FTG Working Paper Series

Principal Trading Arrangements: Optimality under Temporary and Permanent Price Impact

by

Markus Baldauf
Joshua Mollner
christoph frei

Working Paper No. 00095-00

Finance Theory Group

www.financetheory.com

Principal Trading Arrangements: Optimality under Temporary and Permanent Price Impact*

Markus Baldauf[†] Christoph Frei[‡] Joshua Mollner[§]

December 15, 2022

Click here for latest version

Abstract

We study the optimal execution problem in a principal-agent setting. A client (e.g., a pension fund, endowment, or other institution) contracts to purchase a large position from a dealer at a future point in time. In the interim, the dealer acquires the position from the market, choosing how to divide his trading across time. Price impact may have temporary and permanent components. There is hidden action in that the client cannot directly dictate the dealer's trades. Rather, she chooses a contract with the goal of minimizing her expected payment, given the price process and an understanding of the dealer's incentives. Many contracts used in practice prescribe a payment equal to some weighted average of the market prices within the execution window. We explicitly characterize the optimal such weights: they are symmetric and generally U-shaped over time. This U-shape is strengthened by permanent price impact and weakened by both temporary price impact and dealer risk aversion. In contrast, the first-best solution (which reduces to a classical optimal execution problem) is invariant to these parameters. Back-of-the-envelope calculations suggest that switching to our optimal contract could save clients billions of dollars per year.

Keywords: agency conflict, dealer-client relationship, principal trading, price impact
JEL Codes: G11, G14, G23, D82, D86

* *First Version:* February 3, 2021. We thank Torben Andersen, Sandeep Baliga, Eric Budish, Georgy Egorov, Péter Eső, Timothy Feddersen, Michael Fishman, John Kuong (discussant), Nicolas Lambert, Peter Klibanoff, Andrzej Skrzypacz, Duane Seppi, Chester Spatt, Alireza Tahbaz-Salehi, Dimitri Vayanos, Haoxiang Zhu, and various seminar participants for helpful comments. 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. Markus Baldauf thanks the University of Chicago, Booth School of Business, where he was visiting when large parts of this paper were written.

[†]University of British Columbia, Sauder School of Business: baldauf@mail.ubc.ca, https://sites.google.com/site/mbaldauf/.

[‡]University of Alberta: cfrei@ualberta.ca, https://www.math.ualberta.ca/∼cfrei/.

[§]Northwestern University, Kellogg School of Management: joshua.mollner@kellogg.northwestern.edu, https://sites.google.com/site/joshuamollner/.

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 consider how each child order will affect the prices obtained for future child orders. A literature on optimal execution considers that problem, addressing how an institution ought to behave if handling execution in house. Yet, pension funds and other institutions often outsource execution, in which case agency conflicts also become relevant. While such 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 they in fact interact in important and subtle ways.

Specifically, we model a situation in which an institution ('the client' henceforth) contracts with a dealer, agreeing to conduct a block trade: a single, large off-market transaction. The complexities of execution are thus shifted to the dealer, who would then pursue offsetting trades on the market. The client and dealer then need only agree on how their block trade should be priced. In practice, many trading arrangements prescribe a payment equal to some weighted average of the market prices prevailing over the execution window. For example, it is common for the dealer to be paid 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, as in a *guaranteed market-on-close (MOC)* contract. Because these payment rules 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 payment rules 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 the general arrangement 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 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 begin by solving the model in discrete time, then characterize the continuous-time limit. Although the contract that we derive as optimal 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 according to 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 are typically accompanied by 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 the price impact of trading. 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.[1] 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,c,d,e; WSJ, 2022a,b,c,d). These transaction costs might be reduced if prevailing arrangements were modified to more closely resemble the contractual arrangements that we derive.

Model. 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 $(p_1, \ldots, p_T)^\top$. If the dealer accepts the offered contract, then he prepares by pursuing offsetting trades on the market during the trading periods $\{1, \ldots, T\}$. In modeling how these trades affect the dynamics of prices, we assume a canonical market model that allows for price shocks and both permanent and temporary price impact. Finally, at time $T + 1$, the client and dealer conduct their agreed-upon off-market trade.

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 contractual arrangements are optimal in light of them.

Results. What is the optimal contract? Its precise form depends on the parameters of the model: the market parameters that govern price impact and the dealer's degree of risk aversion. Our analysis optimizes over contracts that are weighted averages of market prices. This is for both analytical tractability and realism, as more exotic functions of prices are likely to be prohibitively complex. Indeed, many of the contracts that are commonly 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 provides a solution for the optimal such contract, which we denote $\boldsymbol{\tau}^* = (\tau_1^*, \ldots, \tau_T^*)^\top$, where τ_t^* represents the weight on

[1] For example, in relation to 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).

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).

We also derive a closed-form solution for the trading behavior that the dealer selects in response to the optimal contract $\boldsymbol{\tau}^*$. This behavior can be described by a vector of trades $\boldsymbol{x}^* = (x_1^*, \ldots, x_T^*)^\top$, where x_t^* is the volume traded in period t. Numerical experimentation suggests that $\boldsymbol{x}^*$ is frontloaded in the sense of first-order stochastic dominance (i.e., $\sum_{s=1}^t x_s^* \geq \frac{t}{T}$ for all t), with a severity that is strengthened by permanent price impact, weakened by temporary price impact, and weakened by the dealer's risk aversion.

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 (determined by the contract) and the costs of his on-market 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 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 thought of 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 larger than 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 and the severity of the dealer's frontloading—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, as the result of two opposing forces that 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 that it is possible to improve substantially upon the status quo, even while staying within the class of weighted-average-price contracts.

Related literature. There is a long tradition of models that 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, Vayanos and Woolley, 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. The agent's actions are unobserved by the client, and, therefore, this problem belongs to the large literature on moral hazard.[2]

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, Malenko and Malenko, 2016) or chooses the timing of a disclosure (e.g., Curello and Sinander, 2021) 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

[2] 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, Mariotti, Plantin and Rochet, 2007).

'market model' that governs how her trades affect the dynamics of prices. Solving for the first-best benchmark of our model is equivalent to such an optimal execution 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 to the client.

The most related paper is Baldauf, Frei and Mollner (2022).[3] It begins with a certain commonly-used contract—the 'guaranteed VWAP' contract, in which the client pays the dealer according to the market's volume-weighted average price—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 permanent price impact, which not only allows this paper to speak to a much broader class of settings but also makes the problem conceptually different: it becomes genuinely dynamic in the sense that the ordering of time periods matters.

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 provides 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

A client (the principal) offers her dealer (the agent) a contract regarding a trade between them.[4] 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.

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.

[3] 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.

[4] In assuming a preexisting bilateral relationship between the client and the dealer, we abstract away from the question of how the client should select a dealer. See Baldauf and Mollner (2022) for an analysis of such an issue.

Dealer. The dealer has constant absolute risk aversion (CARA), with coefficient λ. To economize on notation, we use $u(w) = -\exp(-\lambda w)$ to denote the dealer's utility function.

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

Contracts. We focus on contracts that are weighted averages of the market prices. Although these weights will be nonnegative in the optimum, we do not impose this as a constraint. Thus, a contract can be thought of as a vector $\boldsymbol{\tau} \in \mathcal{T} \equiv \left\{(\tau_1, \ldots, \tau_T)^\top \in \mathbb{R}^T \,\middle|\, \sum_{t=1}^T \tau_t = 1\right\}$, which stipulates that the client will pay the dealer $\sum_{t=1}^T \tau_t p_t$. When we wish to highlight its dependence on the number of periods, we sometimes write $\mathcal{T}^T$.

Thus, the client can contract only on prices. 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.

Remark 1. 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. For example, 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 respectively correspond to what are known in practice as a *guaranteed TWAP* contract and a *guaranteed MOC* contract. While a general contract space 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 that they prescribe for certain price-path realizations.[5] We interpret our analysis as a search for the best contract among those comparable in complexity to those already in use.[6]

Remark 2. Of the contracts not of the weighted-average price form and hence outside $\mathcal{T}$, perhaps the most notable are fixed-price contracts (analogous to 'sell-the-firm' contracts in classical contract theory). The absence of such contracts is particularly relevant if the dealer is risk-neutral. (Although our focus is on settings where the dealer is risk averse, our formulation allows for a risk-neutral dealer.) In such cases, it is immediate that a fixed-price contract—if feasible—would give the client her first-best payoff and would be an optimal contract overall. All our analysis can be extended to accommodate fixed-price contracts. In Appendix OA.B, we repeat the analysis for the case in which the feasible contracts are all affine functions of prices; such a contract can

[5] Indeed, Mirrlees has observed that in classic moral hazard settings with normally-distributed noise and where arbitrarily large punishments are possible, the contracting 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 could arise here were we to optimize over a general contract space.

[6] 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.

be thought of as a vector $(\tau_0, \tau_1, \ldots, \tau_T) \in \mathbb{R}^{T+1}$, which stipulates that the client will pay the dealer $\tau_0 + \sum_{t=1}^{T} \tau_t p_t$. However, the optimal weighted-average-price contract is arguably even more interesting.

2.2 Market model

On-market trades. If the dealer accepts the contract, then he must purchase the required share on the market. Effectively, the dealer will intermediate between the client and 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,[7] $\gamma \geq 0$ parametrizes temporary price impact, and p_0 parametrizes the initial price level. Finally, ε_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 θ and γ is strictly positive.

Remark 3. In addition to being simple and tractable, this specification captures many empirical facts about markets, for example that liquidity is limited over the trading horizon (even when trade is known to be 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 these price dynamics are taken as exogenous for the purposes of our analysis, both these and related dynamics can be readily micro-founded (as in, e.g., Gârleanu and Pedersen, 2016; Kyle, Obizhaeva and Wang, 2018).

Information sets. In each period t, the dealer selects x_t with knowledge of the history $h_t = (p_s, x_s)_{s=1}^{t-1}$. Let $\mathcal{H}_t$ be the set of period-t histories. From an ex-ante perspective (i.e., from the moment after accepting the contract), the dealer can equivalently be thought of as choosing a *trading strategy*: a vector of measurable functions $\boldsymbol{x} = (x_1, \ldots, x_T)^\top$ such that $x_t : \mathcal{H}_t \to \mathbb{R}$ and $\sum_{t=1}^{T} x_t = 1$ almost surely. We denote the set of trading strategies by $\mathcal{X}$.

Remark 4. A special class of trading strategies are those in which the dealer does not condition on previously-realized prices 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 *static* trading strategies.

[7] 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.

Remark 5. 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.

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.[8] 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. Note that, although the client can observe prices, these constitute only a noisy signal of the dealer's trades because prices are also affected by shocks. Mathematically, the client solves the following program

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

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

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

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

3 First-Best Benchmark

Before solving the client's problem itself, we begin by solving for the first-best benchmark. For this benchmark, we remove the friction of hidden action—that is, we assume the client can observe the dealer's trades. We therefore modify the set of feasible contracts accordingly: to preserve comparability with the main analysis, we assume that the feasible contracts are weighted averages of the market prices, but—for this section only—where those weights can depend on the realized trajectory of trades. Mathematically, a contract is a function $\boldsymbol{\tau} : \{\boldsymbol{x} \in \mathbb{R}^T : \sum_{t=1}^{T} x_t = 1\} \to \mathcal{T}$, which maps a trajectory of trades into weights for the market prices.

Given concavity of u, it is optimal to satisfy (IR) by choosing $\boldsymbol{\tau}$ to be the identity function when the dealer acts according to the recommended trading strategy. In fact, note that by choosing $\boldsymbol{\tau}$ to

[8] 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, as subsequent analysis reveals that the dealer has a *unique* best response to any contract $\tau \in \mathcal{T}$ (*cf.* Lemma 2).

be the identity function on the entire domain, the dealer is rendered indifferent among all trading strategies, so that (IC) is trivially satisfied.[9] Plugging $\boldsymbol{\tau} \cdot \boldsymbol{p} = \boldsymbol{x} \cdot \boldsymbol{p}$ into her objective, the client's problem reduces to

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

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

In fact, given that 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 is put on the variance. The classic result (also found 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 obtain the following result:

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

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

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 6. Note that this first-best trading strategy is static; that is, the entire trajectory of trades is determined ex ante. As we will see with Lemma 2 in the next section, an analogous result holds for the second-best problem.

Remark 7. 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 $\boldsymbol{x}^{FB}$ herself.[10] Several considerations might motivate this approach. First, the client might lack the dealer's infrastructure, including market access, order-handling capabilities, risk management, and compliance—each of which requires substantial fixed-cost investments. 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 than the client to possess. 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 γ and θ), 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 $\theta^{client} > \theta$).

[9] In canonical hidden-action models, the standard method for solving the first-best problem would have been to point out that the principal can use a forcing contract to give the agent a strict incentive to take the desired action. Given the structure of $\mathcal{T}$, it would be mathematically complex to describe an appropriate forcing contract in our setting. For simplicity, we therefore use this alternative argument in which (IC) is satisfied, albeit only with equality.

[10] For models of endogenous choice between trading on the market and trading with a dealer, see, e.g., Seppi (1990); Lee and Wang (2022).

4 Discrete-Time Solution

Having solved for the first-best benchmark, this section turns to the second-best problem (as formulated in Section 2.3). 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, an attractive feature of the set of weighted-average-price contracts $\mathcal{T}$ is that for any $\boldsymbol{\tau} \in \mathcal{T}$, the dealer can guarantee himself the payoff $u(0)$ by selecting the static trading strategy $\boldsymbol{x} = \boldsymbol{\tau}$. Intuitively, under this choice of $\boldsymbol{x}$, the dealer's costs (from his on-market trades) are the same weighted average of the market prices that determines 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 simplify the (IC) constraint. Lemma 2 states that, for any contract $\boldsymbol{\tau} \in \mathcal{T}$, the dealer has a unique best response. Thus, (IC) simply requires that the recommended trading strategy be this best response.

Lemma 2. *Define $T \times T$ matrices A, E, and F by*

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

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

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

For any $\boldsymbol{\tau} \in \mathcal{T}$, the dealer has a unique best response in $\mathcal{X}$, which is the static trading strategy $\boldsymbol{x} = FA^{-1}E\boldsymbol{\tau}$.

According to the lemma, the dealer's best response is in fact a static trading strategy. To see the intuition, suppose that after accepting a contract $\boldsymbol{\tau} \in \mathcal{T}$, the dealer makes a tentative plan to pursue a particular static trading strategy $\boldsymbol{x}$. After implementing x_1, the dealer observes p_1, which reveals the realization of ε_1. Would he want to re-optimize $(x_2, \dots, x_T)$? The answer is no. The intuition is that ε_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, ε_1 does not affect his terminal wealth, so learning it is irrelevant. More generally, suppose that after implementing x_t, the dealer learns ε_t. Would he want to re-optimize $(x_{t+1}, \dots, x_T)$? Again, the answer is no. This is because ε_t shifts the dealer's terminal wealth by the constant $\varepsilon_t \sum_{s=1}^{t-1}(x_s - \tau_s)$:

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

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 $(x_{t+1}, \dots, x_T)$.[11]

One implication of Lemma 2 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 2, 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.[12]

Proposition 3. *$FA^{-1}E\boldsymbol{\tau}^{TWAP}$ is frontloaded relative to $\boldsymbol{x}^{FB}$, with equality iff $\theta = 0$.*

Proposition 3 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 the prices, so that regardless of the

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

[12] Formally, we define frontloading as follows. Given two T-dimensional vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, each of whose elements sum to one, we say that $\boldsymbol{a}$ is *frontloaded relative to* $\boldsymbol{b}$ if $\sum_{s=1}^{t} a_s \geq \sum_{s=1}^{t} b_s$ for all t. We also say that a vector is *frontloaded* if it is frontloaded relative to $(\frac{1}{T}, \dots, \frac{1}{T})$.

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 volume from a period t'' to a period $t' < t''$. The direct effect of this deviation 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 prices) 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. 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 the risk of tracking error 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 can in fact then be achieved via $\boldsymbol{\tau}^{TWAP}$, the (IC) constraint notwithstanding.[13] 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

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

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 4. *The weights of the optimal contract and the dealer's on-path 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* $\mathbb{1} = (1, 1, \ldots, 1)^\top$ *denotes a* T*-dimensional vector of ones and*

$$M = \theta A^{-1}E + \theta E^\top (A^{-1})^\top + \gamma FA^{-1}E + \gamma E^\top (A^{-1})^\top F^\top. \tag{2}$$

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

To establish this as the form of the optimal contract, the proof shows that the client's expected payment under a contract $\boldsymbol{\tau}$ can be expressed as $\frac{1}{2}\boldsymbol{\tau}^\top M \boldsymbol{\tau}$. By symmetry of M, the optimal contract

[13] 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, Frei and Mollner (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.

weights satisfy $M\boldsymbol{\tau}^* = \mu\mathbb{1}$, where μ 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 2, we also 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 4. Nevertheless, the logic of the solution can be explained through three special cases: (*i*) when permanent price impact is the dominant consideration (i.e., $\theta \to \infty$), (*ii*) when temporary price impact is the dominant consideration (i.e., $\gamma \to \infty$), and (*iii*) when price risk is the dominant consideration (i.e., $\lambda \to \infty$).[14] 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 dominant consideration

For the case in which permanent price impact is the dominant consideration, suppose that there is no temporary price impact and that the dealer is risk-neutral.

Corollary 5. *Assume that there is no temporary price impact ($\gamma = 0$) and that the dealer is risk-neutral ($\lambda = 0$).*

(i) For any $\boldsymbol{\tau} \in \mathcal{T}$, 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 $\boldsymbol{\tau}^* = (\frac{1}{2}, 0, \ldots, 0, \frac{1}{2})^\top$, so that the dealer's on-path trading strategy is $\boldsymbol{x}^* = (\frac{T+1}{2T}, \frac{1}{2T}, \ldots, \frac{1}{2T})^\top$. The client's expected costs of execution are $p_0 + \frac{\theta(3T+1)}{4T}$.

For the arguments 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 5(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 $\boldsymbol{\tau} \in \mathcal{T}$. Beginning from any static 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.* 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 the 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.

[14] 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 (2) is the zero matrix so that the inverse (needed in the formula for $\boldsymbol{\tau}^*$ in Proposition 4) is not well defined. Thus, Section 4.5 maintains the assumption that at least one of θ and γ is strictly positive, instead considering the limit as $\lambda \to \infty$.

- *Indirect effect.* Of course, prices will not hold fixed. In particular, this shift affects $\mathbb{E}[p_t]$, creating the following indirect effect: $(\tau_t - x_t)\Big(\underbrace{\tfrac{\partial \mathbb{E}[p_t]}{\partial x_t}}_{=\theta} - \underbrace{\tfrac{\partial \mathbb{E}[p_t]}{\partial x_{t+1}}}_{=0}\Big) = (\tau_t - x_t)\theta$.[15]

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 τ_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 $\boldsymbol{\tau}$.[16]

Summing both effects, the total derivative is $\theta[x_{t+1} + \tau_t - x_t]$. Thus, if $\boldsymbol{x}$ were a best response to $\boldsymbol{\tau}$, we would have $x_{t+1} = x_t - \tau_t$ for all $t < T$. Having assumed 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

$$x_t = \frac{1}{T} - \sum_{s=1}^{T} \frac{s}{T}\tau_s + \sum_{s=t}^{T} \tau_s, \tag{3}$$

as claimed by Corollary 5(i).

The client's problem. For intuition into why $\boldsymbol{\tau}^* = (\frac{1}{2}, 0, \ldots, 0, \frac{1}{2})^\top$ 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 are advantageous for the client:

- *Direct effect.* 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 $\boldsymbol{\tau}$.

- *Indirect effect.* Of course, prices will not hold fixed, as a change in $\boldsymbol{\tau}$ leads to a change in the dealer's best response $\boldsymbol{x}$, affecting price dynamics. Using (3), we compute

$$\sum_{s=1}^{t} x_s = \frac{t}{T} + \sum_{s=1}^{t} s\left(1 - \frac{t}{T}\right)\tau_s + \sum_{s=t+1}^{T} t\left(1 - \frac{s}{T}\right)\tau_s. \tag{4}$$

Observe that for all t, the coefficients on $(\tau_1, \ldots, \tau_T)^\top$ in this expression form a weakly concave sequence.[17] Thus, a mean-preserving spread of $\boldsymbol{\tau}$ leads the dealer to backload his trading,

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

[16] This is for any $\boldsymbol{\tau}$ and is therefore a stronger conclusion than that of Proposition 3, which applies only if the offered contract is $\boldsymbol{\tau}^{TWAP}$. On the other hand, Proposition 3 holds for general parameters, whereas this section specializes to the case of $\gamma = \lambda = 0$.

[17] More precisely, the coefficients on $(\tau_1, \ldots, \tau_T)^\top$ constitute an inverted-V, which is maximized at the coefficient on τ_t. The intuition is that $\sum_{s=1}^t x_s$ is increasing in each of $(x_1, \ldots, x_t)$, and given the dealer's frontloading motive,

in the sense of first-order stochastic dominance. 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.[18] 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.* On the one hand, given the 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 prefer to put full weight on the first period.

- *Indirect effect.* On the other hand, price dynamics will respond to the contract. All else equal, the client prefers low prices. Given the permanent price impact, each price is lowest when the dealer backloads his trading as much as possible. Taking into account the contract's influence on the dealer's trading strategy (i.e., that his trading will be frontloaded relative to the contract), prices are then 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 $\tau_1 = \tau_T = \frac{1}{2}$.[19]

4.4 When temporary price impact is the dominant consideration

For the case in which temporary price impact is the dominant consideration, suppose that there is no permanent price impact and that the dealer is risk-neutral. In this case, the client optimally offers the guaranteed TWAP contract, which weights each trading 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.

---

an increase in τ_t leads each of $(x_1, \ldots, x_t)$ to increase. Let us contrast that with τ_{t-1} and τ_{t+1}. An increase in τ_{t-1} leads $(x_1, \ldots, x_{t-1})$ to increase but does not lead x_t to increase. An increase in τ_{t+1} also leads $(x_1, \ldots, x_t)$ to increase, but the effect is more muted because τ_{t+1} also works to increase x_{t+1}.

[18] 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 static 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 $\sum_{s=1}^{t} \bar{x}_s - \delta \min\{t, T-t\} \leq \sum_{s=1}^{t} x_s \leq \sum_{s=1}^{t} \bar{x}_s + \delta \min\{t, T-t\}$, so that the dealer can manipulate p_t by $\theta\delta \min\{t, T-t\}$ in either direction.

[19] 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 equation (4) and $\tau_2 = \cdots = \tau_{T-1} = 0$. Thus, the direct effect of perturbing $\boldsymbol{\tau} = (\tau_1, 0, \ldots, 0, \tau_T)^\top$ 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

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

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

(i) For any $\boldsymbol{\tau} \in \mathcal{T}$, 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 on-path 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 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 the dealer is offered the guaranteed TWAP contract $\boldsymbol{\tau}^{TWAP} = \left(\frac{1}{T}, \dots, \frac{1}{T}\right)^\top$, then he selects $\boldsymbol{x} = (\frac{1}{T}, \dots, \frac{1}{T})^\top$, 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, meaning that this contract must be optimal.

4.5 When price risk is the dominant consideration

For the case in which price risk is the dominant consideration, fix θ and γ, and consider the limit as $\lambda \to \infty$. According to claim (ii) of the following result, the outcome is similar to the case in which temporary price impact is the dominant consideration: 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 7. *Consider the limit as the dealer becomes infinitely risk averse ($\lambda \to \infty$).*

(i) For any $\boldsymbol{\tau} \in \mathcal{T}$, the dealer's best response converges to $\boldsymbol{x} = \boldsymbol{\tau}$.

