Journal of Financial Economics 160 (2024) 103901

Contents lists available at ScienceDirect

Journal of Financial Economics

journal homepage: www.elsevier.com/locate/finec

# Block trade contracting[☆]

Markus Baldauf[a], Christoph Frei[b], Joshua Mollner[c,*]

[a] University of British Columbia, Sauder School of Business, 2053 Main Mall, Vancouver, BC, V6T 1Z2, Canada
[b] University of Alberta, 621 Central Academic Building, Edmonton, AB, T6G 2G1, Canada
[c] Northwestern University, Kellogg School of Management, 2211 Campus Drive, Evanston, IL, 60657, USA

ARTICLE INFO

*JEL classification:*
G11
G14
G23
D82
D86

*Keywords:*
Agency conflict
Block trading
Contracting
Dealer–client relationship
Price impact

ABSTRACT

We study the optimal execution problem in a principal–agent setting. A client contracts to purchase from a dealer. The dealer hedges, buying from the market, creating temporary and permanent price impact. The client chooses a contract, which specifies payment as a function of market prices; hidden action precludes conditioning on the dealer's hedging trades. We show the first-best benchmark is theoretically achievable with an unrestricted contract set. We then consider weighted-average-price contracts, which are commonly used. In the continuous-time limit, the optimal weighting entails a constant density at interior times and discrete masses at the extremes.

## 1. Introduction

When trading large volumes in financial markets, two frictions play important roles: price impact and agency conflicts. Owing to price impact, it is typically desirable to split a larger 'parent order' into a number of smaller 'child orders' rather than to trade all at once. Determining precisely how to create that split is a complex problem, as one must assess how each child order will affect prices obtained for future child orders. A literature on optimal execution considers that problem, addressing how an institution ought to proceed if handling execution in house. Yet, pension funds and other institutions often outsource execution, in which case agency conflicts also arise. Although these agency conflicts are deeply appreciated by practitioners and regulators, they have so far received little attention in the literature. Analyzing a setting with both price impact and agency conflict, we show that these frictions interact in important and subtle ways.

Specifically, we model a situation in which an institution ('the client') contracts with a dealer, agreeing to conduct a block trade: a single, large off-market transaction. The dealer would then pursue offsetting trades on the market, effectively assuming the complexities of execution. It remains to determine how the block trade between the client and dealer will be priced. In practice, many trading arrangements use some weighted average of the market prices prevailing over the execution window. For example, block trades are often priced at the time-weighted average price (TWAP) prevailing on the market, as in a *guaranteed TWAP* contract, or at the price prevailing at the end of the execution window,

[☆] Nikolai Roussanov was the editor for this article. An earlier version was circulated under the title "Principal Trading Arrangements: Optimality under Temporary and Permanent Price Impact." For helpful comments, we thank Torben Andersen, Sandeep Baliga, Eric Budish, Georgy Egorov, Péter Eső, Timothy Feddersen, Michael Fishman, Lewen Guo (discussant), John Kuong (discussant), Nicolas Lambert, W. Scott Langford (discussant), Peter Klibanoff, James Schummer, Andrzej Skrzypacz, Duane Seppi, Chester Spatt, Alireza Tahbaz-Salehi, Dimitri Vayanos, Yajun Wang (discussant), Haoxiang Zhu, and various seminar participants, as well as an anonymous referee. We also thank a number of industry participants for sharing their views. Markus Baldauf and Christoph Frei gratefully acknowledge support by the Social Sciences and Humanities Research Council of Canada (grant numbers are IDG 430-2018-00100, IG 435-2018-0049, IG 435-2021-1142, and IG 435-2022-0507). Markus Baldauf thanks the University of Chicago, Booth School of Business, where he was visiting when large parts of this paper were written.

[*] Corresponding author.

*E-mail addresses:* baldauf@mail.ubc.ca (M. Baldauf), cfrei@ualberta.ca (C. Frei), joshua.mollner@kellogg.northwestern.edu (J. Mollner).

*URLs:* https://sites.google.com/site/mbaldauf/ (M. Baldauf), https://www.math.ualberta.ca/~cfrei/ (C. Frei), https://sites.google.com/site/joshuamollner/ (J. Mollner).

https://doi.org/10.1016/j.jfineco.2024.103901
Received 15 June 2023; Received in revised form 20 June 2024; Accepted 21 June 2024
Available online 19 July 2024

as in a *guaranteed market-on-close (MOC)* contract.[1] Because these contracts transfer some of the price risk burden onto the client, one economic justification for them is risk aversion on the part of the dealer. Indeed, it is appropriate to account for risk aversion because these trades are often large, and because dealers may be reluctant to take on risk due to regulation requiring them to hold capital in amounts corresponding to their exposure. Yet, questions remain: Are either of these common contracts optimal for the client—or at least optimal in some class? If not, how could she do better for herself?

To answer these questions, we formulate this interaction as a problem of contracting under moral hazard, with the client as the principal and the dealer as the agent. The friction is that the client cannot directly observe the offsetting on-market trades that the dealer makes, but only the realized time series of market prices. Because the dealer's trading creates price impact, market prices are signals of the dealer's actions, but only noisy ones. We first solve the model in discrete time, then characterize the continuous-time limit. Although the contract that we derive is in general neither of the commonly-used contracts mentioned above, interestingly and perhaps surprisingly, it does incorporate features of both: in the continuous-time limit, the optimal contract puts discrete weights on the initial and terminal prices, and it weights interior prices with a constant density.

These results apply to situations faced by pension funds, mutual funds, endowments, or other institutions when outsourcing execution of their large trades in fixed income, foreign exchange, or equity blocks. These large trades typically entail large transactions costs: for example, Nasdaq (2022) and SIFMA (2021) estimate institutional transaction costs for U.S. equities of around $70 billion per year, nearly all of which are attributable to price impact. Given the complexities of order execution and the sums involved, this setting is rife with potential conflict between the interests of dealers and clients. Cognizant of this, FINRA Rule 5270 prohibits dealers from trading on "non-public market information concerning an imminent block transaction", also called "front-running". However, that same rule provides an exemption "for the purpose of fulfilling, or facilitating the execution of, the customer block order" (FINRA, 2013), leaving ample scope for conflict regarding the timing of trades made for this purpose. The potential for conflict is also recognized by umbrella agreements between dealers and their institutional clients.[2] Furthermore, ample anecdotal evidence highlights that these conflicts of interest have real and sizable implications for transaction costs (e.g., Traders Magazine, 2005a,b; Bloomberg, 2020, 2022a,b, 2024a,b; The Wall Street Journal, 2022a, 2024). These transaction costs might be reduced if prevailing arrangements were modified to more closely resemble the contractual arrangements that we derive.

*Model.* The client seeks to buy a fixed amount of some security. (Symmetric results apply for selling.) At time zero, the client offers a contract to the dealer. A contract is an agreement that the client and dealer will conduct an off-market trade at time $T+1$, the price of which will be a function of market prices $(p_1,\dots,p_T)^\top$. If the dealer accepts the offered contract, then he hedges with offsetting trades on the market during the trading periods $\{1,\dots,T\}$. In modeling how these trades affect the dynamics of prices, we assume a canonical market model: essentially the baseline version of Almgren and Chriss (2001), with additive price shocks and linearity in both permanent and temporary price impact.

Mathematically, the client's problem is to choose a contract and a recommended trading strategy for the dealer to pursue subject to individual rationality and incentive compatibility constraints. The first-best benchmark is what would be optimal in lieu of the hidden-action friction, that is, if the dealer's on-market trades were observable to the client. In that case, the problem in fact reduces to a well-known optimal execution problem whose classic solution entails trading an equal amount in each period. Our main results highlight how outcomes change due to agency conflicts, as well as which contracts perform well in light of them.

*Results.* What is an optimal contract? Our main analysis optimizes over contracts that are weighted averages of market prices. This is for both analytical tractability and realism. Indeed, as a benchmark, we consider the situation without any restrictions on the contract's functional form. In that case, the client could approximate her first-best payoff arbitrarily closely via contracts that impose very large penalties with very low probability. Such contracts, however, seem impractical and especially vulnerable to model misspecification. Weighted-average-price contracts, in contrast, seem more realistic. In fact, many contracts used in practice are in this class. What is not obvious, however, is whether these commonly-used contracts are optimal in this class—and if not, what the optimum is.

Our main result for the discrete-time formulation characterizes the optimal weighted-average-price contract in terms of the parameters of the model: the market parameters that govern price impact and the dealer's degree of risk aversion. We denote this contract $\boldsymbol{\tau}^*=(\tau_1^*,\dots,\tau_T^*)^\top$, where $\tau_t^*$ represents the weight on the period-$t$ price. Although in closed form, the general formula for $\boldsymbol{\tau}^*$ is complicated and difficult to analyze. Nevertheless, numerical experimentation suggests a great deal of interesting structure (all of which is consistent with what we subsequently prove to hold in the continuous-time limit). It suggests that the optimal contract is U-shaped (i.e., $\tau_1^*\geq\tau_2^*\geq\cdots\geq\tau_{\lceil T/2\rceil}^*\leq\cdots\leq\tau_{T-1}^*\leq\tau_T^*$), with a severity that is strengthened by permanent price impact, weakened by temporary price impact, and weakened by the dealer's risk aversion. We also show formally that the optimal contract is symmetric (i.e., $\tau_t^*=\tau_{T+1-t}^*$ for all $t$). Finally, we derive a closed-form solution for the trading behavior that the dealer selects in response to the optimal weighted-average-price contract $\boldsymbol{\tau}^*$. It is described by a vector of trades $\boldsymbol{x}^*=(x_1^*,\dots,x_T^*)^\top$, where $x_t^*$ is the volume traded in period $t$.

To understand the intuition for these patterns, consider first the dealer's trading incentives. His profit is the difference between what he receives from the client (specified by the contract) and the costs of his on-market hedging trades. So, given an offered weighted-average-price contract, he can guarantee himself a profit of zero by selecting trading weights that perfectly mirror the contract weights. But he can do better by shifting some trading volume from periods with high expected prices to periods with low expected prices. Permanent price impact raises later prices relative to earlier ones and consequently generates a frontloading motive for the dealer—an incentive to select a trading strategy that differs from the offered contract by weighting early periods more heavily. This incentive to frontload is consistent with dealer behavior observed in various asset classes, including foreign exchange (Bloomberg, 2016), interest rates swaps (Risk.net, 2021), and options (Bloomberg, 2019).

Turning now to the client's problem, the optimal contract reflects a balance between two forces. On the one hand, permanent price impact leads prices to rise over the trading interval. Thus, if the dealer's trading strategy – and hence price dynamics – were fixed, the client would prefer

[1] In practice, it is sometimes more common to use the volume-weighted average price (VWAP). Our analysis does not consider this, however, as we do not have market volumes in our model. The reason is that, while we build on a canonical, off-the-shelf model of price impact, there is no analogously canonical model to specify how market volumes are determined. Within our model, TWAP is the closest analogue to VWAP.

[2] For example, regarding orders of institutional equities clients, HSBC Securities Inc. (HSI) states that "[p]rior to the execution of a guaranteed price order, HSI may establish a hedge through single or multiple trades that serve to offset HSI market risk associated with facilitating these transactions. This hedge will usually involve principal trades (possibly throughout the day) in the same security…such activity may ultimately affect the agreed guaranteed benchmark price" (HSBC, 2022). Such disclosures are standard (e.g., Goldman Sachs, 2017; Morgan Stanley, 2022).

weighting earlier periods in her contract. On the other hand, the dealer's trading strategy is endogenous. Moreover, permanent price impact means that frontloaded trading strategies raise all prices. The client would therefore prefer for the dealer to use a less frontloaded strategy, but, given the dealer's aforementioned frontloading motive, this requires the client to weight later periods in her contract. The combination of these incentives to weight early and late periods leads to a symmetric and U-shaped optimal contract. Moreover, because permanent price impact drives these incentives, it tends to strengthen the severity of both the U-shape and the dealer's ultimate frontloading. In contrast, temporary price impact and risk aversion induce other motives for the dealer and opposite effects.

Finally, we turn to the continuous-time limit of our discrete-time model. In this limit, the optimal contract takes a strikingly simple form, which can be seen as an extreme U-shape: atoms of equal mass at the two extreme times and a constant density at interior times. The dealer's best response is similarly simple: it entails the same constant density at interior times, as well as atoms at the extreme times where, reflecting his frontloading motive, the initial atom is three times the terminal atom. We also prove comparative statics for this continuous-time limit that are consistent with the aforementioned numerical experimentation for the discrete-time model. The mass at the extreme times – and hence the severity of the optimal contract's U-shape – is increasing in permanent price impact and decreasing in the dealer's risk aversion. Interestingly, temporary price impact does not affect the solution in this limit, because two opposing forces offset each other: on the one hand, temporary price impact raises prices and hence the client's costs (if the dealer's trades are held fixed), but on the other hand, temporary price impact also partially counteracts the dealer's frontloading motive, reducing the client's costs.

To quantify our findings, we perform a back-of-the-envelope calculation in which we compare our optimal contract against the two commonly-used contracts mentioned before. We argue that, for realistic parameters, transaction costs (as measured by implementation shortfall) under our optimal contract are 9.8 percent lower than those under the guaranteed TWAP contract and 40.1 percent lower than those under the guaranteed MOC contract. For a trade valued at $100 million, the cost savings could be hundreds of thousands of dollars. While we hesitate, in this paper, to precisely quantify these gains, this analysis highlights potential for substantial improvement upon the status quo, even while staying within the class of weighted-average-price contracts.

*Related literature.* A long tradition of models study contracting in financial settings. Often studied are delegated portfolio management, where the agent selects a financial portfolio (e.g., Bhattacharya and Pfleiderer, 1985; Carpenter, 2000; Buffa et al., 2022), and delegated asset management, where the agent manages capital invested in a risky asset and can secretly divert returns (e.g., DeMarzo and Fishman, 2007; Di Tella and Sannikov, 2021). In this paper, the agent performs a different financial task—namely, scheduling the execution of a large trade. These actions are unobserved by the client, and, therefore, this problem belongs to the large literature on moral hazard.[3]

Another connection is to principal–agent models in which the agent controls *when* an action is taken. For example, this is the case if the agent makes an irreversible stopping decision (e.g., Kruse and Strack, 2015; Grenadier et al., 2016) or chooses the timing of a disclosure (e.g., Curello and Sinander, 2024) or report (e.g., Madsen, 2022). Such problems also arise in the literature on revenue management (e.g., Board and Skrzypacz, 2016; Garrett, 2016), in which consumers decide when to buy.

The trading aspects of our model closely relate to the literature on optimal execution (e.g., Bertsimas and Lo, 1998; Almgren and Chriss, 2001; Obizhaeva and Wang, 2013). In that literature, a trader solves how to optimally work an order across time, taking as given an exogenously-specified 'market model' that governs how her trades affect price dynamics. Solving for the first-best benchmark of our model is equivalent to such a problem. Moreover, our specification of the market model follows the baseline cases of some of those classic models. Our derivation of the first-best trading strategy therefore replicates classic results from that literature. Nevertheless, we depart from that literature in that our primary focus is on the second-best problem, where the key friction is that the dealer's on-market trades are actions hidden from the client.

The most related paper is Baldauf et al. (2022).[4] It begins with a certain commonly-used contract (the guaranteed VWAP contract, *cf.* footnote 1), then derives conditions on the market model that would rationalize this contract as optimal. Among the conditions required for that contract's optimality is that price impact has no permanent component. This paper takes the opposite approach: it begins instead with a canonical market model that allows for both permanent and temporary price impact, then derives the optimal weighted-average-price contract. Outside of special cases, this optimum is not a commonly-used contract in itself—nevertheless, it suggests simple and useful modifications to prevailing arrangements. The key innovation is allowing for a permanent component of price impact. This makes the problem conceptually different: it becomes genuinely dynamic in the sense that the ordering of time periods matters. Allowing for permanent price impact is also valuable because of its prominent role in theoretical models (e.g., Kyle, 1985) and because of its empirical importance (Biais et al., 2005).

*Outline.* The remainder of the paper is organized as follows. Section 2 formulates the model in discrete time. Section 3 solves for the first-best benchmark. Section 4 derives a general discrete-time solution for the second-best, discusses its comparative statics, and considers several special cases. Section 5 analyzes the continuous-time limit. Section 6 concludes.

## 2. Model

Roughly speaking, our model combines a canonical model of price impact (à la Bertsimas and Lo, 1998; Almgren and Chriss, 2001) with a canonical model of contracting with hidden actions (à la Holmström, 1979). A client (the principal) offers her dealer (the agent) a contract regarding a trade between them.[5] If the dealer accepts, he prepares for the trade by acquiring an offsetting position from the market. The main friction is hidden action: the client cannot observe the dealer's precise sequence of on-market trades.

[3] Particularly related are models set in continuous time (e.g., Holmström and Milgrom, 1987; Sannikov, 2008) and, especially, analyses of the continuous-time limits of discrete-time models (e.g., Hellwig and Schmidt, 2002; Biais et al., 2007).

[4] Edelen and Kadlec (2012) study a related problem involving delegated trading. The primary difference is that they study agency trading (where the client pays the realized execution costs). The friction is that effort, which can lead to a better execution price, is unobservable to the client. In contrast, we study principal trading (where the payment is contracted in advance and need not equate to realized execution costs). The friction is that the on-market trades, which influence the contracted payment, are unobservable to the client.

[5] In modeling only a single dealer and a single client, we abstract away from the possibility that, in practice, a dealer could receive potentially offsetting orders from separate clients. We also abstract away from the question of how the client should select a dealer. See Baldauf and Mollner (2024) for an analysis of the latter issue.

### 2.1. Contracting environment

*Client.* The client needs to trade a fixed quantity of a particular security, which we normalize to a purchase of one share. She is risk-neutral.

*Dealer.* The dealer has constant absolute risk aversion (CARA), with coefficient $\lambda$. To simplify notation, we use $u(w) = -\exp(-\lambda w)$ for the dealer's utility function. In addition to the usual sources, the dealer's risk aversion might also originate from leverage/margin constraints or from constraints on equity capital, as in the literature on limits to arbitrage (surveyed by Gromb and Vayanos, 2010).

*Time.* At time 0, the client offers a contract, which the dealer either accepts or rejects. The contract specifies a price at which the client will purchase one share from the dealer at time $T + 1$. In between are a discrete number of trading periods $t \in \{1, \dots, T\}$, where $p_t$ denotes the market price in period $t$.

*Contracts.* The client can contract only on market prices.[6] In particular, she cannot contract directly on the dealer's trades. This assumption reflects the fact that on-market trading is anonymous in most settings. Moreover, the client could not reasonably ask the dealer to disclose a complete record of his trading, as that would expose the dealer's unmodeled trade secrets and trading relationships.[7] Formally, $\mathcal{T}$ will denote a (possibly strict) subset of the measurable real-valued functions with argument $\boldsymbol{p} = (p_1, \dots, p_T)^\top$. And $\tau \in \mathcal{T}$ will denote a typical contract. We discuss several candidates for $\mathcal{T}$ in Section 2.4.

### 2.2. Market model

*On-market trades.* If the dealer accepts the contract, then he must purchase the required share on the market. Letting $x_t$ denote the number of shares purchased by the dealer in period $t$, we therefore require $\sum_{t=1}^{T} x_t = 1$.

*Price dynamics.* Recalling that $p_t$ denotes the market price in period $t$, we assume the dynamics

$$p_t = p_0 + \gamma x_t + \theta \sum_{s=1}^{t} x_s + \sum_{s=1}^{t} \varepsilon_s.$$

Thus, $\theta \geq 0$ parametrizes permanent price impact,[8] $\gamma \geq 0$ temporary price impact, and $p_0$ the initial price level. Finally, $\varepsilon_s$ represents the price shock in period $s$, which we assume is an independent draw from $N(0, \sigma^2)$, where $\sigma > 0$. To avoid degenerate solutions, we assume throughout that at least one of $\theta$ and $\gamma$ is strictly positive.

*Remark* 1. This type of specification is widely used by practitioners, mainly because it is simple, tractable, and because it captures many empirical facts about markets (e.g., that liquidity is limited over the trading horizon, even when trade is for reasons other than information). This specification is, furthermore, canonical and standard in the literature. For example, it nests the basic case of Bertsimas and Lo (1998), it is nested by Gârleanu and Pedersen (2013), and it closely relates to the linear case of Almgren and Chriss (2001). Finally, although our analysis takes these price dynamics as exogenous, both these and related dynamics can be readily micro-founded (as in, e.g., Gârleanu and Pedersen, 2016; Kyle et al., 2018).[9]

*Trading strategies.* We denote by $\mathcal{F}_t$ the $\sigma$-algebra generated by the price shocks $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_t$. A *trading strategy* is a random vector $\boldsymbol{x} = (x_1, \dots, x_T)^\top$ such that $x_t$ is $\mathcal{F}_{t-1}$-measurable for all $t$ and $\sum_{t=1}^{T} x_t = 1$ almost surely. We denote the set of trading strategies by $\mathcal{X}$. Due to the measurability assumption on $\boldsymbol{x}$, $\mathcal{F}_t$ is also the $\sigma$-algebra generated by $p_1, p_2, \dots, p_t$ and therefore corresponds to the dealer's information at period $t$ available from the prices realized in previous periods.

*Remark* 2. A special class of trading strategies are those in which the dealer does not use any information in selecting his on-market trades, so that the entire trajectory of trades is determined ex ante. Such a trading strategy can also be thought of as a vector $\boldsymbol{x} \in \mathbb{R}^T$. We refer to these as the *deterministic* trading strategies.

*Remark* 3. The requirement $\sum_{t=1}^{T} x_t = 1$ precludes any net change in the dealer's inventory. The dealer merely intermediates between the client and the market, neither trading with the client out of his own inventory nor taking on a proprietary position of his own. We therefore shut down certain dealer misbehavior: the dealer's trading in our model does not meet the definition of illegal front-running, but rather that of permitted transactions for the purpose of fulfilling a client block order, under FINRA Rule 5270. This also distinguishes our model from the literature on dual trading (e.g., Röell, 1990; Fishman and Longstaff, 1992; Bernhardt and Taub, 2008), which considers dealer–client conflicts that arise if the dealer can either front-run or trade alongside a client order. Instead, our analysis focuses on conflicts pertaining to timing of the dealer's hedging trades.

[6] We therefore assume that the client can observe the full sequence of market prices. This is appropriate for modeling asset classes that are transparent, with publicly available trading data (e.g., equities). But other asset classes are more opaque, with less accessible data (e.g., foreign exchange), which precludes contracting on prices in arbitrary ways. Nevertheless, a third party with access to data – perhaps a platform or regulator – could compute a pricing benchmark and make it available for contracting. For such asset classes, our analysis can be reinterpreted to have implications for the design of those benchmarks. See Section 5.1 of Baldauf et al. (2022) for a more detailed discussion of the connection.

[7] In principle, the client might ask the dealer to disclose a *redacted* record of his trading. However, the client should not expect this to help, as the dealer could costlessly generate any disclosure. To make this formal, suppose the trades that the dealer discloses must sum to the one share that he sells to the client, i.e., $\sum_{t=1}^{T} x_t^{disclose} = 1$. And he can disclose only trades that he actually conducted. But he need not disclose a complete record of all trades he conducted. As claimed, the dealer can costlessly generate any disclosure. Indeed, suppose he wishes to pursue trading strategy $\boldsymbol{x}$, yet disclose a trading record $\boldsymbol{x}^{disclose}$. He can do this by making three child trades in each period $t$: (*i*) buy $x_t$ shares, (*ii*) buy $x_t^{disclose}$ shares, and (*iii*) sell $x_t^{disclose}$ shares. He would then disclose only the components (*ii*). Yet, because components (*ii*) and (*iii*) offset, price dynamics and the dealer's net trading are equivalent to those under $\boldsymbol{x}$.

[8] As is common in the literature, we intend "permanent" to refer to whatever price impact does not revert over the trading horizon. For example, if the trading horizon is one day, then price impact that reverts the next morning can be called permanent for our purposes, even though it does not literally last forever.

[9] Other models micro-found only the permanent component of price impact (e.g., Kyle, 1985). Accordingly, purely permanent price impact is a relevant special case (which we analyze in Section 4.3). As we will see, this special case is a theoretically interesting one also, because permanent price impact is what underlies the model's primary economic forces.

### 2.3. The client's problem

The client's problem is to choose a contract and a recommended trading strategy for the dealer to pursue subject to individual rationality and incentive compatibility constraints.[10] There is hidden action in that the client cannot directly observe the dealer's trades; hence, the contract must make the recommended trading strategy incentive compatible. Although the client observes prices, they constitute only a noisy signal of the dealer's trades because they are also affected by shocks. Mathematically, the client solves the program

$$\min_{\tau\in\mathcal{T},\boldsymbol{x}\in\mathcal{X}} \mathbb{E}[\tau(\boldsymbol{p})] \quad \text{subject to}$$

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

$$\boldsymbol{x} \in \arg\max_{\hat{\boldsymbol{x}}\in\mathcal{X}} \mathbb{E}[u(\tau(\boldsymbol{p})-\hat{\boldsymbol{x}}\cdot\boldsymbol{p})]. \tag{IC}$$

The form of the (IR) and (IC) constraints follow from the facts that the dealer's revenue (from the client) is $\tau(\boldsymbol{p})$ and his cost (from on-market trading) is $\boldsymbol{x}\cdot\boldsymbol{p}$.[11]

### 2.4. The contract set

All that remains is to specify the set of contracts over which the client optimizes.

