that. But a slightly more subtle issue is that period t volumes might no longer be homogeneous of degree one in x_t and η_t. For this reason, we have $x^{VWAP} \neq x^{FB}$ even when $\boldsymbol{\varepsilon}^v$ is deterministic and the dealer is risk neutral.[45]

8 Conclusion

Institutional investors often delegate the execution of their trades to dealers. But in many markets, it is difficult for such investors to monitor their dealers throughout the execution

[45] Nevertheless, in this case of deterministic $\boldsymbol{\varepsilon}^v$, the VWAP contract can be modified to restore optimality. Consider the contract

$$\tau^* = \frac{(v_1-\varepsilon_1^v)p_1 + (v_2-\varepsilon_2^v)p_2}{v_1+v_2-\varepsilon_1^v-\varepsilon_2^v},$$

which, like τ^{VWAP}, is a weighted average price, but with weights modified to account for $\boldsymbol{\varepsilon}^v$. When $\boldsymbol{\varepsilon}^v = \mathbf{0}$, τ^* reduces to τ^{VWAP}, and indeed, the optimality of τ^* for $\boldsymbol{\varepsilon}^v \neq \mathbf{0}$ follows by the same arguments. τ^* is, moreover, optimal not only under risk neutrality (as an appropriately-specified fixed price contract would also be) but also under risk aversion.

process. Even though dealers are typically subject to 'best interest' or 'best execution' obligations, these responsibilities are often vague and leave dealers with some leeway to act in ways that may harm their clients. One solution is to search for contractual arrangements that perform well in spite of the constraints imposed by this agency conflict. To make progress, our first contribution is to formulate a general model of this contracting problem, which, to the best of our knowledge, has been heretofore unstudied.

In practice, contracts used in these settings commonly take the MOC form, in which the client and dealer trade at the price prevailing in the market at the end of the execution window. Although such arrangements might seem innocuous on the surface, they often lead to inefficient execution for reasons stemming from the underlying agency conflict that we model and explore. Our first result demonstrates that MOC contracts are suboptimal across all realistic parameterizations of the general model.

To investigate which contracts are in fact optimal, we then proceed to specialize the model. Under a particular set of conditions, we obtain a strikingly simple solution: the VWAP contract is optimal, uniformly across a wide parameter space. This result therefore explains the usage of this contract in practice, at least in settings that are roughly consistent with these conditions. But, conversely, our analysis also provides a reason to question the usage of VWAP contracts in settings that are inconsistent with these conditions.

Another set of applications address questions of benchmark design, which is a particularly relevant issue in markets without public data availability. In such markets, participants typically cannot contract directly on prices and volumes, but they *can* often contract on a benchmark published by a platform or a regulator. In markets such as foreign exchange, prevailing benchmarks more closely resemble the closing price than the VWAP. In consequence, current principal trading arrangements often effectively take the form of MOC contracts. In the specialized model, such contracts may induce the dealer to distort his trading away from the efficient policy, instead trading an overly large quantity at the close. This prediction of the model is consistent with behavior often observed in such markets, including some of the episodes of manipulation mentioned throughout the text. To reduce the distortions stemming from such manipulation, our results recommend that the definitions of these benchmarks should be amended to more closely resemble a volume-weighted average price.

A final point concerns instances of our general model that fall outside the confines of our specialized model (wherein price impact is temporary and volumes are predictable). As we

have shown, VWAP-based contracts remain approximately optimal under small departures from the specialized model. But if the departure is large, then our results do not apply. Nevertheless, our general framework could still be useful in studying optimal contracting under alternative constellations of assumptions. It would be valuable to push this research program further with such analysis in future work.

A Lemmas

We present in this appendix auxiliary results in a general framework so that we can apply them in Appendix B to the proofs of both settings of Sections 5 and 7.2.1.

Lemma 9. *For random variables $\varepsilon_1, \dots, \varepsilon_T$, the following are equivalent:*

(a) $\mathbb{E}[\varepsilon_t - \varepsilon_T | \boldsymbol{\eta}, \varepsilon_1, \dots, \varepsilon_{t-1}] = 0$ *for all* $t = 1, \dots, T-1$,

(b) $\mathbb{E}[\varepsilon_{t+1} - \varepsilon_t | \boldsymbol{\eta}, \varepsilon_1, \dots, \varepsilon_{t-1}] = 0$ *for all* $t = 1, \dots, T-1$.

Proof of Lemma 9. To show that (a) implies (b), we compute

$$\begin{aligned}
\mathbb{E}[\varepsilon_{t+1} - \varepsilon_t | \boldsymbol{\eta}, \varepsilon_1, \dots, \varepsilon_{t-1}] &= \mathbb{E}[\varepsilon_{t+1} - \varepsilon_T + \varepsilon_T - \varepsilon_t | \boldsymbol{\eta}, \varepsilon_1, \dots, \varepsilon_{t-1}] \\
&= \mathbb{E}[\varepsilon_{t+1} - \varepsilon_T | \boldsymbol{\eta}, \varepsilon_1, \dots, \varepsilon_{t-1}] - \underbrace{\mathbb{E}[\varepsilon_t - \varepsilon_T | \boldsymbol{\eta}, \varepsilon_1, \dots, \varepsilon_{t-1}]}_{=0 \text{ by (a)}} \\
&= \mathbb{E}\big[\underbrace{\mathbb{E}[\varepsilon_{t+1} - \varepsilon_T | \boldsymbol{\eta}, \varepsilon_1, \dots, \varepsilon_{t-1}, \varepsilon_t]}_{=0 \text{ by (a)}} \big| \boldsymbol{\eta}, \varepsilon_1, \dots, \varepsilon_{t-1} \big] \\
&= 0.
\end{aligned}$$

Conversely, assume that (b) holds so that

$$\begin{aligned}
\mathbb{E}[\varepsilon_t - \varepsilon_T | \boldsymbol{\eta}, \varepsilon_1, \dots, \varepsilon_{t-1}] &= \mathbb{E}\big[\varepsilon_t - \underbrace{\mathbb{E}[\varepsilon_T | \boldsymbol{\eta}, \varepsilon_1, \dots, \varepsilon_{t-1}, \dots, \varepsilon_{T-2}]}_{= \mathbb{E}[\varepsilon_{T-1} | \boldsymbol{\eta}, \varepsilon_1, \dots, \varepsilon_{t-1}, \dots, \varepsilon_{T-2}] \text{ by (b)}} \big| \boldsymbol{\eta}, \varepsilon_1, \dots, \varepsilon_{t-1} \big] \\
&= \mathbb{E}[\varepsilon_t - \varepsilon_{T-1} | \boldsymbol{\eta}, \varepsilon_1, \dots, \varepsilon_{t-1}].
\end{aligned}$$

From this, we deduce

$$\mathbb{E}[\varepsilon_t - \varepsilon_T | \boldsymbol{\eta}, \varepsilon_1, \dots, \varepsilon_{t-1}] = \mathbb{E}[\varepsilon_t - \varepsilon_{T-1} | \boldsymbol{\eta}, \varepsilon_1, \dots, \varepsilon_{t-1}] = \cdots = \mathbb{E}[\varepsilon_t - \varepsilon_{t+1} | \boldsymbol{\eta}, \varepsilon_1, \dots, \varepsilon_{t-1}] = 0$$

by repeatedly applying (b). □

Lemma 10. *Consider random variables $\boldsymbol{\varepsilon}$ and $\boldsymbol{\eta}$. Assume that either*

1. *there exist functions $\boldsymbol{x}(\cdot)$ depending on $\boldsymbol{\eta}$ such that it holds $\sum_{t=1}^{T} x_t(\boldsymbol{\eta}) = 1$ almost surely, and $\mathbb{E}[\varepsilon_t | \boldsymbol{\eta}] = \mu$ almost surely for all t; or*

2. *there exist functions* $\boldsymbol{x}(\cdot)$ *with each* x_t *depending on* $\boldsymbol{\eta}$ *and* $(\varepsilon_s)_{s=1}^{t-1}$ *such that it holds* $\sum_{t=1}^{T} x_t\big(\boldsymbol{\eta}, (\varepsilon_s)_{s=1}^{t-1}\big) = 1$ *almost surely, and* $\boldsymbol{\varepsilon}$ *satisfies* (1) *for all* t.