(ii) The weights of the optimal contract converge to $\boldsymbol{\tau}^* = \left(\frac{1}{T}, \dots, \frac{1}{T}\right)^\top$, so that the dealer's on-path trading strategy converges to $\boldsymbol{x}^* = \left(\frac{1}{T}, \dots, \frac{1}{T}\right)^\top$. 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 a greater emphasis on insuring himself against price shocks. In fact, he can perfectly insure himself by selecting a static 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, $\boldsymbol{\tau} = (\frac{1}{T}, \ldots, \frac{1}{T})^\top$ 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 λ.

4.6 Discussion of the general solution

The general model can be thought of 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. And the general formula for the optimal contract similarly 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 8. *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 4, Figures 1–3 display $\boldsymbol{\tau}^*$ and $\boldsymbol{x}^*$ for various choices of the parameters θ, γ, and λ.[20] The optimal contract weights are depicted in the left panels of these figures; consistent with Corollary 8, they are indeed symmetric. The right panels depict the dealer's on-path trading strategy.

Figures 1–3 suggest that the general solution exhibits several additional qualitative patterns. First, the optimal contract weights are U-shaped: $\tau_1^* \geq \tau_2^* \geq \cdots \geq \tau_{\lceil T/2 \rceil}^* \leq \cdots \tau_{T-1}^* \leq \tau_T^*$.[21] Second, the dealer responds with a trading strategy that is frontloaded in the sense of first-order stochastic dominance: $\sum_{t=1}^{s} x_t^* \geq \frac{s}{T}$ for all $s = 1, \ldots, T$. Third, the severity of both this U-shape and this frontloading is strengthened by θ (the coefficient of permanent price impact), weakened by γ (the coefficient of temporary price impact), and weakened by λ (the dealer's coefficient of absolute risk aversion).

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

[21] 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 8), this convexity condition implies the U-shape condition $\tau_1^* \geq \tau_2^* \geq \cdots \geq \tau_{\lceil T/2 \rceil}^* \leq \cdots \tau_{T-1}^* \leq \tau_T^*$.

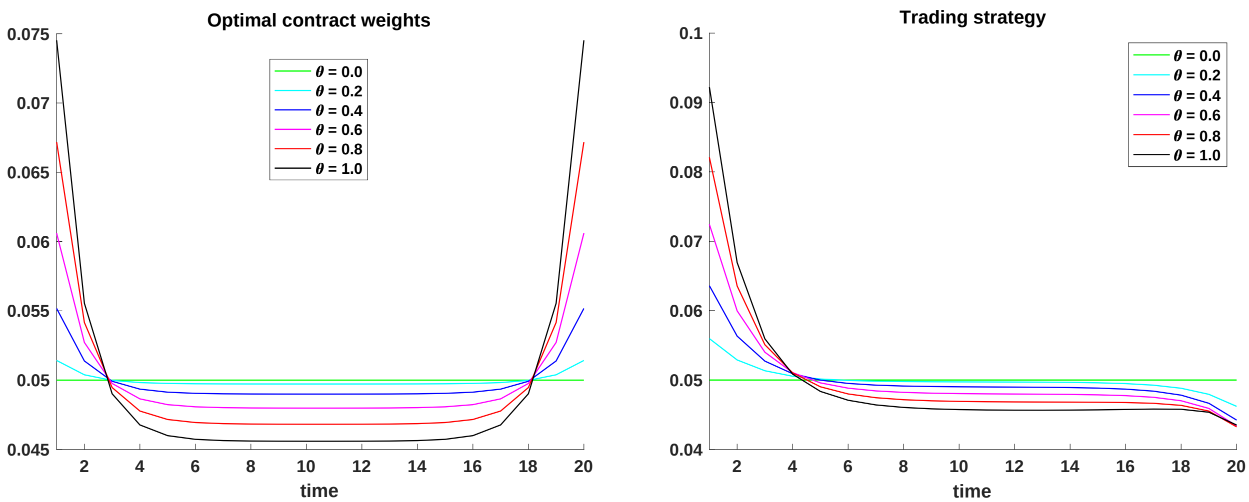

Figure 1: The optimal contract weights and trading strategy for different levels of permanent price impact. When there is no permanent price impact ($\theta = 0$), both the optimal weights and the trading strategy are constant over time. When the permanent price impact becomes larger, the optimal weights become more U-shaped, and the dealer's trading strategy becomes more frontloaded. The other parameters are $\gamma = 1$, $\lambda = 1$, $\sigma = 1$, and $T = 20$.

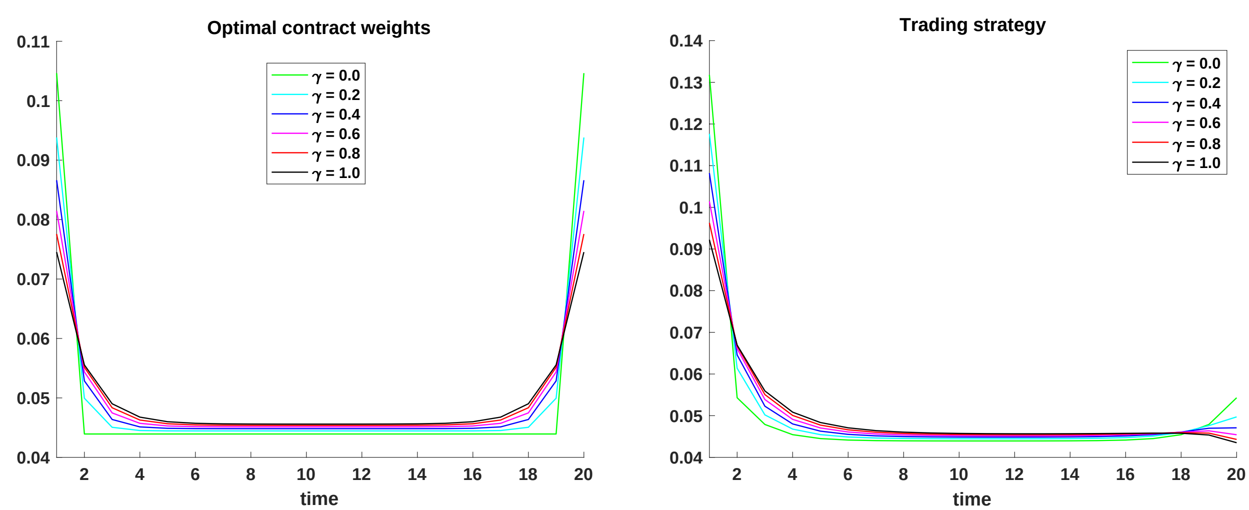

Figure 2: The optimal contract weights and trading strategy for different levels of temporary price impact. When there is no temporary price impact ($\gamma = 0$), the optimal weights are the same for all periods except for the first and last periods, and the dealer's trading strategy is frontloaded. When the temporary price impact becomes larger, the curves for the optimal weights become smoother, and the dealer's trading strategy becomes less frontloaded. The other parameters are $\theta = 1$, $\lambda = 1$, $\sigma = 1$, and $T = 20$.

The intuition for these patterns can be understood through the aforementioned special cases. With permanent price impact as the dominant consideration, we have $\boldsymbol{\tau}^* = (\frac{1}{2}, 0, \ldots, 0, \frac{1}{2})^\top$,

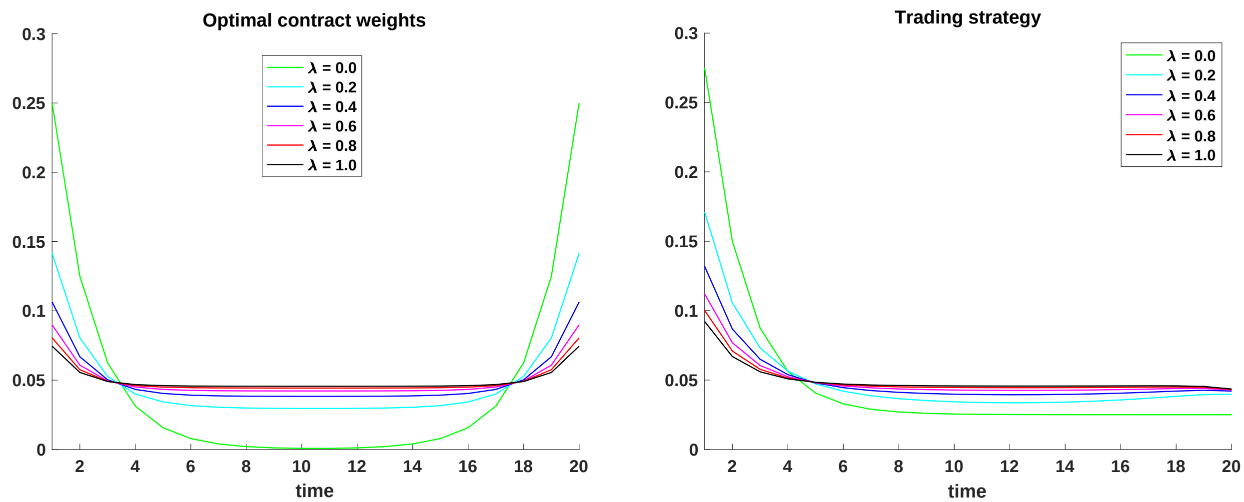

Figure 3: The optimal contract weights and trading strategy for different levels of risk aversion. When risk aversion becomes larger, the optimal weights become less U-shaped, and the dealer's trading strategy becomes less frontloaded. The other parameters are $\theta = 1$, $\gamma = 1$, $\sigma = 1$, and $T = 20$.

which is the maximally-severe U-shape, and $\boldsymbol{x}^* = (\frac{T+1}{2T}, \frac{1}{2T}, \ldots, \frac{1}{2T})^\top$, which is strictly frontloaded. With either temporary price impact or price risk as the dominant consideration, we have $\boldsymbol{\tau}^* = (\frac{1}{T}, \ldots, \frac{1}{T})^\top$, which is the minimally-severe U-shape, and $\boldsymbol{x}^* = (\frac{1}{T}, \ldots, \frac{1}{T})^\top$, which represents minimally-severe frontloading.

The intuition for why the U-shape and frontloading (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 θ 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 γ and λ reduce those severities because they move us toward the limiting cases of Sections 4.4 and 4.5, respectively.

We stress that these observations about the U-shape of the optimal contract and the frontloading of the dealer's on-path 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 consider the continuous-time limit of our discrete-time model. For this limit, we let the number of trading periods diverge (i.e., $T \to \infty$). And at the same time, we also let the distance between consecutive trading periods vanish, so as to hold the execution horizon constant. To capture the latter in this model, we shrink the variance of price shocks to zero (i.e., $\sigma^2 \to 0$) in such a way that $T\sigma^2$ remains constant.

5.1 The optimal contract

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

Proposition 9. *Assume the dealer is strictly risk-averse ($\lambda > 0$). 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 $\boldsymbol{\tau}^{*k} \in \mathcal{T}^{T_k}$ be the associated optimal contract, and let $\boldsymbol{x}^{*k}$ be the dealer strategy that best responds to $\boldsymbol{\tau}^{*k}$. 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 thought of as an extreme U-shape: the 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.[22] 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 larger than the terminal atom.

Comparative statics. Note that this density a is increasing in (λ, T, σ^2) and decreasing in θ. Because a is inversely related to the severity of the optimal contract's U-shape, these relationships are consistent 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 θ implies a larger expected gap and a larger frontloading motive. On the other hand, larger $T\sigma^2$ implies a larger variance for this gap, hence more price risk, and a smaller frontloading motive. Larger λ means less risk-bearing capacity, and 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 could obtain her first-best payoff from a guaranteed TWAP contract (i.e., the case of $a = 1$). By similar logic, smaller (larger) frontloading motives imply larger (smaller) values of a.

[22] 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 superficial. They solve a different problem (a problem of optimal execution rather than one of optimal contracting) under a different set of assumptions.

In the continuous-time limit considered in Proposition 9, the client's expected costs of execution exceed the execution cost of the first best by $\frac{1-a}{4}\theta$, resulting as the difference between $p_0 + \frac{3-a}{4}\theta$ and $p_0 + \frac{1}{2}\theta$.[23] This wedge is decreasing in a, as larger values of a mean smaller frontloading of the optimal contract and a trading strategy that is closer to the first best.[24,25]

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

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 temporary price impact does has no effect on the limit characterized by Proposition 9. 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 γ would raise prices and hence the client's payment. But on the other hand, an increase in γ 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

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

[23] A related observation is that, using the notation introduced in Remark 7, the client benefits from contracting with a dealer (rather than trading directly on the market) if $p_0 + \frac{3-a}{4}\theta < p_0 + \frac{1}{2}\theta^{client}$, or if $\theta^{client} > \frac{3-a}{2}\theta$.

[24] Because a is increasing in (λ, T, σ^2), it follows that this wedge is decreasing in those parameters. Moreover, θ enters the expression for this wedge both directly and through a. Because a is decreasing in θ, both effects go in the same direction: the wedge is increasing in θ.

[25] Appendix A.11 decomposes this wedge as the sum of the dealer's expected profit and an inefficiency due to suboptimal trading (i.e., the fact that $\boldsymbol{x}^* \neq \boldsymbol{x}^{FB}$):

$$\underbrace{\frac{1-a}{4}\theta}_{\substack{\text{wedge between first-best} \\ \text{and second-best payments}}} = \underbrace{\frac{1-a}{2}\theta - \frac{(1-a)^2}{4}\frac{\theta^2}{\theta+2\gamma}}_{\text{dealer's expected profit}} + \underbrace{\frac{(1-a)^2}{4}\frac{\theta^2}{\theta+2\gamma} - \frac{1-a}{4}\theta}_{\text{inefficiency from suboptimal trading}}.$$

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 9 specifies for the jumps of $\lim_{k\to\infty} \sum_{t=1}^{\lceil qT_k \rceil} \tau_t^{*k}$ at $q = 0$ and $q = 1$. Figure 4 illustrates this convergence.[26]

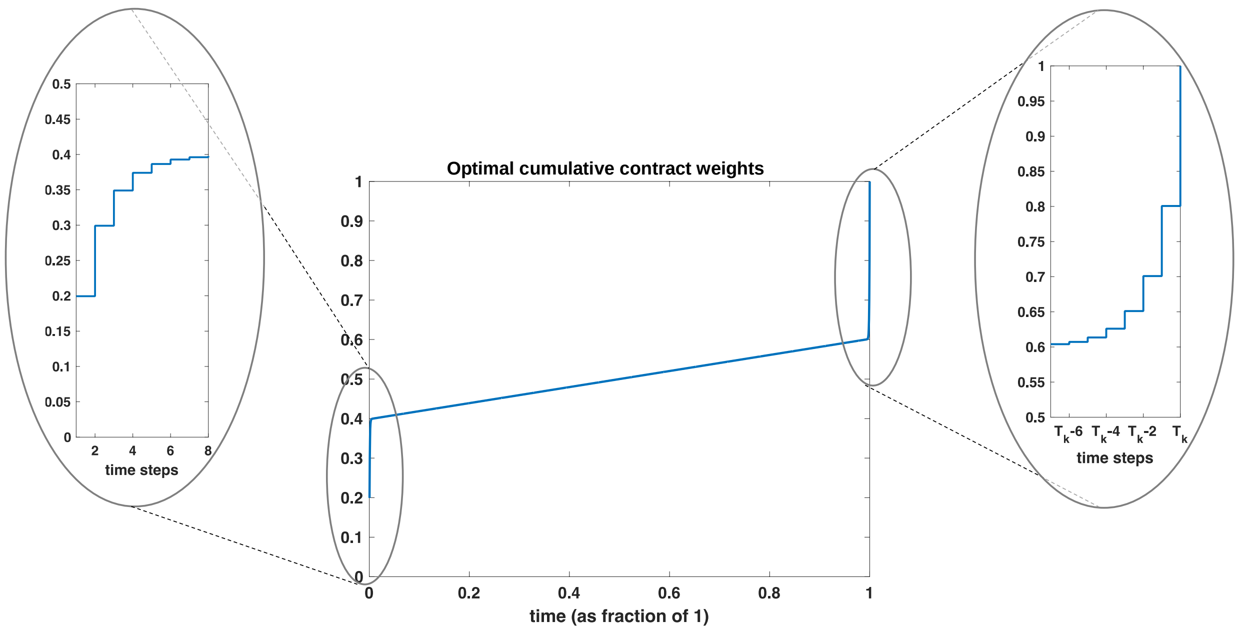

Figure 4: The optimal contract for a large k (such that $T_k = 2{,}000$). There are two sequences of discrete weights at the beginning and end of the trading times while the weights are smooth for interior times. The parameters are $\theta = 1$, $\gamma = 1$, $\lambda = 1$, $T = 1$, and $\sigma = 1$.

Interestingly, the optimal trading strategy has a different form: it has a sequence of discrete weights only at the beginning, but not at the end of the trading times.[27] Without temporary price impact, only the first element of the trading strategy converges to a nonzero value

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

[26] Although λ 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 the purposes of these periods around the boundary times, the dealer behaves in the limit as if he were risk-neutral (regardless of λ). The role of γ is exactly the opposite: it affects the division of weight across the sequences, but not the weight assigned to the sequences in total.

[27] This observation is consistent with the best-response form $\boldsymbol{x} = FA^{-1}E\boldsymbol{\tau}$ in Lemma 2: We can check that $\boldsymbol{\tau}$ given by $\tau_{j+1} = c\frac{\theta\gamma^j}{(\theta+\gamma)^{j+1}}$ for a constant $c > 0$ and all $j = 0, 1, \ldots, T-1$ is an eigenvector of the matrix $FA^{-1}E$ to eigenvalue 1. Hence, $\boldsymbol{x} = \boldsymbol{\tau}$ for such a $\boldsymbol{\tau}$, which explains why the contract and trading strategy have the same sequence of discrete weights at the beginning of the trading times. The same argument does not apply to the weights at the end of the trading times. Indeed, if we set $\boldsymbol{x}^k = FA^{-1}E\boldsymbol{\tau}^k$ for $\boldsymbol{\tau}^k$ given by $\tau_{T_k-j}^k = c\frac{\theta\gamma^j}{(\theta+\gamma)^{j+1}}$ for a constant $c > 0$ and all $j = 0, 1, \ldots, T_k - 1$, we can compute $\lim_{k\to\infty} x_{T_k-j}^k = 0$ for all j.

When there is temporary price impact, we have

$$\lim_{k\to\infty} x^{*k}_{j+1} = \frac{\theta\gamma^j}{(\theta+\gamma)^{j+1}}\frac{1-a}{2}, \quad \lim_{k\to\infty} x^{*k}_{T_k-j} = 0 \text{ for any fixed } j \geq 0. \tag{5}$$

However, the jumps of $\lim_{k\to\infty}\sum_{t=1}^{\lceil qT_k\rceil} x^{*k}_t$ at $q=0$ and $q=1$ are not determined only by these sequences of discrete weights. As stated in Proposition 9, 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 (5), but also another infinite sum whose terms individually converge to zero but whose sum converges to $\frac{1-a}{4}$. This is illustrated in Figure 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.

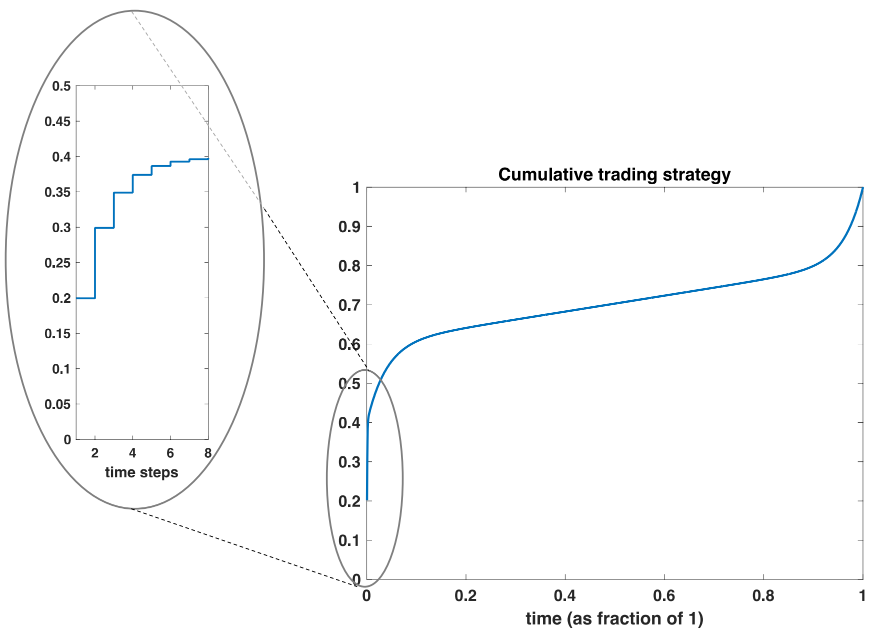

Figure 5: The cumulative trading strategy for a large k (such that $T_k = 2{,}000$). There is a sequence of discrete weights at the beginning of the trading times while the weights are smooth for interior times and towards the end (although part of the smooth curve converges to a vertical line). The parameters are $\gamma=1$, $\theta=1$, $\lambda=1$, $T=1$, and $\sigma=1$.

Proof strategy. We begin by conjecturing three different regions of convergence for the contract: the first S_k periods, the last S_k periods, and the middle $T_k - 2S_k$ periods, where $S_k \to \infty$ and

$\frac{S_k^3}{T_k} \to 0$.[28,29] We demonstrate that the limiting optimal contract entails a constant density in the middle region. Given an arbitrary such density a, we demonstrate that the initial and terminal regions converge to the limits described above. Plugging in, the client's cost in the limit can be expressed as a quadratic function of a, which is minimized at the a reported in the proposition.

To explain in more detail why the limiting optimal contract entails a constant density at interior times, define $X_q = \lim_{k\to\infty} \sum_{t=1}^{\lceil qT_k \rceil} x_t^k$ and $V_q = \lim_{k\to\infty} \sum_{t=1}^{\lceil qT_k \rceil} \tau_t^k$. In the proof, we demonstrate that the continuous-time limit of the dealer's best response function as defined by Lemma 2 is $X_q = \frac{\theta}{\lambda T \sigma^2} \dot{V}_q + V_q$ for all $q \in (0,1)$. This equation reflects the dealer's frontloading motive: provided $\dot{V}_q \geq 0$ (as holds in the optimum), we have $X_q \geq V_q$. It likewise indicates that this frontloading motive is strengthened by θ and weakened by (λ, T, σ^2). Plugging in this best response, the client's expected costs due to permanent price impact at the interior times $q \in (0,1)$ are

$$\theta \int_0^1 X_q \, dV_q = \frac{\theta^2}{\lambda T \sigma^2} \int_0^1 \dot{V}_q^2 \, dq + \underbrace{\theta \int_0^1 V_q \dot{V}_q \, dq}_{= \frac{\theta}{2}\left(V_{1-}^2 - V_{0+}^2\right)} . \tag{6}$$

In fact, permanent price impact is the client's only consideration at interior times—that temporary price impact is avoided follows from continuity of X_q and V_q on $q \in (0,1)$.[30] Given arbitrary values for V_{1-} and V_{0+}, the client's problem for the interior times therefore distills to choosing V_q to minimize the objective (6), which is done by choosing $\dot{V}_q$ to be a constant. This constant density at the interior times is the key for understanding the shape of the optimal contract—intuitively, the atoms at the boundary instants follow because they are then the only way to render a U-shape.

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 common contracts? By how much do they underperform the optimal contract? Under what situations, if any, do they deliver outcomes that are close to the client's second-best payoff? The following result allows us to answer these.

[28] This proof strategy is not fully rigorous because it assumes that the cumulative weights of the optimal contracts converge to a smooth function (except for jumps at 0 and 1) and determines the limit under this assumption. Although this assumption is consistent with numerical experimentation, a fully rigorous proof would also demonstrate the nature of the convergence. For the special case of no temporary price impact, we can produce such a proof, and we include it in Appendix OA.A.2. We also note that this proof strategy relies on a conjecture only about the convergence of the optimal contract—and no analogous conjecture about the dealer's best response, whose convergence behavior is somewhat more complicated (i.e., the convergence of part of the curve to a vertical line illustrated by Figure 5).

