Contracting for Financial Execution*

Markus Baldauf† Christoph Frei‡ Joshua Mollner§

July 20, 2019

Abstract

Financial contracts often specify reference prices whose values are undetermined at the time of contracting, which makes them prone to manipulation. 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. Under certain conditions, a simple contract based on the market volume-weighted average price (VWAP) emerges as the unique optimal solution to this principal-agent problem. This result explains the prevalence of guaranteed VWAP contracts in practice and also suggests considerations for the optimal design of financial benchmarks.

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

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

*We thank Robert Almgren, Hendrik Bessembinder, Piotr Dworczak, Lawrence Harris (discussant), Burton Hollifield (discussant), Pete Kyle, Kasper Larsen, Esen Onur, Andreas Park (discussant), Duane Seppi, Yang Song (discussant), Chester Spatt, and Nicholas Westray 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.

†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/.

Electronic copy available at: https://ssrn.com/abstract=3177283
1 Introduction

Many financial contracts reference benchmark prices whose values are yet to be determined at the time of contracting. For example, a client might agree, in advance, to trade one billion US dollars with her broker-dealer (henceforth, simply “dealer”) in exchange for British pounds at the ‘fix’, a popular end of day benchmark for foreign currency transactions. But because these benchmarks are typically endogenous, the dealer may trade in a way that influences them to his advantage, but to the detriment of the client. Financial markets abound with examples of dealers taking advantage of their clients in this way. Manipulation of the fix has recently led HSBC to agree to pay more than 100 million dollars (as part of a deferred prosecution agreement) and to the imprisonment of the firm’s head of global foreign exchange trading (DOJ, 2018a,b). A number of other examples involve similar manipulation of foreign exchange rates (Bloomberg, 2013; WSJ, 2018), and these publicized cases are likely only the tip of the iceberg. Although this manipulative behavior is often ruled to be illegal, it appears to continue to occur, likely because monitoring is difficult and because of legal gray areas between acceptable and prohibited forms of trading, as between front-running and pre-trade hedging (FINRA, 2013).

Of course, a predetermined, fixed price cannot 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. Moreover, regulatory requirements that dealers hold capital in amounts corresponding to their exposure might render them effectively risk averse.

In this paper, we formulate a principal-agent problem so as to derive a contract that
a client might offer to a dealer that would perform well in spite of both aforementioned frictions: (i) the possibility of hard-to-detect benchmark manipulation, and (ii) risk aversion on the part of the dealer. The aforementioned examples come from foreign exchange, and indeed, we view that market as a leading application for our analysis. However, both the problem of how a client should contract with her dealer and the model that we introduce are general, and our analysis is not restricted to any single asset class.

In the model, the client offers a contract to the dealer at time $0$. The contract specifies that the client and dealer will conduct an over-the-counter trade at time $T$, in which the client will purchase a quantity normalized to one share from the dealer at a price that may depend on the history of market volumes and prices. If the dealer accepts the contract, then he purchases the share on the market and may divide his trades across the intervening trading periods. Trading creates price impact, the severity of which is influenced by market conditions (e.g., the participation of other traders). And because market conditions may differ across periods, the expected cost of trading depends on the dealer’s division of his trades across time.

The client cannot directly observe either market conditions or the dealer’s trades. However, she can observe two outcomes that are influenced by them: market prices and volumes. By conditioning on those prices and volumes in the contract that she offers, the client can therefore shape the dealer’s incentives about how to trade. The client’s problem is then to devise a contract that will minimize her payment to the dealer, accounting for its effect on the dealer’s incentives.

Under certain conditions, the optimal contract is unique and corresponds to what is known in practice as guaranteed VWAP: the agreement that, at a predetermined future time, the dealer and client will trade at the volume-weighted average price (VWAP) prevailing in the market over the intervening time interval. Those conditions include, for example,
the assumptions that price impact is purely temporary and that the dealer’s knowledge of market conditions is sufficient to perfectly forecast volumes. Our initial description of those conditions (in Section 3.2) 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 3.3) a micro-foundation for one particular member of this class.

The first-best trading strategy under these conditions 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.

We allow for a large class of contracts, and it is instructive to consider a few familiar cases in more detail. First, consider a guaranteed market-on-close contract, where the client pays the dealer at the price prevailing at period $T$. This resembles the above foreign exchange example involving HSBC. In the model, this contract is not optimal because the dealer has an incentive to deviate from the first-best trading policy by tilting his trades toward the last period. Because his trades have price impact, this will move the price in period $T$, thereby creating a gap between this price and the average price that he paid to acquire the position. This behavior corresponds to what is known in practice as ‘banging the close.’ Second, consider a contract in which the client pays the dealer at a predetermined, fixed price. Because it is fixed, this price cannot be manipulated by the dealer. However, as alluded to in the second paragraph, risk aversion (or minimum capital requirements) of the dealer imply that he would need to be compensated for assuming the price risk of the position.

In contrast, we establish in Section 4 that, under our assumptions, the optimal contract references the VWAP benchmark. Unlike fixed price contracts, the VWAP contract insures the dealer against price fluctuations, provided that he pursues the first-best trading policy.
And unlike many other contracts that are tied to future prices, the VWAP contract does not incentivize the dealer to deviate from the first-best policy. This result may help to explain the popularity of such contracts in practice.¹

We then turn to a discussion of the extent to which these results can be generalized. The contracts considered in the previous paragraphs are examples of ‘principal trading’ arrangements, in which the dealer is the counterparty of the agent. Such arrangements are our primary focus, and our main result is that the VWAP contract is uniquely optimal in this class. However, in Section 5.1, we also consider ‘agency trading’ arrangements, in which the dealer trades on the client’s behalf, acting as a matchmaker between the client and a third party. 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 5.2 considers an extension of the model in which the timing is adjusted so that the dealer can make trading decisions in a more flexible way. Under certain additional assumptions on other aspects of the model, the VWAP contract remains optimal. Next, Section 5.3 investigates the implications of relaxing the assumptions made in Section 3.2: 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 conditions hold in approximation, then the main result extends in approximation: a VWAP-based contract will be nearly optimal. Nevertheless, it would be interesting, in future research, to investigate optimal contracting in settings where these conditions are violated more substantially.

We also apply our results to the question of benchmark design. This is particularly

¹For example, Nomura (2014) provides some evidence of the prevalence of these contracts in the domain of equity trading.
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, and that benchmark might be observed and contracted on. Our results suggest that it may be desirable to compute this benchmark as the VWAP, since, in that case, it becomes possible for the client to propose the optimal contract by referring to the benchmark. As we describe in the text, additional applications might include how to compute the settlement price of various futures contracts, and the net asset values of various funds.

2 Related literature

Empirics. In addition to the publicized cases mentioned in the introduction, a number of academic studies have demonstrated empirically that the trading decisions of broker-dealers are in fact sometimes distorted. Closely related to our analysis, Henderson, Pearson and Wang (2019) find evidence of distortions similar to those that we model in the trading conducted by issuers of structured equity products. Evidence of related distortions is also found by Barbon, Di Maggio, Franzoni and Landier (2017). Although the latter do not find brokers acting directly on information about an impending client transaction, they provide evidence of their acting indirectly by leaking this information to their other clients. In a similar spirit, others have found evidence suggesting 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, 2018), or to use their own alternative trading systems (Anand, Samadi, Sokobin and Venkataraman, 2019).

Also related, a number of studies have found evidence of trade-based benchmark ma-
nipulation. An important benchmark is the closing price. Hillion and Suominen (2004) find evidence of brokers manipulating closing prices so as to enhance their reputation for execution quality. Of course, other parties beyond brokers may have their own reasons for manipulating closing prices. In line with this, a number of papers have also found evidence of such manipulation (Harris, 1989; Felixson and Pelli, 1999; Carhart, Kaniel, Musto and Reed, 2002; Ben-David, Franzoni, Landier and Moussawi, 2013; Comerton-Forde and Putniņš, 2011, 2014). And, 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. Given its formula, it can be manipulated by options trades, which could in turn create opportunities for the manipulator to profitably trade securities that are tied to the VIX. Indeed, Griffin and Shams (2017) have found evidence of trading patterns in out-of-the-money options that are consistent with such manipulation.

**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, taking on a net position. In contrast, we analyze a setting in which conflict may lead to suboptimal execution, 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. For similar reasons, benchmark choice plays a major role in many other aspects of financial markets. Interest rate benchmarks constitute 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 quality of fund managers constitute another example. Fund managers are incentivized 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 (Goetzmann, Ingersoll, Spiegel and Welch, 2007). 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.

Since the contract that emerges from our framework as optimal, the VWAP contract, is a fairly simple one, this paper relates to a literature on foundations for contracts possessing simple features such as linearity. Holmström and Milgrom (1987) show that when an agent repeatedly performs the same task, the principal optimally provides the same incentives at each point in time, which makes the agent's payment a linear function of output. Carroll (2015) shows that linear contracts are optimal if the principal is uncertain about the actions available to the agent and treats this uncertainty with a worst-case criterion.

While our focus is on the contractual relationship between a client and her dealer, there is also a connection to the literature on optimal trading strategies. In our model, the solution to the first-best benchmark amounts to a volume participation strategy. Similarly, Kato (2015) provides conditions under which such a participation strategy is optimal. 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 trading decisions. One might think that this additional friction would preclude achievement of the first-best benchmark. To the contrary, we show that first-best can be achieved in the model if the dealer is incentivized with an appropriate contract. Our main contribution is to identify the guaranteed VWAP contract as the unique contract to do this.

3 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.\(^2\) If the dealer accepts the contract offered to him, then he purchases from the market the shares that he will subsequently sell to the client. Importantly, the price 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 how the dealer trades while he is acquiring the shares from the market.

In Section 3.1, we formulate a fairly general model of this principal-agent relationship. Though it is difficult to characterize the optimal contract in general, progress can be made in certain specifications of the model. In Section 3.2, we restrict attention to an important class of settings in which, as we later show, a clean solution can be derived. And in Section 3.3, we provide a micro-foundation for one particular element of that class.

\(^2\)Symmetric analysis applies to the case in which the client’s need is to sell.
3.1 General framework

Trading. There are finitely many discrete trading periods \( t \in \{1, 2, \ldots, T\} \).\(^3\) The client contracts to purchase the share from the dealer after time \( T \). In advance of this transaction, the dealer must purchase the required share in the market. We also require the dealer to purchase no more than the required share, so that he ends with an inventory of zero. Letting \( x_t \) denote the number of shares purchased by the dealer in trading period \( t \), we therefore require \( \sum_{t=1}^{T} x_t = 1 \). We also require that all \( x_t \) be nonnegative, so that the dealer does not sell in any period.\(^5\) We refer to a vector \( \mathbf{x} = (x_t)_{t=1}^{T} \) that satisfies both of these conditions as a trading schedule.

Market conditions. Time-varying market conditions both affect total trading volume and determine the price impact of the dealer’s trades. The market condition in period \( t \), denoted \( \eta_t \), is an exogenous random variable that takes positive values. The realization of \( \eta = (\eta_t)_{t=1}^{T} \) is learned by the dealer prior to period \( t = 1 \) (hence, \( \eta \) should be interpreted as what the dealer knows about the market) but the client knows only the distribution. We do not impose any assumptions on the joint distribution of \( \eta \). In particular, its elements do not need to be independent. Nor do they need to be identically distributed; thus, our approach permits the client to have some information about market conditions, albeit less than the dealer.

Our approach is to treat \( \eta_t \) as an abstract object that affects price and volume in ways

---

\(^3\)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).