Then

$$\sum_{t=1}^{T} \mathbb{E}[\varepsilon_t x_t | \boldsymbol{\eta}] = \mu \quad \textit{almost surely,}$$

suppressing in x_t *the arguments* $\boldsymbol{\eta}$ *and* $\big(\boldsymbol{\eta}, (\varepsilon_s)_{s=1}^{t-1}\big)$ *in cases 1 and 2, respectively.*

Proof of Lemma 10. We also suppress the argument in x_t in this proof, and all equalities are meant to hold almost surely. In the first case, we compute

$$\sum_{t=1}^{T} \mathbb{E}[\varepsilon_t x_t | \boldsymbol{\eta}] = \sum_{t=1}^{T} \mathbb{E}\big[x_t \mathbb{E}[\varepsilon_t | \boldsymbol{\eta}] \big| \boldsymbol{\eta}\big] = \mathbb{E}\left[\sum_{t=1}^{T} x_t \mu \middle| \boldsymbol{\eta}\right] = \mu,$$

using $\mathbb{E}[\varepsilon_t | \boldsymbol{\eta}] = \mu$. In the second case, we argue as follows:

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[\varepsilon_t x_t | \boldsymbol{\eta}] &= \mathbb{E}[\varepsilon_T | \boldsymbol{\eta}] + \sum_{t=1}^{T-1} \mathbb{E}[(\varepsilon_t - \varepsilon_T) x_t | \boldsymbol{\eta}] \\
&= \mu + \sum_{t=1}^{T-1} \mathbb{E}\big[\mathbb{E}[\varepsilon_t - \varepsilon_T | \boldsymbol{\eta}, \varepsilon_1, \ldots, \varepsilon_{t-1}] x_t \big| \boldsymbol{\eta}\big] \\
&= \mu,
\end{aligned}$$

where the final step is thanks to (1) and Lemma 9. □

Lemma 11. *Consider random variables* $\boldsymbol{\varepsilon}$ *and* $\boldsymbol{\eta}$*, with* $\boldsymbol{\eta}$ *taking positive values. Assume that either*

1. *there exist nonnegative functions* $\boldsymbol{x}(\cdot)$ *depending on* $\boldsymbol{\eta}$ *and a positive function* v *with domain* $\mathrm{dom}(v) \subseteq \mathbb{R}_+ \times \mathbb{R}_{++}$*, and that* $\mathbb{E}[\varepsilon_t | \boldsymbol{\eta}] = \mu$ *almost surely for all* t*; or*

2. *there exist nonnegative functions* $\boldsymbol{x}(\cdot)$ *with each* x_t *depending on* $\boldsymbol{\eta}$ *and* $(\varepsilon_s)_{s=1}^{t-1}$ *such that it holds* $\sum_{t=1}^{T} x_t\big(\boldsymbol{\eta}, (\varepsilon_s)_{s=1}^{t-1}\big) = 1$ *almost surely,* $v(x, \eta) = x + \eta$*, and that* $\boldsymbol{\varepsilon}$ *satisfies* (1) *for all* t.

Then

$$\mathbb{E}\left[\frac{\sum_{t=1}^{T} \varepsilon_t v(x_t, \eta_t)}{\sum_{s=1}^{T} v(x_s, \eta_s)}\right] = \mu,$$

suppressing in x_t the arguments $\boldsymbol{\eta}$ and $\left(\boldsymbol{\eta}, (\varepsilon_s)_{s=1}^{t-1}\right)$ in cases 1 and 2, respectively.

Proof of Lemma 11. In the first case, we compute

$$\mathbb{E}\left[\frac{\sum_{t=1}^T \varepsilon_t v(x_t,\eta_t)}{\sum_{s=1}^T v(x_s,\eta_s)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{t=1}^T \varepsilon_t v(x_t,\eta_t)}{\sum_{s=1}^T v(x_s,\eta_s)}\middle|\boldsymbol{\eta}\right]\right] = \mathbb{E}\left[\frac{\sum_{t=1}^T \mathbb{E}[\varepsilon_t|\boldsymbol{\eta}] v(x_t,\eta_t)}{\sum_{s=1}^T v(x_s,\eta_s)}\right] = \mu,$$

using $\mathbb{E}[\varepsilon_t|\boldsymbol{\eta}] = \mu$ almost surely. For the second case, we begin by observing that $v(x,\eta) = x + \eta$ implies that daily market volume

$$\sum_{s=1}^T v(x_s,\eta_s) = \sum_{s=1}^T x_s + \sum_{s=1}^T \eta_s = 1 + \sum_{s=1}^T \eta_s$$

depends only on $\boldsymbol{\eta}$. We then compute

$$\begin{aligned}
\mathbb{E}\left[\frac{\sum_{t=1}^T \varepsilon_t v(x_t,\eta_t)}{\sum_{s=1}^T v(x_s,\eta_s)}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{t=1}^T \varepsilon_t v(x_t,\eta_t)}{\sum_{s=1}^T v(x_s,\eta_s)}\middle|\boldsymbol{\eta}\right]\right] \\
&= \mathbb{E}\left[\frac{\sum_{t=1}^T \mathbb{E}[\varepsilon_t x_t|\boldsymbol{\eta}] + \sum_{t=1}^T \mathbb{E}[\varepsilon_t \eta_t|\boldsymbol{\eta}]}{1 + \sum_{s=1}^T \eta_s}\right] \\
&= \mathbb{E}\left[\frac{\mu + \mu\sum_{t=1}^T \eta_t}{1 + \sum_{s=1}^T \eta_s}\right] = \mu,
\end{aligned}$$

using Lemma 10 and $\mathbb{E}[\varepsilon_t|\boldsymbol{\eta}] = \mu$ almost surely. □

Lemma 12. *When $v(x,\eta)$ satisfies the equivalent statements of Proposition 2, then*

$$\left(\sum_{t=1}^T x_t\right)\left(\sum_{t=1}^T h\left(\frac{x_t}{\eta_t}\right) v(x_t,\eta_t)\right) \leq \left(\sum_{t=1}^T v(x_t,\eta_t)\right)\left(\sum_{t=1}^T h\left(\frac{x_t}{\eta_t}\right) x_t\right) \tag{4}$$

for all $T \in \mathbb{N}$ and $(x_1,\eta_1),\ldots,(x_T,\eta_T) \in dom(v)$. For $(x_1,\eta_1),\ldots,(x_T,\eta_T) \in dom(v)$ with $\sum_{t=1}^T x_t = 1$, equality in (4) holds if and only if $x_t = \frac{\eta_t}{\sum_{s=1}^T \eta_s}$ for all $t = 1,\ldots,T$.

Proof of Lemma 12. We prove (4) by induction over T.
Induction base: For $T = 1$, (4) becomes

$$x_1 h\left(\frac{x_1}{\eta_1}\right) v(x_1,\eta_1) \leq v(x_1,\eta_1) h\left(\frac{x_1}{\eta_1}\right) x_1,$$

which holds with equality.

Induction step: We can write (4) as

$$\left(\sum_{t=1}^{T-1} x_t\right)\left(\sum_{t=1}^{T-1} h\left(\frac{x_t}{\eta_t}\right)v(x_t,\eta_t)\right) + h\left(\frac{x_T}{\eta_T}\right)v(x_T,\eta_T)\sum_{t=1}^{T-1} x_t + x_T\sum_{t=1}^{T-1} h\left(\frac{x_t}{\eta_t}\right)v(x_t,\eta_t)$$
$$\leq \left(\sum_{t=1}^{T-1} v(x_t,\eta_t)\right)\left(\sum_{t=1}^{T-1} h\left(\frac{x_t}{\eta_t}\right)x_t\right) + h\left(\frac{x_T}{\eta_T}\right)x_T\sum_{t=1}^{T-1} v(x_t,\eta_t) + v(x_T,\eta_T)\sum_{t=1}^{T-1} h\left(\frac{x_t}{\eta_t}\right)x_t.$$

Using the induction hypothesis, it is enough to show

$$h\left(\frac{x_T}{\eta_T}\right)v(x_T,\eta_T)x_t + x_T h\left(\frac{x_t}{\eta_t}\right)v(x_t,\eta_t) \leq h\left(\frac{x_T}{\eta_T}\right)x_T v(x_t,\eta_t) + v(x_T,\eta_T)h\left(\frac{x_t}{\eta_t}\right)x_t \qquad (5)$$

for every $t = 1, 2, \ldots, T-1$. Rearranging terms, (5) is equivalent to

$$\left(x_t v(x_T,\eta_T) - x_T v(x_t,\eta_t)\right)\left(h\left(\frac{x_T}{\eta_T}\right) - h\left(\frac{x_t}{\eta_t}\right)\right) \leq 0. \qquad (6)$$

If $x_t = 0$ or $x_T = 0$, then (6) holds. In other cases, using $v(x,\eta) = xV(x/\eta)$ for $x \neq 0$, it becomes

$$x_t x_T\left(V\left(\frac{x_T}{\eta_T}\right) - V\left(\frac{x_t}{\eta_t}\right)\right)\left(h\left(\frac{x_T}{\eta_T}\right) - h\left(\frac{x_t}{\eta_t}\right)\right) \leq 0,$$

which is satisfied for all $(x_t,\eta_t), (x_T,\eta_T) \in \mathrm{dom}(v)$ because V is decreasing and h is increasing.

We now turn to the second part. It is straightforward to check that if $x_t = \frac{\eta_t}{\sum_{s=1}^T \eta_s}$ for all $t = 1, \ldots, T$, then (4) holds with equality. For the converse, consider $(x_1,\eta_1), \ldots, (x_T,\eta_T) \in \mathrm{dom}(v)$ with $\sum_{t=1}^T x_t = 1$ and suppose that (4) holds with equality. By the above induction hypothesis, this can be the case only if (6) holds with equality for all $t = 1, \ldots, T-1$. Note that (6) holds with equality if and only if $x_t v(x_T,\eta_T) = x_T v(x_t,\eta_t)$ or $x_T\eta_t = x_t\eta_T$. However, $x_t v(x_T,\eta_T) = x_T v(x_t,\eta_t)$ implies $x_T\eta_t = x_t\eta_T$. To see this, suppose that $x_T\eta_t > x_t\eta_T$. This can be the case only if $x_T \neq 0$. We separately consider two cases. In the first, suppose further that $x_t = 0$. Then we obtain $x_T v(x_t,\eta_t) > 0 = x_t v(x_T,\eta_T)$, where the inequality follows because $v(x_t,\eta_t)$ is positive. In the second case, suppose instead that $x_t \neq 0$. Then we obtain

$$x_T v(x_t,\eta_t) = x_T x_t V(x_t/\eta_t) > x_T x_t V(x_T/\eta_T) = x_t v(x_T,\eta_T),$$

where the inequality follows because V is strictly decreasing. By symmetry, $x_T\eta_t < x_t\eta_T$ implies $x_T v(x_t, \eta_t) < x_t v(x_T, \eta_T)$ so that the equality $x_t v(x_T, \eta_T) = x_T v(x_t, \eta_t)$ can hold only if $x_T\eta_t = x_t\eta_T$. Hence, we can have equality in (4) only if $x_T\eta_t = x_t\eta_T$ for all t, which means $x_t = \frac{\eta_t}{\sum_{s=1}^{T}\eta_s}$ since $\sum_{t=1}^{T} x_t = 1$. □

B Proofs

Throughout this Appendix, we will typically write a trading policy as $\boldsymbol{x}(\boldsymbol{\eta})$ for notational convenience, despite the potential dependence on previous prices (*viz.* when the proofs apply to the version of the model considered in Section 7.2.1).