[29] The reason we require $\frac{S_k^3}{T_k} \to 0$ rather than simply $\frac{S_k}{T_k} \to 0$ will become clear in the proof. In short, it is because we approximate the client's expected costs in the continuous-time limit by terms with errors of order $\frac{S_k^3}{T_k}$, where S_k^3 comes from multiple layers of sums related to the permanent price impact and the contract.

[30] Indeed, both $\max_{1+S_k \leq t \leq T_k - S_k} |x_t^{*k}|$ and $\max_{1+S_k \leq t \leq T_k - S_k} |\tau_t^{*k}|$ are $O(\frac{1}{T_k})$, meaning that the client's expected costs due to temporary price impact at these interior times are $\gamma \sum_{t=1+S_k}^{T_k - S_k} x_t^{*k} \tau_t^{*k} \to 0$.

Proposition 10. *Assume the dealer is strictly risk-averse ($\lambda > 0$). 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.*

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

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

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

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

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.[31] Hence, introducing a terminal atom is one way in 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 that is sometimes referred to as 'banging the close' in practice. Such extraordinarily concentrated trading is inefficient, and one way in 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 the continuous-time limit of our model. Consider a client who desires to trade a position, currently valued at $V = \$100$ million, in a certain stock. Let the parameters be $p_0 = \$100, \theta = 2\times10^{-6}$, $\gamma = 0$, $\lambda = 2\times10^{-6}$, $T = 1$, and $\sigma^2 = 6.1$.[32] Under these parameters, the optimal contract

[31] In terms of the notation introduced earlier, we have $V_q^{TWAP} = q$, so that $X_q^{TWAP} = \frac{\theta}{\lambda T\sigma^2}\dot{V}_q^{TWAP} + V_q^{TWAP} = \frac{\theta}{\lambda T\sigma^2} + q$ for $q \in (0,1)$. Suppose we modify it by collecting contract weights from interior times after $q = 1 - \frac{\theta}{\lambda T\sigma^2}$ into an atom on the terminal instant. This yields an alternative contract defined for $q \in (0,1)$ by

$$V_q = \begin{cases} q & \text{if } q \le 1 - \frac{\theta}{\lambda T\sigma^2} \\ 1 - \frac{\theta}{\lambda T\sigma^2} & \text{if } q > 1 - \frac{\theta}{\lambda T\sigma^2} \end{cases} \quad \text{which induces} \quad X_q = \begin{cases} \frac{\theta}{\lambda T\sigma^2} + q & \text{if } q \le 1 - \frac{\theta}{\lambda T\sigma^2} \\ 1 & \text{if } q > 1 - \frac{\theta}{\lambda T\sigma^2} \end{cases}$$

thereby eliminating overbuying. The resulting outcome is better for the client, because it weights the terminal price rather than interior prices that would have been inflated by the dealer's overbuying.

[32] 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

puts 19.8 percent of its weight on the opening price, 19.8 percent of its weight on the closing price, and 60.4 percent of its weight on the intraday time-weighted average price.

We then use our results to compute model-implied transaction costs (measured by implementation shortfall) under various scenarios. Whereas our theoretical analysis normalized the 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., SEC, 2005; Abel Noser, 2021; WSJ, 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 $\boldsymbol{\tau}^{MOC}$, transaction costs 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. The 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 outcome 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 extrapolate to billions of dollars per year.[33] Of course, additional cost savings could be achieved by optimizing over an even larger set of contracts (e.g., the set of affine contracts considered in Appendix OA.B, or a fully general set as discussed in footnote 5). 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.

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 point in 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 a classic one of optimal execution. Indeed, our analysis of the first-best problem recovers classic results from that literature about the optimality of trading at a constant rate.

results ranging from 0.63×10^{-6} to 2.03×10^{-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 2×10^{-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.

[33] We start with a conservative estimate of institutional transaction costs of \$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.

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: it consists of two 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.

References

Abel Noser Holdings LLC, "Hidden Figures–Introducing a Portfolio Transition & Event Universe," October 2021. https://bit.ly/3FksgpG.

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

Avramov, Doron, Tarun Chordia, and Amit Goyal, "The Impact of Trades on Daily Volatility," *The Review of Financial Studies*, 2006, *19* (4), 1241–1277.

Baldauf, Markus and Joshua Mollner, "Competition and Information Leakage," *Working Paper*, 2022. https://ssrn.com/abstract=3744499.

_ , Christoph Frei, and Joshua Mollner, "Principal Trading Arrangements: When Are Common Contracts Optimal?," *Management Science*, 2022, *68* (4), 3112–3128.

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

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

Bhattacharya, Sudipto and Paul Pfleiderer, "Delegated Portfolio Management," *Journal of Economic Theory*, 1985, *36* (1), 1–25.

Biais, Bruno, Thomas Mariotti, Guillaume Plantin, and Jean-Charles Rochet, "Dynamic Security Design: Convergence to Continuous Time and Asset Pricing Implications," *The Review of Economic Studies*, 2007, *74* (2), 345–390.

Bloomberg, "Cairn Energy Said to Be Victim of HSBC Currency Frontrunning," July 2016. https://bloom.bg/3sBA14k.

_ , "Barclays Trader Beats U.S. Prosecution of Front-Running Charges," March 2019. https://bloom.bg/3Dzbcy0.

_ , "Hedge Funds Press for Crackdown on Front-Running Loophole in EU," August 2020. https://bloom.bg/3Lr0iul.

_ , "Japan's SMBC Nikko, Staff Charged With Market Manipulation," March 2022. https://bloom.bg/3F8zRIJ.

_ , "Morgan Stanley Among Block-Trading Firms Facing U.S. Probe," February 2022. https://bloom.bg/3Lyzb1z.

_ , "Morgan Stanley's Passi Faces U.S. Block-Trading Probe," February 2022. https://bloom.bg/3OJQoXL.

_ , "'The Morgan Stanley Fade': U.S. Probe Dredges Up Years of Animus," March 2022. https://bloom.bg/3ky6FkN.

_ , "What Block Trades Are and Why the SEC's Investigating," February 2022. https://bloom.bg/3y906Ns.

Board, Simon and Andrzej Skrzypacz, "Revenue Management with Forward-Looking Buyers," *Journal of Political Economy*, 2016, *124* (4), 1046–1087.

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

Buffa, Andrea M., Dimitri Vayanos, and Paul Woolley, "Asset Management Contracts and Equilibrium Prices," *Journal of Political Economy*, 2022, *130* (12), 3025–3342.

Campo, Sandra, Emmanuel Guerre, Isabelle Perrigne, and Quang Vuong, "Semiparametric Estimation of First-Price Auctions with Risk-Averse Bidders," *The Review of Economic Studies*, 01 2011, *78* (1), 112–147.

Carpenter, Jennifer N., "Does Option Compensation Increase Managerial Risk Appetite?," *The Journal of Finance*, 2000, *55* (5), 2311–2331.

Cartea, Alvaro and Sebastian Jaimungal, "Incorporating Order-Flow into Optimal Execution," *Mathematics and Financial Economics*, 2016, *10* (3), 339–364.

Curello, Gregorio and Ludvig Sinander, "Screening for Breakthroughs," *Working Paper*, 2021. https://arxiv.org/abs/2011.10090.

DeMarzo, Peter M. and Michael J. Fishman, "Optimal Long-Term Financial Contracting," *The Review of Financial Studies*, 2007, *20* (6), 2079–2128.

Di Tella, Sebastian and Yuliy Sannikov, "Optimal Asset Management Contracts With Hidden Savings," *Econometrica*, 2021, *89* (3), 1099–1139.

Duffie, Darrell and Piotr Dworczak, "Robust Benchmark Design," *Journal of Financial Economics*, 2021, *142* (2), 775–802.

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

Financial Industry Regulatory Authority, "FINRA Rule 5270. Front Running of Block Transactions," September 2013. http://bit.ly/3pgazzw.

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

Gârleanu, Nicolae and Lasse Heje Pedersen, "Dynamic Trading with Predictable Returns and Transaction Costs," *The Journal of Finance*, 2013, *68* (6), 2309–2340.

_ **and** _ , "Dynamic Portfolio Choice with Frictions," *Journal of Economic Theory*, 2016, *165*, 487–516.

Garrett, Daniel F., "Intertemporal Price Discrimination: Dynamic Arrivals and Changing Values," *American Economic Review*, 2016, *106* (11), 3275–3299.

Goldman Sachs & Co. LLC, "Disclosure Regarding FINRA Rule 5270," 2017. https://www.goldmansachs.com/disclosures/disclosure-regarding-finra-rule-5270.pdf.

Grenadier, Steven R., Andrey Malenko, and Nadya Malenko, "Timing Decisions in Organizations: Communication and Authority in a Dynamic Environment," *American Economic Review*, 2016, *106* (9), 2552–2581.

Hellwig, Martin F. and Klaus M. Schmidt, "Discrete-Time Approximations of the Holmström-Milgrom Brownian-Motion Model of Intertemporal Incentive Provision," *Econometrica*, 2002, *70* (6), 2225–2264.

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

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

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

Kruse, Thomas and Philipp Strack, "Optimal Stopping with Private Information," *Journal of Economic Theory*, 2015, *159*, 702–727.

Kyle, Albert S., Anna A. Obizhaeva, and Yajun Wang, "Smooth Trading with Overconfidence and Market Power," *The Review of Economic Studies*, 2018, *85* (1), 611–662.

Lee, Tomy and Chaojun Wang, "Why Trade Over-the-Counter?," *Working Paper*, 2022. https://finance.wharton.upenn.edu/~wangchj/papers/OTC_EX.pdf.

Madsen, Erik, "Designing Deadlines," *American Economic Review*, 2022, *112* (3), 963–997.

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

Nasdaq, Inc., "How Much Does Trading Cost the Buy Side?," February 2022. https://bit.ly/3U1O2Gw.

Obizhaeva, Anna A. and Jiang Wang, "Optimal Trading Strategy and Supply/Demand Dynamics," *Journal of Financial Markets*, 2013, *16* (1), 1–32.

Risk.net, "Page 19901: The Benchmark that Time Forgot," June 2021. https://www.risk.net/7818166.

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

Sannikov, Yuliy, "A Continuous-Time Version of the Principal-Agent Problem," *The Review of Economic Studies*, 2008, *75* (3), 957–984.

Securities Industry and Financial Markets Association, "Global Equity Markets Primer: Market Costs," November 2021. https://bit.ly/3eGr5IL.

Seppi, Duane J., "Equilibrium Block Trading and Asymmetric Information," *The Journal of Finance*, 1990, *45* (1), 73–94.

The Wall Street Journal, "Big Stock Sales Are Supposed to Be Secret. The Numbers Indicate They Aren't.," March 2022. https://on.wsj.com/3LrYQZy.

_ , "How We Analyzed Wall Street Block Trades," March 2022. https://on.wsj.com/3vwaLjP.

_ , "Investment Firm Files Finra Arbitration Demand Against Morgan Stanley," April 2022. https://on.wsj.com/3OMjxlc.

_ , "Regulators Probe Block Trading at Morgan Stanley, Goldman, Other Wall Street Firms," February 2022. https://on.wsj.com/3w7qK71.

Traders Magazine, "NYSE & NASD Investigate Guaranteed VWAP," September 2005. https://bit.ly/3BSsu8f.

_ , "VWAP Debate Divides the Trading Industry," November 2005. https://bit.ly/3vmUspm.

U.S. Securities and Exchange Commission, "Regulation NMS: Final Rules and Amendments to Joint Industry Plans," *Federal Register*, June 29 2005, *70* (124).

A Proofs of Results Stated in the Main Text

Except for the proof of Proposition 9, where σ_k depends on k, we assume for all proofs that $\sigma = 1$ without loss of generality. (For cases of $\sigma \neq 1$, we would simply replace λ by $\lambda\sigma^2$ throughout.)

A.1 Proof of Proposition 1

Lemma 11. *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 11. 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

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

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

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 11. Observe that $\boldsymbol{\varepsilon}$ has fallen out of the expression. Thus, the first-best trading strategy, which minimizes this expression, will not be a function of $\boldsymbol{\varepsilon}$—in other words, it will be static. This first-best trading strategy solves the program

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

Taking the Lagrangian (with μ as the multiplier on the constraint), we obtain

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

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

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

A.2 Proof of Lemma 2

Proof. The dealer's expected utility equals

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

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

$$-\mathbb{E}_{\boldsymbol{x}}[u(\boldsymbol{\tau} \cdot \boldsymbol{p} - \boldsymbol{x} \cdot \boldsymbol{p})] = \mathbb{E}_{\boldsymbol{x}}\left[\exp\left(\lambda \sum_{t=1}^{T} (X_t - X_{t-1} - \tau_t)\big(\theta X_t + \gamma (X_t - X_{t-1})\big) - \lambda \sum_{t=1}^{T} \varepsilon_t \left(X_{t-1} - \sum_{j=1}^{t-1} \tau_j\right)\right)\right] \tag{9}$$

over $X_1, X_2, \ldots, X_{T-1}$. We denote by $\mathcal{F}_t$ the σ-algebra generated by $\varepsilon_1, \ldots, \varepsilon_t$. 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 (9) on $\mathcal{F}_{T-1}$ and will then go backward subsequently. From the law of iterated expectations, we obtain

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

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 ε_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

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

which we rewrite as

$$
-\big(\theta+2\gamma\big)X_T+\big(\lambda+2\theta+4\gamma\big)X_{T-1}-(\theta+2\gamma)X_{T-2}=-\gamma\tau_T+(\theta+\gamma)\tau_{T-1}+\lambda\sum_{j=1}^{T-1}\tau_j, \tag{10}
$$

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

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

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

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

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 (10) 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

$$-(\theta+2\gamma)X_{t+1}+(\lambda+2\theta+4\gamma)X_t-(\theta+2\gamma)X_{t-1}=-\gamma\tau_{t+1}+(\theta+\gamma)\tau_t+\lambda\sum_{j=1}^{t}\tau_j,\quad t=1,2,\dots,T-1. \tag{11}$$

This linear system of equations can be written as $A\boldsymbol{X}=E\boldsymbol{\tau}$ using the $T\times T$ matrices A and E from (1). We conclude that $\boldsymbol{x}=F\boldsymbol{X}=FA^{-1}E\boldsymbol{\tau}$. □

A.3 Proof of Proposition 3

Proof. We can write $A = \tilde{I}\tilde{A}$, where

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

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

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 $E\boldsymbol{\tau}^{TWAP} = \frac{1}{T}(\theta + \lambda\sigma^2, \theta + 2\lambda\sigma^2, \ldots, \theta + (T-1)\lambda\sigma^2, T)^\top$, and $AF^{-1}\boldsymbol{x}^{FB} = \frac{1}{T}(\lambda\sigma^2, 2\lambda\sigma^2, \ldots, (T-1)\lambda\sigma^2, T)^\top$, so $E\boldsymbol{\tau}^{TWAP} \geq AF^{-1}\boldsymbol{x}^{FB}$ (where $\geq$ is in the component-wise sense). Using the fact that A^{-1} is nonnegative,

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

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

A.4 Proof of Proposition 4

Proof. The client's expected cost of a contract $\boldsymbol{\tau} \cdot \boldsymbol{p}$ is

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{\tau} \cdot \boldsymbol{p}] = \mathbb{E}_{\boldsymbol{x}}\left[\sum_{t=1}^{T} \tau_t p_t\right] &= \mathbb{E}_{\boldsymbol{x}}\left[\sum_{t=1}^{T} \tau_t \left(p_0 + \sum_{j=1}^{t}(\theta x_j + \varepsilon_j) + \gamma x_t\right)\right] = p_0 + \sum_{t=1}^{T} \tau_t \left(\sum_{j=1}^{t} \theta x_j + \gamma x_t\right) \\ &= p_0 + \theta \sum_{t=1}^{T} \tau_t X_t + \gamma \sum_{t=1}^{T} \tau_t (X_t - X_{t-1}) = p_0 + \theta \boldsymbol{\tau}^\top A^{-1} E \boldsymbol{\tau} + \gamma \boldsymbol{\tau}^\top F A^{-1} E \boldsymbol{\tau} \\ &= p_0 + \frac{1}{2} \boldsymbol{\tau}^\top M \boldsymbol{\tau} \end{aligned} \tag{12}$$

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

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

hence $\boldsymbol{\tau}^* = \mu M^{-1}\mathbb{1}$ and $\mathbb{1}^\top \boldsymbol{\tau}^* = \mu \mathbb{1}^\top M^{-1} \mathbb{1} = 1$. We obtain $\mu = \frac{1}{\mathbb{1}^\top M^{-1} \mathbb{1}}$ and thus $\boldsymbol{\tau}^* = \frac{1}{\mathbb{1}^\top M^{-1} \mathbb{1}} M^{-1}\mathbb{1}$ and $\boldsymbol{x}^* = F A^{-1} E \boldsymbol{\tau}^* = \frac{1}{\mathbb{1}^\top M^{-1} \mathbb{1}} F A^{-1} E M^{-1} \mathbb{1}$, using Lemma 2. We can compute the client's expected costs of execution under $\boldsymbol{\tau}^* = \frac{1}{\mathbb{1}^\top M^{-1} \mathbb{1}} M^{-1}\mathbb{1}$ as

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

□

A.5 Proof of Corollary 5

Proof. Claim (i): It follows from (11) with $\gamma = \lambda = 0$ that

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

which implies

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

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:

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

Next, define $\boldsymbol{v} = (\frac{1}{2}, 0, \ldots, 0, \frac{1}{2})^\top$. 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

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

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. □

A.6 Proof of Corollary 6

Proof. Claim (i): It follows from (11) with $\theta = \lambda = 0$ that

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

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): To prove $\boldsymbol{\tau}^* = \left(\frac{1}{T}, \ldots, \frac{1}{T}\right)^\top$, 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 4. To aid in showing that, we first define

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

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

$$M\mathbb{1} = \gamma FA^{-1}E\mathbb{1} + \gamma E^\top(A^{-1})^\top F^\top\mathbb{1} = \gamma FA^{-1}\boldsymbol{v}_1 + \gamma E^\top(A^{-1})^\top\boldsymbol{v}_2 = \gamma F\boldsymbol{v}_3 + \gamma E^\top\boldsymbol{v}_2 = \gamma\mathbb{1} + \gamma\mathbb{1} = 2\gamma\mathbb{1}, \tag{13}$$

establishing that $\mathbb{1}$ is an eigenvector of M, as required. For the dealer's trading strategy, we deduce from Proposition 4 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 (13) 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 $p_0 + \frac{1}{2\mathbb{1}^\top M^{-1}\mathbb{1}} = p_0 + \frac{\gamma}{T}$. □

A.7 Proof of Corollary 7

Proof. Claim (i): Dividing (11) by λ and then letting λ go to infinity gives $X_t = \sum_{j=1}^{t} \tau_j$ for $t = 1, 2, \ldots, T-1$, hence $x_t = \tau_t$ for all $t = 1, 2, \ldots, T$.

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

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

We begin by showing $\lim_{\lambda\to\infty} A^{-1}E = Q$. To this end, we note that Λ^{-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

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

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

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

So from (2), we deduce

$$\lim_{\lambda\to\infty} M = \theta\left(\lim_{\lambda\to\infty} A^{-1}E\right) + \theta\left(\lim_{\lambda\to\infty} A^{-1}E\right)^{\top} + \gamma\left(F\lim_{\lambda\to\infty} A^{-1}E\right) + \gamma\left(F\lim_{\lambda\to\infty} A^{-1}E\right)^{\top}$$

$$= \theta Q + \theta Q^{\top} + \gamma FQ + \gamma(FQ)^{\top} = \theta Q + \theta Q^{\top} + 2\gamma I = \begin{pmatrix} 2\theta+2\gamma & \theta & \cdots & \theta \\ \theta & 2\theta+2\gamma & \cdots & \theta \\ \vdots & & \ddots & \vdots \\ \theta & \theta & \cdots & 2\theta+2\gamma \end{pmatrix}.$$

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

$$\lim_{\lambda\to\infty} \boldsymbol{\tau}^* = \lim_{\lambda\to\infty}\left(\frac{1}{\mathbb{1}^{\top}M^{-1}\mathbb{1}}M^{-1}\mathbb{1}\right) = \frac{1}{\lim_{\lambda\to\infty}\mathbb{1}^{\top}M^{-1}\mathbb{1}}\lim_{\lambda\to\infty} M^{-1}\mathbb{1}$$

$$= \frac{2\gamma+\theta(T+1)}{T}\frac{1}{2\gamma+\theta(T+1)}\mathbb{1} = \frac{1}{T}\mathbb{1},$$

$$\lim_{\lambda\to\infty} \boldsymbol{x}^* = \lim_{\lambda\to\infty} FA^{-1}E\boldsymbol{\tau}^* = F\left(\lim_{\lambda\to\infty} A^{-1}E\right)\left(\lim_{\lambda\to\infty}\boldsymbol{\tau}^*\right) = FQ\left(\frac{1}{T}\mathbb{1}\right) = I\left(\frac{1}{T}\mathbb{1}\right) = \frac{1}{T}\mathbb{1},$$

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. □

A.8 Proof of Corollary 8

Proof. Define the $T\times T$ anti-diagonal matrix

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

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

$$E_1=\begin{pmatrix}1&0&\cdots&0\\0&1&&0\\\vdots&&\ddots&\\0&0&\cdots&0\end{pmatrix}\qquad E_2=\begin{pmatrix}0&0&\cdots&0\\1&0&&\vdots\\0&1&\ddots&0\\\vdots&&\ddots&0\\0&0&\cdots&0\end{pmatrix}$$

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

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

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

where $b_{t+1}=\gamma/(\gamma+\theta)b_t$ for all $t\in\{1,2\ldots,T-1\}$.[34] Each of the following matrices is symmetric:

$$A_1(\theta I+\gamma F)^{-1}PE_1=\begin{pmatrix}0&0&\cdots&0&0\\0&0&\cdots&b_1&0\\\vdots&&\iddots&\vdots&\vdots\\0&b_1&\cdots&b_{T-2}&0\\0&0&\cdots&0&0\end{pmatrix}$$

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

$$A_1(\theta I+\gamma F)^{-1}PE_2 = \begin{pmatrix} 0 & \cdots & 0 & b_1 & 0 \\ 0 & \cdots & b_1 & b_2 & 0 \\ \vdots & ⋰ & \vdots & \vdots & \\ b_1 & b_2 & \cdots & b_{T-1} & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}$$

$$A_1(\theta I+\gamma F)^{-1}PE_3 = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & b_1 & \cdots & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & \sum_{t=1}^{T-3} b_t & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}$$

$$A_2(\theta I+\gamma F)^{-1}PE_1 = \begin{pmatrix} 0 & \cdots & & & 0 & b_1 & 0 \\ 0 & \cdots & & & b_1 & b_2 & 0 \\ 0 & \cdots & & & b_2 & b_1+b_3 & 0 \\ 0 & \cdots & & & b_1+b_3 & b_2+b_4 & 0 \\ \vdots & ⋰ & ⋰ & ⋰ & ⋰ & \vdots & \vdots \\ b_1 & b_2 & b_1+b_3 & b_2+b_4 & \cdots & b_{T-3}+b_{T-1} & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}$$

$$A_2(\theta I+\gamma F)^{-1}PE_2 = \begin{pmatrix} 0 & \cdots & 0 & b_1 & b_2 & 0 \\ 0 & & b_1 & b_2 & b_1+b_3 & 0 \\ 0 & ⋰ & b_2 & b_1+b_3 & b_2+b_4 & 0 \\ b_1 & ⋰ & ⋰ & ⋰ & \vdots & \vdots \\ b_2 & b_1+b_3 & b_2+b_4 & \cdots & b_{T-2}+b_T & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}$$

$$A_2(\theta I+\gamma F)^{-1}PE_3 = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & b_1 & 0 \\ \vdots & & ⋰ & \vdots & \vdots \\ 0 & b_1 & \cdots & \sum_{t=1}^{T-2} b_t & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}$$

$$A_3(\theta I+\gamma F)^{-1}PE_4 = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sum_{t=1}^{T} b_t \end{pmatrix}$$