*Unrestricted set.* A potential baseline is the setting in which a contract can be *any* measurable, real-valued function of market prices. We denote this contract set $\mathcal{T}^{all}$. Although this unrestricted set may be interesting from a theoretical perspective, it would permit contracts that are unrealistic, either in their complexity or in the severity of the punishments they impose for certain price-path realizations. Indeed, Mirrlees has observed that in classic moral hazard settings with normally-distributed noise and no restrictions on the contract's functional form, the agency friction essentially disappears, in the sense that the first-best outcome can be approximated arbitrarily closely using contracts that prescribe massive punishments for very low realizations of output (Bolton and Dewatripont, 2004, Sec. 4.3). Given the structure of our model, similar issues arise here when optimizing over an unrestricted set of contracts. See Section 3.2 for details.

*Weighted-average-price contracts.* Motivated by the unrealistic contracts that emerge when optimizing over an unrestricted contract set, it seems appropriate to impose some restrictions. In particular, our main analysis will optimize over contracts that are weighted averages of market prices. Although these weights will be nonnegative in the optimum, we do not constrain them to be. Thus, a contract can be thought of as a vector $\boldsymbol{\tau} \in \left\{(\tau_1,\ldots,\tau_T)^\top \in \mathbb{R}^T \,\middle|\, \sum_{t=1}^T \tau_t = 1\right\}$, which stipulates that the client will pay the dealer $\sum_{t=1}^T \tau_t p_t$. We denote this contract set $\mathcal{T}^{wa}$.

Although it is restrictive to focus only on contracts that are weighted averages of market prices (rather than on arbitrary functions), this does nest some important examples of commonly-used contracts: special cases include $\boldsymbol{\tau}^{TWAP} = \left(\frac{1}{T},\ldots,\frac{1}{T}\right)^\top$ and $\boldsymbol{\tau}^{MOC} = (0,\ldots,0,1)^\top$, which correspond to what are known in practice as a *guaranteed TWAP* contract and a *guaranteed MOC* contract. We interpret our analysis as a search for the best contract among those comparable in complexity to those already in use.

*Affine contracts.* Whereas our baseline analysis optimizes over the set of weighted-average-price contracts, an alternative would have been to instead optimize over the set of affine functions of market prices, which is a superset. We do this in Online Appendix A. We denote this contract set $\mathcal{T}^{affine}$. One reason it is interesting is that it contains fixed-price contracts (i.e., contracts specifying a payment that is constant in the market prices). And fixed-price contracts are interesting because, in the special case where the dealer is risk neutral, some fixed-price contract achieves the client's first-best payoff. This is for the same reason that, in classic models of contracting under moral hazard, a 'sell-the-firm' contract achieves first-best when the agent is risk-neutral. Nevertheless, the optimal affine contract is qualitatively similar in many ways to the optimal weighted-average-price contract.

## 3. First-best benchmark

### 3.1. First-best entails trading at a constant rate

Before solving the client's problem itself, we begin with the first-best benchmark. To that end, we remove the friction of hidden action—that is, we assume the client observes the dealer's trades. We also exclude any frictions that might arise from restrictions on the contract's functional form. Mathematically, a contract is – for this section only – a measurable real-valued function with argument $(\boldsymbol{p},\boldsymbol{x}) \in \mathbb{R}^T \times \mathbb{R}^T$ for $\sum_{t=1}^T x_t = 1$.[12]

In this setting, the client can implement any trading strategy she desires with a 'forcing contract' that imposes a large penalty if the dealer deviates from the recommended trading strategy. And given concavity of $u$, it is optimal to satisfy (IR) by choosing $\tau(\boldsymbol{p},\boldsymbol{x}) = \boldsymbol{x}\cdot\boldsymbol{p}$ when the dealer does follow the recommendation. Plugging $\tau(\boldsymbol{p},\boldsymbol{x}) = \boldsymbol{x}\cdot\boldsymbol{p}$ into her objective, the client's problem becomes

$$\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}[\boldsymbol{x}\cdot\boldsymbol{p}].$$

In short, solving for the first-best trading strategy reduces to a problem of optimal execution.

[10] In allowing the client to recommend a trading strategy to the dealer, this formulation follows classical models of moral hazard (e.g., Holmström, 1979). Effectively, it assumes that the client can break the dealer's indifference however she likes. This assumption is, however, irrelevant for our analysis. In particular, the dealer will have a *unique* best response to most contracts we consider (*cf.* Lemma 3).

[11] Alternatively, we can define a dynamic incentive compatibility constraint: for all $t = 1,2,\ldots,T$ and $A \in \mathcal{F}_{t-1}$ with $P[A] > 0$, $\boldsymbol{x} \in \arg\max_{\hat{\boldsymbol{x}}\in\mathcal{X}_t^{\boldsymbol{x}}} \mathbb{E}[u(\tau(\boldsymbol{p})-\hat{\boldsymbol{x}}\cdot\boldsymbol{p})|A]$ where $\mathcal{X}_t^{\boldsymbol{x}} = \{\hat{\boldsymbol{x}}\in\mathcal{X} : \hat{x}_s = x_s \text{ for } s<t\}$. This is equivalent to (IC). For details, see Online Appendix B.

[12] For situations where the second-best analysis entails optimizing over a restricted contract space $\mathcal{T} \subsetneq \mathcal{T}^{all}$, a perhaps more appropriate benchmark would be to remove the friction of hidden action while preserving any frictions due to restrictions on functional form. Mathematically, a contract (for this benchmark only) would be a measurable real-valued function with argument $(\boldsymbol{p},\boldsymbol{x}) \in \mathbb{R}^T \times \mathbb{R}^T$ for $\sum_{t=1}^T x_t = 1$, where for any fixed $\boldsymbol{x}$, $\tau(\cdot,\boldsymbol{x}) \in \mathcal{T}$. If $\mathcal{T} \supseteq \mathcal{T}^{wa}$, then this benchmark coincides with what we derive below. One way to achieve the first-best in this setting is with a contract that, for any $\boldsymbol{x}$, simply pays the dealer a weighted average of the market prices $\boldsymbol{p}$ with weights that exactly mirror $\boldsymbol{x}$. Under such a contract, $\tau(\boldsymbol{p},\boldsymbol{x}) = \boldsymbol{x}\cdot\boldsymbol{p}$ for all $\boldsymbol{x}$ and $\boldsymbol{p}$, meaning that (IC) is satisfied (albeit only with equality) by rendering the dealer indifferent over all trading policies.

In fact, because our market model is essentially the baseline case considered by Almgren and Chriss (2001), their results apply to this problem. The first-best solution corresponds to what they would derive as the optimal trading strategy in the case where all weight is put on minimizing the mean of implementation shortfall (and no weight on the variance). The classic result (also obtained by others, e.g., Bertsimas and Lo, 1998) is that under these baseline conditions, the optimal strategy is to trade an equal amount in each period. We therefore have the following result:

**Proposition 1.** *The first-best trading strategy is*

$$\boldsymbol{x}^{FB} = \left(\frac{1}{T}, \dots, \frac{1}{T}\right)^{\top}.$$

*In the first best, the client's expected costs of execution are* $p_0 + \frac{\gamma}{T} + \frac{\theta(T+1)}{2T}$.

For completeness, we also include a proof of this classic result in Appendix A.

*Remark* 4. Note that this first-best trading strategy is deterministic; that is, the entire trajectory of trades is determined ex ante. As we will see with Lemma 3 in the next section, an analogous result holds for the second-best problem.

*Remark* 5. In the same way that canonical contracting models take it as an exogenous constraint that the principal cannot herself perform the agent's action, we assume that the client cannot access the market and directly implement $\boldsymbol{x}^{FB}$.[13] Several considerations justify this approach. First, the client might lack the dealer's infrastructure, including market access, order-handling capabilities, low-latency trading technology, subscriptions to exchange direct data feeds, risk management, trade reporting, and compliance—all of which require substantial fixed-cost investments. The client might also lack the dealer's qualification for volume-based discounts on exchange fees. Second, on-market trading might be more complex than its reduced-form representation in our model (e.g., it might entail order splitting across multiple venues in each period), so that optimal trading might not be as simple as the expression for $\boldsymbol{x}^{FB}$ suggests. Rather, optimal trading might depend on specialized knowledge of market structure, which the dealer is more likely to possess. Relative to a typical dealer, a typical client simply will not have invested nearly as many resources into developing or fine-tuning smart-order routers and trading algorithms. One way to cast this idea within the language of the model is to suppose that when the dealer trades, he creates price impact according to the dynamics described above (with price impact coefficients $\gamma$ and $\theta$), but if the client were to trade directly on the market, she would do so less efficiently (with price impact coefficients $\gamma^{client} > \gamma$ and $\theta^{client} > \theta$).

### *3.2. First-best is achievable with an unrestricted contract set*

Having solved for the first-best benchmark, we now turn to the second-best problem (as formulated in Section 2.3). Our first observation is that when the contract set is $\mathcal{T}^{all}$, the client can approximate her first-best payoff arbitrarily well. However, doing so may require contracts that are complex, impractical, and unrealistic.

*A special case.* To illustrate, consider the special case of $\gamma = 0$, $\theta = 1$, $\sigma = 1$, and $T = 2$ periods. In this special case, we have $p_1 - p_0 = x_1 + \varepsilon_1$, which is normally distributed with mean $x_1$ and unit variance. This suggests a contract scheme inspired by arguments Mirrlees has made in settings of classic moral hazard: (*i*) pay the dealer the TWAP, $\frac{1}{2}(p_1 + p_2)$, (*ii*) pay also a flat commission $\delta$, but (*iii*) also impose a penalty of $K$ if $p_1 - p_0$ exceeds a threshold $\zeta$. As derived above, first-best entails a choice of $x_1 = \frac{1}{2}$. Without the penalty, the dealer would frontload his trading by choosing $x_1 > \frac{1}{2}$. (Proposition 4 will show this formally.) The penalty, however, counteracts this frontloading motive. Of course, a dealer who does not frontload is also exposed to the penalty—which is why $\delta > 0$ is needed to ensure that $x_1 = \frac{1}{2}$ is compatible with (IR). But the key is that normal distributions have the property that extreme right-tail realizations are much more likely under a higher mean relative to a lower mean. Consequently, by setting $\zeta$ and $K$ both very large, the penalty can be designed to deter frontloading by imposing a relatively large risk on a dealer who frontloads by choosing $x_1 > \frac{1}{2}$ yet at the same time virtually no risk on a dealer who chooses $x_1 = \frac{1}{2}$ (which permits $\delta$ to be set very close to zero).

Formally, we would proceed as follows in this special case. Consider a class of contracts indexed by $(\zeta, \delta, K)$, specifying payment

$$\tau(\boldsymbol{p}) = \delta + \frac{1}{2}(p_1 + p_2) - K\mathbb{1}[p_1 - p_0 > \zeta].$$

Using $x_2 = 1 - x_1$, the dealer's monetary payoff is

$$\tau(\boldsymbol{p}) - \boldsymbol{x} \cdot \boldsymbol{p} = \delta + \left(\frac{1}{2} - x_1\right)(p_1 - p_2) - K\mathbb{1}[p_1 - p_0 > \zeta] = \delta + \left(\frac{1}{2} - x_1\right)(x_1 - 1 - \varepsilon_2) - K\mathbb{1}[\varepsilon_1 > \zeta - x_1].$$

We then compute the dealer's expected utility, differentiate with respect to $x_1$, and evaluate at $x_1 = \frac{1}{2}$ to obtain

$$\frac{d}{dx_1}\log\big(-\mathbb{E}\big[u\big(\tau(\boldsymbol{p}) - \boldsymbol{x} \cdot \boldsymbol{p}\big)\big]\big)\bigg|_{x_1 = 1/2} = -\frac{\lambda}{2} + \frac{\phi(\zeta - 1/2)}{\frac{1}{1 - \exp(-\lambda K)} - \Phi(\zeta - 1/2)},$$

where $\Phi$ and $\phi$ denote the standard normal CDF and pdf. See Online Appendix C for details.

For this contract to induce first-best trading, $x_1 = \frac{1}{2}$ must satisfy (IC), hence this derivative must equal zero. For all sufficiently large $\zeta$, there exists a $K$ that does this. Formally, choose $\bar{\zeta}$ to ensure $\frac{\phi(\bar{\zeta} - 1/2)}{1 - \Phi(\bar{\zeta} - 1/2)} > \frac{\lambda}{2}$; this is possible because the inverse Mills ratio $\frac{\phi(x)}{1 - \Phi(x)}$ grows without bound as $x$ increases. Let $\zeta \geq \bar{\zeta}$. Setting the above derivative equal to zero and solving for $K$, we obtain

$$K = \frac{1}{\lambda}\log\left(1 + \frac{\lambda/2}{\phi(\zeta - 1/2) - \frac{\lambda}{2}\left[1 - \Phi(\zeta - 1/2)\right]}\right), \tag{1}$$

[13] For a model of endogenous choice between doing a block trade with a dealer versus working an order on the market, see Seppi (1990).

which is well-defined because $\zeta \geq \bar{\zeta}$ ensures that $\frac{\phi(\zeta-1/2)}{1-\Phi(\zeta-1/2)} \geq \frac{\phi(\bar{\zeta}-1/2)}{1-\Phi(\bar{\zeta}-1/2)} > \frac{\lambda}{2}$. Given this value for $K$, choosing $x_1 = \frac{1}{2}$ leaves the dealer with expected utility

$$\mathbb{E}\big[u\big(\tau(\boldsymbol{p}) - \boldsymbol{x} \cdot \boldsymbol{p}\big)\big]\Big|_{x_1=1/2} = -\frac{\exp(-\lambda\delta)}{1 - \frac{\lambda}{2}\frac{1-\Phi(\zeta-1/2)}{\phi(\zeta-1/2)}}.$$

Again, see Online Appendix C for details.

For this contract to satisfy (IR), the dealer's expected utility must be at least $u(0) = -1$. Choosing $\delta$ to make this constraint hold with equality, we obtain

$$\delta = -\frac{1}{\lambda}\log\left[1 - \frac{\lambda}{2}\frac{1-\Phi(\zeta-1/2)}{\phi(\zeta-1/2)}\right]. \tag{2}$$

Finally, note that the client can approximate her first-best payoff arbitrarily closely by setting $\zeta \geq \bar{\zeta}$ sufficiently large, setting $K$ according to (1), and setting $\delta$ according to (2). Indeed, because $\delta$ is set to satisfy (IR), the dealer accepts such a contract. The client's expected costs of execution $\mathbb{E}\big[\tau(\boldsymbol{p})\big]$ are bounded above by $\mathbb{E}\big[\delta + \frac{1}{2}(p_1 + p_2)\big] = \delta + p_0 + \frac{1}{2}(x_1 + 1)$. Because $K$ is set so that $x_1 = \frac{1}{2}$ satisfies (IC), this bound becomes $\delta + p_0 + \frac{3}{4}$. Applying Proposition 1, the client's first-best costs are exceeded by at most $\delta$. However, according to (2), $\delta$ can be made arbitrarily close to zero by setting $\zeta$ sufficiently large.

As we demonstrate in the proof of the following result, this construction can be extended to accommodate the model in its full generality.

**Proposition 2.** *When the set of feasible contracts is $\mathcal{T}^{all}$, the client's expected costs of execution can be made arbitrarily close to her first-best costs of $p_0 + \frac{\gamma}{T} + \frac{\theta(T+1)}{2T}$.*

*Discussion.* Although we have derived contracts that, within this model, permit the client to approximate her first-best payoff arbitrarily closely, we do not view these contracts as realistic or practical. One issue is that the penalty size $K$ and penalty threshold $\zeta$ must be fine-tuned to the economic environment with an extraordinary degree of accuracy, leaving these contracts extremely susceptible to model misspecification.[14] Moreover, to our knowledge, these contracts do not bear even a remote resemblance to arrangements used in practice, which suggests that they might be prohibitively complex to write or to enforce.

We therefore think it more meaningful to search for the optimum within a restricted contract set. As to how, precisely, to restrict the contract set, it may be useful to return to the question of which arrangements are used in practice. Many contracts use some weighted average of the market prices within the execution window (e.g., the guaranteed TWAP and guaranteed MOC contracts mentioned earlier). This motivates us to optimize over weighted-average-price contracts, to see if we can improve on contracts that are in use while staying within this class of relatively simple contracts.[15]

## 4. Discrete-time solution

We now turn to the optimal weighted-average-price contract. Although we are predominantly interested in the continuous-time limit, we find it helpful to begin by deriving the general discrete-time solution and discussing its features. To that end, we consider several special cases, which illuminate the economic forces underpinning the comparative statics of this general solution.

### *4.1. The dealer's best response*

Our first step in solving the client's problem is to note that the (IR) constraint can be eliminated. Indeed, for any weighted-average-price contract $\boldsymbol{\tau} \in \mathcal{T}^{wa}$, the dealer can guarantee himself the payoff $u(0)$ by selecting the deterministic trading strategy $\boldsymbol{x} = \boldsymbol{\tau}$. Intuitively, under this choice of $\boldsymbol{x}$, the dealer's costs (his on-market trades) are the same weighted average of market prices that determines his revenue (his payment from the client). He would then obtain a profit of zero—regardless of the realized price shocks. Because the dealer can in this way always guarantee himself his outside option, it follows that (IC) actually implies (IR).

Our second step is to characterize the (IC) constraint. Lemma 3 states that, for any contract $\boldsymbol{\tau} \in \mathcal{T}^{wa}$, the dealer has a unique best response. Thus, (IC) simply requires the recommended trading strategy to be this best response.

**Lemma 3.** *For any contract $\boldsymbol{\tau} \in \mathcal{T}^{wa}$, the dealer has a unique best response in $\mathcal{X}$, which is the deterministic trading strategy $\boldsymbol{x} = FA^{-1}E\boldsymbol{\tau}$, where $A$, $E$, and $F$ are $T \times T$ matrices explicitly given in terms of the parameters $\lambda$, $\theta$, $\gamma$ and $\sigma$, as stated in Eq. (A.8) of Appendix A.*

For the special case of $T = 2$ periods, Lemma 3 simplifies considerably. Using $\tau_2 = 1 - \tau_1$, the dealer's best response is

$$x_1 = \tau_1 + \frac{(1-\tau_1)\theta + (1-2\tau_1)\gamma}{\lambda\sigma^2 + 2\theta + 4\gamma},$$

[14] Discussing essentially the same point, Holmström and Milgrom (1987) write:

> [These] near-optimal [contracts] are disturbing. In practice, one feels, schemes that adjust compensation only when rare events occur are not likely to provide correct incentives for the agent in ordinary, high probability circumstances. Moreover, to construct the scheme, the principal requires very precise knowledge about the agent's preferences and beliefs, and about the technology that he controls. The [. . . ] scheme performs ideally if the model's assumptions are precisely met, but can be made to perform quite poorly if small deviations in the assumptions about the variance or (especially) about the agent's ability to control the probability of rare events are introduced. (p. 305)

[15] Studying a different problem – how to formulate a manipulation-resistant benchmark price from a set of transactions – Duffie and Dworczak (2021) take a related approach, restricting attention to benchmarks that are weighted averages of transaction prices.

and $x_2 = 1 - x_1$. We can see the following in this special case: (*i*) $\lim_{\theta\to\infty} x_1 = \frac{1}{2}(\tau_1 + 1)$, which for $\tau_1 \in [0,1]$ exceeds $\tau_1$, which we interpret to mean that permanent price impact creates a frontloading motive for the dealer; (*ii*) $\lim_{\gamma\to\infty} x_1 = \frac{1}{2}\left(\tau_1 + \frac{1}{2}\right)$, which we interpret to mean that temporary price impact creates a smoothing motive for the dealer; and (*ii*) $\lim_{\lambda\to\infty} x_1 = \tau_1$, which we interpret to mean that risk aversion creates a mirroring motive for the dealer. Sections 4.3–4.5 show that these forces generalize beyond the special case of $T = 2$.

According to the lemma, the dealer's best response is in fact a deterministic trading strategy. To see the intuition, suppose that after accepting a contract $\boldsymbol{\tau} \in \mathcal{T}^{wa}$, the dealer makes a tentative plan to pursue a particular deterministic trading strategy $\boldsymbol{x}$. After implementing $x_1$, the dealer observes $p_1$, which reveals the realized $\varepsilon_1$. Would he want to re-optimize $(x_2, \dots, x_T)$? The answer is no. The intuition is that $\varepsilon_1$ affects not only $p_1$ but also every future price. Then, given that the dealer's revenue $\boldsymbol{\tau}\cdot\boldsymbol{p}$ and costs $\boldsymbol{x}\cdot\boldsymbol{p}$ are both weighted averages of the prices, $\varepsilon_1$ does not affect his terminal wealth, so learning it is irrelevant. More generally, suppose that after implementing $x_t$, the dealer learns $\varepsilon_t$. Would he want to re-optimize $(x_{t+1}, \dots, x_T)$? Again, no. This is because $\varepsilon_t$ shifts the dealer's terminal wealth by the constant $\varepsilon_t \sum_{s=1}^{t-1}(x_s - \tau_s)$:

$$\sum_{t=1}^{T}(\tau_t - x_t)p_t = \sum_{t=1}^{T}(\tau_t - x_t)\left(p_0 + \gamma x_t + \theta\sum_{s=1}^{t} x_s + \sum_{s=1}^{t}\varepsilon_s\right) = \sum_{t=1}^{T}\varepsilon_t \underbrace{\sum_{s=t}^{T}(\tau_s - x_s)}_{=\sum_{s=1}^{t-1}(x_s-\tau_s)} + \sum_{t=1}^{T}(\tau_t - x_t)\left(p_0 + \gamma x_t + \theta\sum_{s=1}^{t} x_s\right).$$

And because the dealer has CARA utility, this constant shift in the distribution of his terminal wealth does not affect his preferences over his remaining choices $(x_{t+1}, \dots, x_T)$.[16]

One implication of Lemma 3 is that the (IC) constraint generally renders the first-best unachievable. Indeed, it follows from the analysis in Section 3 that if the client could choose a contract and a recommended trading policy free of the (IC) constraint, then she would select $\boldsymbol{\tau}^{TWAP}$ and $\boldsymbol{x}^{FB}$, both of which are the equally-weighted vectors $\left(\frac{1}{T}, \dots, \frac{1}{T}\right)^\top$. These choices implement the efficient action, while also leaving the dealer perfectly insured and with zero surplus. Unfortunately for the client, it is not generally true that $FA^{-1}E\boldsymbol{\tau}^{TWAP} = \boldsymbol{x}^{FB}$, so that by Lemma 3, these choices are inconsistent with (IC). In particular, inequality obtains whenever $\theta > 0$, and the departure from equality has a particular structure: it is frontloaded in the sense of first-order stochastic dominance.

**Definition.** Given two $T$-dimensional vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, each of whose elements sum to 1, $\boldsymbol{a}$ is *frontloaded relative to* $\boldsymbol{b}$ if $\sum_{s=1}^{t} a_s \geq \sum_{s=1}^{t} b_s$ for all $t$.

**Proposition 4.** *$FA^{-1}E\boldsymbol{\tau}^{TWAP}$ is frontloaded relative to $\boldsymbol{x}^{FB}$, with equality iff $\theta = 0$.*

Proposition 4 implies that permanent price impact creates a frontloading motive for the dealer, in the sense that if $\theta > 0$, then the client cannot obtain her first-best payoff. The intuition is as follows. Suppose the dealer is offered $\boldsymbol{\tau}^{TWAP}$. If he selects $\boldsymbol{x}^{FB}$, then his trading costs and his payment from the client are both a simple average of prices, so that regardless of the realized prices, he is guaranteed a profit of zero. But he can do better by frontloading his trading. Suppose the dealer deviates from $\boldsymbol{x}^{FB}$ by shifting a small amount of volume from a period $t''$ to a period $t' < t''$. The direct effect of this deviation (i.e., holding prices fixed) is to reduce his expected costs at the rate $\mathbb{E}[p_{t''}] - \mathbb{E}[p_{t'}] = \theta\sum_{t=t'+1}^{t''} x_t^{FB} + \gamma(x_{t''}^{FB} - x_{t'}^{FB}) = \theta\frac{t''-t'}{T}$. And the indirect effect (through price changes) is vanishing, because as we have noted, $\boldsymbol{x}^{FB}$ insures the dealer against price fluctuations. Moreover, the effect on the variance of his profit is second-order for small deviations. It follows that when $\theta > 0$, some sufficiently small deviation from $\boldsymbol{x}^{FB}$ allows the dealer to make himself better off. In general, the dealer's best response trades off this incentive to frontload against increased exposure to price shocks and excess temporary-impact costs.

In the special case of $\theta = 0$, this incentive to frontload does not arise, and the proposition implies that the first-best outcome is in fact achieved via $\boldsymbol{\tau}^{TWAP}$, the (IC) constraint notwithstanding.[17] Hence, our model predicts that $\boldsymbol{\tau}^{TWAP}$ might yield reasonably good outcomes when applied to settings or securities for which permanent price impact is relatively small. However, when permanent price impact is a major factor then $\boldsymbol{\tau}^{TWAP}$ ought not be expected to perform as well, which is what motivates the subsequent analysis.