Proof of Proposition 1. By Condition 1, there exist $s \in \{1, 2, \ldots, T-1\}$ and a set Ω with strictly positive probability such that $\mathbb{E}[\bar{p}_T|\boldsymbol{\eta}] > \mathbb{E}[\bar{p}_s|\boldsymbol{\eta}]$ on Ω. For a vector $\hat{\boldsymbol{x}}$ with $T-1$ entries, we define

$$f(\hat{\boldsymbol{x}}) = \mathbb{E}\left[u\left(p_T - \sum_{t=1}^{T-1} p_t\hat{x}_t - p_T\left(1 - \sum_{t=1}^{T-1}\hat{x}_t\right)\right)\middle|\boldsymbol{\eta}\right] = \mathbb{E}\left[u\left(\sum_{t=1}^{T-1}(p_T - p_t)\hat{x}_t\right)\middle|\boldsymbol{\eta}\right].$$

We recall that f depends on $\hat{\boldsymbol{x}}$ also through p_T and $\boldsymbol{p}$, and note that $f(\hat{\mathbf{0}}) = u(0)$ for $\hat{\mathbf{0}}$ the vector of $T-1$ zeros. Thus, we compute

$$\begin{aligned}
\frac{\partial f}{\partial \hat{x}_s}(\hat{\mathbf{0}}) &= \mathbb{E}\left[u'\left(\sum_{t=1}^{T-1}(p_T - p_t)\hat{x}_t\right)\left(\sum_{t=1}^{T-1}\hat{x}_t\frac{\partial}{\partial \hat{x}_s}(p_T - p_t) + (p_T - p_s)\right)\middle|\boldsymbol{\eta}\right]\Bigg|_{\hat{\boldsymbol{x}}=\hat{\mathbf{0}}} \\
&= \mathbb{E}[u'(0)(p_T - p_s)|\boldsymbol{\eta}]\big|_{\hat{\boldsymbol{x}}=\hat{\mathbf{0}}} \\
&= u'(0)\mathbb{E}[\bar{p}_T - \bar{p}_s|\boldsymbol{\eta}] \\
&> 0
\end{aligned}$$

on Ω. The properties $f(\hat{\mathbf{0}}) = u(0)$ and $\frac{\partial f}{\partial x_s}(\hat{\mathbf{0}}) > 0$ on Ω imply that there exists an $\boldsymbol{\eta}$-measurable $\hat{\boldsymbol{x}}(\boldsymbol{\eta})$ with $\hat{x}_t(\boldsymbol{\eta}) = 0$ for all $t \neq s$, $\hat{x}_s(\boldsymbol{\eta}) \in (0, 1)$ and $f(\hat{\boldsymbol{x}}(\boldsymbol{\eta})) > u(0)$ on Ω. We define $\boldsymbol{x}(\boldsymbol{\eta})$ in the following way. On Ω, let $x_t(\boldsymbol{\eta}) = \hat{x}_t(\boldsymbol{\eta})$ for all $t < T$ and

$x_T(\boldsymbol{\eta}) = 1 - \sum_{t=1}^{T-1} \hat{x}_t(\boldsymbol{\eta})$. On $\Omega^{\complement}$, let $x_t(\boldsymbol{\eta}) = (0, \ldots, 0, 1)$. This yields

$$\begin{aligned}
\mathbb{E}\big[u\big(p_T - \boldsymbol{p}\cdot\boldsymbol{x}(\boldsymbol{\eta})\big)\big] &= \mathbb{E}\Big[\mathbb{E}\big[u\big(p_T - \boldsymbol{p}\cdot\boldsymbol{x}(\boldsymbol{\eta})\big)\big|\boldsymbol{\eta}\big]\mathbb{1}_{\Omega} + \mathbb{E}\big[u\big(p_T - \boldsymbol{p}\cdot\boldsymbol{x}(\boldsymbol{\eta})\big)\big|\boldsymbol{\eta}\big]\mathbb{1}_{\Omega^{\complement}}\Big] \\
&= \mathbb{E}\big[f(\hat{\boldsymbol{x}}(\boldsymbol{\eta}))\mathbb{1}_{\Omega} + u(0)\mathbb{1}_{\Omega^{\complement}}\big] \\
&> \mathbb{E}\big[u(0)\mathbb{1}_{\Omega} + u(0)\mathbb{1}_{\Omega^{\complement}}\big] \\
&= u(0),
\end{aligned}$$

so that the constraint (IR) is slack. This immediately implies that $\tau^{MOC} = p_T$ is not optimal. Indeed, there must then exist some constant $c > 0$ such that $\hat{\tau} = \tau^{MOC} - c$ also satisfies (IR). Moreover, any such contract provides the dealer with the same incentives as τ^{MOC} and is therefore strictly cheaper for the client. □

Proof of Proposition 2. It is straightforward to check that (i) implies (ii). For the converse, we define a function g by $g(x,\eta) = v(x,\eta)/x$ for $(x,\eta) \in \mathrm{dom}(v)$ with $x \neq 0$. For $(x',\eta'), (x'',\eta'') \in \mathrm{dom}(v)$ with $x'x'' \neq 0$ and $x'/\eta' = x''/\eta''$, we deduce

$$g(x',\eta') = \frac{v(x',\eta')}{x'} = \frac{v\big(\frac{\eta'}{\eta''}x'', \frac{\eta'}{\eta''}\eta''\big)}{\frac{\eta'}{\eta''}x''} = \frac{\frac{\eta'}{\eta''}v(x'',\eta'')}{\frac{\eta'}{\eta''}x''} = \frac{v(x'',\eta'')}{x''} = g(x'',\eta''), \tag{7}$$

where the third equality uses the fact that $v(x,\eta)$ is homogeneous of degree one. We partition $\mathrm{dom}(v)$ into sets $D_y = \{(x,\eta) \in \mathrm{dom}(v) : x/\eta = y\}$ for $y \in \mathbb{R}_+$. If there are no $(x,\eta) \in \mathrm{dom}(v)$ with $x/\eta = y$, we set $D_y = \emptyset$. Note that $\mathrm{dom}(v) = \bigcup_{y\in\mathbb{R}_+} D_y$ and $D_y \cap D_z = \emptyset$ for $y \neq z$. For every y with $D_y \neq \emptyset$ with $y \neq 0$, (7) implies that $g(x,\eta)$ takes the same value for all $(x,\eta) \in D_y$. Therefore, we can write $g(x,\eta) = V\big(\frac{x}{\eta}\big)$ for a function V and $(x,\eta) \in \mathrm{dom}(v)$ with $x \neq 0$, so that $v(x,\eta) = xV\big(\frac{x}{\eta}\big)$ for all $(x,\eta) \in \mathrm{dom}(v)$. Because $v(x,\eta)$ is strictly increasing in η for $(x,\eta) \in \mathrm{dom}(v)$ with $x \neq 0$, we obtain that $V(y)$ is strictly decreasing for $y \neq 0$. □

Proof of Lemma 3. Plugging in $\boldsymbol{p}$, a trading policy $\boldsymbol{x}(\cdot)$ is first best if for all $\boldsymbol{\eta}$, $\boldsymbol{x}(\boldsymbol{\eta})$ minimizes the following objective subject to the constraint $\sum_{t=1}^{T} x_t = 1$:

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(h\left(\frac{x_t}{\eta_t}\right)x_t + \varepsilon_t x_t\right)\middle|\boldsymbol{\eta}\right] = \sum_{t=1}^{T}\mathbb{E}\left[h\left(\frac{x_t}{\eta_t}\right)x_t\middle|\boldsymbol{\eta}\right] + \sum_{t=1}^{T}\mathbb{E}[\varepsilon_t x_t|\boldsymbol{\eta}].$$