To show that $A(\theta I+\gamma F)^{-1}PE^\top$ is indeed symmetric, it only remains to show that $[(\lambda+2\theta+4\gamma)A_1-(\theta+2\gamma)A_2](\theta I+\gamma F)^{-1}PE_4$ 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 $[(\lambda+2\theta+4\gamma)A_1-(\theta+2\gamma)A_2](\theta I+\gamma F)^{-1}PE_4$ 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:

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

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

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

which implies

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

Letting $M_1=(\theta I+\gamma F)A^{-1}E$, we can rewrite this as $M_1P=PM_1^\top$. Because $P^{-1}=P$, we also have $PM_1=M_1^\top P$. Together, these imply $(M_1+M_1^\top)P=P(M_1+M_1^\top)$. 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 $j=1,\dots,T$. □

A.9 Proof of Proposition 9

Proof. In the following analysis, we verify the statements about the limit of the optimal contract and trading strategy made earlier in this section. Throughout, we assume $\theta>0$ since the results for $\theta=0$ follow from Corollary 6. 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 λ is replaced by $\frac{\lambda T\sigma^2}{T_k}$. Therefore, (11) becomes

$$X_q^k+\frac{\theta+2\gamma}{\lambda T\sigma^2}\frac{2X_q^k-X_{q+\frac{1}{T_k}}^k-X_{q-\frac{1}{T_k}}^k}{1/T_k}=\frac{\gamma}{\lambda T\sigma^2}\frac{2V_q^k-V_{q+\frac{1}{T_k}}^k-V_{q-\frac{1}{T_k}}^k}{1/T_k}+\frac{\theta}{\lambda T\sigma^2}\frac{V_q^k-V_{q-\frac{1}{T_k}}^k}{1/T_k}+V_q^k \quad (14)$$

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 $q \in (0,1)$,

$$\lim_{k\to\infty} \frac{V_q^k - V_{q-\frac{1}{T_k}}^k}{1/T_k} = \lim_{\epsilon\to 0} \frac{V_q - V_{q-\epsilon}}{\epsilon} = \dot{V}_q$$

$$\lim_{k\to\infty} \frac{2V_q^k - V_{q+\frac{1}{T_k}}^k - V_{q-\frac{1}{T_k}}^k}{1/T_k} = \lim_{\epsilon\to 0} \left(\frac{V_q - V_{q-\epsilon}}{\epsilon} - \frac{V_{q+\epsilon} - V_q}{\epsilon} \right) = \dot{V}_q - \dot{V}_q = 0,$$

$$\lim_{k\to\infty} \frac{2X_q^k - X_{q+\frac{1}{T_k}}^k - X_{q-\frac{1}{T_k}}^k}{1/T_k} = \lim_{\epsilon\to 0} \frac{-(X_{q+\epsilon} - X_q) + (X_q - X_{q-\epsilon})}{\epsilon} = -\dot{X}_q + \dot{X}_q = 0.$$

Thus, it follows from (14) that

$$X_q = \frac{\theta}{\lambda T \sigma^2} \dot{V}_q + V_q \tag{15}$$

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

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

$$V_1 - V_{1-} = V_1 - \lim_{q \nearrow 1} V_q = \lim_{k\to\infty} \sum_{t=1}^{S_k} \tau_{T_k-(t-1)}^k = \lim_{k\to\infty} \sum_{t=1}^{S_k} \left(V_{1-\frac{t-1}{T_k}}^k - V_{1-\frac{t}{T_k}}^k \right),$$

where S_k is a sequence chosen so that both $S_k \to \infty$ and $S_k^3/T_k \to 0$. From (14), it follows that

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

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

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

$$x_1^k - x_2^k = \frac{\gamma}{\theta + 2\gamma} \left(\tau_1^k - \tau_2^k \right) + \frac{\theta}{\theta + 2\gamma} \tau_1^k + O\left(\frac{1}{T_k}\right),$$

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

Assuming $x^k_{S_k} = O\big(\frac{1}{T_k}\big)$ and $\tau^k_{S_k} = O\big(\frac{1}{T_k}\big)$, we obtain

$$\begin{aligned} x^k_t &= x^k_{t+1} + \frac{\gamma}{\theta + 2\gamma}\big(\tau^k_t - \tau^k_{t+1}\big) + \frac{\theta}{\theta+2\gamma}\tau^k_t + O\bigg(\frac{1}{T_k}\bigg) \\ &= \frac{\gamma}{\theta+2\gamma}\sum_{j=t}^{S_k}\big(\tau^k_j - \tau^k_{j+1}\big) + \frac{\theta}{\theta+2\gamma}\sum_{j=t}^{S_k}\tau^k_j + O\bigg(\frac{S_k}{T_k}\bigg) \\ &= \frac{\gamma}{\theta+2\gamma}\tau^k_t + \frac{\theta}{\theta+2\gamma}\sum_{j=t}^{S_k}\tau^k_j + O\bigg(\frac{S_k}{T_k}\bigg), \end{aligned} \tag{16}$$

and likewise if $x^k_{T_k-S_k} = O\big(\frac{1}{T_k}\big)$ and $\tau^k_{T_k-S_k} = O\big(\frac{1}{T_k}\big)$,

$$\begin{aligned} x^k_{T_k-t} &= x^k_{T_k-(t+1)} - \frac{\gamma}{\theta+2\gamma}\big(\tau^k_{T_k-(t+1)} - \tau^k_{T_k-t}\big) - \frac{\theta}{\theta+2\gamma}\tau^k_{T_k-(t+1)} + O\bigg(\frac{1}{T_k}\bigg) \\ &= -\frac{\gamma}{\theta+2\gamma}\sum_{j=t}^{S_k}\big(\tau^k_{T_k-(j+1)} - \tau^k_{T_k-j}\big) - \frac{\theta}{\theta+2\gamma}\sum_{j=t+1}^{S_k}\tau^k_{T_k-j} + O\bigg(\frac{S_k}{T_k}\bigg) \\ &= \frac{\gamma}{\theta+2\gamma}\tau^k_{T_k-t} - \frac{\theta}{\theta+2\gamma}\sum_{j=t+1}^{S_k}\tau^k_{T_k-j} + O\bigg(\frac{S_k}{T_k}\bigg). \end{aligned} \tag{17}$$

Therefore, the expected costs for the client are

$$\begin{aligned} \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{\tau}^k \cdot \boldsymbol{p}] &= p_0 + \theta\sum_{t=1}^{T_k} X^k_{\frac{t}{T_k}}\Big(V^k_{\frac{t}{T_k}} - V^k_{\frac{t-1}{T_k}}\Big) + \gamma\sum_{t=1}^{T_k}\Big(X^k_{\frac{t}{T_k}} - X^k_{\frac{t-1}{T_k}}\Big)\Big(V^k_{\frac{t}{T_k}} - V^k_{\frac{t-1}{T_k}}\Big) \\ &= p_0 + \theta\int_0^1 X_q\,dV_q + \theta\sum_{t=1}^{S_k} X^k_{\frac{t}{T_k}}\tau^k_t + \gamma\sum_{t=1}^{S_k} x^k_t\tau^k_t + \theta\sum_{t=1}^{S_k} X^k_{1-\frac{t-1}{T_k}}\tau^k_{T_k-(t-1)} \\ &\quad + \gamma\sum_{t=1}^{S_k} x^k_{T_k-(t-1)}\tau^k_{T_k-(t-1)} + O\bigg(\frac{1}{T_k}\bigg). \end{aligned} \tag{18}$$

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 (15) to write

$$\int_0^1 X_q\,dV_q = \int_0^1\bigg(\frac{\theta}{\lambda T\sigma^2}\dot V_q + V_q\bigg)\dot V_q\,dq = \frac{\theta}{\lambda T\sigma^2}\int_0^1 \dot V_q^2\,dq + \frac{1}{2}V_{1-}^2 - \frac{1}{2}V_{0+}^2, \tag{19}$$

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 8, 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, (19) in the optimum becomes

$$\int_0^1 X_q \, dV_q = \frac{\theta}{\lambda T \sigma^2} \int_0^1 \dot{V}_q^2 \, dq + \frac{1}{2} V_{1-}^2 - \frac{1}{2} V_{0+}^2 = \frac{\theta^2}{\lambda T \sigma^2} a^2 + \frac{\theta(1+a)^2}{8} - \frac{\theta(1-a)^2}{8}. \tag{20}$$

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

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

Next, we analyze the minimization of

$$\theta \sum_{t=1}^{S_k} X_{\frac{t}{T_k}}^k \tau_t^k + \gamma \sum_{t=1}^{S_k} x_t^k \tau_t^k. \tag{21}$$

subject to $\sum_{t=1}^{S_k} \tau_t^k = \frac{1-a}{2}$. Using (16), we write

$$\begin{aligned}
&\theta \sum_{t=1}^{S_k} X_{\frac{t}{T_k}}^k \tau_t^k + \gamma \sum_{t=1}^{S_k} x_t^k \tau_t^k \\
&= \theta \sum_{t=1}^{S_k} \sum_{\ell=1}^{t} \left(\frac{\gamma}{\theta + 2\gamma} \tau_\ell^k + \frac{\theta}{\theta + 2\gamma} \sum_{j=\ell}^{S_k} \tau_j^k \right) \tau_t^k + \gamma \sum_{t=1}^{S_k} \left(\frac{\gamma}{\theta + 2\gamma} \tau_t^k + \frac{\theta}{\theta + 2\gamma} \sum_{j=t}^{S_k} \tau_j^k \right) \tau_t^k + O\left(\frac{S_k^3}{T_k}\right) \\
&= \theta \sum_{t=1}^{S_k} \left(\frac{\gamma}{\theta + 2\gamma} \sum_{\ell=1}^{t} \tau_\ell^k + \frac{\theta}{\theta + 2\gamma} \sum_{j=1}^{S_k} \min\{j, t\} \tau_j^k \right) \tau_t^k + \gamma \sum_{t=1}^{S_k} \left(\frac{\gamma}{\theta + 2\gamma} \tau_t^k + \frac{\theta}{\theta + 2\gamma} \sum_{j=t}^{S_k} \tau_j^k \right) \tau_t^k + O\left(\frac{S_k^3}{T_k}\right).
\end{aligned}$$

We can simplify two terms

$$\theta \sum_{t=1}^{S_k} \frac{\gamma}{\theta+2\gamma} \sum_{\ell=1}^{t} \tau_\ell^k \tau_t^k + \gamma \sum_{t=1}^{S_k} \frac{\theta}{\theta+2\gamma} \sum_{j=t}^{S_k} \tau_j^k \tau_t^k = \frac{\theta\gamma}{\theta+2\gamma} \sum_{t=1}^{S_k} \sum_{\ell=1}^{S_k} \tau_\ell^k \tau_t^k + \frac{\theta\gamma}{\theta+2\gamma} \sum_{t=1}^{S_k} \left(\tau_t^k\right)^2$$
$$= \frac{\theta\gamma}{\theta+2\gamma} \sum_{t=1}^{S_k} \frac{1-a}{2} \tau_t^k + \frac{\theta\gamma}{\theta+2\gamma} \sum_{t=1}^{S_k} \left(\tau_t^k\right)^2$$
$$= \frac{\theta\gamma}{\theta+2\gamma} \frac{(1-a)^2}{4} + \frac{\theta\gamma}{\theta+2\gamma} \sum_{t=1}^{S_k} \left(\tau_t^k\right)^2.$$

Therefore, the optimization problem (21) becomes

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

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

$$\begin{aligned} 0 &= \frac{\partial}{\partial \tau_s^k} \left(\theta^2 \sum_{t=1}^{S_k} \tau_t^k \sum_{j=1}^{S_k} \min\{j,t\} \tau_j^k + (\theta\gamma+\gamma^2) \sum_{t=1}^{S_k} \left(\tau_t^k\right)^2 \right) + \lambda_1 \\ &= \theta^2 \sum_{j=1}^{S_k} \min\{j,s\} \tau_j^k + \theta^2 \sum_{t=1}^{S_k} \tau_t^k \min\{s,t\} + 2(\theta\gamma+\gamma^2) \tau_s^k + \lambda_1 \\ &= 2\theta^2 \sum_{j=1}^{S_k} \min\{j,s\} \tau_j^k + 2(\theta\gamma+\gamma^2) \tau_s^k + \lambda_1 \end{aligned} \tag{22}$$

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

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

Its solution is

$$\tau_j^k = \frac{\theta\gamma^{j-1}}{(\theta+\gamma)^j}\frac{1-a}{2}, \quad j = 1, 2, \ldots, \tag{23}$$

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 (16), (17), and (23) that

$$\begin{aligned}
x_t^k &= \frac{\gamma}{\theta+2\gamma}\tau_t^k + \frac{\theta}{\theta+2\gamma}\sum_{j=t}^{S_k}\tau_j^k + O\left(\frac{S_k}{T_k}\right) \\
&= \frac{\gamma}{\theta+2\gamma}\frac{\theta\gamma^{t-1}}{(\theta+\gamma)^t}\frac{1-a}{2} + \frac{\theta}{\theta+2\gamma}\sum_{j=t}^{S_k}\frac{\theta\gamma^{j-1}}{(\theta+\gamma)^j}\frac{1-a}{2} + O\left(\frac{S_k}{T_k}\right) \\
&= \frac{\gamma}{\theta+2\gamma}\frac{\theta\gamma^{t-1}}{(\theta+\gamma)^t}\frac{1-a}{2} + \frac{\theta}{\theta+2\gamma}\frac{\theta\gamma^{t-1}}{(\theta+\gamma)^t}\frac{1}{1-\frac{\gamma}{\theta+\gamma}}\frac{1-a}{2} + O\left(\frac{S_k}{T_k}\right) \\
&= \frac{\theta\gamma^{t-1}}{(\theta+\gamma)^t}\frac{1-a}{2} + O\left(\frac{S_k}{T_k}\right), \\
x_{T_k-t}^k &= \frac{\gamma}{\theta+2\gamma}\tau_{T_k-t}^k - \frac{\theta}{\theta+2\gamma}\sum_{j=t+1}^{S_k}\tau_{T_k-j}^k + O\left(\frac{S_k}{T_k}\right) \\
&= \frac{\gamma}{\theta+2\gamma}\tau_{t+1}^k - \frac{\theta}{\theta+2\gamma}\sum_{j=t+2}^{S_k}\tau_j^k + O\left(\frac{S_k}{T_k}\right) \\
&= \frac{\gamma}{\theta+2\gamma}\frac{\theta\gamma^t}{(\theta+\gamma)^{t+1}}\frac{1-a}{2} - \frac{\theta}{\theta+2\gamma}\sum_{j=t+2}^{S_k}\frac{\theta\gamma^{j-1}}{(\theta+\gamma)^j}\frac{1-a}{2} + O\left(\frac{S_k}{T_k}\right) \\
&= \frac{\gamma}{\theta+2\gamma}\frac{\theta\gamma^t}{(\theta+\gamma)^{t+1}}\frac{1-a}{2} - \frac{\theta}{\theta+2\gamma}\frac{\theta\gamma^{t+1}}{(\theta+\gamma)^{t+2}}\frac{1}{1-\frac{\gamma}{\theta+\gamma}}\frac{1-a}{2} + O\left(\frac{S_k}{T_k}\right) \\
&= O\left(\frac{S_k}{T_k}\right).
\end{aligned} \tag{24}$$

For $s = 1$, (22) 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 (22) 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

$$\begin{aligned}
&\frac{\theta\gamma}{\theta + 2\gamma}\frac{(1-a)^2}{4} + \frac{\theta^2}{\theta + 2\gamma}\sum_{t=1}^{S_k} \tau_t^k \sum_{j=1}^{S_k} \min\{j,t\}\tau_j^k + \frac{\theta\gamma + \gamma^2}{\theta + 2\gamma}\sum_{t=1}^{S_k} \left(\tau_t^k\right)^2 \\
&= \frac{\theta\gamma}{\theta + 2\gamma}\frac{(1-a)^2}{4} - \frac{\theta\gamma + \gamma^2}{\theta + 2\gamma}\sum_{t=1}^{S_k} \left(\tau_t^k\right)^2 + \frac{1}{\theta + 2\gamma}\sum_{t=1}^{S_k} \tau_t^k(\theta^2 + \gamma\theta)\frac{1-a}{2} + \frac{\theta\gamma + \gamma^2}{\theta + 2\gamma}\sum_{t=1}^{S_k} \left(\tau_t^k\right)^2 \\
&= \frac{\theta\gamma}{\theta + 2\gamma}\frac{(1-a)^2}{4} + \frac{1}{\theta + 2\gamma}\sum_{t=1}^{S_k} \tau_t^k(\theta^2 + \gamma\theta)\frac{1-a}{2} \\
&= \frac{\theta\gamma}{\theta + 2\gamma}\frac{(1-a)^2}{4} + \frac{\theta^2 + \gamma\theta}{\theta + 2\gamma}\frac{(1-a)^2}{4} \\
&= \theta\frac{(1-a)^2}{4}.
\end{aligned} \tag{25}$$

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 (24), the jumps of the trading strategy in the limit disappear so that $x^k_{T_k-(t-1)} = O\big(\frac{S_k}{T_k}\big)$ and $X^k_{1-\frac{t-1}{T_k}} = 1 + O\big(\frac{S_k^2}{T_k}\big)$. Therefore, the value of the minimization problem becomes

$$\theta \sum_{t=1}^{S_k} X^k_{1-\frac{t-1}{T_k}} \tau^k_{T_k-(t-1)} + \gamma \sum_{t=1}^{S_k} x^k_{T_k-(t-1)} \tau^k_{T_k-(t-1)} = \theta \sum_{t=1}^{S_k} \tau^k_{T_k-(t-1)} + O\left(\frac{S_k^3}{T_k}\right) = \theta\frac{1-a}{2} + O\left(\frac{S_k^3}{T_k}\right). \tag{26}$$

In summary, using (20), (25), and (26), the expected costs from (18) in the optimum are

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{\tau}^k \cdot \boldsymbol{p}] &= p_0 + \theta \int_0^1 X_q \, dV_q + \theta \sum_{t=1}^{S_k} X^k_{\frac{t}{T_k}} \tau^k_t + \gamma \sum_{t=1}^{S_k} x^k_t \tau^k_t + \theta \sum_{t=1}^{S_k} X^k_{1-\frac{t-1}{T_k}} \tau^k_{T_k-(t-1)} \\
&\quad + \gamma \sum_{t=1}^{S_k} x^k_{T_k-(t-1)} \tau^k_{T_k-(t-1)} + O\left(\frac{S_k^3}{T_k}\right) \\
&= p_0 + \frac{\theta^2}{\lambda T \sigma^2} a^2 + \frac{\theta(1+a)^2}{8} - \frac{\theta(1-a)^2}{8} + \theta \frac{(1-a)^2}{4} + \theta \frac{1-a}{2} + O\left(\frac{S_k^3}{T_k}\right) \\
&= p_0 + \frac{\theta^2}{\lambda T \sigma^2} a^2 + \theta \frac{(1-a)^2}{4} + \frac{\theta}{2} + O\left(\frac{S_k^3}{T_k}\right),
\end{aligned} \tag{27}$$

minimized over a. The first-order condition gives

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

so that

$$a = \frac{\frac{\theta}{2}}{\frac{\theta}{2} + \frac{2\theta^2}{\lambda T \sigma^2}} = \frac{1}{1 + \frac{4\theta}{\lambda T \sigma^2}}. \tag{28}$$

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 (15), which implies the formulas for the limits of the optimal contract and dealer's trading strategy. Thanks to (27) and (28), the client's expected costs converge to

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

A.10 Proof of Proposition 10

Proof. This proof builds on the first half of the proof of Proposition 9.

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 (15) 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 (18) that the expected costs for the client under a TWAP contract are

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

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 (15) 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

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

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 (17). □

A.11 Decomposition of the wedge between the first-best and second-best

As observed in the text, the wedge between the first-best and second-best payments (in the continuous-time limit) is $\frac{1-a}{4}\theta$, where $a = \frac{1}{1+\frac{4\theta}{\lambda T\sigma^2}}$. Footnote 25 claims that this wedge can be decomposed in the following way:

$$\underbrace{\frac{1-a}{4}\theta}_{\substack{\text{wedge between first-best} \\ \text{and second-best payments}}} = \underbrace{\frac{1-a}{2}\theta - \frac{(1-a)^2}{4}\frac{\theta^2}{\theta+2\gamma}}_{\text{dealer's expected profit}} + \underbrace{\frac{(1-a)^2}{4}\frac{\theta^2}{\theta+2\gamma} - \frac{1-a}{4}\theta}_{\text{inefficiency from suboptimal trading}}.$$

Formally, this decomposition follows from the subsequent result. As an aside, the presence of γ in this decomposition highlights that—although it does not affect the client's execution costs when the contract is chosen optimally—temporary price impact does not vanish and does play a role in the continuous-time limit.