### *4.2. The general solution*

Having eliminated the (IR) constraint and characterized the (IC) constraint, the client's problem reduces to

$$\min_{\boldsymbol{\tau}\in\mathcal{T}^{wa}} \mathbb{E}[\boldsymbol{\tau}\cdot\boldsymbol{p}] \quad \text{subject to} \quad \boldsymbol{x} = FA^{-1}E\boldsymbol{\tau}.$$

Our next result concerns the solution to this problem. It provides explicit formulas for the optimal weighted-average-price contract $\boldsymbol{\tau}^*$ (henceforth, simply the "optimal contract") and the incentive-compatible trading strategy $\boldsymbol{x}^* = FA^{-1}E\boldsymbol{\tau}^*$ that the client recommends to the dealer. The formulas are complicated, but they are fully explicit and easy to compute.

**Proposition 5.** *The weights of the optimal contract and the dealer's best-responding trading strategy are given by $\boldsymbol{\tau}^* = \frac{1}{\mathbb{1}^\top M^{-1}\mathbb{1}}M^{-1}\mathbb{1}$ and $\boldsymbol{x}^* = \frac{1}{\mathbb{1}^\top M^{-1}\mathbb{1}}FA^{-1}EM^{-1}\mathbb{1}$, where $\mathbb{1} = (1, 1, \dots, 1)^\top$ denotes a $T$-dimensional vector of ones and*

$$M = \theta A^{-1}E + \theta E^\top(A^{-1})^\top + \gamma FA^{-1}E + \gamma E^\top(A^{-1})^\top F^\top. \tag{3}$$

*The client's expected costs of execution are $p_0 + \frac{1}{2\mathbb{1}^\top M^{-1}\mathbb{1}}$.*

[16] Several model components therefore combine to imply optimality of a deterministic trading policy. For example, the dealer's best response might not be deterministic if he had non-constant absolute risk aversion, if he were facing a nonlinear contract, or if the random walk component of the price process were instead an AR(1).

[17] This aspect of the result and the economic forces behind it are similar to why the guaranteed VWAP contract – under which the client pays the dealer at the market's volume-weighted average price (VWAP) – is optimal in the setting of Baldauf et al. (2022). One subtlety is that in that paper, each trading period has an associated 'market condition,' about which the dealer has superior information. The optimal contract weights prices by market volume so as to incentivize the dealer to properly trade on his information about market conditions. In contrast, this paper uses a canonical market model in which such market conditions do not feature (or do not differ across periods). Hence, the optimal contract need not weight by volumes, and a simple average of prices achieves the optimum.

To establish this as the optimal contract, the proof shows that the client's expected payment under a contract $\boldsymbol{\tau} \in \mathcal{T}^{wa}$ can be expressed as $\frac{1}{2}\boldsymbol{\tau}^\top M \boldsymbol{\tau}$. By symmetry of $M$, the optimal contract weights satisfy $M\boldsymbol{\tau}^* = \mu\mathbb{1}$, where $\mu$ is the Lagrange multiplier on the constraint $\boldsymbol{\tau}^\top \mathbb{1} = 1$. The constraint then implies $\boldsymbol{\tau}^* = \frac{1}{\mathbb{1}^\top M^{-1}\mathbb{1}} M^{-1}\mathbb{1}$, as the proposition says. Applying Lemma 3, we obtain $\boldsymbol{x}^* = \frac{1}{\mathbb{1}^\top M^{-1}\mathbb{1}} F A^{-1} E M^{-1}\mathbb{1}$.

The problem and its solution are mathematically complex, and so it is difficult to provide intuition for the precise form of the general solution exhibited in Proposition 5. Nevertheless, the logic of the solution can be explained through three special cases: (*i*) when permanent price impact is the only influence (i.e., $\theta \to \infty$), (*ii*) when temporary price impact is the only influence (i.e., $\gamma \to \infty$), and (*iii*) when price risk is the only influence (i.e., $\lambda \to \infty$).[18] We next consider each of these special cases in turn, then build upon them to explain the features and the comparative statics of the general solution.

### *4.3. When permanent price impact is the only influence*

For the case in which permanent price impact is the only influence, assume away temporary price impact, and make the dealer risk-neutral.

**Corollary 6.** *Assume that there is no temporary price impact ($\gamma = 0$) and that the dealer is risk-neutral ($\lambda = 0$).*

*(i) For any $\boldsymbol{\tau} \in \mathcal{T}^{wa}$, the dealer's best response is $x_t = \frac{1}{T} - \sum_{s=1}^{T} \frac{s}{T}\tau_s + \sum_{s=t}^{T} \tau_s$.*

*(ii) The weights of the optimal contract are $\boldsymbol{\tau}^* = \left(\frac{1}{2}, 0, \dots, 0, \frac{1}{2}\right)^\top$, so that the dealer's best-responding trading strategy is $\boldsymbol{x}^* = \left(\frac{T+1}{2T}, \frac{1}{2T}, \dots, \frac{1}{2T}\right)^\top$. The client's expected costs of execution are $p_0 + \frac{\theta(3T+1)}{4T}$.*

Let us highlight two differences relative to the first-best solution given by Proposition 1. First, the client must now respect the (IC) constraint, which implies a particular relationship between the contract and the dealer's trading strategy that in this case is given by Corollary 6*(i)*. Second, the client must now sacrifice some rent to the dealer. As a result, her expected costs increase from $p_0 + \frac{\theta(T+1)}{2T}$ to $p_0 + \frac{\theta(3T+1)}{4T}$.

For the discussion below, let us assume that both contract weights $\boldsymbol{\tau}$ and the dealer's trading strategy $\boldsymbol{x}$ are restricted to entail nonnegative weights. This is only for simplicity of the exposition. Indeed, given Corollary 6*(ii)*, this restriction does not bind, and in fact, many of the arguments below could be formulated in reverse to rule out putative solutions entailing negative weights. With this in hand, we proceed by backward induction:

*The dealer's problem.* Consider how the dealer would respond to an arbitrary contract $\boldsymbol{\tau} \in \mathcal{T}^{wa}$. Beginning from any deterministic trading strategy $\boldsymbol{x}$, consider a perturbation that shifts volume from $x_{t+1}$ to $x_t$. The dealer's expected profit $\mathbb{E}[\boldsymbol{\tau} \cdot \boldsymbol{p} - \boldsymbol{x} \cdot \boldsymbol{p}]$ is affected in two ways:

- *Direct effect (holding prices fixed).* The direct effect is positive: $\mathbb{E}[p_{t+1}] - \mathbb{E}[p_t] = \theta x_{t+1}$. Intuitively, prices tend to increase over time because of permanent price impact of the dealer's trades. Thus, if prices were held fixed, the dealer would reduce the cost of his on-market trading by shifting volume to earlier periods.
- *Indirect effect (through price changes).* Of course, prices will not hold fixed. In particular, this shift affects $\mathbb{E}[p_t]$, creating the following indirect effect: $(\tau_t - x_t)\Big(\underbrace{\frac{\partial \mathbb{E}[p_t]}{\partial x_t}}_{=\theta} - \underbrace{\frac{\partial \mathbb{E}[p_t]}{\partial x_{t+1}}}_{=0}\Big) = (\tau_t - x_t)\theta$.[19]

Note that if $x_t = \tau_t$, then the indirect effect is zero – intuitively, the dealer is perfectly insured with respect to $p_t$ if $x_t = \tau_t$ – leaving the positive direct effect to dominate. It follows that the optimal $x_t$ must exceed $\tau_t$. This argument applies for any $t < T$, implying that the dealer has a *frontloading motive* in this case: his best response is to choose an $\boldsymbol{x}$ that is frontloaded relative to the offered $\boldsymbol{\tau}$.[20]

Summing both effects, the total derivative is $\theta[x_{t+1} + \tau_t - x_t]$. Thus, if $\boldsymbol{x}$ best responds to $\boldsymbol{\tau}$, we must have $x_{t+1} = x_t - \tau_t$ for all $t < T$. Having assumed that all entries of $\boldsymbol{\tau}$ are nonnegative, we conclude from these first-order conditions that $(x_t)_{t=1}^T$ is a weakly decreasing sequence. These conditions moreover imply

$$x_t = \frac{1}{T} - \sum_{s=1}^{T} \frac{s}{T}\tau_s + \sum_{s=t}^{T} \tau_s, \tag{4}$$

as claimed by Corollary 6*(i)*.

*The client's problem.* For intuition into why $\boldsymbol{\tau}^* = \left(\frac{1}{2}, 0, \dots, 0, \frac{1}{2}\right)^\top$ is optimal in this case of a risk-neutral dealer and no temporary price impact, we first explain why the optimal contract puts weight only on the extremal prices. Starting from an arbitrary $\boldsymbol{\tau}$, consider a perturbation that implements a mean-preserving spread of the contract weights. Both the direct and indirect effects of this perturbation benefit the client:

- *Direct effect (holding prices fixed).* As mentioned while analyzing the dealer's problem, $(x_t)_{t=1}^T$ is a weakly decreasing sequence. As a positive affine transformation of the partial sums of $(x_t)_{t=1}^T$, $(\mathbb{E}[p_t])_{t=1}^T$ is therefore a weakly concave sequence. Thus, if the prices were held fixed, the client's payment would be weakly lower under a mean-preserving spread of $\boldsymbol{\tau}$.

[18] Case (*i*) is equivalent to what obtains if $\theta > 0$ and $\gamma = \lambda = 0$. Likewise, case (*ii*) is equivalent to what obtains if $\gamma > 0$ and $\theta = \lambda = 0$. Because the exposition is simpler if limits are avoided, this is what Sections 4.3 and 4.4 consider. On the other hand, Section 4.5 does treat the limiting case of $\lambda \to \infty$. Although $\lambda > 0$ and $\theta = \gamma = 0$ leads to the same dealer's best response, it does not lead to a unique optimal contract, as without price impact, all contracts lead to identical outcomes for the dealer. Mathematically, if $\theta = \gamma = 0$, the matrix $M$ in (3) is the zero matrix so that the inverse (needed in the formula for $\tau^*$ in Proposition 5) is not well defined. Thus, Section 4.5 maintains the assumption that at least one of $\theta$ and $\gamma$ is strictly positive, instead considering the limit as $\lambda \to \infty$.

[19] There are no other indirect effects: (*i*) this shift does not affect the earlier prices $p_1, \dots, p_{t-1}$; and (*ii*) because price impact is purely permanent, it also does not affect the later prices $p_{t+1}, \dots, p_T$.

[20] This is for any $\boldsymbol{\tau}$ and is therefore a stronger conclusion than that of Proposition 4, which applies only if the offered contract is $\boldsymbol{\tau}^{TWAP}$. On the other hand, Proposition 4 holds for general parameters, whereas this section specializes to the case of $\gamma = \lambda = 0$.

- *Indirect effect (through price changes).* Of course, prices will not hold fixed, as a change in $\boldsymbol{\tau}$ changes the dealer's best response $\boldsymbol{x}$, affecting price dynamics. Using (4), we compute

$$\sum_{s=1}^{t} x_s = \frac{t}{T} + \sum_{s=1}^{t} s\left(1-\frac{t}{T}\right)\tau_s + \sum_{s=t+1}^{T} t\left(1-\frac{s}{T}\right)\tau_s. \tag{5}$$

  Observe that for all $t$, the coefficients on $(\tau_1,\dots,\tau_T)^\top$ in this expression form a weakly concave sequence.[21] Thus, a mean-preserving spread of $\boldsymbol{\tau}$ leads the dealer to backload his trading, in the sense of first-order stochastic dominance (i.e., reduce $\sum_{s=1}^{t} x_s$ for all $t$). Given that $\mathbb{E}[p_t] = p_0 + \theta\sum_{s=1}^{t} x_s$, such backloading weakly reduces each price, which benefits the client.

The client therefore unambiguously benefits from mean-preserving spreads of the contract weights. It follows that $\tau_2^* = \cdots = \tau_{T-1}^* = 0$, so that the optimal contract is a U-shape.[22] To see that it is also symmetric, begin from an arbitrary contract whose interior weights are all zero and consider a perturbation that shifts weight from period $T$ to period 1. Unlike the mean-preserving perturbation considered above, here the direct and indirect effects have opposite signs:

- *Direct effect (holding prices fixed).* On the one hand, owing to permanent price impact, prices are expected to rise over the trading interval. Thus, if the dealer's trading strategy – and hence price dynamics – were fixed, the client would put full weight on the first period.
- *Indirect effect (through price changes).* On the other hand, price dynamics will respond to the contract. All else equal, the client prefers low prices. Owing to permanent price impact, each price is lowest when the dealer backloads his trading as much as possible. Given the contract's influence on the dealer's trading strategy (i.e., that his trading will be frontloaded relative to the contract), prices are lowest when the client puts full weight on the last period.

The optimal contract must balance these two considerations. Due to the linearity of price impact, these two effects offset when $\tau_1 = \tau_T = \frac{1}{2}$.[23]

### 4.4. When temporary price impact is the only influence

For the case in which temporary price impact is the only influence, assume away permanent price impact, and make the dealer risk-neutral. In this case, the client optimally offers the guaranteed TWAP contract, which weights each period equally. It induces the dealer to use the first-best trading strategy, which similarly puts equal weight on each period. And in this case, the client obtains her first-best payoff.

**Corollary 7.** *Assume that there is no permanent price impact ($\theta = 0$) and that the dealer is risk-neutral ($\lambda = 0$).*

*(i) For any $\boldsymbol{\tau} \in \mathcal{T}^{wa}$, the dealer's best response is $\boldsymbol{x} = \frac{1}{2}\boldsymbol{\tau} + \frac{1}{2}\left(\frac{1}{T},\dots,\frac{1}{T}\right)^\top$.*

*(ii) The weights of the optimal contract are $\boldsymbol{\tau}^* = \left(\frac{1}{T},\dots,\frac{1}{T}\right)^\top$, so that the dealer's best-responding trading strategy is $\boldsymbol{x}^* = \left(\frac{1}{T},\dots,\frac{1}{T}\right)^\top$. The client's expected costs of execution are $p_0 + \frac{\gamma}{T}$.*

Claim *(i)* says that the dealer has a *smoothing motive* in this case: his best response is to choose an $\boldsymbol{x}$ that partially smooths the offered $\boldsymbol{\tau}$. To understand this, suppose that the dealer did not smooth at all, selecting the trading strategy $\boldsymbol{x} = \boldsymbol{\tau}$. His trading costs and his payment from the client would both therefore be the same weighted average of the prices, so that he would be guaranteed a profit of zero. But he can do better by smoothing his trading. Suppose the dealer deviates from $\boldsymbol{x} = \boldsymbol{\tau}$ by shifting volume from a period $t''$ to a period $t'$ where $\tau_{t'} < \tau_{t''}$. The direct effect of this deviation is to reduce his expected costs at the rate $\mathbb{E}[p_{t''}] - \mathbb{E}[p_{t'}] = \gamma(x_{t''} - x_{t'}) = \gamma(\tau_{t''} - \tau_{t'}) > 0$. And the indirect effect (through prices) is vanishing, because as we have noted, $\boldsymbol{x} = \boldsymbol{\tau}$ insures the dealer against price fluctuations. That a risk-neutral dealer optimally smooths precisely one half of the variation in $\boldsymbol{\tau}$ is due to the linearity of price impact.

In particular, if offered the guaranteed TWAP contract $\boldsymbol{\tau}^{TWAP} = \left(\frac{1}{T},\dots,\frac{1}{T}\right)^\top$, the dealer selects $\boldsymbol{x} = \left(\frac{1}{T},\dots,\frac{1}{T}\right)^\top$, which is in fact the efficient action (i.e., $\boldsymbol{x}^{FB}$). This outcome also leaves the dealer with zero surplus. It follows that $\boldsymbol{\tau}^{TWAP}$ gives the client her first-best payoff. Clearly, nothing can do better than that, and this contract must be optimal.

[21] More precisely, the coefficients on $(\tau_1,\dots,\tau_T)^\top$ constitute an inverted-$V$, which is maximized at the coefficient on $\tau_t$. The intuition is that $\sum_{s=1}^{t} x_s$ is increasing in each of $(x_1,\dots,x_t)$, and given the dealer's frontloading motive, an increase in $\tau_t$ leads each of $(x_1,\dots,x_t)$ to increase. Let us contrast that with $\tau_{t-1}$ and $\tau_{t+1}$. An increase in $\tau_{t-1}$ leads $(x_1,\dots,x_{t-1})$ to increase but does not lead $x_t$ to increase. An increase in $\tau_{t+1}$ also leads $(x_1,\dots,x_t)$ to increase, but the effect is more muted because $\tau_{t+1}$ also works to increase $x_{t+1}$.

[22] That the optimal contract puts relatively less weight on prices of interior periods is also intuitive because these prices are the easiest for the dealer to manipulate, in the following sense. Fixing any deterministic trading strategy $\bar{\boldsymbol{x}}$ as a baseline, imagine that in choosing his trading strategy $\boldsymbol{x}$, the dealer is constrained not only by $\sum_{s=1}^{T} x_s = 1$ but also by $x_t \in [\bar{x}_t - \delta, \bar{x}_t + \delta]$ for all $t$. This means that $\sum_{s=1}^{t} \bar{x}_s - \delta\min\{t, T-t\} \le \sum_{s=1}^{t} x_s \le \sum_{s=1}^{t} \bar{x}_s + \delta\min\{t, T-t\}$, so that the dealer can manipulate $p_t$ by $\theta\delta\min\{t, T-t\}$ in either direction.

[23] Indeed, we have $\mathbb{E}[p_t] = p_0 + \theta\sum_{s=1}^{t} x_s = p_0 + \theta\left[\frac{t}{T} + \left(1-\frac{t}{T}\right)\tau_1\right]$, using Eq. (5) and $\tau_2 = \cdots = \tau_{T-1} = 0$. Thus, the direct effect of perturbing $\boldsymbol{\tau} = (\tau_1, 0, \dots, 0, \tau_T)^\top$ so as to shift weight from period $T$ to period 1 is

$$\mathbb{E}[p_1] - \mathbb{E}[p_T] = \theta\left[\frac{1}{T} + \left(1-\frac{1}{T}\right)\tau_1 - 1\right] = -\theta\left(1-\frac{1}{T}\right)(1-\tau_1).$$

And the indirect effect is

$$\tau_1\left(\frac{d\mathbb{E}[p_1]}{d\tau_1} - \underbrace{\frac{d\mathbb{E}[p_1]}{d\tau_T}}_{=0}\right) + \tau_T\left(\underbrace{\frac{d\mathbb{E}[p_T]}{d\tau_1}}_{=0} - \underbrace{\frac{d\mathbb{E}[p_1]}{d\tau_T}}_{=0}\right) = \tau_1\theta\left(1-\frac{1}{T}\right).$$

### 4.5. When price risk is the only influence

For the case in which price risk is the only influence, fix $\theta$ and $\gamma$, and consider the limit as $\lambda \to \infty$. According to claim *(ii)* of the following result, the outcome resembles the case in which temporary price impact is the only influence: the guaranteed TWAP contract is optimal, it induces the dealer to use the first-best trading strategy, and the client obtains her first-best payoff. But it is for a different reason, as according to claim *(i)*, there is a difference in the dealer's best response function.

**Corollary 8.** *Consider the limit as the dealer becomes infinitely risk averse ($\lambda \to \infty$).*

*(i) For any $\boldsymbol{\tau} \in \mathcal{T}^{wa}$, the dealer's best response converges to $\boldsymbol{x} = \boldsymbol{\tau}$.*
*(ii) The weights of the optimal contract converge to $\boldsymbol{\tau}^* = \left(\frac{1}{T}, \ldots, \frac{1}{T}\right)^\top$, so that the dealer's best-responding trading strategy converges to $\boldsymbol{x}^* = \left(\frac{1}{T}, \ldots, \frac{1}{T}\right)^\top$. The client's expected costs of execution converge to $p_0 + \frac{\gamma}{T} + \frac{\theta(T+1)}{2T}$.*

Claim *(i)* says that the dealer has a *mirroring motive* in this case: his best response is to choose an $\boldsymbol{x}$ equal to the offered $\boldsymbol{\tau}$. The intuition is the following. As the dealer becomes more risk averse, he places greater emphasis on insuring himself against price shocks. In fact, he can perfectly insure himself by selecting a deterministic trading strategy with weights that mirror the contract he is offered. In the limit of infinite risk aversion, this is exactly what he does. In particular, $\boldsymbol{\tau}^{TWAP} = \left(\frac{1}{T}, \ldots, \frac{1}{T}\right)^\top$ induces the dealer to select the efficient action $\boldsymbol{x}^{FB} = \left(\frac{1}{T}, \ldots, \frac{1}{T}\right)^\top$. This outcome moreover leaves the dealer perfectly insured and with zero surplus. It follows that $\boldsymbol{\tau}^{TWAP}$ gives the client her first-best payoff, and must therefore be optimal.

This result reflects an interesting contrast relative to classical models of moral hazard (e.g., Holmström, 1979). Those classical models feature an insurance-incentives tradeoff: the agent can be induced to take the efficient action (i.e., high effort) only if he is exposed to risk. And if the agent is very risk averse, then he must be paid a significant risk premium for that. The principal's payoff then typically declines as the agent becomes more risk averse. In contrast, given the special structure of our setting, inducing the efficient action (i.e., $\boldsymbol{x}^{FB}$) does not always require exposing the dealer to risk. In fact, in this limit of an infinitely risk averse dealer, $\boldsymbol{\tau} = \left(\frac{1}{T}, \ldots, \frac{1}{T}\right)^\top$ induces the dealer to select $\boldsymbol{x}^{FB}$ without exposing him to any risk at all. In consequence, the client's payoff is not monotonically decreasing in $\lambda$.

### 4.6. Discussion of the general solution

The general model can be viewed as a combination of the three aforementioned special cases. Accordingly, the general formula for the dealer's best response reflects a mixture of the frontloading, smoothing, and mirroring motives respectively discussed in the previous sections. Likewise, the general formula for the optimal contract combines the features of the optimal contracts from those special cases. One notable feature shared by all three special cases is that the optimal contract is symmetric, in the sense that $\tau_j^* = \tau_{T+1-j}^*$ for all $j$. In fact, such symmetry holds in general.

**Corollary 9.** *The optimal contract weights are symmetric: $\tau_j^* = \tau_{T+1-j}^*$ for all $j = 1, \ldots, T$.*

The intuition for why symmetry obtains in general can be thought of as a combination of the various reasons for why it obtains in each of the three special cases discussed before.

To illustrate the general solution provided by Proposition 5, Figs. 1–3 display $\boldsymbol{\tau}^*$ and $\boldsymbol{x}^*$ for various choices of the parameters $\theta$, $\gamma$, and $\lambda$.[24] The left panels of these figures depict the optimal contract weights; consistent with Corollary 9, they are indeed symmetric. The right panels depict the dealer's best-responding trading strategy.

Figs. 1–3 suggest that the general solution exhibits several additional qualitative patterns. First, the optimal contract weights are U-shaped: $\tau_1^* \geq \tau_2^* \geq \cdots \geq \tau_{\lceil T/2 \rceil}^* \leq \cdots \tau_{T-1}^* \leq \tau_T^*$.[25] Second, the dealer responds with a trading strategy that is frontloaded relative to the contract: $\sum_{s=1}^t x_t^* \geq \sum_{s=1}^t \tau_s^*$ for all $t = 1, \ldots, T$. Third, the severity of both this U-shape and this frontloading is strengthened by $\theta$ (the coefficient of permanent price impact), weakened by $\gamma$ (the coefficient of temporary price impact), and weakened by $\lambda$ (the dealer's coefficient of absolute risk aversion).

These patterns can be understood through the aforementioned special cases. With permanent price impact as the only influence, we have $\boldsymbol{\tau}^* = \left(\frac{1}{2}, 0, \ldots, 0, \frac{1}{2}\right)^\top$, which is the maximally-severe U-shape, and $\boldsymbol{x}^* = \left(\frac{T+1}{2T}, \frac{1}{2T}, \ldots, \frac{1}{2T}\right)^\top$, which is strictly frontloaded (relative to the contract). With either temporary price impact or price risk as the only influence, we have $\boldsymbol{\tau}^* = \left(\frac{1}{T}, \ldots, \frac{1}{T}\right)^\top$, which is the minimally-severe U-shape, and $\boldsymbol{x}^* = \left(\frac{1}{T}, \ldots, \frac{1}{T}\right)^\top$, which represents minimally-severe frontloading. The intuition for why the U-shape in $\boldsymbol{\tau}^*$ and frontloading in $\boldsymbol{x}^*$ (weakly) obtain in general can be thought of as a combination of the different reasons for why they obtain in each of the three special cases discussed before. The comparative statics can also be understood in these terms. An increase in $\theta$ moves us toward the limiting case of Section 4.3, so it increases both the severity of the U-shape and the severity of the frontloading. Increases in $\gamma$ and $\lambda$ reduce those severities because they move us toward the limiting cases of Sections 4.4 and 4.5, respectively.

[24] Note that $\lambda$ and $\sigma$ affect the solution only through the quantity $\lambda\sigma^2$. Hence, Fig. 3, which depicts how the solution changes with $\lambda$, speaks also to how the solution changes with $\sigma$.