By Lemma 10, the last term equates to μ almost surely. We therefore find that this objective, the expected trading cost conditional on $\boldsymbol{\eta}$, is

$$\begin{aligned}
&\mu + \mathbb{E}\left[\sum_{t=1}^{T} h\left(\frac{x_t}{\eta_t}\right) x_t \middle| \boldsymbol{\eta}\right] \\
&= \mu + \mathbb{E}\left[\left(\sum_{s=1}^{T} \eta_s\right)\left(\frac{1}{\sum_{s=1}^{T}\eta_s}\sum_{t=1}^{T}\frac{x_t}{\eta_t} h\left(\frac{x_t}{\eta_t}\right)\eta_t\right)\middle|\boldsymbol{\eta}\right] \\
&\geq \mu + \mathbb{E}\left[\left(\sum_{s=1}^{T} \eta_s\right)\left(\frac{1}{\sum_{s=1}^{T}\eta_s}\sum_{t=1}^{T}\frac{x_t}{\eta_t}\eta_t\right) h\left(\frac{1}{\sum_{s=1}^{T}\eta_s}\sum_{t=1}^{T}\frac{x_t}{\eta_t}\eta_t\right)\middle|\boldsymbol{\eta}\right] \\
&= \mu + \mathbb{E}\left[h\left(\frac{1}{\sum_{s=1}^{T}\eta_s}\sum_{t=1}^{T} x_t\right)\left(\sum_{t=1}^{T} x_t\right)\middle|\boldsymbol{\eta}\right] \\
&= \mu + h\left(\frac{1}{\sum_{t=1}^{T}\eta_t}\right)
\end{aligned}$$

almost surely, where the second step in the above uses Jensen's inequality applied to the convex function $yh(y)$, and the final step uses $\sum_{t=1}^{T} x_t = 1$. Equality in the above holds if and only if $x_1/\eta_1 = x_2/\eta_2 = \cdots = x_T/\eta_T$. Since we must have $\sum_{t=1}^{T} x_t = 1$, the expected trading cost conditional on $\boldsymbol{\eta}$ is minimized if and only if the trading schedule is $\boldsymbol{x} = \left(\frac{\eta_t}{\sum_{s=1}^{T}\eta_s}\right)_{t=1}^{T}$. Thus, $\boldsymbol{x}^{FB}(\cdot)$ is the first-best trading policy, and it results in the unconditional expected trading cost

$$\mu + \mathbb{E}\left[h\left(\frac{1}{\sum_{t=1}^{T}\eta_t}\right)\right].$$

The last statement of Lemma 3 follows from

$$\begin{aligned}
\frac{v(x_t^{FB}(\boldsymbol{\eta}),\eta_t)}{\sum_{s=1}^{T} v(x_s^{FB}(\boldsymbol{\eta}),\eta_s)} &= \frac{x_t^{FB}(\boldsymbol{\eta}) V\left(\frac{x_t^{FB}(\boldsymbol{\eta})}{\eta_t}\right)}{\sum_{s=1}^{T} x_s^{FB}(\boldsymbol{\eta}) V\left(\frac{x_s^{FB}(\boldsymbol{\eta})}{\eta_s}\right)} \\
&= \frac{\frac{\eta_t}{\sum_{r=1}^{T}\eta_r} V\left(\frac{1}{\sum_{r=1}^{T}\eta_r}\right)}{\sum_{s=1}^{T}\frac{\eta_s}{\sum_{r=1}^{T}\eta_r} V\left(\frac{1}{\sum_{r=1}^{T}\eta_r}\right)} = \frac{\eta_t}{\sum_{r=1}^{T}\eta_r} = x_t^{FB}(\boldsymbol{\eta}),
\end{aligned}$$

where the first equality uses that $v(x,\eta) = xV\left(\frac{x}{\eta}\right)$ for all $(x,\eta) \in \mathrm{dom}(v)$ by assumption. □

Proof of Lemma 4. Sufficiency. If τ satisfies condition (i), then $(\tau, \boldsymbol{x}^{FB}(\cdot))$ satisfies (IC). Similarly, if τ satisfies condition (ii), then $(\tau, \boldsymbol{x}^{FB}(\cdot))$ satisfies (IR). Furthermore, condition (ii) also implies

$$\mathbb{E}\big[\tau\big(\boldsymbol{p}, \boldsymbol{v}(\boldsymbol{x}^{FB}(\boldsymbol{\eta}), \boldsymbol{\eta})\big)\big] = \mathbb{E}\big[\boldsymbol{p} \cdot \boldsymbol{x}^{FB}(\boldsymbol{\eta})\big] = \mu + \mathbb{E}\left[h\left(\frac{1}{\sum_{t=1}^{T} \eta_t}\right)\right]$$

by Lemma 3. Moreover, no pair $(\tau', \boldsymbol{x}(\cdot))$ satisfying (IR) can better this objective. To see this, first note that (IR) requires

$$\mathbb{E}\big[u\big(\tau'\left(\boldsymbol{p}, \boldsymbol{v}(\boldsymbol{x}(\boldsymbol{\eta}), \boldsymbol{\eta})\right) - \boldsymbol{p} \cdot \boldsymbol{x}(\boldsymbol{\eta})\big)\big] \geq u(0).$$

Since u is concave and strictly increasing, this requires

$$\mathbb{E}\left[\tau'(\boldsymbol{p}, \boldsymbol{v}(\boldsymbol{x}(\boldsymbol{\eta}), \boldsymbol{\eta}))\right] \geq \mathbb{E}\left[\boldsymbol{p} \cdot \boldsymbol{x}(\boldsymbol{\eta})\right] \geq \mu + \mathbb{E}\left[h\left(\frac{1}{\sum_{t=1}^{T} \eta_t}\right)\right],$$

where the last step follows from Lemma 3.

Necessity. Now assume that there exists τ satisfying the two conditions and let τ' also be an optimal contract. Then there must exist some $\boldsymbol{x}(\cdot)$ such that $(\tau', \boldsymbol{x}(\cdot))$ satisfies (IR) and (IC) and where

$$\mathbb{E}\left[\tau'(\boldsymbol{p}, \boldsymbol{v}(\boldsymbol{x}(\boldsymbol{\eta}), \boldsymbol{\eta}))\right] = \mu + \mathbb{E}\left[h\left(\frac{1}{\sum_{t=1}^{T} \eta_t}\right)\right]. \tag{8}$$

First, we claim that $\boldsymbol{x}(\cdot) = \boldsymbol{x}^{FB}(\cdot)$ almost surely. Suppose by way of contradiction that this is not the case. Then Lemma 3 implies

$$\mathbb{E}\left[\boldsymbol{p} \cdot \boldsymbol{x}(\boldsymbol{\eta})\right] > \mu + \mathbb{E}\left[h\left(\frac{1}{\sum_{t=1}^{T} \eta_t}\right)\right]. \tag{9}$$

Combining (8) and (9),

$$\mathbb{E}\left[\tau'(\boldsymbol{p}, \boldsymbol{v}(\boldsymbol{x}(\boldsymbol{\eta}), \boldsymbol{\eta}))\right] < \mathbb{E}\left[\boldsymbol{p} \cdot \boldsymbol{x}(\boldsymbol{\eta})\right].$$

Because u is concave and strictly increasing, this implies that

$$\mathbb{E}\left[u\left(\tau'(\boldsymbol{p}, \boldsymbol{v}(\boldsymbol{x}(\boldsymbol{\eta}), \boldsymbol{\eta})) - \boldsymbol{p} \cdot \boldsymbol{x}(\boldsymbol{\eta})\right)\right] < u(0),$$

which violates (IR). Next, observe that because $\boldsymbol{x}(\cdot) = \boldsymbol{x}^{FB}(\cdot)$ almost surely, $(\tau', \boldsymbol{x}(\cdot))$ satisfying (IC) implies that $(\tau', \boldsymbol{x}^{FB}(\cdot))$ satisfies it as well, which implies condition (i).

Finally, suppose by way of contradiction that condition (ii) is violated. Because $\boldsymbol{x}(\cdot) = \boldsymbol{x}^{FB}(\cdot)$ almost surely, this implies it is not the case that $\tau'(\boldsymbol{p}, \boldsymbol{v}(\boldsymbol{x}(\boldsymbol{\eta}), \boldsymbol{\eta})) = \boldsymbol{p} \cdot \boldsymbol{x}(\boldsymbol{\eta})$ almost surely. If u is strictly concave, then Jensen's inequality implies that

$$\mathbb{E}\left[u\left(\tau'(\boldsymbol{p}, \boldsymbol{v}(\boldsymbol{x}(\boldsymbol{\eta}), \boldsymbol{\eta})) - \boldsymbol{p} \cdot \boldsymbol{x}(\boldsymbol{\eta})\right)\right] < u\left(\mathbb{E}\left[\tau'(\boldsymbol{p}, \boldsymbol{v}(\boldsymbol{x}(\boldsymbol{\eta}), \boldsymbol{\eta})) - \boldsymbol{p} \cdot \boldsymbol{x}(\boldsymbol{\eta})\right]\right).$$

Because $(\tau', \boldsymbol{x}(\cdot))$ satisfies (IR), the left-hand side is bounded below by $u(0)$. Because u is increasing, this implies

$$\mathbb{E}\left[\tau'(\boldsymbol{p}, \boldsymbol{v}(\boldsymbol{x}(\boldsymbol{\eta}), \boldsymbol{\eta})) - \boldsymbol{p} \cdot \boldsymbol{x}(\boldsymbol{\eta})\right] > 0. \tag{10}$$