Proposition 12. *Assume the dealer is strictly risk-averse ($\lambda > 0$). 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 $\boldsymbol{\tau}^{*k} \in \mathcal{T}^{T_k}$ be the associated optimal contract, and let $\boldsymbol{x}^{*k}$ be the dealer strategy that best responds to $\boldsymbol{\tau}^{*k}$. The dealer's expected profit under $\boldsymbol{\tau}^{*k}$ and $\boldsymbol{x}^{*k}$ converges to*

$$\frac{1-a}{2}\theta - \frac{(1-a)^2}{4}\frac{\theta^2}{\theta + 2\gamma}, \tag{29}$$

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

Proof. To simplify notation throughout this proof, we write simply $\boldsymbol{\tau}^k$ and $\boldsymbol{x}^k$ instead of $\boldsymbol{\tau}^{*k}$ and $\boldsymbol{x}^{*k}$. The dealer's expected profit is given by

$$\mathbb{E}_{\boldsymbol{x}^k}\left[\boldsymbol{\tau}^k \cdot \boldsymbol{p}^k - \boldsymbol{x}^k \cdot \boldsymbol{p}^k\right] = \sum_{t=1}^{T_k}\left(\tau_t^k - x_t^k\right)\left(p_0 + \theta X^k_{\frac{t}{T_k}} + \gamma x_t^k\right) = \sum_{t=1}^{T_k}\left(\tau_t^k - x_t^k\right)\left(\theta X^k_{\frac{t}{T_k}} + \gamma x_t^k\right),$$

using for the second equality that $\sum_{t=1}^{T_k}\tau_t^k = \sum_{t=1}^{T_k} x_t^k = 1$. Arguing similarly to (18), this expression converges to

$$\lim_{k\to\infty}\sum_{t=1}^{T_k}\left(\tau_t^k - x_t^k\right)\left(\theta X^k_{\frac{t}{T_k}} + \gamma x_t^k\right) = \theta\int_0^1 X_q\, d(V_q - X_q) + \lim_{k\to\infty}\sum_{t=1}^{S_k}\left(\tau_t^k - x_t^k\right)\left(\theta X^k_{\frac{t}{T_k}} + \gamma x_t^k\right) \tag{30}$$
$$+ \lim_{k\to\infty}\sum_{t=1}^{S_k}\left(\tau^k_{T_k-(t-1)} - x^k_{T_k-(t-1)}\right)\left(\theta X^k_{1-\frac{t-1}{T_k}} + \gamma x^k_{T_k-(t-1)}\right).$$

We analyze the terms on the right-hand side of (30). The first term can be written as

$$\begin{aligned}
\theta\int_0^1 X_q\, d(V_q - X_q) &= \theta a\int_0^1 X_q\, dq - \frac{\theta}{2}X_{1-}^2 + \frac{\theta}{2}\lim_{k\to\infty}\left(X_{0+} - \sum_{t=1}^{S_k} x^k_{T_k-(t-1)}\right)^2 \\
&= \theta a\int_0^1\left(\frac{3(1-a)}{4} + aq\right)dq - \frac{\theta}{2}\left(\frac{3+a}{4}\right)^2 + \frac{\theta}{2}\left(\frac{3(1-a)}{4} - \frac{1-a}{2}\right)^2 \\
&= \theta a\frac{3(1-a)}{4} + \theta a^2\frac{1}{2} - \frac{\theta}{2}\left(\frac{3+a}{4}\right)^2 + \frac{\theta}{2}\left(\frac{1-a}{4}\right)^2 \\
&= \frac{\theta}{4}(3a - a^2 - 1 - a) \\
&= -\frac{\theta}{4}(1-a)^2.
\end{aligned} \tag{31}$$

Using (5) and (25), the second term on the right-hand side of (30) can be computed as

$$\begin{aligned}
&\lim_{k\to\infty}\sum_{t=1}^{S_k}\left(\tau_t^k - x_t^k\right)\left(\theta X^k_{\frac{t}{T_k}} + \gamma x_t^k\right)\\
&= \lim_{k\to\infty}\left(\theta\sum_{t=1}^{S_k} X^{*,k}_{\frac{t}{T_k}}\tau_t^{*,k} + \gamma\sum_{t=1}^{S_k} x_t^{*,k}\tau_t^{*,k} - \theta\sum_{t=1}^{S_k} X^{*,k}_{\frac{t}{T_k}} x_t^{*,k} - \gamma\sum_{t=1}^{S_k}\left(x_t^{*,k}\right)^2\right)\\
&= \theta\frac{(1-a)^2}{4} - \theta\sum_{j=1}^{\infty}\sum_{\ell=1}^{j}\frac{\theta\gamma^{\ell-1}}{(\theta+\gamma)^\ell}\frac{1-a}{2}\frac{\theta\gamma^{j-1}}{(\theta+\gamma)^j}\frac{1-a}{2} - \gamma\sum_{j=1}^{\infty}\left(\frac{\theta\gamma^{j-1}}{(\theta+\gamma)^j}\frac{1-a}{2}\right)^2\\
&= \frac{(1-a)^2}{4}\left(\theta - \frac{\theta^3}{\gamma^2}\frac{\gamma}{\theta+\gamma}\sum_{j=1}^{\infty}\frac{1-\left(\frac{\gamma}{\theta+\gamma}\right)^j}{1-\frac{\gamma}{\theta+\gamma}}\frac{\gamma^j}{(\theta+\gamma)^j} - \frac{\theta^2}{\gamma}\sum_{j=1}^{\infty}\left(\frac{\gamma}{\theta+\gamma}\right)^{2j}\right)\\
&= \frac{(1-a)^2}{4}\left(\theta - \frac{\theta^3}{\gamma^2}\frac{\frac{\gamma}{\theta+\gamma}}{1-\frac{\gamma}{\theta+\gamma}}\left(\frac{\frac{\gamma}{\theta+\gamma}}{1-\frac{\gamma}{\theta+\gamma}} - \left(\frac{\gamma}{\theta+\gamma}\right)^2\frac{1}{1-\left(\frac{\gamma}{\theta+\gamma}\right)^2}\right) - \frac{\theta^2}{\gamma}\left(\frac{\gamma}{\theta+\gamma}\right)^2\frac{1}{1-\left(\frac{\gamma}{\theta+\gamma}\right)^2}\right)\\
&= \frac{(1-a)^2}{4}\left(\theta - \frac{\theta^3}{\gamma^2}\frac{\left(\frac{\gamma}{\theta+\gamma}\right)^2}{1-\frac{\gamma}{\theta+\gamma}}\frac{1}{1-\left(\frac{\gamma}{\theta+\gamma}\right)^2} - \frac{\theta^2}{\gamma}\left(\frac{\gamma}{\theta+\gamma}\right)^2\frac{1}{1-\left(\frac{\gamma}{\theta+\gamma}\right)^2}\right)\\
&= \frac{(1-a)^2}{4}\left(\theta - \frac{\theta^2}{\theta+2\gamma}\right). \qquad (32)
\end{aligned}$$

Because $\boldsymbol{x}^k$ has no persistent jumps at the end time by (24), the third term on the right-hand side of (30) becomes

$$\lim_{k\to\infty}\sum_{t=1}^{S_k}\tau^k_{T_k-(t-1)}\left(\theta X^k_{1-\frac{t-1}{T_k}} + \gamma x^k_{T_k-(t-1)}\right) = \theta\frac{1-a}{2} \qquad (33)$$

by (26). Substituting (31)–(33) into (30), we obtain that the dealer's expected profit converges to (29). □

Online Appendix to

"Principal Trading Arrangements: Optimality under Temporary and Permanent Price Impact"

Markus Baldauf Christoph Frei Joshua Mollner

OA.A Additional Special Cases

With various sections of the main text, we have already considered several special cases of the model. This appendix considers four more special cases. Appendix OA.A.1 considers the case of two trading periods (i.e., $T = 2$), and provides further intuition for the optimality of symmetric contract weights. Appendix OA.A.2 considers the case of no temporary price impact (i.e., $\gamma = 0$), demonstrating that the optimal contract in this case is a U-shape with a 'flat bottom,' in that all interior weights are the same. Finally, Appendix OA.A.3 considers the limiting case of an infinite execution horizon (i.e., $T \to \infty$), demonstrating that the client obtains her first-best payoff in this limit.

OA.A.1 Two trading periods

To provide additional intuition for the optimal contract's symmetry, this appendix considers the version of the model with $T = 2$ trading periods. Our approach is to provide intuition by connecting our setup to a rudimentary microeconomics-based calculation.

Monopolist with linear demand. Consider the problem of a monopolist facing a linear inverse demand curve $P = a - bQ$ and a marginal cost c. Following the standard derivation, the monopolist optimally sets

$$Q^* = \frac{a}{2b} - \frac{c}{2b} \qquad \text{and} \qquad P^* = \frac{a}{2} - \frac{c}{2}.$$

Let us use Q_0^* and P_0^* to denote these optimal values in the special case of $c = 0$. In particular, it holds that

$$Q_0^* = \frac{a}{b}\frac{dP^*}{dc}.$$

As we demonstrate below, this relationship is a consequence of the linear nature of demand; $\frac{a}{b}$ is the appropriate scaling factor, which can be interpreted as the quantity demanded at a price of zero.

In fact, an analogous relationship holds even more generally. Suppose the monopolist faces an additional cost that is quadratic in $P - c$, so that her objective is $(P - c)Q - \frac{\lambda}{2}(P - c)^2$, which can be written

$$(a - bQ - c)Q - \frac{\lambda}{2}(a - bQ - c)^2.$$

In the case of $\lambda = 0$, this reduces to the classic problem mentioned above. We obtain

$$Q^* = \frac{a-c}{b} \underbrace{\frac{1+b\lambda}{2+b\lambda}}_{\equiv \phi}. \tag{34}$$

The interpretation of the first factor in (34) is the quantity demanded at marginal-cost pricing. The precise form of the second factor does not matter for the subsequent arguments, so we simply label it ϕ; it can be interpreted as the monopolist's optimal amount of quantity reduction. That this is a constant is due to the linear nature of demand. It follows from (34) that $Q_0^* = \frac{a}{b}\phi$. The intuition is that Q_0^* must equal the quantity demanded at a price of zero (which, as mentioned is $\frac{a}{b}$) times quantity reduction from marginal-cost pricing (which we have denoted ϕ). It also follows from (34) that

$$P^* = a - bQ^* = a(1-\phi) + \phi c.$$

That P^* should be this weighted average of a and c is intuitive because $\phi = 1$ corresponds to no quantity reduction (i.e., marginal-cost pricing) and $\phi = 0$ corresponds to full quantity reduction (i.e., pricing so that zero quantity is demanded). Hence, $\frac{dP^*}{dc} = \phi$. We conclude that

$$Q_0^* = \frac{a}{b}\phi = \frac{a}{b}\frac{dP^*}{dc},$$

as claimed.

The dealer's problem. To see the connection to our setting, suppose that $T = 2$, and parametrize

$$(x_1, x_2) = \left(\frac{1}{2} + x, \frac{1}{2} - x\right) \qquad \text{and} \qquad (\tau_1, \tau_2) = \left(\frac{1}{2} + \tau, \frac{1}{2} - \tau\right).$$

Given any τ, the dealer's problem is to choose x to maximize

$$(x - \tau)\big(\mathbb{E}[p_2] - \mathbb{E}[p_1]\big) - \frac{\lambda\sigma^2}{2}(\tau - x)^2.$$

Because $\mathbb{E}[p_2] - \mathbb{E}[p_1] = \frac{\theta}{2} - (\theta + 2\gamma)x$, this objective is of the same form that was analyzed above, with the roles of (P, Q, c, a, b, λ) now played by $\big(x, \mathbb{E}[p_2] - \mathbb{E}[p_1], \tau, \frac{\theta}{2\theta+4\gamma}, \frac{1}{\theta+2\gamma}, \lambda\sigma^2\big)$. We conclude that when $\tau = 0$, it will hold that $\mathbb{E}[p_2] - \mathbb{E}[p_1] = \frac{\theta}{2}\frac{dx^*}{d\tau}$.

The client's problem. The client's problem is to choose τ to minimize

$$\left(\frac{1}{2} + \tau\right)\mathbb{E}[p_1] + \left(\frac{1}{2} - \tau\right)\mathbb{E}[p_2],$$

taking into account how τ influences x^*, and hence prices. Suppose we begin at $\tau = 0$ and perturb τ upward. This produces two effects:

- *The direct effect.* The direct effect of this perturbation is to increase the objective at the rate

$$\mathbb{E}[p_1] - \mathbb{E}[p_2],$$

which is equal to $-\frac{\theta}{2}\frac{dx^*}{d\tau}$ by the above observation. Intuitively, permanent price impact leads us to have $\mathbb{E}[p_2] > \mathbb{E}[p_1]$ (given the dealer's best response to $\tau = 0$), and so the direct effect of increasing τ is to put more weight on the lower price, which is beneficial to the client.

- *The indirect effect.* The indirect effect of this perturbation (through its effect on expected prices) is to increase the objective at the rate

$$\frac{1}{2}\left(\underbrace{\frac{d\mathbb{E}[p_1]}{dx}}_{=\gamma+\theta} + \underbrace{\frac{d\mathbb{E}[p_2]}{dx}}_{=-\gamma}\right)\frac{dx^*}{d\tau}.$$

Intuitively, this increase in τ leads the dealer to trade more in the first period, which—because of permanent price impact—leads p_1 to increase more than p_2 decreases. At $\tau = 0$, the client's payment is an equally-weighted average of these two prices, so the indirect effect of this perturbation is harmful to the client.

By the arguments above, these two effects cancel out, so it follows that it is optimal for the client to select $\tau = 0$.

OA.A.2 No temporary price impact

This appendix considers the case in which price impact has no temporary component (i.e., is purely permanent). We begin by considering the optimal contract in discrete time, demonstrating that it is a U-shape with a 'flat bottom,' in that all interior weights are the same. We then consider the continuous-time limit, producing a more rigorous proof of the result of Proposition 9 for the case without temporary price impact.

OA.A.2.1 The optimal contract in discrete time

Without temporary component (i.e., when $\gamma = 0$), we obtain the following as a corollary of the general solution given by Proposition 4.

Corollary 13. *Assume that there is no temporary price impact ($\gamma = 0$). Then the weights of the optimal contract satisfy $\tau_i^* = \tau_j^*$ for all $i, j \in \{2, 3, \ldots, T-1\}$. Moreover, the solution depends on $(\lambda, \sigma, \theta)$ only through $\frac{\lambda\sigma^2}{\theta}$.*

Corollary 13 says that in this case of only permanent price impact, the optimal contract puts the same weight on all interior periods. From Proposition 4, we know that the extremal elements of $\boldsymbol{\tau}^*$ are also equal, but they differ from the interior weights, which can also be seen in the left panel of Figure 2 for the curve $\gamma = 0.0$. From Figure 6, we observe that the extremal weights of the optimal contract are greater than $1/T$ and decreasing in $\alpha \equiv \frac{\lambda\sigma^2}{\theta}$ so that the dominance of the extremal weights over the inner weights is stronger for small values of α. In fact, it follows from Corollaries 5 and 7 that the extremal weights converge to $1/2$ for $\alpha \to 0$ and to $1/T$ for $\alpha \to \infty$.

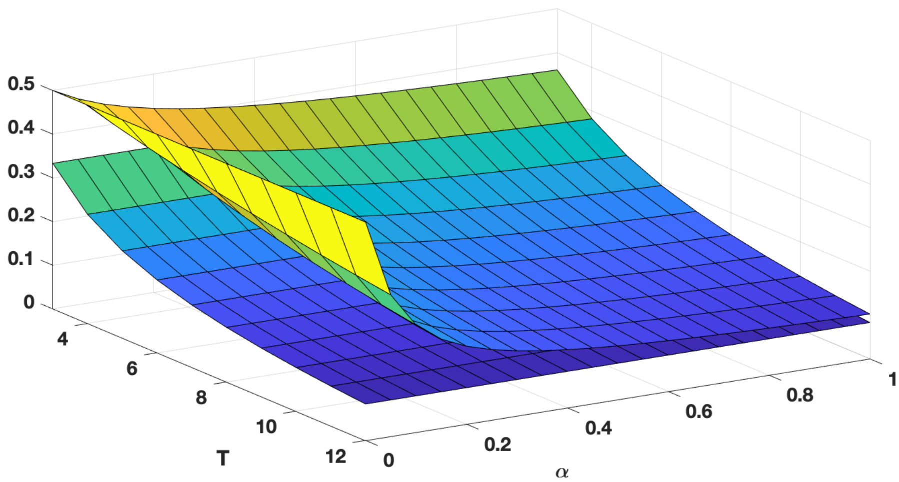

Figure 6: The upper surface shows the extremal weights of the optimal contract as a function of T and $\alpha = \frac{\lambda\sigma^2}{\theta}$ when there is no temporary price impact. As a visual comparison, the lower surface represents $1/T$. The extremal weights are greater than $1/T$ and decreasing in α.

As mentioned, numerical experimentation suggests that the dealer's best response to the optimal contract is frontloaded relative to the contract itself. In this case of no temporary price impact, we can formally prove a limited version of this general conjecture: that the first component of the dealer's trading strategy dominates the first weight of the optimal contract. Indeed, without temporary price impact, the dealer's best response $\boldsymbol{x}$ to *any* symmetric contract with equal inner weights; that is, of the form $\left(\frac{1-a(T-2)}{2}, a, \ldots, a, \frac{1-a(T-2)}{2}\right)$ for $0 \leq a \leq \frac{1}{T-2}$ and with $T \geq 3$ satisfies

$$x_1 > \frac{1-a(T-2)}{2}, \tag{35}$$

as we show in Appendix OA.A.2.3.

Proof of Corollary 13. We start by writing

$$\begin{aligned} M^{-1} &= \left(\theta A^{-1}E + \theta E^\top (A^{-1})^\top\right)^{-1} = \frac{1}{\theta}\left(A^{-1}E + E^\top (A^{-1})^\top\right)^{-1} = \frac{1}{\theta}\left(A^{-1}\left(E + AE^\top (A^{-1})^\top\right)\right)^{-1} \\ &= \frac{1}{\theta}\left(E + AE^\top (A^{-1})^\top\right)^{-1} A = \frac{1}{\theta}\left((EA^\top + AE^\top)(A^{-1})^\top\right)^{-1} A \\ &= \frac{1}{\theta} A^\top \left(EA^\top + AE^\top\right)^{-1} A = \frac{1}{\theta} A^\top \left(EA^\top + AE^\top\right)^{-1} A \end{aligned} \tag{36}$$

and

$$EA^\top + AE^\top = B + \boldsymbol{v}\boldsymbol{v}^\top \tag{37}$$

where $\boldsymbol{v} = \lambda(1, 1, \ldots, 1)^\top$ and

$$B = \begin{pmatrix} 4\theta^2 + \lambda^2 + 2\lambda\theta & -2\theta^2 & \lambda\theta & \cdots & -\lambda^2 + \lambda + \theta \\ -2\theta^2 & 4\theta^2 + \lambda^2 & -2\theta^2 - \theta\lambda & 0 & -\lambda^2 + \lambda \\ \lambda\theta & -2\theta^2 - \theta\lambda & 4\theta^2 + \lambda^2 & -2\theta^2 - \theta\lambda & -\lambda^2 + \lambda \\ \lambda\theta & 0 & -2\theta^2 - \theta\lambda & 4\theta^2 + \lambda^2 & \vdots \\ \vdots & \vdots & \ddots & \ddots & -\lambda^2 + \lambda \\ -\lambda^2 + \lambda + \theta & -\lambda^2 + \lambda & \cdots & -\lambda^2 + \lambda & 2 - \lambda^2 \end{pmatrix},$$

which is a tridiagonal matrix except for the first and last rows and the first and last columns. The Sherman-Morrison formula implies

$$\left(B + \boldsymbol{v}\boldsymbol{v}^\top\right)^{-1} = B^{-1} - \frac{B^{-1}\boldsymbol{v}\boldsymbol{v}^\top B^{-1}}{1 + \boldsymbol{v}^\top B^{-1}\boldsymbol{v}}. \tag{38}$$

We also note that $B = (2\theta + \lambda)A + V$, where

$$V = \begin{pmatrix} 2\lambda\theta & \lambda\theta & \cdots & -\lambda^2 + \lambda + \theta \\ \lambda\theta & 0 & \cdots & -\lambda^2 + \lambda \\ \lambda\theta & 0 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \lambda - \lambda^2 + \theta(\lambda + 2\theta) \\ -\lambda^2 + \lambda + \theta & -\lambda^2 + \lambda & \cdots & 2 - \lambda^2 - \lambda - 2\theta \end{pmatrix},$$

hence

$$A = \frac{1}{2\theta + \lambda}(B - V). \tag{39}$$

Combining (36)–(39) yields

$$\begin{aligned}(\mathbb{1}^\top M^{-1}\mathbb{1})\boldsymbol{\tau}^* &= M^{-1}\mathbb{1} = \frac{1}{\theta}A^\top\left(EA^\top + AE^\top\right)^{-1}A\mathbb{1} = \frac{1}{\theta}A^\top\left(B + \boldsymbol{v}\boldsymbol{v}^\top\right)^{-1}A\mathbb{1} \\ &= \frac{1}{\theta}A^\top\left(B^{-1} - \frac{B^{-1}\boldsymbol{v}\boldsymbol{v}^\top B^{-1}}{1 + \boldsymbol{v}^\top B^{-1}\boldsymbol{v}}\right)A\mathbb{1} = \frac{1}{\theta(2\theta + \lambda)}(B - V^\top)\left(B^{-1} - \frac{B^{-1}\boldsymbol{v}\boldsymbol{v}^\top B^{-1}}{1 + \boldsymbol{v}^\top B^{-1}\boldsymbol{v}}\right)A\mathbb{1} \\ &= \frac{1}{\theta(2\theta + \lambda)}\left(A\mathbb{1} - \frac{\boldsymbol{v}\boldsymbol{v}^\top B^{-1}}{1 + \boldsymbol{v}^\top B^{-1}\boldsymbol{v}}A\mathbb{1} - V^\top\left(B^{-1} - \frac{B^{-1}\boldsymbol{v}\boldsymbol{v}^\top B^{-1}}{1 + \boldsymbol{v}^\top B^{-1}\boldsymbol{v}}\right)A\mathbb{1}\right).\end{aligned}$$

Let us now consider a T vector $\boldsymbol{\nu}$ which equals $\boldsymbol{\tau}^*$, except that elements $j, i \in \{2, 3, \ldots, T-1\}$ are flipped. We write $\boldsymbol{\nu} = P\boldsymbol{\tau}^*$ for a permutation matrix P, which is a diagonal matrix with 1 on the diagonal except for the i^{th} and j^{th} elements on the diagonal which are zero, and $P(i, j) = 1$ and $P(j, i) = 1$. We check directly that

$$PA\mathbb{1} = A\mathbb{1}, \qquad PV^\top = V^\top, \qquad P\boldsymbol{v} = \boldsymbol{v}$$

so that

$$\begin{aligned}\boldsymbol{\nu} = P\boldsymbol{\tau}^* &= \frac{1}{\theta(2\theta+\lambda)(\mathbb{1}^\top M^{-1}\mathbb{1})}\left(PA\mathbb{1} - \frac{P\boldsymbol{v}\boldsymbol{v}^\top B^{-1}}{1+\boldsymbol{v}^\top B^{-1}\boldsymbol{v}}A\mathbb{1} - PV^\top\left(B^{-1} - \frac{B^{-1}\boldsymbol{v}\boldsymbol{v}^\top B^{-1}}{1+\boldsymbol{v}^\top B^{-1}\boldsymbol{v}}\right)A\mathbb{1}\right)\\ &= \frac{1}{\theta(2\theta+\lambda)(\mathbb{1}^\top M^{-1}\mathbb{1})}\left(A\mathbb{1} - \frac{\boldsymbol{v}\boldsymbol{v}^\top B^{-1}}{1+\boldsymbol{v}^\top B^{-1}\boldsymbol{v}}A\mathbb{1} - V^\top\left(B^{-1} - \frac{B^{-1}\boldsymbol{v}\boldsymbol{v}^\top B^{-1}}{1+\boldsymbol{v}^\top B^{-1}\boldsymbol{v}}\right)A\mathbb{1}\right) = \boldsymbol{\tau}^*,\end{aligned}$$

which shows $\tau_i^* = \tau_j^*$ and thus proves $\tau_i^* = \tau_j^*$ for all $i, j \in \{2, 3, \ldots, T-1\}$.

For the second statement that the weights of the optimal contract and the dealer's trading strategy depend on $(\lambda, \sigma, \theta)$ only through $\alpha = \frac{\lambda\sigma^2}{\theta}$, we recall the comment at the beginning of this Appendix that the case of a general σ is equivalent to replacing λ by $\lambda\sigma^2$ and then setting $\sigma = 1$. For the dependence structure on θ, it can be seen in the case $\gamma = 0$ that minimizing the client's expected costs $\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{\tau} \cdot \boldsymbol{p}] = p_0 + \theta \sum_{t=1}^T \tau_t X_t$ in (12) is equivalent to minimizing $\sum_{t=1}^T \tau_t X_t$, which does not directly depend on θ. Moreover, for $\gamma = 0$ and $\theta \neq 0$, the constraint (11) can be rewritten as

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

which depends on θ only through $\frac{\lambda}{\theta}$, hence so must the optimizers. □

OA.A.2.2 Alternative proof of Proposition 9 for the case of no temporary price impact

We now consider the continuous-time limit of the model in this case with no temporary price impact. In this case, we can build upon Corollary 13 to provide a more rigorous proof of the result of Proposition 9 than the general proof given in Appendix A.

As in Proposition 9, we 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. In this case of no temporary price impact, Corollary 13 implies that the optimal contract in discrete time is a U-shape with a 'flat bottom,' in that all interior weights are the same. Thus, without temporary price impact, we know for each fixed k the optimal contract up to a single parameter, say a_k. We then derive a recursive system of linear equations to characterize the trading policy that the dealer selects in response to any a_k. Solving this system for the trading policy and plugging that in, the client's optimization problem becomes quadratic in a_k, so that we can readily solve for the optimal a_k. The formula for the optimal a_k is complicated, but it simplifies in the limit as $k \to \infty$. That expression in turn allows us to compute the limits of the optimal contract and trading policy, which are indeed as stated in Proposition 9.