[25] In fact, a stronger property appears to hold. The figures suggest that the optimal contract weights are convex in the sense that $\tau_1^* - \tau_2^* \geq \tau_2^* - \tau_3^* \geq \cdots \geq \tau_{T-2}^* - \tau_{T-1}^* \geq \tau_{T-1}^* - \tau_T^*$. Given that the weights are symmetric (*cf.* Corollary 9), this convexity condition implies the U-shape condition $\tau_1^* \geq \tau_2^* \geq \cdots \geq \tau_{\lceil T/2 \rceil}^* \leq \cdots \tau_{T-1}^* \leq \tau_T^*$.

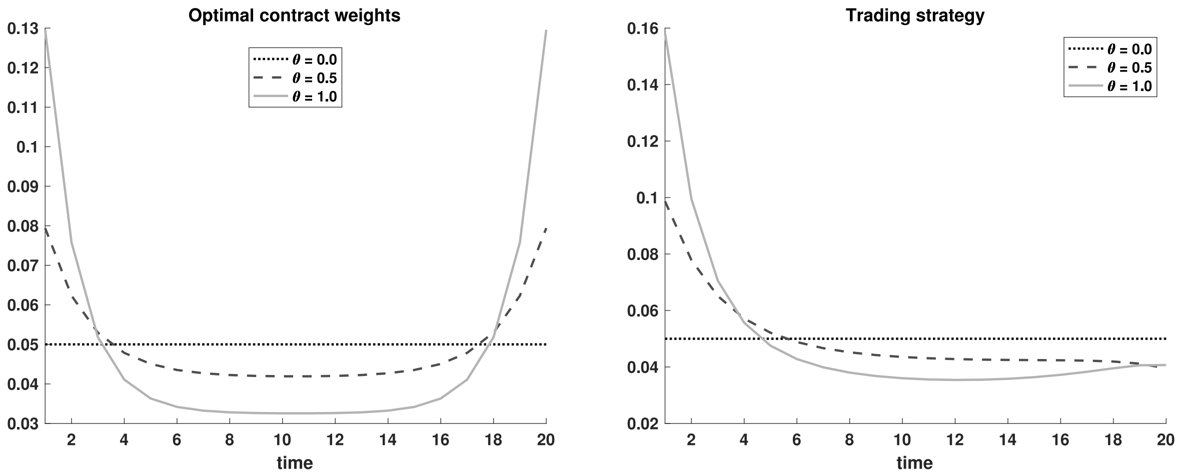

**Fig. 1.** The optimal contract weights and best-responding trading strategy for different levels of permanent price impact. Without permanent price impact ($\theta = 0$), both the optimal weights and the trading strategy are constant over time. As permanent price impact grows, the optimal weights become more U-shaped, and the dealer's trading strategy becomes more frontloaded. The other parameters are $\gamma = 1$, $\lambda = 1$, $\sigma = 0.5$, and $T = 20$.

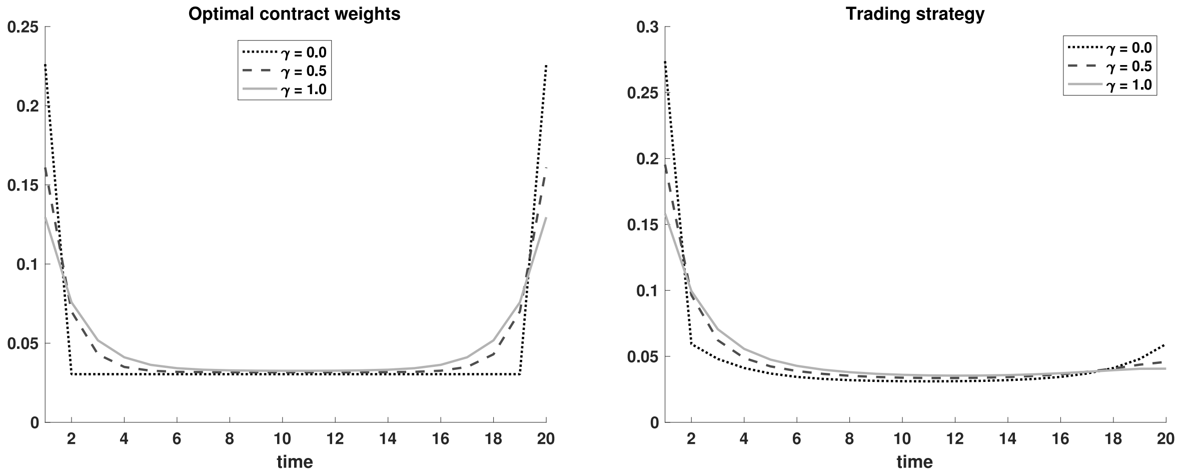

**Fig. 2.** The optimal contract weights and best-responding trading strategy for different levels of temporary price impact. Without temporary price impact ($\gamma = 0$), the optimal weights are the same for all periods except for the first and last periods, and the dealer's trading strategy is frontloaded. As temporary price impact grows, the curves for the optimal weights become smoother, and the dealer's trading strategy becomes less frontloaded. The other parameters are $\theta = 1$, $\lambda = 1$, $\sigma = 0.5$, and $T = 20$.

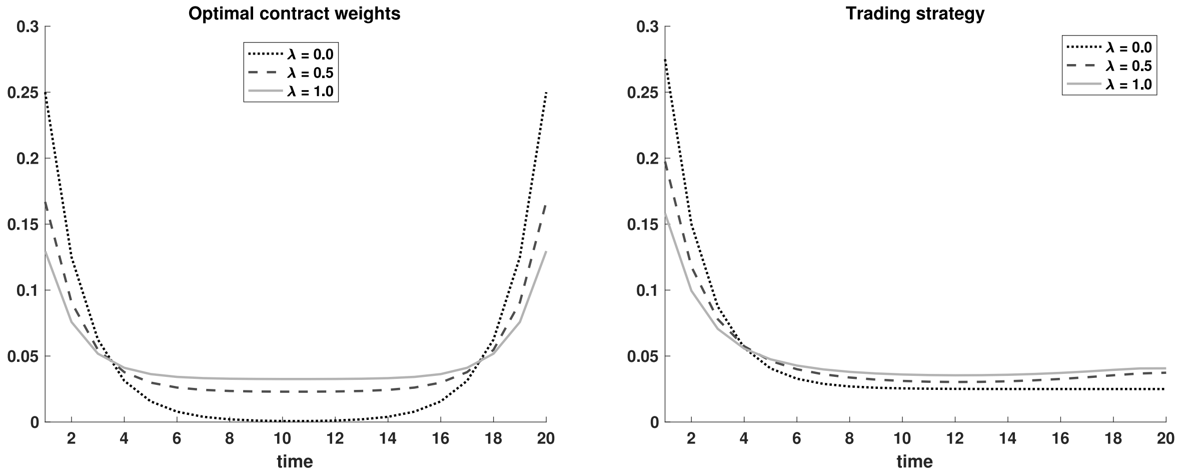

**Fig. 3.** The optimal contract weights and best-responding trading strategy for different levels of risk aversion. As risk aversion grows, the optimal weights become less U-shaped, and the dealer's trading strategy becomes less frontloaded. The other parameters are $\theta = 1$, $\gamma = 1$, $\sigma = 0.5$, and $T = 20$.

We stress that these observations about the U-shape of the optimal contract and the frontloading of the dealer's trading come only from numerical experimentation and do not correspond to any formal result that we have been able to derive from our closed-form solution to the general discrete-time model. We do, however, prove analogues of these observations for the continuous-time limit analyzed in the next section.

## 5. Continuous-time limit

In light of ambiguity regarding what precisely a trading period represents, as well as recent trends toward progressively high-frequency trading, we are motivated to study the continuous-time limit of our discrete-time model. For this limit, we consider a sequence of models, indexed by $k$. Along this sequence, we let the number of periods diverge (i.e., $T_k \to \infty$). At the same time, the distance between consecutive periods should vanish, so as to hold the execution horizon constant. To capture this, we shrink the variance of price shocks to zero (i.e., $\sigma_k^2 \to 0$) in such a way that $T_k\sigma_k^2$ remains equal to a constant, which we denote $T\sigma^2$.[26]

### 5.1. The optimal contract

To illuminate the underlying patterns, we state the following result in terms of cumulative values through quantiles $q$ of the execution period: $\sum_{t=1}^{\lceil qT \rceil} \tau_t^*$ and $\sum_{t=1}^{\lceil qT \rceil} x_t^*$. To ensure that the convergence is well behaved, we focus on the case of a strictly risk-averse dealer.

**Proposition 10.** *Consider a sequence of execution horizons $(T_k)_{k=1}^{\infty}$ and a sequence of price-shock variances $(\sigma_k^2)_{k=1}^{\infty}$ such that $\lim_{k\to\infty} T_k = \infty$ and $T_k\sigma_k^2 = T\sigma^2$ for all $k$. For each $k$, let $\boldsymbol{\tau}^{*k}$ be the associated optimal contract, and let $\boldsymbol{x}^{*k}$ be the dealer strategy that best responds to $\boldsymbol{\tau}^{*k}$. Assume also that the dealer is strictly risk-averse ($\lambda > 0$). For all $q \in [0,1]$,*

$$\lim_{k\to\infty} \sum_{t=1}^{\lceil qT_k \rceil} \tau_t^{*k} = \begin{cases} 0 & \text{if } q = 0 \\ \frac{1-a}{2} + aq & \text{if } q \in (0,1) \\ 1 & \text{if } q = 1 \end{cases} \quad \text{and} \quad \lim_{k\to\infty} \sum_{t=1}^{\lceil qT_k \rceil} x_t^{*k} = \begin{cases} 0 & \text{if } q = 0 \\ \frac{3(1-a)}{4} + aq & \text{if } q \in (0,1) \\ 1 & \text{if } q = 1 \end{cases}$$

*where $a = \frac{1}{1+\frac{4\theta}{\lambda T\sigma^2}}$. The client's expected costs of execution converge to $p_0 + \frac{3-a}{4}\theta$.*

The optimal contract in the continuous-time limit takes a surprisingly simple form, which can in fact be viewed as an extreme U-shape: interior times are weighted with a constant density of $a$, and the two boundary instants are weighted with atoms of $\frac{1-a}{2}$ each.[27] For the dealer's best response, interior times are also weighted with a constant density of $a$, but there is frontloading in terms of the boundary weights: the initial atom is three times the terminal atom.

*Comparative statics.* This density $a$ is increasing in $(\lambda, T\sigma^2)$ and decreasing in $\theta$. As $a$ relates inversely to the severity of the optimal contract's U-shape, these relationships accord with what the earlier numerical experimentation suggests holds in general. To explain the intuition, note that permanent price impact generates an expected gap between the initial and terminal prices, creating a frontloading motive for the dealer: by frontloading, the dealer expects to buy low and sell high. A larger $\theta$ implies a larger expected gap, hence a larger frontloading motive. On the other hand, frontloading exposes the dealer to price risk owing to variance in this gap. Larger $T\sigma^2$ implies a larger variance, hence a smaller frontloading motive. Larger $\lambda$ means less risk-bearing capacity, hence a similarly smaller frontloading motive. Finally, to see the connection between the frontloading motive and $a$, consider what would happen if the frontloading motive were to disappear entirely so that the dealer's trades perfectly mirrored the weights of the offered contract. In that case, the client obtains her first-best payoff from a guaranteed TWAP contract (i.e., with $a = 1$). Analogously, smaller (larger) frontloading motives imply larger (smaller) values of $a$.

*Temporary price impact.* If trading were everywhere sufficiently diffuse, then temporary price impact would vanish in the limit. Indeed, temporary price impact costs are $\gamma \sum_{t=1}^{T_k} (x_t^k)^2$, which, for example, vanish under the first-best trading policy, $\boldsymbol{x}^{FB,k} = \left(\frac{1}{T_k}, \dots, \frac{1}{T_k}\right)^\top$. More generally, a sufficient condition for vanishing temporary price impact is that $\max_{1\le t\le T_k} |x_t^k|$ is $o\left(1/\sqrt{T_k}\right)$.

However, under the best response to the optimal contract, trading is not everywhere diffuse in this sense (unless $\theta = 0$), and temporary price impact does not vanish. So it is for subtle reasons that $\gamma$ does not affect the limit characterized by Proposition 10. This invariance obtains because temporary price impact creates two effects. On the one hand, if the dealer's trading schedule were held fixed, then an increase in $\gamma$ would raise prices and hence the client's payment. But on the other hand, an increase in $\gamma$ creates a smoothing motive for the dealer, which reduces the extent of the dealer's frontloading and hence the client's payment. In the continuous-time limit, these two considerations offset under the optimal contract.

*Convergence.* Although temporary price impact has no effect on the continuous-time limit, it does affect convergence to this limit. Without temporary price impact, the first and last contract weights converge to the atoms of the continuous-time limits so that

$$\lim_{k\to\infty} \tau_1^{*k} = \lim_{k\to\infty} \tau_{T_k}^{*k} = \frac{1-a}{2} \quad \text{and} \quad \lim_{k\to\infty} \tau_{j+1}^{*k} = \lim_{k\to\infty} \tau_{T_k-j}^{*k} = 0 \quad \text{for any fixed } j \ge 1.$$

In contrast, with temporary price impact, we have a sequence of discrete weights

$$\lim_{k\to\infty} \tau_{j+1}^{*k} = \lim_{k\to\infty} \tau_{T_k-j}^{*k} = \frac{\theta\gamma^j}{(\theta+\gamma)^{j+1}} \frac{1-a}{2} \quad \text{for any fixed } j \ge 0.$$

[26] This type of limit is classical in the study of continuous-time limits of discrete-time models, going back to Cox et al. (1979).

[27] Studying an optimal execution problem, Obizhaeva and Wang (2013) derive a very similar form for the optimal trading strategy: interior times weighted with a constant density and boundary instants weighted with equal atoms. But the similarity is only coincidental. They solve a different problem (a problem of optimal execution rather than one of optimal contracting) under a different set of assumptions.

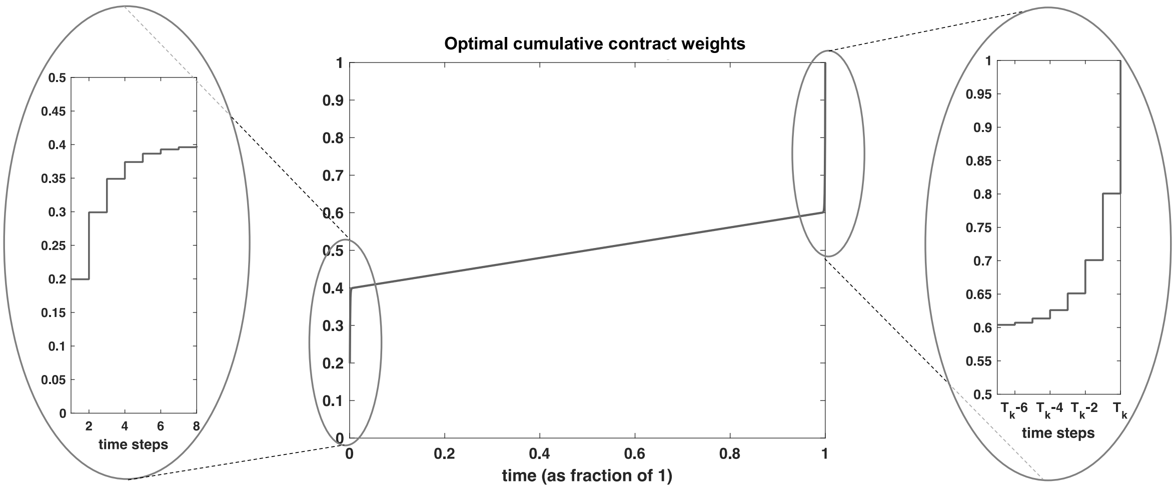

**Fig. 4.** The optimal contract for a large $k$ (such that $T_k = 2{,}000$). There are two sequences of discrete weights at the beginning and end, while the weights are smooth for interior times. Parameters are $\theta = 1$, $\gamma = 1$, $\lambda = 1$, and $T\sigma^2 = 1$. The limit features a constant density at interior times of $a = 1/\big(1 + \frac{4\theta}{\lambda T\sigma^2}\big) = 0.2$, as well as initial and terminal atoms of $\frac{1-a}{2} = 0.4$ each.

Note that the sum of each of the two sequences equals

$$\sum_{j=0}^{\infty} \frac{\theta\gamma^j}{(\theta+\gamma)^{j+1}} \frac{1-a}{2} = \frac{\theta}{\theta+\gamma} \frac{1}{1-\frac{\gamma}{\theta+\gamma}} \frac{1-a}{2} = \frac{1-a}{2},$$

which coincides with what Proposition 10 specifies for the jumps of $\lim_{k\to\infty} \sum_{t=1}^{\lceil qT_k \rceil} \tau_t^{*k}$ at $q = 0$ and $q = 1$. Fig. 4 illustrates this convergence.[28]

Interestingly, the optimal trading strategy has a different form: a sequence of discrete weights only at the beginning, but not at the end of the trading times. Without temporary price impact, only the first element of the trading strategy converges to a nonzero value

$$\lim_{k\to\infty} x_1^{*k} = \frac{1-a}{2}, \quad \lim_{k\to\infty} x_{j+2}^{*k} = 0, \quad \lim_{k\to\infty} x_{T_k-j}^{*k} = 0 \;\text{ for any fixed } j \ge 0.$$

With temporary price impact, we have

$$\lim_{k\to\infty} x_{j+1}^{*k} = \frac{\theta\gamma^j}{(\theta+\gamma)^{j+1}} \frac{1-a}{2}, \quad \lim_{k\to\infty} x_{T_k-j}^{*k} = 0 \;\text{ for any fixed } j \ge 0. \tag{6}$$

However, the jumps of $\lim_{k\to\infty} \sum_{t=1}^{\lceil qT_k \rceil} x_t^{*k}$ at $q = 0$ and $q = 1$ are not determined only by these sequences of discrete weights. As stated in Proposition 10, the jump at $q = 0$ is $\frac{3(1-a)}{4}$, consisting of not only $\sum_{j=0}^{\infty} \frac{\theta\gamma^j}{(\theta+\gamma)^{j+1}} \frac{1-a}{2} = \frac{1-a}{2}$ from (6), but also another infinite sum whose terms individually converge to zero but whose sum converges to $\frac{1-a}{4}$. This is illustrated in Fig. 5, where we see both the sequence of discrete weights at zero and a piece of the curve near zero that converges to a vertical line as $k \to \infty$. Likewise, the jump at $q = 1$ is $\frac{1-a}{4}$, which comes entirely from an infinite sum of terms that individually all converge to zero.

*Supplemental material.* See Online Appendix D for supplemental details on this continuous-time limit, including an overview of our strategy for proving Proposition 10 and a discussion of the wedge between the first-best and second-best.

### 5.2. Discussion of outcomes under common contracts

Although not optimal in our model, two contracts that are nevertheless commonly used are $\boldsymbol{\tau}^{TWAP}$ and $\boldsymbol{\tau}^{MOC}$. Natural questions include: What trading behavior is induced by these contracts? By how much do they underperform the optimal contract? Under what situations, if any, do they deliver outcomes close to the client's second-best payoff? The following result provides answers.

**Proposition 11.** *Consider a sequence of execution horizons $(T_k)_{k=1}^{\infty}$ and a sequence of price-shock variances $(\sigma_k^2)_{k=1}^{\infty}$ such that $\lim_{k\to\infty} T_k = \infty$ and $T_k\sigma_k^2 = T\sigma^2$ for all $k$. Assume also that the dealer is strictly risk-averse ($\lambda > 0$).*

*(i) For each $k$, let $\boldsymbol{x}^{TWAP,k}$ be the dealer strategy that best responds to $\boldsymbol{\tau}^{TWAP,k}$. For all $q \in [0,1]$,*

$$\lim_{k\to\infty} \sum_{t=1}^{\lceil qT_k \rceil} x_t^{TWAP,k} = \begin{cases} 0 & \text{if } q = 0 \\ \frac{\theta}{\lambda T\sigma^2} + q & \text{if } q \in (0,1) \\ 1 & \text{if } q = 1 \end{cases} \tag{7}$$

[28] Although $\lambda$ affects $a$, and hence the total amount of weight in these sequences, it does not affect how this total is divided across the elements of the sequences (in the limit). This is intuitive because when the time periods become shorter, price fluctuations between consecutive periods become smaller, so that for purposes of these periods around the boundary times, the dealer behaves in the limit as if he were risk-neutral (regardless of $\lambda$). The role of $\gamma$ is exactly the opposite: it affects the division of weight across the sequences, but not the weight assigned to the sequences in total.

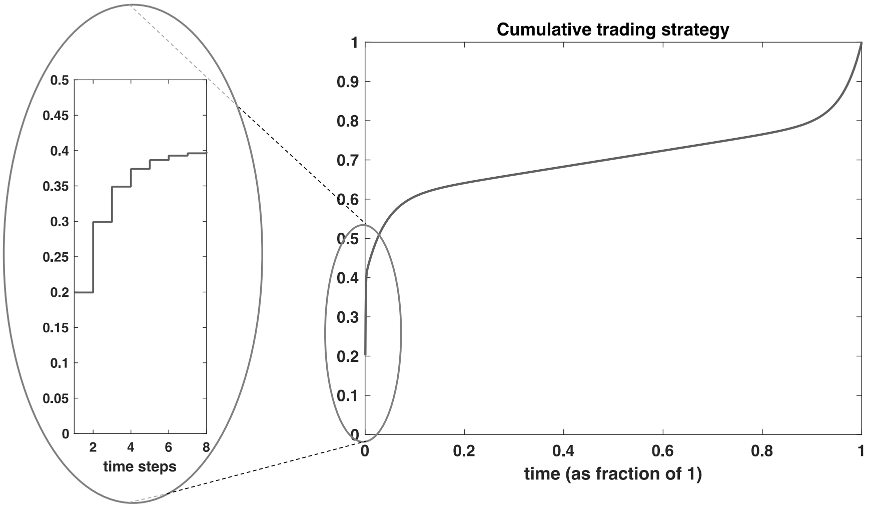

**Fig. 5.** The dealer's best-responding trading strategy for a large $k$ (such that $T_k = 2{,}000$). There is a sequence of discrete weights at the beginning, while the weights are smooth for interior times and toward the end (although part of the smooth curve converges to a vertical line). Parameters are $\gamma = 1$, $\theta = 1$, $\lambda = 1$, and $T\sigma^2 = 1$. The limit features a constant density at interior times of $a = 1/\left(1 + \frac{4\theta}{\lambda T \sigma^2}\right) = 0.2$, an initial atom of $\frac{3(1-a)}{4} = 0.6$, and a terminal atom of $\frac{1-a}{4} = 0.2$.

*The client's expected costs of execution converge to* $p_0 + \frac{1}{2}\theta + \frac{\theta^2}{\lambda T \sigma^2}$.

*(ii) For each* $k$, *let* $\mathbf{x}^{MOC,k}$ *be the dealer strategy that best responds to* $\boldsymbol{\tau}^{MOC,k}$. *For all* $q \in [0, 1]$,

$$\lim_{k\to\infty} \sum_{t=1}^{\lceil qT_k \rceil} x_t^{MOC,k} = \begin{cases} 0 & \textit{if } q \in [0,1) \\ 1 & \textit{if } q = 1 \end{cases} \tag{8}$$

*The client's expected costs of execution converge to* $p_0 + \theta + \frac{\gamma^2}{\theta + 2\gamma}$.

According to part *(i)* of the proposition, $\boldsymbol{\tau}^{TWAP}$ leads the dealer to frontload his trading so much that he actually overbuys, before selling a discrete amount at the terminal instant. The client can deter this overbuying – and consequently do better – by collecting contract weights from interior times near the end of the window into an atom on the terminal instant. Hence, its terminal atom is one way by which the optimal contract improves upon $\boldsymbol{\tau}^{TWAP}$. According to part *(ii)* of the proposition, $\boldsymbol{\tau}^{MOC}$ leads the dealer to concentrate all his trading at the terminal instant, behavior sometimes termed 'banging the close.' Such extraordinarily concentrated trading is inefficient, and one way by which the optimal contract improves upon $\boldsymbol{\tau}^{MOC}$ is to deter it.

*Back-of-the-envelope calculation.* To quantify our findings, we consider a reasonable parametrization for our model's continuous-time limit. Consider a client who desires to trade a position, currently valued at $V$ = \$100 million, in a certain stock. Let the parameters be $p_0$ = \$100, $\theta = 2 \times 10^{-6}$, $\gamma = 0$, $\lambda = 2 \times 10^{-6}$, $T = 1$, and $\sigma^2 = 6.1$.[29] Under these parameters, the optimal contract puts 19.8 percent of its weight on the opening price, 19.8 percent on the closing price, and 60.4 percent on the intraday time-weighted average price.