Combining (8) and (10), we obtain

$$\mathbb{E}\left[\boldsymbol{p} \cdot \boldsymbol{x}(\boldsymbol{\eta})\right] < \mu + \mathbb{E}\left[h\left(\frac{1}{\sum_{t=1}^{T} \eta_t}\right)\right],$$

which is impossible by Lemma 3. □

Proof of Proposition 5. To show that τ^{VWAP} is an optimal contract, it suffices to establish that it satisfies the two conditions of Lemma 4. Applying Jensen's inequality to the concave function u, we obtain that the dealer's expected utility from pursuing a trading schedule $\boldsymbol{x}$

is

$$\begin{aligned}
\mathbb{E}\big[u\big(\tau^{VWAP} - \boldsymbol{p}\cdot\boldsymbol{x}\big)\big] &= \mathbb{E}\left[u\left(\sum_{t=1}^{T}\frac{\big(h\big(\frac{x_t}{\eta_t}\big)+\varepsilon_t\big)v(x_t,\eta_t)}{\sum_{s=1}^{T}v(x_s,\eta_s)} - \sum_{t=1}^{T}\left(h\left(\frac{x_t}{\eta_t}\right)+\varepsilon_t\right)x_t\right)\right] \\
&\leq \mathbb{E}\left[u\left(\sum_{t=1}^{T}\frac{\varepsilon_t v(x_t,\eta_t)}{\sum_{s=1}^{T}v(x_s,\eta_s)} - \sum_{t=1}^{T}\varepsilon_t x_t\right)\right] \qquad (11)\\
&\leq u\left(\mathbb{E}\left[\sum_{t=1}^{T}\frac{\varepsilon_t v(x_t,\eta_t)}{\sum_{s=1}^{T}v(x_s,\eta_s)} - \sum_{t=1}^{T}\varepsilon_t x_t\right]\right) \qquad (12)\\
&= u(0),
\end{aligned}$$

where (11) follows from the first part of Lemma 12; and the last equality is implied by Lemmas 10 and 11. Equality in (11) holds if and only if (4) holds almost surely, hence if and only if $x_t = \frac{\eta_t}{\sum_{s=1}^{T}\eta_s}$ almost surely, by the second part of Lemma 12. Note that in this case, we also have $x_t = \frac{v(x_t,\eta_t)}{\sum_{s=1}^{T}v(x_s,\eta_s)}$ almost surely by Lemma 3, so that there is equality in (12) as well. Thus, a trading policy $\boldsymbol{x}(\cdot)$ maximizes $\mathbb{E}\big[u\big(\tau^{VWAP}(\boldsymbol{p},\boldsymbol{v}(\boldsymbol{x}(\boldsymbol{\eta}),\boldsymbol{\eta}))-\boldsymbol{p}\cdot\boldsymbol{x}(\boldsymbol{\eta})\big)\big]$ if and only if it implies that $x_t = \frac{\eta_t}{\sum_{s=1}^{T}\eta_s}$ almost surely, or equivalently, if and only if it equals $\boldsymbol{x}^{FB}(\cdot)$ almost surely.

We therefore conclude that $(\tau^{VWAP}, \boldsymbol{x}^{FB}(\cdot))$ satisfies (IC), which implies condition (i) of Lemma 4. But in fact, we also obtain the stronger conclusion that for all trading policies $\hat{\boldsymbol{x}}(\cdot)$ not equal to $\boldsymbol{x}^{FB}(\cdot)$ almost surely, (IC) holds with strict inequality:

$$\mathbb{E}\big[u\big(\tau^{VWAP}\big(\boldsymbol{p},\boldsymbol{v}(\boldsymbol{x}^{FB}(\boldsymbol{\eta}),\boldsymbol{\eta})\big)-\boldsymbol{p}\cdot\boldsymbol{x}^{FB}(\boldsymbol{\eta})\big)\big] > \mathbb{E}\big[u\big(\tau^{VWAP}(\boldsymbol{p},\boldsymbol{v}(\hat{\boldsymbol{x}}(\boldsymbol{\eta}),\boldsymbol{\eta}))-\boldsymbol{p}\cdot\hat{\boldsymbol{x}}(\boldsymbol{\eta})\big)\big].$$

The same computation reveals that $\tau^{VWAP}\big(\boldsymbol{p},\boldsymbol{v}(\boldsymbol{x}^{FB}(\boldsymbol{\eta}),\boldsymbol{\eta})\big)-\boldsymbol{p}\cdot\boldsymbol{x}^{FB}(\boldsymbol{\eta}) = 0$. We therefore obtain condition (ii) of Lemma 4. □

Proof of Proposition 6. Suppose that u is strictly concave and that the distributions of $\boldsymbol{\varepsilon}$ and $\boldsymbol{\eta}$ have full support over $\mathbb{R}^T$ and $\mathbb{R}^T_{++}$, respectively. Suppose that τ is an optimal contract. In proving Proposition 5, we established that τ^{VWAP} satisfies the conditions of Lemma 4. Therefore, the second half of that lemma requires that τ does the same. Condition (ii) of

that lemma requires that both of the following hold almost surely:

$$\tau\big(\boldsymbol{p}, \boldsymbol{v}(\boldsymbol{x}^{FB}(\boldsymbol{\eta}), \boldsymbol{\eta})\big) = \boldsymbol{p} \cdot \boldsymbol{x}^{FB}(\boldsymbol{\eta})$$
$$\tau^{VWAP}\big(\boldsymbol{p}, \boldsymbol{v}(\boldsymbol{x}^{FB}(\boldsymbol{\eta}), \boldsymbol{\eta})\big) = \boldsymbol{p} \cdot \boldsymbol{x}^{FB}(\boldsymbol{\eta})$$

Using $\boldsymbol{\iota}$ to denote a vector of ones, we conclude that the following holds almost surely:

$$\tau\left(h\left(\frac{1}{\sum_{t=1}^{T}\eta_t}\right)\boldsymbol{\iota} + \boldsymbol{\varepsilon}, \boldsymbol{v}\left(\boldsymbol{x}^{FB}(\boldsymbol{\eta}), \boldsymbol{\eta}\right)\right) = \tau^{VWAP}\left(h\left(\frac{1}{\sum_{t=1}^{T}\eta_t}\right)\boldsymbol{\iota} + \boldsymbol{\varepsilon}, \boldsymbol{v}\left(\boldsymbol{x}^{FB}(\boldsymbol{\eta}), \boldsymbol{\eta}\right)\right)$$

By the full-support assumptions on $\boldsymbol{\varepsilon}$ and $\boldsymbol{v}(\boldsymbol{x}^{FB}(\boldsymbol{\eta}), \boldsymbol{\eta})$, this requires that $\tau = \tau^{VWAP}$ almost everywhere on its domain. □

Proof of Proposition 8. To simplify the notation, we write $x = x_1$ so that $x_2 = 1 - x$. We also choose $f = \gamma/2$.

Step 1: we show $\mathcal{X}(\tau^{VWAP} + f) \neq \emptyset$ for $c \leq f$.
To this end, we compute

$$\begin{aligned}
&\mathbb{E}\big[\tau^{VWAP}(\boldsymbol{p}, \boldsymbol{x}(\boldsymbol{\eta}) + \boldsymbol{\eta} + \boldsymbol{\varepsilon}^v) - \boldsymbol{p} \cdot \boldsymbol{x}(\boldsymbol{\eta})\big|\boldsymbol{\eta}, \boldsymbol{\varepsilon}^v\big] \\
&= \mathbb{E}\Bigg[\frac{(cx + x/\eta_1 + \varepsilon_1)(x + \eta_1 + \varepsilon_1^v)}{1 + \eta_1 + \eta_2 + \varepsilon_1^v + \varepsilon_2^v} - (cx + x/\eta_1 + \varepsilon_1)x \\
&\qquad + \frac{(c + (1-x)/\eta_2 + \varepsilon_2)(1 - x + \eta_2 + \varepsilon_2^v)}{1 + \eta_1 + \eta_2 + \varepsilon_1^v + \varepsilon_2^v} - (c + (1-x)/\eta_2 + \varepsilon_2)(1-x)\Bigg|\boldsymbol{\eta}, \boldsymbol{\varepsilon}^v\Bigg] \\
&= \frac{(cx + x/\eta_1)(\eta_1 + \varepsilon_1^v - x(\eta_1 + \eta_2 + \varepsilon_1^v + \varepsilon_2^v))}{1 + \eta_1 + \eta_2 + \varepsilon_1^v + \varepsilon_2^v} \\
&\quad + \frac{(c + (1-x)/\eta_2)(-\eta_1 - \varepsilon_1^v + x(\eta_1 + \eta_2 + \varepsilon_1^v + \varepsilon_2^v))}{1 + \eta_1 + \eta_2 + \varepsilon_1^v + \varepsilon_2^v} \\
&= \frac{(cx - c + x/\eta_1 - (1-x)/\eta_2)(\eta_1 + \varepsilon_1^v - x(\eta_1 + \eta_2 + \varepsilon_1^v + \varepsilon_2^v))}{1 + \eta_1 + \eta_2 + \varepsilon_1^v + \varepsilon_2^v}.
\end{aligned} \tag{13}$$

Choosing $x = \frac{\eta_1}{\eta_1 + \eta_2}$, this expression simplifies to

$$\mathbb{E}\big[\tau^{VWAP}(\boldsymbol{p}, \boldsymbol{x}(\boldsymbol{\eta}) + \boldsymbol{\eta} + \boldsymbol{\varepsilon}^v) - \boldsymbol{p} \cdot \boldsymbol{x}(\boldsymbol{\eta})\big|\boldsymbol{\eta}, \boldsymbol{\varepsilon}^v\big] = \frac{c\frac{-\eta_2}{\eta_1+\eta_2}\left(\varepsilon_1^v\frac{\eta_2}{\eta_1+\eta_2} - \frac{\eta_1}{\eta_1+\eta_2}\varepsilon_2^v\right)}{1 + \eta_1 + \eta_2 + \varepsilon_1^v + \varepsilon_2^v}$$

so that

$$\mathbb{E}\big[\tau^{VWAP}(\boldsymbol{p}, \boldsymbol{x}(\boldsymbol{\eta}) + \boldsymbol{\eta} + \boldsymbol{\varepsilon}^v) - \boldsymbol{p} \cdot \boldsymbol{x}(\boldsymbol{\eta}) \big| \boldsymbol{\eta}, \boldsymbol{\varepsilon}^v\big] \geq \frac{-c\big(\frac{\eta_2}{\eta_1+\eta_2}\big)^2 \varepsilon_1^v}{1+\eta_1+\eta_2+\varepsilon_1^v+\varepsilon_2^v} \geq -c.$$

We deduce

$$\mathbb{E}\big[\tau^{VWAP}(\boldsymbol{p}, \boldsymbol{x}(\boldsymbol{\eta}) + \boldsymbol{\eta} + \boldsymbol{\varepsilon}^v) + f - \boldsymbol{p} \cdot \boldsymbol{x}(\boldsymbol{\eta})\big] \geq f - c \geq 0,$$

which concludes the proof of $\mathcal{X}(\tau^{VWAP} + f) \neq \emptyset$ for $c \leq f$.