\(^4\)Alternatively to this discrete-time model, we could consider a continuous-time model with trading during an interval \([0, T]\); see the remark at the end of Section 5.2.1.

\(^5\)This restriction is only to simplify the presentation; with suitable adjustments to the model (e.g., to distinguish between signed and unsigned quantities), the same results would hold even without the restriction. In particular, the optimal contract will lead the dealer to optimally choose nonnegative \( x_t \) in all periods, even when he is not restricted to doing so.
specified below. Nevertheless, one concrete example is that in which \( \eta_t \) represents the unsigned volume traded on the market by outside traders in period \( t \). Another example will be pursued in Section 3.3, where \( \eta_t \) denotes the number of outside traders.

**Prices and volumes.** Let \( p = (p_t)_{t=1}^T \) denote the sequence of prices, and let \( v = (v_t)_{t=1}^T \) denote the sequence of volumes. In general, both of these will be linked to \( x \) and \( \eta \), although that dependence will often be suppressed in the notation. Section 3.2 imposes some specific assumptions on that dependence under which the model can be cleanly solved.

**Contracts.** In specifying the set of feasible contracts, we imagine that the client can observe the sequence of prices \( p \) and the sequence of total volumes \( v \) at time \( T \). We allow the client to offer contracts that are arbitrary functions of these market outcomes.\(^6\) 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 \( \tau(p, v) \) to the dealer at time \( T \) in exchange for one share of the security.

This formulation permits the client to propose a wide variety of trading arrangements including many familiar ones. Many contracts observed in practice specify that the client pay the dealer according to a particular benchmark price. Common benchmarks include

1. the closing price, which corresponds to \( \tau = p_T \),
2. a predetermined fixed price, which corresponds to \( \tau = \tau_0 \) for some constant \( \tau_0 \),
3. the time-weighted average price (TWAP), which corresponds to \( \tau = \frac{1}{T} \sum_{t=1}^T p_t \), and
4. the VWAP, which corresponds to \( \tau = \frac{\sum_{t=1}^T p_t v_t}{\sum_{s=1}^T v_s} \).

Our approach focuses on what are known in practice as principal trading arrangements,
wherein the dealer acts as the client’s counterparty. In contrast, agency trading arrangements are those in which the dealer merely acts as a matchmaker in locating a counterparty. Agency trading can be thought of within our framework as a contract specifying that the client reimburses the trading costs $p \cdot x$ in exchange for the share. The set of feasible contracts described above does not allow for this payment rule. However, in Section 5.1, we discuss how our results would be affected if such contracts were available.

**Timing.** The timing of events is as follows. Prior to trading, the client offers a contract $\tau$ 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 the contract, then he learns $\eta$ and chooses a trading schedule $x$. Given $\eta$ and $x$, prices $p$ and volumes $v$ are realized. After trading, the client pays the dealer as specified by $\tau$.

Note that the choice of $x_t$ can depend only on $\eta$ and not, for example, on unobserved shocks that might also influence $p_t$. Thus, the dealer should be thought of as conducting his trades using market orders. Nor may $x_t$ depend upon $(p_s)_{s=1}^{t-1}$. In the interpretation of the model in which $t$ indexes venues (cf. footnote 3), this restriction would be the natural assumption. But for the interpretation in which $t$ indexes time, this restriction may artificially constrain the actions available to the dealer. However, in Section 5.2, we consider a version of the model in which the dealer can modify the trading schedule dynamically, so that $x_t$ can depend not only on $\eta$ but also on information observed in previous periods.

**The dealer’s payoffs.** Since the dealer may schedule trades in a way that depends on $\eta$, he can be thought of as choosing a trading policy, which is a measurable function $x(\cdot)$ that maps each $\eta$ into a trading schedule. The dealer’s utility function over money is some

---

8The results that we subsequently derive would remain unchanged if the dealer were to learn $\eta$ before deciding whether to accept or reject the contract.
function \( u \), which is assumed to be strictly increasing and weakly concave. From accepting a contract \( \tau \) and choosing a trading policy \( x(\cdot) \), the dealer receives expected utility

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

**The client’s payoffs.** We assume that if the dealer rejects the offered contract, then the client receives a payoff of negative infinity.\(^9\) 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 \( \tau \) and chooses a trading policy \( x(\cdot) \), then the client receives expected utility

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

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 generates a significant amount of temporary price impact and because of our assumption that the client is risk neutral. And ‘unsophisticated’ because the client chooses to access the market through an external dealer rather than one in house and because the client has no information about fundamentals. As a concrete example, consider Cairn Energy Plc, a British oil explorer that had contracted with HSBC in the episode alluded to in the introduction. In connection with selling part of its ownership interest in a subsidiary company, it realized a one-time need to convert $3.5 billion USD into British pounds.

\(^9\)Assuming that the client receives infinite disutility from rejection is simply to ensure 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 quantity that exceeds her expected payment to the dealer under the optimal contract.
3.2 Additional conditions

The general model laid out above does not lend itself to a solution unless additional structure is imposed. One special case is that in which the dealer is risk neutral. In that case, and by standard arguments, an appropriately-specified fixed price (i.e., ‘sell-the-firm’) contract would be optimal. In that sense, our model is consistent with the usage of such arrangements in practice.

More interestingly, however, our model can also explain the prevalence of guaranteed VWAP contracts: those contracts attain optimality under the additional conditions we state below. The primary substance of these conditions is to require (i) that price impact is temporary and (ii) that volumes can be perfectly forecast given knowledge of market conditions.\(^{10}\) These assumptions are strong, but they allow us to obtain this sharp result in spite of making relatively weak assumptions about other aspects of the model.

**Price.** 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; \(\varepsilon = (\varepsilon_t)_{t=1}^T\) are random variables that all have the same expectation conditional on \(\eta\), namely, \(\mathbb{E}[\varepsilon_t | \eta] = \mu\) almost surely for all \(t\) and some constant \(\mu\). However, we impose no further assumptions on the elements of \(\varepsilon\): they do not need to be independent of each other or \(\eta\), and neither do they need to be identically distributed. Note that \(\varepsilon\) 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.

\(^{10}\)In Section 5.3.1, we revisit these two assumptions, showing a continuity result: if the assumptions hold in approximation, then our main result also holds in approximation.
For a given number of shares $x_t$, the impact on price is influenced by the market conditions prevailing in the period. Therefore, the price impact depends on $x_t$ measured relative to $\eta_t$, and not on $x_t$ itself. In light of this, the market condition $\eta_t$ can be thought of as parametrizing how steep price impact will be in period $t$.\textsuperscript{11} 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).\textsuperscript{12}

\textit{Remark.} Note also that the price depends only on contemporaneous values of $x_t$ and $\eta_t$. In that sense, our baseline model should be interpreted as one of temporary price impact. This is not inconsistent with the literature, where both empirically and theoretically, different forms of price impact have been analyzed without clear conclusions as to its functional form (or how that form might vary with the setting). Nevertheless, see Section 5.3.2 for a discussion of the issues that arise when price impact has a permanent component.

\textbf{Volume.} We assume the total unsigned 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 $\eta_t$ 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 total volume $v(x, \eta)$ is increasing in $\eta$. This is natural: in light of the fact that $\eta_t$ can be thought of as parametrizing the steepness of price impact

\textsuperscript{11}In the case where $\eta_t$ represents outside volume, using precisely the ratio $x_t/\eta_t$ makes the price impact dimensionless (Almgren, Thum, Hauptmann and Li, 2005).

\textsuperscript{12}Indeed, linear price impact is consistent with the price impact models based on adverse selection by Kyle (1985) and Kyle, Obizhaeva and Wang (2018). 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).
in period $t$, we obtain the intuitive relationship that, holding fixed the dealer’s volume, price impact is smaller when volume is larger. The assumption also implies that total 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 $\eta$ correspond to an increase in scale of the market: both the dealer’s volume and total 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 1.** 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 $\eta$ for $(x, \eta) \in \text{dom}(v)$ with $x \neq 0$.

**Remark.** Whereas we have modeled prices as stochastic, our baseline formulation assumes that volumes are a deterministic function of the dealer’s trading schedule and market conditions.\textsuperscript{13} This is, of course, not completely realistic: even the most knowledgeable traders are unable to perfectly forecast volumes. However, this formulation is consistent with the existence of known empirical regularities in trading volume (e.g., the so-called ‘liquidity smile’), whereas there are fewer such patterns in prices (and in fact no such patterns under

---

\textsuperscript{13}Our later arguments go through if the price impact and total volume are of the forms $h\left(\frac{x}{\sum \eta_i}\right)$ and $v(x_t, \eta_t) = x_t V\left(\frac{x_t}{\Xi \sum \eta_t}\right)$, respectively, for a random variable $\Xi$ that is independent from $\varepsilon$. Under these forms of price impact and total volume, the true market conditions would be $\left(\Xi \eta_t\right)_{t=1}^T$. The dealer observes $(\eta_t)_{t=1}^T$, which therefore provides only a noisy signal of market conditions, yet does nevertheless fully reveal the relative market conditions $\left(\frac{\Xi \eta_t}{\sum \eta_t}\right)_{t=1}^T$.  

Electronic copy available at: https://ssrn.com/abstract=3177283
the weak-form efficient market hypothesis). Nevertheless, see Section 5.3.3 for a discussion of the issues that arise when the volume profile is only imperfectly predictable.

3.3 Micro-foundation

In the baseline model presented in Sections 3.1 and 3.2, prices and volumes in a period $t$ are jointly influenced by both the dealer’s trades $x_t$ and market conditions $\eta_t$. 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 forms.

Suppose that the dealer trades using market orders $x_t$, as before. 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 $\eta_t$, 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, as we demonstrate below, $p_t$ and $v_t$ will depend on $x_t$ and $\eta_t$ in a way that is nested by our previous analysis.

The client and the dealer are precisely as before. What is different is that we more precisely specify the other traders in the market and the nature of trading.

Outside traders. In each period $t$, a positive measure $\eta_t$ of outside traders arrive, who then depart the market after the period has ended. Consistent with our previous notation, the realization of $\eta$ is learned prior to the start of trading by the dealer, but it is not learned by the client, and we do not impose any assumptions on the joint distribution of its elements.

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 $\theta_i = \psi_i + \varepsilon_i.$\footnote{There might be many ways to micro-found this 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 $v_i(y, p_t) = \mathbb{E}[\exp(-[V(y - \theta_i) - p_t y])] = -\exp(p_t y + (y - \theta_i)^2/2)$. Since there is 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 given in the text.} The idiosyncratic component $\psi_i$ is an independent draw from a standard normal distribution, and the common components $(\varepsilon_i)_{t=1}^T$ are random variables that all have the same expectation conditional on $(\eta_t)_{t=1}^T$, which is also consistent with our previous assumption. The security is in zero net supply, and the price $p_t$ is chosen to clear the market. All trades take place at that price.