We then use our results to compute model-implied transaction costs under various scenarios. Our metric is implementation shortfall: the client's expected costs net of what it would have cost to trade her entire volume at $p_0$ per share. Whereas our theoretical analysis normalized trade size to one share, we are now contemplating a trade of $V/p_0 = 1$ million shares. In our model, transaction costs grow with the square of volume, so we therefore scale up by a factor of one trillion. Doing so, we find the following. First-best transaction costs are $10^{12}\left(\frac{1}{2}\theta\right)$ = \$1 million, or 100 bps of the value of the trade, which is the correct order of magnitude for trades of block sizes (e.g., US Securities and Exchange Commission, 2005; Abel Noser, 2021; The Wall Street Journal, 2022a,b). Second-best transaction costs are $10^{12}\left(\frac{1}{2}\theta + \frac{\theta^2}{4\theta + \lambda T \sigma^2}\right)$ = \$1.2 million (or 120 bps). Under $\boldsymbol{\tau}^{TWAP}$, transaction costs are $10^{12}\left(\frac{1}{2}\theta + \frac{\theta^2}{\lambda T \sigma^2}\right)$ = \$1.33 million (or 133 bps). Under $\boldsymbol{\tau}^{MOC}$, they are $10^{12}\left(\theta + \frac{\gamma^2}{\theta + 2\gamma}\right)$ = \$2 million (or 200 bps).

The calculations reported in the previous paragraph indicate that switching to the optimal contract from $\boldsymbol{\tau}^{TWAP}$ would reduce transaction costs by 13 bps. Gains of a switch from $\boldsymbol{\tau}^{MOC}$ would be even larger, 80 bps. In either case, such a switch closes a sizable portion of the gap relative to the first-best and represents a cost saving on the order of hundred(s) of thousands of dollars per trade. Scaling up by the market-wide volume of such trades, these cost savings could extrapolate to billions of dollars per year.[30] Of course, additional savings might be possible with more complex contracts (e.g., the set of affine contracts as in Online Appendix A, or the fully general set as in Section 3.2). Nevertheless, it is striking that such substantial cost savings can be obtained, even while staying within the relatively simple class of weighted-average-price contracts.

[29] These parameter values are consistent with the following facts. Abel Noser (2021) describes a dataset of portfolio transitions, with a median size of \$145 million. The median S&P 500 stock price was \$112 as of April 27, 2022. Cartea and Jaimungal (2016, Tables 7 and 8) estimate the coefficient of permanent price impact for 17 stocks, with results ranging from $0.63 \times 10^{-6}$ to $2.03 \times 10^{-4}$. Choosing $\gamma = 0$ is to be maximally conservative, biasing our analysis in favor of finding a small difference between the performance of $\boldsymbol{\tau}^{MOC}$ and our optimal contract. Campo, Guerre, Perrigne, and Vuong (2011, Table 2) estimate a coefficient of absolute risk aversion of $2 \times 10^{-6}$. $T = 1$ reflects an execution window of one day. Avramov, Chordia, and Goyal (2006, Table 1) find that the standard deviation of daily returns is 2.47%, which for a \$100 stock equates to a variance of 6.10.

[30] A conservative estimate of institutional transaction costs is \$70 billion per year (Nasdaq, 2022; SIFMA, 2021). Assuming that this figure represents 133 bps (200 bps) of the traded value and that a switch to our optimal contract would save 13 bps (80 bps) of that value implies total cost savings of \$6.8 billion (\$28 billion) per year.

## 6. Conclusion

This paper formulates a contracting problem in which a client (the principal) contracts to purchase a position from a dealer (the agent) at some future time. In the interim, the dealer acquires the position from the market. The friction is hidden action, in that the client cannot observe the dealer's on-market trades, but only the evolution of market prices, so that the dealer has an incentive to frontload his trading. Eliminating this friction and solving for the first-best benchmark, the problem becomes one of optimal execution. Indeed, our analysis of the first-best problem recovers classic results from that literature about optimality of trading at a constant rate.

However, we depart from the optimal execution literature by analyzing the implications of these agency conflicts. Focusing on contracts that are weighted averages of market prices, we characterize the second-best solution in discrete time, then take the continuous-time limit. The optimal contract in this limit is an extreme U-shape: atoms of equal mass at the two extreme times and a constant density at interior times. The mass at the extreme times – and hence the severity of the U-shape – is increasing in permanent price impact, decreasing in the dealer's risk aversion, and constant in temporary price impact.

These results shed light on the interplay between price impact and agency conflicts in financial markets. They could also aid in reducing the transaction costs of pension funds, endowments, or other institutional traders who sometimes outsource the execution of large trades. In particular, guaranteed TWAP contracts (and similar guaranteed VWAP contracts) are common in practice. Although our results rationalize the practice of putting equal weight on interior prices, they also indicate that these contracts themselves are unlikely to be optimal unless price impact is predominantly temporary or the dealer is highly risk averse. Guaranteed MOC contracts, which put full weight on the closing price, are also commonly used. Although our results rationalize the practice of putting substantial weight on the closing price, they also recommend that the opening price receive equally substantial weight, so that the contract more closely resembles the U-shape that is optimal in the model. As regulators review best practices in relation to over-the-counter block trading, they may revisit the wisdom of various pricing benchmarks in light of our analysis.

## CRediT authorship contribution statement

**Markus Baldauf:** Writing – review & editing, Writing – original draft, Formal analysis, Conceptualization. **Christoph Frei:** Writing – review & editing, Writing – original draft, Formal analysis, Conceptualization. **Joshua Mollner:** Writing – review & editing, Writing – original draft, Formal analysis, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Block Trade Contracting (code for reproducing figures) https://data.mendeley.com/datasets/5gp82wtfjc/1.

## Appendix A. Proofs of results stated in the main text

*How the parameter $\sigma$ is handled in the proofs.* Except for the proofs of Propositions 10 and 11, where $\sigma_k$ depends on $k$, we assume for all proofs that $\sigma = 1$ without loss of generality. Indeed, any parametrization of the model with $\sigma \neq 1$ is economically-equivalent to an alternative parametrization in which the price-shock variance is 1 (instead of $\sigma^2$) and in which the dealer's risk aversion is $\lambda\sigma^2$ (instead of $\lambda$).

We begin with a lemma that will be used in the proof of Proposition 1.

**Lemma 12.** *For any trading strategy $\boldsymbol{x} \in \mathcal{X}$,*

$$\mathbb{E}\left[\sum_{t=1}^{T} x_t \sum_{s=1}^{t} \varepsilon_s\right] = 0.$$

**Proof of Lemma 12.** We start by writing

$$\mathbb{E}\left[\sum_{t=1}^{T} x_t \sum_{s=1}^{t} \varepsilon_s\right] = \mathbb{E}\left[\sum_{t=1}^{T} x_t \left(\sum_{s=1}^{T} \varepsilon_s - \sum_{s=t+1}^{T} \varepsilon_s\right)\right] = \mathbb{E}\left[\sum_{t=1}^{T} x_t \sum_{s=1}^{T} \varepsilon_s\right] - \mathbb{E}\left[\sum_{t=1}^{T} x_t \sum_{s=t+1}^{T} \varepsilon_s\right].$$

We complete the proof by showing that each of these two terms evaluates to zero:

$$\mathbb{E}\left[\sum_{t=1}^{T} x_t \sum_{s=1}^{T} \varepsilon_s\right] = \mathbb{E}\left[\sum_{s=1}^{T} \varepsilon_s \sum_{t=1}^{T} x_t\right] = \mathbb{E}\left[\sum_{s=1}^{T} \varepsilon_s\right] = 0,$$

$$\mathbb{E}\left[\sum_{t=1}^{T} x_t \sum_{s=t+1}^{T} \varepsilon_s\right] = \mathbb{E}\left[\sum_{t=1}^{T} x_t \sum_{s=t+1}^{T} \mathbb{E}\left[\varepsilon_s | x_t\right]\right] = 0. \quad \square$$

**Proof of Proposition 1.** Given an arbitrary trading strategy $\boldsymbol{x} \in \mathcal{X}$, the expected costs of execution are

$$\mathbb{E}\left[\sum_{t=1}^{T} x_t p_t\right] = \mathbb{E}\left[\sum_{t=1}^{T} x_t \left(p_0 + \gamma x_t + \theta \sum_{s=1}^{t} x_s + \sum_{s=1}^{t} \varepsilon_s\right)\right] = p_0 \mathbb{E}\left[\sum_{t=1}^{T} x_t\right] + \gamma \mathbb{E}\left[\sum_{t=1}^{T} x_t^2\right] + \theta \mathbb{E}\left[\sum_{t=1}^{T} x_t \sum_{s=1}^{t} x_s\right] + \mathbb{E}\left[\sum_{t=1}^{T} x_t \sum_{s=1}^{t} \varepsilon_s\right].$$

The first term evaluates to $p_0$. The last term evaluates to zero by Lemma 12. Observe that $\varepsilon$ has fallen out of the expression. Thus, the first-best trading strategy, which minimizes this expression, will not be a function of $\varepsilon$—in other words, it will be deterministic. This first-best trading strategy solves the program

$$\min_{(x_1,\dots,x_T)^\top \in \mathbb{R}^T} p_0 + \gamma \sum_{t=1}^{T} x_t^2 + \theta \sum_{t=1}^{T} x_t \sum_{s=1}^{t} x_s \quad \text{subject to} \quad \sum_{t=1}^{T} x_t = 1.$$

Taking the Lagrangian (with $\mu$ as the multiplier on the constraint), we obtain

$$2\gamma x_t^{FB} + 2\theta x_t^{FB} + \theta \sum_{s \neq t} x_s^{FB} = \mu \quad \text{for all } t = 1,\dots,T.$$

These imply $x_1^{FB} = \cdots = x_T^{FB}$. And from the constraint, we must therefore have $x_1^{FB} = \cdots = x_T^{FB} = \frac{1}{T}$. To obtain the expected costs of execution under this strategy, we compute

$$\mathbb{E}\left[\sum_{t=1}^{T} x_t^{FB} p_t\right] = p_0 + \gamma \sum_{t=1}^{T} \left(\frac{1}{T}\right)^2 + \theta \sum_{t=1}^{T} \frac{1}{T} \sum_{s=1}^{t} \frac{1}{T} = p_0 + \frac{\gamma}{T} + \frac{\theta(T+1)}{2T}. \quad \square$$

**Proof of Proposition 2.** Consider a class of contracts indexed by $(\zeta, \delta, K_1, \dots, K_{T-1})$, where the payment is

$$\tau(\boldsymbol{p}) = \delta + \frac{1}{T} \sum_{t=1}^{T} p_t - \sum_{t=1}^{T-1} K_t \mathbb{1}[p_t - p_{t-1} > \zeta].$$

The dealer's monetary payoff is

$$\tau(\boldsymbol{p}) - \boldsymbol{x} \cdot \boldsymbol{p} = \delta + \sum_{t=1}^{T} p_t \left(\frac{1}{T} - x_t\right) - \sum_{t=1}^{T-1} K_t \mathbb{1}[p_t - p_{t-1} > \zeta].$$

The dealer's expected utility is therefore

$$\begin{aligned}
&\mathbb{E}\big[u\big(\tau(\boldsymbol{p}) - \boldsymbol{x} \cdot \boldsymbol{p}\big)\big] \\
&= \mathbb{E}\left[u\left(\delta + \sum_{t=1}^{T} p_t \left(\frac{1}{T} - x_t\right) - \sum_{t=1}^{T-1} K_t \mathbb{1}\Big[p_t - p_{t-1} > \zeta\Big]\right)\right] \\
&= -\mathbb{E}\left[\exp\left(-\lambda\delta - \lambda \sum_{t=1}^{T} p_t \left(\frac{1}{T} - x_t\right) + \lambda \sum_{t=1}^{T-1} K_t \mathbb{1}[p_t - p_{t-1} > \zeta]\right)\right] \\
&= -\mathbb{E}\left[\exp\left(-\lambda\delta - \lambda \sum_{t=1}^{T}\left(p_0 + \gamma x_t + \theta \sum_{s=1}^{t} x_s + \sum_{s=1}^{t} \varepsilon_s\right)\left(\frac{1}{T} - x_t\right) + \lambda \sum_{t=1}^{T-1} K_t \mathbb{1}[\gamma(x_t - x_{t-1}) + \theta x_t + \varepsilon_t > \zeta]\right)\right] \\
&= -\exp\left(-\lambda\delta - \lambda \sum_{t=1}^{T}\left(p_0 + \gamma x_t + \theta \sum_{s=1}^{t} x_s\right)\left(\frac{1}{T} - x_t\right)\right)\mathbb{E}\left[\mathbb{E}\left[\exp\left(-\lambda \sum_{t=2}^{T} \varepsilon_t \left(\sum_{s=1}^{t-1} x_s - \frac{t-1}{T}\right) + \lambda \sum_{t=1}^{T-1} K_t \mathbb{1}[\gamma(x_t - x_{t-1}) + \theta x_t + \varepsilon_t > \zeta]\right)\middle|\mathcal{F}_{T-1}\right]\right] \\
&= -\exp\left(-\lambda\delta - \lambda \sum_{t=1}^{T}\left(p_0 + \gamma x_t + \theta \sum_{s=1}^{t} x_s\right)\left(\frac{1}{T} - x_t\right) + \frac{\lambda^2}{2}\left(\sum_{s=1}^{T-1} x_s - \frac{T-1}{T}\right)^2\right) \\
&\quad \times \mathbb{E}\left[\exp\left(-\lambda \sum_{t=2}^{T-1} \varepsilon_t \left(\sum_{s=1}^{t-1} x_s - \frac{t-1}{T}\right) + \lambda \sum_{t=1}^{T-1} K_t \mathbb{1}[\gamma(x_t - x_{t-1}) + \theta x_t + \varepsilon_t > \zeta]\right)\right] \\
&= -\exp\left(-\lambda\delta - \lambda \sum_{t=1}^{T-1}\left(p_0 + \gamma x_t + \theta \sum_{s=1}^{t} x_s\right)\left(\frac{1}{T} - x_t\right) - \lambda\left(p_0 + \gamma\left(1 - \sum_{s=1}^{T-1} x_s\right) + \theta\right)\left(\frac{1}{T} - 1 + \sum_{s=1}^{T-1} x_s\right) + \frac{\lambda^2}{2}\left(\sum_{s=1}^{T-1} x_s - \frac{T-1}{T}\right)^2\right) \\
&\quad \times \prod_{t=1}^{T-1} \mathbb{E}\left[\exp\left(-\lambda\varepsilon_t \left(\sum_{s=1}^{t-1} x_s - \frac{t-1}{T}\right) + \lambda K_t \mathbb{1}\Big[\gamma(x_t - x_{t-1}) + \theta x_t + \varepsilon_t > \zeta\Big]\right)\right] \\
&= -\exp\left(-\lambda\delta - \lambda \sum_{t=1}^{T-1}\left(p_0 + \gamma x_t + \theta \sum_{s=1}^{t} x_s\right)\left(\frac{1}{T} - x_t\right) - \lambda\left(p_0 + \gamma\left(1 - \sum_{s=1}^{T-1} x_s\right) + \theta\right)\left(\frac{1}{T} - 1 + \sum_{s=1}^{T-1} x_s\right) + \frac{\lambda^2}{2} \sum_{t=1}^{T}\left(\sum_{s=1}^{t-1} x_s - \frac{t-1}{T}\right)^2\right) \\
&\quad \times \prod_{t=1}^{T-1}\left(\exp(\lambda K_t) - (\exp(\lambda K_t) - 1)\Phi\left(\zeta - \gamma(x_t - x_{t-1}) - \theta x_t + \lambda\left(\sum_{s=1}^{t-1} x_s - \frac{t-1}{T}\right)\right)\right),
\end{aligned}$$

using for the last equality that $\varepsilon_t$ is normally distributed with standard deviation one and mean $-\lambda\big(\sum_{s=1}^{t-1} x_s - \frac{t-1}{T}\big)$ under the probability measure $Q_t$ given by $\frac{dQ_t}{dP} = \exp\big(-\lambda\varepsilon_t\big(\sum_{s=1}^{t-1} x_s - \frac{t-1}{T}\big) - \frac{\lambda^2}{2}\big(\sum_{s=1}^{t-1} x_s - \frac{t-1}{T}\big)^2\big)$. Taking the logarithm and the derivative with respect to $x_j$ at $x_s = 1/T$ for all $s$ gives

$$\lambda\left(p_0 + \gamma\frac{1}{T} + \theta\frac{j}{T}\right) - \lambda\left(p_0 + \gamma\frac{1}{T} + \theta\right)$$

$$
-\sum_{t=1}^{T-1} \frac{\frac{d}{dx_j}\Phi\left(\zeta-\gamma(x_t-x_{t-1})-\theta x_t+\lambda\left(\sum_{s=1}^{t-1}x_s-\frac{t-1}{T}\right)\right)\Big|_{x_s=1/T\,\forall s}}{\frac{1}{1-\exp(-\lambda K_t)}-\Phi(\zeta-\gamma/T\mathbb{1}[t=1]-\theta/T)}
$$

$$
=\lambda\theta\frac{j-T}{T}+\frac{(\gamma+\theta)\phi(\zeta-\gamma/T\mathbb{1}[j=1]-\theta/T)}{\frac{1}{1-\exp(-\lambda K_j)}-\Phi(\zeta-\gamma/T\mathbb{1}[j=1]-\theta/T)}-\frac{(\gamma+\lambda)\phi(\zeta-\theta/T)}{\frac{1}{1-\exp(-\lambda K_{j+1})}-\Phi(\zeta-\theta/T)}\mathbb{1}[j<T-1]-\sum_{t=j+2}^{T-1}\frac{\lambda\phi(\zeta-\theta/T)}{\frac{1}{1-\exp(-\lambda K_t)}-\Phi(\zeta-\theta/T)}.
$$

We set these expressions to zero and solve for $K_j$ iteratively for $j = T-1, T-2, \ldots, 1$. To this end, we write

$$
a_n=\frac{(\gamma+\theta)\phi(\zeta-\gamma/T\mathbb{1}[n=T-1]-\theta/T)}{\frac{1}{1-\exp(-\lambda K_{T-n})}-\Phi(\zeta-\gamma/T\mathbb{1}[n=T-1]-\theta/T)} \tag{A.1}
$$

for $n = 1, 2, \ldots, T-1$, and find

$$
\begin{aligned}
a_1&=\frac{\lambda\theta}{T}\\
a_2&=\frac{\lambda\theta}{T}\left(2+\frac{\gamma+\lambda}{\gamma+\theta}\right)\\
a_3&=\frac{\lambda\theta}{T}\left(3+\frac{\gamma+\lambda}{\gamma+\theta}\left(2+\frac{\gamma+\lambda}{\gamma+\theta}\right)+\frac{\lambda}{\gamma+\theta}\right)\\
a_n&=\frac{\lambda\theta}{T}n+\frac{\gamma+\lambda}{\gamma+\theta}a_{n-1}+\frac{\lambda}{\gamma+\theta}\sum_{j=1}^{n-2}a_j\\
a_{n-1}&=\frac{\lambda\theta}{T}(n-1)+\frac{\gamma+\lambda}{\gamma+\theta}a_{n-2}+\frac{\lambda}{\gamma+\theta}\sum_{j=1}^{n-3}a_j.
\end{aligned}
$$

Combining the last two formulae yields

$$
\begin{aligned}
a_n&=\frac{\lambda\theta}{T}n+\frac{\gamma+\lambda}{\gamma+\theta}a_{n-1}+\frac{\lambda}{\gamma+\theta}a_{n-2}+a_{n-1}-\frac{\lambda\theta}{T}(n-1)-\frac{\gamma+\lambda}{\gamma+\theta}a_{n-2}\\
&=\frac{\lambda\theta}{T}+\frac{2\gamma+\lambda+\theta}{\gamma+\theta}a_{n-1}-\frac{\gamma}{\gamma+\theta}a_{n-2},
\end{aligned} \tag{A.2}
$$

which we can write as

$$
\begin{aligned}
\begin{pmatrix} a_n \\ a_{n-1}\end{pmatrix}&=\begin{pmatrix}\frac{\lambda\theta}{T}\\ 0\end{pmatrix}+\begin{pmatrix}\frac{2\gamma+\lambda+\theta}{\gamma+\theta} & -\frac{\gamma}{\gamma+\theta}\\ 1 & 0\end{pmatrix}\begin{pmatrix}a_{n-1}\\ a_{n-2}\end{pmatrix}\\
&=\sum_{j=0}^{n-3}\begin{pmatrix}\frac{2\gamma+\lambda+\theta}{\gamma+\theta} & -\frac{\gamma}{\gamma+\theta}\\ 1 & 0\end{pmatrix}^j\begin{pmatrix}\frac{\lambda\theta}{T}\\ 0\end{pmatrix}+\begin{pmatrix}\frac{2\gamma+\lambda+\theta}{\gamma+\theta} & -\frac{\gamma}{\gamma+\theta}\\ 1 & 0\end{pmatrix}^{n-2}\begin{pmatrix}\frac{\lambda\theta}{T}\left(2+\frac{\gamma+\lambda}{\gamma+\theta}\right)\\ \frac{\lambda\theta}{T}\end{pmatrix}.
\end{aligned}
$$

This shows that $a_n$ is uniquely defined in terms of $n$, $T$, $\lambda$, $\gamma$, and $\theta$. Note that $a_n$ is strictly positive and increasing in $n$, which follows from (A.2) by induction.

Let $\bar{\zeta}$ be such that $\frac{\phi(\bar{\zeta}-\gamma/T-\theta/T)}{1-\Phi(\bar{\zeta}-\gamma/T-\theta/T)} > \frac{a_{T-1}}{\gamma+\theta}$; this is possible because the inverse Mills ratio $\frac{\phi(x)}{1-\Phi(x)}$ grows without bound as $x$ increases. It implies $\frac{\phi(\bar{\zeta}-\theta/T)}{1-\Phi(\bar{\zeta}-\theta/T)} > \frac{a_n}{\gamma+\theta}$ for all $n = 1, 2, \ldots, T-1$. Let $\zeta \geq \bar{\zeta}$. Solving (A.1) for $K_{T-n}$ gives

$$
K_{T-n}=\frac{1}{\lambda}\log\left(1+\frac{a_n}{(\gamma+\theta)\phi(\zeta-\gamma/T\mathbb{1}[n=T-1]-\theta/T)-a_n\big(1-\Phi(\zeta-\gamma/T\mathbb{1}[n=T-1]-\theta/T)\big)}\right). \tag{A.3}
$$

This choice leaves the dealer with expected utility

$$
\mathbb{E}\big[u\big(\tau(\boldsymbol{p})-\boldsymbol{x}\cdot\boldsymbol{p}\big)\big]\Big|_{x_s=1/T\,\forall s}=-\exp(-\lambda\delta)\prod_{t=1}^{T-1}\frac{1}{1-\frac{a_{T-t}}{\gamma+\theta}\frac{1-\Phi(\zeta-\gamma/T\mathbb{1}[t=1]-\theta/T)}{\phi(\zeta-\gamma/T\mathbb{1}[t=1]-\theta/T)}}.
$$

For this contract to satisfy (IR), this must be at least $u(0) = -1$. Choosing $\delta$ to make this constraint hold with equality, we obtain

$$
\delta=-\frac{1}{\lambda}\sum_{t=1}^{T-1}\log\left(1-\frac{a_{T-t}}{\gamma+\theta}\frac{1-\Phi(\zeta-\gamma/T\mathbb{1}[t=1]-\theta/T)}{\phi(\zeta-\gamma/T\mathbb{1}[t=1]-\theta/T)}\right). \tag{A.4}
$$

Finally, we argue that the client can approximate her first-best payoff arbitrarily closely by setting $\zeta \geq \bar{\zeta}$ sufficiently large, setting $K_{T-n}$ according to (A.3), and setting $\delta$ according to (A.4). Indeed, because $\delta$ is set to satisfy (IR), the client accepts such a contract. Hence, the client's expected costs of execution $\mathbb{E}\big[\tau(\boldsymbol{p})\big]$ are bounded above by $\mathbb{E}\big[\delta+\frac{1}{T}\sum_{t=1}^{T}p_t\big]=\delta+p_0+\frac{\gamma}{T}+\frac{\theta}{T}\sum_{t=1}^{T}\sum_{s=1}^{t}x_s$. Because $K_1, \ldots, K_{T-1}$ are set so that $\boldsymbol{x}=\left(\frac{1}{T},\ldots,\frac{1}{T}\right)^{\top}$ satisfies (IC), this bound becomes $\delta+p_0+\frac{\gamma}{T}+\frac{\theta(T+1)}{2T}$. Applying Proposition 1, the client's first-best costs are exceeded by at most $\delta$. However, according to (A.4), $\delta$ can be made arbitrarily close to zero by setting $\zeta$ sufficiently large. □

**Proof of Lemma 3.** The dealer's expected utility equals

$$
\begin{aligned}
\mathbb{E}[u(\boldsymbol{\tau}\cdot\boldsymbol{p}-\boldsymbol{x}\cdot\boldsymbol{p})]&=-\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{T}(x_t-\tau_t)p_t\right)\right]\\
&=-\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{T}(x_t-\tau_t)\left(\sum_{j=1}^{t}(\theta x_j+\varepsilon_j)+\gamma x_t\right)\right)\right]
\end{aligned}
$$

$$= -\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{T}(x_t-\tau_t)\left(\sum_{j=1}^{t}\theta x_j+\gamma x_t\right)+\lambda\sum_{j=1}^{T}\varepsilon_j\sum_{t=j}^{T}(x_t-\tau_t)\right)\right].$$