Step 2: lower bound for the right-hand side of (3).
Because the second-best outcome cannot be better than that of the first best, we have

$$\inf_{\tau, \boldsymbol{x} \in \mathcal{X}(\tau)} \mathbb{E}\big[\tau\big(\boldsymbol{p}, \boldsymbol{x}(\boldsymbol{\eta}) + \boldsymbol{\eta} + \boldsymbol{\varepsilon}^v\big)\big] + \gamma \geq \mathbb{E}\big[\boldsymbol{p} \cdot \boldsymbol{x}^{FB}(\boldsymbol{\eta})\big] + \gamma.$$

We can further compute

$$\begin{aligned}
&\mathbb{E}\big[\boldsymbol{p} \cdot \boldsymbol{x}^{FB}(\boldsymbol{\eta})\big] \\
&= \mu + \mathbb{E}\left[c(x^{FB})^2 + \frac{(x^{FB})^2}{\eta_1} + c(1-x^{FB}) + \frac{(1-x^{FB})^2}{\eta_2}\right] \\
&= \mu + \mathbb{E}\left[(c + 1/\eta_1 + 1/\eta_2)\left(x^{FB} - \frac{c/2 + 1/\eta_2}{c + 1/\eta_1 + 1/\eta_2}\right)^2 - \frac{(c/2+1/\eta_2)^2}{c+1/\eta_1+1/\eta_2} + c + 1/\eta_2\right] \\
&= \mu + \mathbb{E}\left[\frac{(c+1/\eta_1+1/\eta_2)(c+1/\eta_2) - (c/2+1/\eta_2)^2}{c+1/\eta_1+1/\eta_2}\right] \\
&= \mu + \mathbb{E}\left[\frac{3c^2/4 + c/\eta_1 + c/\eta_2 + 1/(\eta_1\eta_2)}{c+1/\eta_1+1/\eta_2}\right] \\
&\geq \mu + \mathbb{E}\left[\frac{1/(\eta_1\eta_2)}{1/\eta_1+1/\eta_2}\right],
\end{aligned}$$

where in the third step we deduced that $x^{FB} = \frac{c/2+1/\eta_2}{c+1/\eta_1+1/\eta_2}$, which minimizes the expression,

and in the last step we used that $\frac{3c^2/4+c/\eta_1+c/\eta_2+1/(\eta_1\eta_2)}{c+1/\eta_1+1/\eta_2}$ is an increasing function in c, since

$$\begin{aligned}
&\frac{\partial}{\partial c}\frac{3c^2/4+c/\eta_1+c/\eta_2+1/(\eta_1\eta_2)}{c+1/\eta_1+1/\eta_2}\\
&=\frac{(c+1/\eta_1+1/\eta_2)(3c/2+1/\eta_1+1/\eta_2)-(3c^2/4+c/\eta_1+c/\eta_2+1/(\eta_1\eta_2))}{(c+1/\eta_1+1/\eta_2)^2}\\
&=\frac{(1/\eta_1+1/\eta_2)(3c/2+1/\eta_1+1/\eta_2)+3c^2/4-1/(\eta_1\eta_2)}{(c+1/\eta_1+1/\eta_2)^2}\\
&>0.
\end{aligned}$$

Therefore, we find a lower bound for the right-hand side of (3), namely,

$$\inf_{\tau,\boldsymbol{x}\in\mathcal{X}(\tau)}\mathbb{E}\big[\tau\big(\boldsymbol{p},\boldsymbol{x}(\boldsymbol{\eta})+\boldsymbol{\eta}+\boldsymbol{\varepsilon}^v\big)\big]+\gamma\geq\mu+\mathbb{E}\left[\frac{1}{\eta_1+\eta_2}\right]+\gamma. \tag{14}$$

Step 3: we choose $\boldsymbol{x}\in\mathcal{X}(\tau^{VWAP}+f)$ and show

$$\mathbb{E}\big[\tau^{VWAP}\big(\boldsymbol{p},\boldsymbol{x}(\boldsymbol{\eta})+\boldsymbol{\eta}+\boldsymbol{\varepsilon}^v\big)\big]\leq\mu+\mathbb{E}\left[\frac{1}{\eta_1+\eta_2}\right]+\gamma/2 \tag{15}$$

for all $c\in[0,\overline{c}]$ and all $\boldsymbol{\varepsilon}^v\in[0,\overline{\varepsilon}^v]^2$ almost surely, where $\overline{c}>0$ and $\overline{\varepsilon}^v>0$ will be specified in the following. Note that this will complete the proof, as (15) implies (3) thanks to (14) and $f=\gamma/2$.

We require that $\overline{c}>0$ and $\overline{\varepsilon}^v>0$ satisfy $\overline{c}\leq f$ and

$$\mathbb{E}\left[\frac{1}{\eta_1+\eta_2}\max\left\{\eta_1\overline{c}+\max\{\overline{\varepsilon}^v/\eta_1,\overline{c}\eta_1\eta_2\}(\eta_1\overline{c}+1),\overline{c}(\eta_1+\eta_2)+\overline{\varepsilon}^v/\eta_2\right\}\right]\leq\gamma/2. \tag{16}$$

Note that such $\overline{c}>0$ and $\overline{\varepsilon}^v>0$ exist because

$$\lim_{\overline{c}\searrow0,\overline{\varepsilon}^v\searrow0}\mathbb{E}\left[\frac{1}{\eta_1+\eta_2}\max\left\{\eta_1\overline{c}+\max\{\overline{\varepsilon}^v/\eta_1,\overline{c}\eta_2\}(\eta_1\overline{c}+1),\overline{c}(\eta_1+\eta_2)+\overline{\varepsilon}^v/\eta_2\right\}\right]=0 \tag{17}$$

thanks to the assumption $\mathbb{E}\left[\frac{\eta_1\eta_2+1/\min\{\eta_1,\eta_2\}}{\eta_1+\eta_2}\right]<\infty$. Indeed, to show (17), we interchange limit and expectation and note that the expression within the expectation converges to zero as $\overline{c}\searrow0$ and $\overline{\varepsilon}^v\searrow0$. To be able to interchange limit and expectation, we need that the

expectation is finite, which can be shown using $\mathbb{E}\left[\frac{\eta_1\eta_2+1/\min\{\eta_1,\eta_2\}}{\eta_1+\eta_2}\right] < \infty$.

To prove (15), we first show

$$\frac{\eta_1 - \overline{\varepsilon}^v}{\eta_1+\eta_2} \leq x(\boldsymbol{\eta}) \leq \frac{\eta_1 + \max\left\{\overline{\varepsilon}^v, \overline{c}\eta_1\eta_2\right\}}{\eta_1+\eta_2} \tag{18}$$

almost surely. We achieve this by showing that if (18) did not hold almost surely, there would be a failure of (IC′), which refers to the fact that $\boldsymbol{x}(\cdot)$ maximizes $\mathbb{E}\big[\tau^{VWAP}(\boldsymbol{p}, \hat{\boldsymbol{x}}(\boldsymbol{\eta}) + \boldsymbol{\eta} + \boldsymbol{\varepsilon}^v) - \boldsymbol{p}\cdot\hat{\boldsymbol{x}}(\boldsymbol{\eta})\big]$ over $\hat{\boldsymbol{x}}(\cdot)$. And to show that, it suffices to demonstrate that if (18) is violated for some value of $\boldsymbol{\eta}$, then either (i) for all $\boldsymbol{\varepsilon}^v \in [0, \overline{\varepsilon}^v]^2$,

$$\frac{\partial}{\partial x}\mathbb{E}\big[\tau^{VWAP}(\boldsymbol{p}, \boldsymbol{x}(\boldsymbol{\eta}) + \boldsymbol{\eta} + \boldsymbol{\varepsilon}^v) - \boldsymbol{p}\cdot\boldsymbol{x}(\boldsymbol{\eta})\big|\boldsymbol{\eta}, \boldsymbol{\varepsilon}^v\big] > 0$$

or (ii) for all $\boldsymbol{\varepsilon}^v \in [0, \overline{\varepsilon}^v]^2$,

$$\frac{\partial}{\partial x}\mathbb{E}\big[\tau^{VWAP}(\boldsymbol{p}, \boldsymbol{x}(\boldsymbol{\eta}) + \boldsymbol{\eta} + \boldsymbol{\varepsilon}^v) - \boldsymbol{p}\cdot\boldsymbol{x}(\boldsymbol{\eta})\big|\boldsymbol{\eta}, \boldsymbol{\varepsilon}^v\big] < 0,$$