**Solution.** In period $t$, $\eta_t$ outside traders are active. Substituting for $\theta_i = \psi_i + \varepsilon_i$ 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 total volume traded) will be

$$v(x_t, \eta_t) = x_t + \eta_t \int_{\frac{x_t}{\eta_t}}^{\infty} \left( z - \frac{x_t}{\eta_t} \right) \phi(z) \, dz$$

$$= x_t \Phi \left( \frac{x_t}{\eta_t} \right) + \eta_t \phi \left( \frac{x_t}{\eta_t} \right).$$

This function is positive since $\eta_t > 0$ and of the form $v(x_t, \eta_t) = x_t V \left( \frac{X}{\eta_t} \right)$ 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

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

### 4 Main results

Our first main result, Theorem 4, states that the guaranteed VWAP contract is optimal. Our second main result, Theorem 5, states a sense in which the VWAP form is necessary for optimality under certain conditions. Before coming to these main results, we characterize the first-best trading policy in Lemma 2. Proofs are relegated to the Appendix.
4.1 The first-best trading policy

We begin by defining and characterizing the first-best trading policy. Given that the client is risk neutral, a trading policy $x(\cdot)$ is first best if, for all $\eta$,

$$x(\eta) \in \arg \min_x \mathbb{E}[p \cdot x|\eta].$$

Thus, a first-best trading policy minimizes expected trading cost conditional on all realizations of the market conditions $\eta$. Consequently, such a policy also minimizes the (unconditional) expected trading cost $\mathbb{E}[p \cdot x]$. This is also how the client would trade herself if she were to possess the dealer’s knowledge of market conditions.

Given the structure of the model, there is a unique first-best trading policy, which Lemma 2 characterizes as the policy under which the dealer’s trades $x_t$ are proportional to market conditions $\eta_t$. This trading policy is efficient because it 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 total volume profile of the market.

**Lemma 2.** The first-best trading policy is

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

The expected trading cost incurred by this policy is $\mathbb{E}[p \cdot x^{FB}(\eta)] = \mu + \mathbb{E} \left[ h \left( \frac{1}{\sum_{l=1}^{T} \eta_l} \right) \right].$

---

15We remind the reader that the dependence of $p$ on $x$, $\eta$, and $\varepsilon$ is suppressed in the notation. Similarly, $v$ depends on $x$ and $\eta$, and we will make that dependence explicit with the notation $v(x, \eta) = (v(x_t, \eta_t))_{t=1}^{T}$ where it helps to clarify the dependence structure.
Moreover, the following equality holds

\[ x^{FB}(\eta) = \left( \frac{v(x^{FB}_t(\eta), \eta_t)}{\sum_{s=1}^{T} v(x^{FB}_s(\eta), \eta_s)} \right)^T. \]

The optimality of volume participation strategies in the 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 trading cost incurred by this policy, \( p \cdot x^{FB}(\eta) \), will always be equal to the market VWAP.

### 4.2 The client’s problem

Having derived the first-best trading policy, we now turn our attention to the second best. The client chooses a contract \( \tau \), as well as a ‘recommended trading policy’ \( x(\cdot) \) to maximize her expected utility (equivalently, to minimize her expected payment) subject to individual rationality and incentive compatibility constraints of the dealer:

\[
\min_{\tau, x(\cdot)} \mathbb{E}\left[ \tau(\mathbf{p}, v(x(\eta), \eta)) \right] \quad \text{subject to} \\
\mathbb{E}\left[ u(\tau(\mathbf{p}, v(x(\eta), \eta)) - p \cdot x(\eta)) \right] \geq u(0) \quad \text{(IR)} \\
\forall x(\cdot) : \mathbb{E}\left[ u(\tau(\mathbf{p}, v(\hat{x}(\eta), \eta)) - p \cdot \hat{x}(\eta)) \right] \geq \mathbb{E}\left[ u(\tau(\mathbf{p}, v(x(\eta), \eta)) - p \cdot x(\eta)) \right] \quad \text{(IC)}
\]

As is the standard approach in contract theory, we presume that the principal can choose the agent’s action so long as (IR) and (IC) are satisfied. In particular, if the dealer is indifferent among several policies—that is, if (IC) holds with equality for a certain alternative trading policy—then this approach presumes that the dealer will resolve his indifference in favor of the policy recommended by the client.
A contract $\tau$ is *optimal* if it is part of a solution to this problem. Following from Lemma 2, we can derive useful intermediate results about properties of optimal contracts, which are stated in Lemma 3.

**Lemma 3.** The following conditions are together sufficient for a contract $\tau$ to be optimal:

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

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

(ii) $\tau(p, v(x^{FB}(\eta), \eta)) = p \cdot x^{FB}(\eta)$ 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 result, notice that condition (i) is equivalent to $(\tau, x^{FB}(\cdot))$ satisfying (IC). And condition (ii) implies that under $(\tau, x^{FB}(\cdot))$, the dealer is fully insured and (IR) is satisfied with equality. Thus, $(\tau, x^{FB}(\cdot))$ implements the efficient outcome—both the efficient trading policy and efficient risk sharing—and leaves the dealer with zero surplus. 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.
4.3 Optimality of VWAP

Building on Lemma 3, we now proceed to state our main results, both of which pertain to the guaranteed VWAP contract

\[ \tau^{VWAP}(p, v) = \frac{\sum_{t=1}^{T} p_t v_t}{\sum_{s=1}^{T} v_s} \]

The following two theorems concern the optimality of this contract. Theorem 4 states that the VWAP contract is optimal. Theorem 5 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.

**Theorem 4.** The contract \( \tau^{VWAP} \) is optimal.

**Theorem 5.** If \( u \) is strictly concave and the distributions of \( \varepsilon \) and \( v(x^{FB}(\eta), \eta) \) have full support over \( \mathbb{R}^T \) and \( \mathbb{R}^T_+ \), respectively, then a contract \( \tau \) is optimal only if \( \tau = \tau^{VWAP} \) almost everywhere on its domain.

To prove Theorem 4, we establish that \( \tau^{VWAP} \) satisfies the conditions of Lemma 3. The result is then immediately implied by that lemma. To see that condition (ii) of Lemma 3 is satisfied, recall that, by the last part of Lemma 2, the first-best trading policy corresponds to a volume participation strategy. As a result, the costs incurred by the policy, \( p \cdot x^{FB}(\eta) \), are always equal to the market VWAP, and thus always equal to the payment \( \tau^{VWAP}(p, v(x^{FB}(\eta), \eta)) \). The meat of the argument lies in establishing that condition (i) of Lemma 3 is satisfied.

To see why condition (i) of the lemma is satisfied, suppose that a dealer who is compensated according to \( \tau^{VWAP} \) considers deviating from \( x^{FB}(\cdot) \) to trade \( \delta > 0 \) fewer shares at time \( t \) and \( \delta \) more shares at time \( t' \). By the definition of \( x^{FB}(\cdot) \), this will raise the dealer’s expected
costs for acquiring the position from the market, $\mathbb{E}[p \cdot x(\eta)]$. 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, $\mathbb{E} \left[ \frac{1}{\sum_{s=1}^{T} v(x_s(\eta), \eta_s)} p \cdot v(x(\eta), \eta) \right]$. 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 $\delta$. But what matters for the change in revenue is the fraction of total volume that is shifted. This is a smaller fraction than $\delta$ for the reason 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 3 holds. That condition requires a contract to provide weak incentives for the dealer to pursue the first-best policy. But the VWAP contract in fact provides strict incentives for doing so. Mathematically, we show in the proof of Theorem 4 that for all trading policies $\hat{x}(\cdot)$ not equal to $x^{FB}(\cdot)$ almost surely, (IC) holds with strict inequality:

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

This is true not only when the dealer is risk averse but also when he is risk neutral.

For proving the uniqueness result of Theorem 5, we build on the conclusion that $\tau^{VWAP}$ satisfies the conditions of Lemma 3 to deduce that all optimal contracts must satisfy those same conditions. These conditions are demanding and severely restrict the possibilities for $\tau$. With the full support assumptions, they in fact pin down $\tau$ to equal $\tau^{VWAP}$ almost everywhere.\(^{16}\)

\(^{16}\)The full-support assumption could be abandoned if the uniqueness statement were weakened to $\tau(p, v(x^{FB}(\eta), \eta)) = \tau^{VWAP}(p, v(x^{FB}(\eta), \eta))$ almost surely, where $p$ are the prices corresponding to the first-best trading policy $x^{FB}(\eta)$.
Some intuition for the uniqueness result in Theorem 5 can be gleaned by studying why contracts tied to other common benchmark prices fail to be optimal in our model. First, consider a contract benchmarked to the closing price: $\tau = p_T$. This contract, which corresponds to what is known in practice as a guaranteed market-on-close contract, 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.’ For example, a risk-neutral dealer would trade so as to maximize the expected gap between the closing price and his average price paid. At the other extreme, an infinitely risk-averse dealer would only trade in the last period and thereby completely ignore his knowledge of market conditions. This failure to induce the first-best trading policy destroys optimality. Next, consider a contract benchmarked to the TWAP: $\tau = \frac{1}{T} \sum_{t=1}^{T} p_t$. This contract also incentivizes the dealer to deviate from the first-best trading policy—in this case by smoothing trading across time periods more than the first-best policy prescribes—and so it similarly fails to be optimal. Finally, consider fixed price contracts. By paying the dealer a pre-determined amount, 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, this constitutes inefficient risk sharing, which destroys optimality.

The previous 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 $\tau$ in which the contract could be altered without affecting optimality.

We would stress that, beyond its optimality in our model, an additional virtue of the

\footnote{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.}
guaranteed 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 $\eta$ and $\varepsilon$, 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 guaranteed VWAP contract might be expected to hold only in approximation. And in particular, it might be necessary to add a small commission in order to ensure that the (IR) constraint is satisfied. Consistent with this, such ‘VWAP plus commission’ contracts are commonly observed in practice.

5 Extensions

We now turn to the question of when our main results generalize beyond our baseline model. We introduced two conditions in Section 3.2: temporary price impact and forecastable volumes. We first show that if those conditions are maintained, then certain other aspects of the model can be modified without affecting our main result on the optimality of the VWAP contract. In Section 5.1, we enrich the set of contracts available to the client so as to include agency trading in addition to principal trading. The VWAP contract remains optimal; however, it is no longer uniquely so: the agency contract is also optimal within the model when it is available. In Section 5.2, we enrich the set of deviations available to the dealer by allowing him to condition his trading decisions on the realizations of previous prices. Under additional assumptions on the volume function $v(x, \eta)$ and the dependence structure of price shocks $\varepsilon$, we show that the VWAP contract again remains optimal.
Then in Section 5.3, we investigate the implications of relaxing the two aforementioned conditions of Section 3.2 themselves (viz. temporary price impact and forecastable volumes). We begin by presenting a continuity result: if those conditions hold in approximation, then VWAP-based contracts remain approximately optimal. But under larger violations of the conditions, VWAP contracts may be far from optimal.

5.1 Agency trading

In the baseline model, the feasible contracts were all functions of the sequence of prices and total volumes. This would be the appropriate set of contracts to consider if that were all the client could observe and contract on at time $T$.

However, agency trading constitutes another commonly-observed set of trading arrangements in practice. Under such arrangements, the client reimburses the dealer for all trading costs that he incurs. In the language of our model, this corresponds to reimbursement in the amount of $\tau^A = p \cdot x$. This contract is not included in the set of feasible contracts considered in the baseline model, for the reason that $x$ is typically not observed by the client in practice. However, in many asset classes, there exist regulatory bodies who can observe $x$ and do enforce contracts of the form $\tau^A$ (though typically they do not also enforce contracts based on arbitrary functions involving $x$).

In such asset classes, it may therefore be natural to extend the class of feasible contracts to include $\tau^A$. Because the VWAP contract $\tau^{VWAP}$ satisfies the conditions of Lemma 3, it

---

18For principal trading arrangements, which are the main focus of our paper, the broker-dealer acts as a dealer, which is why we have referred to him as a “dealer” throughout. But for agency trading arrangements, the broker-dealer acts as a broker. In spite of that, we continue to refer to him a “dealer” throughout this section to maintain consistent terminology.