Instead of maximizing this expression over $x_t$ subject to $\sum_{t=1}^{T}x_t=1$, we set $X_j=\sum_{t=1}^{j}x_t$ with $X_0=0$ and $X_T=1$, and minimize

$$-\mathbb{E}[u(\boldsymbol{\tau}\cdot\boldsymbol{p}-\boldsymbol{x}\cdot\boldsymbol{p})]=\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{T}(X_t-X_{t-1}-\tau_t)\big(\theta X_t+\gamma(X_t-X_{t-1})\big)-\lambda\sum_{t=1}^{T}\varepsilon_t\left(X_{t-1}-\sum_{j=1}^{t-1}\tau_j\right)\right)\right] \tag{A.5}$$

over $X_1,X_2,\dots,X_{T-1}$. Because $X_t$ needs to be chosen before the price in period $t$ is observable, $X_t$ is $\mathcal{F}_{t-1}$-measurable. We start by conditioning (A.5) on $\mathcal{F}_{T-1}$ and will then go backward subsequently. From the law of iterated expectations, we obtain

$$\begin{aligned}
&-\mathbb{E}[u(\boldsymbol{\tau}\cdot\boldsymbol{p}-\boldsymbol{x}\cdot\boldsymbol{p})]\\
&=\mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{T}(X_t-X_{t-1}-\tau_t)\big(\theta X_t+\gamma(X_t-X_{t-1})\big)-\lambda\sum_{t=1}^{T-1}\varepsilon_{t+1}\left(X_t-\sum_{j=1}^{t}\tau_j\right)\right)\middle|\mathcal{F}_{T-1}\right]\right]\\
&=\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{T}(X_t-X_{t-1}-\tau_t)\big(\theta X_t+\gamma(X_t-X_{t-1})\big)-\lambda\sum_{t=1}^{T-2}\varepsilon_{t+1}\left(X_t-\sum_{j=1}^{t}\tau_j\right)\right)\mathbb{E}\left[\exp\left(-\lambda\varepsilon_T\left(X_{T-1}-\sum_{j=1}^{T-1}\tau_j\right)\right)\middle|\mathcal{F}_{T-1}\right]\right]\\
&=\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{T}(X_t-X_{t-1}-\tau_t)\big(\theta X_t+\gamma(X_t-X_{t-1})\big)-\lambda\sum_{t=1}^{T-2}\varepsilon_{t+1}\left(X_t-\sum_{j=1}^{t}\tau_j\right)+\frac{\lambda^2}{2}\left(X_{T-1}-\sum_{j=1}^{T-1}\tau_j\right)^2\right)\right]\\
&=\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{T-2}(X_t-X_{t-1}-\tau_t)\big(\theta X_t+\gamma(X_t-X_{t-1})\big)-\lambda\sum_{t=1}^{T-2}\varepsilon_{t+1}\left(X_t-\sum_{j=1}^{t}\tau_j\right)\right)\right.\\
&\qquad\left.\times\exp\left(\lambda\sum_{t=T-1}^{T}(X_t-X_{t-1}-\tau_t)\big(\theta X_t+\gamma(X_t-X_{t-1})\big)+\frac{\lambda^2}{2}\left(X_{T-1}-\sum_{j=1}^{T-1}\tau_j\right)^2\right)\right],
\end{aligned}$$

where we used that $X_{T-1}$ is $\mathcal{F}_{T-2}$-measurable and thus also $\mathcal{F}_{T-1}$-measurable, along with the fact that $\varepsilon_T$ is independent of $\mathcal{F}_{T-1}$. We note that $X_{T-1}$ appears only in the last line, but not in the penultimate line, and the dependence on $X_{T-1}$ is quadratic. Therefore, the optimal $X_{T-1}$ is given by the first-order condition

$$-\lambda\big(\theta X_T+\gamma(X_T-X_{T-1})\big)-\lambda\gamma(X_T-X_{T-1}-\tau_T)+\lambda(\theta X_{T-1}+\gamma X_{T-1}-\gamma X_{T-2})+\lambda(\theta+\gamma)(X_{T-1}-X_{T-2}-\tau_{T-1})+\lambda^2\left(X_{T-1}-\sum_{j=1}^{T-1}\tau_j\right)=0,$$

which we rewrite as

$$-\big(\theta+2\gamma\big)X_T+\big(\lambda+2\theta+4\gamma\big)X_{T-1}-(\theta+2\gamma)X_{T-2}=-\gamma\tau_T+(\theta+\gamma)\tau_{T-1}+\lambda\sum_{j=1}^{T-1}\tau_j, \tag{A.6}$$

This implies that $X_{T-1}$ is $\mathcal{F}_{T-3}$-measurable because so is $X_{T-2}$ and all other terms are deterministic. Next, we condition on $\mathcal{F}_{T-2}$ to obtain

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{T-2}(X_t-X_{t-1}-\tau_t)\big(\theta X_t+\gamma(X_t-X_{t-1})\big)-\lambda\sum_{t=1}^{T-2}\varepsilon_{t+1}\left(X_t-\sum_{j=1}^{t}\tau_j\right)\right)\right.\\
&\qquad\left.\times\exp\left(\lambda\sum_{t=T-1}^{T}(X_t-X_{t-1}-\tau_t)\big(\theta X_t+\gamma(X_t-X_{t-1})\big)+\frac{\lambda^2}{2}\left(X_{T-1}-\sum_{j=1}^{T-1}\tau_j\right)^2\right)\right]\\
&=\mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{T-2}(X_t-X_{t-1}-\tau_t)\big(\theta X_t+\gamma(X_t-X_{t-1})\big)-\lambda\sum_{t=1}^{T-2}\varepsilon_{t+1}\left(X_t-\sum_{j=1}^{t}\tau_j\right)\right)\middle|\mathcal{F}_{T-2}\right]\right.\\
&\qquad\left.\times\exp\left(\lambda\sum_{t=T-1}^{T}(X_t-X_{t-1}-\tau_t)\big(\theta X_t+\gamma(X_t-X_{t-1})\big)+\frac{\lambda^2}{2}\left(X_{T-1}-\sum_{j=1}^{T-1}\tau_j\right)^2\right)\right]\\
&=\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{T}(X_t-X_{t-1}-\tau_t)\big(\theta X_t+\gamma(X_t-X_{t-1})\big)-\lambda\sum_{t=1}^{T-3}\varepsilon_{t+1}\left(X_t-\sum_{j=1}^{t}\tau_j\right)\right)\right.\\
&\qquad\left.\times\exp\left(\frac{\lambda^2}{2}\left(X_{T-2}-\sum_{j=1}^{T-2}\tau_j\right)^2+\frac{\lambda^2}{2}\left(X_{T-1}-\sum_{j=1}^{T-1}\tau_j\right)^2\right)\right].
\end{aligned}$$

The terms within the exponential function that depend on $X_{T-2}$ are

$$\lambda\sum_{t=T-2}^{T-1}(X_t-X_{t-1}-\tau_t)\big(\theta X_t+\gamma(X_t-X_{t-1})\big)+\frac{\lambda^2}{2}\left(X_{T-2}-\sum_{j=1}^{T-2}\tau_j\right)^2$$

so that the first-order condition implies

$$-(\theta+2\gamma)X_{T-1}+(\lambda+2\theta+4\gamma)X_{T-2}-(\theta+2\gamma)X_{T-3}=-\gamma\tau_{T-1}+(\theta+\gamma)\tau_{T-2}+\lambda\sum_{j=1}^{T-2}\tau_j.$$

Because $X_{T-1}$ is a function of $X_{T-2}$ in the optimum by (A.6) and $X_{T-3}$ is $\mathcal{F}_{T-4}$-measurable while all other terms are deterministic, this implies that $X_{T-2}$ is $\mathcal{F}_{T-4}$-measurable. And, using again that $X_{T-1}$ is a function of $X_{T-2}$, this implies that $X_{T-1}$ is $\mathcal{F}_{T-4}$-measurable as well. Continuing this procedure, we obtain in the end that all $X_t$ are deterministic and satisfy

$$-(\theta+2\gamma)X_{t+1}+(\lambda+2\theta+4\gamma)X_t-(\theta+2\gamma)X_{t-1}=-\gamma\tau_{t+1}+(\theta+\gamma)\tau_t+\lambda\sum_{j=1}^{t}\tau_j,\quad t=1,2,\dots,T-1. \tag{A.7}$$

This linear system of equations can be written as $AX = E\tau$, hence $x = FX = FA^{-1}E\tau$, using the $T\times T$ matrices $A$, $E$, and $F$ (written here with general $\sigma$, for reference in the main text) given by

$$A=\begin{pmatrix} \lambda\sigma^2+2\theta+4\gamma & -(\theta+2\gamma) & 0 & 0 & \cdots \\ -(\theta+2\gamma) & \lambda\sigma^2+2\theta+4\gamma & -(\theta+2\gamma) & 0 & \cdots \\ 0 & -(\theta+2\gamma) & \lambda\sigma^2+2\theta+4\gamma & -(\theta+2\gamma) & \\ \vdots & & \ddots & \ddots & \ddots \\ & & -(\theta+2\gamma) & \lambda\sigma^2+2\theta+4\gamma & -(\theta+2\gamma) \\ 0 & \cdots & & 0 & 1 \end{pmatrix},$$

$$E=\begin{pmatrix} \theta+\gamma+\lambda\sigma^2 & -\gamma & 0 & 0 & 0 & \cdots \\ \lambda\sigma^2 & \theta+\gamma+\lambda\sigma^2 & -\gamma & 0 & 0 & \cdots \\ \lambda\sigma^2 & \lambda\sigma^2 & \theta+\gamma+\lambda\sigma^2 & -\gamma & 0 & \cdots \\ \vdots & \vdots & & \ddots & \ddots & \\ \lambda\sigma^2 & \lambda\sigma^2 & & & \theta+\gamma+\lambda\sigma^2 & -\gamma \\ 1 & \cdots & & & 1 & 1 \end{pmatrix}, \tag{A.8}$$

$$F=\begin{pmatrix} 1 & 0 & 0 & 0 & \cdots \\ -1 & 1 & 0 & 0 & \cdots \\ 0 & -1 & 1 & 0 & \cdots \\ \vdots & & \ddots & \ddots & \\ 0 & \cdots & & -1 & 1 \end{pmatrix}. \quad \square$$

**Proof of Proposition 4.** We can write $A=\tilde I\tilde A$, where

$$\tilde A=\begin{pmatrix} \lambda\sigma^2+2\theta+4\gamma & -(\theta+2\gamma) & 0 & 0 & \cdots \\ -(\theta+2\gamma) & \lambda\sigma^2+2\theta+4\gamma & -(\theta+2\gamma) & 0 & \cdots \\ 0 & -(\theta+2\gamma) & \lambda\sigma^2+2\theta+4\gamma & -(\theta+2\gamma) & \\ \vdots & & \ddots & \ddots & \ddots \\ & & -(\theta+2\gamma) & \lambda\sigma^2+2\theta+4\gamma & -(\theta+2\gamma) \\ 0 & \cdots & & 0 & \lambda\sigma^2+2\theta+4\gamma \end{pmatrix},\quad \tilde I=\begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & \frac{1}{\lambda\sigma^2+2\theta+4\gamma} \end{pmatrix}.$$

Note that $\tilde A$ is a Z-matrix (i.e., a square matrix where all off-diagonal entries are nonpositive). In fact, $\tilde A$ is an M-matrix. Indeed, we can express $\tilde A=(\lambda\sigma^2+2\theta+4\gamma)I-\tilde B$, where

$$\tilde B=\begin{pmatrix} 0 & \theta+2\gamma & 0 & 0 & \cdots \\ \theta+2\gamma & 0 & \theta+2\gamma & 0 & \cdots \\ 0 & \theta+2\gamma & 0 & \theta+2\gamma & \\ \vdots & & \ddots & \ddots & \ddots \\ & & \theta+2\gamma & 0 & \theta+2\gamma \\ 0 & \cdots & & 0 & 0 \end{pmatrix}$$

is a matrix whose eigenvalues (i.e., $\pm(\theta+2\gamma)\sqrt{2}$ and 0) are bounded in magnitude by $\lambda\sigma^2+2\theta+4\gamma$. Because $\tilde A$ is an M-matrix, its inverse is a nonnegative matrix. Hence, $A^{-1}=\tilde A^{-1}\tilde I^{-1}$ is also nonnegative. Next, observe that $E\tau^{TWAP}=\frac{1}{T}(\theta+\lambda\sigma^2,\theta+2\lambda\sigma^2,\dots,\theta+(T-1)\lambda\sigma^2,T)^\top$, and $AF^{-1}x^{FB}=\frac{1}{T}(\lambda\sigma^2,2\lambda\sigma^2,\dots,(T-1)\lambda\sigma^2,T)^\top$, so $E\tau^{TWAP}\geq AF^{-1}x^{FB}$ (where $\geq$ is in the component-wise sense). Using the fact that $A^{-1}$ is nonnegative,

$$F^{-1}FA^{-1}E\tau^{TWAP}=A^{-1}E\tau^{TWAP}\geq F^{-1}x^{FB},$$

which is precisely what it means for $FA^{-1}E\tau^{TWAP}$ to be frontloaded relative to $x^{FB}$. For the final claim, note that all the inequalities can be replaced with equalities if and only if $\theta=0$. □

**Proof of Proposition 5.** The client's expected cost under a contract $\tau$ is

$$\begin{aligned}\mathbb{E}[\tau\cdot p]&=\mathbb{E}\left[\sum_{t=1}^{T}\tau_t p_t\right]=\mathbb{E}\left[\sum_{t=1}^{T}\tau_t\left(p_0+\sum_{j=1}^{t}(\theta x_j+\varepsilon_j)+\gamma x_t\right)\right]=p_0+\sum_{t=1}^{T}\tau_t\left(\sum_{j=1}^{t}\theta x_j+\gamma x_t\right)\\ &=p_0+\theta\sum_{t=1}^{T}\tau_t X_t+\gamma\sum_{t=1}^{T}\tau_t(X_t-X_{t-1})=p_0+\theta\tau^\top A^{-1}E\tau+\gamma\tau^\top FA^{-1}E\tau=p_0+\frac{1}{2}\tau^\top M\tau\end{aligned} \tag{A.9}$$

where $F$ and $M$ are defined in (A.8) and (3), respectively. Therefore, we minimize $\frac{1}{2}\tau^\top M\tau$ subject to $\tau^\top\mathbb{1}=1$, where $\mathbb{1}=(1,1,\dots,1)^\top$ denotes a $T$-dimensional vector of ones. From the Lagrange method (and using the symmetry of $M$), it follows that

$$M\tau^*-\mu\mathbb{1}=0,$$

hence $\tau^*=\mu M^{-1}\mathbb{1}$ and $\mathbb{1}^\top\tau^*=\mu\mathbb{1}^\top M^{-1}\mathbb{1}=1$. We obtain $\mu=\frac{1}{\mathbb{1}^\top M^{-1}\mathbb{1}}$ and thus $\tau^*=\frac{1}{\mathbb{1}^\top M^{-1}\mathbb{1}}M^{-1}\mathbb{1}$ and $x^*=FA^{-1}E\tau^*=\frac{1}{\mathbb{1}^\top M^{-1}\mathbb{1}}FA^{-1}EM^{-1}\mathbb{1}$, using Lemma 3. We can compute the client's expected costs of execution under $\tau^*=\frac{1}{\mathbb{1}^\top M^{-1}\mathbb{1}}M^{-1}\mathbb{1}$ as

$$p_0+\frac{1}{2}(\tau^*)^\top M\tau^*=p_0+\frac{1}{2}(\mathbb{1}^\top M^{-1}\mathbb{1})^{-2}\mathbb{1}^\top(M^{-1})^\top MM^{-1}\mathbb{1}=p_0+\frac{1}{2\mathbb{1}^\top M^{-1}\mathbb{1}}.\quad\square$$

**Proof of Corollary 6.** Claim *(i)*: It follows from (A.7) with $\gamma = \lambda = 0$ that

$$-x_{t+1} + x_t = \tau_t \quad \text{for } t = 1, 2, \dots, T-1,$$

which implies

$$x_t = x_{t+1} + \tau_t = x_{t+2} + \tau_t + \tau_{t+1} = \cdots = c + \sum_{s=t}^{T} \tau_s$$

for some constant $c$ and all $t$. To determine $c$, we use that $\sum_{t=1}^{T} x_t = 1$, hence

$$c = \frac{1}{T} - \frac{1}{T}\sum_{t=1}^{T}\sum_{s=t}^{T} \tau_s = \frac{1}{T} - \sum_{s=1}^{T} \frac{s}{T}\tau_s.$$

Claim *(ii)*: Define $G$ as the $T \times T$ matrix whose first column and last row are all 1, and otherwise the $ij$ entry is $i(T+1-j)/T$ for $j \geq i$ and $i(T+1-j)/T - i + j$ for $j < i$. In this case of $\gamma = \lambda = 0$, it can be checked that $E = AG$. It follows that $A^{-1}E = G$. Note that for all $t$:

$$G_{t1} + G^{\top}_{t1} + G_{tT} + G^{\top}_{tT} = G_{t1} + G_{1t} + G_{tT} + G_{Tt} = 1 + (T+1-t)/T + t/T + 1 = 3 + 1/T.$$

Next, define $\boldsymbol{v} = \left(\frac{1}{2}, 0, \dots, 0, \frac{1}{2}\right)^{\top}$. We compute

$$M\boldsymbol{v} = \theta(A^{-1}E + E^{\top}(A^{-1})^{\top})\boldsymbol{v} = \theta(G + G^{\top})\boldsymbol{v} = \frac{\theta}{2}(3 + 1/T)\mathbb{1}.$$

This implies that $M^{-1}\mathbb{1} = \frac{2T}{\theta(3T+1)}\boldsymbol{v}$ and $\mathbb{1}^{\top}M^{-1}\mathbb{1} = \frac{2T}{\theta(3T+1)}$. Thus, $\boldsymbol{\tau}^* = \frac{1}{\mathbb{1}^{\top}M^{-1}\mathbb{1}}M^{-1}\mathbb{1} = \boldsymbol{v}$, as claimed. We also compute

$$\boldsymbol{x}^* = FA^{-1}E\boldsymbol{v} = FG\boldsymbol{v} = \frac{1}{2}F(1 + 1/T, 1 + 2/T, \dots, 2)^{\top} = \frac{1}{2}(1 + 1/T, 1/T, \dots, 1/T)^{\top},$$

as claimed. Finally, the client's expected costs of execution are $p_0 + \frac{1}{2\mathbb{1}^{\top}M^{-1}\mathbb{1}} = p_0 + \frac{\theta(3T+1)}{4T}$, as claimed. □

**Proof of Corollary 7.** *Claim (i):* It follows from (A.7) with $\theta = \lambda = 0$ that

$$-2x_{t+1} + 2x_t = -\tau_{t+1} + \tau_t \quad \text{for } t = 1, 2, \dots, T-1,$$

which implies

$$x_t = x_{t+1} - \frac{1}{2}\tau_{t+1} + \frac{1}{2}\tau_t = x_{t+2} - \frac{1}{2}\tau_{t+2} + \frac{1}{2}\tau_t = \cdots = c + \frac{1}{2}\tau_t$$

for some constant $c$ and all $t$. To determine $c$, we use that $\sum_{t=1}^{T} x_t = 1$, hence

$$c = \frac{1}{T} - \frac{1}{2T}\sum_{t=1}^{T}\tau_t = \frac{1}{2T}.$$

*Claim (ii):* For additional generality in this part of the proof, we deliberately do not use the assumption that $\lambda = 0$. To prove $\boldsymbol{\tau}^* = \left(\frac{1}{T}, \dots, \frac{1}{T}\right)^{\top}$, it is enough to show that $\mathbb{1}$ is an eigenvector of $M$ because $M\boldsymbol{\tau}^* = \frac{1}{\mathbb{1}^{\top}M^{-1}\mathbb{1}}\mathbb{1}$ by Proposition 5. To aid in showing that, we first define

$$\boldsymbol{v}_1 = \left(\lambda, 2\lambda, \dots, (T-1)\lambda, T\right)^{\top}, \ \boldsymbol{v}_2 = (0, \dots, 0, 1)^{\top} \ \text{ and } \ \boldsymbol{v}_3 = (1, 2, \dots, T)^{\top}.$$

In this case of $\theta = 0$, observe that $\boldsymbol{v}_1 = A\boldsymbol{v}_3$, which implies $A^{-1}\boldsymbol{v}_1 = \boldsymbol{v}_3$. Observe also that $\boldsymbol{v}_2^{\top} = \boldsymbol{v}_2^{\top}A$, which implies $(A^{-1})^{\top}\boldsymbol{v}_2 = \boldsymbol{v}_2$. We then compute

$$M\mathbb{1} = \gamma FA^{-1}E\mathbb{1} + \gamma E^{\top}(A^{-1})^{\top}F^{\top}\mathbb{1} = \gamma FA^{-1}\boldsymbol{v}_1 + \gamma E^{\top}(A^{-1})^{\top}\boldsymbol{v}_2 = \gamma F\boldsymbol{v}_3 + \gamma E^{\top}\boldsymbol{v}_2 = \gamma\mathbb{1} + \gamma\mathbb{1} = 2\gamma\mathbb{1}, \tag{A.10}$$

establishing that $\mathbb{1}$ is an eigenvector of $M$, as required. For the dealer's trading strategy, we deduce from Proposition 5 that

$$\boldsymbol{x}^* = \frac{1}{\mathbb{1}^{\top}M^{-1}\mathbb{1}}FA^{-1}EM^{-1}\mathbb{1} = \frac{2\gamma}{T}FA^{-1}E\frac{1}{2\gamma}\mathbb{1} = \frac{1}{T}\mathbb{1},$$

where the second equality uses (A.10) to obtain $\frac{1}{\mathbb{1}^{\top}M^{-1}\mathbb{1}} = \frac{2\gamma}{T}$ and $M^{-1}\mathbb{1} = \frac{1}{2\gamma}\mathbb{1}$. Finally, we compute the client's expected costs of execution as $p_0 + \frac{1}{2\mathbb{1}^{\top}M^{-1}\mathbb{1}} = p_0 + \frac{\gamma}{T}$. □

**Proof of Corollary 8.** *Claim (i):* Dividing (A.7) by $\lambda$ and then letting $\lambda$ go to infinity gives $X_t = \sum_{j=1}^{t}\tau_j$ for $t = 1, 2, \dots, T-1$, hence $x_t = \tau_t$ for all $t = 1, 2, \dots, T$.