so that the first-order condition is not satisfied at $\boldsymbol{\eta}$. Using x to denote $x(\boldsymbol{\eta})$, we begin by using (13) to compute

$$\begin{aligned}
&\frac{\partial}{\partial x}\mathbb{E}\big[\tau^{VWAP}(\boldsymbol{p}, \boldsymbol{x}(\boldsymbol{\eta}) + \boldsymbol{\eta} + \boldsymbol{\varepsilon}^v) - \boldsymbol{p}\cdot\boldsymbol{x}(\boldsymbol{\eta})\big|\boldsymbol{\eta}, \boldsymbol{\varepsilon}^v\big] \\
&= \frac{\partial}{\partial x}\frac{(cx - c + x/\eta_1 - (1-x)/\eta_2)(\eta_1 + \varepsilon_1^v - x(\eta_1+\eta_2+\varepsilon_1^v+\varepsilon_2^v))}{1+\eta_1+\eta_2+\varepsilon_1^v+\varepsilon_2^v} \\
&= \frac{\partial}{\partial x}\frac{-x^2(c+1/\eta_1+1/\eta_2)(\eta_1+\eta_2+\varepsilon_1^v+\varepsilon_2^v) + x(c+1/\eta_1+1/\eta_2)(\eta_1+\varepsilon_1^v)}{1+\eta_1+\eta_2+\varepsilon_1^v+\varepsilon_2^v} \\
&\quad + \frac{\partial}{\partial x}\frac{x(c+1/\eta_2)(\eta_1+\eta_2+\varepsilon_1^v+\varepsilon_2^v) - (c+1/\eta_2)(\eta_1+\varepsilon_1^v)}{1+\eta_1+\eta_2+\varepsilon_1^v+\varepsilon_2^v} \\
&= \frac{-2x(c+1/\eta_1+1/\eta_2)(\eta_1+\eta_2+\varepsilon_1^v+\varepsilon_2^v)}{1+\eta_1+\eta_2+\varepsilon_1^v+\varepsilon_2^v} \\
&\quad + \frac{(c+1/\eta_1+1/\eta_2)(\eta_1+\varepsilon_1^v) + (c+1/\eta_2)(\eta_1+\eta_2+\varepsilon_1^v+\varepsilon_2^v)}{1+\eta_1+\eta_2+\varepsilon_1^v+\varepsilon_2^v}.
\end{aligned}$$

In the event $x > \frac{\eta_1+\max\{\overline{\varepsilon}^v,\overline{c}\eta_1\eta_2\}}{\eta_1+\eta_2}$, we have

$$\begin{aligned}
&-2x(c+1/\eta_1+1/\eta_2)(\eta_1+\eta_2+\varepsilon_1^v+\varepsilon_2^v)+(c+1/\eta_1+1/\eta_2)(\eta_1+\varepsilon_1^v)\\
&\quad+(c+1/\eta_2)(\eta_1+\eta_2+\varepsilon_1^v+\varepsilon_2^v)\\
&<(c+1/\eta_1+1/\eta_2)\left(\eta_1+\varepsilon_1^v-\frac{\eta_1+\overline{\varepsilon}^v}{\eta_1+\eta_2}(\eta_1+\eta_2+\varepsilon_1^v+\varepsilon_2^v)\right)\\
&\quad+(\eta_1+\eta_2+\varepsilon_1^v+\varepsilon_2^v)\left(c+1/\eta_2-\frac{1/\eta_2+\overline{c}}{1/\eta_1+1/\eta_2}(c+1/\eta_1+1/\eta_2)\right)\\
&<0
\end{aligned}$$

so that

$$\frac{\partial}{\partial x}\mathbb{E}\big[\tau^{VWAP}(\boldsymbol{p},\boldsymbol{x}(\boldsymbol{\eta})+\boldsymbol{\eta}+\boldsymbol{\varepsilon}^v)-\boldsymbol{p}\cdot\boldsymbol{x}(\boldsymbol{\eta})\big|\boldsymbol{\eta},\boldsymbol{\varepsilon}^v\big]<0\ \text{ on } x>\frac{\eta_1+\max\{\overline{\varepsilon}^v,\overline{c}\eta_1\eta_2\}}{\eta_1+\eta_2}.$$

Similarly, we can show

$$\frac{\partial}{\partial x}\mathbb{E}\big[\tau^{VWAP}(\boldsymbol{p},\boldsymbol{x}(\boldsymbol{\eta})+\boldsymbol{\eta}+\boldsymbol{\varepsilon}^v)-\boldsymbol{p}\cdot\boldsymbol{x}(\boldsymbol{\eta})\big|\boldsymbol{\eta},\boldsymbol{\varepsilon}^v\big]>0\ \text{ on } x<\frac{\eta_1-\overline{\varepsilon}^v}{\eta_1+\eta_2}.$$

Thus, we have shown that (18) holds almost surely. From (18), we deduce that

$$\begin{aligned}
&\mathbb{E}\big[\tau^{VWAP}(\boldsymbol{p},\boldsymbol{x}(\boldsymbol{\eta})+\boldsymbol{\eta}+\boldsymbol{\varepsilon}^v)\big]\\
&=\mu+\mathbb{E}\left[\frac{(cx+x/\eta_1)(x+\eta_1+\varepsilon_1^v)+(c+(1-x)/\eta_2)(1-x+\eta_2+\varepsilon_2^v)}{1+\eta_1+\eta_2+\varepsilon_1^v+\varepsilon_2^v}\right]\\
&\le\mu+\mathbb{E}\big[\max\{\overline{c}x+x/\eta_1,\overline{c}+(1-x)/\eta_2\}\big]\\
&\le\mu+\mathbb{E}\left[\max\left\{\frac{1+\max\{\overline{\varepsilon}^v/\eta_1,\overline{c}\eta_2\}}{\eta_1+\eta_2}(\eta_1\overline{c}+1),\overline{c}+\frac{1+\overline{\varepsilon}^v/\eta_2}{\eta_1+\eta_2}\right\}\right]\\
&\le\mu+\mathbb{E}\left[\frac{1}{\eta_1+\eta_2}\right]+\mathbb{E}\left[\frac{1}{\eta_1+\eta_2}\max\left\{\eta_1\overline{c}+\max\{\overline{\varepsilon}^v/\eta_1,\overline{c}\eta_2\}(\eta_1\overline{c}+1),\overline{c}(\eta_1+\eta_2)+\overline{\varepsilon}^v/\eta_2\right\}\right]\\
&\le\mu+\mathbb{E}\left[\frac{1}{\eta_1+\eta_2}\right]+\gamma/2,
\end{aligned}$$

using (16) for the last inequality. This shows (15) and concludes the proof. □

References

Almgren, Robert and Neil Chriss, "Optimal Execution of Portfolio Transactions," *Journal of Risk*, 2001, *3*, 5–40.

_ , Chee Thum, Emmanuel Hauptmann, and Hong Li, "Direct Estimation of Equity Market Impact," *Risk*, 2005, *18* (7), 58–62.

Anand, Amber, Mehrdad Samadi, Jonathan Sokobin, and Kumar Venkataraman, "Institutional Order Handling and Broker-Affiliated Trading Venues," *Working Paper*, 2019. https://www.finra.org/sites/default/files/OCE_WP_jan2019.pdf.

Barbon, Andrea, Marco Di Maggio, Francesco Franzoni, and Augustin Landier, "Brokers and Order Flow Leakage: Evidence from Fire Sales," *The Journal of Finance*, 2019, *74* (6), 2707–2749.

Basak, Suleyman and Anna Pavlova, "Asset Prices and Institutional Investors," *American Economic Review*, 2013, *103* (5), 1728–1758.

Battalio, Robert H., Brian C. Hatch, and Mehmet Sağlam, "The Cost of Routing Orders to High Frequency Traders," *Working Paper*, 2019. https://papers.ssrn.com/abstract_id=3281324.

Battalio, Robert, Shane A. Corwin, and Robert Jennings, "Can Brokers Have It All? On the Relation between Make-Take Fees and Limit Order Execution Quality," *The Journal of Finance*, 2016, *71* (5), 2193–2238.

Ben-David, Itzhak, Francesco Franzoni, Augustin Landier, and Rabih Moussawi, "Do Hedge Funds Manipulate Stock Prices?," *The Journal of Finance*, 2013, *68* (6), 2383–2434.

Berkman, Henk, Tim Brailsford, and Alex Frino, "A Note on Execution Costs for Stock Index Futures: Information versus Liquidity Effects," *Journal of Banking & Finance*, 2005, *29* (3), 565–577.

Berkowitz, Stephen A., Dennis E. Logue, and Eugene A. Noser, "The Total Cost of Transactions on the NYSE," *The Journal of Finance*, 1988, *43* (1), 97–112.

Bernhardt, Dan and Bart Taub, "Front-Running Dynamics," *Journal of Economic Theory*, 2008, *138* (1), 288–296.

Bertsimas, Dimitris and Andrew W. Lo, "Optimal Control of Execution Costs," *Journal of Financial Markets*, 1998, *1* (1), 1–50.

Bloomberg, "Traders Said to Rig Currency Rates to Profit Off Clients," June 2013. https://www.bloomberg.com/news/articles/2013-06-11/traders-said-to-

rig-currency-rates-to-profit-off-clients Accessed: 2018-04-25.