19In the case of equities, many regulations detail a broker’s obligations for agency trades. For example, for NMS securities, SEC Rule 34-43590 stipulates certain rights of customers that pertain to information about order routing by their broker. Moreover, the existence of a consolidated tape as well as a consolidated audit trail enables regulatory bodies to enforce these regulations. But in the case of foreign exchange trading, no such regulation is currently in place.
remains optimal regardless of which other contracts are feasible. In other words, Theorem 4 extends to the version of the model in which the agency contract is available. However, Theorem 5, which states conditions under which the VWAP contract is uniquely optimal, does not extend to this version of the model. Indeed, the agency contract also solves the client’s problem described in Section 4.2, with \( x^{FB}(\cdot) \) as the corresponding recommended trading policy. Consequently, we have the following result, which is an immediate corollary of Lemma 3.

**Corollary 6.** If the agency contract \( \tau^A \) is feasible, then it is optimal.

Thus, two contracts perform particularly well in the model: the VWAP contract and the agency contract. However, the model highlights one potentially important difference between these two optima. On one hand, the agency contract makes the dealer indifferent between all trading policies, and its optimality relies upon this indifference being broken in favor of the first-best policy.\(^{20}\) On the other hand, and as observed in Section 4.3, the VWAP contract provides the dealer with a strict incentive to pursue the first-best policy, which makes 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, if a risk-averse dealer might have imperfect knowledge of \( \eta \), then the VWAP contract would require the dealer to bear risk. As a result it 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, the agency contract completely insulates the

\(^{20}\)Thus, the agency contract fails to be robust to many perturbations of the model. 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 the first-best policy would not minimize these effort costs. In this perturbation, the first-best policy would then no longer be incentive compatible under the agency contract.
dealer from risk, even under imperfect knowledge of $\eta$. And mainly for that reason, the agency contract would continue to be optimal.

As another example of how these two contracts differ in robustness to unmodeled elements, suppose there are many dealers who have the same level of risk aversion but 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 $\eta$, as in the baseline model, and (ii) low-skill dealers, who have imperfect knowledge. Then the VWAP contract has the added advantage that it may serve as a screening device: high-skill dealers would accept it, while low-skill dealers would not. The agency contract, on the other hand, would be equally acceptable to all types of dealers, which could prove expensive to the client if accepted by a low-skill dealer.\footnote{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).}

\section*{5.2 Dynamic trading policies}

In the baseline model, the trading policies available to the dealer were functions that mapped market conditions into trading schedules. In this section, we permit a more general class of trading policies in which the dealer may also condition his trading decisions on previous prices. For this section (Section 5.2) only, a trading policy is a measurable vector $(x_t(\cdot))_{t=1}^T$, such that (i) for all $t$, $x_t$ depends on $\eta$ and $(p_s)_{s=1}^{t-1}$, (ii) for all $t$, $x_t(\eta, (p_s)_{s=1}^{t-1}) \geq 0$ almost surely, and (iii) $\sum_{t=1}^T x_t(\eta, (p_s)_{s=1}^{t-1}) = 1$ almost surely.

\subsection*{5.2.1 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 5.2.1) only,
we assume that $v(x, \eta) = x + \eta$. In Section 5.2.2, 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.

For our previous results to extend, we must also impose an additional restriction on the dependence structure of $\varepsilon$. For this section (Section 5.2.1) only, we make the mild assumption that $\varepsilon$ satisfies

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

for all $t = 1, \ldots, T - 1$. Condition (1) implies that, given $(\varepsilon_s)_{s=0}^{t-1}$, the price shocks $\varepsilon_{t+1}$ and $\varepsilon_t$ have the same predicted mean. 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 condition (1) implies the assumption of the baseline model: that all $\varepsilon_t$ have the same expectation conditional on $\eta$, which we continue to denote by $\mu$. Note also that given the market conditions $\eta$, 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 $\eta, p_1, \ldots, p_{t-1}$, which represents the information available to the dealer when making 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 $t$.

All results stated in Section 4 extend to this version of the model. Most proofs remain unchanged. The only difference lies in one part of the proof of Theorem 4. Nevertheless, the same result obtains:

**Theorem 4'.** In the alternate version of the model described in this section, the contract $\tau^{VWAP}$ is optimal.

**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 \([1, 2, \ldots, T]\). In such a continuous-time model, a trading policy prescribes a trading rate. Our results continue to hold for continuous-time versions of both our baseline model and its extensions, provided that the setting is suitably adjusted. One subtlety is that for the continuous-time version of the dynamic trading policies extension considered in this section, condition (1) becomes the martingale property for \(\varepsilon\) (with respect to the \(\sigma\)-algebras generated by itself and the entire process \(\eta\)).

By contrast, condition (1) in discrete time is slightly more general than the martingale property because the conditioning is only over \(\eta, \varepsilon_1, \ldots, \varepsilon_{t-1}\) and not \(\eta, \varepsilon_1, \ldots, \varepsilon_{t-1}, \varepsilon_t\). 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.

### 5.2.2 Why VWAP optimality may fail without those conditions

The baseline model allows for a very general class of volume functions \(v(x, \eta)\), but restricts the set of trading policies to functions of \(\eta\). In contrast, the version of the model considered just above (in Section 5.2.1) allows for a broader class of trading policies, allowing the dealer to adjust his strategy based on prices he observes in the course of trading, but restricts to \(v(x, \eta) = x + \eta\). We next explain why the VWAP contract may no longer be optimal if both \(v\) and the set of trading policies are general.\(^{22}\)

Assume that after trading according to \(x^{FB}\) in the first trading period, the dealer observes a \(p_1\) that implies the realization of \(\varepsilon_1\) was much greater than the expected value \(\mu\). If the dealer will be compensated according to a VWAP contract, he may then want to deviate from the first-best trading policy if it is possible that by doing so he can distort the total

\(^{22}\)Our focus in this section is on the necessity of having required the additional assumption \(v(x, \eta) = x + \eta\). We do not discuss the additional assumption that \(\varepsilon\) satisfies (1) for the reason that this additional condition is mild. For example, it is implied by the weak-form efficient market hypothesis. Nevertheless, it is fairly easy to see that this condition is important for our results. For example, if dynamic trading policies are possible, then without this condition, even the first-best such policy might not be as characterized by Lemma 2.
daily 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 $\tau^{\text{VWAP}}$ above the weight placed on $p_1$ in the dealer’s trading costs $p \cdot x$. If $\varepsilon_1$ is sufficiently high, this deviation would be profitable.

Such a deviation is not possible in the baseline model. There, the dealer must commit to a complete trading policy prior to observing any information about the realized prices, and so he cannot condition his trading decisions on $p_1$. Such a deviation is also not possible if $v(x, \eta) = x + \eta$, as in Section 5.2.1. Although trading decisions can depend on $p_1$ in that version of the model, they cannot create the requisite distortion, since the total daily 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 the VWAP contract would no longer be optimal.\footnote{To elaborate, $\tau^{\text{VWAP}}$ would not be optimal because it would be dominated by a ‘VWAP minus commission’ contract $\tau^{\text{VWAP}} - c$ for an appropriately chosen commission $c$. Indeed, if $(\tau^{\text{VWAP}}, x(\cdot))$ satisfies (IC), then so does $(\tau^{\text{VWAP}} - c, x(\cdot))$. Furthermore, the existence of the profitable deviation described in the text means that $(\tau^{\text{VWAP}}, x(\cdot))$ satisfies (IR) with slack. Thus, for an appropriately chosen $c$, $(\tau^{\text{VWAP}} - c, x(\cdot))$ also satisfies (IR) while being cheaper for the client. Note that while this establishes that $\tau^{\text{VWAP}}$ is suboptimal, it does not solve for the optimal contract, which could be an even more substantial deviation from VWAP.}

### 5.3 Permanent price impact and stochastic volumes

The model, as described in Sections 3.1 and 3.2, allows for a great deal of generality in (i) the price impact function $h$, (ii) the volume function $v$, (iii) the dealer utility function $u$, (iv) the distribution of market conditions $\eta$, and (v) the distribution of price shocks $\varepsilon$. Furthermore, the previous subsections describe two modifications of the model, both of which preserve optimality of the VWAP contract. But the generality of that result is limited by two strong assumptions: that price impact is purely temporary and that knowledge of market conditions allows the dealer to perfectly forecast volumes.

In this section, we analyze the contractual 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 baseline model. The setting is as follows: (i) there are only two periods i.e., \( T = 2 \); (ii) total unsigned volume at time \( t \) is given by \( x_t + \eta_t + \delta_t \) for an independent, nonnegative random variable \( \delta_t \); (iii) both the dealer and the client are risk neutral i.e., \( u(w) = w \); and (iv) the price impact function allows for temporary and permanent price impact of the form

\[
p_t = c \sum_{s \leq t} x_s + \frac{x_t}{\eta_t} + \varepsilon_t.
\]

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 \( \delta_1 = \delta_2 = 0 \), the setting is nested by our previous analysis, for the case in which \( h(y) = y \) and \( v(x, \eta) = x + \eta \).

5.3.1 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 \( \gamma \), some VWAP-plus-fee contract gives the client a payoff that is not worse by \( \gamma \) than the optimal payoff when the permanent price impact is small enough and volumes are close enough to deterministic in a suitable sense.

When there is permanent price impact, the guaranteed VWAP contract \( \tau^{VWAP} \) 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 \( \tau \), denote by \( \mathcal{X}(\tau) \) the set of trading policies \( x(\cdot) \) that, together with \( \tau \), satisfy both individual rationality and
incentive compatibility, which become here

\[\mathbb{E}[\tau(p, x(\eta) + \eta + \delta) - p \cdot x(\eta)] \geq 0\]  \hspace{1cm} (IR')

\[\forall \hat{x}(\cdot) : \mathbb{E}[\tau(p, \hat{x}(\eta) + \eta + \delta) - p \cdot \hat{x}(\eta)] \geq \mathbb{E}[\tau(p, x(\eta) + \eta + \delta) - p \cdot x(\eta)].\]  \hspace{1cm} (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 7. Assume that \(\mathbb{E}\left[\frac{\eta_1 \eta_2 + 1}{\min(\eta_1, \eta_2)}\right] < \infty\). Then for all \(\gamma > 0\), there exist \(f > 0\), \(\bar{c} > 0\) and \(\delta > 0\) such that, for all \(c \in [0, \bar{c}]\) and all \(\delta \in [0, \bar{\delta}]^2\) almost surely, we have \(\mathcal{X}(\tau^{VWAP} + f) \neq \emptyset\) and

\[\sup_{x \in \mathcal{X}(\tau^{VWAP} + f)} \mathbb{E}[\tau^{VWAP}(p, x(\eta) + \eta + \delta) + f] \leq \inf_{\tau, x \in \mathcal{X}(\tau)} \mathbb{E}[\tau(p, x(\eta) + \eta + \delta)] + \gamma.\]  \hspace{1cm} (2)

The left-hand side of (2) 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 \(\gamma\)) is a lower bound on her payment in the optimum. The result says that if \(c\) and \(\delta\) are small, then these two bounds are close, so that \(\tau^{VWAP} + f\) is approximately optimal.

The proof of this result 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 \(\delta\), the right-hand side of (2) is close to the client’s utility under the second-best solution to the baseline model. Third, for small \(c\) and \(\delta\), 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 4.3, such a contract provides the dealer with a strict incentive to pursue
the first-best policy in the baseline model), which implies that the left-hand side of (2) is also close to the client’s utility under the second-best solution to the baseline model.