*Claim (ii):* Let $Q$ be the lower-triangular matrix with all entries of 1 on and below the diagonal; let $\Lambda$ be the diagonal matrix that has $\lambda$ everywhere on its diagonal except for the last entry, which equals 1:

$$Q = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}, \quad \Lambda = \begin{pmatrix} \lambda & 0 & \cdots & 0 \\ 0 & \lambda & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

We begin by showing $\lim_{\lambda\to\infty} A^{-1}E = Q$. To this end, we note that $\Lambda^{-1}$ is a diagonal matrix that has $1/\lambda$ everywhere on its diagonal except for the last entry which equals 1, and we then compute

$$\lim_{\lambda\to\infty} \Lambda^{-1}E = \lim_{\lambda\to\infty} \begin{pmatrix} (\theta+\gamma+\lambda)/\lambda & -\gamma/\lambda & 0 & 0 & 0 & \cdots \\ 1 & (\theta+\gamma+\lambda)/\lambda & -\gamma/\lambda & 0 & 0 & \cdots \\ 1 & 1 & (\theta+\gamma+\lambda)/\lambda & -\gamma/\lambda & 0 & \cdots \\ \vdots & \vdots & & \ddots & \ddots & \\ 1 & 1 & & & (\theta+\gamma+\lambda)/\lambda & -\gamma/\lambda \\ 1 & \cdots & & & 1 & 1 \end{pmatrix} = Q,$$

$$\lim_{\lambda\to\infty} \Lambda^{-1}A = \lim_{\lambda\to\infty} \begin{pmatrix} (\lambda+2\theta+4\gamma)/\lambda & -(\theta+2\gamma)/\lambda & 0 & 0 & \cdots \\ -(\theta+2\gamma)/\lambda & (\lambda+2\theta+4\gamma)/\lambda & -(\theta+2\gamma)/\lambda & 0 & \cdots \\ 0 & -(\theta+2\gamma)/\lambda & (\lambda+2\theta+4\gamma)/\lambda & -(\theta+2\gamma)/\lambda & \\ \vdots & & \ddots & \ddots & \ddots \\ & & -(\theta+2\gamma)/\lambda & (\lambda+2\theta+4\gamma)/\lambda & -(\theta+2\gamma)/\lambda \\ 0 & \cdots & & 0 & 1 \end{pmatrix} = I,$$

where $I$ denotes the $T\times T$ identity matrix. The latter implies $\lim_{\lambda\to\infty} A^{-1}\Lambda = \lim_{\lambda\to\infty}\left(\Lambda^{-1}A\right)^{-1} = I$. Thus, we obtain

$$\lim_{\lambda\to\infty} A^{-1}E = \lim_{\lambda\to\infty} A^{-1}\Lambda\Lambda^{-1}E = \left(\lim_{\lambda\to\infty} A^{-1}\Lambda\right)\left(\lim_{\lambda\to\infty} \Lambda^{-1}E\right) = IQ = Q.$$

So from (3), we deduce

$$\begin{aligned} \lim_{\lambda\to\infty} M &= \theta\left(\lim_{\lambda\to\infty} A^{-1}E\right) + \theta\left(\lim_{\lambda\to\infty} A^{-1}E\right)^{\top} + \gamma\left(F\lim_{\lambda\to\infty} A^{-1}E\right) + \gamma\left(F\lim_{\lambda\to\infty} A^{-1}E\right)^{\top} \\ &= \theta Q + \theta Q^{\top} + \gamma FQ + \gamma(FQ)^{\top} = \theta Q + \theta Q^{\top} + 2\gamma I = \begin{pmatrix} 2\theta+2\gamma & \theta & \cdots & \theta \\ \theta & 2\theta+2\gamma & \cdots & \theta \\ \vdots & & \ddots & \vdots \\ \theta & \theta & \cdots & 2\theta+2\gamma \end{pmatrix}. \end{aligned}$$

Thus, $\left(\lim_{\lambda\to\infty} M\right)\mathbb{1} = [2\gamma + \theta(T+1)]\mathbb{1}$, which implies that

$$\lim_{\lambda\to\infty} M^{-1}\mathbb{1} = \frac{1}{2\gamma+\theta(T+1)}\left(\lim_{\lambda\to\infty} M^{-1}\right)\left(\lim_{\lambda\to\infty} M\right)\mathbb{1} = \frac{1}{2\gamma+\theta(T+1)}\left(\lim_{\lambda\to\infty} M^{-1}M\right)\mathbb{1} = \frac{1}{2\gamma+\theta(T+1)}\mathbb{1},$$

and hence $\lim_{\lambda\to\infty} \mathbb{1}^{\top}M^{-1}\mathbb{1} = \frac{T}{2\gamma+\theta(T+1)}$. We can then compute

$$\begin{aligned} \lim_{\lambda\to\infty} \boldsymbol{\tau}^* &= \lim_{\lambda\to\infty}\left(\frac{1}{\mathbb{1}^{\top}M^{-1}\mathbb{1}}M^{-1}\mathbb{1}\right) = \frac{1}{\lim_{\lambda\to\infty}\mathbb{1}^{\top}M^{-1}\mathbb{1}}\lim_{\lambda\to\infty} M^{-1}\mathbb{1} = \frac{2\gamma+\theta(T+1)}{T}\frac{1}{2\gamma+\theta(T+1)}\mathbb{1} = \frac{1}{T}\mathbb{1}, \\ \lim_{\lambda\to\infty} \boldsymbol{x}^* &= \lim_{\lambda\to\infty} FA^{-1}E\boldsymbol{\tau}^* = F\left(\lim_{\lambda\to\infty} A^{-1}E\right)\left(\lim_{\lambda\to\infty}\boldsymbol{\tau}^*\right) = FQ\left(\frac{1}{T}\mathbb{1}\right) = I\left(\frac{1}{T}\mathbb{1}\right) = \frac{1}{T}\mathbb{1}, \end{aligned}$$

each of which is as claimed. Finally, the client's expected costs of execution converge to

$$\lim_{\lambda\to\infty}\left(p_0 + \frac{1}{2\mathbb{1}^{\top}M^{-1}\mathbb{1}}\right) = p_0 + \frac{1}{2}\frac{2\gamma+\theta(T+1)}{T} = p_0 + \frac{\gamma}{T} + \frac{\theta(T+1)}{2T},$$

which is also as claimed. □

**Proof of Corollary 9.** Define the $T\times T$ anti-diagonal matrix

$$P = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 1 & 0 \\ & \ddots & & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix},$$

and we begin by showing that $A(\theta I + \gamma F)^{-1}PE^{\top}$ is symmetric. To that end, define $T\times T$ matrices

$$A_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & & \vdots \\ 0 & 1 & 0 & \ddots & 0 \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \cdots & & 0 & 0 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & 1 \end{pmatrix},$$

so that $A = (\lambda + 2\theta + 4\gamma)A_1 - (\theta + 2\gamma)A_2 + A_3$. Define also the $T\times T$ matrices

$$E_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & 0 \end{pmatrix}, \quad E_2 = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 1 & 0 & & \vdots \\ 0 & 1 & \ddots & 0 \\ \vdots & & \ddots & 0 \\ 0 & 0 & \cdots & 0 \end{pmatrix}, \quad E_3 = \begin{pmatrix} 0 & 1 & 1 & \cdots & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix}, \quad E_4 = \begin{pmatrix} 0 & 0 & \cdots & 1 \\ 0 & 0 & & 1 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix},$$

so that $E^{\top} = (\lambda + \gamma + \theta)E_1 - \gamma E_2 + \lambda E_3 + E_4$. Observe that we can also write

$$(\theta I + \gamma F)^{-1}P = \begin{pmatrix} 0 & \cdots & 0 & b_1 \\ 0 & & b_1 & b_2 \\ & \ddots & & \vdots \\ b_1 & \cdots & b_{T-1} & b_T \end{pmatrix},$$

where $b_{t+1} = \gamma/(\gamma+\theta)b_t$ for all $t \in \{1, 2 \dots, T-1\}$.[31] Each of the following matrices is symmetric:

$$A_1(\theta I + \gamma F)^{-1} P E_1 = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & b_1 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & b_1 & \cdots & b_{T-2} & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}$$

$$A_1(\theta I + \gamma F)^{-1} P E_2 = \begin{pmatrix} 0 & \cdots & 0 & b_1 & 0 \\ 0 & \cdots & b_1 & b_2 & 0 \\ \vdots & \ddots & \vdots & \vdots & \\ b_1 & b_2 & \cdots & b_{T-1} & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}$$

$$A_1(\theta I + \gamma F)^{-1} P E_3 = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & b_1 & \cdots & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & \sum_{t=1}^{T-3} b_t & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}$$

$$A_2(\theta I + \gamma F)^{-1} P E_1 = \begin{pmatrix} 0 & \cdots & & & 0 & b_1 & 0 \\ 0 & \cdots & & & b_1 & b_2 & 0 \\ 0 & \cdots & & & b_2 & b_1 + b_3 & 0 \\ 0 & \cdots & & & b_1 + b_3 & b_2 + b_4 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ b_1 & b_2 & b_1 + b_3 & b_2 + b_4 & \cdots & b_{T-3} + b_{T-1} & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}$$

$$A_2(\theta I + \gamma F)^{-1} P E_2 = \begin{pmatrix} 0 & \cdots & 0 & b_1 & b_2 & 0 \\ 0 & & b_1 & b_2 & b_1 + b_3 & 0 \\ 0 & \ddots & b_2 & b_1 + b_3 & b_2 + b_4 & 0 \\ b_1 & \ddots & \ddots & \ddots & \vdots & \vdots \\ b_2 & b_1 + b_3 & b_2 + b_4 & \cdots & b_{T-2} + b_T & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}$$

$$A_2(\theta I + \gamma F)^{-1} P E_3 = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & b_1 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & b_1 & \cdots & \sum_{t=1}^{T-2} b_t & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}$$

$$A_3(\theta I + \gamma F)^{-1} P E_4 = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sum_{t=1}^{T} b_t \end{pmatrix}$$

To show that $A(\theta I + \gamma F)^{-1} P E^\top$ is indeed symmetric, it only remains to show that $[(\lambda + 2\theta + 4\gamma)A_1 - (\theta + 2\gamma)A_2](\theta I + \gamma F)^{-1} P E_4$ is the transpose of $A_3(\theta I + \gamma F)^{-1} P[(\lambda + \gamma + \theta)E_1 - \gamma E_2 + \lambda E_3]$. The former has nonzero entries only in the first $T-1$ entries of its last column, while the latter has nonzero entries only in the first $T-1$ entries of its last row. We check these nonzero entries. For $t < T$, the $(t, T)$-element of $[(\lambda + 2\theta + 4\gamma)A_1 - (\theta + 2\gamma)A_2](\theta I + \gamma F)^{-1} P E_4$ is

$$(\lambda + 2\theta + 4\gamma) \sum_{s=1}^{t} b_s - (\theta + 2\gamma) \left[ \sum_{s=1}^{t+1} b_s + \sum_{s=1}^{t-1} b_s \right] = -(\theta + 2\gamma) b_{t+1} + (\lambda + \theta + 2\gamma) b_t + \lambda \sum_{s=1}^{t-1} b_s.$$

For $t < T$, the $(T, t)$-element of $A_3(\theta I + \gamma F)^{-1} P[(\lambda + \gamma + \theta)E_1 - \gamma E_2 + \lambda E_3]$ is

$$(\lambda + \gamma + \theta) b_t - \gamma b_{t+1} + \lambda \sum_{s=1}^{t-1} b_s.$$

Computing the difference:

$$\left[ (\lambda + \gamma + \theta) b_t - \gamma b_{t+1} + \lambda \sum_{s=1}^{t-1} b_s \right] - \left[ -(\theta + 2\gamma) b_{t+1} + (\lambda + \theta + 2\gamma) b_t + \lambda \sum_{s=1}^{t-1} b_s \right] = (\theta + \gamma) b_{t+1} - \gamma b_t,$$

which equals zero because $b_{t+1} = \gamma/(\gamma+\theta)b_t$. We conclude that $A(\theta I + \gamma F)^{-1} P E^\top$ is symmetric, as claimed. Mathematically,

$$A(\theta I + \gamma F)^{-1} P E^\top = E P^\top (\theta I^\top + \gamma F^\top)^{-1} A^\top = E P (\theta I + \gamma F^\top)^{-1} A^\top,$$

which implies

$$P E^\top (A^{-1})^\top (\theta I + \gamma F^\top) = (\theta I + \gamma F) A^{-1} E P.$$

[31] We also have $b_1 = 1/(\gamma + \theta)$, although that will not be relevant for the following arguments.

Letting $M_1 = (\theta I + \gamma F)A^{-1}E$, we can rewrite this as $M_1 P = P M_1^\top$. Because $P^{-1} = P$, we also have $P M_1 = M_1^\top P$. Together, these imply $\big(M_1 + M_1^\top\big)P = P\big(M_1 + M_1^\top\big)$. Then using $M = M_1 + M_1^\top$, we conclude $MP = PM$, hence $MPM^{-1}\mathbb{1} = \mathbb{1}$, and hence $PM^{-1}\mathbb{1} = M^{-1}\mathbb{1}$. Therefore, we conclude

$$P\boldsymbol{\tau}^* = \frac{1}{\mathbb{1}^\top M^{-1}\mathbb{1}} PM^{-1}\mathbb{1} = \frac{1}{\mathbb{1}^\top M^{-1}\mathbb{1}} M^{-1}\mathbb{1} = \boldsymbol{\tau}^*,$$

hence $\tau_j = \tau_{T+1-j}$ for all $j = 1, \dots, T$. □

**Proof of Proposition 10.** Throughout, we assume $\theta > 0$ since the results for $\theta = 0$ follow from Corollary 7*(ii)*, whose proof does not require assuming $\lambda = 0$. We note that a model with price-shock variance $\sigma_k^2 = \frac{T\sigma^2}{T_k}$ is equivalent to a model with price-shock variance normalized to 1 while $\lambda$ is replaced by $\frac{\lambda T\sigma^2}{T_k}$. Therefore, (A.7) becomes

$$X_q^k + \frac{\theta + 2\gamma}{\lambda T\sigma^2} \frac{2X_q^k - X_{q+\frac{1}{T_k}}^k - X_{q-\frac{1}{T_k}}^k}{1/T_k} = \frac{\gamma}{\lambda T\sigma^2} \frac{2V_q^k - V_{q+\frac{1}{T_k}}^k - V_{q-\frac{1}{T_k}}^k}{1/T_k} + \frac{\theta}{\lambda T\sigma^2} \frac{V_q^k - V_{q-\frac{1}{T_k}}^k}{1/T_k} + V_q^k \tag{A.11}$$

for all $q \in (0,1)$, where $X_q^k = \sum_{t=1}^{\lceil qT_k \rceil} x_t^k$ and $V_q^k = \sum_{t=1}^{\lceil qT_k \rceil} \tau_t^k$ for $\boldsymbol{\tau}^k \in \Delta^{T_k}$ and a dealer strategy $\boldsymbol{x}^k$ in the $k$th model. Throughout this analysis, we assume that the limiting processes $X_q = \lim_{k\to\infty} X_q^k$ and $V_q = \lim_{k\to\infty} V_q^k$ exist and are continuously differentiable, except for jumps at 0 and 1. Thus, given any $q \in (0,1)$,

$$\begin{aligned}
\lim_{k\to\infty} \frac{V_q^k - V_{q-\frac{1}{T_k}}^k}{1/T_k} &= \lim_{\epsilon\to 0} \frac{V_q - V_{q-\epsilon}}{\epsilon} = \dot V_q \\
\lim_{k\to\infty} \frac{2V_q^k - V_{q+\frac{1}{T_k}}^k - V_{q-\frac{1}{T_k}}^k}{1/T_k} &= \lim_{\epsilon\to 0} \left( \frac{V_q - V_{q-\epsilon}}{\epsilon} - \frac{V_{q+\epsilon} - V_q}{\epsilon} \right) = \dot V_q - \dot V_q = 0, \\
\lim_{k\to\infty} \frac{2X_q^k - X_{q+\frac{1}{T_k}}^k - X_{q-\frac{1}{T_k}}^k}{1/T_k} &= \lim_{\epsilon\to 0} \frac{-(X_{q+\epsilon} - X_q) + (X_q - X_{q-\epsilon})}{\epsilon} = -\dot X_q + \dot X_q = 0.
\end{aligned}$$

Thus, it follows from (A.11) that

$$X_q = \frac{\theta}{\lambda T\sigma^2} \dot V_q + V_q \tag{A.12}$$

for all $q \in (0,1)$. Furthermore, the jumps of $V_q$ at 0 and 1 are given by

$$\begin{aligned}
V_{0+} &= \lim_{q\searrow 0} V_q = \lim_{k\to\infty} \sum_{t=1}^{S_k} \tau_t^k = \lim_{k\to\infty} \sum_{t=1}^{S_k} \left( V_{\frac{t}{T_k}}^k - V_{\frac{t-1}{T_k}}^k \right), \\
V_1 - V_{1-} = V_1 - \lim_{q\nearrow 1} V_q &= \lim_{k\to\infty} \sum_{t=1}^{S_k} \tau_{T_k-(t-1)}^k = \lim_{k\to\infty} \sum_{t=1}^{S_k} \left( V_{1-\frac{t-1}{T_k}}^k - V_{1-\frac{t}{T_k}}^k \right),
\end{aligned}$$

where $S_k$ is a sequence chosen so that both $S_k \to \infty$ and $S_k^3/T_k \to 0$. From (A.11), it follows that

$$\begin{aligned}
2X_{\frac{1}{T_k}}^k - X_{\frac{2}{T_k}}^k &= \frac{\gamma}{\theta + 2\gamma}\left( 2V_{\frac{1}{T_k}}^k - V_{\frac{2}{T_k}}^k \right) + \frac{\theta}{\theta + 2\gamma} V_{\frac{1}{T_k}}^k + O\left(\frac{1}{T_k}\right), \\
2X_{\frac{t}{T_k}}^k - X_{\frac{t+1}{T_k}}^k - X_{\frac{t-1}{T_k}}^k &= \frac{\gamma}{\theta + 2\gamma}\left( 2V_{\frac{t}{T_k}}^k - V_{\frac{t+1}{T_k}}^k - V_{\frac{t-1}{T_k}}^k \right) + \frac{\theta}{\theta + 2\gamma}\left( V_{\frac{t}{T_k}}^k - V_{\frac{t-1}{T_k}}^k \right) + O\left(\frac{1}{T_k}\right),
\end{aligned}$$

where $a^k = b^k + O\big(\frac{1}{T_k}\big)$ means $\limsup_{k\to\infty} \frac{|a^k - b^k|}{1/T_k} < \infty$, and thus

$$\begin{aligned}
x_1^k - x_2^k &= \frac{\gamma}{\theta + 2\gamma}\left(\tau_1^k - \tau_2^k\right) + \frac{\theta}{\theta + 2\gamma}\tau_1^k + O\left(\frac{1}{T_k}\right), \\
x_t^k - x_{t+1}^k &= \frac{\gamma}{\theta + 2\gamma}\left(\tau_t^k - \tau_{t+1}^k\right) + \frac{\theta}{\theta + 2\gamma}\tau_t^k + O\left(\frac{1}{T_k}\right).
\end{aligned}$$

Assuming $x_{S_k}^k = O\big(\frac{1}{T_k}\big)$ and $\tau_{S_k}^k = O\big(\frac{1}{T_k}\big)$, we obtain

$$\begin{aligned}
x_t^k &= x_{t+1}^k + \frac{\gamma}{\theta + 2\gamma}\left(\tau_t^k - \tau_{t+1}^k\right) + \frac{\theta}{\theta + 2\gamma}\tau_t^k + O\left(\frac{1}{T_k}\right) \\
&= \frac{\gamma}{\theta + 2\gamma} \sum_{j=t}^{S_k} \left(\tau_j^k - \tau_{j+1}^k\right) + \frac{\theta}{\theta + 2\gamma} \sum_{j=t}^{S_k} \tau_j^k + O\left(\frac{S_k}{T_k}\right) \\
&= \frac{\gamma}{\theta + 2\gamma}\tau_t^k + \frac{\theta}{\theta + 2\gamma} \sum_{j=t}^{S_k} \tau_j^k + O\left(\frac{S_k}{T_k}\right),
\end{aligned} \tag{A.13}$$

and likewise if $x_{T_k-S_k}^k = O\big(\frac{1}{T_k}\big)$ and $\tau_{T_k-S_k}^k = O\big(\frac{1}{T_k}\big)$,

$$x_{T_k-t}^k = x_{T_k-(t+1)}^k - \frac{\gamma}{\theta + 2\gamma}\left(\tau_{T_k-(t+1)}^k - \tau_{T_k-t}^k\right) - \frac{\theta}{\theta + 2\gamma}\tau_{T_k-(t+1)}^k + O\left(\frac{1}{T_k}\right)$$

$$
\begin{aligned}
&= -\frac{\gamma}{\theta+2\gamma}\sum_{j=t}^{S_k}\left(\tau^k_{T_k-(j+1)}-\tau^k_{T_k-j}\right)-\frac{\theta}{\theta+2\gamma}\sum_{j=t+1}^{S_k}\tau^k_{T_k-j}+O\left(\frac{S_k}{T_k}\right)\\
&= \frac{\gamma}{\theta+2\gamma}\tau^k_{T_k-t}-\frac{\theta}{\theta+2\gamma}\sum_{j=t+1}^{S_k}\tau^k_{T_k-j}+O\left(\frac{S_k}{T_k}\right). \qquad \text{(A.14)}
\end{aligned}
$$

Therefore, the expected costs for the client are

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{\tau}^k\cdot\boldsymbol{p}] &= p_0+\theta\sum_{t=1}^{T_k}X^k_{\frac{t}{T_k}}\left(V^k_{\frac{t}{T_k}}-V^k_{\frac{t-1}{T_k}}\right)+\gamma\sum_{t=1}^{T_k}\left(X^k_{\frac{t}{T_k}}-X^k_{\frac{t-1}{T_k}}\right)\left(V^k_{\frac{t}{T_k}}-V^k_{\frac{t-1}{T_k}}\right)\\
&= p_0+\theta\int_0^1 X_q\,dV_q+\theta\sum_{t=1}^{S_k}X^k_{\frac{t}{T_k}}\tau^k_t+\gamma\sum_{t=1}^{S_k}x^k_t\tau^k_t+\theta\sum_{t=1}^{S_k}X^k_{1-\frac{t-1}{T_k}}\tau^k_{T_k-(t-1)}+\gamma\sum_{t=1}^{S_k}x^k_{T_k-(t-1)}\tau^k_{T_k-(t-1)}+O\left(\frac{1}{T_k}\right). \qquad \text{(A.15)}
\end{aligned}
$$

We next analyze each of the following three terms:

1. $\theta\int_0^1 X_q\,dV_q$,
2. $\theta\sum_{t=1}^{S_k}X^k_{\frac{t}{T_k}}\tau^k_t+\gamma\sum_{t=1}^{S_k}x^k_t\tau^k_t$,
3. $\theta\sum_{t=1}^{S_k}X^k_{1-\frac{t-1}{T_k}}\tau^k_{T_k-(t-1)}+\gamma\sum_{t=1}^{S_k}x^k_{T_k-(t-1)}\tau^k_{T_k-(t-1)}$.

For the first term, we use (A.12) to write

$$
\int_0^1 X_q\,dV_q=\int_0^1\left(\frac{\theta}{\lambda T\sigma^2}\dot V_q+V_q\right)\dot V_q\,dq=\frac{\theta}{\lambda T\sigma^2}\int_0^1\dot V_q^2\,dq+\frac{1}{2}V_{1-}^2-\frac{1}{2}V_{0+}^2, \qquad \text{(A.16)}
$$

where the second equality is implied by

$$
\int_0^1 V_q\dot V_q\,dq=\frac{1}{2}V_{1-}^2-\frac{1}{2}V_{0+}^2,
$$

which in turn follows from integration by parts

$$
\int_0^1 V_q\dot V_q\,dq=V_{1-}^2-V_{0+}^2-\int_0^1\dot V_q V_q\,dq.
$$

Set $a=V_{1-}-V_{0+}$. By Corollary 9, optimal contract weights are symmetric, so we have $V_{0+}=1-V_{1-}=(1-a)/2$. Moreover, the minimizer of $\int_0^1\dot V_q^2\,dq$ subject to $a=V_{1-}-V_{0+}$ is $\dot V_q=a$ almost everywhere on $(0,1)$ by Jensen's inequality. Therefore, (A.16) in the optimum becomes

$$
\int_0^1 X_q\,dV_q=\frac{\theta}{\lambda T\sigma^2}\int_0^1\dot V_q^2\,dq+\frac{1}{2}V_{1-}^2-\frac{1}{2}V_{0+}^2=\frac{\theta^2}{\lambda T\sigma^2}a^2+\frac{\theta(1+a)^2}{8}-\frac{\theta(1-a)^2}{8}. \qquad \text{(A.17)}
$$

Using $V_{0+}=1-V_{1-}=(1-a)/2$, we also have

$$
\sum_{t=1}^{S_k}\tau^k_t=\frac{1-a}{2}+O\left(\frac{1}{T_k}\right),\quad \sum_{t=1}^{S_k}\tau^k_{T_k-(t-1)}=\frac{1-a}{2}+O\left(\frac{1}{T_k}\right).
$$

Next, we analyze the minimization of

$$
\theta\sum_{t=1}^{S_k}X^k_{\frac{t}{T_k}}\tau^k_t+\gamma\sum_{t=1}^{S_k}x^k_t\tau^k_t. \qquad \text{(A.18)}
$$

subject to $\sum_{t=1}^{S_k}\tau^k_t=\frac{1-a}{2}$. Using (A.13), we write

$$
\begin{aligned}
\theta\sum_{t=1}^{S_k}X^k_{\frac{t}{T_k}}\tau^k_t+\gamma\sum_{t=1}^{S_k}x^k_t\tau^k_t &= \theta\sum_{t=1}^{S_k}\sum_{\ell=1}^{t}\left(\frac{\gamma}{\theta+2\gamma}\tau^k_\ell+\frac{\theta}{\theta+2\gamma}\sum_{j=\ell}^{S_k}\tau^k_j\right)\tau^k_t+\gamma\sum_{t=1}^{S_k}\left(\frac{\gamma}{\theta+2\gamma}\tau^k_t+\frac{\theta}{\theta+2\gamma}\sum_{j=t}^{S_k}\tau^k_j\right)\tau^k_t+O\left(\frac{S_k^3}{T_k}\right)\\
&= \theta\sum_{t=1}^{S_k}\left(\frac{\gamma}{\theta+2\gamma}\sum_{\ell=1}^{t}\tau^k_\ell+\frac{\theta}{\theta+2\gamma}\sum_{j=1}^{S_k}\min\{j,t\}\tau^k_j\right)\tau^k_t+\gamma\sum_{t=1}^{S_k}\left(\frac{\gamma}{\theta+2\gamma}\tau^k_t+\frac{\theta}{\theta+2\gamma}\sum_{j=t}^{S_k}\tau^k_j\right)\tau^k_t+O\left(\frac{S_k^3}{T_k}\right).
\end{aligned}
$$

We can simplify two terms

$$
\begin{aligned}
\theta\sum_{t=1}^{S_k}\frac{\gamma}{\theta+2\gamma}\sum_{\ell=1}^{t}\tau^k_\ell\tau^k_t+\gamma\sum_{t=1}^{S_k}\frac{\theta}{\theta+2\gamma}\sum_{j=t}^{S_k}\tau^k_j\tau^k_t &= \frac{\theta\gamma}{\theta+2\gamma}\sum_{t=1}^{S_k}\sum_{\ell=1}^{S_k}\tau^k_\ell\tau^k_t+\frac{\theta\gamma}{\theta+2\gamma}\sum_{t=1}^{S_k}\left(\tau^k_t\right)^2\\
&= \frac{\theta\gamma}{\theta+2\gamma}\sum_{t=1}^{S_k}\frac{1-a}{2}\tau^k_t+\frac{\theta\gamma}{\theta+2\gamma}\sum_{t=1}^{S_k}\left(\tau^k_t\right)^2\\
&= \frac{\theta\gamma}{\theta+2\gamma}\frac{(1-a)^2}{4}+\frac{\theta\gamma}{\theta+2\gamma}\sum_{t=1}^{S_k}\left(\tau^k_t\right)^2.
\end{aligned}
$$