Proof. Assume $\gamma = 0$. We consider a fixed k and 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 λ is replaced by

$\frac{\lambda T\sigma^2}{T_k}$. From Corollary 13, it follows that the optimal contract is of the form

$$\boldsymbol{\tau}^k = \left(\frac{1-(T_k-2)a_k}{2}, a_k, a_k, \ldots, a_k, \frac{1-(T_k-2)a_k}{2}\right)^\top \tag{40}$$

for some constant a_k. Therefore, (11) with $\gamma = 0$ becomes

$$\begin{aligned} X_1^k + \frac{T_k\theta}{\lambda T\sigma^2}\big(2X_1^k - X_2^k\big) &= \left(\frac{T_k\theta}{\lambda T\sigma^2}+1\right)\frac{1-(T_k-2)a_k}{2}, \\ X_t^k + \frac{T_k\theta}{\lambda T\sigma^2}\big(2X_t^k - X_{t+1}^k - X_{t-1}^k\big) &= \frac{1-(T_k-2)a_k}{2} + \frac{T_k\theta}{\lambda T\sigma^2}a_k + a_k(t-1), \quad t = 2,3,\ldots,T_k-1. \end{aligned}$$

Let us define

$$Y_t^k = X_t^k - \frac{1-(T_k-2)a_k}{2} - \frac{T_k\theta}{\lambda T\sigma^2}a_k - a_k(t-1), \quad t = 1,2,\ldots,T_k, \tag{41}$$

which satisfies

$$\begin{aligned} Y_1^k + \frac{T_k\theta}{\lambda T\sigma^2}\big(2Y_1^k - Y_2^k\big) &= -\frac{T_k^2\theta^2 a_k}{\lambda^2T^2\sigma^4}, \\ Y_t^k + \frac{T_k\theta}{\lambda T\sigma^2}\big(2Y_t^k - Y_{t+1}^k - Y_{t-1}^k\big) &= 0, \quad t = 2,3,\ldots,T_k-1. \end{aligned} \tag{42}$$

We set $Z_t^k = Y_{t+1}^k$ for $t = 0,1,\ldots,T_k-1$ so that

$$Z_t^k = \left(\frac{\lambda T\sigma^2}{T_k\theta}+2\right)Z_{t-1}^k - Y_{t-1}^k \ \text{ and } \ Y_t^k = Z_{t-1}^k, \quad t = 2,3,\ldots,T_k-1,$$

which we can write as

$$\begin{pmatrix} Z_t^k \\ Y_t^k \end{pmatrix} = M_k \begin{pmatrix} Z_{t-1}^k \\ Y_{t-1}^k \end{pmatrix} \ \text{ for } \ M_k = \begin{pmatrix} \frac{\lambda T\sigma^2}{T_k\theta}+2 & -1 \\ 1 & 0 \end{pmatrix}, \quad t = 2,3,\ldots,T_k-1.$$

By iteration, we obtain

$$\begin{pmatrix} Z_t^k \\ Y_t^k \end{pmatrix} = \underbrace{M_k\cdots M_k}_{t-1 \text{ times}}\begin{pmatrix} Z_1^k \\ Y_1^k \end{pmatrix} = M_k^{t-1}\begin{pmatrix} Y_2^k \\ Y_1^k \end{pmatrix}, \quad t = 1,2,\ldots,T_k-1.$$

Next, we note that M_k is diagonalizable with eigendecomposition $M_k = V_kD_kV_k^{-1}$ for

$$D_k = \begin{pmatrix} 1+\alpha_k-\beta_k & 0 \\ 0 & 1+\alpha_k+\beta_k \end{pmatrix} \ \text{ and } \ V_k = \begin{pmatrix} 1+\alpha_k-\beta_k & 1+\alpha_k+\beta_k \\ 1 & 1 \end{pmatrix},$$

where $\alpha_k = \frac{\lambda T \sigma^2}{2T_k\theta}$ and $\beta_k = \sqrt{\frac{\lambda T\sigma^2}{T_k\theta}\left(\frac{\lambda T\sigma^2}{4T_k\theta}+1\right)}$. We can compute

$$
\begin{aligned}
&M_k^{t-1} = V_k D_k^{t-1} V_k^{-1} \\
&= \frac{1}{2\beta_k}\begin{pmatrix} (1+\alpha_k+\beta_k)^t - (1+\alpha_k-\beta_k)^t & (1+\alpha_k-\beta_k)^t(1+\alpha_k+\beta_k) - (1+\alpha_k+\beta_k)^t(1+\alpha_k-\beta_k) \\ (1+\alpha_k+\beta_k)^{t-1} - (1+\alpha_k-\beta_k)^{t-1} & (1+\alpha_k-\beta_k)^{t-1}(1+\alpha_k+\beta_k) - (1+\alpha_k+\beta_k)^{t-1}(1+\alpha_k-\beta_k) \end{pmatrix},
\end{aligned}
$$

so that we obtain

$$
\begin{aligned}
Z_{T_k-1}^k = {} & \frac{1}{2\beta_k}\left((1+\alpha_k+\beta_k)^{T_k-1} - (1+\alpha_k-\beta_k)^{T_k-1}\right)Y_2^k \\
& + \frac{1}{2\beta_k}\left((1+\alpha_k-\beta_k)^{T_k-1}(1+\alpha_k+\beta_k) - (1+\alpha_k+\beta_k)^{T_k-1}(1+\alpha_k-\beta_k)\right)Y_1^k.
\end{aligned}
$$

From (41), we also have

$$
Z_{T_k-1}^k = Y_{T_k}^k = 1 - \frac{1-(T_k-2)a_k}{2} - \frac{T_k\theta}{\lambda T\sigma^2}a_k - a_k(T_k-1) = \frac{1}{2} - \frac{1}{2}\left(T_k - \frac{1}{\alpha_k}\right)a_k.
$$

Hence,

$$
\begin{aligned}
\beta_k - \beta_k\left(T_k - \frac{1}{\alpha_k}\right)a_k = {} & \left((1+\alpha_k+\beta_k)^{T_k-1} - (1+\alpha_k-\beta_k)^{T_k-1}\right)Y_2^k \\
& + \left((1+\alpha_k-\beta_k)^{T_k-1}(1+\alpha_k+\beta_k) - (1+\alpha_k+\beta_k)^{T_k-1}(1+\alpha_k-\beta_k)\right)Y_1^k
\end{aligned}
$$

Combining this with (42), which can be written as

$$
\left(1+\frac{1}{a_k}\right)Y_1^k - \frac{1}{2\alpha_k}Y_2^K = -\frac{a_k}{4\alpha_k^2},
$$

we obtain

$$
A_k\begin{pmatrix} Y_1^k \\ Y_2^k \end{pmatrix} = \begin{pmatrix} -\beta_k T_k - \frac{\beta_k}{\alpha_k} \\ -\frac{1}{4\alpha_k^2} \end{pmatrix} a_k + \begin{pmatrix} \beta_k \\ 0 \end{pmatrix},
$$

where

$$
\begin{aligned}
A_k &= \begin{pmatrix} A_k^{11} & A_k^{12} \\ 1+\frac{1}{\alpha_k} & -\frac{1}{2\alpha_k} \end{pmatrix} \\
A_k^{11} &= (1+\alpha_k-\beta_k)^{T_k-1}(1+\alpha_k+\beta_k) - (1+\alpha_k+\beta_k)^{T_k-1}(1+\alpha_k-\beta_k) \\
A_k^{12} &= (1+\alpha_k+\beta_k)^{T_k-1} - (1+\alpha_k-\beta_k)^{T_k-1}
\end{aligned}
$$

so that

$$
Y_t^k = \left(\boldsymbol{v}_t^k\right)^\top A_k^{-1}\begin{pmatrix} -\beta_k T_k - \frac{\beta_k}{\alpha_k} \\ -\frac{1}{4\alpha_k^2} \end{pmatrix} a_k + \left(\boldsymbol{v}_t^k\right)^\top A_k^{-1}\begin{pmatrix} \beta_k \\ 0 \end{pmatrix}, \quad t = 1, 2, \ldots, T_k
$$

for

$$
\begin{aligned}
\left(\boldsymbol{v}_t^k\right)^\top &= \begin{pmatrix} 0 & 1 \end{pmatrix} M_k^{t-1} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \\
&= \frac{1}{2\beta_k} \begin{pmatrix} (1+\alpha_k-\beta_k)^{t-1}(1+\alpha_k+\beta_k) - (1+\alpha_k+\beta_k)^{t-1}(1+\alpha_k-\beta_k) \\ (1+\alpha_k+\beta_k)^{t-1} - (1+\alpha_k-\beta_k)^{t-1} \end{pmatrix}^\top .
\end{aligned}
$$

For future reference, we also compute

$$
\begin{aligned}
A_k^{-1} &= \frac{2\alpha_k}{A_k^{11} + 2A_k^{12}(\alpha_k+1)} \begin{pmatrix} \frac{1}{2\alpha_k} & A_k^{12} \\ 1+\frac{1}{\alpha_k} & -A_k^{11} \end{pmatrix} \\
A_k^{11} + 2A_k^{12}(\alpha_k+1) &= (1+\alpha_k+\beta_k)^{T_k} - (1+\alpha_k-\beta_k)^{T_k}.
\end{aligned}
$$

Thus, we deduce from (41) that

$$
X_t^k = \eta_t^k a_k + \kappa_t^k, \quad t = 1, 2, \ldots, T_k, \tag{43}
$$

where

$$
\eta_t^k = \left(\boldsymbol{v}_t^k\right)^\top A_k^{-1} \begin{pmatrix} -\beta_k T_k - \frac{\beta_k}{\alpha_k} \\ -\frac{1}{4\alpha_k^2} \end{pmatrix} - \frac{T_k-2}{2} + \frac{1}{2\alpha_k} + t - 1, \tag{44}
$$

$$
\kappa_t^k = \left(\boldsymbol{v}_t^k\right)^\top A_k^{-1} \begin{pmatrix} \beta_k \\ 0 \end{pmatrix} + \frac{1}{2}. \tag{45}
$$

By (12) with $\gamma = 0$, the client's expected cost of the contract is

$$
\begin{aligned}
&p_0 + \theta \sum_{t=1}^{T_k} \tau_t^k X_t^k = p_0 + \theta a_k \sum_{t=1}^{T_k} \tau_t^k \eta_t^k + \theta \sum_{t=1}^{T_k} \tau_t^k \kappa_t^k \\
&= p_0 + \theta a_k \left(\tau_1^k \eta_1^k + \tau_{T_k}^k \eta_{T_k}^k + \sum_{t=2}^{T_k-1} \tau_t^k \eta_t^k \right) + \theta \left(\tau_1^k \kappa_1^k + \tau_{T_k}^k \kappa_{T_k}^k + \sum_{t=2}^{T_k-1} \tau_t^k \kappa_t^k \right) \\
&= p_0 + \theta a_k \left(\frac{1-(T_k-2)a_k}{2} \left(\eta_1^k + \eta_{T_k}^k\right) + a_k \sum_{t=2}^{T_k-1} \eta_t^k \right) + \theta \left(\frac{1-(T_k-2)a_k}{2} \left(\kappa_1^k + \kappa_{T_k}^k\right) + a_k \sum_{t=2}^{T_k-1} \kappa_t^k \right) \\
&= p_0 + \theta a_k^2 \left(-\frac{T_k-2}{2} \left(\eta_1^k + \eta_{T_k}^k\right) + \sum_{t=2}^{T_k-1} \eta_t^k \right) + \theta a_k \left(\frac{\eta_1^k + \eta_{T_k}^k}{2} - \frac{T_k-2}{2} \left(\kappa_1^k + \kappa_{T_k}^k\right) + \sum_{t=2}^{T_k-1} \kappa_t^k \right) \\
&\quad + \theta \frac{\kappa_1^k + \kappa_{T_k}^k}{2}.
\end{aligned} \tag{46}
$$

The first-order condition yields

$$a_k = \frac{\frac{\eta_1^k+\eta_{T_k}^k}{2} - \frac{T_k-2}{2}(\kappa_1^k + \kappa_{T_k}^k) + \sum_{t=2}^{T_k-1} \kappa_t^k}{(T_k-2)(\eta_1^k + \eta_{T_k}^k) - 2\sum_{t=2}^{T_k-1} \eta_t^k}. \tag{47}$$

We will next show

$$\lim_{k\to\infty} \frac{\eta_{\lceil qT_k \rceil}^k}{T_k} = \begin{cases} -\frac{1}{2} + \frac{\theta}{\lambda T\sigma^2} + q & \text{if } q \in (0,1), \\ 0 & \text{if } q = 1, \end{cases} \tag{48}$$

$$\lim_{k\to\infty} \kappa_{\lceil qT_k \rceil}^k = \begin{cases} \frac{1}{2} & \text{if } q \in (0,1), \\ 1 & \text{if } q = 1, \end{cases} \tag{49}$$

$$\lim_{k\to\infty} \frac{\eta_1^k}{T_k} = -\frac{1}{2}, \tag{50}$$

$$\lim_{k\to\infty} \kappa_1^k = \frac{1}{2}. \tag{51}$$

In the following limit computations, we will repeatedly use the fact that

$$\lim_{n\to\infty}\left(1+\frac{c}{\sqrt{n}}\right)^n = \lim_{m\to\infty}\left(1+\frac{c}{m}\right)^{m^2} = \lim_{m\to\infty}\left(\left(1+\frac{c}{m}\right)^m\right)^m = \lim_{m\to\infty} \mathrm{e}^{cm} = \begin{cases} 0 & \text{if } c < 0, \\ 1 & \text{if } c = 0, \\ \infty & \text{if } c > 0, \end{cases}$$

and that exponential convergence is faster than any power convergence.

To illustrate how we use this fact, first note that we have

$$1 + \alpha_k + \beta_k = 1 + \frac{\lambda T\sigma^2}{2T_k\theta} + \sqrt{\frac{\lambda T\sigma^2}{T_k\theta}\left(\frac{\lambda T\sigma^2}{4T_k\theta} + 1\right)} \geq 1 + \sqrt{\frac{\lambda T\sigma^2}{T_k\theta}}$$

and that for sufficiently large T_k, we also have

$$1 + \alpha_k - \beta_k = 1 + \frac{\lambda T\sigma^2}{2T_k\theta} - \sqrt{\frac{\lambda T\sigma^2}{T_k\theta}\left(\frac{\lambda T\sigma^2}{4T_k\theta} + 1\right)} \leq 1 - \frac{1}{2}\sqrt{\frac{\lambda T\sigma^2}{T_k\theta}}.$$

So the fact implies that for any $C \in (0,1]$, we have both $(1+\alpha_k+\beta_k)^{CT_k} \to \infty$ and $(1+\alpha_k-\beta_k)^{CT_k} \to 0$. The reason is that β_k converges to zero only like $\frac{1}{\sqrt{T_k}}$ while α_k converges to zero like $\frac{1}{T_k}$. Because of the slow convergence of β_k, it dominates the other terms.

To show (48), we first compute

$$\lim_{k\to\infty} \frac{1}{T_k}\left(\boldsymbol{v}_{\lceil qT_k \rceil}^k\right)^\top A_k^{-1} \begin{pmatrix} -\beta_k T_k - \frac{\beta_k}{\alpha_k} \\ -\frac{1}{4\alpha_k^2} \end{pmatrix}$$

$$
\begin{aligned}
&= \lim_{k\to\infty} \frac{1}{T_k}\big(\boldsymbol{v}^k_{\lceil qT_k\rceil}\big)^\top \frac{2\alpha_k}{A_k^{11}+2A_k^{12}(\alpha_k+1)} \begin{pmatrix} \frac{1}{2\alpha_k} & A_k^{12} \\ 1+\frac{1}{\alpha_k} & -A_k^{11} \end{pmatrix} \begin{pmatrix} -\beta_k T_k - \frac{\beta_k}{\alpha_k} \\ -\frac{1}{4\alpha_k^2} \end{pmatrix} \\
&= \lim_{k\to\infty} \frac{1}{T_k}\big(\boldsymbol{v}^k_{\lceil qT_k\rceil}\big)^\top \frac{1}{A_k^{11}+2A_k^{12}(\alpha_k+1)} \begin{pmatrix} -\beta_k T_k - \frac{\beta_k}{\alpha_k} - \frac{A_k^{12}}{2\alpha_k} \\ 2(1+\alpha_k)\big(-\beta_k T_k - \frac{\beta_k}{\alpha_k}\big) + \frac{A_k^{11}}{2\alpha_k} \end{pmatrix} \\
&= \lim_{k\to\infty} \frac{1}{2T_k\beta_k(A_k^{11}+2A_k^{12}(\alpha_k+1))} \Bigg(\Big((1+\alpha_k-\beta_k)^{\lceil qT_k\rceil-1}(1+\alpha_k+\beta_k) \\
&\quad - (1+\alpha_k+\beta_k)^{\lceil qT_k\rceil-1}(1+\alpha_k-\beta_k)\Big)\Big(-\beta_k T_k - \frac{\beta_k}{\alpha_k} - \frac{A_k^{12}}{2\alpha_k}\Big) \\
&\quad + \Big((1+\alpha_k+\beta_k)^{\lceil qT_k\rceil-1} - (1+\alpha_k-\beta_k)^{\lceil qT_k\rceil-1}\Big)\Big(2(1+\alpha_k)\Big(-\beta_k T_k - \frac{\beta_k}{\alpha_k}\Big) + \frac{A_k^{11}}{2\alpha_k}\Big)\Bigg) \\
&= \lim_{k\to\infty} \frac{1}{2T_k\beta_k(A_k^{11}+2A_k^{12}(\alpha_k+1))} \Bigg((1+\alpha_k-\beta_k)^{\lceil qT_k\rceil-1}\Big((1+\alpha_k+\beta_k) \\
&\quad \times \Big(-\beta_k T_k - \frac{\beta_k}{\alpha_k} - \frac{A_k^{12}}{2\alpha_k}\Big) - 2(1+\alpha_k)\Big(-\beta_k T_k - \frac{\beta_k}{\alpha_k}\Big) + \frac{A_k^{11}}{2\alpha_k}\Big) \\
&\quad + (1+\alpha_k+\beta_k)^{\lceil qT_k\rceil-1}\Big(-(1+\alpha_k-\beta_k)\Big(-\beta_k T_k - \frac{\beta_k}{\alpha_k} - \frac{A_k^{12}}{2\alpha_k}\Big) \\
&\quad + 2(1+\alpha_k)\Big(-\beta_k T_k - \frac{\beta_k}{\alpha_k}\Big) + \frac{A_k^{11}}{2\alpha_k}\Big)\Bigg) \\
&= \lim_{k\to\infty} \frac{(1+\alpha_k+\beta_k)^{\lceil qT_k\rceil-1}}{2T_k\beta_k((1+\alpha_k+\beta_k)^{T_k} - (1+\alpha_k-\beta_k)^{T_k})} \\
&\quad \times \Big(-(1+\alpha_k-\beta_k)\Big(-\beta_k T_k - \frac{\beta_k}{\alpha_k} + \frac{(1+\alpha_k-\beta_k)^{T_k-1}}{2\alpha_k}\Big) \\
&\quad + 2(1+\alpha_k)\Big(-\beta_k T_k - \frac{\beta_k}{\alpha_k}\Big) + \frac{(1+\alpha_k-\beta_k)^{T_k-1}(1+\alpha_k+\beta_k)}{2\alpha_k}\Big) \\
&= \lim_{k\to\infty} \frac{(1+\alpha_k+\beta_k)^{\lceil qT_k\rceil-1}\big((1+\alpha_k+\beta_k)\big(-\beta_k T_k - \frac{\beta_k}{\alpha_k}\big) + (1+\alpha_k-\beta_k)^{T_k-1}\frac{\beta_k}{\alpha_k}\big)}{2T_k\beta_k((1+\alpha_k+\beta_k)^{T_k} - (1+\alpha_k-\beta_k)^{T_k})} \\
&= \lim_{k\to\infty} \frac{(1+\alpha_k+\beta_k)^{\lceil qT_k\rceil-T_k}\big(-\beta_k T_k - \frac{\beta_k}{\alpha_k} + (1+\alpha_k-\beta_k)^{T_k-1}\frac{\beta_k}{\alpha_k(1+\alpha_k+\beta_k)}\big)}{2T_k\beta_k\Big(\frac{(1+\alpha_k-\beta_k)^{T_k}}{(1+\alpha_k+\beta_k)^{T_k}}+1\Big)} \\
&= \begin{cases} 0 & \text{if } q\in(0,1), \\ \lim\limits_{k\to\infty} \frac{-\beta_k T_k - \frac{\beta_k}{\alpha_k}}{2T_k\beta_k} = \lim\limits_{k\to\infty}\big(-\frac{1}{2} - \frac{1}{2T_k\alpha_k}\big) = -\frac{1}{2} - \frac{\theta}{\lambda T\sigma^2} & \text{if } q=1. \end{cases}
\end{aligned}
\tag{52}
$$

Therefore, we deduce from (44) that

$$\lim_{k\to\infty}\frac{\eta^k_{\lceil qT_k\rceil}}{T_k} = \lim_{k\to\infty}\left(\frac{1}{T_k}\big(\boldsymbol{v}^k_{\lceil qT_k\rceil}\big)^\top A_k^{-1}\begin{pmatrix} -\beta_k T_k - \frac{\beta_k}{\alpha_k} \\ -\frac{1}{4\alpha_k^2}\end{pmatrix} - \frac{1}{2} + \frac{1}{2T_k\alpha_k} + q\right)$$
$$= \begin{cases} -\frac{1}{2} + \frac{\theta}{\lambda T\sigma^2} + q & \text{if } q\in(0,1), \\ 0 & \text{if } q = 1,\end{cases}$$

which shows (48). Similarly, for (49), we first compute

$$\begin{aligned}
&\lim_{k\to\infty}\big(\boldsymbol{v}^k_{\lceil qT_k\rceil}\big)^\top A_k^{-1}\begin{pmatrix}\beta_k\\ 0\end{pmatrix} \\
&= \lim_{k\to\infty}\big(\boldsymbol{v}^k_{\lceil qT_k\rceil}\big)^\top \frac{2\alpha_k}{A_k^{11}+2A_k^{12}(\alpha_k+1)}\begin{pmatrix}\frac{1}{2\alpha_k} & A_k^{12} \\ 1+\frac{1}{\alpha_k} & -A_k^{11}\end{pmatrix}\begin{pmatrix}\beta_k\\ 0\end{pmatrix} \\
&= \lim_{k\to\infty}\big(\boldsymbol{v}^k_{\lceil qT_k\rceil}\big)^\top \frac{1}{A_k^{11}+2A_k^{12}(\alpha_k+1)}\begin{pmatrix}\beta_k \\ 2\beta_k(1+\alpha_k)\end{pmatrix} \\
&= \lim_{k\to\infty}\Bigg(\frac{\frac{1}{2}\big((1+\alpha_k-\beta_k)^{\lceil qT_k\rceil-1}(1+\alpha_k+\beta_k)-(1+\alpha_k+\beta_k)^{\lceil qT_k\rceil-1}(1+\alpha_k-\beta_k)\big)}{A_k^{11}+2A_k^{12}(\alpha_k+1)} \\
&\qquad + \frac{\big((1+\alpha_k+\beta_k)^{\lceil qT_k\rceil-1}-(1+\alpha_k-\beta_k)^{\lceil qT_k\rceil-1}\big)(1+\alpha_k)}{A_k^{11}+2A_k^{12}(\alpha_k+1)}\Bigg) \\
&= \lim_{k\to\infty}\frac{-\frac{1}{2}(1+\alpha_k+\beta_k)^{\lceil qT_k\rceil-1}(1+\alpha_k-\beta_k)+(1+\alpha_k+\beta_k)^{\lceil qT_k\rceil-1}(1+\alpha_k)}{(1+\alpha_k+\beta_k)^{T_k}} \\
&= \lim_{k\to\infty}\frac{1}{2}(1+\alpha_k+\beta_k)^{\lceil qT_k\rceil-T_k} \\
&= \begin{cases} 0 & \text{if } q\in(0,1), \\ \frac{1}{2} & \text{if } q=1,\end{cases}
\end{aligned} \tag{53}$$

so that by (45) we obtain

$$\lim_{k\to\infty}\kappa^k_{\lceil qT_k\rceil} = \begin{cases}\frac{1}{2} & \text{if } q\in(0,1), \\ 1 & \text{if } q=1,\end{cases}$$

which proves (49). To show (50), we begin by computing from (44)

$$
\begin{aligned}
\eta_1^k &= \left(\boldsymbol{v}_1^k\right)^\top A_k^{-1} \begin{pmatrix} -\beta_k T_k - \frac{\beta_k}{\alpha_k} \\ -\frac{1}{4\alpha_k^2} \end{pmatrix} - \frac{T_k - 2}{2} + \frac{1}{2\alpha_k} \\
&= (1,0) \frac{2\alpha_k}{A_k^{11} + 2A_k^{12}(\alpha_k+1)} \begin{pmatrix} \frac{1}{2\alpha_k} & A_k^{12} \\ 1 + \frac{1}{\alpha_k} & -A_k^{11} \end{pmatrix} \begin{pmatrix} -\beta_k T_k - \frac{\beta_k}{\alpha_k} \\ -\frac{1}{4\alpha_k^2} \end{pmatrix} - \frac{T_k}{2} + 1 + \frac{1}{2\alpha_k} \\
&= \frac{2\alpha_k}{A_k^{11} + 2A_k^{12}(\alpha_k+1)} \begin{pmatrix} \frac{1}{2\alpha_k} & A_k^{12} \end{pmatrix} \begin{pmatrix} -\beta_k T_k - \frac{\beta_k}{\alpha_k} \\ -\frac{1}{4\alpha_k^2} \end{pmatrix} - \frac{T_k}{2} + 1 + \frac{1}{2\alpha_k} \\
&= \frac{-\beta_k T_k - \frac{\beta_k}{\alpha_k} - \frac{A_k^{12}}{2\alpha_k}}{A_k^{11} + 2A_k^{12}(\alpha_k+1)} - \frac{T_k}{2} + 1 + \frac{1}{2\alpha_k} \\
&= \frac{-\beta_k T_k - \frac{\beta_k}{\alpha_k} - \frac{1}{2\alpha_k}\left((1+\alpha_k+\beta_k)^{T_k-1} - (1+\alpha_k-\beta_k)^{T_k-1}\right)}{(1+\alpha_k+\beta_k)^{T_k} - (1+\alpha_k-\beta_k)^{T_k}} - \frac{T_k}{2} + 1 + \frac{1}{2\alpha_k}.
\end{aligned} \tag{54}
$$