It is true that the path of volumes—or, following footnote 13, the path of relative volumes—is not perfectly predictable. However, in many markets, a sophisticated trader would be able to forecast it with a fair degree of accuracy (e.g., Exhibit 9 of Satish, Saxena and Palmer, 2014). As remarked in Section 3.2, this stands in stark contrast to the unpredictability of the price path, which is what underlies our motivation for modeling prices as stochastic but volumes as deterministic in the baseline model. Moreover, this observation suggests that, to the extent volumes are not perfectly predictable, the relevant cases involve small deviations from perfect predictability. For that reason, the above limiting result for $\delta$ is a valuable one. We interpret it to mean that our baseline assumption of deterministic volumes—though strong—is not indispensable for our conclusions about the performance of VWAP-based contracts.

For completeness, Proposition 7 also contains an analogous limiting result for price impact: if the permanent component of price impact is sufficiently small, then a VWAP-based contract remains approximately optimal. For an example of a market to which this aspect of the result is relevant, consider stock index futures. Berkman, Brailsford and Frino (2005) quantify permanent price impact in that market, 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).
5.3.2 Non-negligible permanent price impact

The previous analysis indicates that VWAP-based contracts remain approximately optimal when 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, the VWAP contract does not possess this property, due to the commutative property of the weighted average. Therefore, when offered a VWAP 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 $\delta_1 = \delta_2 = 0$ so that market volumes are forecastable, as in the baseline. Recall that price impact is of the form

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

For the purposes of this analysis, we use the notation $x_1 = x$ and $x_2 = 1 - x$. We begin by deriving the first-best trading policy:

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

which yields the optimal number of shares in the first period

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

The effect of the permanent component of price impact, parametrized by $c$, is that the first-best trading policy becomes smoothed toward a uniform rate of trading, regardless of how
market conditions may vary across time: \( \lim_{c \to \infty} x^{FB} = \frac{1}{2} \).

We next derive the dealer’s profit-maximizing trading strategy under a guaranteed VWAP contract. Under that contract, the payment made by the client to the dealer is \( \tau^{VWAP} = \frac{(x + \eta_1)p_1 + (1-x + \eta_2)p_2}{\eta_1 + \eta_2 + 1} \), so, recalling that we have here assumed the dealer to be risk-neutral, he optimizes

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

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

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

As in our baseline analysis, if the dealer were to trade \( \frac{\eta_1}{\eta_1 + \eta_2} \) in the first period, then his compensation would exactly offset his costs, leaving him with zero profits. Without permanent price impact, that is the best he can do. But with permanent price impact he can do better. From the expression for \( x^{VWAP} \) given above, we can see that \( \frac{d}{dx}(x^{VWAP}) > 0 \), so that permanent price impact leads the dealer to trade more than \( \frac{\eta_1}{\eta_1 + \eta_2} \) in the first period. This ‘manipulation’ leads to a slightly higher price in the second period (compared to the case in which \( x = \frac{\eta_1}{\eta_1 + \eta_2} \)). The higher price in the second period affects both the market VWAP and the dealer’s costs. However, because the dealer traded slightly more in the first period, he is less affected on a volume-weighted basis than what is reflected in the market VWAP, hence can beat the market VWAP overall. Since the dealer can therefore earn a positive expected profit, the VWAP contract cannot be optimal for the reasons laid out in footnote 23.

Moreover, if \( c > 0 \) so that there exists permanent price impact, then we also have \( x^{VWAP} \neq x^{FB} \), meaning that the VWAP contract induces the dealer to distort his trading

\[\text{As an aside, note that if } \eta_1 = \eta_2, \text{ then as in the baseline specification of } \text{Bertsimas and Lo (1998), first best corresponds to trading at a uniform rate over time: } x^{FB} = \frac{1}{2}.\]
decision away from the first-best trading policy.\textsuperscript{25} For that reason, even a ‘VWAP-minus-commission’ contract chosen to make (IR’) bind would not be optimal for the reason that it would likewise not induce first-best trading.\textsuperscript{26} Note that in this case of a risk-neutral dealer, an appropriately-specified fixed price contract would achieve the optimum (and, in fact, achieve the first best).\textsuperscript{27}

5.3.3 Non-negligible stochastic volumes

Previous analysis indicated that VWAP-based contracts remain approximately optimal when the stochastic component of market 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 the way in which 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 the section, but where $c = 0$ so that price impact is purely temporary (just as in the baseline). We again use the notation $x_1 = x$ and $x_2 = 1 - x$. Just as in the baseline analysis, the number of shares traded

\textsuperscript{25}In fact, in this case of a risk-neutral dealer, it is also possible to sign the direction of the distortion. We have $x^{VWAP} > x^{FB}$, so that the dealer trades more in the first period than he would under first best. But this need not generalize: under a sufficient amount of risk aversion, the distortion could have either sign.

\textsuperscript{26}Although the guaranteed VWAP contract is not in itself optimal, we can show that, under the above assumptions, it is unambiguously better for the client than other alternatives, such as guaranteed market-on-close.

\textsuperscript{27}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 $\tau(p, v) = \tau_0 + (\tau_1 v_1 + \tau_2) p_1 + (\tau_1 v_2 + \tau_2) p_2$ for constants $(\tau_0, \tau_1, \tau_2)$. 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.
in the first period under the first-best policy is

\[ x^{FB} = \frac{\eta_1}{\eta_1 + \eta_2}. \]

We next derive the dealer’s profit-maximizing trading strategy under a guaranteed VWAP contract, which, in contrast to the first-best policy, is affected by the presence of \( \delta \). Under that contract, the payment made by the client to the dealer is

\[ \tau^{VWAP} = \frac{(x + \eta_1 + \delta_1)p_1 + (1-x + \eta_2 + \delta_2)p_2}{\delta_1 + \delta_2 + 1 + \eta_1 + \eta_2}, \]

and, recalling that we have assumed the dealer to be risk-neutral, he optimizes

\[ \max_{x \in [0,1]} E \left[ \tau^{VWAP} - p_1 x - p_2 (1 - x) \mid \eta_1, \eta_2 \right]. \]

Assuming now that \( \delta \) is independent of \( \epsilon \) (though not necessarily from \( \eta \)), the number of shares traded by the dealer in the first period is

\[ x^{VWAP} = \frac{E \left[ \frac{\delta_1}{\eta_1} + \frac{\delta_2}{\eta_2} + 2\frac{\delta_1}{\eta_1} + 2\frac{\eta_1}{\eta_2} + 2 \right] \mid \eta_1, \eta_2}{2E \left[ \frac{1}{\eta_1} + 1/\eta_2 \right] \left( \frac{\delta_1 + \delta_2 + \eta_1 + \eta_2}{\delta_1 + \delta_2 + 1 + \eta_1 + \eta_2} \right) \mid \eta_1, \eta_2}. \]

Note that we in general have \( x^{VWAP} \neq x^{FB} \), meaning that the VWAP contract induces the dealer to distort his trading decision away from the first-best trading policy. Then, as in Section 5.3.2, neither the guaranteed VWAP contract nor even a ‘VWAP-minus-commission’ contract chosen to make (IR’) bind would be optimal.

Stochastic volumes create several issues. An obvious one is that, under a guaranteed VWAP 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 volumes might no longer be homogeneous of degree one in the dealer’s trades and market conditions. For this reason, we
have \( x^{VWAP} \neq x^{FB} \) even when \( \delta \) is deterministic and the dealer is risk neutral.\(^{28}\)

6 Applications to benchmark design

To this point, we have focused on questions of bilateral contracting, investigating how a client should contract with her dealer. We concluded that it might be desirable to tie the dealer’s compensation to a particular benchmark price: the VWAP. But play a role in many other settings, suggesting further applications for our results. This section highlights three. Although the applications are somewhat more speculative than what we have analyzed so far, for reasons that we describe in the text, many of the same economic arguments can be made in favor of VWAP benchmarks.

6.1 Benchmark computation in opaque markets

In our baseline model, we assumed that any measurable function of prices and total 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, such as 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 total 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

\(^{28}\)Nevertheless, in this case of deterministic \( \delta \), the guaranteed VWAP contract can be modified to restore optimality. Consider the contract

\[
\tau^* = \frac{(v_1 - \delta_1)p_1 + (v_2 - \delta_2)p_2}{v_1 + v_2 - \delta_1 - \delta_2},
\]

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

39
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’). In recent years, industry participants have begun to rethink how this benchmark ought to be computed. For example, in 2015, WM/Reuters widened the relevant window for collecting prices from one minute to five (e.g., Michelberger and Witte, 2016). 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, a concern shared by the members of the Foreign Exchange Joint Standing Committee (FXJSC, 2008).

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 provide an answer to this question is as follows. Add a benchmark administrator whose role is to compute and publish a benchmark \( b(p, v) \). 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 our previous results immediately imply that an optimal benchmark is the VWAP

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

By choosing \( \tau(b) = b \), the client would effectively be writing a VWAP contract, so that Theorem 4 implies that the optimum is being achieved. Likewise, if the benchmark adminis-
trator’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.

**Discussion.** According to this analysis, an optimal benchmark is the price that would be referenced in the optimal contract under full data availability for prices and total volumes. With such a benchmark, clients can overcome their data limitations and propose the optimal contract. Though not captured directly in our model, reducing transaction costs of clients in this way would, presumably, allow more gains from trade to be realized. Traded volume would likely increase as well, and for that reason, platforms might also have an incentive to publish such a benchmark. To the extent that the foreign exchange market resembles the setting of our model, our analysis suggests that the definition of the fix should be amended to more closely resemble a VWAP. Consistent with this, the Financial Stability Board as well as the FXSCJ have in fact recommended modifying prevailing definitions to include the use of volumes (FSB, 2014; FXJSC, 2008).

One caveat that should be mentioned is that ours is a partial equilibrium model in that we assume market conditions 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 as benchmark design. In the language of our model, the choice of benchmark may affect the distribution over $\eta$ 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 realiza-
tion of the benchmark beyond the trades that they will directly conduct at that benchmark. 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. For example, the LIBOR scandal constitutes a particularly stark illustration of the power of these incentives. 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 Trading at settlement in futures markets

Another application of our model is to ‘Trading At Settlement’ (TAS), which is an order matching procedure that is available for some commodity futures. It allows market participants to trade futures contracts at a particular benchmark: the yet-to-be-determined daily settlement price. In such markets, a question may concern how best to compute this settlement price. In such markets, a question may concern how best to compute this settlement price.

Our analysis suggests that the daily settlement price is susceptible to a certain type of manipulation when it is computed in a manner other than VWAP.29 For example, if the settlement price were the closing price, then a trader might achieve a positive expected profits.

29This 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.
profit through the following scheme: conduct a TAS trade, then 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. 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. 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 manipulation of the kind described above (CFTC, 2008, 2013; WSJ, 2011). Similar considerations apply to transaction types other than TAS that are also executed at a differential from a yet-to-be-determined benchmark price, such as BTIC (Basis Trade at Index Close), TACO (Basis Trade At Cash Open), and TAM (Trading At Marker); see Rule 524 (CME Group, 2018) for details on these transaction types.