Therefore, the optimization problem (A.18) becomes

$$
\frac{\theta\gamma}{\theta+2\gamma}\frac{(1-a)^2}{4}+\frac{\theta^2}{\theta+2\gamma}\sum_{t=1}^{S_k}\tau^k_t\sum_{j=1}^{S_k}\min\{j,t\}\tau^k_j+\frac{\theta\gamma+\gamma^2}{\theta+2\gamma}\sum_{t=1}^{S_k}\left(\tau^k_t\right)^2
$$

subject to $\sum_{t=1}^{S_k} \tau_t^k = \frac{1-a}{2}$. The first-order condition from the Lagrange multiplier method implies

$$\begin{aligned} 0 &= \frac{\partial}{\partial \tau_s^k}\left( \theta^2 \sum_{t=1}^{S_k} \tau_t^k \sum_{j=1}^{S_k} \min\{j,t\}\tau_j^k + (\theta\gamma+\gamma^2)\sum_{t=1}^{S_k}\left(\tau_t^k\right)^2 \right) + \lambda_1 \\ &= \theta^2 \sum_{j=1}^{S_k} \min\{j,s\}\tau_j^k + \theta^2 \sum_{t=1}^{S_k} \tau_t^k \min\{s,t\} + 2(\theta\gamma+\gamma^2)\tau_s^k + \lambda_1 \\ &= 2\theta^2 \sum_{j=1}^{S_k} \min\{j,s\}\tau_j^k + 2(\theta\gamma+\gamma^2)\tau_s^k + \lambda_1 \end{aligned} \tag{A.19}$$

for all $s = 1,2,\ldots$ Specifically, for $\gamma = 0$, this implies with $s = 1$ that $\lambda_1 = -\theta^2(1-a)$; with $s = 2$ that $\lambda_1 = -\theta^2(1-a) - \sum_{j=2}^{S_k} \tau_j$, implying $\sum_{j=2}^{S_k} \tau_j$; and iteratively comparing different $s$, we obtain $\tau_1^k = \frac{1-a}{2}$ and $\tau_j^k = 0$ for all $j > 1$. For $\gamma > 0$, we deduce

$$\begin{aligned} 0 &= 2\theta^2 \sum_{j=1}^{S_k}\big(\min\{j,s+1\} - \min\{j,s\}\big)\tau_j^k + 2(\theta\gamma+\gamma^2)\big(\tau_{s+1}^k - \tau_s^k\big) \\ &= 2\theta^2 \sum_{j=s+1}^{S_k} \tau_j^k + 2(\theta\gamma+\gamma^2)\big(\tau_{s+1}^k - \tau_s^k\big) \\ &= 2\theta^2\left(\frac{1-a}{2} - \sum_{j=1}^{s}\tau_j^k\right) + 2(\theta\gamma+\gamma^2)\big(\tau_{s+1}^k - \tau_s^k\big) \end{aligned}$$

for all $s = 1,2,\ldots$, so that

$$\tau_{s+1}^k = \tau_s^k - \frac{\theta^2}{\gamma\theta+\gamma^2}\left(\frac{1-a}{2} - \sum_{j=1}^{s}\tau_j^k\right), \qquad s = 1,2,\ldots$$

Its solution is

$$\tau_j^k = \frac{\theta\gamma^{j-1}}{(\theta+\gamma)^j}\frac{1-a}{2}, \quad j = 1,2,\ldots, \tag{A.20}$$

which satisfies

$$\sum_{j=1}^{s}\tau_j^k = \sum_{j=1}^{s}\frac{\theta\gamma^{j-1}}{(\theta+\gamma)^j}\frac{1-a}{2} = \frac{\theta}{\theta+\gamma}\frac{1-\frac{\gamma^s}{(\theta+\gamma)^s}}{1-\frac{\gamma}{\theta+\gamma}}\frac{1-a}{2} \overset{s\to\infty}{\to} \frac{1-a}{2}.$$

We also note that it follows from (A.13), (A.14), and (A.20) that

$$\begin{aligned} x_t^k &= \frac{\gamma}{\theta+2\gamma}\tau_t^k + \frac{\theta}{\theta+2\gamma}\sum_{j=t}^{S_k}\tau_j^k + O\left(\frac{S_k}{T_k}\right) \\ &= \frac{\gamma}{\theta+2\gamma}\frac{\theta\gamma^{t-1}}{(\theta+\gamma)^t}\frac{1-a}{2} + \frac{\theta}{\theta+2\gamma}\sum_{j=t}^{S_k}\frac{\theta\gamma^{j-1}}{(\theta+\gamma)^j}\frac{1-a}{2} + O\left(\frac{S_k}{T_k}\right) \\ &= \frac{\gamma}{\theta+2\gamma}\frac{\theta\gamma^{t-1}}{(\theta+\gamma)^t}\frac{1-a}{2} + \frac{\theta}{\theta+2\gamma}\frac{\theta\gamma^{t-1}}{(\theta+\gamma)^t}\frac{1}{1-\frac{\gamma}{\theta+\gamma}}\frac{1-a}{2} + O\left(\frac{S_k}{T_k}\right) \\ &= \frac{\theta\gamma^{t-1}}{(\theta+\gamma)^t}\frac{1-a}{2} + O\left(\frac{S_k}{T_k}\right), \\ x_{T_k-t}^k &= \frac{\gamma}{\theta+2\gamma}\tau_{T_k-t}^k - \frac{\theta}{\theta+2\gamma}\sum_{j=t+1}^{S_k}\tau_{T_k-j}^k + O\left(\frac{S_k}{T_k}\right) \\ &= \frac{\gamma}{\theta+2\gamma}\tau_{t+1}^k - \frac{\theta}{\theta+2\gamma}\sum_{j=t+2}^{S_k}\tau_j^k + O\left(\frac{S_k}{T_k}\right) \\ &= \frac{\gamma}{\theta+2\gamma}\frac{\theta\gamma^{t}}{(\theta+\gamma)^{t+1}}\frac{1-a}{2} - \frac{\theta}{\theta+2\gamma}\sum_{j=t+2}^{S_k}\frac{\theta\gamma^{j-1}}{(\theta+\gamma)^j}\frac{1-a}{2} + O\left(\frac{S_k}{T_k}\right) \\ &= \frac{\gamma}{\theta+2\gamma}\frac{\theta\gamma^{t}}{(\theta+\gamma)^{t+1}}\frac{1-a}{2} - \frac{\theta}{\theta+2\gamma}\frac{\theta\gamma^{t+1}}{(\theta+\gamma)^{t+2}}\frac{1}{1-\frac{\gamma}{\theta+\gamma}}\frac{1-a}{2} + O\left(\frac{S_k}{T_k}\right) \\ &= O\left(\frac{S_k}{T_k}\right). \end{aligned} \tag{A.21}$$

For $s = 1$, (A.19) simplifies to

$$0 = 2\theta^2\sum_{j=1}^{S_k}\tau_j^k + 2(\theta\gamma+\gamma^2)\tau_1^k + \lambda_1,$$

which implies

$$\lambda_1 = -2\theta^2\sum_{j=1}^{S_k}\tau_j^k - 2(\theta\gamma+\gamma^2)\tau_1^k = -2\theta^2\frac{1-a}{2} - 2(\theta\gamma+\gamma^2)\frac{\theta}{\theta+\gamma}\frac{1-a}{2} = -2(\theta^2+\gamma\theta)\frac{1-a}{2}.$$

We also derive from (A.19) that

$$2\theta^2 \sum_{j=1}^{S_k} \min\{j,t\}\tau_j^k = -2(\theta\gamma+\gamma^2)\tau_t^k - \lambda_1 = -2(\theta\gamma+\gamma^2)\tau_t^k + 2(\theta^2+\gamma\theta)\frac{1-a}{2},$$

hence the optimization problem becomes

$$\begin{aligned}
&\frac{\theta\gamma}{\theta+2\gamma}\frac{(1-a)^2}{4} + \frac{\theta^2}{\theta+2\gamma}\sum_{t=1}^{S_k}\tau_t^k\sum_{j=1}^{S_k}\min\{j,t\}\tau_j^k + \frac{\theta\gamma+\gamma^2}{\theta+2\gamma}\sum_{t=1}^{S_k}\left(\tau_t^k\right)^2 \\
&= \frac{\theta\gamma}{\theta+2\gamma}\frac{(1-a)^2}{4} - \frac{\theta\gamma+\gamma^2}{\theta+2\gamma}\sum_{t=1}^{S_k}\left(\tau_t^k\right)^2 + \frac{1}{\theta+2\gamma}\sum_{t=1}^{S_k}\tau_t^k(\theta^2+\gamma\theta)\frac{1-a}{2} + \frac{\theta\gamma+\gamma^2}{\theta+2\gamma}\sum_{t=1}^{S_k}\left(\tau_t^k\right)^2 \\
&= \frac{\theta\gamma}{\theta+2\gamma}\frac{(1-a)^2}{4} + \frac{1}{\theta+2\gamma}\sum_{t=1}^{S_k}\tau_t^k(\theta^2+\gamma\theta)\frac{1-a}{2} \\
&= \frac{\theta\gamma}{\theta+2\gamma}\frac{(1-a)^2}{4} + \frac{\theta^2+\gamma\theta}{\theta+2\gamma}\frac{(1-a)^2}{4} \\
&= \theta\frac{(1-a)^2}{4}.
\end{aligned} \tag{A.22}$$

Finally, we analyze the minimization problem

$$\theta\sum_{t=1}^{S_k} X^k_{1-\frac{t-1}{T_k}}\tau^k_{T_k-(t-1)} + \gamma\sum_{t=1}^{S_k} x^k_{T_k-(t-1)}\tau^k_{T_k-(t-1)}$$

subject to $\sum_{t=1}^{S_k}\tau^k_{T_k-(t-1)} = \frac{1-a}{2}$. However, as per (A.21), the jumps of the trading strategy in the limit disappear so that $x^k_{T_k-(t-1)} = O\left(\frac{S_k}{T_k}\right)$ and $X^k_{1-\frac{t-1}{T_k}} = 1 + O\left(\frac{S_k^2}{T_k}\right)$. Therefore, the value of the minimization problem becomes

$$\theta\sum_{t=1}^{S_k} X^k_{1-\frac{t-1}{T_k}}\tau^k_{T_k-(t-1)} + \gamma\sum_{t=1}^{S_k} x^k_{T_k-(t-1)}\tau^k_{T_k-(t-1)} = \theta\sum_{t=1}^{S_k}\tau^k_{T_k-(t-1)} + O\left(\frac{S_k^3}{T_k}\right) = \theta\frac{1-a}{2} + O\left(\frac{S_k^3}{T_k}\right). \tag{A.23}$$

In summary, using (A.17), (A.22), and (A.23), the expected costs from (A.15) in the optimum are

$$\begin{aligned}
\mathbb{E}[\tau^k\cdot p] &= p_0 + \theta\int_0^1 X_q\,dV_q + \theta\sum_{t=1}^{S_k} X^k_{\frac{t}{T_k}}\tau_t^k + \gamma\sum_{t=1}^{S_k} x_t^k\tau_t^k + \theta\sum_{t=1}^{S_k} X^k_{1-\frac{t-1}{T_k}}\tau^k_{T_k-(t-1)} + \gamma\sum_{t=1}^{S_k} x^k_{T_k-(t-1)}\tau^k_{T_k-(t-1)} + O\left(\frac{S_k^3}{T_k}\right) \\
&= p_0 + \frac{\theta^2}{\lambda T\sigma^2}a^2 + \frac{\theta(1+a)^2}{8} - \frac{\theta(1-a)^2}{8} + \theta\frac{(1-a)^2}{4} + \theta\frac{1-a}{2} + O\left(\frac{S_k^3}{T_k}\right) \\
&= p_0 + \frac{\theta^2}{\lambda T\sigma^2}a^2 + \theta\frac{(1-a)^2}{4} + \frac{\theta}{2} + O\left(\frac{S_k^3}{T_k}\right),
\end{aligned} \tag{A.24}$$

minimized over $a$. The first-order condition gives

$$\frac{\theta^2}{\lambda T\sigma^2}2a - \theta\frac{1-a}{2} = 0$$

so that

$$a = \frac{\frac{\theta}{2}}{\frac{\theta}{2} + \frac{2\theta^2}{\lambda T\sigma^2}} = \frac{1}{1 + \frac{4\theta}{\lambda T\sigma^2}}. \tag{A.25}$$

We conclude for $q \in (0,1)$ that

$$\begin{aligned}
V_q &= V_{0+} + (V_q - V_{0+}) = \frac{1-a}{2} + aq, \\
X_q &= \frac{\theta}{\lambda T\sigma^2}\dot{V}_q + V_q = \frac{\theta a}{\lambda T\sigma^2} + \frac{1-a}{2} + aq,
\end{aligned}$$

using (A.12), which implies the formulas for the limits of the optimal contract and dealer's trading strategy. Thanks to (A.24) and (A.25), the client's expected costs converge to

$$p_0 + \frac{\theta^2}{\lambda T\sigma^2}a^2 + \theta\frac{(1-a)^2}{4} + \frac{\theta}{2} = p_0 + \frac{\theta}{4}\left(1 + \frac{4\theta}{\lambda T\sigma^2}\right)a^2 - \frac{a}{2}\theta + \frac{3}{4}\theta = p_0 + \frac{3-a}{4}\theta. \quad \square$$

**Proof of Proposition 11.** This proof builds on the first half of the proof of Proposition 10.

*Claim (i):* In the case of a TWAP contract, we have

$$V_q^{TWAP} = \lim_{k\to\infty} V_q^{TWAP,k} = \lim_{k\to\infty}\sum_{t=1}^{\lceil qT_k\rceil}\tau_t^{TWAP,k} = q$$

for all $q \in (0,1)$ so that (A.12) becomes

$$X_q^{TWAP} = \frac{\theta}{\lambda T\sigma^2}\dot{V}_q^{TWAP} + V_q^{TWAP} = \frac{\theta}{\lambda T\sigma^2} + q$$

for all $q \in (0,1)$. Along with the conditions $X_0^{TWAP,k} = 0$ and $X_1^{TWAP,k} = 1$ for all $k$, this shows (7). Because $V_q^{TWAP}$ does not have any jumps, it follows from (A.15) that the expected costs for the client under a TWAP contract are

$$\lim_{k\to\infty} \mathbb{E}[\boldsymbol{\tau}^{TWAP,k} \cdot \boldsymbol{p}] = p_0 + \theta \int_0^1 X_q^{TWAP}\, dV_q^{TWAP} = p_0 + \theta \int_0^1 \left( \frac{\theta}{\lambda T \sigma^2} + q \right) dq = p_0 + \frac{\theta^2}{\lambda T \sigma^2} + \frac{1}{2}\theta.$$

*Claim (ii):* In the case of a MOC contract, we have

$$V_q^{MOC} = \lim_{k\to\infty} V_q^{MOC,k} = \lim_{k\to\infty} \sum_{t=1}^{\lceil qT_k \rceil} \tau_t^{MOC,k} = 0$$

for all $q \in (0,1)$ so that (A.12) becomes

$$X_q^{MOC} = \frac{\theta}{\lambda T \sigma^2} \dot{V}_q^{MOC} + V_q^{MOC} = 0$$

for all $q \in (0,1)$. Along with the conditions $X_0^{MOC} = 0$ and $X_1^{MOC} = 1$, this shows (8). The expected costs for the client are

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\tau}^{MOC,k} \cdot \boldsymbol{p}] &= p_0 + \theta \sum_{t=1}^{T_k} X_{\frac{t}{T_k}}^{MOC,k} \left( V_{\frac{t}{T_k}}^{MOC,k} - V_{\frac{t-1}{T_k}}^{MOC,k} \right) + \gamma \sum_{t=1}^{T_k} \left( X_{\frac{t}{T_k}}^{MOC,k} - X_{\frac{t-1}{T_k}}^{MOC,k} \right) \left( V_{\frac{t}{T_k}}^{MOC,k} - V_{\frac{t-1}{T_k}}^{MOC,k} \right) \\
&= p_0 + \theta X_1^{MOC,k} + \gamma \left( X_1^{MOC,k} - X_{\frac{T_k-1}{T_k}}^{MOC,k} \right) \\
&= p_0 + \theta + \frac{\gamma^2}{\theta + 2\gamma} + O\left( \frac{S_k}{T_k} \right),
\end{aligned}$$

using that

$$X_1^{MOC,k} - X_{\frac{T_k-1}{T_k}}^{MOC,k} = x_{T_k}^{MOC,k} = \frac{\gamma}{\theta + 2\gamma} + O\left( \frac{S_k}{T_k} \right)$$

by (A.14). □

## Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jfineco.2024.103901.

## References

Abel Noser, 2021. Hidden figures–Introducing a portfolio transition & event universe. https://bit.ly/3zE5Lhl.

Almgren, Robert, Chriss, Neil, 2001. Optimal execution of portfolio transactions. J. Risk 3, 5–40.

Avramov, Doron, Chordia, Tarun, Goyal, Amit, 2006. The impact of trades on daily volatility. Rev. Financ. Stud. 19 (4), 1241–1277.

Baldauf, Markus, Frei, Christoph, Mollner, Joshua, 2022. Principal trading arrangements: When are common contracts optimal? Manage. Sci. 68 (4), 3112–3128.

Baldauf, Markus, Mollner, Joshua, 2024. Competition and information leakage. J. Polit. Econ. 132 (5), 1603–1641.

Bernhardt, Dan, Taub, Bart, 2008. Front-running dynamics. J. Econom. Theory 138 (1), 288–296.

Bertsimas, Dimitris, Lo, Andrew W., 1998. Optimal control of execution costs. J. Financial Mark. 1 (1), 1–50.

Bhattacharya, Sudipto, Pfleiderer, Paul, 1985. Delegated portfolio management. J. Econom. Theory 36 (1), 1–25.

Biais, Bruno, Glosten, Larry, Spatt, Chester, 2005. Market microstructure: A survey of microfoundations, empirical results, and policy implications. J. Financ. Mark. 8 (2), 217–264.

Biais, Bruno, Mariotti, Thomas, Plantin, Guillaume, Rochet, Jean-Charles, 2007. Dynamic security design: Convergence to continuous time and asset pricing implications. Rev. Econ. Stud. 74 (2), 345–390.

Bloomberg, 2016. Cairn energy said to be victim of HSBC currency frontrunning. https://bloom.bg/3sBA14k.

Bloomberg, 2019. Barclays trader beats U.S. prosecution of front-running charges. https://bloom.bg/3Dzbcy0.

Bloomberg, 2020. Hedge funds press for crackdown on front-running loophole in EU. https://bloom.bg/3LrOiul.

Bloomberg, 2022a. Japan's SMBC Nikko, staff charged with market manipulation. https://bloom.bg/3F8zRIJ.

Bloomberg, 2022b. Morgan Stanley's Passi faces U.S. block-trading probe. https://bloom.bg/3OJQoXL.

Bloomberg, 2024a. Citadel among hedge funds that got Morgan Stanley's block-trading leaks. https://bloom.bg/4bBj93z.

Bloomberg, 2024b. Morgan Stanley to pay $249 million to end block trade probes. https://bloom.bg/44Edqrg.

Board, Simon, Skrzypacz, Andrzej, 2016. Revenue management with forward-looking buyers. J. Polit. Econ. 124 (4), 1046–1087.

Bolton, Patrick, Dewatripont, Mathias, 2004. Contract Theory. MIT Press.

Buffa, Andrea M., Vayanos, Dimitri, Woolley, Paul, 2022. Asset management contracts and equilibrium prices. J. Polit. Econ. 130 (12), 3025–3342.

Campo, Sandra, Guerre, Emmanuel, Perrigne, Isabelle, Vuong, Quang, 2011. Semiparametric estimation of first-price auctions with risk-averse bidders. Rev. Econ. Stud. 78 (1), 112–147.

Carpenter, Jennifer N., 2000. Does option compensation increase managerial risk appetite? J. Finance 55 (5), 2311–2331.

Cartea, Alvaro, Jaimungal, Sebastian, 2016. Incorporating order-flow into optimal execution. Math. Financ. Econ. 10 (3), 339–364.

Cox, John C., Ross, Stephen A., Rubinstein, Mark, 1979. Option pricing: A simplified approach. J. Financ. Econ. 7 (3), 229–263.

Curello, Gregorio, Sinander, Ludvig, 2024. Screening for breakthroughs. Working Paper, https://arxiv.org/abs/2011.10090.

DeMarzo, Peter M., Fishman, Michael J., 2007. Optimal long-term financial contracting. Rev. Financ. Stud. 20 (6), 2079–2128.

Di Tella, Sebastian, Sannikov, Yuliy, 2021. Optimal asset management contracts with hidden savings. Econometrica 89 (3), 1099–1139.

Duffie, Darrell, Dworczak, Piotr, 2021. Robust benchmark design. J. Financ. Econ. 142 (2), 775–802.

Edelen, Roger M., Kadlec, Gregory B., 2012. Delegated trading and the speed of adjustment in security prices. J. Financ. Econ. 103 (2), 294–307.

Financial Industry Regulatory Authority, 2013. FINRA rule 5270. Front running of block transactions. https://www.finra.org/rules-guidance/rulebooks/finra-rules/5270.

Fishman, Michael J., Longstaff, Francis A., 1992. Dual trading in futures markets. J. Finance 47 (2), 643–671.

Gârleanu, Nicolae, Pedersen, Lasse Heje, 2013. Dynamic trading with predictable returns and transaction costs. J. Finance 68 (6), 2309–2340.

Gârleanu, Nicolae, Pedersen, Lasse Heje, 2016. Dynamic portfolio choice with frictions. J. Econom. Theory 165, 487–516.

Garrett, Daniel F., 2016. Intertemporal price discrimination: Dynamic arrivals and changing values. Amer. Econ. Rev. 106 (11), 3275–3299.

Goldman Sachs, 2017. Disclosure regarding FINRA rule 5270. https://www.goldmansachs.com/disclosures/disclosure-regarding-finra-rule-5270.pdf.

Grenadier, Steven R., Malenko, Andrey, Malenko, Nadya, 2016. Timing decisions in organizations: Communication and authority in a dynamic environment. Amer. Econ. Rev. 106 (9), 2552–2581.

Gromb, Denis, Vayanos, Dimitri, 2010. Limits of arbitrage. Annu. Rev. Finan. Econ. 2 (1), 251–275.

Hellwig, Martin F., Schmidt, Klaus M., 2002. Discrete-time approximations of the Holmström-Milgrom Brownian-motion model of intertemporal incentive provision. Econometrica 70 (6), 2225–2264.

Holmström, Bengt, 1979. Moral hazard and observability. Bell J. Econ. 10 (1), 74–91.

Holmström, Bengt, Milgrom, Paul, 1987. Aggregation and linearity in the provision of intertemporal incentives. Econometrica 55 (2), 303–328.

HSBC Global Banking and Markets, 2022. Equities. https://www.gbm.hsbc.com/solutions/markets/equities.

Kruse, Thomas, Strack, Philipp, 2015. Optimal stopping with private information. J. Econom. Theory 159, 702–727.

Kyle, Albert S., 1985. Continuous auctions and insider trading. Econometrica 53 (6), 1315–1335.

Kyle, Albert S., Obizhaeva, Anna A., Wang, Yajun, 2018. Smooth trading with overconfidence and market power. Rev. Econ. Stud. 85 (1), 611–662.

Madsen, Erik, 2022. Designing deadlines. Amer. Econ. Rev. 112 (3), 963–997.

Morgan Stanley, 2022. Wealth management disclosures. https://www.morganstanley.com/wealth-disclosures/disclosures#5.

Nasdaq, 2022. How much does trading cost the buy side? https://bit.ly/3U1O2Gw.

Obizhaeva, Anna A., Wang, Jiang, 2013. Optimal trading strategy and supply/demand dynamics. J. Financial Mark. 16 (1), 1–32.

Risk.net, 2021. Page 19901: The benchmark that time forgot. https://www.risk.net/7818166.

Röell, Ailsa, 1990. Dual-capacity trading and the quality of the market. J. Financ. Intermediat. 1 (2), 105–124.

Sannikov, Yuliy, 2008. A continuous-time version of the principal-agent problem. Rev. Econ. Stud. 75 (3), 957–984.

Securities Industry and Financial Markets Association, 2021. Global equity markets primer: Market costs. https://bit.ly/3eGr5IL.

Seppi, Duane J., 1990. Equilibrium block trading and asymmetric information. J. Finance 45 (1), 73–94.

The Wall Street Journal, 2022a. Big stock sales are supposed to be secret. The numbers indicate they aren't. https://on.wsj.com/3LrYQZy.

The Wall Street Journal, 2022b. How we analyzed Wall Street block trades. https://on.wsj.com/3vwaLjP.

The Wall Street Journal, 2024. Morgan Stanley agrees to pay $249 million to settle block-trading probes. https://on.wsj.com/3UIn6N5.

Traders Magazine, 2005a. NYSE & NASD investigate guaranteed VWAP. https://bit.ly/3BSsu8f.

Traders Magazine, 2005b. VWAP debate divides the trading industry. https://bit.ly/3vmUspm.

US Securities and Exchange Commission, 2005. Regulation NMS: Final rules and amendments to joint industry plans. Federal Register 70, 124.