We then compute the limit

$$
\begin{aligned}
\lim_{k\to\infty} \frac{\eta_1^k}{T_k} &= \frac{-\beta_k - \frac{\beta_k}{\alpha_k T_k} - \frac{1}{2\alpha_k T_k}\left((1+\alpha_k+\beta_k)^{T_k-1} - (1+\alpha_k-\beta_k)^{T_k-1}\right)}{(1+\alpha_k+\beta_k)^{T_k} - (1+\alpha_k-\beta_k)^{T_k}} - \frac{1}{2} + \frac{1}{T_k} + \frac{1}{2\alpha_k T_k} \\
&= \lim_{k\to\infty} \frac{-\frac{\beta_k}{(1+\alpha_k+\beta_k)^{T_k-1} - (1+\alpha_k-\beta_k)^{T_k-1}}\left(1 + \frac{2\theta}{\lambda T \sigma^2}\right) - \frac{\theta}{\lambda T\sigma^2}}{\frac{(1+\alpha_k+\beta_k)^{T_k-1}(1+\alpha_k+\beta_k) - (1+\alpha_k-\beta_k)^{T_k-1}(1+\alpha_k-\beta_k)}{(1+\alpha_k+\beta_k)^{T_k-1} - (1+\alpha_k-\beta_k)^{T_k-1}}} - \frac{1}{2} + \frac{\theta}{\lambda T\sigma^2} \\
&= \lim_{k\to\infty} \frac{-\frac{\theta}{\lambda T\sigma^2}}{1+\alpha_k+\beta_k} - \frac{1}{2} + \frac{\theta}{\lambda T\sigma^2} = -\frac{1}{2},
\end{aligned}
$$

which proves (50). Finally, for (51), we begin by computing from (45)

$$
\begin{aligned}
\kappa_1^k &= \left(\boldsymbol{v}_1^k\right)^\top A_k^{-1} \begin{pmatrix} \beta_k \\ 0 \end{pmatrix} + \frac{1}{2} \\
&= (1,0) \frac{2\alpha_k}{A_k^{11} + 2A_k^{12}(\alpha_k+1)} \begin{pmatrix} \frac{1}{2\alpha_k} & A_k^{12} \\ 1 + \frac{1}{\alpha_k} & -A_k^{11} \end{pmatrix} \begin{pmatrix} \beta_k \\ 0 \end{pmatrix} + \frac{1}{2} \\
&= \frac{\beta_k}{(1+\alpha_k+\beta_k)^{T_k} - (1+\alpha_k-\beta_k)^{T_k}} + \frac{1}{2},
\end{aligned} \tag{55}
$$

so that $\lim_{k\to\infty} \kappa_1^k = \frac{1}{2}$, which concludes the proof of (47)–(51). Additionally, we note that the convergence in (52) and (53) is uniform in q on compact sets in $(0,1)$. To see this, note that the expressions in the limits (52) and (53) depend on q only through $(1+\alpha_k+\beta_k)^{\lceil qT_k \rceil - T_k}$ and $(1+\alpha_k-\beta_k)^{\lceil qT_k \rceil - 1}$, which both converge to zero uniformly in q on compact sets in $(0,1)$. Indeed, for every $0 < q_0 < Q_0 < 1$, we have

$$0 \le \lim_{k\to\infty} \sup_{q\in[q_0,Q_0]} (1+\alpha_k+\beta_k)^{\lceil qT_k\rceil - T_k} \le \lim_{k\to\infty} (1+\alpha_k+\beta_k)^{\lceil Q_0T_k\rceil - T_k} = 0,$$
$$0 \le \lim_{k\to\infty} \sup_{q\in[q_0,Q_0]} (1+\alpha_k-\beta_k)^{\lceil qT_k\rceil - 1} \le \lim_{k\to\infty} (1+\alpha_k-\beta_k)^{\lceil Q_0T_k\rceil - 1} = 0,$$

which shows

$$\lim_{k\to\infty} \sup_{q\in[q_0,Q_0]} (1+\alpha_k+\beta_k)^{\lceil qT_k\rceil - T_k} = 0 \text{ and } \lim_{k\to\infty} \sup_{q\in[q_0,Q_0]} (1+\alpha_k-\beta_k)^{\lceil qT_k\rceil - 1} = 0.$$

This implies that for every $0 < q_0 < Q_0 < 1$ and every $\epsilon > 0$, there exists k_0 such that

$$\left|\frac{\eta^k_{\lceil qT_k\rceil}}{T_k} + \frac{1}{2} - \frac{\theta}{\lambda T\sigma^2} - q\right| < \epsilon \quad\text{and}\quad \left|\kappa^k_{\lceil qT_k\rceil} - \frac{1}{2}\right| < \epsilon$$

for all $k > k_0$ and $q \in [q_0, Q_0]$. Note that k_0 may depend on q_0, Q_0 and ϵ, but it does not depend on q because of the uniform convergence. We can also deduce from (52) and (53) that there exists k_0 such that $\left|\frac{\eta^k_t}{T_k}\right| \le 1 + \frac{\theta}{\lambda T\sigma^2}$ and $|\kappa^k_t| \le \frac{3}{2}$ for all $k \ge k_0$ and $t = 1, 2, \dots, T_k$. Therefore, for every $0 < q_0 < Q_0 < 1$ and every $\epsilon > 0$, there exists k_0 such that

$$\begin{aligned}
&\left|\frac{1}{T_k}\sum_{t=2}^{T_k-1}\kappa^k_t - \frac{T_k-2}{2T_k}\right| = \frac{1}{T_k}\left|\sum_{t=2}^{T_k-1}\left(\kappa^k_t - \frac{1}{2}\right)\right| \\
&\le \frac{1}{T_k}\Big(2\big(\lceil q_0T_k\rceil - 1\big) + (T_k-2)\epsilon + 2\big(\lceil (1-Q_0)T_k\rceil - 1\big)\Big) \\
&< 2q_0 + \epsilon + 2(1-Q_0), \\
&\left|\frac{1}{T_k^2}\sum_{t=2}^{T_k-1}\eta^k_t + \frac{T_k-2}{2T_k} - \frac{\theta(T_k-2)}{\lambda T\sigma^2 T_k} - \frac{(T_k-2)(T_k-1)}{2T_k^2}\right| = \frac{1}{T_k}\left|\sum_{t=2}^{T_k-1}\left(\frac{\eta^k_t}{T_k} + \frac{1}{2} - \frac{\theta}{\lambda T\sigma^2} - \frac{t-1}{T_k}\right)\right| \\
&\le \frac{1}{T_k}\left(\left(3 + \frac{2\theta}{\lambda T\sigma^2}\right)\big(\lceil q_0T_k\rceil - 1\big) + (T_k-2)\epsilon + \left(3 + \frac{2\theta}{\lambda T\sigma^2}\right)\big(\lceil (1-Q_0)T_k\rceil - 1\big)\right) \\
&< \left(3 + \frac{2\theta}{\lambda T\sigma^2}\right)q_0 + \epsilon + \left(3 + \frac{2\theta}{\lambda T\sigma^2}\right)(1-Q_0)
\end{aligned}$$

for all $k > k_0$. Letting $q_0 \to 0$, $Q_0 \to 1$ and $\epsilon \to 0$, we deduce

$$\lim_{k\to\infty}\frac{1}{T_k}\sum_{t=2}^{T_k-1}\kappa^k_t = \frac{1}{2} \quad\text{and}\quad \lim_{k\to\infty}\frac{1}{T_k^2}\sum_{t=2}^{T_k-1}\eta^k_t = -\frac{1}{2} + \frac{\theta}{\lambda T\sigma^2} + \frac{1}{2} = \frac{\theta}{\lambda T\sigma^2}. \tag{56}$$

Together with (47)–(51), this implies

$$\lim_{k\to\infty} T_k a_k = \lim_{k\to\infty}\frac{\frac{\eta^k_1+\eta^k_{T_k}}{2T_k} - \frac{1}{2}(\kappa^k_1 + \kappa^k_{T_k}) + \frac{1}{T_k}\sum_{t=2}^{T_k-1}\kappa^k_t}{\frac{\eta^k_1+\eta^k_{T_k}}{T_k} - \frac{2}{T_k^2}\sum_{t=2}^{T_k-1}\eta^k_t} = \frac{-\frac{1}{4} - \frac{3}{4} + \frac{1}{2}}{-\frac{1}{2} - \frac{2\theta}{\lambda T\sigma^2}} = \frac{1}{1 + \frac{4\theta}{\lambda T\sigma^2}}. \tag{57}$$

Combining this with (40) and setting $a = \frac{1}{1+\frac{4\theta}{\lambda T\sigma^2}}$, we find

$$\lim_{k\to\infty}\sum_{t=1}^{\lceil qT_k\rceil}\tau_t^{*k} = \begin{cases} 0 & \text{if } q=0,\\ \lim_{k\to\infty}\Big(\frac{1-(T_k-2)a_k}{2} + \big(\lceil qT_k\rceil - 1\big)a_k\Big) = \frac{1-a}{2} + qa & \text{if } q\in(0,1),\\ 1 & \text{if } q=1.\end{cases}$$

Similarly, using (43), (48), (49), and (57), we derive

$$\lim_{k\to\infty} X^k_{\lceil qT_k\rceil} = \begin{cases} 0 & \text{if } q=0,\\ \lim_{k\to\infty}\Big(\eta^k_{\lceil qT_k\rceil}a_k + \kappa^k_{\lceil qT_k\rceil}\Big) = \big(-\frac{1}{2} + \frac{\theta}{\lambda T\sigma^2} + q\big)a + \frac{1}{2} = \frac{\theta a}{\lambda T\sigma^2} + \frac{1-a}{2} + aq & \text{if } q\in(0,1),\\ \lim_{k\to\infty}\big(\eta^k_{T_k}a_k + \kappa^k_{T_k}\big) = 0+1=1 & \text{if } q=1.\end{cases}$$

Finally, the client's expected costs of execution converge to

$$\begin{aligned}
&\lim_{k\to\infty}\Bigg(p_0 + \theta a_k^2\Bigg(-\frac{T_k-2}{2}\big(\eta_1^k+\eta_{T_k}^k\big) + \sum_{t=2}^{T_k-1}\eta_t^k\Bigg) + \theta a_k\Bigg(\frac{\eta_1^k+\eta_{T_k}^k}{2} - \frac{T_k-2}{2}\big(\kappa_1^k+\kappa_{T_k}^k\big) + \sum_{t=2}^{T_k-1}\kappa_t^k\Bigg)\\
&\qquad + \theta\frac{\kappa_1^k+\kappa_{T_k}^k}{2}\Bigg)\\
&= \lim_{k\to\infty}\Bigg(p_0 + \theta\underbrace{a_k^2T_k^2}_{\to\frac{1}{(1+\frac{4\theta}{\lambda T\sigma^2})^2}}\Bigg(-\underbrace{\frac{1-2/T_k}{2}}_{\to\frac{1}{2}}\underbrace{\frac{\eta_1^k+\eta_{T_k}^k}{T_k}}_{\to-\frac{1}{2}} + \underbrace{\frac{1}{T_k^2}\sum_{t=2}^{T_k-1}\eta_t^k}_{\to\frac{\theta}{\lambda T\sigma^2}}\Bigg)\\
&\qquad + \theta\underbrace{a_kT_k}_{\to\frac{1}{1+\frac{4\theta}{\lambda T\sigma^2}}}\Bigg(\underbrace{\frac{\eta_1^k+\eta_{T_k}^k}{2T_k}}_{\to-\frac{1}{4}} - \underbrace{\frac{1-2/T_k}{2}}_{\to\frac{1}{2}}\underbrace{\big(\kappa_1^k+\kappa_{T_k}^k\big)}_{\to\frac{3}{2}} + \underbrace{\frac{1}{T_k}\sum_{t=2}^{T_k-1}\kappa_t^k}_{\to\frac{1}{2}}\Bigg) + \theta\underbrace{\frac{\kappa_1^k+\kappa_{T_k}^k}{2}}_{\to\frac{3}{4}}\Bigg)\\
&= p_0 + \frac{\theta}{\big(1+\frac{4\theta}{\lambda T\sigma^2}\big)^2}\Bigg(-\frac{1}{2}\cdot\Big(-\frac{1}{2}\Big) + \frac{\theta}{\lambda T\sigma^2}\Bigg) + \frac{\theta}{1+\frac{4\theta}{\lambda T\sigma^2}}\Bigg(-\frac{1}{4} - \frac{1}{2}\cdot\frac{3}{2} + \frac{1}{2}\Bigg) + \theta\frac{3}{4}\\
&= p_0 + \frac{\theta}{4\big(1+\frac{4\theta}{\lambda T\sigma^2}\big)} - \frac{\theta}{2\big(1+\frac{4\theta}{\lambda T\sigma^2}\big)} + \frac{3\theta}{4}\\
&= p_0 + \frac{1+\frac{6\theta}{\lambda T\sigma^2}}{2\big(1+\frac{4\theta}{\lambda T\sigma^2}\big)}\theta\\
&= p_0 + \frac{3-a}{4}\theta
\end{aligned}$$

for $a = \frac{1}{1+\frac{4\theta}{\lambda T\sigma^2}}$, where we used (46), (48)–(51), (56), and (57). □

OA.A.2.3 Proof of equation (35)

Proof of (35). According to (43), $X_1^k = \eta_1^k a_k + \kappa_1^k$, which by (54) and (55) can be computed as

$$\begin{aligned}
X_1^k &= \eta_1^k a_k + \kappa_1^k \\
&= \left(\frac{-\beta_k T_k - \frac{\beta_k}{\alpha_k} - \frac{1}{2\alpha_k}((1+\alpha_k+\beta_k)^{T_k-1} - (1+\alpha_k-\beta_k)^{T_k-1})}{(1+\alpha_k+\beta_k)^{T_k} - (1+\alpha_k-\beta_k)^{T_k}} - \frac{T_k}{2} + 1 + \frac{1}{2\alpha_k} \right) a_k \\
&\quad + \frac{\beta_k}{(1+\alpha_k+\beta_k)^{T_k} - (1+\alpha_k-\beta_k)^{T_k}} + \frac{1}{2} \\
&= \frac{1-(T_k-2)a_k}{2} + \frac{\beta_k(1-a_kT_k) - \frac{a_k\beta_k}{\alpha_k} - \frac{a_k}{2\alpha_k}((1+\alpha_k+\beta_k)^{T_k-1} - (1+\alpha_k-\beta_k)^{T_k-1})}{(1+\alpha_k+\beta_k)^{T_k} - (1+\alpha_k-\beta_k)^{T_k}} + \frac{a_k}{2\alpha_k}.
\end{aligned}$$

To conclude the proof, we need to show that

$$\frac{\beta_k(1-a_kT_k) - \frac{a_k\beta_k}{\alpha_k} - \frac{a_k}{2\alpha_k}((1+\alpha_k+\beta_k)^{T_k-1} - (1+\alpha_k-\beta_k)^{T_k-1})}{(1+\alpha_k+\beta_k)^{T_k} - (1+\alpha_k-\beta_k)^{T_k}} + \frac{a_k}{2\alpha_k} > 0.$$

For $a_k = 0$, this inequality is satisfied, and for $a_k > 0$, it is equivalent to

$$(1+\alpha_k+\beta_k)^{T_k-1}(\alpha_k+\beta_k) - (1+\alpha_k-\beta_k)^{T_k-1}(\alpha_k-\beta_k) > -2\alpha_k\beta_k(1/a_k - T_k) + 2\beta_k. \tag{58}$$

Because of $a_k \leq \frac{1}{T_k-2}$, equation (58) follows from

$$\begin{aligned}
&(1+\alpha_k+\beta_k)^{T_k-1}(\alpha_k+\beta_k) + (1+\alpha_k-\beta_k)^{T_k-1}(-\alpha_k+\beta_k) \\
&= (\alpha_k+\beta_k)\sum_{j=0}^{T_k-1}\binom{T_k-1}{j}\beta_k^j(1+\alpha_k)^{T_k-1-j} + (-\alpha_k+\beta_k)\sum_{j=0}^{T_k-1}\binom{T_k-1}{j}(-\beta_k)^j(1+\alpha_k)^{T_k-1-j} \\
&= 2\alpha_k\sum_{j=0}^{\lfloor (T_k-1)/2\rfloor}\binom{T_k-1}{2j+1}\beta_k^{2j+1}(1+\alpha_k)^{T_k-2j-2} + 2\beta_k\sum_{j=0}^{\lfloor (T_k-1)/2\rfloor}\binom{T_k-1}{2j}\beta_k^{2j}(1+\alpha_k)^{T_k-1-2j} \\
&\geq 2\beta_k(1+\alpha_k)^{T_k-1} \\
&> 2\beta_k + 2(T_k-1)\alpha_k\beta_k \\
&\geq 2\beta_k + 4\alpha_k\beta_k,
\end{aligned}$$

where we used that $T_k \geq 3$. □

OA.A.3 Infinite execution horizon

Besides the continuous-time limit considered in Section 5, another limit of economic interest concerns what happens as the execution horizon T diverges to infinity, holding all other parameters fixed. In particular, σ would be held fixed, so the distance between consecutive trading periods should not be interpreted as going to zero, distinguishing this from the continuous-time limit. Rather, this limit can be interpreted as speaking to how the optimal contract changes as the client becomes progressively more patient.

As we did when analyzing the continuous-time limit, the following result is stated in terms of the cumulative values through quantiles q of the execution period, and we focus on the case of a strictly risk-averse dealer to ensure well-behaved convergence.

Proposition 14. *Assume $\gamma = 0$ and that the dealer is strictly risk-averse ($\lambda > 0$). For each execution horizon $T \in \mathbb{N}$, let $\boldsymbol{\tau}^{*T} \in \mathcal{T}^T$ be the optimal contract, and let $\boldsymbol{x}^{*T}$ be the dealer strategy that best responds to $\boldsymbol{\tau}^{*T}$. For all $q \in [0, 1]$,*

$$\lim_{T\to\infty} \sum_{t=1}^{\lceil qT \rceil} \tau_t^{*T} = q \quad \textit{and} \quad \lim_{T\to\infty} \sum_{t=1}^{\lceil qT \rceil} x_t^{*T} = q$$

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

This result states a sense in which, as the execution horizon T diverges, (i) the optimal contract $\boldsymbol{\tau}^{*T}$ converges to guaranteed TWAP, and (ii) the dealer's trading strategy $\boldsymbol{x}^{*T}$ converges to the first best. It follows that the client receives her first-best payoff in the limit.[35]

For the intuition, note that in general, the dealer's frontloading motive reflects a balance between two considerations: given an offered contract $\boldsymbol{\tau}$, if the dealer deviates from $\boldsymbol{x} = \boldsymbol{\tau}$ to an alternative that differs from $\boldsymbol{\tau}$, then he may increase his expected profit, but he also exposes himself to price risk. As $T \to \infty$, the dealer must take on more and more price risk in order to create the same increase in expected profit, so that such deviations become progressively less attractive, and the dealer's best response converges to a trading strategy that mirrors the contract weights. Thus, the limit resembles the case in which the dealer's best response is to choose $\boldsymbol{x} = \boldsymbol{\tau}$. And in that case, it is optimal for the client to induce the first-best policy, which she can achieve with a guaranteed TWAP contract.

Proof of Proposition 14. In this setting, (11) becomes

$$X_q^k + \frac{\theta + 2\gamma}{\lambda\sigma^2}\left(2X_q^k - X_{q+\frac{1}{T_k}}^k - X_{q-\frac{1}{T_k}}^k\right) = \frac{\gamma}{\lambda T\sigma^2}\left(2V_q^k - V_{q+\frac{1}{T_k}}^k - V_{q-\frac{1}{T_k}}^k\right) + \frac{\theta}{\lambda T\sigma^2}\left(V_q^k - V_{q-\frac{1}{T_k}}^k\right) + V_q^k \tag{59}$$

for all $q \in (0, 1)$, where $X_q^k = \sum_{t=1}^{\lceil qT_k \rceil} x_t^k$ and $V_q^k = \sum_{j=1}^{\lceil qT_k \rceil} \tau_j^k$ for $\boldsymbol{\tau}^k \in \Delta^{T_k}$ and a dealer strategy $\boldsymbol{x}^k$ in the k^{th} model. Assuming 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 continuous, we have

$$X_q = \lim_{k\to\infty} X_q^k = \lim_{k\to\infty} X_{q+\frac{1}{T_k}}^k = \lim_{k\to\infty} X_{q-\frac{1}{T_k}}^k \text{ and } V_q = \lim_{k\to\infty} V_q^k = \lim_{k\to\infty} V_{q+\frac{1}{T_k}}^k = \lim_{k\to\infty} V_{q-\frac{1}{T_k}}^k,$$

hence we obtain from (59) by letting k go to ∞ that $X_q = V_q$ for all $q \in [0, 1]$, and the client's expected costs converge to $p_0 + \frac{\theta}{2}$. Consequently, the client chooses the first-best strategy as the

[35] Indeed, Proposition 1 states that, under the first best, the client's expected costs of execution are $p_0 + \frac{\gamma}{T} + \frac{\theta(T+1)}{2T}$. Taking the limit as $T \to \infty$, we obtain $p_0 + \frac{\theta}{2}$, as in Proposition 14.

contract weights in the limit. Formally, this can be shown by analysis similar to that in the proof of Proposition 9 in Appendix A. One would follow the computations related to the client's expected costs in (18), while using $X_q = V_q$ for all $q \in [0, 1]$. $\square$

OA.B Affine Contracts

Our main result, Proposition 4, provides explicit formulas for the optimal contract and the dealer's trading strategy when the feasible contracts are weighted averages of the market prices. When the feasible contracts are affine functions of the market prices, we can still give explicit results, although the formulas become more complicated.