Approximately the same economics applies to on-close orders offered by equities exchanges. And indeed, the SEC has also sued traders for similar types of manipulation (SEC, 2014). But for this application to equities, one caveat that should be mentioned is that equity closing 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 closing price special: in particular, the market is typically deeper during the closing auction. To the extent that this difference is encapsulated with a very high value for $\eta_T$ relative to previous periods, then our analysis applies. But if the price impact function $h$ and/or the volume function $v$ themselves differ in period $T$ relative to previous periods, then our analysis may
not carry over directly, and the application becomes more speculative. \(^{30}\)

### 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. \(^{31}\) 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. \(^{32}\)

\(^{30}\)A similar caveat pertains to Section 6.3, below, as regards applying our analysis to the computation of NAV for equity funds.

\(^{31}\)Whereas in the baseline model, the client is hurt by higher payments to the dealer, in this fund application, the remaining shareholders are hurt by a dilution of the fund’s value.

\(^{32}\)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.
7 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 process. Even though dealers are typically bound by ‘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. Potential solutions to this conflict of interest include requiring more transparency, so that dealers could be more easily monitored, as well as more vigorous enforcement of the obligations of dealers. But another solution, on which we focus in this paper, is to search for contractual arrangements that mitigate the conflict itself by aligning the dealer’s interests with those of the client.

We see two applications for our analysis. The most direct application pertains to the question of what sort of contract a client should push for in markets with public data availability, since in such markets, it is possible to contract on prices and volumes, as our model presumes. Solving the model under a particular set of conditions, we obtain a strikingly simple solution: the guaranteed VWAP contract is optimal for the client. 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.

A second application is to the question of how to design a benchmark, which is particularly relevant in markets without public data availability. In such markets, participants cannot contract in arbitrary ways 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, the prevailing benchmark more closely resembles the closing price than the VWAP. In consequence, current principal trading arrangements often effectively take the form of guaranteed market-on-close...
contracts. However, 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 in the introduction. 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 settings that are inconsistent with our framework, most notably with the conditions laid out in Section 3.2: that price impact is temporary and that volumes are predictable. As we have shown, if deviations from these conditions are small, then VWAP-based contracts remain approximately optimal. But if deviations are large, then our results do not apply. Nevertheless, the general framework that we lay out in Section 3.1 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 Auxiliary lemmas

Lemma 8. For random variables $\varepsilon_1, \ldots, \varepsilon_T$, the following are equivalent:

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

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

Proof of Lemma 8. To show that (a) implies (b), we compute

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

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

$$
= \mathbb{E}[\varepsilon_{t+1} - \varepsilon_T | \eta, \varepsilon_1, \ldots, \varepsilon_{t-1}] - \mathbb{E}[\varepsilon_{t-1} - \varepsilon_T | \eta, \varepsilon_1, \ldots, \varepsilon_{t-1}]
$$

$$
= 0.
$$

Conversely, assume that (b) holds so that

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

$$
= \mathbb{E}[\varepsilon_t - \varepsilon_{T-2} | \eta, \varepsilon_1, \ldots, \varepsilon_{t-1}] = \cdots = \mathbb{E}[\varepsilon_t - \varepsilon_t | \eta, \varepsilon_1, \ldots, \varepsilon_{t-1}]
$$

From this, we deduce

$$
\mathbb{E}[\varepsilon_t - \varepsilon_{T-1} | \eta, \varepsilon_1, \ldots, \varepsilon_{t-1}] = \mathbb{E}[\varepsilon_t - \varepsilon_{T-2} | \eta, \varepsilon_1, \ldots, \varepsilon_{t-1}] = \cdots = \mathbb{E}[\varepsilon_t - \varepsilon_t | \eta, \varepsilon_1, \ldots, \varepsilon_{t-1}]
$$

by repeatedly applying (b).

Lemma 9. Consider random variables $\varepsilon$ and $\eta$. Assume that either
1. there exist functions $x(\cdot)$ depending on $\eta$ such that it holds $\sum_{t=1}^{T} x_t(\eta) = 1$ almost surely, and $\mathbb{E}[\varepsilon_t|\eta] = \mu$ almost surely for all $t$, or

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

Then

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

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

**Proof of Lemma 9.** 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

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

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

$$\sum_{t=1}^{T} \mathbb{E}[\varepsilon_t x_t|\eta] = \mathbb{E}[\varepsilon_T|\eta] + \sum_{t=1}^{T-1} \mathbb{E}[(\varepsilon_t - \varepsilon_T)x_t|\eta]$$

$$= \mu + \sum_{t=1}^{T-1} \mathbb{E}[\varepsilon_t - \varepsilon_T |\eta, \varepsilon_1, \ldots, \varepsilon_{t-1}, x_t|\eta]$$

$$= \mu,$$

where the final step is thanks to (1) and Lemma 8.

**Lemma 10.** When $v(x, \eta)$ satisfies the equivalent statements of Proposition 1, then

$$\left( \sum_{t=1}^{T} x_t \right) \left( \sum_{t=1}^{T} h\left( \frac{x_t}{\eta_t} \right) v(x_t, \eta_t) \right) \leq \left( \sum_{t=1}^{T} v(x_t, \eta_t) \right) \left( \sum_{t=1}^{T} h\left( \frac{x_t}{\eta_t} \right) x_t \right)$$  \hspace{1cm} (3)
for all \( T \in \mathbb{N} \) and \((x, \eta), \ldots, (x_T, \eta_T) \in \text{dom}(v)\). For \((x, \eta), \ldots, (x_T, \eta_T) \in \text{dom}(v)\) with \(\sum_{t=1}^{T} x_t = 1\), equality in (3) 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 10.** We prove (3) by induction over \( T \).

**Induction base:** For \( T = 1 \), (3) 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 (3) 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

\[
h\left(\frac{x_T}{\eta_T}\right) v(x_T, \eta_T) x_T + x_T h\left(\frac{x_T}{\eta_T}\right) v(x_T, \eta_T) \leq h\left(\frac{x_T}{\eta_T}\right) x_T v(x_T, \eta_T) + v(x_T, \eta_T) h\left(\frac{x_T}{\eta_T}\right) x_T
\]

for every \( t = 1, 2, \ldots, T - 1 \). Rearranging terms, (4) is equivalent to

\[
(x_t v(x_T, \eta_T) - x_T v(x_t, \eta_t)) \left( h\left(\frac{x_T}{\eta_T}\right) - h\left(\frac{x_t}{\eta_t}\right)\right) \leq 0. \tag{5}
\]

If \( x_t = 0 \) or \( x_T = 0 \), then (5) holds. In other cases, using \( v(x, \eta) = x V(x/\eta) \) 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 \((x, \eta), (x_T, \eta_T) \in \text{dom}(v)\) because \( V \) is decreasing and \( h \) is increasing.
We now turn to the second part. It is straightforward to check that if \( x_t = \frac{\eta}{\sum_{s=1}^{T} \eta_s} \) for all \( t = 1, \ldots, T \), then (3) holds with equality. For the converse, consider \((x_1, \eta_1), \ldots, (x_T, \eta_T) \in \text{dom}(v)\) with \( \sum_{t=1}^{T} x_t = 1 \) and suppose that (3) holds with equality. By the above induction hypothesis, this can be the case only if (5) holds with equality for all \( t = 1, \ldots, T - 1 \). Note that (5) 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 (3) only if \( x_T \eta_t = x_t \eta_T \) for all \( t \), which means \( x_t = \frac{\eta}{\sum_{s=1}^{T} \eta_s} \) since \( \sum_{t=1}^{T} x_t = 1 \).

B Proofs

Throughout this Appendix, we will typically write a trading policy as \( x(\eta) \) for notational convenience, despite the potential dependence on previous prices (viz. when the proofs apply to the version of the model considered in Section 5.2.1).

Proof of Proposition 1. 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 \( (x, \eta) \in \text{dom}(v) \) with \( x \neq 0 \). For
\((x', \eta'), (x'', \eta'') \in \text{dom}(v)\) with \(x'x'' \neq 0\) and \(x'/\eta' = x''/\eta''\), we deduce
\[
g(x', \eta') = \frac{v(x', \eta')}{x'} = \frac{v(\frac{\eta'}{\eta}x', \frac{\eta'}{\eta} \eta'')}{\frac{\eta'}{\eta}x'} = \frac{\eta''}{\eta'} \frac{v(x'', \eta'')}{x''} = \frac{v(x'', \eta'')}{x''} = g(x'', \eta''), \tag{6}
\]
where the third equality uses the fact that \(v(x, \eta)\) is homogeneous of degree one. We partition \(\text{dom}(v)\) into sets \(D_y = \{(x, \eta) \in \text{dom}(v) : x/\eta = y\}\) for \(y \in \mathbb{R}_+\). If there are no \((x, \eta) \in \text{dom}(v)\) with \(x/\eta = y\), we set \(D_y = \emptyset\). Note that \(\text{dom}(v) = \bigcup_{y \in \mathbb{R}_+} D_y\) and \(D_y \cap D_z = \emptyset\) for \(y \neq z\). For every \(y\) with \(D_y \neq \emptyset\) with \(y \neq 0\), (6) implies that \(g(x, \eta)\) takes the same value for all \((x, \eta) \in D_y\). Therefore, we can write \(g(x, \eta) = V(\frac{x}{\eta})\) for a function \(V\) and \((x, \eta) \in \text{dom}(v)\) with \(x \neq 0\), so that \(v(x, \eta) = xV(\frac{x}{\eta})\) for all \((x, \eta) \in \text{dom}(v)\). Because \(v(x, \eta)\) is strictly increasing in \(\eta\) for \((x, \eta) \in \text{dom}(v)\) with \(x \neq 0\), we obtain that \(V(y)\) is strictly decreasing for \(y \neq 0\).

Proof of Lemma 2. Plugging in \(p\), a trading policy \(x(\cdot)\) is first best if for all \(\eta, x(\eta)\) minimizes the following objective subject to the constraint \(\sum_{t=1}^T x_t = 1\):
\[
\mathbb{E} \left[ \sum_{t=1}^T \left( h\left(\frac{x_t}{\eta_t}\right) x_t + \varepsilon_t x_t\right) \right| \eta] = \sum_{t=1}^T \mathbb{E} \left[ h\left(\frac{x_t}{\eta_t}\right) x_t \right| \eta] + \sum_{t=1}^T \mathbb{E}[\varepsilon_t x_t | \eta].
\]
By Lemma 9, the last term equates to \(\mu\) almost surely. We therefore find that this objective,
the expected trading cost conditional on $\eta$, is

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

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

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

$$
= \mu + \mathbb{E} \left[ h \left( \frac{1}{\sum_{s=1}^{T} \eta_s} \sum_{t=1}^{T} x_t \right) \left( \sum_{t=1}^{T} x_t \right) | \eta \right]
$$

$$
= \mu + h \left( \frac{1}{\sum_{t=1}^{T} \eta_t} \right)
$$

almost surely, where the second step in the above uses Jensen’s inequality applied to the convex function $y h(y)$, 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 $\eta$ is minimized if and only if the trading schedule is $x = \left( \frac{\eta}{\sum_{s=1}^{T} \eta_s} \right)_{t=1}^{T}$. Thus, $x^{FB}(\cdot)$ 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 2 follows from

$$
\frac{v(x^{FB}(\eta), \eta_t)}{\sum_{s=1}^{T} v(x^{FB}(\eta), \eta_s)} = \frac{x^{FB}(\eta)V \left( \frac{x^{FB}(\eta)}{\eta_t} \right)}{\sum_{s=1}^{T} x^{FB}(\eta)V \left( \frac{x^{FB}(\eta)}{\eta_s} \right)} = \frac{\frac{\eta_t}{\sum_{r=1}^{T} \eta_r} V \left( \frac{1}{\sum_{r=1}^{T} \eta_r} \right)}{\frac{\eta_t}{\sum_{s=1}^{T} \eta_s} V \left( \frac{1}{\sum_{s=1}^{T} \eta_s} \right)} = \frac{\eta_t}{\sum_{r=1}^{T} \eta_r} = x^{FB}(\eta),
$$

where the first equality uses that $v(x, \eta) = x V \left( \frac{x}{\eta} \right)$ for all $(x, \eta) \in \text{dom}(v)$ by assumption. □
Proof of Lemma 3. Sufficiency. If \( \tau \) satisfies condition (i), then \((\tau, x^{FB}(\cdot))\) satisfies (IC). Similarly, if \( \tau \) satisfies condition (ii), then \((\tau, x^{FB}(\cdot))\) satisfies (IR). Furthermore, condition (ii) also implies

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

by Lemma 2. Moreover, no pair \((\tau', x(\cdot))\) satisfying (IR) can better this objective. To see this, first note that (IR) requires

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

Since \( u \) is concave and strictly increasing, this requires

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

where the last step follows from Lemma 2.