_ , "Cairn Energy Said to Be Victim of HSBC Currency Frontrunning," July 2016. https://www.bloomberg.com/news/articles/2016-07-20/cairn-energy-said-to-be-victim-of-hsbc-currency-frontrunning Accessed: 2019-11-15.

Brennan, Michael, "Agency and Asset Pricing," *Working Paper*, 1993. https://www.anderson.ucla.edu/documents/areas/fac/finance/6-93.pdf.

Buffa, Andrea M., Dimitri Vayanos, and Paul Woolley, "Asset Management Contracts and Equilibrium Prices," *Working Paper*, 2019. https://www.nber.org/papers/w20480.

Carhart, Mark M., Ron Kaniel, David K. Musto, and Adam V. Reed, "Leaning for the Tape: Evidence of Gaming Behavior in Equity Mutual Funds," *The Journal of Finance*, 2002, *57* (2), 661–693.

Carroll, Gabriel, "Robustness and Linear Contracts," *The American Economic Review*, 2015, *105* (2), 536–563.

Cartea, Álvaro and Sebastian Jaimungal, "A Closed-Form Execution Strategy to Target Volume Weighted Average Price," *SIAM Journal on Financial Mathematics*, 2016, *7* (1), 760–785.

CME Group, "Market Regulation Advisory Notice," August 2018. https://www.cmegroup.com/rulebook/files/cme-group-Rule-524.pdf Accessed: 2019-07-20.

Comerton-Forde, Carole and Tālis J Putniņš, "Measuring Closing Price Manipulation," *Journal of Financial Intermediation*, 2011, *20* (2), 135–158.

_ **and** _ , "Stock Price Manipulation: Prevalence and Determinants," *Review of Finance*, 2014, *18* (1), 23–66.

Coulter, Brian, Joel Shapiro, and Peter Zimmerman, "A Mechanism for LIBOR," *Review of Finance*, 2018, *22* (2), 491–520.

Cuoco, Domenico and Ron Kaniel, "Equilibrium Prices in the Presence of Delegated Portfolio Management," *Journal of Financial Economics*, 2011, *101* (2), 264–296.

Duffie, Darrell and Piotr Dworczak, "Robust Benchmark Design," *Working Paper*, 2018. https://www.nber.org/papers/w20540.

_ , _ , **and Haoxiang Zhu**, "Benchmarks in Search Markets," *The Journal of Finance*, 2017, *72* (5), 1983–2044.

Edelen, Roger M. and Gregory B. Kadlec, "Delegated Trading and the Speed of Adjustment in Security Prices," *Journal of Financial Economics*, 2012, *103* (2), 294–307.

Felixson, Karl and Anders Pelli, "Day End Returns—Stock Price Manipulation," *Journal of Multinational Financial Management*, 1999, *9* (2), 95–127.

Financial Industry Regulatory Authority, "FINRA Rule 5270. Front Running of Block Transactions," September 2013. http://finra.complinet.com/en/display/display_main.html?rbid=2403&element_id=10860 Accessed: 2019-07-20.

Financial Stability Board, "Foreign Exchange Benchmarks," *Final Report*, September 2014. https://www.fsb.org/2014/09/r_140930/ Accessed: 2019-11-13.

Fishman, Michael J. and Francis A. Longstaff, "Dual Trading in Futures Markets," *The Journal of Finance*, 1992, *47* (2), 643–671.

Foreign Exchange Joint Standing Committee Chief Dealers Sub-Group, "Draft Minutes of the 13th Meeting," July 2008. https://bit.ly/2NVdRrq Accessed: 2019-11-17.

Frei, Christoph and Nicholas Westray, "Optimal Execution of a VWAP Order: A Stochastic Control Approach," *Mathematical Finance*, 2015, *25* (3), 612–639.

Griffin, John M. and Amin Shams, "Manipulation in the VIX?," *The Review of Financial Studies*, 2017, *31* (4), 1377–1417.

Harris, Lawrence, "A Day-End Transaction Price Anomaly," *Journal of Financial and Quantitative Analysis*, 1989, *24* (1), 29–45.

Henderson, Brian J., Neil D. Pearson, and Li Wang, "Pre-Trade Hedging: Evidence from the Issuance of Retail Structured Products," *Working Paper*, 2019. http://ssrn.com/abstract=3068903.

Hillion, Pierre and Matti Suominen, "The Manipulation of Closing Prices," *Journal of Financial Markets*, 2004, *7* (4), 351–375.

Hölmstrom, Bengt, "Moral Hazard and Observability," *The Bell Journal of Economics*, 1979, *10* (1), 74–91.

Holmström, Bengt and Paul Milgrom, "Aggregation and Linearity in the Provision of Intertemporal Incentives," *Econometrica*, 1987, *55* (2), 303–328.

Humphery-Jenner, Mark, "Optimal VWAP Trading under Noisy Conditions," *Journal of Banking & Finance*, 2011, *35* (9), 2319–2329.

Ingersoll, Jonathan, William Goetzmann, Matthew Spiegel, and Ivo Welch, "Portfolio Performance Manipulation and Manipulation-Proof Performance Measures," *The Review of Financial Studies*, 2007, *20* (5), 1503–1546.

Kashyap, Anil K., Natalia Kovrijnykh, Jian Li, and Anna Pavlova, "The Benchmark Inclusion Subsidy," *Working Paper*, 2019. https://www.nber.org/2019LTAM/f117870.

pdf.

Kato, Takashi, "VWAP Execution as an Optimal Strategy," *JSIAM Letters*, 2015, *7*, 33–36.

Kumar, Praveen and Duane J. Seppi, "Futures Manipulation with "Cash Settlement"," *The Journal of Finance*, 1992, *47* (4), 1485–1502.

Mastromatteo, Iacopo, Bence Tóth, and Jean-Philippe Bouchaud, "Agent-Based Models for Latent Liquidity and Concave Price Impact," *Physical Review E*, 2014, *89*, 042805.

Mirrlees, James A., "The Optimal Structure of Incentives and Authority within an Organization," *The Bell Journal of Economics*, 1976, *7* (1), 105–131.

Nomura Research Institute, Ltd., "Asset Management Companies' Evaluation of Brokers – State of Asset Management Companies' Trading in 2014," April 2014. https://bit.ly/30vskOw Accessed: 2019-07-12.

Röell, Ailsa, "Dual-Capacity Trading and the Quality of the Market," *Journal of Financial Intermediation*, 1990, *1* (2), 105–124.

Saakvitne, Jo, "'Banging the Close': Price Manipulation or Optimal Execution?," *Working Paper*, 2016. http://ssrn.com/abstract=2753080.

Satish, Venkatesh, Abhay Saxena, and Max Palmer, "Predicting Intraday Trading Volume and Volume Percentages," *The Journal of Trading*, 2014, *9* (3), 15–25.

Subrahmanyam, Avanidhar, "A Theory of Trading in Stock Index Futures," *The Review of Financial Studies*, 1991, *4* (1), 17–51.

TheTrade, "The 2019 Algorithmic Trading Survey," 2019. https://www.thetradenews.com/surveys/algorithmic-trading-survey-long-results-2019/ Accessed: 2019-07-20.

The United States Commodity Futures Trading Commission, "CFTC v. Optiver US LLC, Optiver Holding BV, Optiver VOF, Christopher Dowson, Bastiaan van Kempen, and Randal Meijer," July 2008. https://www.cftc.gov/sites/default/files/idc/groups/public/@lrenforcementactions/documents/legalpleading/enfoptiveruscomplaint072408.pdf Accessed: 2018-08-09.

_ , "CFTC v. Daniel Shak and SHK Management LLC," November 2013. http://dodd-frank.com/wp-content/uploads/2013/12/Shak-SHK-banging-the-close.pdf Accessed: 2018-08-09.

_ , "CFTC v. Christophe Rivoire," December 2019. https://www.cftc.gov/media/3266/enfrivoirecomplaint122019/download Accessed: 2020-03-26.

The United States Department of Justice, "Former Head of HSBC's Global Foreign

Exchange Cash-Trading Sentenced to Prison for Multimillion-Dollar Front-Running Scheme," April 2018. https://www.justice.gov/opa/pr/former-head-hsbc-s-global-foreign-exchange-cash-trading-sentenced-prison-multimillion-dollar Accessed: 2018-08-08.

_ , "HSBC Holdings Plc Agrees to Pay More Than $100 Million to Resolve Fraud Charges," January 2018. https://www.justice.gov/opa/pr/hsbc-holdings-plc-agrees-pay-more-100-million-resolve-fraud-charges Accessed: 2018-04-25.

The United States Securities and Exchange Commission, "SEC Order: Athena Capital Research, LLC, Release No. 34-73369," October 16 2014. https://www.sec.gov/litigation/admin/2014/34-73369.pdf Accessed: 2018-11-18.

The Wall Street Journal, "Big Fine in Metals Case," July 2011. https://www.wsj.com/articles/SB10001424053111904772304576468530892198942 Accessed: 2018-08-06.

_ , "Deutsche Bank to Pay $205 Million Fine to End N.Y. Currency-Trading Probe," June 2018. https://www.wsj.com/articles/deutsche-bank-to-pay-205-million-fine-to-end-n-y-currency-trading-probe-1529517428 Accessed: 2018-08-06.

Zhang, Anthony Lee, "Competition and Manipulation in Derivative Contract Markets," *Working Paper*, 2020. https://anthonyleezhang.github.io/pdfs/cmdcm.pdf.