For the purposes of this appendix, a contract is a vector $(\tau_0, \tau_1, \ldots, \tau_T) \in \mathbb{R}^{T+1}$, which stipulates that the client will pay the dealer $\tau(\boldsymbol{p}) = \tau_0 + \sum_{t=1}^{T} \tau_t p_t$. For notational convenience, we write $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_T)^\top$, excluding the constant τ_0. The weighted average price contracts considered in the main text correspond to restricting to such contracts $(\tau_0, \tau_1, \ldots, \tau_T)$ with $\tau_0 = 0$ and $\sum_{t=1}^{T} \tau_t = 1$.

We provide a mathematical characterization of the optimal affine contract in Appendix OA.B.1. Then we illustrate with numerical examples and discuss the solution in Appendix OA.B.2. Of special interest is the case of a risk-neutral dealer, which we discuss in Appendix OA.B.3.

OA.B.1 Characterization of the optimal affine contract

The problem that we analyze here is the same as the formulation in Section 2.3, except that the contract space is different. The proof of the following result goes along the same lines as that of Lemma 2 and is therefore omitted.

Lemma 15. *Recall the $T \times T$ matrix F from (1) and define $T \times T$ matrices $\widetilde{A}$ and $\widetilde{E}$ by*

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

$$\widetilde{E} = \begin{pmatrix} \theta + \gamma & -\gamma - \lambda\sigma^2 & -\lambda\sigma^2 & \cdots & & \\ 0 & \theta + \gamma & -\gamma - \lambda\sigma^2 & -\lambda\sigma^2 & \cdots & \\ 0 & 0 & \theta + \gamma & -\gamma - \lambda\sigma^2 & -\lambda\sigma^2 & \cdots \\ \vdots & \vdots & & \ddots & \ddots & \\ 0 & 0 & & & \theta + \gamma & -\gamma - \lambda\sigma^2 \\ 0 & \cdots & & & 0 & 0 \end{pmatrix}.$$

For any affine contract with $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_T)^\top$, the dealer has a unique best response in $\mathcal{X}$, which is the static trading strategy $\boldsymbol{x} = F\widetilde{A}^{-1}\widetilde{E}\boldsymbol{\tau} + \lambda\sigma^2 F\widetilde{A}^{-1}\mathbb{1}$, where $\mathbb{1} = (1, 1, \ldots, 1)^\top$ denotes a T-dimensional vector of ones.

Likewise, the following result is the analogue to Proposition 4 for the space of affine contracts. While the formulas look complicated, they are fully explicit and can be computed directly.

Proposition 16. *We define*

$$
\begin{aligned}
G &= \begin{pmatrix} 1 & 1 & 1 & 1 & \cdots \\ 0 & 1 & 1 & 1 & \cdots \\ 0 & 0 & 1 & 1 & \cdots \\ \vdots & & \ddots & \ddots & \\ 0 & \cdots & & 0 & 1 \end{pmatrix}, \\
M &= \theta F^\top + \theta F + 2\gamma F^\top F + \lambda\sigma^2 F^\top G^\top G F, \\
N &= \widetilde{E}^\top \widetilde{A}^{-\top} M \widetilde{A}^{-1} \widetilde{E} + \lambda\sigma^2 G^\top G - \lambda\sigma^2 G^\top G F \widetilde{A}^{-1} \widetilde{E} - \lambda\sigma^2 \widetilde{E}^\top \widetilde{A}^{-\top} F^\top G^\top G, \\
\boldsymbol{v} &= \lambda^2\sigma^4 G^\top G F \widetilde{A}^{-1} \mathbb{1} - \lambda\sigma^2 \widetilde{E}^\top \widetilde{A}^{-\top} M \widetilde{A}^{-1} \mathbb{1}.
\end{aligned} \tag{60}
$$

The optimal affine contract and the dealer's trading strategy are given by

$$
\begin{aligned}
\boldsymbol{\tau}^* &= N^{-1}\boldsymbol{v}, \\
\boldsymbol{x}^* &= F\widetilde{A}^{-1}\widetilde{E}N^{-1}\boldsymbol{v} + \lambda\sigma^2 F\widetilde{A}^{-1}\mathbb{1}, \\
\tau_0^* &= p_0(1 - \mathbb{1}^\top\boldsymbol{\tau}^*) + (\boldsymbol{x}^* - \boldsymbol{\tau}^*)^\top(\theta F^{-1}\boldsymbol{x}^* + \gamma\boldsymbol{x}^*) + \frac{\lambda\sigma^2}{2}(\boldsymbol{x}^* - \boldsymbol{\tau}^*)^\top G^\top G(\boldsymbol{x}^* - \boldsymbol{\tau}^*).
\end{aligned} \tag{61}
$$

Under them, the client's expected costs of execution are

$$
p_0 + \frac{1}{2}\lambda^2\sigma^4 \mathbb{1}^\top \widetilde{A}^{-\top} M \widetilde{A}^{-1} \mathbb{1} - \frac{1}{2}\boldsymbol{v}^\top (N^{-1})^2 \boldsymbol{v}.
$$

The proof of Proposition 16 is deferred to Appendix OA.B.4.

OA.B.2 Illustration and discussion of the optimal affine contract

Illustration. To illustrate the solution provided by Proposition 16, Figures 7–9 display the optimal contract and dealer's trading strategy for various choices of the parameters θ, γ, and λ. For the left panels, note that τ_t^* corresponds to the weight on p_t for $t \in \{1, \dots, T\}$ and to the additive constant for $t = 0$.

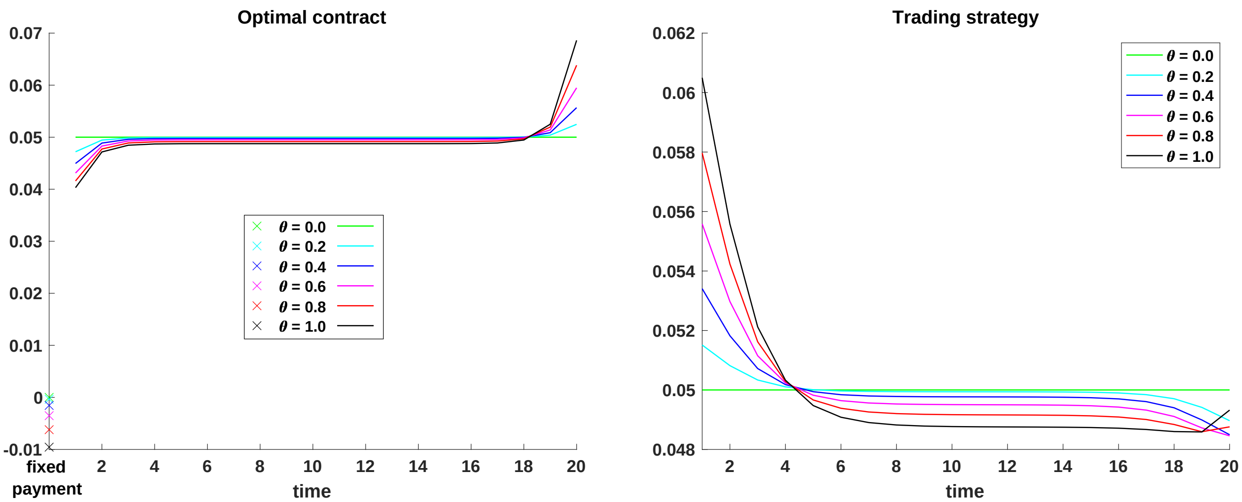

Figure 7: The optimal affine contract and trading strategy for different levels of permanent price impact. When there is no permanent price impact ($\theta = 0$), the first best is achieved. In this case, the optimal contract satisfies $\tau_j = 1/T$ for all $j = 1, \ldots, T$ and $\tau_0 = 0$, while the trading strategy is constant over time. The other parameters are $\gamma = 1$, $\lambda = 1$, $\sigma = 1$, $p_0 = 0$, and $T = 20$.

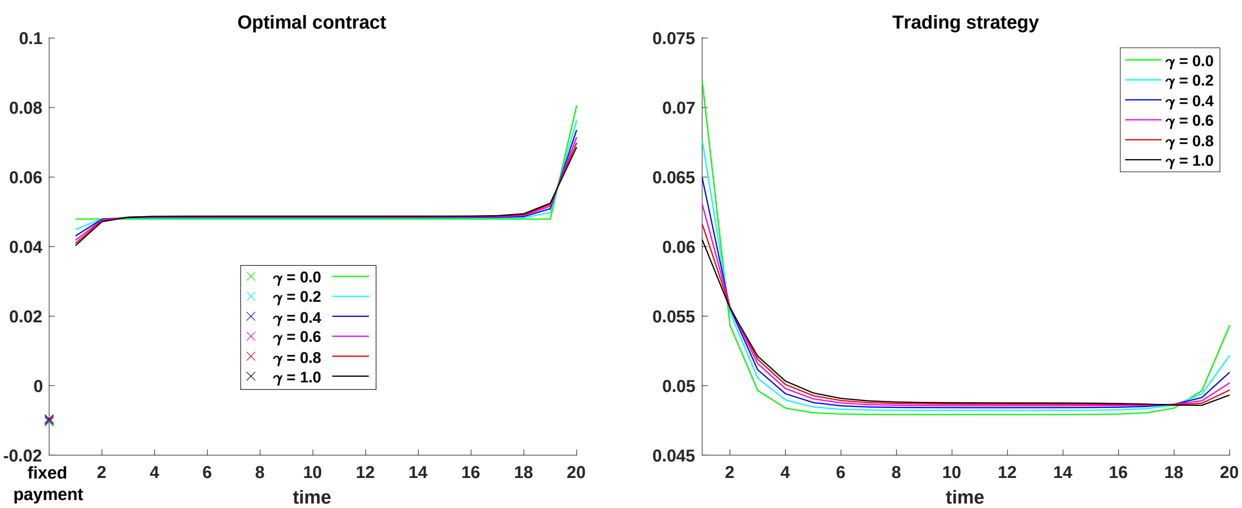

Figure 8: The optimal affine contract and trading strategy for different levels of temporary price impact. The other parameters are $\theta = 1$, $\lambda = 1$, $\sigma = 1$, $p_0 = 0$, and $T = 20$.

Comparison to the baseline analysis. Several features of our baseline analysis (with weighted-average-price contracts) extend to the case of affine contracts. One similarity is that the dealer's best response continues to reflect a frontloading motive. In the main text, this frontloading motive was formalized by Proposition 3. The availability of additional contracts is irrelevant to that result, so that the proposition *carries over unchanged*. Likewise, the right panels of Figures 7–9 suggest that the dealer's trading policy in response to the optimal contract is frontloaded (in the sense that $\sum_{s=1}^{t} x_s^* \geq \frac{t}{T}$ for all t), just as in our baseline analysis (*cf.* the right panels of Figures 1–3).

In terms of the optimal contract, the main qualitative similarity to our baseline analysis is that

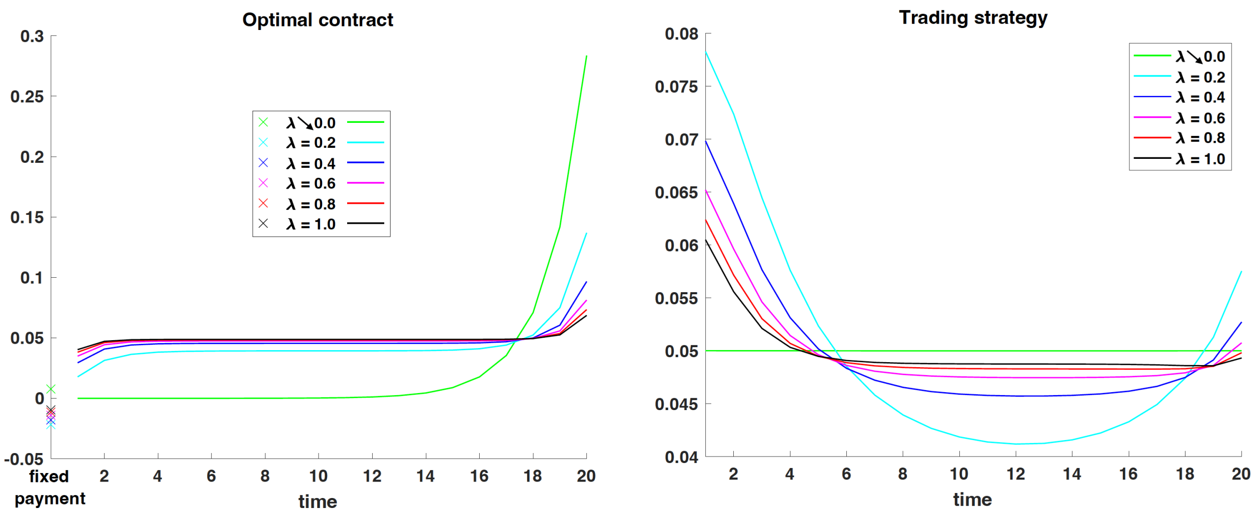

Figure 9: The optimal affine contract and trading strategy for different levels of risk aversion. The other parameters are $\theta = 1$, $\sigma = 1$, $p_0 = 0$, and $T = 20$. For $\lambda = 0$, the trading strategy is flat, coinciding with the first best. Interestingly, the shape of the trading strategy changes significantly when comparing $\lambda = 0.0$ and $\lambda = 0.2$.

there continue to be large weights on late prices, as illustrated by the left panels of Figures 7–9. The main qualitative difference—besides the obvious presence of the constant—is that there are no longer large weights on early prices. Here is the intuition for those patterns. In the baseline, the reason to weight earlier prices was because, holding prices fixed, the client can reduce her payment by shifting weight from later prices (which are high in expectation) to earlier prices (which are low in expectation). With affine contracts, the client has two additional tools for reducing her payment: (*i*) reducing the total weight put on prices (as $\sum_{t=1}^{T} \tau_t$ is no longer fixed at one), and (*ii*) reducing the fixed-payment component τ_0.As a result, she no longer has the same reason to weight early prices. However, her rationale for weighting later periods remains—and for the same reasons as discussed before.

OA.B.3 The case of a risk-neutral dealer

A notable case is that in which the dealer is risk-neutral (i.e., $\lambda = 0$), in which case the client obtains her first-best payoff. For example, she can achieve this by offering the dealer a suitable fixed-price contract (analogous to a 'sell-the-firm' contract in classical models of moral hazard), which she can do because the set of affine contracts includes all fixed-price contracts. However, such a contract is not *unique* in implementing first best; many contracts are optimal when $\lambda = 0$. In fact, we prove the following proposition in Appendix OA.B.4.

Proposition 17. *When $\lambda = 0$, an affine contract $(\tau_0, \boldsymbol{\tau})$ is optimal if and only if (i) $\boldsymbol{\tau} = c\left(\frac{\theta+\gamma}{\gamma}, \frac{(\theta+\gamma)^2}{\gamma^2}, \ldots, \frac{(\theta+\gamma)^T}{\gamma^T}\right)^\top$ for some $c \in \mathbb{R}$, and (ii) τ_0 is chosen to make the (IR) constraint bind.*

In the language of Proposition 17, the optimal fixed-price contract entails $c = 0$. As illustrated in Figure 9, the optimal contract identified by Proposition 16 entails a $c > 0$ when λ converges to zero.[36]

OA.B.4 Proofs of results in Appendix OA.B

Proof of Proposition 16. It follows from (IR) that

$$\tau_0 \geq p_0 - p_0 \sum_{t=1}^{T} \tau_t + \sum_{t=1}^{T} (X_t - X_{t-1} - \tau_t)\big(\theta X_t + \gamma(X_t - X_{t-1})\big) + \frac{\lambda\sigma^2}{2} \sum_{t=1}^{T} \left(X_{t-1} - 1 + \sum_{j=t}^{T} \tau_j \right)^2, \quad (62)$$

with equality if (IR) is binding. The client's expected cost of a contract $\tau(\boldsymbol{p})$ is

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}}[\tau(\boldsymbol{p})] &= \tau_0 + \mathbb{E}_{\boldsymbol{x}}\left[\sum_{t=1}^{T} \tau_t p_t \right] \\
&= \tau_0 + \mathbb{E}_{\boldsymbol{x}}\left[\sum_{t=1}^{T} \tau_t \left(p_0 + \sum_{j=1}^{t} (\theta x_j + \varepsilon_j) + \gamma x_t \right) \right] \\
&= \tau_0 + \sum_{t=1}^{T} \tau_t \left(p_0 + \sum_{j=1}^{t} \theta x_j + \gamma x_t \right) \\
&\geq p_0 + \sum_{t=1}^{T} (X_t - X_{t-1})\big(\theta X_t + \gamma(X_t - X_{t-1})\big) + \frac{\lambda\sigma^2}{2} \sum_{t=1}^{T} \left(X_{t-1} - 1 + \sum_{j=t}^{T} \tau_j \right)^2, \quad (63)
\end{aligned}$$

with equality if (IR) is binding. Using matrices, we can rewrite this when (IR) is binding as

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}}[\tau(\boldsymbol{p})] &= p_0 + \boldsymbol{X}^\top F^\top (\theta \boldsymbol{X} + \gamma F \boldsymbol{X}) + \frac{\lambda\sigma^2}{2} (\boldsymbol{x} - \boldsymbol{\tau})^\top G^\top G (\boldsymbol{x} - \boldsymbol{\tau}) \\
&= p_0 + \frac{1}{2} \boldsymbol{X}^\top M \boldsymbol{X} + \frac{\lambda\sigma^2}{2} \boldsymbol{\tau}^\top G^\top G \boldsymbol{\tau} - \lambda\sigma^2 \boldsymbol{\tau}^\top G^\top G \boldsymbol{x} \\
&= p_0 + \frac{1}{2} \boldsymbol{\tau}^\top \widetilde{E}^\top \widetilde{A}^{-\top} M \widetilde{A}^{-1} \widetilde{E} \boldsymbol{\tau} + \frac{\lambda\sigma^2}{2} \boldsymbol{\tau}^\top G^\top G \boldsymbol{\tau} - \lambda\sigma^2 \boldsymbol{\tau}^\top G^\top G F \widetilde{A}^{-1} \widetilde{E} \boldsymbol{\tau} - \lambda^2\sigma^4 \boldsymbol{\tau}^\top G^\top G F \widetilde{A}^{-1} \mathbb{1} \\
&\quad + \lambda\sigma^2 \mathbb{1}^\top \widetilde{A}^{-\top} M \widetilde{A}^{-1} \widetilde{E} \boldsymbol{\tau} + \frac{\lambda^2\sigma^4}{2} \mathbb{1}^\top \widetilde{A}^{-\top} M \widetilde{A}^{-1} \mathbb{1},
\end{aligned}$$

where F is defined in (1), and G and M are given in (60). We can write this as

$$\mathbb{E}_{\boldsymbol{x}}[\tau(\boldsymbol{p})] = \frac{1}{2} \boldsymbol{\tau}^\top N \boldsymbol{\tau} - \boldsymbol{v}^\top \boldsymbol{\tau} + p_0 + \frac{\lambda^2\sigma^4}{2} \mathbb{1}^\top \widetilde{A}^{-\top} M \widetilde{A}^{-1} \mathbb{1},$$

where N and $\boldsymbol{v}$ are defined in (60). From the first-order condition, the minimizer is given by $\boldsymbol{\tau}^* = N^{-1}\boldsymbol{v}$. It now follows from Lemma 15 that $\boldsymbol{x}^* = F\widetilde{A}^{-1}\widetilde{E}N^{-1}\boldsymbol{v} + \lambda\sigma^2 F\widetilde{A}^{-1}\mathbb{1}$. We deduce (61)

[36] If we plug $\lambda = 0$ in the formula of Proposition 16, it will not be well defined because $\widetilde{A}$ is not invertible. Proposition 16 gives a unique optimal contract for every $\lambda > 0$, converging to a unique limiting contract as $\lambda \searrow 0$.

from (62). For the expected costs under this contract, we write

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}[\tau(\boldsymbol{p})] &= \frac{1}{2}(\boldsymbol{\tau}^*)^\top N\boldsymbol{\tau}^* - \boldsymbol{v}^\top\boldsymbol{\tau}^* + p_0 + \frac{\lambda^2\sigma^4}{2}\mathbb{1}^\top\widetilde{A}^{-\top}M\widetilde{A}^{-1}\mathbb{1} \\ &= \frac{1}{2}(\boldsymbol{\tau}^* - N^{-1}\boldsymbol{v})^\top N(\boldsymbol{\tau}^* - N^{-1}\boldsymbol{v}) - \frac{1}{2}\boldsymbol{v}^\top(N^{-1})^2\boldsymbol{v} + p_0 + \frac{\lambda^2\sigma^4}{2}\mathbb{1}^\top\widetilde{A}^{-\top}M\widetilde{A}^{-1}\mathbb{1} \\ &= -\frac{1}{2}\boldsymbol{v}^\top(N^{-1})^2\boldsymbol{v} + p_0 + \frac{\lambda^2\sigma^4}{2}\mathbb{1}^\top\widetilde{A}^{-\top}M\widetilde{A}^{-1}\mathbb{1},\end{aligned}$$

which concludes the proof. □

Proof of Proposition 17. By Proposition 1, the expected costs of execution are $p_0 + \frac{\gamma}{T} + \frac{\theta(T+1)}{2T}$ under the first-best trading strategy $\boldsymbol{x}^{FB} = \left(\frac{1}{T}, \ldots, \frac{1}{T}\right)^\top$. We will find all $\boldsymbol{\tau}$ with

$$\mathbb{E}_{\boldsymbol{x}}[\tau(\boldsymbol{p})] = p_0 + \frac{\gamma}{T} + \frac{\theta(T+1)}{2T}.$$

To do so, we first note that the best response to all such $\boldsymbol{\tau}$ must be $\boldsymbol{x}^{FB}$. Indeed, (63) for $\lambda = 0$ gives

$$\mathbb{E}_{\boldsymbol{x}}[\tau(\boldsymbol{p})] \geq p_0 + \gamma\sum_{t=1}^{T} x_t^2 + \theta\sum_{t=1}^{T} x_t\sum_{s=1}^{t} x_s \geq p_0 + \frac{\gamma}{T} + \frac{\theta(T+1)}{2T}, \tag{64}$$

where the second inequality follows from the proof of Proposition 1. Equality holds in the first part of (64) if and only if (IR) is binding. There is equality in the second part of (64) if and only if $\boldsymbol{x} = \boldsymbol{x}^{FB}$. Next, we rewrite the best-response formula in Lemma 15 for $\lambda = 0$ as $\overline{A}F^{-1}\boldsymbol{x} = \overline{E}\boldsymbol{\tau} + (0, \ldots, 0, 1)^\top$, where

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

$$\overline{E} = \begin{pmatrix} \theta+\gamma & -\gamma & 0 & \cdots & & \\ 0 & \theta+\gamma & -\gamma & 0 & \cdots & \\ 0 & 0 & \theta+\gamma & -\gamma & 0 & \cdots \\ \vdots & \vdots & & \ddots & \ddots & \\ 0 & 0 & & & \theta+\gamma & -\gamma \\ 0 & \cdots & & & 0 & 0 \end{pmatrix}.$$

We compute $\overline{A}F^{-1}\boldsymbol{x}^{FB} = (0, \ldots, 0, 1)^\top$. Therefore, a contract is optimal if and only if (i) $\overline{E}\boldsymbol{\tau} = 0$ and (ii) τ_0 is chosen to make the (IR) constraint bind. The condition $\overline{E}\boldsymbol{\tau} = 0$ is equivalent to $(\theta+\gamma)\tau_t = \gamma\tau_{t+1}$ for all $t = 1, \ldots, T-1$, or $\boldsymbol{\tau} = c\left(\frac{\theta+\gamma}{\gamma}, \frac{(\theta+\gamma)^2}{\gamma^2}, \ldots, \frac{(\theta+\gamma)^T}{\gamma^T}\right)^\top$ for some $c \in \mathbb{R}$. □