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

\[
\mathbb{E}[\tau'(p, v(x(\eta), \eta))] = \mu + \mathbb{E}\left[h\left(\frac{1}{\sum_{t=1}^{T} \eta_t}\right)\right]. \tag{7}
\]
First, we claim that $x(\cdot) = x^{FB}(\cdot)$ almost surely. Suppose by way of contradiction that this is not the case. Then Lemma 2 implies

$$\mathbb{E}[p \cdot x(\eta)] > \mu + \mathbb{E}\left[h\left(\frac{1}{\sum_{t=1}^{T} \eta_t}\right)\right]. \quad (8)$$

Combining (7) and (8),

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

Because $u$ is concave and strictly increasing, this implies that

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

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

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

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

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

$$\mathbb{E}[\tau'(p, v(x(\eta), \eta)) - p \cdot x(\eta)] > 0. \quad (9)$$

Combining (7) and (9), we obtain

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

54

Electronic copy available at: https://ssrn.com/abstract=3177283
which is impossible by Lemma 2.

\[ \square \]

**Proof of Theorem 4.** To show that \( \tau^{VWAP} \) is an optimal contract, it suffices to establish that it satisfies the two conditions of Lemma 3. We begin by observing that

\[
E \left[ \sum_{t=1}^{T} \epsilon_t v(x_t, \eta_t) \right] = E \left[ \frac{\sum_{t=1}^{T} \epsilon_t v(x_t, \eta_t)}{\sum_{s=1}^{T} v(x_s, \eta_s)} \right] = E \left[ \sum_{t=1}^{T} E \left[ \frac{\epsilon_t v(x_t, \eta_t)}{v(x_s, \eta_s)} \bigg| \eta \right] \right] = \mu, \quad (10)
\]

using \( E[\epsilon_t \eta] = \mu \) almost surely.

Applying Jensen’s inequality to the concave function \( u \), we obtain that the dealer’s expected utility from pursuing a trading schedule \( x \) is

\[
E\left[u\left(\tau^{VWAP} - p \cdot x\right)\right] = E\left[u\left(\sum_{t=1}^{T} \left(h\left(\frac{x_t}{\eta_t}\right) + \epsilon_t\right) v(x_t, \eta_t) - \sum_{t=1}^{T} \left(h\left(\frac{x_t}{\eta_t}\right) + \epsilon_t\right) x_t\right)\right] \\
\leq E\left[u\left(\sum_{t=1}^{T} \epsilon_t v(x_t, \eta_t) - \sum_{t=1}^{T} \epsilon_t x_t\right)\right] \quad (11) \\
\leq u\left(E\left[\sum_{t=1}^{T} \epsilon_t v(x_t, \eta_t) - \sum_{t=1}^{T} \epsilon_t x_t\right]\right) \quad (12) \\
= u(0),
\]

where (11) follows from the first part of Lemma 10; and the last equality is implied by Lemma 9 and equation (10). Equality in (11) holds if and only if (3) holds almost surely, hence if and only if \( x_t = \frac{n_t}{\sum_{s=1}^{T} \eta_s} \) almost surely, by the second part of Lemma 10. 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 2, so that there is equality in (12) as well. Thus, a trading policy \( x(\cdot) \) maximizes \( E\left[u\left(\tau^{VWAP} (p, v(x(\eta), \eta)) - p \cdot x(\eta)\right)\right] \) if and only if it implies that \( x_t = \frac{n_t}{\sum_{s=1}^{T} \eta_s} \) almost surely, or equivalently, if and only if it equals \( x^{FB}(\cdot) \) almost surely.

We therefore conclude that \( (\tau^{VWAP}, x^{FB}(\cdot)) \) satisfies (IC), which implies condition (i) of
Lemma 3. But in fact, we also obtain the stronger conclusion that for all trading policies  \( \hat{x}(\cdot) \) not equal to  \( x^{FB}(\cdot) \) almost surely, (IC) holds with strict inequality:

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

The same computation reveals that  \( \tau^{VWAP}(p, v(x^{FB}(\eta), \eta)) - p \cdot x^{FB}(\eta) = 0 \). We therefore obtain condition (ii) of Lemma 3.

**Proof of Theorem 5.** Suppose that  \( u \) is strictly concave and that the distributions of  \( \varepsilon \) and  \( \eta \) have full support over  \( \mathbb{R}^T \) and  \( \mathbb{R}^{T,+} \), respectively. Suppose that  \( \tau \) is an optimal contract.

In proving Theorem 4, we established that  \( \tau^{VWAP} \) satisfies the conditions of Lemma 3. Therefore, the second half of that lemma requires that  \( \tau \) does the same. Condition (ii) of that lemma requires that both of the following hold almost surely:

\[
\tau(p, v(x^{FB}(\eta), \eta)) = p \cdot x^{FB}(\eta)
\]

\[
\tau^{VWAP}(p, v(x^{FB}(\eta), \eta)) = p \cdot x^{FB}(\eta)
\]

Using  \( \mathbf{1} \) to denote a vector of ones, we conclude that the following holds almost surely:

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

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

**Proof of Theorem 4’.** Since  \( v(x, \eta) = x + \eta \), the total 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 $\eta$ so that
\[
\mathbb{E}\left[\frac{\sum_{t=1}^{T} \varepsilon_t v(x_t, \eta_t)}{\sum_{s=1}^{T} v(x_s, \eta_s)}\right] = \mathbb{E}\left[\frac{\sum_{t=1}^{T} \varepsilon_t v(x_t, \eta_t)}{\sum_{s=1}^{T} v(x_s, \eta_s)} \mid \eta\right]
= \mathbb{E}\left[\frac{\sum_{t=1}^{T} \varepsilon_t v(x_t, \eta_t)}{\sum_{s=1}^{T} v(x_s, \eta_s)} \mid \eta\right] \frac{\sum_{t=1}^{T} \varepsilon_t v(x_t, \eta_t)}{1 + \sum_{s=1}^{T} \eta_s}
= \mathbb{E}\left[\mu + \mu \sum_{t=1}^{T} \varepsilon_t \frac{x_t}{\eta_t} \right] = \mu,
\]
(13)

using Lemma 9 and $\mathbb{E}[\varepsilon_t \mid \eta] = \mu$ almost surely.

Applying Jensen’s inequality to the concave function $u$, we obtain that the dealer’s expected utility from pursuing a trading schedule $x$ is
\[
\mathbb{E}[u(\tau \text{VWAP} - \mathbf{p} \cdot \mathbf{x})]
= \mathbb{E}\left[u\left(\sum_{t=1}^{T} \frac{h\left(\frac{x_t}{\eta_t}\right) + \varepsilon_t v(x_t, \eta_t)}{\sum_{s=1}^{T} v(x_s, \eta_s)} - \sum_{t=1}^{T} \left(h\left(\frac{x_t}{\eta_t}\right) + \varepsilon_t\right)x_t\right)\right]
\leq \mathbb{E}\left[u\left(\sum_{t=1}^{T} \frac{\varepsilon_t v(x_t, \eta_t)}{\sum_{s=1}^{T} v(x_s, \eta_s)} - \sum_{t=1}^{T} \varepsilon_t x_t\right)\right]
\leq u\left(\mathbb{E}\left[\sum_{t=1}^{T} \frac{\varepsilon_t v(x_t, \eta_t)}{\sum_{s=1}^{T} v(x_s, \eta_s)} - \sum_{t=1}^{T} \varepsilon_t x_t\right]\right)
= u(0),
\]
(14)

where (14) follows from the first part of Lemma 10; and the last equality is implied by Lemma 9 and equation (13). The remainder of the argument proceeds as in the proof of Theorem 4.

Proof of Proposition 7. To simplify the notation, we write $x = x_1$ so that $x_2 = 1 - x$. We also choose $f = \gamma/2$. 

57
Step 1: we show $\mathcal{X}(\tau^{VWAP} + f) \neq \emptyset$ for $c \leq f$.

To this end, we compute

$$
\mathbb{E}[\tau^{VWAP}(p, x(\eta) + \eta + \delta) - p \cdot x(\eta) | \eta, \delta] \\
= \mathbb{E} \left[ \frac{(cx + x/\eta_1 + \varepsilon_1)(x + \eta_1 + \delta_1)}{1 + \eta_1 + \eta_2 + \delta_1 + \delta_2} - \frac{(cx + x/\eta_1 + \varepsilon_1)x}{1 + \eta_1 + \eta_2 + \delta_1 + \delta_2} \\
+ \frac{(c + (1 - x)/\eta_2 + \varepsilon_2)(1 - x + \eta_2 + \delta_2)}{1 + \eta_1 + \eta_2 + \delta_1 + \delta_2} - \frac{(c + (1 - x)/\eta_2 + \varepsilon_2)(1 - x)}{1 + \eta_1 + \eta_2 + \delta_1 + \delta_2} \bigg| \eta, \delta \right] \\
= \mathbb{E} \left[ \frac{(cx + x/\eta_1)(\eta_1 + \delta_1 - x(\eta_1 + \eta_2 + \delta_1 + \delta_2))}{1 + \eta_1 + \eta_2 + \delta_1 + \delta_2} \\
+ \frac{(c + (1 - x)/\eta_2)(\eta_1 + \delta_1 - x(\eta_1 + \eta_2 + \delta_1 + \delta_2))}{1 + \eta_1 + \eta_2 + \delta_1 + \delta_2} \bigg| \eta, \delta \right] \\
= \mathbb{E} \left[ \frac{(cx - c + x/\eta_1 - (1 - x)/\eta_2)(\eta_1 + \delta_1 - x(\eta_1 + \eta_2 + \delta_1 + \delta_2))}{1 + \eta_1 + \eta_2 + \delta_1 + \delta_2} \right].
$$

(15)

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

$$
\mathbb{E}[\tau^{VWAP}(p, x(\eta) + \eta + \delta) - p \cdot x(\eta) | \eta, \delta] = \frac{c - \eta_2}{\eta_1 + \eta_2} \left( \frac{\eta_2}{\eta_1 + \eta_2} - \frac{\eta_1}{\eta_1 + \eta_2} \delta_2 \right) \\
\mathbb{E}[\tau^{VWAP}(p, x(\eta) + \eta + \delta) - p \cdot x(\eta) | \eta, \delta] \geq \frac{-c(\frac{\eta_2}{\eta_1 + \eta_2})^2 \delta_1}{1 + \eta_1 + \eta_2 + \delta_1 + \delta_2} \geq -c.
$$

We deduce

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

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 (2).
Because the second-best outcome cannot be better than that of the first best, we have

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

We can further compute

\[
\mathbb{E}[p \cdot x^{FB}(\eta)] = \mu + \mathbb{E}\left[\frac{c(x^{FB})^2 + (x^{FB})^2}{\eta_1} + c(1 - x^{FB}) + \frac{(1 - x^{FB})^2}{\eta_2}\right]
\]

\[
= \mu + \mathbb{E}\left[(c + 1/\eta_1 + 1/\eta_2)\left(x^{FB} - \frac{c/2 + 1/\eta_2}{c + 1/\eta_1 + 1/\eta_2}\right)^2 - \frac{(c/2 + 1/\eta_2)^2}{c + 1/\eta_1 + 1/\eta_2} + c + 1/\eta_2\right]
\]

\[
= \mu + \mathbb{E}\left[(c + 1/\eta_1 + 1/\eta_2)(c + 1/\eta_2) - (c/2 + 1/\eta_2)^2\right]
\]

\[
= \mu + \mathbb{E}\left[\frac{3c^2/4 + c/\eta_1 + c/\eta_2 + 1/(\eta_1\eta_2)}{c + 1/\eta_1 + 1/\eta_2}\right]
\]

\[
\geq \mu + \mathbb{E}\left[\frac{1/(\eta_1\eta_2)}{1/\eta_1 + 1/\eta_2}\right],
\]

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

\[
\frac{\partial}{\partial c} \frac{3c^2/4 + c/\eta_1 + c/\eta_2 + 1/(\eta_1\eta_2)}{c + 1/\eta_1 + 1/\eta_2}
\]

\[
= \frac{(c + 1/\eta_1 + 1/\eta_2)(3c/2 + 1/\eta_1 + 1/\eta_2) - (3c^2/4 + c/\eta_1 + c/\eta_2 + 1/(\eta_1\eta_2))}{(c + 1/\eta_1 + 1/\eta_2)^2}
\]

\[
= \frac{(1/\eta_1 + 1/\eta_2)(3c/2 + 1/\eta_1 + 1/\eta_2) + 3c^2/4 - 1/(\eta_1\eta_2)}{(c + 1/\eta_1 + 1/\eta_2)^2}
\]

\[
> 0.
\]
Therefore, we find a lower bound for the right-hand side of (2), namely,

\[
\inf_{\tau,x}\mathbb{E}\left[\tau (p, x(\eta) + \eta + \delta)\right] + \gamma \geq \mu + \mathbb{E}\left[\frac{1}{\eta_1 + \eta_2}\right] + \gamma. \tag{16}
\]

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

\[
\mathbb{E}\left[\tau^{VWAP} (p, x(\eta) + \eta + \delta)\right] \leq \mu + \mathbb{E}\left[\frac{1}{\eta_1 + \eta_2}\right] + \gamma/2 \tag{17}
\]

for all \(c \in [0, \overline{c}]\) and all \(\delta \in [0, \overline{\delta}]^2\) almost surely, where \(\overline{c} > 0\) and \(\overline{\delta} > 0\) will be specified in the following. Note that this will complete the proof, as (17) implies (2) thanks to (16) and \(f = \gamma/2\).

We require that \(\overline{c} > 0\) and \(\overline{\delta} > 0\) satisfy \(\overline{c} \leq f\) and

\[
\mathbb{E}\left[\frac{1}{\eta_1 + \eta_2} \max \left\{\eta_1 \overline{c} + \max \{\delta/\eta_1, \overline{c}\eta_1\eta_2\}(\eta_1 \overline{c} + 1), \overline{c}(\eta_1 + \eta_2) + \delta/\eta_2\right\}\right] \leq \gamma/2. \tag{18}
\]

Note that such \(\overline{c} > 0\) and \(\overline{\delta} > 0\) exist because

\[
\lim_{\tau \searrow 0, \delta \searrow 0} \mathbb{E}\left[\frac{1}{\eta_1 + \eta_2} \max \left\{\eta_1 \overline{c} + \max \{\delta/\eta_1, \overline{c}\eta_1\eta_2\}(\eta_1 \overline{c} + 1), \overline{c}(\eta_1 + \eta_2) + \delta/\eta_2\right\}\right] = 0 \tag{19}
\]

thanks to the assumption \(\mathbb{E}\left[\frac{\eta_1\eta_2 + 1}{\eta_1 + \eta_2}\right] < \infty\). Indeed, to show (19), we interchange limit and expectation and note that the expression within the expectation converges to zero as \(\overline{c} \searrow 0\) and \(\overline{\delta} \searrow 0\). To be able to interchange limit and expectation, we need that the expectation is finite, which can be shown using \(\mathbb{E}\left[\frac{\eta_1\eta_2 + 1}{\eta_1 + \eta_2}\right] < \infty\).

To prove (17), we first show

\[
\frac{\eta_1 - \overline{\delta}}{\eta_1 + \eta_2} \leq x(\eta) \leq \frac{\eta_1 + \max \{\overline{c}, \overline{c}\eta_1\eta_2\}}{\eta_1 + \eta_2} \tag{20}
\]
almost surely. We achieve this by showing that if (20) did not hold almost surely, there would be a failure of \((IC')\), which refers to the fact that \(x(\cdot)\) maximizes \(E[\tau^{VWAP}(p, \hat{x}(\eta) + \eta + \delta) - p \cdot \hat{x}(\eta)]\) over \(\hat{x}(\cdot)\). And to show that, it suffices to demonstrate that if (20) is violated for some value of \(\eta\), then either (i) for all \(\delta \in [0, \overline{\delta}]^2\),

\[
\frac{\partial}{\partial x} E[\tau^{VWAP}(p, x(\eta) + \eta + \delta) - p \cdot x(\eta) | \eta, \delta] > 0
\]

or (ii) for all \(\delta \in [0, \overline{\delta}]^2\),

\[
\frac{\partial}{\partial x} E[\tau^{VWAP}(p, x(\eta) + \eta + \delta) - p \cdot x(\eta) | \eta, \delta] < 0,
\]

so that the first-order condition is not satisfied at \(\eta\). Using \(x\) to denote \(x(\eta)\), we begin by using (15) to compute

\[
\frac{\partial}{\partial x} E[\tau^{VWAP}(p, x(\eta) + \eta + \delta) - p \cdot x(\eta) | \eta, \delta] = \\
\frac{\partial}{\partial x} (cx - c + x/\eta_1 - (1 - x)/\eta_2)(\eta_1 + \delta_1 - x(\eta_1 + \eta_2 + \delta_1 + \delta_2)) \\
= \frac{\partial}{\partial x} \frac{-x^2(c + 1/\eta_1 + 1/\eta_2)(\eta_1 + \eta_2 + \delta_1 + \delta_2) + x(c + 1/\eta_1 + 1/\eta_2)(\eta_1 + \delta_1)}{1 + \eta_1 + \eta_2 + \delta_1 + \delta_2} \\
+ \frac{\partial}{\partial x} \frac{x(c + 1/\eta_1 + 1/\eta_2)(\eta_2 + \delta_1 + \delta_2) - (c + 1/\eta_2)(\eta_1 + \delta_1)}{1 + \eta_1 + \eta_2 + \delta_1 + \delta_2} \\
- 2\frac{x(c + 1/\eta_1 + 1/\eta_2)(\eta_1 + \eta_2 + \delta_1 + \delta_2)}{1 + \eta_1 + \eta_2 + \delta_1 + \delta_2} \\
+ \frac{(c + 1/\eta_1 + 1/\eta_2)(\eta_1 + \delta_1) + (c + 1/\eta_2)(\eta_1 + \eta_2 + \delta_1 + \delta_2)}{1 + \eta_1 + \eta_2 + \delta_1 + \delta_2}.
\]
In the event \( x > \frac{n_1 + \max(\delta, c\eta_1, \eta_2)}{\eta_1 + \eta_2} \), we have

\[
-2x(c + 1/\eta_1 + 1/\eta_2)(\eta_1 + \eta_2 + \delta_1 + \delta_2) + (c + 1/\eta_1 + 1/\eta_2)(\eta_1 + \delta_1) \\
+ (c + 1/\eta_2)(\eta_1 + \eta_2 + \delta_1 + \delta_2) \\
< (c + 1/\eta_1 + 1/\eta_2) \left( \eta_1 + \delta_1 - \frac{\eta_1 + \delta}{\eta_1 + \eta_2} (\eta_1 + \eta_2 + \delta_1 + \delta_2) \right) \\
+ (\eta_1 + \eta_2 + \delta_1 + \delta_2) \left( c + 1/\eta_2 - \frac{1/\eta_2 + \varepsilon}{1/\eta_1 + 1/\eta_2} (c + 1/\eta_1 + 1/\eta_2) \right) \\
< 0
\]

so that

\[
\frac{\partial}{\partial x} \mathbb{E} \left[ r_{VWAP} (p, x(\eta) + \eta + \delta) - p \cdot x(\eta) \mid \eta, \delta \right] < 0 \text{ on } x > \frac{n_1 + \max(\delta, c\eta_1, \eta_2)}{\eta_1 + \eta_2}.
\]

Similarly, we can show

\[
\frac{\partial}{\partial x} \mathbb{E} \left[ r_{VWAP} (p, x(\eta) + \eta + \delta) - p \cdot x(\eta) \mid \eta, \delta \right] > 0 \text{ on } x < \frac{\eta_1 - \delta}{\eta_1 + \eta_2}.
\]
Thus, we have shown that (20) holds almost surely. From (20), we deduce that

\[
\mathbb{E}[\tau^{VWAP}(\mathbf{p}, \mathbf{x}(\mathbf{\eta}) + \mathbf{\eta} + \delta)] = \mu + \mathbb{E} \left[ \frac{(c x + x/\eta_1)(x + \eta_1 + \delta_1) + (c + (1 - x)/\eta_2)(1 - x + \eta_2 + \delta_2)}{1 + \eta_1 + \eta_2 + \delta_1 + \delta_2} \right] \\
\leq \mu + \mathbb{E} \left[ \max\{c x + x/\eta_1, c + (1 - x)/\eta_2\} \right] \\
\leq \mu + \mathbb{E} \left[ \max \left\{ \frac{1 + \max\{\bar{\delta}/\eta_1, \bar{\delta}/\eta_2\}}{\eta_1 + \eta_2} (\eta_1 \bar{c} + 1), \bar{c} + \frac{1 + \bar{\delta}/\eta_2}{\eta_1 + \eta_2} \right\} \right] \\
\leq \mu + \mathbb{E} \left[ \frac{1}{\eta_1 + \eta_2} \right] + \mathbb{E} \left[ \frac{1}{\eta_1 + \eta_2} \max \left\{ \eta_1 \bar{c} + \max\{\bar{\delta}/\eta_1, \bar{\delta}/\eta_2\} (\eta_1 \bar{c} + 1), \bar{c} (\eta_1 + \eta_2) + \bar{\delta}/\eta_2 \right\} \right] \\
\leq \mu + \mathbb{E} \left[ \frac{1}{\eta_1 + \eta_2} \right] + \gamma/2,
\]

using (18) for the last inequality. This shows (17) and concludes the proof.

\[\square\]

References


_ , Anna A. Obizhaeva, and Yajun Wang, “Smooth Trading with Overconfidence and

Electronic copy available at: https://ssrn.com/abstract=3177283


_ , “HSBC Holdings Plc Agrees to Pay More Than $100 Million to Resolve Fraud Charges,”

